Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 22 (2017) 430–439

Transportation Research Procedia

www.elsevier.com/locate/procedia

19th EURO Working Group on Transportation Meeting, EWGT2016, 5-7 September 2016, Istanbul, Turkey

Location analysis of emergency vehicles using an approximate queueing model

M. Altan Akdogan [a,*], Z. Pelin Bayındır [a], Cem Iyigun [a]

[a]Middle East Technical University, Dumlupınar Bulvarı No:1, Ankara 06800, Turkey

Abstract

In this study, location analysis of emergency system vehicles is discussed using an approximate queueing model (AQM) on fully connected networks. We relax single vehicle restriction in each vehicle location which is commonly employed in the literature. Order of districting structure is used to transition rates in AQM. Service rates are computed regarding the location of the demand call and the vehicle, which provides higher resolution than inter-intra district service rates. Generic rate formulations are delivered for service rate calculations. A coverage threshold is used and all location solutions for a problem instance having a higher coverage than the threshold are analyzed. Mean response time for these solutions are found using AQM and an exact simulation model. Mean absolute percent error for mean response time is reported against different order of districting levels and network specific variables such as distribution of demand over the area, traffic intensity or the number of vehicles.

Keywords: location analysis; queueing model; simulation

1. Introduction

Emergency system (ES) is a system of several components including organizations, transportation and communication networks, trained professionals, administrators aiming to work in coordination for success. Having the primary concern as emergencies, planning of this system requires significant work to ensure serving the public at its best. Other than administrative decisions, planning of physical infrastructure, primarily locating emergency vehicles constitutes a major part in the performance of the system. Various criteria could be important in deciding the locations of emergency vehicle server locations. In addition to location of the vehicles, strategic decisions such as dispatch policies, types or number of vehicles are subject to planning of these systems as well.

In our study, the problem of location analysis of ES vehicles such as fire trucks, ambulances, patrols as a server-to-customer service is discussed based on Markovian processes. The performance of the system is assessed using both Markovian analysis and discrete event simulation. The region where ES vehicles would be located is divided into demand regions and some of those regions are considered as candidate vehicle locations. For the given number of

[*] Corresponding author. Tel.: +90-312-210-4790.
E-mail address: aaltan@metu.edu.tr

2214-241X © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Scientific Committee of EWGT2016.
10.1016/j.trpro.2017.03.018

vehicles, different location solutions are analyzed in terms of mean response time to the incidents while a predefined coverage requirement is satisfied. A demand region is considered covered if it is closer to a vehicle location than a threshold travel time. In assessing the performance of a location solution, it is required that frequency of total demand generated from covered regions is greater than a threshold coverage ratio. In the problem, demand is assumed to be independent, non-identical and it follows a time homogeneous Poisson process. Service time to an incident is assumed exponentially distributed and its mean is known.

2. Literature Review and Motivation of the Study

Deterministic and probabilistic studies are reported on both location analysis of ES in the literature. Hakimi (1964) first proposed p-Median model for the locations of switching centers in telephone networks as a deterministic model. Berman et al. (2003) used coverage decay function in a generalization of maximal coverage and median-based models. Kunkel et al. (2013) used weighted p-median problem in assigning medical assistants to population centers and then Capacitated Facility Location Problem (CFLP) was solved to assign these medical assistants to resupply centers.

In probabilistic stream, Daskin (1983) introduced Maximum Expected Covering Problem for location analysis of public service facilities. Maximum Availability Location Problem(MALP) was developed by Revelle and Hogan (1989) incorporating vehicle availability into the location problem. Beraldi and Bruni (2009) located ambulances to cover demand within a specified reliance by using Two Stage Stochastic Programming.

Regarding the uncertainties inherit to these systems, queueing theory is utilized to analyze the performance of these systems. Hypercube Queueing Model (HQM) by Larson (1974) is the first model using queueing theory to achieve performance measures which are significant to the location decision of ES vehicles. With the help of HQM, an existing system can be analyzed regarding various performance measures such as utilization of vehicles, mean response time and probability of loss. Iannoni and Morabito (2007), Takeda et al. (2007) used HQM as a base to measure performance of EMS systems. In these studies, service time variations resulting from the variations in the travel times were taken of second order and service rates were defined specific to the servers not to the call itself. Morabito et al. (2008) used HQM by defining non-homogeneous service rates for servers and compares results with homogeneous service rate assumption. Iannoni et al. (2008) used a genetic algorithm to find the optimal location for EMS servers allowing only one server in a single location while using service rates specific to servers. Iannoni et al. (2009), Iannoni et al. (2011) used HQM in an optimization environment for location and districting decision of EMS servers on high ways with alternative objectives.

As an extension of HQM, Spatial Queueing Model (SQM) by Geroliminis et al. (2009) was proposed for spatial networks. SQM is introduced as an optimization problem where mean response time of the system is minimized. In a different study, Geroliminis et al. (2011) worked on a large scale system to deploy emergency response mobile units with server specific service rates and single vehicle restriction for each location. Boyaci and Geroliminis (2015) proposed approximation algorithms for large scale networks with spatially distributed demand. In this study, multiple vehicles were allowed and service rates for servers were differentiated regarding inter or intra-district service. They proposes Aggregate Hypercube Queueing Model that is used with a partitioning algorithm to find the optimal server locations.

