DATA DRIVEN APPROACHES TO IMPROVE OPERATIONAL EFFICIENCY OF EMERGENCY MEDICAL SERVICES

BY

KAUSHIK KRISHNAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Industrial Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Assistant Professor Lavanya Marla

Abstract

We study data-driven approaches to maximize the service level of Emergency Medical Services (EMS) in emerging economies. These systems usually operate under heavy resource constraints and face significant operational challenges, making them structurally and operationally different from systems in developed countries. In this thesis we study two specific issues - (i) modeling human behavior, and (ii) accounting for risk metrics due to tail behavior.

First, we address the issue of ambulance abandonment that occurs when a patient's willingness to wait is less than the ambulance response time resulting in the vehicle not being utilized. We present a maximum likelihood estimation approach to estimate willingness to wait for different types of patients. We then use the estimate of waiting times in a greedy simulation based optimization model to redesign the EMS network to maximize the number of patients served within their waiting time thresholds. Computational experiments using data from an Indian metropolitan city show that our proposed resource allocation model reduces abandonment by approximately 2 percentage points with the current ambulance fleet size, 5 percentage points by doubling the fleet size and 6 percentage points by tripling the fleet size.

Next, we present a risk-based optimization approach to make the EMS network robust to unexpected changes in demand patterns. This is motivated by the fact that when few parts of the network face heavy-tailed demand patterns, the demand for entire network under the resource constrained setting behaves in a heavy-tailed manner. To achieve a robust location strategy we include risk metrics, specifically the Conditional Value at Risk, that focus on tail behavior in addition to average case performance metrics. Computational experiments show that planning with a view of minimizing risk leads to solutions that perform well in heavy-tailed settings.

To Mom, Dad and Deepak

Acknowledgments

I extend my sincere thanks to my advisor, Professor Lavanya Marla, for giving me the opportunity to work on challenging research problems during my graduate studies. This thesis would not have been possible without her motivation and support for me at all times. I would also like to thank her for showing faith in myself by encouraging me to present our work in two prestigious conferences within a year and a half of beginning my studies.

I thank Professor James Leake, Professor Ramavarapu Sreenivas and Ms. Holly Kizer along with my advisor for ensuring that I had financial support for the entire course of my studies.

Finally, I thank Mom, Dad and Deepak for having been an immense pillar of support for me at all times.

Table of Contents

List of Abbreviations

EMS	Emergency Medical Services
MLE	Maximum Likelihood Estimation
NPMLE	Non Parametric Maximum Likelihood Estimation
CVaR	Conditional Value at Risk

Chapter 1

Introduction

Emergency Medical Services form a crucial component of public transportation infrastructure, and involve immediate response to critical events. The problem of timely EMS response is important to developing countries where high population densities and heavy traffic conditions impose strategic and operational challenges, causing travel (and hence response) times to be high. The critical nature of the events and limited resource availability warrants the use of sophisticated analytical tools for the design and operation of Emergency Medical Systems. Problems that have been extensively studied in literature include ambulance base positioning, fleet allocation and dispatch policies.

In this thesis, we focus on two problems primarily motivated by emerging economies. The first issue is that of customer abandonment of ambulances. The motivation to study this aspect arises due to the fact that we observe abandonment of ambulances by patients in the system. We hypothesize that there is a threshold for the time a patient is willing to wait for service. When an ambulance does not respond to a patient within the patient's waiting threshold, the patient chooses to abandon the ambulance and resort to other ways to reach the hospital. We refer to this phenomenon as in-service ambulance abandonment. Abandonment decreases the efficiency of the system as ambulances spend time in travel without serving patients. We therefore study policy measures to improve system efficiency.

The second question we study in this thesis is that of incorporating risk metrics in ambulance allocation decisions. Typically, allocation decisions of ambulance location and redeployment are made primarily based on expected value metrics but not on other metrics that account for risk. It is, however, well-known that in stochastic systems, it is useful to consider risk metrics in addition to expected-value metrics. We address this as the second issue in this thesis.

1.1 Outline of the thesis

The thesis is structured as follows.

- In chapter 2, we describe the setting of abandonment by customers that is faced by operators in emerging economy settings. We present it as a question of first estimating the willingness to wait of customers, and then that of network design by the operator to decrease the inefficiencies due to customer abandonment. We begin by describing the data set and providing summary statistics for the important variables. Then we explain the concept of *interval censored data* (Huang and Wellner, 1997) and describe models for estimation of such data. We apply the Proportional Hazards model (Cox, 1992) to estimate patient waiting times. Then we describe a simulation-based greedy optimization model for the emergency medical resource allocation problem. We perform experiments pertaining to existing resource optimization, resource expansion and network redesign, and demonstrate that we can obtain the highest benefits in the system using network redesign.

- In chapter 3, we introduce the notion of optimization using risk-based metrics. Specifically, we describe minimizing the total coverage loss defined by a linear combination of expected value of loss and the Conditional-Value-at-Risk of that loss to achieve a robust allocation of ambulances across bases. We use a modified version of the data from (Yue et al., 2012a) for this chapter. This chapter has already been published as (Krishnan et al., 2016) . We demonstrate that the robust allocation has lower levels of cascading effects resulting in higher service levels, than an allocation based on expected loss metrics.

Chapter 2

Network Design to Mitigate In-Service Ambulance Abandonment

We consider a setting with novel customer behavior that, to our knowledge, has not been studied in EMS systems. Our problem is motivated by settings in emerging economies such as India and South Africa, where ambulance systems are in early stages of development. These systems experience high resource constraints on ambulance fleet sizes and heavy road traffic conditions, causing travel (and consequently, response) times to be large. Patients who call for ambulances during emergencies may typically have a limit on the willingness to wait, which becomes apparent under these resource constraints; and results in *in-service abandonment* of ambulances.

We define in-service abandonment in the EMS setting as follows. We hypothesize that patients calling the call center have a ceiling on the amount of time they are willing to wait for the ambulance to arrive at the scene of the incident. If an ambulance is available and is dispatched to serve the caller, the caller is often provided an estimated time of arrival. However, upon arrival at the caller's location, the ambulance may find that the caller has left the scene by another mode of travel. This can happen when the caller's willingness to wait exceeds the base-to-scene travel time of the dispatched ambulance. This leads to the vehicles that are dispatched to this call not being used for service, and potentially, other calls being adversely affected as the abandoned ambulance may be better employed elsewhere.

Abandonment occurs in many queueing systems, such as those in call centers, amusement parks and emergency rooms. This has been studied from the queueing perspective, both analytically and empirically. (Gans et al., 2003) provides a survey of the analytical research on call center modeling, including abandonment. More recently, empirical studies with real-world data on abandonment describe the behavior of abandoning callers. These empirical studies have focused on abandonment in call centers (Aksin et al., 2013, Hathaway et al., 2017, Yu et al., 2017, Aksin et al., 2017), and more

recently, emergency departments (Dong et al., 2017, Batt and Terwiesch, 2015).

The setting here is distinctly different. First, here, there is loss in server (ambulance) utilization despite a server being assigned, because the server spends time being busy until it discovers the patient has abandoned. Second, a caller's willingness to wait is *censored*, as the operator cannot empirically observe when a patient left service. In fact, an ambulance arriving at the scene observes either that (i) the caller is waiting at the scene, indicating the willingness to wait is greater than the call-to-scene time, or (ii) the caller is no longer at the scene and has abandoned the ambulance, indicating that the willingness to wait is less then the call-to-scene time.

