LOCATING EMS VEHICLES IN GENERAL NETWORKS WITH AN APPROXIMATE QUEUEING MODEL AND A METAHEURISTIC SOLUTION APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MUHARREM ALTAN AKDOĞAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

SEPTEMBER 2015

Approval of the thesis:

LOCATING EMS VEHICLES IN GENERAL NETWORKS WITH AN APPROXIMATE QUEUEING MODEL AND A METAHEURISTIC SOLUTION APPROACH

submitted by **MUHARREM ALTAN AKDOĞAN** in partial fulfillment of the requirements for the degree of **Master of Science in Industrial Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Murat Köksalan
Head of Department, **Industrial Engineering** _____

Assoc. Prof. Dr. Pelin Bayındır
Supervisor, **Industrial Engineering Department, METU** _____

Assoc. Prof. Dr. Cem İyigün
Co-supervisor, **Industrial Engineering Department, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Canan Sepil
Industrial Engineering Department, METU _____

Assoc. Prof. Dr. Pelin Bayındır
Industrial Engineering Department, METU _____

Assoc. Prof. Dr. Cem İyigün
Industrial Engineering Department, METU _____

Assist. Prof. Dr. Özgen Karaer
Industrial Engineering Department, METU _____

Assist. Prof. Dr. Mustafa Alp Ertem
Industrial Engineering Department, Çankaya University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: MUHARREM ALTAN AKDOĞAN

Signature :

ABSTRACT

LOCATING EMS VEHICLES IN GENERAL NETWORKS WITH AN APPROXIMATE QUEUEING MODEL AND A METAHEURISTIC SOLUTION APPROACH

Akdoğan, Muharrem Altan
M.S., Department of Industrial Engineering
Supervisor : Assoc. Prof. Dr. Pelin Bayındır
Co-Supervisor : Assoc. Prof. Dr. Cem İyigün

September 2015, 59 pages

In this study, problem of optimal location decision of ambulances as an server-to-customer Emergency Medical Service (EMS) vehicle is discussed. Hypercube queueing models (HQM) are employed to achieve performance measures significant to the location decision of EMS vehicles in spatial networks. Geroliminis et al(2009) extends HQM and propose Spatial Queueing Model(SQM). Our study proposes a generalization of SQM to be used in general networks. Quality of approximations inherit to SQM is questioned, and reported against various system spesific variables such as distribution of demand among regions, traffic intensity or distribution of demand regions over the area. Restriction of number of servers to be located per each location as one in the SQM model is relaxed. Effect of allowing multi servers per each location in general networks are reported. A metaheuristic algorithm (genetic algorithm) is proposed to solve the model for which no closed form expression exists and its performance is reported.

Keywords: Emergency Medical Services, Ambulance Location, Hypercube Queueing Model, Approximate Queueing Model, Server-to Customer Models

ÖZ

ACİL TIBBİ YARDIM ARAÇLARININ GENEL AĞLARDA YAKLAŞIK BİR KUYRUK MODELİ VE SEZGİSEL BİR ÇÖZÜM YAKLAŞIMI İLE LOKASYONUNUN BELİRLENMESİ

Akdoğan, Muharrem Altan
Yüksek Lisans, Endüstri Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Pelin Bayındır
Ortak Tez Yöneticisi : Doç. Dr. Cem İyigün

Eylül 2015 , 59 sayfa

Bu çalışmada, müşteriye hizmet götüren acil tıbbi yardım servisleri olarak ambulansların en iyi lokasyonlarına karar verme problemi tartışılmıştır. Uzamsal ağlarda verilen ambulans lokasyon kararları sonucunda ortaya çıkan sistemin performans parametrelerini hesaplamak için Hiperküp Kuyruk Modeli (HKM) kullanılır. Geroliminis ve ekibi, HKM modelini genişletir ve Uzamsal Kuyruk Modeli (UKM) önerir. Bizim çalışmamız, UKM'nin genel ağlarda kullanılabilmesi için bir genişletme önerir. UKM modelinde kullanılan yakınsama yaklaşımlarının kalitesi sorgulanır, ve bu kalitenin değişimi, talep miktarının alanlara göre dağılımı, kuyruk sisteminin trafik yoğunluğu ve talep alanlarının ağ üzerindeki dağılımı gibi farklı ağ parametrelerine göre raporlanır. Bu çalışma her alana en fazla bir tane araç yerleştirme kısıtlamasını kaldırır. Genel ağlarda bunun etkisini raporlar. Kapalı bir ifadesi olmayan matematiksel model için sezgisel bir çözüm algoritması önerilir ve performansı raporlanır.

Anahtar Kelimeler: Acil Tıbbi Servisleri, Ambulans Lokasyonu Problemi, Hiperküp Kuyruk Modeli, Yaklaşık Kuyruk Modeli

To the young souls;

who passed away before getting a chance to fulfill their dreams...

ACKNOWLEDGMENTS

This thesis has become possible with not only my efforts but the help of many others on the way. I would like to give them the credits for helping this study happen.

First of all, I would like to thank to my advisors, Dr. Pelin Bayındır and Dr. Cem İyigün for their academic guidance with their knowledge, small talks and supports throughout the process and for their encouragements.

I would also like to thank to examining committee members Dr. Canan Sepil, Dr. Özgen Karaer and Dr. Mustafa Alp Ertem for their time in reviewing this work.

I thank to my mother and father for always believing in me and expressing their support every time whatever I choose to do in my life. My not-so-little sister also deserves to be thanked since she is always there to support me.

I would like to thank to my friends with whom I spend more time than my family in the distance: to Ezgi Evrim Özkol and Tuğçe Kırbaş for their motivational talks; to Melike İşbilir for being my thesis-comrade and for the discussions we made; to Elçin Ergin for being always ready to my questions and pokes over overseas; to Derya Kılınç for her unique sense of humor; to Emel Ekici for her never ending optimism and lastly to Gökhan Sarı although he always tries to dissuade me with his entrepreneurship.

Last but not least, I would like to thank my family-to-be, Deniz Coşkun for believing in me at my every step. I always have her support and life-guidance when I let my mood down. She always succeeds making me believe that I can get it done whatever it is, and I appreciate it.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vi

ACKNOWLEDGMENTS viii

TABLE OF CONTENTS ix

LIST OF TABLES xi

LIST OF FIGURES xiii

CHAPTERS

1 INTRODUCTION 1

2 LITERATURE REVIEW 3

3 PROBLEM DEFINITION AND MATHEMATICAL FORMULATION 7

3.1 The Environment 8

3.2 Problem Definition 9

3.3 Mathematical Formulation 12

3.3.1 Underlying Queuing System 14

3.3.2 Order of Districting 17

3.3.3 Calculation of Rates in Queuing System 19

4 SOLUTION APPROACH 27

4.1 Genetic Algorithm 27

5 COMPUTATIONAL STUDY 33

5.1 Problem Instances 33

5.2 Discrete Event Simulation 35

5.3 Comparison of Alternative Downward Transition Rate Formulations and Levels of Order of Districting Combinations . 36

5.4 Design of Experiment for Genetic Algorithm 44

5.5 Comparison of Single and Multi Server Model 46

6 CONCLUSION 51

REFERENCES 55

APPENDICES

A 59

LIST OF TABLES

TABLES

Table 3.1 Notation used for Environment 9

Table 3.2 Notation used for Mathematical Formulation 12

Table 3.3 Notation used for Service Rate and Mean Total Service Time 14

Table 3.4 Notation used for Sub Area Definition 17

Table 3.5 Sub Area Sets for Problem Instance 3.1 19

Table 3.6 Notation added for Transition Rate Calculation 19

Table 3.7 Notation added for Alternative I 22

Table 3.8 Notation added for Alternative II 23

Table 3.9 Notation added for Alternative IV 25

Table 4.1 Notation used for Genetic Algorithm 27

Table 5.1 Notation used for Problem Instance Parameters 34

Table 5.2 Levels of parameters for problem instances 37

Table 5.3 MAPE across DoD 38

Table 5.4 Fraction of Time Minimum Found across DoD 39

Table 5.5 MAPE across RoA 40

Table 5.6 Fraction of Time Minimum Found across RoA 41

Table 5.7 MAPE across VoD 42

Table 5.8 Fraction of Time Minimum Found across VoD 42

Table 5.9 MAPE across TI 43

Table 5.10 Fraction of Time Minimum Found across TI 43

Table 5.11 Overall MAPE 44

Table 5.12 DOE Setup 45

Table 5.13 Fraction of Time Optimum Found with GA 45

Table 5.14 Overall Performance 46

Table 5.15 Levels of Parameters for Problem Instances 47

Table 5.16 Comparison of MSM and SSM solutions 48

Table A.1 MAPE across DoD without problem instances with RoA 0.1 & 0.2 . 59

LIST OF FIGURES

FIGURES

Figure 3.1 A problem instance 18

Figure 5.1 Uniform DoD with 50 demand regions 34

Figure 5.2 Circular DoD with 50 demand regions 34

CHAPTER 1

INTRODUCTION

Emergency medical services (EMS) are defined by by Washington, DC, Department of Health [30] as prehospital services providing transportation to definitive care and on sight medical care to patients. EMS is a system of components which includes organizations, transportation and communication networks, hospitals, trauma centers, trained professionals, physicians, nurses, administrators helping this system working in success. Having the primary concern as emergencies, planing of this system requires significant work to ensure serving the public at its best. Other than administrative decisions, planing of physical infrastructure constitutes a major part in the performance of the system. This study specifically concerns a strategic problem where we determine locations of given number of ambulances in predefined possible regions.

While locating ambulances, there should be more than one concern to be taken into consideration, each targeting different performance measures. These could be quantitative or qualitative concerns which regard the view of the decision maker and or regulations. Some criteria can be listed as follows;

- utilization of ambulances
- fraction of demand lost
- mean response time
- investment and operational cost related to the ambulances

In this study, a model with the objective of mean response time is studied while conceiving other criteria listed above could be other focuses for an EMS system location

problem. Costs are not considered in the scope of this study. The problem is defined as locating given number of ambulances to predefined candidate ambulance locations which are also demand regions, by achieving minimum mean system response time under a coverage criterion.