We base our study on SQM proposed by Geroliminis et al. (2009). SQM is not an exact system. Mean response time is calculated by constructing a queueing system for a location solution. Only busy-available information of the vehicles defines the states of the queueing system. No prior information is considered about which region a busy vehicle is serving. Hereby, districting is used to approximate the transition rates in the queueing system. Although the order of districting is not limited, it is studied as three in the literature by Geroliminis et al. (2009). With this, a demand call would be served by at most the third nearest vehicle to this region. After generating the transition rates, fractions of dispatch of vehicles to regions are found with an approximate formulation from steady state probabilities by solving the balance equations. Then, these fractions are used to calculate the mean response time in the system. In SQM, at most one vehicle is allowed in a candidate location. This restriction is appropriate when one considers the distribution of demand over a spatial network. Service rate is not different for every demand region but it is differentiated for sets of demand regions emerging in the calculation of arrival rates of the queueing system. However, in the study no closed form expression is delivered for calculation of these rates.

We propose a more general model to represent emergency systems with queueing models by relaxing the restriction of single vehicle in a location, searching for the effect of order of districting. Differently from the literature, we define service rates specific to the location of the demand call and the location of the server and defining generic formulations for service rate calculations in the queueing system. Therefore, we offer more resolution for the service rates for demand calls than inter-intra district service rates.

We propose Approximate Queueing Model (AQM) where multiple vehicles in a candidate location are allowed based on the structure of SQM. Demand regions can be defined both on Euclidean networks and on spatial networks. Accordingly, the proposed model is not necessarily a hypercube queueing model. The state of the queueing system is defined as the number of busy vehicles in a location. Then, a state is an n-dimensional array where all n vehicles are located in different locations, resulting in a state space of unit hypercube. If different number of vehicles are located in the locations, then state space cannot be defined as a hypercube. AQM is called to be an approximate queueing model since the location of a busy vehicle to which it is serving is not tracked in the state definition. Therefore, in the calculation of transition rates of the queueing system, districting is used similar to SQM. Different from SQM, service times are differentiated regarding the travel times to the location of each incident.Service time for an incident is composed of the travel time from the vehicle location to the demand region, the time to clear the incident and the travel time back to the vehicle location. In our study, it is assumed that the time for clearing the incident and the travel time between regions follow exponential distribution with known rates. Although the sum of these three random variables does not follow exponential distribution, it is assumed so with rate equal to the sum of rates of three components in order to utilize Markovian property. Four explicit alternative formulations are proposed to calculate the service rates in AQM. The order of districting is also under question in terms of its effect on the mean response time. In a spatial network, 3^{rd} order of districting could be justified however on a Euclidean network, assuming at most the third nearest vehicle would serve the incident, could not be justified regarding the traffic intensity or distribution of demand regions over the area. Therefore, the mean time of the location solutions is assessed in three different settings as 3^{rd}, 4^{th} and 5^{th} order of districting. All possible location solutions satisfying the coverage requirement are analyzed in terms of mean response time under four different service rate formulations and three levels of order of districting, resulting 12 different mean response time for every solution.

Organization of the paper is as follows: environment for our study is defined in Section 3. In Section 4, Approximate Queueing Model is defined in detail and experimental results is given in Section 5. Lastly, we conclude in Section 6.

3. The Environment

The entire region, where vehicles are to be located, is divided into demand regions. Some of these demand regions are listed as candidate vehicle locations. Number of vehicles to be located is predetermined.

Demand is defined as the call for an emergency service and occurs randomly at the center of each demand region. It is assumed that the demand in the regions are independent and non identical and follows a time homogeneous Poisson process. The mean number of calls for a vehicle in a unit time is known for each demand region.

For each region, a list of vehicle locations sorted with respect to the proximity is available. When a call is received from a certain region, vehicle locations are checked for availability in the order of proximity to demand. If there is an available vehicle, the service starts. Otherwise, the demand is lost.

Travel times between the regions are assumed exponentially distributed with known mean.

A service is composed of travel time to the demand region from the vehicle location, service time for the demand and travel time back to the vehicle location. Total service time is random, dependent on the location of both vehicle and the demand, and is assumed to be exponentially distributed with mean as equal to the sum of mean of travel time to demand location, mean of service time and mean of travel time back to vehicle location. Since the queueing model is not exact, order of districting is used in the queueing system. For every transition, a set of demand regions which can cause to this transition in the queueing system is defined as an area and service rate for this area is calculated.

The fraction of total demand generated from a demand region is calculated from the demand rates. All solutions having a coverage of larger than a coverage threshold are analyzed in this study. A demand region is said to be covered if it is closer to a vehicle location than a given travel time threshold. Coverage of a solution is found by summing fractions of demand of such regions.

The quality of the service is defined as mean response time. The performance measure is selected as mean response time which is one of the most important performance measures for emergency systems.

Notation used is as follows: Q is the set of demand regions, R is the set of candidate vehicle locations, N is the number of vehicles to be located t_{qr} defines mean travel time between demand region q and vehicle location r in minutes. ω_q is demand rate (number of demand calls per hour) for demand region $q \in Q$ and f_q is the frequency of demand generated from demand region $q \in Q$ and it is equal to $\frac{\omega_q}{\sum_{q \in Q} \omega_q}$. T stands for the travel time threshold for coverage and α for the required minimum coverage.

4. Approximate Queueing Model

An approximate queuing model is generated to obtain the performance measure. Notation and definitions in addition to Section 3 are given in Table 1. In the rest of the study, $\bar{x}$ is dropped from the notation for the sake of brevity.