Several questions arise in understanding and managing such systems, from both the supplier (operator's) perspective and the demand (caller's) perspective. From the callers' perspective, abandonment could be due to impatience caused by the situation or a more fundamental lack of confidence in the system. From the operator's perspective, can we find strategies and policies that take abandonment into account, and identify if the resource-constraints in the system are the primary drivers of ambulance abandonment? If resource-constraints were increased to the level of those in more developed countries, would abandonment decrease?

We therefore divide this into five sets of research explorations. The first is to understand the human behavior underlying abandonment and see what factors may influence such behavior. Second is to estimate the callers' willingness to wait, depending on the relevant factors identified. The third question, is to find if the operator can potentially allocate ambulances differently in the current system to maximize the number of callers that can be served (fraction of callers to whom ambulance is sent and is not abandoned). If the number of ambulances at these bases were increased, what happens to abandonment behavior? The fourth relates to understanding if ambulance dispatch policies can designed to account for abandonment, by modifying the typical nearest-free-ambulance dispatch policy. For example, we examine selective (hypothetical) dispatch policies that do not serve customers that are most likely to abandon. The fifth question relates to identifying the need for, and re-designing the network formed by the ambulance configuration in the system to decrease abandonment.

Contributions

We use empirical data from an operator in India to identify, measure and optimize for the phenomenon of abandonment. Historical call logs available from the operator are used to understand the spatio-temporal processes of call arrivals and customer behavior. We build a discrete-event simulation of the system to evaluate the impact of measures proposed, as well as combine it wih heuristics in a simulation-optimization framework for resource allocation decisions.

Our first contribution is to identify the factors affecting callers' abandonment rates in the EMS system of interest. We identify key factors as the call-to-scene travel time of the dispatched ambulance, the location of the caller (semi-urban or urban area) and the severity of the emergency.

Second, we build a semi-parametric model of the waiting times of the patients to estimate their willingness to wait based on the influencing factors identified in the data. We build survival curves for each customer group to understand their abandonment behavior and the probability of abandonment relative to the service (call-to-scene) time.

Third, we establish that re-allocation of ambulances at the current set of bases, provides very small improvement. With the same fleet size of ambulances as currently, there is a minimal change in the percentage of calls not served (either no ambulance was available or caller abandoned the ambulance). Surprisingly, we show that this result does not change significantly even when the fleet size is increased considerably.

Fourth, we introduce a notion of 'selective dispatch', where ambulances are dispatched to callers only if the abandonment probability is less then a threshold value. We show that given the design of the system, selective dispatch has no benefit.

Fifth, we consider the case of a major strucural re-design of the system by considering (approximately) every 'street corner' as a posible base and choosing the best set of bases to locae ambulances at, for varying budgets. This results in a huge combinatorial optimization problem, about two orders of magnitude greater than allocation problems studied in the EMS literature. To address this, we use a heuristic developed in the literature to identify bases and allocate ambulances in a near-optimal manner. Using this approach, we demonstrate that re-structing the base locations can decrease abandonment

can drop by 2 percentage points with the current fleet size, by 5 percentage points by doubling the fleet size and 6 percentage points using three times the current fleet size (when the ambulance fleet size comparable to developed economies).

2.1 Data Description

As this is a data-driven study we describe the data set in detail. The study is set in a large Indian metropolitan city having a population of 7.7 million people. There are 58 existing ambulance bases distributed across the city and the surrounding suburbs. With a baseline allocation of one ambulance placed at every base, the city has an ambulance-population ratio one ambulance for every 128,000 people. To provide some context, the city of Chicago in the United States has a total of 65 ambulances [1] for a population of 2.7 Million people [2], and hence an ambulance-population ratio in our data set of one ambulance per 36,000 people. Hence, the ambulance-population ratio is less than one-thirds of the that in the United States. Therefore in this extremely resource constrained setting, it becomes very important to understand customer responses to service, characterize abandonment and use ambulances efficiently. We wish to see if efficiency of the EMS system can be improved with the existing resources. If the current resources seem inadequate, we wish to determine how many additional resources are required to match EMS service levels of those in developed economies.

We have data on 18,525 emergency calls for the month of November 2011 from the city and surrounding semi-urban areas. Corresponding to every call in the log, we have geographical and operational data. Geographical data consists of the location of the caller, the city, district, sub-city district, incident landmark and the nearest ambulance that can serve that call. Operational data consists of the call arrival time, type of emergency, triage information, patient characteristics, ambulance dispatch time, arrival time of ambulance at the scene, emergency type and timestamps of various touch-points along the way. Of all the calls recorded, 9599 unique calls needed an ambulance to be dispatched to, and are considered for our analysis.

[1] https://www.cityofchicago.org/city/en/depts/cfd/provdrs/ops.html
[2] https://www.census.gov/quickfacts/fact/table/chicagocityillinois

We discuss a few aspects in the data that are crucial for our analysis. Of all the emergency calls received, 78.5% of all the calls were from urban areas. The lower number of calls from semi-urban areas is in accordance with the lesser population in those areas. *Call-to-scene time* or *response time* is the time elapsed between the time when the emergency call was made and the time when the ambulance reached the scene. We will use these two terms interchangeably in this thesis. The average call-to-scene time in our data set is 21.5 minutes. As shown in Figure 2.1, the average call-to-scene times in semi-urban areas is 10 minutes higher than those in the urban areas. This can be attributed to higher population density, more number of resources and better transportation infrastructure in the urban areas compared to the semi-urban areas.

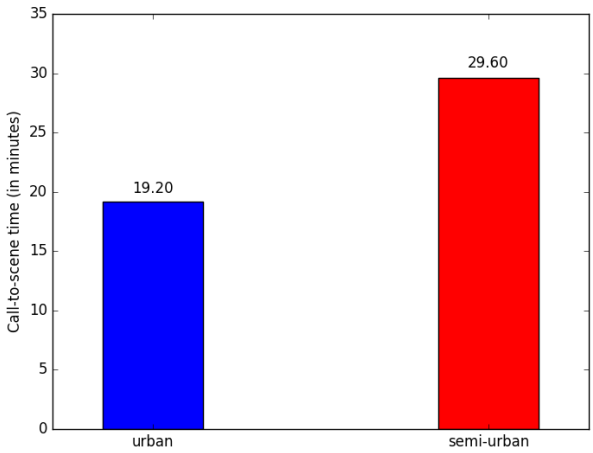

Figure 2.1: call-to-scene time by caller location

The operator classifies a call as 'abandoned' if, upon reaching the scene of the caller, the patient is found to have already moved by other means. During this period, 83.42% of all dispatched ambulances were utilized and not abandoned. 16.58% of dispatched ambulances were abandoned.