Regarding the nature of the problem, queueing theory is used to analyze the performance of the candidate solutions to the problem. Spatial Queueing Model defined by Geroliminis et al [10] is studied to be applied on networks with dispersed regions other than regions with Manhattan distances as in their study.

SQM represents the system with an approximate queueing model. The quality of approximations in SQM is questioned for general networks and reported against various network specific parameters such as distribution of demand, traffic intensity, distribution of demand regions over the area and number of ambulances to be located. Restriction of the number of ambulances to be located per location to one is relaxed by allowing multiple ambulances per location. Effect of this relaxation is explored.

In the scope of the study, a metaheuristic algorithm is proposed to solve the model. Experimental results for this algorithm is provided.

In Chapter 2, a brief summary of studies on EMS location problems is provided. Problem definition, research questions and mathematical model is delivered in Chapter 3. Solution approach is explained in Chapter 4. In Chapter 5, the experimental results for research questions is analyzed. The study is concluded in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

In this chapter, a brief review of studies in the literature on emergency vehicle location models is provided.

Emergency vehicle locations are determined under strategic and operational considerations. Response time and the coverage are the main issues for emergency vehicle location problems.

Early studies model location problems in a static and deterministic way ignoring stochasticity of components such as demand and travel time. These models are categorized as *median* and *coverage* models by Geroliminis et al. [10]. *Median* models work on the average travel time or distance weighted with respect to the demand of the areas in responsible districts. Hakimi [26] formulates a p-median problem for locations of switching centers on a telephone network. Later, ReVelle and Swain [24] model the p-median problem as an integer programming model, allowing only locating facilities on the nodes of the network. For emergency vehicle location problems, Volz [20] locates and relocates ambulances in semi-rural areas regarding the minimum response time and uses a coverage constraint as a predefined coverage standard in minutes. Kunkel et al [6] uses weighted p-median problem to assign medical assistants to population centers and then Capacitated Facility Location Problem (CFLP) is solved to assign these medical assistants to resupply centers. *Coverage* models takes the covered demand or number of regions into account in terms of proximity to the vehicles. Coverage is defined as being closer to the demand areas than a predefined travel time. Hakimi [27] introduces Set Covering Location Problem (SCLP) to find minimum number of policemen to be located on a highway network. Toregas

et al [25] formulates the same problem as an integer programming problem to locate emergency vehicles. Maximal Covering Location Problem (MCLP) is introduced by Church and ReVelle [17] to locate emergency units in a way that every unit would cover the maximum possible part of demand area. Daskin and Stern [15] modify the original MCLP model and introduces the Backup Coverage Problem model which maximizes demand areas covered more than once. Later, Gendreau et al. [9] introduce the Double Standard Model (DSM) which includes two different time threshold for coverage consideration. Berman et al. [5] use coverage decay function in a generalization of MCLP and median-based models.

In emergency vehicle location problems, stochasticity of the problem inputs and change of inputs in time begin to take more attention, regarding the nature of the problem. In this context, *dynamic* and *probabilistic* models emerge. *Dynamic* models work on relocation of vehicles, facilities over a planning horizon. Ballou [22] puts an emphasis on limits of deterministic and static models and locates a single facility to maximize profit over a finite horizon. Scott [1] extends location-allocation problems dynamically by locating multiple facilities in discrete, equally spaced periods. Gendreau [8] extends his DSM for relocation of ambulances considering the nature of the problem. Degel et al [23] extends MCLP with time-dependent variations for ambulance location problem. Probabilistic models are developed to consider changing condition or inputs as the availability of vehicles, changing travel times. Daskin [14] introduces Maximum Expected Covering Problem for location analysis of public service facilities. Maximum Availability Location Problem(MALP) is developed by Revelle and Hogan [3] incorporating vehicle availability into the location problem. Beraldi and Bruni [16] use Two Stage Stochastic Programming in locating ambulances to cover demand with a specified reliance.

Queueing theory is embedded in probabilistic facility location-allocation models firstly in the study of Larson [21]. Hyper Queueing Model (HQM) proposed by Larson, analyzes vehicle location-allocation and districting in emergency response services operating as server-to-customer manner (such as ambulances). HQM model is used to observe steady-state probabilities of the system and get various performance measures such as travel times, work-loads, based on Markovian analysis and Queueing Theory. The developed model is not used in an optimization approach. Sacks and

Grief [28], Brandeau and Larson [13], Iannoni and Morabito [2], Takeda et al [12] use HQM model as a base to measure performance of EMS systems. Marianov and Revelle [29] extend the MALP with HQM and developes Queueing Maximum Availability Location Problem (QMALP). Mendonca and Morabito [7] extends HQM for different service rates for each server in the system but same for intra-district or inter-district responses for the same servers, where HQM is constructed with independent service times from locations of the calls and responding server. Halpern [11] states the estimations for service times in the study of Mendonca and Morabito [7] give questionable results where travel time is a significant part of the service time. Iannoni et al [18], [19] uses hypercube model in an optimization environment for location and districting on high ways with alternative objectives.

Geroliminis et al [10] extends HQM and develops Spatial Queueing Model (SQM) by defining non-identical service times for servers regarding the demand call's characteristics (inter-district or intra-district response). SQM relaxes the predefined server location in HQM. It is used in an optimization algorithm to deliver the optimal location solution for predefined number of service patrol vehicles to be located in possible server locations on a freeway. Dispatching preferences are also not predefined in advance. An heuristic solution approach consisting of random search followed by steepest decent method is used to find near optimal location solution while minimizing response time of the system.

CHAPTER 3

PROBLEM DEFINITION AND MATHEMATICAL FORMULATION

The problem studied is explained and the underlying mathematical formulation is delivered in this chapter. Problem to the center of this study is introduced. Problem environment is explained in Section 3.1. Problem and the research questions is defined in Section 3.2. Mathematical formulation is delivered in Section 3.3

This study aims to develop a mathematical formulation and a solution procedure that locates given number of ambulances to achieve minimum mean response time for the emergency medical service system. SQM defined for emergency vehicle services by Geroliminis et al. [10] constitutes the base for our study. Geroliminis et al. [10] model the problem as locating given number of vehicles in alternative locations , at which at most one vehicle is allowed.The objective is to minimize mean response time while the vehicle locations in the solution should be reachable by at least a given percentage of the regions within a given travel time threshold.

In this model, as explained in Chapter 2, for a given vehicle location solution, dispatch fractions of the vehicles to the regions are taken from a queuing system. These fractions are used to calculate mean response time to emergency calls under the given locations. SQM is an approximate queueing system. Order of districting concept is used to indicate the required level of approximation to be able to achieve the behavior of the exact queueing system. Transition rates between states of the approximate system are calculated regarding the order of districting decision. Travel distances between location of vehicle and location of a demand call is taken into consideration while calculating the service rate for that specific call. However, in the study no

closed form expression is delivered for travel distance consideration.

The model is applied to real life data to locate Free Service Patrol vehicles on a free-way in San Francisco Bay Area. Service rates for the approximate queueing system are generated from processing of this data. In that study, at most one ambulance is allowed to be located per each candidate ambulance location. Level of order of districting is set to 3 by stating calculations for further levels are tedious . The solution approach for the problem modeled is stated as random search followed by steepest decent method and it is tested against the real life data. Accident statistics on a 72 miles section of I-80 free-way in San Francisco Bay Area is used. 7 instances are solved with different demand rates. 500 randoms server locations were calculated for each instance and iterated in the algorithm.

3.1 The Environment

The environment in the study of Geroliminis et al. [10] as follows.

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

Demand is defined as the call for ambulance services 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 an ambulance in a unit time is known for each demand region.

For each region, a list of ambulance locations sorted with respect to the proximity is available. When a call is received from a certain region, ambulance locations are checked for availability in the order of proximity to demand. When an available ambulance is found, the service starts. If an available ambulance in the system cannot be found, the demand is lost.

Travel times are defined between regions. It is assumed exponentially distributed with known mean.

A service is composed of travel time to the demand region from the ambulance location, travel time to the health care center from the demand region and travel time back to the ambulance location. Total service time is random, dependent on the location of both ambulance and the patient and assumed to be exponentially distributed. Since order of districting is used in the queueing system, for every transition a set of demand regions subject to this transition is defined as an area and service rate for this area is calculated. As mentioned before, no formulation for this calculation is delivered in the study.

The fraction of total demand generated from a demand region is calculated from the demand rates. Any solution to the problem should cover regions of which demand add up to a given percentage of total demand, within a given travel time threshold.

The quality of the service is defined as mean response time. The objective function is selected as mean response time since it is one of the most important performance measures for EMS systems.

Notation used is given in Table 3.1.

Table 3.1: Notation used for Environment

Sets	
Q	Set of demand regions
Subsets	
R	Set of candidate ambulance locations, $R \subseteq Q$
Parameters	
N	Number of ambulances to be located
t_{qr}	Mean travel time between demand region q and ambulance location r in minutes
ω_q	Demand rate (number of demand calls per hour) for demand region $q \in Q$
f_q	Frequency of demand generated from demand region $q \in Q$, $= \frac{\omega_q}{\sum_{q \in Q} \omega_q}$
α	Required minimum coverage

3.2 Problem Definition

The problem is defined as follows;

"Locating given number of ambulances to predefined possible ambulance locations, which are also demand regions, under a coverage criterion by minimizing mean response time of the system"

SQM to be used is an approximate model to the exact queuing system which generates a wider state space than SQM. Approximation in the model appears in transition rate calculations for the queueing system and in the order of districting decision. Since no generic expression exists for the service rates used in transition rate calculation, we aim to deliver a generic expression for the service rate calculations for the approximate model.

Order of districting is another major point that determines the quality of the approximate model. The order of districting term indicates the size of the list to be used in ordering ambulances in terms of proximity to a region. If we were to work with third order of districting, it states that for a demand call from any region in the system subject to the problem, at most the third nearest ambulance would be needed to be assigned to this call in exact queueing system, meaning fourth nearest ambulance would never be needed in that system. Thus, three as the level of order of districting is adequate to give information to the model about the assignment scheme of ambulances in the exact queueing system. However, this scheme could be subject to change depending on system parameters as demand, service time patterns, location distribution of regions in the entire area.