Table 1. Notation used for Mathematical Formulation

Notation	Definition
x_r	The number of vehicles located in location $r \in R$
$\bar{x}$	Vector of x_r in a location solution
$B(\bar{x})$	State space of the queuing system generated under $\bar{x}$
$B_i(\bar{x})$	i^{th} member of state space $B(\bar{x})$
$E_{nq}(\bar{x})$	Set of states which vehicle location n has the nearest available vehicle for region q under $\bar{x}$, $1 \leq n \leq \lvert B_i(\bar{x}) \rvert$
$r_n(\bar{x})$	Vehicle location corresponding to the n^{th} entry of the state $B_i(\bar{x}) : r_n(\bar{x}) \in R$ and $1 \leq n \leq \lvert B_i(\bar{x}) \rvert$
$\lambda_{ij}(\bar{x})$	Upward transition rate with $d^+_{ij} = 1$ from state i to j under $\bar{x}$
$\mu_{ij}(\bar{x})$	Downward transition rate with $d^-_{ji} = 1$ from state j to i under $\bar{x}$
$P(B_i(\bar{x}))$	Steady-state probability of state $B_i(\bar{x})$
$\rho_{r_n(\bar{x})q}$	Fraction of dispatch of vehicles in location $r_n(\bar{x})$ to region q

To represent a solution to the problem in exact queueing system, we have to track where a vehicle is located in, if it is busy or not and which region it serves if it is busy. Then, the exact queueing system can be represented by an N-dimensional state as follows: $B_i = (b_1, b_2, ..., b_N)$, $b_i = \{s, r\}$, $i = 1, ..., N$ where $s \in (0 \bigcup Q)$, r states in which location the vehicle represented in i^{th} entry of the state definition located. $b_i = \{0, r\}$ represents the state that the vehicle is free and $b_i = \{q, r\}$ represents the state where the vehicle is busy serving demand region $q \in Q$. This would result in a state space with $(|Q| + 1)^N$ states and its size increases exponentially with increasing number of vehicles.

In the approximate queueing model, system can be modeled by an n-dimensional state indicating the number of busy vehicles on each location selected in the solution. $B_i = (b_1, b_2, ..., b_k, ..., b_n)$ where $1 \leq n \leq |R|$ is the number of locations selected in the solution in concern and b_k is the number of busy vehicles at location $r_k \in R$ in state i

The approximate queueing system, generate a state space with $\prod_{r \in V}(x_r+1)$ states where $V = \{r \in R, x_r \in \bar{x} : x_r > 0\}$. The cardinality of the state space of the approximate system is bounded by 2^N. The maximum size of the state space of solutions to the problem increases exponentially but slower than the maximum size of the exact queueing system with increasing number of vehicles.

For clarification of the approximate identification of the queueing system worked with, it is, explicitly, resulted from not tracking which region a busy vehicle is serving in the state definition. We need to track where an busy vehicle is serving as in the exact queueing system to find exact mean response time but this increases state space and make it computationally inefficient to solve. For this reason, we use an approximate queueing system.

In this system, only one step transition between states is allowed, meaning only one vehicle can become idle or busy in a transition. Hamming distance d_{ij} between state i and j is used to express this behavior in the model. Hamming distance is the number of digits between two states, B_i and B_j , of the system that is different from each other. For states (2,1,0,3,2,0) and (2,1,1,3,2,0), Hamming distance is equal to 1. Therefore, only transition with Hamming distance equal to 1 is allowed.

A transition is classified as upward or downward regarding the characteristic of Hamming distance between states. Upward and downward Hamming distances are introduced as d^+_{ij} and d^-_{ij} by Larson (1974) where d^+_{ij} represents the number of digits with increase in entries in the transition from state i to j while the latter indicates the number of digits with decrease in the entries in the transition from state i to j. Allowing only transition with Hamming distance of one results in a system having upward and downward Hamming distances equal to again only one between transitions.

By use of the definitions, upward transition refers to transitions with upward Hamming distance $d^+_{ij} = 1$ while downward to ones with $d^-_{ij} = 1$. Followingly, rates for the queueing system are identified as upward transition rate and downward transition rate.

λ_{ij} represents the rate of demand call resulting in a transition from state i to j while μ_{ij} is the service rate for this demand call, namely the downward transition rate from state j to i where $B_i = (b_1, ..., 0_k, ..., b_{|B_s|-1}, b_{|B_s|})$ and $B_j = (b_1, ..., 1_k, ..., b_{|B_s|-1}, b_{|B_s|})$.

Under this setting, balance equations for the queuing system is as in (1) and (2).

$$P(B_j) * \left[\sum_{B_i \in B: d^-_{ij}=1} \lambda_{ij} + \sum_{B_i \in B: d^+_{ij}=1} \mu_{ij} \right] = \sum_{B_i \in B: d^-_{ij}=1} \mu_{ij} * P(B_i) + \sum_{B_i \in B: d^+_{ij}=1} \lambda_{ij} * P(B_i), \quad j = 0, 1, ..., |B| - 1. \tag{1}$$

$$\sum_{i=0}^{|B|-1} P(B_i) = 1. \tag{2}$$

Followingly, fraction of dispatches from vehicle location r_n to region q can be stated as:

$$\rho_{r_n q} = f_q \frac{\sum_{B_i \in E_{nq}} P(B_i)}{(1 - P(B_{|B|-1}))}, \quad \forall n : 1 \leq n \leq |B_i| \text{ and } \forall q \in Q. \tag{3}$$

Numerator in (3) is the sum of steady state probabilities of states where n is the nearest available server for demand region q while denominator is the fraction of total demand that is met. The result of the division multiplied by fraction of demand originated from demand region q gives the fraction of dispatch of vehicles from vehicle location r_n to demand region q. Then, mean response time of the system ($\bar{T}$) is equal to $\sum_{n=1}^{|B_i|} \sum_{q \in Q} \left(\rho_{r_n q} t_{r_n q} \right)$.