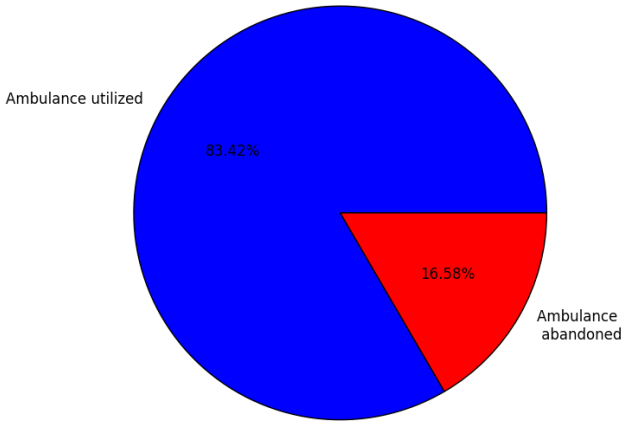

Figure 2.2: Utilization of dispatched vehicles by patients

The share of abandoned calls by location of the caller is shown in Figure 2.3.

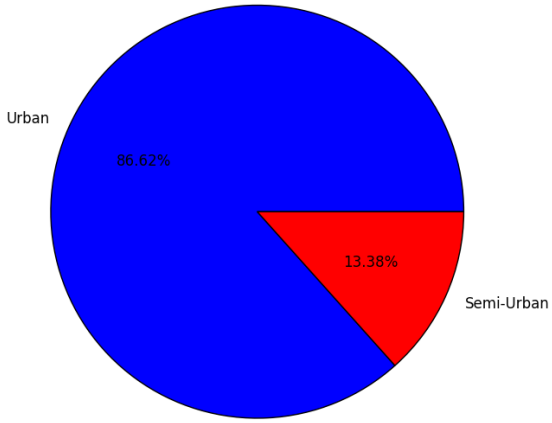

Figure 2.3: Abandonment by urban and semi-urban callers

We note that the higher proportion of abandoned calls are from urban areas in spite of lower call-to-scene times. This implies a higher tendency to wait for semi-urban callers in comparison with urban callers.

Calls are also classified based on the description of the emergency. The reason for every emergency call is listed in the data and we use the Emergency Severity Index (ESI Index) classification from the Agency of Healthcare Research and Quality of the U.S. Department of Health and Human Services

(Gilboy et al., 2012) to triage the calls based on emergency type. (Gilboy et al., 2012) outlines the conditions, based on the reasons for emergency calls, to triage calls into ESI levels 1 through 5, where ESI level 1 is the most severe and ESI level 5 is the least severe. We also consider if a call requires an Advanced Life Support (ALS) ambulance or a Basic Life Support (BLS) ambulance [3]. The description of the various emergency types and the frequencies of their occurrences in the data set are detailed in Table 2.1. While the operator in the systems operates with a homogeneous fleet of ambulances, the ALS and BLS classifications are merely used as a proxy for the severity of the incident.

Emergency Type	Basis of Classification	% Occurences
1	Require ALS and ESI Levels 1 or 2	95.01%
2	Require BLS or other ESI Levels	4.99%

Table 2.1: Emergency Type Classification

2.2 Model Formulation - Estimating Patient Waiting Times

We hypothesize that patients have a ceiling on the time they are willing to wait for service after they have made an emergency call and after they become aware that an ambulance has been dispatched to serve them. We refer to this period of time as the *waiting time* of a patient. Ambulance abandonment results if the patient's waiting time is lesser than the corresponding call-to-scene time. The patients' waiting times are not directly observed and they need to be estimated. We now describe in detail the procedure for waiting time estimation.

2.2.1 Interval censored data

As we had described in the beginning of this chapter, abandonment by a patient cannot be observed until an ambulance reaches the spot to serve a

[3]https://www.nemsis.org/v2/downloads/documents/NEMSIS_Data_Dictionary_v2.2.pdf

patient. If, at the time of ambulance arrival, it is observed that the patient has already left the spot then the ambulance has been abandoned. Else if the patient is still awaiting service, then the ambulance has not yet been abandoned and could have been abandoned if the ambulance had arrived at a later time.

Let us assume that the time taken for an ambulance to reach the patient, or the call-to-scene time, is t. If the patient had abandoned the ambulance, we know that the time of abandonment lies in the interval $[0, t)$. If the patient had not abandoned the ambulance, we can only say that the time of abandonment will lie in the interval $[t, \infty)$. Hence the waiting time data is truncated or 'censored' in the intervals described above. This type of censored data is called *type I interval censored data* or *current status data* (Huang and Wellner, 1997).

We now discuss the special case when patients are served within their waiting thresholds and hence did not abandon the ambulance. We defined that their waiting times lie in the interval $[t, \infty)$. According to this assumption, extremely high values for waiting times can also be considered as feasible. However in reality, this is not a sound assumption to make as patients calling for emergency services do not wait indefinitely. Any model we use for estimating the waiting times will consider a feasibility range of $[t, \infty)$ for those patients whose abandonment was not observed. Hence we define an upper bound or ceiling on the waiting time of every patient so that we make an assumption which is more realistic. We now define this mathematically.

Let $i = 1, 2...n$ denote the number of patients. The corresponding call-to-scene times are denoted by t_i and waiting times are denoted by w_i. The corresponding upper bound on the waiting time of every patient is denoted by u_i. Let δ_i where

$$\delta_i = \begin{cases} 1, \text{if patient } i \text{ had abandoned the ambulance} \\ 0, \text{if patient } i \text{ had not abandoned the ambulance} \end{cases} \tag{2.1}$$

denote the abandonment indicator for every patient.

The waiting time intervals are now defined as

$$w_i \in \begin{cases} [0, t_i), \forall i \in \{i | \delta_i = 1\} \\ [t_i, u_i], \forall i \in \{i | \delta_i = 0\} \end{cases} \tag{2.2}$$

This type of interval censoring where the intervals are finitely bounded is defined as *bivariate interval censored* data or *type 2 interval censored* data (Huang and Wellner, 1997). We now discuss the estimation procedures for interval censored data in detail.

2.2.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is the process of finding the set of parameters of a statistical model given a data set, that maximizes the *likelihood* of observing the data assuming that the phenomena occur as described by the statistical model. In this context, we assume that the waiting times of patients follow a certain probability distribution function, and aim to estimate the parameters of the underlying distribution. We introduce some notation to describe maximum likelihood estimation in detail.

Let W denote the random variable representing the waiting times of patients and $f(W)$ be the corresponding density function representing the patient waiting times. For $i = 1, 2...n$, where n is the number of patients, let L_i and E_i denote the corresponding location and emergency type of caller i. We mathematically define these variables as follows:

$$L_i = \begin{cases} 1, \text{if patient } i \text{ had called from an urban area} \\ 0, \text{if patient } i \text{ had called from an semi-urban area} \end{cases} \tag{2.3}$$

$$E_i = \begin{cases} 1, \text{if patient } i \text{ had called for a Type 1 emergency} \\ 2, \text{if patient } i \text{ had called for a Type 2 emergency} \end{cases} \tag{2.4}$$

The description for the two types of emergencies had been outlined in Table 2.1.

We are also interested in the hazard rate $\lambda(W)$ and survival function $S(W)$ of the distribution. Hazard rate is defined in this context as the probability of instantaneous abandonment or in other words, the likelihood that a patient will abandon an ambulance in the immediate future given that they have waited until time T. Mathematically, it can be written as

$$\lambda(t) = P(W \in (t, t + dt)|W \geq t) \tag{2.5}$$

Hazard rate and abandonment behavior

The hazard rate of the underlying probability distribution of the patient waiting times can describe the patient waiting behavior. If

$$\lambda(t_2) \geq \lambda(t_1), \quad t_2 \geq t_1 \tag{2.6}$$

then the hazard rate is increasing with time. This implies that the probability of abandonment increases when the call-to-scene time increases. On the contrary, if

$$\lambda(t_2) \leq \lambda(t_1), \quad t_2 \leq t_1 \tag{2.7}$$

then the hazard rate is decreasing with time. This implies that the probability of abandonment decreases when the call-to-scene time increases.