The current study works on order of districting to observe the effect of changing orders in approximating the mean response time of the exact queueing system. Minimum mean response time reported by using the approximate model is compared to the mean response time of the simulation of the exact queueing system for location solution of approximate model. With the guidance of this comparisons, we aim to discuss the effect of order of districting in the quality of approximation to the exact queueing system. In the scope of the current study, we do not discuss whether approximate model solution is optimal for the model constructed with exact queueing system.

Other than quality of the approximations, the base model of Geroliminis et al. [10] allows locating at most one ambulance in one region. This restriction can be reason-

able for networks like high ways. In the study by Geroliminis et al. [10], there is no further comment on networks with dispersed regions. This approach can result in missing better solutions, unless it is required by the decision maker as a physical constraint. Hence, relaxing this restrictions and considering the effect of locating multiple ambulances at a single location is included in our study.

We also aim to deliver a solution approach for the model developed which includes terms for which no closed form expression exists. In the study by Geroliminis et al [10], solution approach proposed for the model is not tested against different problem instances and its performance is not clearly tested with respect to the optimal solution. Hence, a genetic algorithm is developed, tested under various problem parameters and its performance is reported.

Therefore, the main contributions of our study can be listed as follows;

- Developing a generic expression and looking into the effect of different service rate formulations for the approximate queueing model on the approximation to the exact queueing system.
- The effect of order of districting level on the quality of approximation.
- Effect of locating multiple in stead of single ambulance at a single location on minimum mean response time of the system.
- Developing a solution approach and assessing its performance under different system parameters.

3.3 Mathematical Formulation

In this section, mathematical model proposed by Geroliminis et al. [10] is delivered regarding the notation given in Table 3.2. Then, the changes required to include into this base model is explained.

Table 3.2: Notation used for Mathematical Formulation

Notation	
Q	Set of demand regions
R	Set of candidate ambulance locations, $R \subseteq Q$
W_q	Set of possible ambulance locations covering demand region q : $\{r \in R : t_{rq} \leq T\}, \forall q \in Q$
N	Parameter indicating the number of ambulances to be located
T	Parameter indicating travel time threshold for coverage
t_{qr}	Parameter indicating the mean travel time between location q and r in minutes
f_q	Parameter indicating the frequency of demand generated from demand region $q \in Q, = \frac{\omega_q}{\sum_{q \in Q} \omega_q}$
α	Parameter indicating the required minimum coverage
x_r	Decision variable indicating the number of ambulances located in location $r \in R$
y_q	Decision variable indicating whether demand region q is covered by an ambulance or not 1, if demand region q is covered by an ambulance 0, otherwise
$\bar{x}$	Vector of x_r in a 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 ambulance location n has the nearest available ambulance for region q under $\bar{x}$, $1 \leq n \leq \lvert B_i(\bar{x})\rvert$
$r_n(\bar{x})$	Ambulance 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 ambulances in location $r_n(\bar{x})$ to region q

Mathematical model to find ambulance locations to minimize mean response time can be expressed as follows:

$$\text{Minimize} \quad \bar{T} = \sum_{n=1}^{|B_i(\bar{x})|} \sum_{q=1}^{Q} \left(\rho_{r_n(\bar{x})q} t_{r_n q}\right) \tag{3.1}$$

$$\text{subject to:} \quad \sum_{q \in Q} f_q y_q \geq \alpha, \tag{3.2}$$

$$\sum_{r \in W_q} x_r \geq y_q, \forall q \in Q \tag{3.3}$$

$$\sum_{r \in R} x_r = N, \tag{3.4}$$

$$x_r \in \{0,1\}, \forall r \in R \tag{3.5}$$

$$y_q \in \{0,1\}, \forall q \in Q \tag{3.6}$$

$$\begin{aligned} \rho_{r_n(\bar{x})q} = f_q \frac{\sum_{B_i(\bar{x}) \in E_{nq}(\bar{x})} P(B_i(\bar{x}))}{\left(1 - P(B_{|B(\bar{x})|-1}(\bar{x}))\right)}, \quad r_n(\bar{x}) \in R, \\ 1 \leq n \leq |B_i(\bar{x})|, \\ \forall q \in Q \end{aligned} \tag{3.7}$$

$$\begin{aligned} & P(B_j(\bar{x})) * \left[\sum_{B_i(\bar{x}) \in B(\bar{x}): d_{ij}^- = 1} \lambda_{ij}(\bar{x}) + \sum_{B_i(\bar{x}) \in B(\bar{x}): d_{ij}^+ = 1} \mu_{ij}(\bar{x}) \right] \\ & = \sum_{B_i(\bar{x}) \in B(\bar{x}): d_{ij}^- = 1} \mu_{ij}(\bar{x}) * P(B_i(\bar{x})) \\ & \quad + \sum_{B_i(\bar{x}) \in B(\bar{x}): d_{ij}^+ = 1} \lambda_{ij}(\bar{x}) * P(B_i(\bar{x})), \\ & \quad j = 0, 1, ..., |B(\bar{x})| - 1, \end{aligned} \tag{3.8}$$

$$\sum_{i=0}^{|B(\bar{x})|-1} P(B_i(\bar{x})) = 1, \tag{3.9}$$

The model is up to minimize the mean system response time while using queueing models to calculate necessary parameter for mean response time measure. Constraint 3.2 is minimum coverage restriction stating a solution should have a coverage score larger then a predefined value $\alpha \leq 1$. Constraints in Equations 3.3 - 3.6 are added for estimation of coverage. Constraint 3.3 forces y_q taking value of zero if there are no ambulances located covering region q. Constraint 3.4 implies the number of ambulances located should be equal to the predefined number N. Constraints 3.5 and 3.6 stand for decision variables x and y taking integer values respectively. Constraint

3.7 is to calculate fraction of dispatch of ambulances with respect to locations. Constraints 3.8 and 3.9 are balance equations regarding the solution stated by $\bar{x}$.

This model allows locating at most one ambulance on a candidate location. We relax this restriction and replace constraint 3.5 with Equation 3.10

$$x_r \geq 0, \forall r \in R \tag{3.10}$$

We explained that the service rates for transitions are data-driven in the study of Geroliminis et al. [10]. We propose generic formulations using predefined service rates to be used in transition rates. Mean total service time is assumed to be known only for the case that a demand region is served by an ambulance located in that region. Therefore, travel times should be included into calculation of the total service time if an ambulance from a different region is serving the call. Notation used for service rate and mean total service time is given in Table 3.3.

Table 3.3: Notation used for Service Rate and Mean Total Service Time

Parameters	
ν_q	Mean total service time in minutes for demand region $q \in Q$
ϕ_q	Service rate per hour ($\frac{60}{\nu_q}$) for demand region $q \in Q$,

In the rest of the study, $\bar{x}$ is dropped from the notation given in Table 3.2 for the sake of brevity.

3.3.1 Underlying Queuing System

An approximate queuing model is generated to obtain the performance measures and accordingly to calculate the objective function value under a solution as mentioned previously. Notation and definitions used are similar to those by Geroliminis et al. [10]

To represent a solution to the problem in exact queueing system, we have to track where an ambulance is located in, if it is busy or not and which region it serves if it is busy. Then, the underlying queueing system can be represented by an N-dimensional

state as follows:

$$B_i = (b_1, b_2, ..., b_N), \tag{3.11}$$

$$b_i = \{s, r\}, \qquad i = 1, ..., N \tag{3.12}$$

where $s \in (0 \bigcup Q)$, r states in which location the ambulance represented in i^{th} entry of the state definition located. $b_i = \{0, r\}$ represents the state that the ambulance is free and $b_i = \{q, r\}$ represents the state where the ambulance 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 ambulances.

In the approximate queueing model, system can be modeled by an n-dimensional state indicating the number of busy ambulances on each location selected in the solution.

$$B_i = (b_1, b_2, ..., b_k, ..., b_n) \tag{3.13}$$

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 ambulances 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 ambulances.

For clarification of the approximate identification of the queueing system worked with, it is, explicitly, resulted from not tracking which region a busy ambulance is serving in the state definition. We need to track where an busy ambulance 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 stystem, only one step transition between states is allowed, meaning only one

ambulance 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 [21] 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 $B_j = (b_1, ..., 1_k, ..., b_{|B_s|-1}, b_{|B_s|})$

Recall that balance equations for the queuing system exist in the form of constraints in the mathematical model, namely Constraint 3.8 and 3.9 as follows:

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

In the mathematical model, the objective is to minimize the mean system response

time. Fraction of dispatches of ambulances to every region is used to express the objective function in Equation 3.1.

Fraction of dispatches from ambulance location r_n to region q can be stated in Constraint 3.7 in the mathematical model:

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

Numerator in 3.7 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 ambulances from ambulance location r_n to demand region q.

3.3.2 Order of Districting

Order of districting is introduced to estimate the upward and downward transition rates in the approximate queuing system. n^{th} order of districting indicates that demand in every region is satisfied by at most n^{th} closest ambulance 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 defined in Table 3.4

Table 3.4: Notation used for Sub Area Definition

Parameters	
O	Maximum level of order of districting
Sets	
D^n_{kl}	Set of regions belonging to the 1^{st} order of district of server k and the n^{th} order of district of server l, $n = 1...O$
U_q	Ordered set of ambulance locations selected in the solution with respect to mean travel times to demand region q, closest first

Regarding the notation, sub areas are defined as follows:

$$D_{kl}^{n} = \{q \in Q : U_q(1) = r_k \text{ and } U_q(n) = r_l\} \tag{3.14}$$

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

To clarify the order of districting structure, a problem instance is used. In Figure 3.1, red dots represent the demand regions without an ambulance and blue dots represent the locations with at least an ambulance. According to the figure, location 9 has two ambulances, location 10 has one ambulance and location 14 has one ambulance.