4.1. Order of districting

In SQM, order of districting is introduced to estimate the upward and downward transition rates. n^{th} order of districting indicates that demand in every region is satisfied by at most n^{th} closest vehicle location. It ignores the possibility that when a demand occurs all n locations' vehicles are busy.

According to order of districting structure, sub areas are introduced for transition rate calculations. The notation used is as follows, O stands for the maximum level of order of districting, D^n_{kl} states the set of regions belonging to the 1^{st} order of district of server k and the n^{th} order of district of server l where $n = 1...O$. U_q defines the ordered set of vehicle locations selected in the solution with respect to mean travel times to demand region q in ascending order.

Regarding the notation, sub areas are defined as $D^n_{kl} = \left\{ q \in Q : U_q(1) = r_k \text{ and } U_q(n) = r_l \right\}$.

By definition, D^n_{kl} is the sub area consisting of the set of regions to which k is the nearest vehicle location and l is the n^{th} nearest vehicle location, and at both of which at least one vehicle is located.

4.2. Calculation of rates in queueing system

Sub areas defined in Section 4.1 are used in the calculation of upward and downward transition rates since the exact region to which a vehicle is sent for a demand call is not tracked. Added notation for this calculations is as follows: L^n_{kl} is the set of regions in the sub area D^n_{kl}. Ω^n_{kl} is the total demand of the regions in the sub area D^n_{kl} and equal to $\sum_{q \in L^n_{kl}} \omega_q$. Λ is the total demand of the system and equal to $\sum_{q \in Q} \omega_q$. For every order of districting, it is obvious that demand will be completely covered:

$$\sum_{k \in |B_i|} \sum_{l \in |B_i|} \Omega^n_{kl} = \Lambda, \forall n = 1, 2, ..., min(O, |B_i|). \tag{4}$$

Upward transition rate λ_{ij} for a transition from $B_i = (b_1, ..., 0_k, ..., b_{|B_i|-1}, b_{|B_i|})$ to $B_j = (b_1, ..., 1_k, ..., b_{|B_i|-1}, b_{|B_i|})$ is calculated as follows:

$$\lambda_{ij} \quad = \quad \Omega^1_{kk} \quad + \sum_{1\le l_1 \le |B_i|:b_{l_1}=x_{r_{l_1}}} \Omega^2_{l_1 k} \quad + \sum_{m=3}^{M} \quad \sum_{1\le l_1,\ldots,l_{m-1}\le |B_i|:b_{l_i}=x_{r_{l_i}}} \Omega^m_{l_1 k} \bigcap \Omega^{m-1}_{l_1 l_{m-1}} \bigcap \ldots \bigcap \Omega^2_{l_1 l_2} \tag{5}$$

where $M = \min\{A, O\}$, $A = \left|\left\{b_s \in B_j : 1 \le s \le |B_i| \ and \ b_s = x_{r_s}\right\}\right|$ as the number of locations with all of their servers busy in state j and $\Omega^m_{lk} \bigcap \Omega^{m'}_{l'k'}$ stands for the sum of the demand of the regions in the intersection; $D^m_{lk} \bigcap D^{m'}_{l'k'}$ with the same indices.

(5) declares that in the transition from state i to j, an available vehicle in the location indexed k will respond to the demand call if the call is originated from D^1_{kk} or; if it is from sub area D^m_{lk} and from first to $(m-1)^{th}$ nearest vehicles are busy.

As an example, for states $B_i = (2, 1, 0, 3, 2, 0)$, $B_j = (2, 1, 1, 3, 2, 0)$, assume that the number of vehicles to be located is $N = 11$, the solution at hand is $\bar{x} = (2, 1, 2, 3, 2, 1, 0, 0, 0, 0)$, $R = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and $O = 4$. Then, we have $l_i = \{1, 2, 4, 5 : b_{l_i} = x_{l_i}\}$ to be used in (5), $k = 3$ and $M = 4$. Accordingly, λ_{ij} is calculated as

$$\lambda_{ij} = \tau_1 + \tau_2 + \tau_3 + \tau_4 \tag{6}$$

where

$$\tau_1 = \ \Omega^1_{33} \tag{7}$$

$$\tau_2 = \ \Omega^2_{13} + \Omega^2_{23} + \Omega^2_{43} + \Omega^2_{53} \tag{8}$$

$$\begin{aligned}\tau_3 = {} & \Omega^3_{13} \bigcap \Omega^2_{12} + \Omega^3_{13} \bigcap \Omega^2_{14} + \Omega^3_{13} \bigcap \Omega^2_{15} + \Omega^3_{23} \bigcap \Omega^2_{21} + \Omega^3_{23} \bigcap \Omega^2_{24} + \Omega^3_{23} \bigcap \Omega^2_{25} \\ & + \Omega^3_{43} \bigcap \Omega^2_{41} + \Omega^3_{43} \bigcap \Omega^2_{42} + \Omega^3_{43} \bigcap \Omega^2_{45} + \Omega^3_{53} \bigcap \Omega^2_{51} + \Omega^3_{53} \bigcap \Omega^2_{52} + \Omega^3_{53} \bigcap \Omega^2_{54}\end{aligned} \tag{9}$$