An increasing hazard rate is more intuitive as one would assume that the tendency to become impatient would increase with time. A decreasing hazard rate is counter-intuitive in this scenario.

Survival Probability

The survival probability at time t for a patient is defined as the probability that the patient would have waited for a period of time greater than t.

$$S(t) = P(W > t) \tag{2.8}$$

The survival function at time t $S(t)$ is the complement of the cumulative distribution function $F(t)$ at time t.

$$S(t) = 1 - P(W \leq t) = 1 - F(t) \tag{2.9}$$

We now outline various maximum likelihood estimation procedures for estimating waiting time distributions using interval censored data.

Parametric Maximum Likelihood Estimation

Parametric MLE assumes that the data follows a specific parametric probability distribution (e.g. Exponential, Weibull distributions) and is aimed at

estimating the parameters that define the assumed distribution. Mathematically, it can be detailed as follows.

If a call has been abandoned ($\delta_i = 1$) then $w_i \leq t_i$. The contribution of this observation to the likelihood function is $P(W \leq t_i) = F(t_i)$. If a call has not been abandoned ($\delta_i = 0$) then $w_i > t_i$ and the contribution of this observation to the likelihood function is $P(W > t_i) = 1 - F(t_i) = S(t_i)$, the survival function. Let θ be the set of parameters in the distribution that need to be estimated.

Hence the Likelihood function is

$$L(\theta|T) = \prod_{i=1}^{n} [F(t_i)^{\delta_i} * S(t_i)^{(1-\delta_i)}] \tag{2.10}$$

Maximizing the logarithm of eq. (2.10) also known as the *log-likelihood function* will provide an estimate of the parameter vector θ. The log-likelihood function is defined as:

$$\log L(\theta|T) = \sum_{i=1}^{n} [\delta_i \log(F(t_i)) + (1 - \delta_i) \log(S(t_i))] \tag{2.11}$$

This technique requires that we estimate the probability distributions separately for all four classes of patients. If our model has more categories of patients, then this procedure of estimating a distribution for every category separately becomes inefficient. Also estimating the distributions separately would render it difficult to understand how the distribution of one class of patients compares to another distribution. To include patient specific *covariates* or explanatory variables in our model, we use a semi-parametric maximum likelihood estimation approach.

MLE with covariates - Semi-parametric Maximum Likelihood Estimation

In parametric MLE, the waiting times are just a function of the call-to-scene times T_i. We observed in section 2.1 that call-to-scene times and abandonment tendencies are not same across urban and semi-urban areas. It is hence obvious that a single distribution for waiting time cannot be a representative for the whole population. There is a need to account for the difference in

waiting tendencies between callers from different locations and callers with different types of emergencies. It becomes imperative that we use explanatory variables also in addition to the call-to-scene times. Semi-parametric MLE becomes useful when there are explanatory variables in addition to the observed data.

We use a conditional distribution $f(T|Z)$, where Z is the vector of the explanatory variables or *covariates*. The log-likelihood function (eq. 2.11) becomes

$$\log L(\theta|t) = \sum_{i=1}^{n} [\delta_i \log F(t_i|z_i) + (1 - \delta_i) \log S(t_i|z_i)] \tag{2.12}$$

The Proportional Hazards model by Cox (Cox, 1992), which we refer to as the Cox-PH model is used to solve a semi-parametric model. According to this model, the survival probability of the distribution at time t given the covariate of interest z is defined as

$$S(t|z) = S_0(t)^{e^{\beta z}} \tag{2.13}$$

where $S_0(t)$ is the baseline survival probability distribution. $S(t|z)$ can be estimated without making a parametric assumption for the baseline distribution for right-censored data using a Cox-PH model. However it is not possible to solve the Cox-PH model without making a parametric assumption for the baseline distribution in the case of interval censored data (Huang and Wellner, 1997). In our experiments, we make a Weibull assumption for the baseline distribution. We describe this in detail in section 2.2.3.

Non-parametric Maximum Likelihood Estimation

A Non-parametric MLE (NPMLE) approach is used to estimate an empirical probability distribution of the data, thus eliminating any need for prior distributional assumptions. We get a discrete probability distribution as a result of this NPMLE procedure.

A number of studies describe solving Maximum Likelihood models using a non-parametric assumption. The Expectation- Maximization algorithm (Dempster et al., 1977) is a popular approach used to solve NPMLE models. For the special case of type 1 interval censored data, (Turnbull, 1976)

proposed a self-consistency algorithm, which is an E-M algorithm. This algorithm divides the waiting time distribution into a number of discrete non-overlapping intervals and estimates the proportion of observations lying within every interval.

(Groeneboom, 1995) proposes an iterative convex minorant algorithm, which maximizes a concave maximum likelihood objective function. (Huang et al., 1996) proposes using a Maximum Profile Likelihood Estimation Approach to solve NPMLE. The log likelihood equation is written in terms of the cumulative hazard function $\Lambda(t)$. For $i = 1, 2...n$ where n is the number of observations, the log-likelihood is

$$l_n(\theta, \Lambda) = \sum_{i=1}^{n} \delta_i \log(1 - e^{-\Lambda(Y_i)e^{\theta' Z_i}} - (1 - \delta_i)e^{\theta' Z_i}\Lambda(Y_i) \tag{2.14}$$

(Huang et al., 1996) has also shown that the log-likelihood function is concave with respect to the cumulative hazard function. His main argument is as follows. The values of the cumulative hazard Λ at the observation points $i = 1, 2...n$ alone matter, hence it could be discretized. The cumulative hazard function can be thought of as a right continuous step function with jump points only at $T_(i)$ where $T_(i)s$ are the order statistics of the call to scene times.The cumulative hazard function is a non-decreasing function and hence this assumption can be valid. Maximizing this log-likelihood can now be formulated as maximizing a non-linear function subject to linear constraints i.e.,

$$Maximize \qquad \phi(\theta, \tilde{x}) = \sum_{i=1}^{n} \delta_i \log(1 - e^{-e^{\theta' Z_i} x_i} - (1 - \delta_i)e^{\theta' Z_i} x_i \tag{2.15}$$

$$s.t. \qquad \theta \in \Theta \qquad and \qquad 0 \leq x_1 \leq x_2 ..... \leq x_n \tag{2.16}$$

This formulation forms the basis of the NPMLE approach. Computing the Non Parametric Maximum Likelihood Estimator is relatively simple and computationally less expensive than the parametric and non-parametric models.

However there are a number of limitations in using a non-parametric model. According to this assumption, the maximum likelihood estimators are computed only at discrete instances in time where ambulance arrivals at the scene are observed. For other instances in time, we cannot compute the density

function making it difficult for generalization. Also since the model only estimates the observed proportion in every waiting time interval, there is a problem of over-fitting the model to the observations in the data set. Perturbing the observations by a small amount will produce a significant change in the estimated distribution, and hence we resort to using a proportional hazards model to estimate the waiting times.