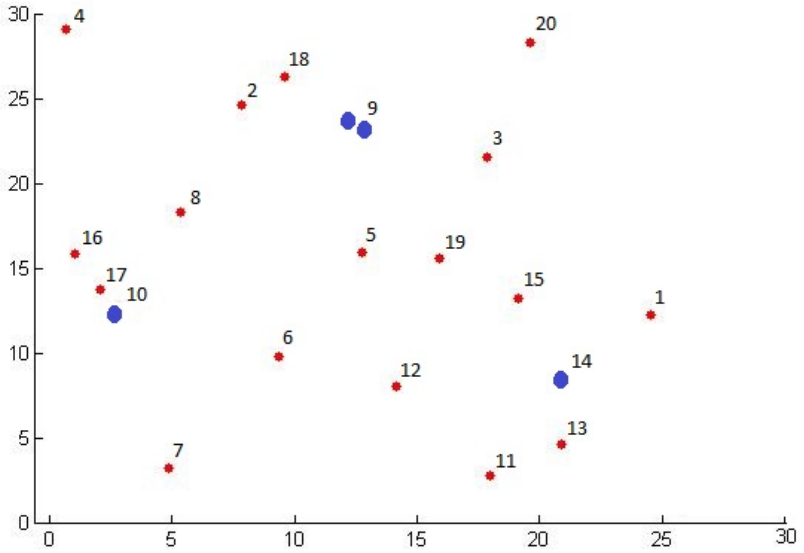

Figure 3.1: A problem instance

For this instance, assume $r_1 = 9, r_2 = 10$ and $r_3 = 14$. Then, D_{kl}^{n} sets occurs as in Table 3.5 for $O = 3$.

Table 3.5: Sub Area Sets for Problem Instance 3.1

Sets	
$D^1_{1,1}$	$\{2, 3, 4, 5, 9, 18, 19, 20\}$
$D^1_{2,2}$	$\{6, 7, 8, 10, 16, 17\}$
$D^1_{3,3}$	$\{1, 11, 12, 13, 14, 15\}$
$D^2_{1,2}$	$\{2, 4, 5, 9, 18\}$
$D^2_{1,3}$	$\{3, 19, 20\}$
$D^2_{2,1}$	$\{8, 10, 16, 17\}$
$D^2_{2,3}$	$\{6, 7\}$
$D^2_{3,1}$	$\{1, 14, 15\}$
$D^2_{3,2}$	$\{11, 12, 13\}$
$D^3_{1,2}$	$\{3, 19, 20\}$
$D^3_{1,3}$	$\{2, 4, 5, 9, 18\}$
$D^3_{2,1}$	$\{6, 7\}$
$D^3_{2,3}$	$\{8, 10, 16, 17\}$
$D^3_{3,1}$	$\{11, 12, 13\}$
$D^3_{3,2}$	$\{1, 14, 15\}$

3.3.3 Calculation of Rates in Queuing System

Sub areas defined in Section 3.3.2 are used in the calculation of upward and downward transition rates since the exact region to which an ambulance is sent for a demand call is not tracked. Added notation for this calculations is given in Table 3.6

Table 3.6: Notation added for Transition Rate Calculation

Sets	
L^n_{kl}	Set of regions in the sub area D^n_{kl}
Parameters	
Ω^n_{kl}	Total demand of the regions in the sub area D^n_{kl} $= \sum_{q \in L^n_{kl}} \omega_q$
Λ	Total demand of the system $= \sum_{q \in Q} \omega_q$

For every order of districting, it is obvious that demand will be completely covered.

$$\sum_{k \in |B_i|} \sum_{l \in |B_i|} \Omega^n_{kl} = \Lambda, \forall n = 1, 2, ..., min(O, |B_i|) \tag{3.15}$$

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;

$$\lambda_{ij} = \Omega^1_{kk} + \sum_{1 \le l_1 \le |B_i| : b_{l_1} = x_{r_{l_1}}} \Omega^2_{l_1 k} + \sum_{m=3}^{M} \sum_{1 \le l_1, ..., l_{m-1} \le |B_i| : b_{l_i} = x_{r_{l_i}}} \Omega^m_{l_1 k} \bigcap \Omega^{m-1}_{l_1 l_{m-1}} \bigcap ... \bigcap \Omega^2_{l_1 l_2} \quad (3.16)$$

where $M = \min\{A, O\}$, $A = |\{b_s \in B_j : 1 \le s \le |B_i| \ and \ b_s = x_{r_s}\}|$ 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.

Equation 3.16 declares that in the transition from state i to j, an available ambulance 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 ambulances are busy.

Example

For states $B_i = (2, 1, 0, 3, 2, 0)$, $B_j = (2, 1, 1, 3, 2, 0)$, assume that the number of ambulances 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 Equation 3.16 , $k = 3$ and $M = 4$. Accordingly, λ_{ij} is calculated as follows;

$$\lambda_{ij} = \tau_1 + \tau_2 + \tau_3 + \tau_4 \quad (3.17)$$

where

$$\tau_1 = \ \Omega^1_{33} \quad (3.18)$$

$$\tau_2 = \ \Omega^2_{13} + \Omega^2_{23} + \Omega^2_{43} + \Omega^2_{53} \quad (3.19)$$

$$\begin{aligned}\tau_3 = &\Omega^3_{13}\bigcap\Omega^2_{12} + \Omega^3_{13}\bigcap\Omega^2_{14} + \Omega^3_{13}\bigcap\Omega^2_{15}\\ &+\Omega^3_{23}\bigcap\Omega^2_{21} + \Omega^3_{23}\bigcap\Omega^2_{24} + \Omega^3_{23}\bigcap\Omega^2_{25}\\ &+\Omega^3_{43}\bigcap\Omega^2_{41} + \Omega^3_{43}\bigcap\Omega^2_{42} + \Omega^3_{43}\bigcap\Omega^2_{45}\\ &+\Omega^3_{53}\bigcap\Omega^2_{51} + \Omega^3_{53}\bigcap\Omega^2_{52} + \Omega^3_{53}\bigcap\Omega^2_{54}\end{aligned} \tag{3.20}$$

$$\begin{aligned}\tau_4 = &+\Omega^4_{13}\bigcap\Omega^3_{12}\bigcap\Omega^2_{14} + \Omega^4_{13}\bigcap\Omega^3_{12}\bigcap\Omega^2_{15}\\ &+\Omega^4_{13}\bigcap\Omega^3_{14}\bigcap\Omega^2_{12} + \Omega^4_{13}\bigcap\Omega^3_{14}\bigcap\Omega^2_{15}\\ &+\Omega^4_{13}\bigcap\Omega^3_{15}\bigcap\Omega^2_{12} + \Omega^4_{13}\bigcap\Omega^3_{15}\bigcap\Omega^2_{14}\\ &+\Omega^4_{23}\bigcap\Omega^3_{21}\bigcap\Omega^2_{24} + \Omega^4_{23}\bigcap\Omega^3_{21}\bigcap\Omega^2_{25}\\ &+\Omega^4_{23}\bigcap\Omega^3_{24}\bigcap\Omega^2_{21} + \Omega^4_{23}\bigcap\Omega^3_{24}\bigcap\Omega^2_{25}\\ &+\Omega^4_{23}\bigcap\Omega^3_{25}\bigcap\Omega^2_{21} + \Omega^4_{23}\bigcap\Omega^3_{25}\bigcap\Omega^2_{24}\\ &+\Omega^4_{43}\bigcap\Omega^3_{41}\bigcap\Omega^2_{42} + \Omega^4_{43}\bigcap\Omega^3_{41}\bigcap\Omega^2_{45}\\ &+\Omega^4_{43}\bigcap\Omega^3_{42}\bigcap\Omega^2_{41} + \Omega^4_{43}\bigcap\Omega^3_{42}\bigcap\Omega^2_{45}\\ &+\Omega^4_{43}\bigcap\Omega^3_{45}\bigcap\Omega^2_{41} + \Omega^4_{43}\bigcap\Omega^3_{45}\bigcap\Omega^2_{42}\\ &+\Omega^4_{53}\bigcap\Omega^3_{51}\bigcap\Omega^2_{52} + \Omega^4_{53}\bigcap\Omega^3_{51}\bigcap\Omega^2_{54}\\ &+\Omega^4_{53}\bigcap\Omega^3_{52}\bigcap\Omega^2_{51} + \Omega^4_{53}\bigcap\Omega^3_{52}\bigcap\Omega^2_{54}\\ &+\Omega^4_{53}\bigcap\Omega^3_{54}\bigcap\Omega^2_{51} + \Omega^4_{53}\bigcap\Omega^3_{54}\bigcap\Omega^2_{52}\end{aligned} \tag{3.21}$$

τ_2 is the term representing the explicit form of the summation in the second term of λ_{ij} in Equation 3.16, 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.

Alternative I

Notation added for Alternative I is given in Table 3.7.

Table 3.7: Notation added for Alternative I

Parameters	
Ω_p	p_{th} term of Equation 3.17 in explicit form
Φ_p	Sum of service rates of the regions in the sub area corresponding to the p_{th} term of Equation 3.17 in explicit form
P	Set of order of terms in Equation 3.17 in explicit form

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

$$\mu_{ij} = \frac{\lambda_{ij}}{\sum_{p=1}^{P} \frac{\Omega_p}{\Phi_p}} \tag{3.22}$$

For the example of which λ_{ij} is given in Equation 3.17, term $\frac{\Omega_p}{\Phi_p}$ of Equation 3.22 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 of the Equation 3.22 occurs as the traffic intensity of the system consisting of location indexed with k in state definition and sub areas in Equation 3.17. It should also be stated that denominator forms as the sum of the traffic intensity of the sub areas; similar to the previous definition as systems consisting of individual terms of Equation 3.17 and location indexed with k in state definition.

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

Alternative II

In the first formulation of μ_{ij} given in Equation 3.22, service rates are used directly without any consideration of travel time to the demand region. If there is no ambulance located in the region, it is obvious that service rate for transition occurring from serving this region from another ambulance location would be different from serving with an ambulance from inside the region. Therefore, it is considered to be meaningful processing service rates with respect to the location of the ambulance 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 ambulance serving in that transition and summed to calculate the downward transition rate. The notation for formulation is delivered in Table 3.8

Table 3.8: Notation added for Alternative II

Sets	
L_{ij}	set of regions included in the sub areas resulting of transition from state i to j
Parameters	
$\phi^{'}_{qr}$	Service rate per hour for demand region $q \in Q$ when it is served from ambulance location $r \in R$

For 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 Equation 3.17 and L_{ij} as the set of regions in the corresponding sub areas for transition from state i to j. The location of ambulance 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 ambulance location for k^{th} entry in the state space representation as defined in Table 3.2 Then, for this transition, service rates ($\phi^{'}_{lr_k}$) of regions that could be served in this transition can be calculated as in Equation 3.23.

$$\phi^{'}_{lr_k} = \frac{60}{\frac{60}{\phi_{lr_k}} + 2 * t_{lr_k}}, \ \forall l \in L_{ij} \tag{3.23}$$

where ϕ_{lr_k} is defined as per hour and t_{lr_k} is given in minutes.