$$\begin{aligned}\tau_4 = {} & +\Omega^4_{13} \bigcap \Omega^3_{12} \bigcap \Omega^2_{14} + \Omega^4_{13} \bigcap \Omega^3_{12} \bigcap \Omega^2_{15} + \Omega^4_{13} \bigcap \Omega^3_{14} \bigcap \Omega^2_{12} + \Omega^4_{13} \bigcap \Omega^3_{14} \bigcap \Omega^2_{15} \\ & +\Omega^4_{13} \bigcap \Omega^3_{15} \bigcap \Omega^2_{12} + \Omega^4_{13} \bigcap \Omega^3_{15} \bigcap \Omega^2_{14} + \Omega^4_{23} \bigcap \Omega^3_{21} \bigcap \Omega^2_{24} + \Omega^4_{23} \bigcap \Omega^3_{21} \bigcap \Omega^2_{25} \\ & +\Omega^4_{23} \bigcap \Omega^3_{24} \bigcap \Omega^2_{21} + \Omega^4_{23} \bigcap \Omega^3_{24} \bigcap \Omega^2_{25} + \Omega^4_{23} \bigcap \Omega^3_{25} \bigcap \Omega^2_{21} + \Omega^4_{23} \bigcap \Omega^3_{25} \bigcap \Omega^2_{24} \\ & +\Omega^4_{43} \bigcap \Omega^3_{41} \bigcap \Omega^2_{42} + \Omega^4_{43} \bigcap \Omega^3_{41} \bigcap \Omega^2_{45} + \Omega^4_{43} \bigcap \Omega^3_{42} \bigcap \Omega^2_{41} + \Omega^4_{43} \bigcap \Omega^3_{42} \bigcap \Omega^2_{45} \\ & +\Omega^4_{43} \bigcap \Omega^3_{45} \bigcap \Omega^2_{41} + \Omega^4_{43} \bigcap \Omega^3_{45} \bigcap \Omega^2_{42} + \Omega^4_{53} \bigcap \Omega^3_{51} \bigcap \Omega^2_{52} + \Omega^4_{53} \bigcap \Omega^3_{51} \bigcap \Omega^2_{54} \\ & +\Omega^4_{53} \bigcap \Omega^3_{52} \bigcap \Omega^2_{51} + \Omega^4_{53} \bigcap \Omega^3_{52} \bigcap \Omega^2_{54} + \Omega^4_{53} \bigcap \Omega^3_{54} \bigcap \Omega^2_{51} + \Omega^4_{53} \bigcap \Omega^3_{54} \bigcap \Omega^2_{52}.\end{aligned} \tag{10}$$

τ_2 is the term representing the explicit form of the summation in the second term of λ_{ij} in (5), τ_3 for the third summation with $m = 3$ and τ_4 with $m = 4$. Notice that τ_1 only includes demand coming from the regions where 3^{rd} server is nearest server, τ_2 from regions where 3^{rd} is the second nearest server and 1^{st} is the nearest, 3^{rd} is the second nearest and 2^{nd} is the nearest, 3^{rd} is the second nearest and 4^{th} is the nearest, and finally 3^{rd} is the second nearest and 5^{th} is the nearest.

Downward transition rate is calculated regarding the upward rate λ_{ij}. In the formulation of upward rate, sub areas in which a demand call could result in the transition from state i to j are taken into account. Then, in the downward rate, these sub areas should be included into the formulation of μ_{ij} since we need to work with the service rates of these sub areas. Four different downward transition rate formulations are proposed as follows:

Alternative I : Notation added for Alternative I is as follows: P is the set of order of terms in (6) in explicit form, Ω_p is the p_{th} term in (6) in explicit form and Φ_p is the sum of service rates of the regions in the sub area corresponding to the p_{th} term in (6) in explicit form.

With respect to the notation given, μ_{ij} is calculated as follows for downward transition from state j to i,

$$\mu_{ij} = \frac{\lambda_{ij}}{\sum_{p\in P} \frac{\Omega_p}{\Phi_p}}. \tag{11}$$

For the example of which λ_{ij} is given in (6), term $\frac{\Omega_p}{\Phi_p}$ in (11) is; $\frac{\Omega^1_{33}}{\sum_{m\in L^1_{33}}(\phi_m)}$ for $p = 1$, $\frac{\Omega^2_{13}}{\sum_{m\in L^2_{13}}(\phi_m)}$ for $p = 2$ and $\frac{\Omega^3_{13}\bigcap\Omega^2_{12}}{\sum_{m\in L^3_{13}\bigcap L^2_{12}}(\phi_m)}$ for $p = 6$ and likewise for $p \in P$.

The denominator in (11) denotes the traffic intensity of the system consisting of location indexed with k in state definition and sub areas in (6). It should also be stated that denominator shows the sum of the traffic intensity of the sub areas in a similar approach.

In this alternative, service rates of the regions are directly used in the calculations without considering the travel time to the demand region and travel time for going back to the server location. It results in a higher downward transition rate. This alternative leads to a higher rate as demand call is served by many vehicles simultaneously due to the sum of service rates in denominator.