2.2.3 Estimating Waiting Times

We first define an upper bound for the waiting time of every patient in the data set. u_i is defined based on the knowledge of the dispatch policies in the system. Typically a dispatch officer finds the nearest available ambulance for each call, in the region he operates in. If no ambulance is found to be available, the call is dropped. Thus, u_i is chosen as the call-to-scene time of the farthest ambulance that can be dispatched to call i without the call being dropped. This helps us practically bound the waiting time, as waiting times higher than this will not be practically observed in the data.

We use a two-parameter Weibull distribution for the baseline as we need not make a prior assumption on the nature of the hazard rate. The shape parameter k defines the nature of the hazard rate. If the estimated shape parameter is strictly less than 1, it indicates the hazard rate is decreasing with increasing call-to-scene time. If the estimated hazard rate parameter k is equal to 1, it indicates a constant hazard rate. If the estimated hazard rate parameter is greater than 1, it means that the hazard rate (probability of abandonment) is increasing with increasing call-to-scene time. Hence the estimated shape parameter defines the patient waiting behavior without a need for assuming it prior to estimation.

We implement the proportional hazards model in R using the package 'icenReg' (Anderson-Bergman, 2017), assuming a Weibull baseline distribution. The estimated function for the waiting time distribution, after solving the MLE for the proportional hazards model is as follows.

$$S(t|z) = (e^{(\frac{-t}{50.25})^{1.699}})^{\beta} \tag{2.17}$$

where the regression coefficient

$$\beta \begin{cases} = 1, \text{if the patient calls for type 1 emergency from an urban area} \\ = 0.7577, \text{if the patient calls for type 2 emergency from an urban area} \\ = 0.5716, \text{if the patient calls for type 1 emergency from a semi-urban area} \\ = 0.5647, \text{if the patient calls for type 2 emergency from a semi-urban area} \end{cases} \tag{2.18}$$

Since the estimated shape parameter k is greater than 1, we deduce that the hazard rate is increasing or in other words, the patients' tendency to abandon increases with time.

The estimated survival curves are shown in Figure 2.4.

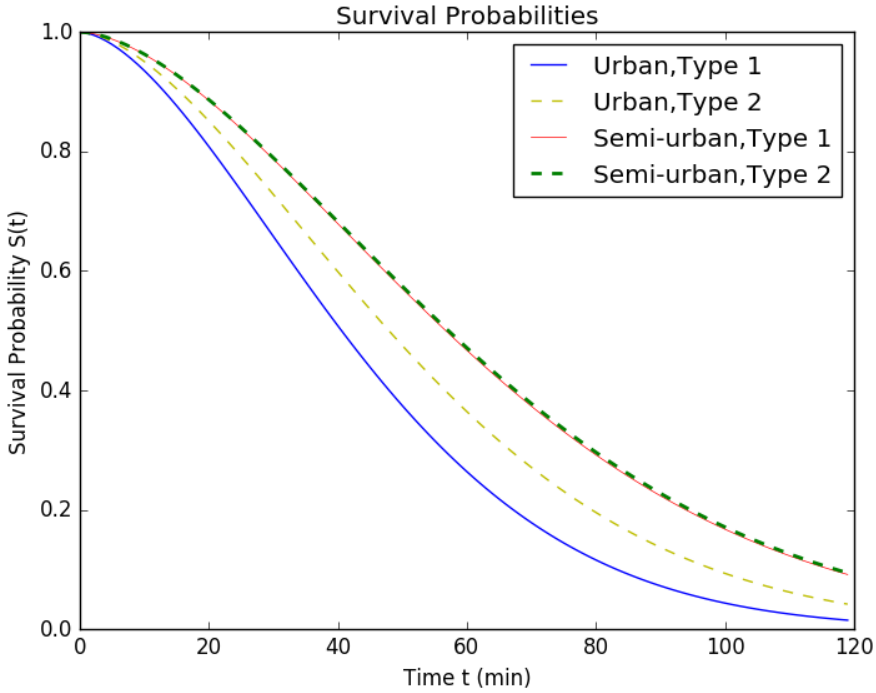

Figure 2.4: Estimated Survival Curves

To understand the survival curves better, we plot the hazard rate of the distribution as a function of time. The hazard rate curves are shown in Figure 2.5.

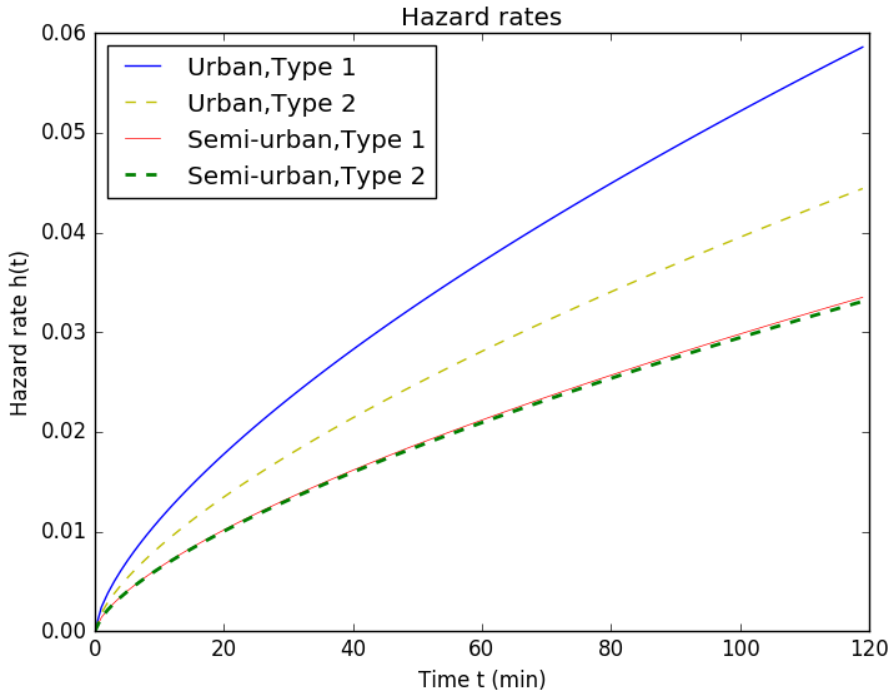

Figure 2.5: Hazard rate curves

The hazard rate of urban patients is more than the hazard rate of semi-urban patients, as we notice that the hazard rate curves for urban patients rise more steeply as compared to the semi-urban patients. In easier terms this can be interpreted as follows: Given that a patient has waited until time T, the probability that a patient from an urban locality will immediately abandon after time T is greater than the probability that a patient from a semi-urban locality will immediately abandon after time T. This observation stems from the fact that the average call-to-scene times in the semi-urban areas is on an average 10 minutes higher than the call-to-scene times in the urban areas.

Within urban and semi-urban cases, the curves for type 1 emergency cases rise more steeply than the curves for type 2 emergencies. Using the same argument described as above, we can conclude that given a certain call to scene time, the patients with type 1 emergencies have the highest hazard rate.

2.3 Resource Allocation

Our principal aim is to maximize the number of calls successfully served by the EMS to achieve service levels comparable to those of developed economies. We first attempt to study if optimizing the use of current resources and operational changes will lead to better service levels, before moving on to studying the effect of additional resources. Section 2.3 has three major components. First, we first aim to achieve a better allocation of the existing set of ambulances across the existing set of bases. Then we look at modifying the current ambulance dispatch policy by utilizing our knowledge of patient waiting times that were estimated in section 2.2. Thirdly, we redesign the existing EMS network by increasing the number of available ambulances, changing the base locations and studying different dispatch policies.