Followingly, μ_{ij} can be written as;

$$\mu_{ij} = \sum_{l \in L_{ij}} \phi^{'}_{lr_k} \tag{3.24}$$

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 also considered.

For a transition, it is obvious that only one ambulance 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 ambulances are serving simultaneously in this transition which multiplies the real service rate.

Based on this consideration, downward transition rate is formulated as the weighted average recalculated service rates ($\phi^{'}_{lr_k}$) of the regions with respect to demand frequencies of these regions in Alternative II.

$$\mu_{ij} = \sum_{l \in L_{ij}} \left(\frac{f_l}{\sum_{m \in L_{ij}} f_m} \right) * \phi^{'}_{lr_k} \tag{3.25}$$

Alternative IV

This alternative is based upon the first formulation 3.22. Traffic intensity approach is used. Differently, processed service rates and frequencies are included. Demand frequencies is calculated for every term of the Equation 3.17 regarding the total demand of the sub area corresponding to each term.

Table 3.9: Notation added for Alternative IV

Sets	
L_{ijp}	Set of regions included in the sub area corresponding p^{th} term of Equation 3.17 in explicit form, resulting from transition from state i to j, $p \in P$
Parameters	
F_p	Weighted demand frequency of the sub area corresponding to p^{th} term of Equation 3.17 in explicit form, $p \in P$
$\Phi_p^{'}$	Sum of recalculated service rates ($\phi_{ln_k^{'}}^{'}$) of the regions in the sub area corresponding to the p_{th} term of Equation 3.17 in explicit form

Regarding the notation, F_p is calculated from Equation 3.26.

$$F_p = \frac{\sum_{k \in L_{ijp}} f_k}{\sum_{p \in P} \sum_{k \in L_{ijp}} f_k} \tag{3.26}$$

Frequencies defined for terms of Equation 3.17, F_p , are used inversely proportional to traffic intensities of the corresponding terms. Accordingly, Alternative IV for downward transition rate is formulated as the following;

$$\mu_{ij} = \frac{\lambda_{ij}}{\sum_{p=1}^{P} \frac{\Omega_p}{\Phi_p^{'} * F_p}} \tag{3.27}$$

CHAPTER 4

SOLUTION APPROACH

In this chapter, solution approach for the model given in Chapter 3 is delivered. In Section 4.1, algorithm is described and pseudo code for the algorithm is delivered.

4.1 Genetic Algorithm

The model developed includes terms for which no closed form expression exists. Therefore, a genetic algorithm (GA) is constructed to solve the problem.

Genetic algorithm uses chromosome structure to encode different solutions and compares their fitness function values, i.e objective function values. It is designed to generate a population of initial solutions and evolve toward better ones in terms of the fitness function value. Evolution is realized through reproduction of the population using two main genetic operators; crossover and mutation operator which create next generation for the algorithm.

Table 4.1: Notation used for Genetic Algorithm

Parameters	
S_{Pop}	Size of the population and the mating pool
fit_i	Fitness value of individual i of the population
$P_s(i)$	Probability of selection of individual i
P_c	Probability of crossover
c_p	Crossover point
P_m	Probability of mutation

A solution to the model is simply stating the number of ambulances located in given

locations. In the chromosome structure, a solution is represented by an N-dimensional array. Each entry of this array is called as gene of the chromosome and shows the location of an ambulance in the solution. Since the model allows multiple ambulances in one location, genes with same locations show up in chromosomes. The number of genes with the same location indicates the number of ambulances located in that location.

For an instance where $N = 11$, the solution at hand $\bar{x} = (2, 1, 2, 3, 2, 1, 0, 0, 0, 0)$ and $I = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, chromosome is encoded as $\langle 1|1|2|3|3|4|4|4|5|5|6\rangle$. Each chromosome is called an individual.

Fitness value of an individual is simply the objective function value. Equation 3.1 is used to calculate the objective function value and is recorded as the fitness value of the chromosome.

A multiple number of S_{Pop} many random individuals are generated to form the initial population. Each individual should represent a feasible solution for the model to be solved. For this reason, the coverage values for these solutions are calculated and the infeasible individuals are discarded. Then, best S_{Pop} feasible individuals with respect to coverage values are selected as the initial population. If the number of individuals is less than S_{Pop}, then feasible ones are chosen randomly with equal probabilities and multiplied to generate S_{Pop} individuals for the initial population. This procedure guarantees working with feasible solutions for the model.

In reproduction process, predefined number of individuals, S_{Pop}, are copied to a mating pool to reproduce the next generation population. Selection of the individuals is based on their fitness value. A larger probability is assigned to an individual with smaller fitness value.

The probability of selection for each individual is given by:

$$P_s(i) = \frac{\frac{1}{fit_i}}{\sum_{j=1}^{S_{Pop}} \frac{1}{fit_j}} \tag{4.1}$$

After constructing the mating pool with individuals from the population, parents are selected in pairs for reproduction. This selection is random with equal probabilities

for all individuals in the mating pool.

Crossover operator is used to transfer genes from parents to children. Crossover operation is applied to selected pairs of individuals (as parents) with a predefined probability P_c. If the probability fails, crossover operation is not applied, the parents are duplicated, i.e, resulting children are the same with the parents. If crossover would be applied, one point crossover is used. A crossover point, c_p, is selected randomly between 1 and N. First c_p genes of parent 1 is copied to child 1, and of parent 2 to child 2. Genes after crossover point, c_p, are copied from parent 1 to child 2 and from parent 2 to child 1.

As an example consider Parent 1, Parent 2 which are given below and $c_p = 5$. Child 1 and Child 2 are reproduced with crossover as follows;

Parent 1	Parent 2
$\langle 1\|1\|2\|3\|3\|4\|4\|4\|5\|5\|6\rangle$	$\langle 1\|2\|3\|3\|3\|4\|6\|6\|7\|7\|10\rangle$
Child 1	Child 2
$\langle 1\|1\|2\|3\|3\|4\|6\|6\|7\|7\|10\rangle$	$\langle 1\|2\|3\|3\|3\|4\|4\|4\|5\|5\|6\rangle$

After reproducing children from crossover operation, every gene of a child is mutated to diversify the solutions in the population and better search the solution space by not restricting the search to solutions only with genes represented in a generation. Mutation is realized with probability P_m for every gene in a child. If probability succeeds, the gene of the child is overridden by a random location from set I with equal probabilities.

After crossover and mutation operations, $2 * S_{Pop}$ individuals exist in the mating pool including the children. Infeasible individuals in terms of coverage are discarded. Next generation is constructed from the best S_{Pop} feasible individuals in terms of the fitness function value from the mating pool. If the number of feasible individuals in the mating pool is less than S_{Pop}, feasible ones are chosen randomly with equal probabilities and multiplied to have S_{Pop} individuals in the next generation.

When the next generation produced from a population of individuals consists of a

single chromosome represented S_{Pop} times, it is stated that the population converges. This individual is taken as the best solution suggested by genetic algorithm and iteration is terminated.

The pseudo code of the GA is given in Algorithm 4.1.

Algorithm 4.1: Pseudo Code

1: $RP = \{rp_1, ..., rp_{S_{Pop}*15}\}$:Generate RP with $S_{Pop} * 15$ solutions
2: $Pop = \{p_1, ..., p_{S_{Pop}}\}$: INITIAL POPULATION $(rp_1, ..., rp_{S_{Pop}*15})$
3: **repeat**
4: **for** $i = 1$ to S_{pop} **do**
5: $p = rand(0, 1)$
6: $m_i = p_i$ where $\sum_{j=1}^{i-1} P_s(j) \leq p \leq \sum_{j=1}^{i} P_s(j)$ and $MatPool = \{m_1, ..., m_i\}$:
7: **end for**
8: **for** $i = 1$ to S_{pop}/2 **do**
9: Generate $i = rand(1, m_{S_{Pop}})$ and $j = rand(1, m_{S_{Pop}})$
10: $Parent_1 := m_i$ and $Parent_2 := m_j$
11: $p_{=}rand(0, 1)$
12: **if** $p \leq P_c$ **then**
13: $Child_1 :=$ CROSSOVER$(Parent_1, Parent_2)$
14: $Child_2 :=$ CROSSOVER$(Parent_1, Parent_2)$
15: **else**
16: $Child_1 := Parent_1$
17: $Child_2 := Parent_2$
18: **end if**
19: **for** $j = 1$ to 2 **do**
20: $p = rand(0, 1)$
21: **if** $p \leq P_m$ **then**
22: $Child_j :=$ MUTATION$(Child_j)$
23: **end if**
24: **end for**
25: Pop = NEXT GEN $(Parent_1, ..., Parent_{S_{Pop}}, Child_1, ..., Child_{S_{Pop}})$
26: **end for**
27: **until** Termination condition is satisfied
28: **return** The best individual

CHAPTER 5

COMPUTATIONAL STUDY

In this chapter, computational study is provided. Problem instance generation is explained in Section 5.1. Discrete event simulation study to achieve the exact performance measures of a solution to the mathematical model in 3.3 is provided in Section 5.2.

Experimental results for generic service rate formulations proposed in Chapter 3 for the approximate queueing system and effect of order of districting levels is provided in Section 5.3. Genetic algorithm proposed to solve the problem is tested. Analysis of parameter setting for the algorithm is reported in Section 5.4. The effect of relaxing single server restriction for locations on mean system response time is provided in Section 5.5.

5.1 Problem Instances

For this study, problem instances are generated with different specifications. 6 parameters are considered significant to the problem environment. They are given in Table 5.1 with different levels to be used in this study.