Alternative II : In the first formulation of μ_{ij} given in (11), service rates are used directly without any consideration of travel time to the demand region. If there is no vehicle located in the region, it is obvious that service rate occurring by serving this region from another vehicle location would be different than serving with a vehicle from inside the region. Therefore, it would be meaningful to process service rates with respect to the location of the vehicle that will serve the demand call for that transition. Differently from the first formulation, traffic intensity is not used. Service rates of the regions in set L_{ij} are recalculated regarding the location of vehicle serving in that transition and summed to calculate the downward transition rate. The notation in the formulation is as follows: L_{ij} is set of regions included in the sub areas resulting of transition from state i to j and $\phi^{'}_{qr}$ is service rate per hour for demand region $q \in Q$ when it is served from vehicle location $r \in R$.

For the transition from $B_i(b_1, ..., 0_k, ..., b_{|B_i|-1}, b_{|B_i|})$ to $B_j(b_1, ..., 1_k, ..., b_{|B_i|-1}, b_{|B_i|})$, we have λ_{ij} as in (6) and L_{ij} as the set of regions in the corresponding sub areas for transition from state i to j. The location of vehicle serving the call is the location corresponding to the k^{th} entry of the state space B_i. Then $r_k \in R$ is the corresponding vehicle location for k^{th} entry in the state space representation as defined in Table 1 Then, for this transition, service rates ($\phi^{'}_{lr_k}$) of regions that could be served in this transition can be calculated as $\phi^{'}_{lr_k} = \frac{60}{\frac{60}{\phi_{lr_k}}+2*t_{lr_k}}$, $\forall l \in L_{ij}$ where ϕ_{lr_k} is defined as per hour and t_{lr_k} is given in minutes, and $\mu_{ij} = \sum_{l\in L_{ij}} \phi^{'}_{lr_k}$.

Alternative III : This alternative is constructed upon Alternative II. Apart from considering travel time to demand region, frequency of demand generated from regions in set L_{ij} is considered.

For a transition, it is obvious that only one vehicle will serve only one of the regions in set L_{ij}. Summing up service rates of the regions (as traffic intensities of the sub areas in Alternative I) is considered resulting in a downward transition rate such that $|L_{ij}|$ (P in Alternative I) number of vehicles are serving simultaneously in this transition which amplifies the real service rate.

Based on this consideration, downward transition rate is formulated as $\mu_{ij} = \sum_{l\in L_{ij}}\left(\frac{f_l}{\sum_{m\in L_{ij}} f_m}\right) * \phi^{'}_{lr_k}$ which is the weighted average service rates ($\phi^{'}_{lr_k}$) of the regions with respect to demand frequencies of these regions in Alternative II.

Alternative IV : This alternative is based on (11) and the traffic intensity approach. Differently, processed service rates and frequencies are included. Demand frequencies is calculated for every term in (6) regarding the total demand of the sub area corresponding to each term. The notation introduced is as follows: L_{ijp}, $p \in P$, is set of regions included in the sub area corresponding p^{th} term of (6) in explicit form, resulting from transition from state i to j. F_p, $p \in P$, is weighted demand frequency of the sub area corresponding to p^{th} term of (6) in explicit form and $\Phi^{'}_p$ is sum of service rates ($\phi^{'}_{ln^{'}_k}$) of the regions in the sub area corresponding to the p_{th} term of (6) in explicit form.

Following the notation, F_p is calculated as $F_p = \frac{\sum_{k\in L_{ijp}} f_k}{\sum_{s\in P}\sum_{k\in L_{ijs}} f_k}$.

F_p , are used inversely proportional to traffic intensities of the corresponding terms. Accordingly, Alternative IV for downward transition rate is formulated as $\mu_{ij} = \frac{\lambda_{ij}}{\sum_{p\in P}\frac{\Omega_p}{\Phi^{'}_p * F_p}}$.

5. COMPUTATIONAL STUDY

In this section, generic service rate formulations proposed in Section 4.2 for AQM and effect of order of districting levels are tested using a simulation study. Experimental results are reported.

5.1. Problem instances and discrete event simulation

For the experimental study, problem instances are generated with considering different specifications. Two different distribution of demand regions (DoD) are defined over the plane as uniform and circular. In uniform DoD instances, demand regions are uniformly distributed. In circular DoD instances, demand regions are distributed around a center such that 40 percent of the regions is placed in a inner circle, 30 percent in the middle and rest in the outer circle. Representative instances for the two case with 50 demand regions can be seen in Figure 1.

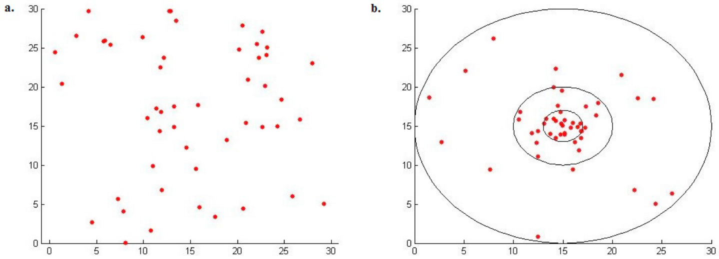

Fig. 1. (a) Uniform DoD with 50 demand regions; (b) Circular DoD with 50 demand regions

Demand of the regions is assumed the same with 1 unit per hour to observe demand pooling due to multiple vehicle relaxation. Service rates for the regions is assumed equal. Total demand is divided into traffic intensity(TI). This division is again divided to the number of vehicles and is written as the service rate for all regions. Traffic intensity is set to 0.4 and 0.8 for problem instances. For all the problem instances, minimum required coverage, α, is taken as 0.90 and travel time threshold, T, as 10 minutes. Only location solutions having a coverage of greater than or equal to 0.90 are analyzed.