2.3.1 Ambulance Allocation

We first focus on achieving a better allocation of the existing fleet of ambulances across the existing bases. The solution space for this problem is exponentially large and the approach to arrive at the optimal solution is NP-hard. To this end, we employ a greedy optimization algorithm described in (Yue et al., 2012a).

This approach first builds a simulator to evaluate the cost of each possible allocation. It uses as input sets of call logs from historical or simulated data. the approach we use is a simulation-based optimization approach. Given the sets of call logs, a dispatch policy and an allocation (configuration) of ambulances at bases, the simulator finds the cost function related to the allocation.

Our cost function for this problem is defined as follows. Each call has a cost of 1 if unsuccessfully served, that is, no ambulance could be assigned to the call as all ambulances were busy, or the call was abandoned after an ambulance was assigned.

Algorithm 1 DISPATCH: First-come First-served Dispatch Policy

input: current request r, available ambulances W, priority queue q_r for request r
for $a \in q_r$ in decreasing preference order **do**
  if $a \in W$ **then**
    return: a
  end if
end for
return: $\perp$

To evaluate the utility of an allocation, we use the **data-driven simulator** described in Algorithm 5. This is similar to the simulator described in (Yue et al., 2012a), but in addition we account for heterogeneity among emergency callers and ambulance abandonment as well.

The following are the inputs to the simulator. Let $R = \{r_1, \ldots, r_N\}$ represent a request log with a sequence of requests. A is the allocation vector of ambulances to bases. y_r denotes the base of ambulance dispatched to service request r ($\perp$ if no ambulance was dispatched). $r(y_r)$ denotes active call r to which ambulance y_r is dispatched and $\bar{t}_r(y_r)$ denotes the completion time of request r. t_r denotes call arrival time for request r, $c_r(y_r)$ denotes the call-to-scene time of the ambulance dispatched for request r and w_r denotes the estimated patient wait time (willingness to wait) of request r.

Algorithm 2 SIMULATOR: Data-driven Simulator Method

input: (R, A), DISPATCH
$W \leftarrow A$ *//keeps track of which ambulances are free*
$\hat{R} \leftarrow \emptyset$ *//keeps track of active requests*
initialize $Y = \{y_r\}_{r \in R}$ such that $y_r \leftarrow \perp$
initialize events $\mathcal{E} \leftarrow R$ sorted in arrival order
while $|\mathcal{E}| > 0$ **do**
  remove next arriving event e from $\mathcal{E}$
  if $e =$ new request r **then**
    $y_r \leftarrow$ DISPATCH(r, W, R) *//dispatch policy*
    if $y_r \neq \perp$ **then**
      $\hat{R} \leftarrow \hat{R} + r(y_r)$ *//updating active requests*
      $W \leftarrow W - y_r$ *//updating free ambulances*
      if $c_r(y_r) > w_r$ **then**
        $\bar{t}_r(y_r) \leftarrow t_r + 2 * c_r(y_r)$ *//call abandoned; updating request completion time*
      end if
      insert job completion event at time $\bar{t}_r(y_r)$ into $\mathcal{E}$
    end if
  else if $e =$ job completion event $\bar{t}_r(y_r)$ **then**
    $\hat{R} \leftarrow \hat{R} - r(y_r)$ *//updating active requests*
    $W \leftarrow W + y_r$ *//updating free ambulances*
  end if
end while
return: Processed assignments of ambulances to requests Y

2.3.2 The Greedy Algorithm

We employ a greedy algorithm to obtain the allocation of a set of m ambulances over n base locations across the city. We describe the procedure in detail below.

The greedy algorithm employed by (Yue et al., 2012a) is used to solve the problem of ambulance allocation. The algorithm is described in Algorithm 3. (Yue et al., 2012a) employ the data driven simulator subroutine described in section 2.3.1, in the greedy algorithm. The algorithm iteratively selects the ambulance that has maximal incremental gain to the current solution until all the ambulances ambulances have been allocated. To employ a greedy

algorithm the submodularity conditions have to be satisfied and (Yue et al., 2012a) show that this problem is approximately submodular.

Algorithm 3 Greedy Ambulance Allocation

1: **input**: F, K
2: $A \leftarrow \emptyset$
3: **for** $\ell = 1, \ldots, K$ **do**
4: $\quad \hat{a} \leftarrow \arg\max_a \delta_F(a|A)$
5: $\quad A \leftarrow A + \hat{a}$
6: **end for**
7: **return**: A

The solution provided by the greedy algorithm is not guaranteed to be optimal. But the the solution is bound to be at least $(1 - \frac{1}{e})$ of the optimum when the property of submodularity holds((Nemhauser et al., 1978)). Yue et al (2012a) show that the greedy algorithm can be close to optimal using a bounding procedure. therefore we employ the same algorithm in this work.

2.4 Computational Experiments

The greedy algorithm described in Section 4 is used to obtain an allocation of m ambulances over n bases. For the purpose of simulating an allocation of ambulances, we require a set of sampled requests R to evaluate an allocation. We sample calls according to a sampling procedure, as described in Algorithm 4. The following estimations have been included in the model.

We fit the call arrival times to a spatio-temporal Poisson process. For the purpose of our analysis, we estimate the Poisson rate parameter separately for each sub-city district within the city. The time elapsed since an ambulance arrives at the scene until it reaches its base again is called as back to base time. We find all historical travel times by learning from real-world traffic data. To find the travel times based on distance between any two points in a map, , we fit a simple regression model based on the historically experienced travel times. We use this model to eliminate the need for large volume GIS data where there is a need to estimate distances between every pair of points for tens of thousands of points.

Algorithm 4 Sampling Procedure

input: t_{start}, t_{end}
$R \leftarrow \emptyset, t \leftarrow t_{start}$
while $t < t_{end}$ **do**
    Sample $r \leftarrow \mathcal{P}(r|t)$ *// Starting request sampling at time t*
    $t \leftarrow t_r$ *// Incrementing time counter*
    $R \leftarrow R \cup \{r\}$ *// Adding a sampled request to the collection*
end while
return: R

Our sampling procedure assumes that the emergency requests are independent of each other and the dispatch behavior of the EMS. Here, $\mathcal{P}(r|t)$ denotes the distribution of the next arriving request starting at time t, and t_r denotes the arrival time of request r.

For the ambulance allocation problem, we sample a total of $N_{total} = 1500$ call logs, where each call log is worth a week and use the sampled data for training, validation and testing. In our experiments, we use $N_{train} = 500$, $N_{valid} = 500$ and $N_{test} = 500$ call logs. In addition, the set of 500 training call logs is divided into $M = 10$ disjoint subsets. We first obtain M allocations $A_1, A_2 ... A_M$ by using the M disjoint subsets of the N_{train} requests. We then evaluate these allocations on N_{valid} validation requests. The allocation for which we observe the maximum number of calls served is chosen, and evaluated on the N_{test} requests. We then report these results as the performance of an allocation. The greedy algorithm minimizes the following cost objective function:

$$Cost = \begin{cases} 0, \text{if a call is served and not abandoned} \\ 1, \text{otherwise} \end{cases}$$

This cost function penalizes any call for which either an ambulance is sent but abandoned or an ambulance is not dispatched for a call. It does not differentiate between the calls that were served. Hence this cost function solely aims to maximize the number of instances where an ambulance was dispatched for service and it was utilized by the patient.