Table 5.1: Notation used for Problem Instance Parameters

Parameters		Levels
DoD	Distribution of demand regions over the area	uniform circular
NoD	Number of demand regions	-
RoL	Ratio of number of possible ambulance locations to number of demand regions	0.3 0.4
RoA	Ratio of number of ambulances to be located to number of demand regions	0.2 0.3
VoD	Variance of demand	low high
TI	Traffic intensity	low medium high

In uniform DoD, demand regions are uniformly distributed over the area as given in Figure 5.1

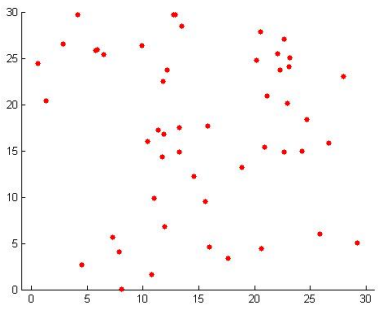

Figure 5.1: Uniform DoD with 50 demand regions

In circular DoD, 40 percent of the regions is placed in a inner circle, 30 percent in the middle and rest in the outer circle as can be seen in Figure 5.2

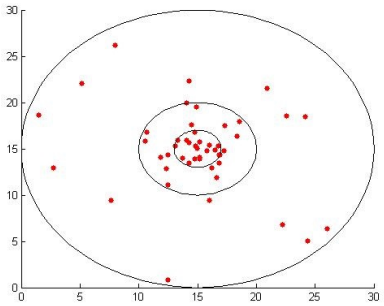

Figure 5.2: Circular DoD with 50 demand regions

Demand of the regions is generated from uniform distribution with mean equal to 4 and, variance equal to 0.33 for low VoD and to 3 for high VoD levels. Service rates for these regions is assumed equal. Total demand is divided into traffic intensity. This division is again divided to the number of regions and is written as the service rate for all regions.

For all the problem instances, minimum required coverage, α, is taken as 0.90 and travel time threshold, T, as 10 minutes.

5.2 Discrete Event Simulation

To be able to compare the results of different models, it is necessary to observe the objective function value of the exact system for the solution delivered by these models. To observe the behavior of the underlying exact queueing system, a simulation model is constructed and coded in Matlab environment to simulate the exact queueing system.

In the simulation model, environment is as defined in 3.1. 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 ambulance from location r is assumed the sum of three exponentially distributed random variables as service time for this region (ν_q), travel time from ambulance location to region (t_{rq}) and travel time from region to ambulance location back (t_{rq}). Simulation is ended when steady state is reached. Steady state is searched for the objective function value in Equation 3.1 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 [4]. These bathes are treated as independent runs and used to construct a confidence interval.

In our study, steady state is assumed achieved by comparing mean of first batch

against 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 . Pseudo code of the procedure to construct a CI for the objective function value under the given solution using the simulation model is given in Algorithm 5.1. This procedure is defined to be sure that we are over warm up period 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 to be collected is set to 10 with 5000 demand calls in each of them.

Algorithm 5.1: Pseudo Code

1: Warm up : 30000 demand calls processed
2: Initialization : 50000 demand calls processed
3: Construct CI : 10 recent consecutive batches with 5000 demand calls
4: Check; if Mean of first batch $\in$ CI, go to Step 7. Otherwise, go to Step 5.
5: Iteration : 5000 more demand calls processed
6: Construct new CI : 10 recent consecutive batches with 5000 demand calls. Go to Step 4.
7: Take mean of CI as objective function value of the solution simulated and stop

5.3 Comparison of Alternative Downward Transition Rate Formulations and Levels of Order of Districting Combinations

In this part of the study, approximations in the SQM is questioned for different networks. For different levels of order of districting, errors in the objective function value calculated from performance measures of the simulated exact queueing system

is found. We aim to show order of districting decision is sensitive to network specific parameters in general networks. Alternative transition rate formulations are also tested in this part with order of districting decision.

Complete enumeration of solutions to the problem instances is made to find the optimal solution. Thereby, problem sizes are kept small due to computational time required to solve the problems with complete enumeration. Instances are generated as follows:

Number of demand region is set to 10. All of demand regions are taken as candidate ambulance locations. Number of ambulances located is changed from 1 to 7. As stated in Chapter 3, it is observed that increasing number of ambulances increases state space of the approximate queueing system exponentially and becomes intractable to solve. In an instance with uniform DoD and RoA 0.8, it takes more than 20 hours in real time on a computer with Intel Core i7 processor and 16 GB of RAM to make complete enumeration for 12 combinations of alternative rate formulation and order of districting levels. Concerning this outcome, a maximum of 7 ambulances are located in this study.

84 problem setting are generated with the combination of 6 parameters in the setting in Table 5.2. For every problem setting, 5 instances are generated. A total of 420 problem instances are solved for this part of the study.

Table 5.2: Levels of parameters for problem instances

Parameters	DoD	NoD	RoL	RoA	VoD	TI
Levels	Uniform	10	1.0	0.1	0.33	0.4
	Circular			0.2	3	0.6
				0.3		0.8
				0.4		
				0.5		
				0.6		
				0.7		

Order of districting levels are determined as 3,4 and 5. Every problem instance is solved for every combination of order of districting level and alternative rate formulation delivered in 3.3.3. In this part, optimal solution to the problem instances are found with complete enumeration. After finding the approximate evaluation of

the optimal solution, this solution is simulated to observe the performance measures of the exact queueing system. Exact evaluation of the objective function value of the optimum solution is calculated from the results of the simulation. Two objective function values, one from mathematical model and other from simulation model, is compared to find the effect of changing levels of order of districting and alternative rate formulations on approximation to the objective function value of the exact queueing system.

Mean absolute percent error (MAPE) of objective function value given by mathematical model for each combination of order of districting level and alternative rate formulations is calculated and reported across different parameters.

Table 5.3: MAPE across DoD

Alternative	Order of Districting Level	Uniformly	Circularly
I	3	0.35	0.24
	4	0.34	0.23
	5	0.34	0.23
II	3	0.39	0.32
	4	0.38	0.31
	5	0.38	0.31
III	3	0.13	0.08
	4	0.10	0.06
	5	0.10	0.05
IV	3	0.13	0.12
	4	0.13	0.12
	5	0.13	0.12

In Table 5.3, it is seen that downward transition rate formulation is more significant for the quality of approximation than order of districting. If the rate formulation is poorly performing, changing levels of order of districting cannot improve approximation quality. Among alternatives, Alternative III is performing best in circular DoD. Alternative III and IV perform equally better than others in uniform DoD. This can be concluded as follows: Alternative IV works with traffic intensities of subareas weighted according to the total demand of the subareas. When demand regions are dispersed (uniform DoD), this approach gets closer to working with weighted averages of service rates of individual demand regions in that transition (Alternative III). Quality of approximation with Alternative III is better in systems with circular DoD

than the ones with uniform DoD for all levels of order of districting. Other alternatives performs slightly better in circular DoD but not as significant as Alternative III. This is considered to result from problem instances with no feasible solutions for smaller RoA levels, with Uniform DoD. So, instances with RoA levels 0.1 and 0.2 is excluded from the analysis. It is seen that for Alternative II and IV difference in quality across DoD vanish while Alternative I and III are still better in Circular DoD. MAPE results without RoA 0.1 and 0.2 is given in Appendix A, Table A.1

For Alternative III, we see that increasing order of districting tends to increase quality of approximation by decreasing MAPE.

Every problem instance is solved for 12 different alternative formulation and order of districting level. We want to show that better performing combinations would consequently find better performing solutions among other combinations. To show this, fraction of time that an order of districting and rate formulation combination finds a solution giving the minimum (out of 12 results with different combinations) mean response time reported by simulation model for a problem instance is given in Table 5.4. In this performance measure, while comparing combinations, mean response time reported using simulation model is used. Objective function value calculated from approximate queueing system with different combinations are considered incomparable since objective function values for a given location solution calculated with different combinations can be different from each other.

Table 5.4: Fraction of Time Minimum Found across DoD

Alternative	Order of Districting Level	Uniformly	Circularly
I	3	0.28	0.23
	4	0.32	0.22
	5	0.34	0.22
II	3	0.21	0.21
	4	0.20	0.21
	5	0.20	0.21
III	3	0.36	0.61
	4	0.45	0.66
	5	0.52	0.66
IV	3	0.36	0.34
	4	0.46	0.33
	5	0.49	0.33

Experimental results in Table 5.4 is used to confirm that approximating the objective function value better increases the possibility of finding a better performing solution for the problem. It is seen that Alternative III is also better in finding the better performing solution among other alternatives in circular DoD and equal with Alternative IV in uniform DoD.

In Table 5.5, MAPE for different levels of RoA is given for combinations of order of districting and alternative formulations. In Table 5.6 , fraction of time minimum found with combinations is reported for levels of RoA.

Table 5.5: MAPE across RoA

Alternative	Order of Districting Level	0.1	0.2	0.3	0.4	0.5	0.6	0.7
I	3	0.01	0.16	0.22	0.29	0.34	0.41	0.47
	4	0.01	0.16	0.22	0.29	0.34	0.38	0.42
	5	0.01	0.16	0.22	0.29	0.34	0.38	0.42
II	3	0.01	0.10	0.22	0.33	0.44	0.53	0.60
	4	0.01	0.10	0.22	0.34	0.43	0.52	0.59
	5	0.01	0.10	0.22	0.34	0.43	0.51	0.58
III	3	0.01	0.03	0.05	0.04	0.08	0.17	0.27
	4	0.01	0.03	0.05	0.07	0.06	0.10	0.18
	5	0.01	0.03	0.05	0.07	0.09	0.08	0.13
IV	3	0.01	0.10	0.11	0.11	0.15	0.15	0.19
	4	0.01	0.10	0.11	0.13	0.15	0.14	0.17
	5	0.01	0.10	0.11	0.13	0.16	0.14	0.17

MAPE values under Alternative III are often better than MAPE under other alternative formulations. MAPE values for RoA 0.1 are ignorable. For RoA 0.1, resulting approximate system is actually the exact queueing system. So, for every combination formulation, dispatch fraction of this one ambulance ($\rho_{r_n q}$) to demand regions are simply equal to the fraction of demand (f_q) generated from these regions, resulting from the Equation 3.7. So, the MAPE values calculated for this RoA level result from the slight difference between objective function value of the model with the approximate queueing system and objective function value of simulation model, due to steady-state analysis of simulation. For other levels of RoA Alternative III always performs better. Increasing level of order of districting does not affect the quality in Alternative I, II and IV as much as in Alternative III. This result is also supported by

the MAPE values in Table 5.3.