To asses the performance of AQM, it is necessary to observe the mean response time of the exact system for the solution. A simulation model is constructed and coded in Matlab environment to simulate the exact queueing system.

In the simulation model, demand call arrives to the system and it is served from the nearest available server. If there is no available server, demand is considered lost. Service time for a call from region q and a responding vehicle from location r is assumed the sum of three exponentially distributed random variables as service time for this region (v_q), travel time from vehicle location to region (t_{rq}) and travel time from region to vehicle location back (t_{rq}). Simulation is ended when steady state is reached. Steady state is searched for the mean response time in ($\bar{T}$) which is calculated from the performance measures at the end of each iteration of the simulation.

Confidence interval (CI) for assessing steady state behavior of objective function value is constructed with a number of consecutive, non-overlapping batches taken from a single run after a warm-up period by referring the study of Steiger et al. (2005). These bathes are treated as independent runs and used to construct a confidence interval.

In our study, reaching the steady state is decided by comparing mean of first batch against the CI constructed. If the mean of first batch falls in CI, the system is assumed to be at steady state. Mean of CI is reported as the objective function value of the solution. If it is not in the interval, simulation is continued for another one-batch-long-many demand calls. New CI is constructed in a rolling horizon by discarding the first batch and adding the period of last one-batch-long-many demand calls as the new batch to end. Again, mean of new first (old second) batch is checked against CI. Simulation continues till having the first batch in CI constructed recently. This procedure is defined to guarantee that the warm up period is over and there is no trend in the performance measures over iterations.

In the experiments, warm up period is determined as 30000 demand calls. Number of batches is set to 10 with 5000 demand calls in each of them.

5.2. Experimental results

Two different traffic intensity as 0.4 and 0.8 and number of vehicles to be located from 1 to 7 are used for both circular and uniform distribution of demand regions on the plane, and in total 28 different problem settings are formed. Each setting has 10 demand regions where all of them are defined as candidate vehicle locations. From each setting, 5 instances are generated. Each instance is solved by setting number of vehicles from 1 to 7, making 980 problems in total. Any solution satisfying minimum coverage of 0.90 for a problem is solved with AQM and simulation model which means number of alternative solutions evaluated is depended both on the number of vehicles and also on the specific problem regarding coverage criterion. Mean response time of solutions are compared and mean absolute percent errors(MAPE) are reported for different settings. Results are given in Table 2. In this table, NoV stands for the number of vehicles, TI denotes the traffic intensity and O is the order of districting level. Instances with up to 2 vehicles are discarded since most of them do not have any solution satisfying coverage criterion.

Table 2. MAPE for comparison of mean response time of each alternative location solution for a problem instance

Alternative	O	DoD		NoV					TI	
		Uni	Circ	3	4	5	6	7	0.4	0.8
I	3	0.40	0.32	0.30	0.38	0.38	0.37	0.35	0.39	0.32
	4	0.39	0.31	0.30	0.38	0.38	0.37	0.34	0.39	0.32
	5	0.39	0.31	0.30	0.38	0.38	0.37	0.34	0.39	0.32
II	3	0.41	0.34	0.27	0.37	0.40	0.41	0.41	0.38	0.36
	4	0.41	0.34	0.27	0.37	0.40	0.41	0.41	0.38	0.36
	5	0.41	0.34	0.27	0.37	0.40	0.41	0.41	0.38	0.36
III	3	0.14	0.10	0.11	0.13	0.12	0.12	0.11	0.15	0.09
	4	0.14	0.09	0.11	0.13	0.13	0.11	0.09	0.15	0.09
	5	0.14	0.09	0.11	0.13	0.13	0.11	0.09	0.15	0.09
IV	3	0.15	0.19	0.21	0.16	0.16	0.14	0.14	0.18	0.16
	4	0.17	0.21	0.21	0.17	0.17	0.16	0.20	0.21	0.17
	5	0.17	0.18	0.21	0.17	0.17	0.13	0.15	0.20	0.14

According to these results, service rate formulations I and II perform very poor performance in approximating the exact queueing system. Alternative III and IV result in fairly good approximations than I and II. Among different network specifications, Alternative III performs better than Alternative IV in instances with circular demand regions and high traffic intensity. In instances with higher number of vehicles, increasing order of districting seems not to have significant effect on the approximations in the comparison of all solutions. Better performance in circular instances can be due to decreasing distances between regions which leads to decreasing difference in mean response time.

For each problem, the minimum response time is also studied as a performance measure for comparing the alternative approximations in AQM and the simulation model. First, the solution with the minimum mean response time with AQM is found. Then, mean response time of this solution in simulation model is calculated. MAPE are reported in Table 3 for different settings. According to the results of this comparison, service rate formulation III and IV still perform better. For the instances with higher number of vehicles, MAPE for minimum response time seems to get worse than overall MAPE in Table 2. When the traffic intensity is high, MAPE for minimum response time is close to overall MAPE and with increasing order of districting Alternative III performs better in approaching to the solution with minimum response time. In these results, it is also seen that increasing order of districting decreases MAPE in most of the problem settings when Alternative III is used. For different environments, it can be possible to use different formulations and order of districting levels to approximate the exact queueing system.