2.4.1 Experimental Setup

The current action space consists of 58 bases and an ambulance located per base. The dispatch policy is to send the nearest free ambulance to every call. In this experimental setup, we allocate ambulances to bases by varying a combination of the following three parameters: the budget of ambulances, the dispatch policy and the base locations themselves. The base locations are shown in the map in Figure 2.6

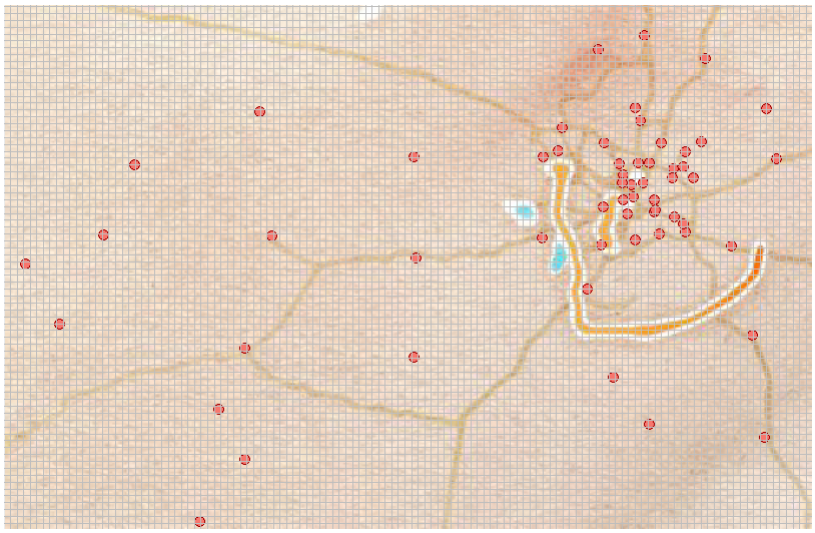

Figure 2.6: Map of existing base Locations with a latitude-longitude grid superimposed on it

2.4.2 Greedy Allocation Solution

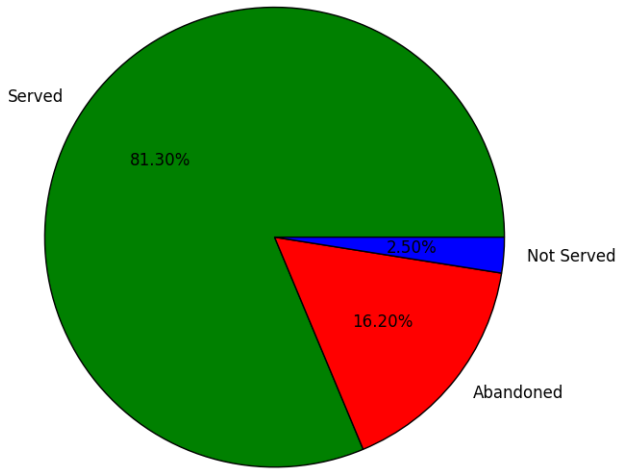

Figure 2.7: Performance of the greedy allocation

The greedy algorithm involves a large number of function evaluations, hence we use a lazy variant of the algorithm described in (Leskovec et al., 2007). This lazy greedy algorithm produces nearly identical results with much lesser function evaluations. We also use Sample Average Approximation ((Verweij et al., 2003)) for model selection. The lazy greedy solution improves upon the baseline performance by less than 1% with the existing budget of 58 ambulances. Hence we seek to achieve a significant jump in the number of calls served by increasing the ambulance budget.

2.4.3 Varying the ambulance budget

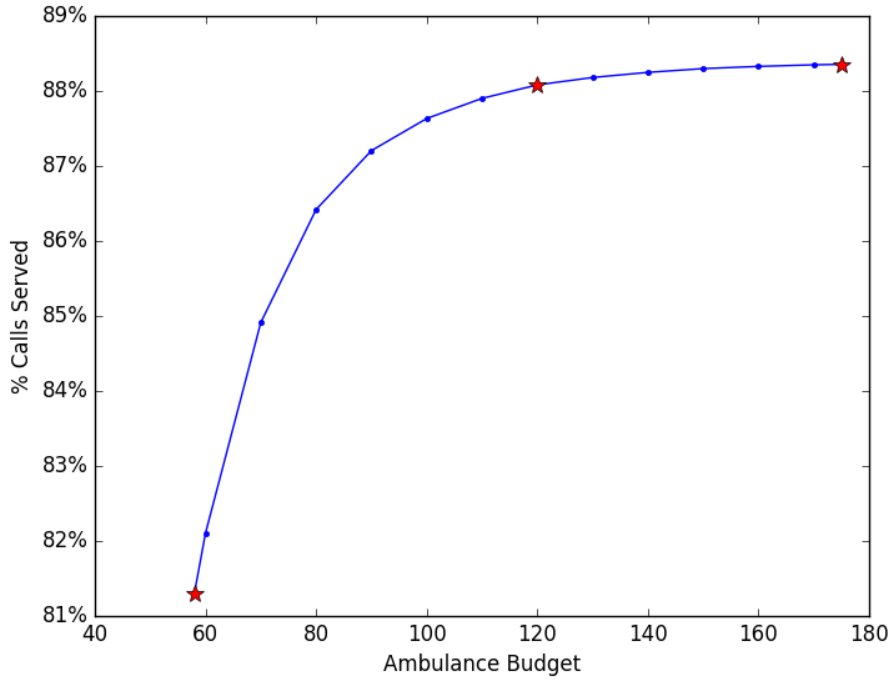

Figure 2.8: Calls Served and not abandoned (values at 58, 120 and 175 ambulances indicated in red)

The allocation obtained from the greedy algorithm does not increase the calls served by a significant amount. Because we have a limited budget of ambulances to service a large geographical area, the question of whether the resource constraint i.e., the limited budget of ambulances is the primary cause of abandonment, is the next question we wish to answer. To tackle this, we increase the budget of ambulances to three times its current budget so that the ambulance-population ratio is comparable to U.S. standards. We then repeat the same experiment as above under similar conditions, but we increase the number of ambulances to as high as 175 from 58 to measure the impact of an enhanced budget on the number of calls served. Figure 2.8 shows that the number of calls served increases monotonically in a non linear fashion with respect to the calls successfully served without being abandoned and becomes nearly constant after 130 ambulances. We get an improvement of 4 percentage points decrease in the calls abandoned by tripling the ambulance budget.

Figure 2.9 illustrates a similar trend in decreasing fashion with respect to abandonment and we see an 18% reduction in abandonment as a result of tripling the ambulance budget.