Table 5.6: Fraction of Time Minimum Found across RoA

Alternative	Order of Districting Level	0.1	0.2	0.3	0.4	0.5	0.6	0.7
I	3	1.00	0.37	0.36	0.26	0.12	0.08	0.02
	4	1.00	0.37	0.36	0.28	0.13	0.07	0.07
	5	1.00	0.37	0.36	0.28	0.13	0.08	0.12
II	3	1.00	0.47	0.30	0.21	0.02	0.00	0.00
	4	1.00	0.47	0.30	0.16	0.02	0.00	0.00
	5	1.00	0.47	0.30	0.16	0.02	0.00	0.00
III	3	1.00	0.89	0.80	0.62	0.37	0.20	0.10
	4	1.00	0.89	0.80	0.48	0.60	0.43	0.17
	5	1.00	0.89	0.80	0.62	0.37	0.55	0.45
IV	3	1.00	0.47	0.56	0.21	0.27	0.20	0.17
	4	1.00	0.47	0.56	0.22	0.30	0.10	0.38
	5	1.00	0.47	0.56	0.22	0.32	0.18	0.55

Results in Table 5.6 support the observations for combinations from MAPE values. Alternative III and order of districting level 5 finds the minimum solution out of 12 combinations with a higher fraction than the others in most of the instances. For instances with higher number of ambulances on the other hand, Alternative IV and order of districting level 5 seems performing better. But, the MAPE values for this combinations is not better than Alternative III and order of districting level 5. So it is concluded that Alternative III and order of districting level 5 would be better in finding better performing solution in the long run.

It should be stated that quality of approximations for all combinations decreases with increasing number of ambulances.

In Table 5.7, MAPE values are reported for combinations and different levels of variance (VoD) in demand across regions. Percent of time finding the minimum response time among combinations is reported in Table 5.8.

Table 5.7: MAPE across VoD

Alternative	Order of Districting Level	Low	High
I	3	0.30	0.29
	4	0.28	0.28
	5	0.28	0.28
II	3	0.34	0.35
	4	0.34	0.35
	5	0.34	0.35
III	3	0.10	0.10
	4	0.08	0.08
	5	0.07	0.07
IV	3	0.13	0.12
	4	0.13	0.12
	5	0.13	0.12

Table 5.8: Fraction of Time Minimum Found across VoD

Alternative	Order of Districting Level	Low	High
I	3	0.24	0.26
	4	0.25	0.27
	5	0.26	0.29
II	3	0.23	0.19
	4	0.22	0.19
	5	0.22	0.19
III	3	0.53	0.47
	4	0.58	0.56
	5	0.64	0.55
IV	3	0.34	0.36
	4	0.38	0.38
	5	0.39	0.39

Alternative III again performs better than other alternatives for both levels of VoD. Increasing order of districting level does not change the quality of Alternative I,II and IV but of Alternative III , same as stated previously for other network parameters. Results in Table 5.8 also shows Alternative III is better in finding better performing solutions.

In Table 5.9 and Table 5.10, MAPE results and fraction of finding the solution with minimum mean response time for combinations is given across 3 different traffic intensity levels.

Table 5.9: MAPE across TI

Alternative	Order of Districting Level	Low	Medium	High
I	3	0.39	0.29	0.20
	4	0.39	0.27	0.19
	5	0.39	0.27	0.19
II	3	0.42	0.35	0.28
	4	0.42	0.34	0.28
	5	0.41	0.34	0.27
III	3	0.15	0.09	0.06
	4	0.12	0.07	0.04
	5	0.11	0.06	0.03
IV	3	0.18	0.12	0.07
	4	0.18	0.12	0.08
	5	0.18	0.13	0.08

Table 5.10: Fraction of Time Minimum Found across TI

Alternative	Order of Districting Level	Low	Medium	High
I	3	0.26	0.24	0.26
	4	0.22	0.25	0.31
	5	0.22	0.27	0.33
II	3	0.21	0.20	0.23
	4	0.20	0.19	0.22
	5	0.20	0.19	0.22
III	3	0.49	0.53	0.50
	4	0.56	0.56	0.59
	5	0.54	0.59	0.67
IV	3	0.38	0.33	0.35
	4	0.38	0.34	0.41
	5	0.40	0.34	0.43

Results in Table 5.9 and Table 5.10 shows that Alternative III with different order of districting levels performs better than any other combination. Quality of approximation increases with increasing level of traffic intensity.

Overall MAPE for alternative formulation and order of districting level combinations is reported in Table 5.11.

Table 5.11: Overall MAPE

Alternative	Order of Districting Level	Overall
I	3	0.29
	4	0.28
	5	0.28
II	3	0.35
	4	0.34
	5	0.34
III	3	0.10
	4	0.08
	5	0.07
IV	3	0.13
	4	0.12
	5	0.13

Regarding Table 5.11, and analysis of performance in the level of network parameters, Alternative III is seen as the best performing rate calculation formulation. It is seen that generating a transition rate formulation is not trivia. Best alternative at hand still tends to worsen by increasing the number of ambulances to be located and requires further study to find out whether it is resulting from rate formulation, from sub-area structure or from calculation of fractions of dispatches ($\rho_{r_n q}$).

Order of districting level is found significant for the quality of approximation to the exact queueing system. Increasing order of districting levels increases the quality of approximation in meaningful rate formulations.

In the rest of the study, rate formulation Alternative III and order of districting level 5 are considered to be used.

5.4 Design of Experiment for Genetic Algorithm

In this section, we set up a design of experiment for the genetic algorithm presented in Chapter 4 to see the effect of different algorithm parameters on finding the optimal solution to the problem. These parameters are determined as initial population size (S_{pop}), probability of crossover(P_c) and probability of mutation (P_m). For each parameter, two levels are defined as in Table

Table 5.12: DOE Setup

Parameter	S_{pop}	P_c	P_m
Levels	50	0.80	0.05
	100	0.90	0.10

Problem instances which are solved with complete enumeration in Section 5.3 are used. To have problem instances with higher number of feasible solutions, problem settings with RoA 0.5, 0.6 and 0.7 are selected. From resulting 36 different problem settings, only one instance is selected, meaning 36 different problem instances are solved using GA. Each problem instance is solved by starting GA five times. Percent of time that GA finds the optimal solution for a problem instance is calculated using the result of these five independent runs. This measure is reported across different network parameters in Table 5.13

Table 5.13: Fraction of Time Optimum Found with GA

S_{pop}	P_c	P_m	DoD		RoA			VoD		TI		
			Uni	Circ	0.5	0.6	0.7	0.33	3	0.4	0.6	0.8
50	0.80	0.05	0.63	0.70	0.63	0.80	0.57	0.71	0.62	0.62	0.70	0.68
		0.10	0.79	0.87	0.80	0.88	0.80	0.80	0.86	0.83	0.82	0.83
	0.90	0.05	0.66	0.80	0.67	0.85	0.67	0.70	0.76	0.70	0.72	0.77
		0.10	0.81	0.86	0.83	0.87	0.80	0.84	0.82	0.82	0.87	0.82
100	0.80	0.05	0.92	0.93	0.87	0.98	0.93	0.96	0.90	0.97	0.97	0.85
		0.10	0.97	0.92	0.92	0.95	0.97	0.94	0.94	0.97	0.98	0.88
	0.90	0.05	0.96	0.91	0.90	0.98	0.92	0.97	0.90	0.98	0.93	0.88
		0.10	0.93	0.90	0.92	0.95	0.88	0.92	0.91	0.97	0.90	0.88

To show the overall performance, fraction of time (OF) that GA found optimum solution in all problem instances, fraction of time (BF) that GA found a solution performing better than the optimum solution (regarding simulation mean with CI), average percent of the difference (AW) between optimum solution and worser solutions in the five runs, maximum percent of the difference (MW) between optimum solution and worst solution in the five runs, maximum percent of the difference (MB) between optimum solution and the solution better than optimal solution in the five runs is reported in Table 5.14.

In the fitness value calculations of GA, mathematical model including approximate queueing system is used. This is the reason why GA could converge to better perform-

ing solution than the optimal solution given by complete enumeration of mathematical model.

Table 5.14: Overall Performance

S_{pop}	P_c	P_m	OF	BF	AW(%)	MW(%)	MB(%)
50	0.80	0.05	0.67	0.11	1.7	4.3	4.4
		0.10	0.83	0.05	1.4	3.7	4.4
	0.90	0.05	0.73	0.07	1.5	6.4	4.4
		0.10	0.83	0.05	1.8	6.1	4.4
100	0.80	0.05	0.93	0.01	1.2	2.9	4.2
		0.10	0.94	0.01	0.4	0.8	0.4
	0.90	0.05	0.93	0.02	1.9	5.3	3.5
		0.10	0.92	0.03	1.0	1.6	4.2

From the results in Table 5.14, it is seen that GA finds optimal or a better performing solution (OF + BF) more likely with ($S_{pop} = 100$, $P_c = 0.80$, $P_m = 0.10$), ($S_{pop} = 100$, $P_c = 0.90$, $P_m = 0.05$) and ($S_{pop} = 100$, $P_c = 0.90$, $P_m = 0.10$). In addition to likelihood of finding optimal or better solutions, AW and MW values are considered. It is seen that these values are smaller for $S_{pop} = 100$, $P_c = 0.80$, $P_m = 0.10$. We also consider fractions in Table 5.13 by checking the minimum fraction among levels of a network variable. For $S_{pop} = 100$, $P_c = 0.80$, $P_m = 0.10$, minimum for DoD, RoA, VoD and TI occurs as 0.92, 0.92, 0.94 and 0.88 respectively. This values are equal to or better than the minimums with $S_{pop} = 100$, $P_c = 0.90$, $P_m = 0.05$ or $S_{pop} = 100$, $P_c = 0.90$, $P_m = 0.10$.

Hence, parameters are set as $S_{pop} = 100$, $P_c = 0.80$, $P_m = 0.10$ for the GA.

5.5 Comparison of Single and Multi Server Model

In this section, effect of allowing more than one ambulance at a single location is explored. Mathematical model given in Chapter 3 is model as allowing multiple ambulances per each location and referred as Multi Server Model (MSM). For this part of the study, a Single Server Model (SSM) is constructed by adding the following

constraint in Equation 5.1 to the model.

$$x_i \in \{0, 1\}, \forall i \in I \tag{5.1}$$

Crossover and mutation operations of GA is changed to be able to reproduce single server solutions. For crossover operation, an exchange vector for each parent which showing the genes only exist in the other parent is constructed. One point crossover is applied to each parent with corresponding exchange vector. This procedure guarantees having single server solutions in the mating pool. For mutation operation, a mutation pool with locations non existent in the current child is constructed. If mutation occurs for a gene, previous location in the gene that is mutated now is added to the mutation pool, allowing it appears in the next genes of the child by mutation.