6. Conclusion

In this study, an approximate queueing model is proposed to analyze performance of ES vehicle locations. AQM is able to model any complete network without differentiating the type of the network. The quality of approximations are reported for different levels of order of districting and different service rate formulations. The effect of order of districting is analyzed in different network settings. It is showed that order of districting affects approximations

Table 3. MAPE for comparison of minimum mean response time that can be obtained for each problem instance

Alternative	O	DoD		NoV					TI	
		Uni	Circ	3	4	5	6	7	0.4	0.8
I	3	0.42	0.37	0.30	0.39	0.42	0.43	0.44	0.48	0.31
	4	0.41	0.36	0.30	0.39	0.42	0.41	0.40	0.47	0.30
	5	0.41	0.36	0.30	0.39	0.42	0.41	0.39	0.47	0.29
II	3	0.45	0.44	0.26	0.39	0.49	0.53	0.56	0.50	0.39
	4	0.45	0.44	0.26	0.39	0.48	0.52	0.56	0.49	0.39
	5	0.45	0.44	0.26	0.39	0.48	0.52	0.56	0.49	0.39
III	3	0.17	0.15	0.09	0.08	0.15	0.22	0.28	0.21	0.11
	4	0.14	0.12	0.09	0.08	0.12	0.16	0.21	0.19	0.08
	5	0.13	0.11	0.09	0.08	0.13	0.15	0.16	0.18	0.07
IV	3	0.20	0.19	0.17	0.22	0.24	0.19	0.17	0.25	0.14
	4	0.19	0.18	0.17	0.21	0.24	0.19	0.14	0.24	0.14
	5	0.19	0.18	0.17	0.21	0.24	0.19	0.14	0.24	0.14

and it should be treated as a problem specific parameter. Analytical results of AQM are compared to discrete event simulation outputs. Initial findings on the deviations from the optimal solution for an optimization approach is given by comparing the solution with minimum mean response time in AQM and the simulation model by complete enumeration of the location solutions satisfying coverage criterion. In the further studies, effect of relaxing single vehicle restriction on minimum mean response time could be useful to justify this relaxation in Euclidean networks. Another direction is to construct a solution algorithm for an optimization approach, where AQM defines a combinatorial problem and has no closed form formulation to be solved with package solvers in an optimization environment.

References

Beraldi, P., Bruni, M. E., 2009. A probabilistic model applied to emergency service vehicle location. European Journal of Operational Research 196(1), 323–331.

Berman, O., Krass, D., Drezner, Z., 2003. The gradual covering decay location problem on a network. European Journal of Operational Research 151(3), 474–480.

Boyaci, B., Geroliminis, N., 2015. Approximation methods for large-scale spatial queueing systems. Transportation Research Part B 74, 151–181.

Daskin, M., 1983. A maximum expected location model: formulation, properties and heuristic solution. Transportation Science 7(1), 48–70.

Geroliminis, N., Karlaftis, M. G., Skabardonis, A., 2009. A spatial queueing model for the emergency vehicle districting and location problem. Transportation Research Part B 43, 798–811.

Geroliminis, N., Kepaptsoglou, K., Karlaftis, M. G., 2011. A hybrid hypercube Genetic algorithm approach for deploying many emergency response mobile units in an urban network. European Journal of Operational Research 210, 287–300.

Hakimi, S., 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. Operations Research 13, 462–475.

Iannoni, A.P., Morabito, R., 2007. A multiple dispatch and partial backup hypercube queueing model to analyze emergency medical systems on higways". Transportation Research part E: logistics and transportation review 43.6 , 755–771.

Iannoni, A.P., Morabito, R., Saydam, C., 2008. A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways". Annals of Operations Research 157, 207–224.

Iannoni, A.P., Morabito, R., Saydam, C., 2009. An optimization approach for ambulance location an the districting of the response segments on highways". Eurpoean Journal of Operational Research 195(2), 528–542.

Iannoni, A. P., Morabito, R., Saydam, C., 2011. Optimizing large-scale emergency medical system operations on highways using the hypercube queueing model. Socio-Economic Planning Sciences 45(3), 105–117.

Kunkel, A. G., Van Itallie, E. S., Wu, D., 2013. Optimal distribution of medical backpacks and health surveillance assistants in Malawi. Health Care Management Science, 1–15.

Larson, R., 1974. A hypercube queueing model for facility location and redistricting in urban facility service. Computers & Operations Research 1(1), 67–95.

Morabito, F. C., Chiyoshi, F., Galvao, R. D., 2008. Non-homogeneous servers in emergency medical systems: Practical applications using the hypercube queueing model. Socio-Economic Planning Sciences 42, 255–270.

Revelle, C., Hogan, K., 1989. The maximum availability location problem. Transportation Science, 23(3), 192–200.

Steiger, N. M., Lada, E.K., Wilson, J. R., Joines, J. A., Alexopoulos, C., Goldsman, D., 2005. Asap3:a batch means procedure fo steady-state simulation analysis. ACM Transaction on Modeling and Computer Simulation 15(1),39-73.

Takeda, R. A., Widmer, J. A., Morabito, R., 2007. Analysis of ambulance decentralization in an urban emergency medical service using hypercube queueing model. Computers & Operations Research 34(3), 727–741.