Using GA algorithm with the parameters set in Section 5.4, following problem instances are solved.

Number of demand regions (NoD) in the problem instances is determined as 20. 48 different problem settings are constructed with the combination of network parameters levels in Table 5.15. For each setting, 5 problem instances are generated resulting in 240 problem instances. Each problem instance is solved by restarting GA 5 times with MSM and 5 times for SSM.

Table 5.15: Levels of Parameters for Problem Instances

Parameters	DoD	NoD	RoL	RoA	VoD	TI
Levels	Uniformly	20	0.3	0.2	0.33	0.4
	Circularly		0.4	0.3	3	0.6
						0.8

Each instance has 5 independent solutions with MSM and 5 with SSM. In 33 instances out of 240, GA could not find a feasible solution in the random search to start the iterations. These are instances with uniform DoD and most of them with RoA 0.2, which makes finding solutions satisfying coverage constraint harder. These problem instances are discarded.

It is not possible to compare MSM and SSM solutions one by one. Average of mean response time of solutions are found for MSM results and SSM results of each in-

stance. Then, the fraction of difference (PD) between them is calculated by taking SSM average as base. Positive levels of PD implies that mean response time would be improved if MSM is used to locate ambulances. Mean response time values are taken from simulation study for this measure. The solution handed by GA is simulated and the mean response time is used in the calculation of PD.

Another performance measure as fraction of time that GA with MSM gives a solution with multiple ambulances at least in one locations (FQ) is calculated. While PD shows the amount of improvement obtained from relaxing single server restriction, FQ gives information about frequency of solutions with multiple ambulances at least in one location.

These performance measures are reported across network parameters in Table 5.16.

Table 5.16: Comparison of MSM and SSM solutions

	DoD		ROL		RoA		VoD		TI			Overall
Levels	Uni	Circ	0.3	0.4	0.2	0.3	0.33	3	0.4	0.6	0.8	
PD	0.03	0.13	0.12	0.06	0.06	0.12	0.09	0.09	0.07	0.10	0.10	0.09
FQ	0.57	0.96	0.83	0.76	0.60	0.97	0.77	0.83	0.76	0.82	0.81	0.80

MSM improves mean response time in Circular DoD more than in Uniform. It also gives more solutions with multi servers at least in one location in Circular DoD. In Circular DoD, model locates ambulances to areas with high demand accumulation in multiple numbers rather than distributing them one by one which increases mean response time of the system.

Candidate ambulance location number generated from RoL also affects the improvements with MSM. As the number of candidate locations increases, number of feasible solutions increases. It is seen that FQ decreases with increasing RoL level. But still MSM gives solutions with smaller mean response time (positive PD) in both levels of RoL.

Number of ambulances to be dispatched (computed from RoA) is another factor affecting locating multiple ambulances. As RoA increases, both PD and FQ increases. For different levels of VoD, PD and FQ stay fairly stable. This is because locating multiple ambulances is more related to high demand around an area than in an indi-

vidual region. There is still a slight increases in FQ with increasing levels of VoD. If some demand region generate very significant fraction of demand in the system (resulting from generating problem instances with higher levels of VoD than 3), then it will be more likely MSM will locate multiple ambulances in nearest candidate location to that region. These type of regions would accumulate enough demand which would increase FQ.

Different TI levels do not change PD and FQ to much extent. Since service times are taken equal among regions, changing the traffic intensity does not create significant difference in PD or FQ.

Overall, allowing multiple ambulances per location outperforms SSM in mean response time in general networks. It is also seen that among different network parameters with different levels, MSM locates multiple ambulances at least in one location, in more than 50% of the instances.

CHAPTER 6

CONCLUSION

In this study, an EMS location problem is studied for locating ambulances. Previous studies on locating EMS vehicles are reviewed. A mathematical model using queueing theory from the literature is employed. Approximate queueing model by Geroliminis et al. [10] is used. This model is analyzed for quality of approximations in general networks. In the approximate queueing model, generic service rate formulations for state transitions are developed and tested. A relaxation to the model of Geroliminis et al is proposed and reported for its effect on mean response time of the system. Discrete event simulation is used to observe the performance measures of exact queueing system which is infeasible to be solved analytically due to exponentially increasing size of the state space, more rapid than approximate queueing system. Mathematical model proposed uses queueing theory to calculate the objective function value, for which no closed form expression exists. A metaheuristic algorithm (Genetic Algorithm) is proposed to solve the problem. Performance of this algorithm is questioned and reported.

It is showed that developing a generic formulation for service rates in the approximate queueing model is not trivial. Four different formulations are proposed. One of the alternative formulations gives promising results. However, quality of this formulation is not determined by comparing objective function value from simulation of every feasible solution to the problem, with objective function value of mathematical model. Only optimal solutions are compared with their simulation results. This gives us an initial idea about the performance of the alternatives but it requires further study to fully understand it.

Analysis of alternative service rate formulations is made based on the objective function value. Approximations of dispatch frequencies can also be analyzed against frequencies in simulation of exact system. This would give us the errors in individual terms, which are summed with a multiplier (mean travel time) in the objective function. Steady state probability of having all ambulances busy can also be compared to the one in simulation study. Only the steady state probability of this state can be comparable since we need again approximation to compare another state of the approximate system which are represented by multiple number of states in exact system.

Order of districting decision is found significant to the quality of approximation. Geroliminis et al [10] uses third order of districting in a planar network but it is showed that higher order of districting increases the quality in general networks. Yet, we do not know where quality of approximation becomes insensitive to increasing order of districting level, from the experimental results with ambulances up to 7. It is required to solve problem instances with higher number of ambulances than 7 and higher order of districting. However a weakness of the mathematical model is becoming inefficient to solve computationally by increasing number of ambulances as mentioned in Chapter 5, Section 5.3. As a further study, a partitioning of the problem into subproblems can be explored by decreasing the state space of the systems that show up in the subproblems.

Allowing multiple ambulances at a single location has a positive effect on the objective function value. We know that after relaxing single server restriction, we can still come up with results which at most one ambulance is located in each location in the solution. In the computational study, it is also shown that multiple ambulances at a single location are preferred with significant percentages and improved objective function value. This implies that it is worth modeling the problem without single server restriction in general networks, as long as it is not a constraint by the decision maker.

A weakness of the mathematical model proposed is taking only mean response time into account. When we work with queueing theory, other performance measures such as utilization of ambulances or fraction of lost demand is inherited in the problem.

These performance measures can also be used as objective function values in a single objective manner, or as constraints to the problem, or as multiple objectives for the problem. This can be an extension resulting a problem formulation with more solid ground for EMS systems.

REFERENCES

[1] Scott A.J. Dynamic location-allocation systems: some basic planning strategies. *Environment and Planning*, 3(1):73–82, 1971.

[2] Iannoni A.P. and Morabito R. 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, 2007.

[3] ReVelle C. and Hogan K. The maximum availability location problem. *Transportation Science*, 23(3):192–200, 1989.

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

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

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

[7] Mendonca F.C. and Morabiot R. Analysing emergency medical service ambulance deployment on a brazilian highway using the hypercube model. *Journal of the Operational Research Society*, 52:261–268, 2001.

[8] Gendreau M. Laporte G. and Semet F. A dynamic model and parallel tabu search heuristic for real-time ambulanca relocation. *Environment and Planning*, 3(1):73–82, 1971.

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

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

[11] Halpern J. The accuracy of estimates of the performance criteria in certain emergency service queueing systems. *Transportation Science*, 11(3):223–241, 1977.

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

[13] Brandeau M. and Larson R.C. Extending and applying the hypercube queueing model to deploy ambulances in boston. *TIMS Studies in the Management Sciences*, 22:121–153, 1986.

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

[15] Daskin M.S. and Stern E.H. A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152, 1981.

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

[17] Church R. and ReVelle C. S. The maximal covering location problem. *Papers of the Regional Science Association*, 32:101–118, 1974.

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

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

[20] Volz R.A. Optimum ambulance location in semi-rural areas. *Transportation Science*, 5(2):193–203, 1971.

[21] Larson R.C. A hypercube queueing model for facility location and redistricting in urban facility service. *Computers & Operations Research 1*, 1:67–95, 1974.

[22] Ballou R.H. Dynamic warehouse location analysis. *Journal of Marketing Research*, 5(3):271–276, 1968.

[23] Degel D. Wiesche L. Rachuba S. and Werners B. Time-dependent ambulance allocation considerig data-driven empricially required coverage. *Health Care Management Science*, pages 1–15, 2004.

[24] ReVelle C. S. and Swain R. W. Central facilities location. *Geographical Analysis*, 2:30–42, 1970.

[25] Toregas C. Swain R. W. ReVelle C. S. and Bergman L. The location of emergency service facilities. *Operations Research,*, 19:1363–1373, 1971.

[26] Hakimi S.L. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 13:462–475, 1964.

[27] Hakimi S.L. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13(3):450–459, 1965.

[28] Sacks S.R. and Grief S. Orlando police department uses or/ms methodology new software to design patrol district. *OR/MS Today*, pages 30–42, 1994.

[29] Marianov V. and Revelle C. The queueing maximal avaliability location problem: a model for the siting of emergency vehicles. *Eurpoean Journal of Operational Research*, 93(1):110–120, 1996.

[30] Department of Health Washington, DC. What is ems? http://doh.dc.gov/service/what-ems, last access date:27.09.2015.

APPENDIX A

Table A.1: MAPE across DoD without problem instances with RoA 0.1 & 0.2

Alternative	Order of Disticting Level	Uniformly	Circularly
I	3	0.38	0.31
	4	0.36	0.30
	5	0.36	0.30
II	3	0.42	0.42
	4	0.42	0.42
	5	0.42	0.42
III	3	0.14	0.11
	4	0.11	0.07
	5	0.10	0.06
IV	3	0.14	0.15
	4	0.13	0.15
	5	0.14	0.15