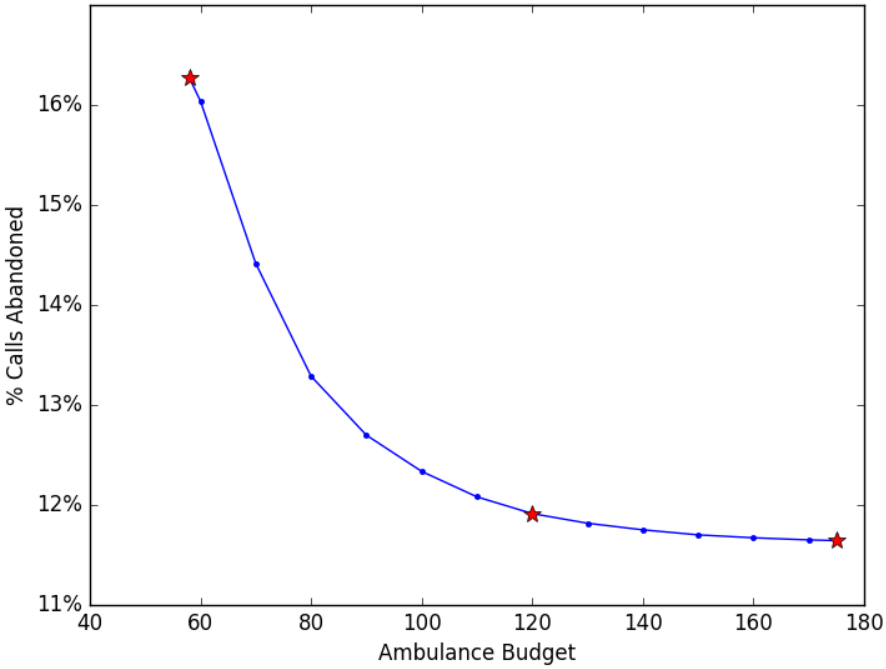

Figure 2.9: Calls abandoned(values at 58, 120 and 175 ambulances indicated in red)

2.4.4 Varying the dispatch policy and ambulance budget

Because increasing the number of ambulances at the current set of bases does not change the abandonment significantly, we explore a different operational strategy, that of changing the dispatch policy. The idea of modifying the dispatch policy is motivated by the following assumption: If we know with a high probability that a patient is going to abandon, the ambulance meant to serve that patient can be re-routed to someone else, improving the utilization and hence the service level of the EMS system.

We utilize our estimate of a patient's waiting time to modify the dispatch policy of ambulances. We use the patient waiting times to compute the

probability of abandonment. The probability of abandonment is defined as

$$P(abandonment) = P(\text{Waiting Time} \leq \text{call-to-scene Time})$$

.

We use a threshold parameter $\alpha, \alpha \in \{0, 0.1, 0.2, 0.3...0.9, 1\}$ for making a dispatch decision.

$$P(abandonment) \begin{cases} \leq \alpha, \text{ then dispatch an ambulance} \\ > \alpha, \text{then do not dispatch an ambulance} \end{cases}$$

If $\alpha = 0$ then no ambulance will be dispatched for any call in the system. If $\alpha = 1$ then ambulances will be dispatched for all the calls in the system irrespective of the probability of abandonment.

In our experiments, we also vary the number of ambulances as a parameter, keeping the number of bases and base locations fixed. The results are summarized in Figure 2.10.

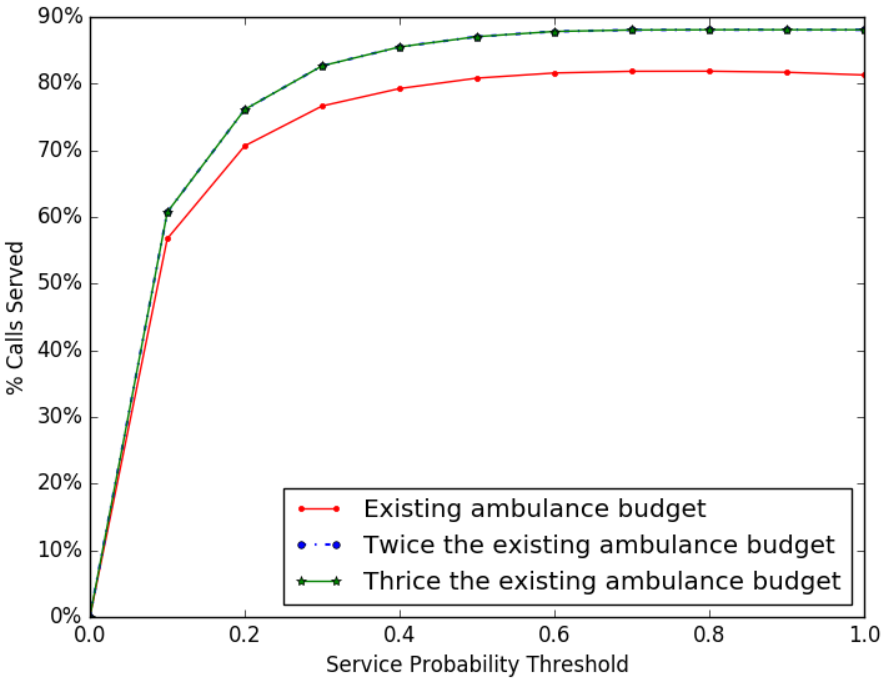

Figure 2.10: Varying dispatch policy

From the results, we can see that the number of calls served follows an increasing trend with α. However, for all the three fleet sizes, we see that the

number of calls served when $\alpha = 0.9$ is marginally higher than the number of calls served when $\alpha = 1.0$. The difference is marginally visible with the current fleet size, but not seen when the fleet size is doubled or tripled.

This leads to the policy implication that selective non-dispatch of ambulances when there is strong evidence of abandonment leads to better performance of the EMS network. However this policy would face operational issues as there will be reluctance from the ambulance providers to not dispatch ambulances to distress calls. However the improvement obtained as a result of this exercise is less than one-thousandth of a percent when the fleet size is increased. Since we see a significant reduction in abandonment when the fleet size is doubled or tripled and the current dispatch policy performs equally good, the policy implication would be to dispatch all ambulances irrespective of abandonment probability, to maximize service level and to encourage implementation compatibility with the ambulance provider.

2.4.5 Network redesign

Since the ambulance budget and dispatch policy together do not significantly increase the calls served, we attempt to redesign the network by changing the locations of the ambulance bases themselves. A grid is superimposed on the map of the city (Figure 2.6), and every intersection of latitudes and longitudes on the grid is taken a candidate base location resulting in a base location at every street corner, or closer.

For every possible call location, we consider all the grid locations within a radius of 30 minutes from the call location as candidate bases. locations for ambulances. We consider each latitude and longitude intersection, with a distance of 0.01 degrees (approximately 5 min travel time) between each other, as possible bases. Combining all such candidate base locations for all possible call locations, we arrive at a total of 13,644 candidate bases in the system, at which ambulances could possibly be located. Hence, this is a large scale experiment where the optimization process not only finds a good set of base locations but also a good allocation of ambulances at these new bases.

Ambulance allocation across bases has been studied previously in literature, but an experiment of this magnitude where resource allocation across more than 10,000 bases has not been done previously. For example, (Re-

strepo, 2008) considers an allocation of 97 ambulances across 88 bases in the city. (McCormack and Coates, 2015) perform a very similar study where they allocate ambulances to 70 bases. Our study is unique because of the sheer magnitude of the action space we consider for optimization, which is two orders of magnitude higher than the size of problems studied in existing literature.

We use this modified action space and run the greedy algorithm to allocate the current fleet of 58 ambulances in the 13,644 bases across the city. The result of this experiment is shown in Figure 2.11.

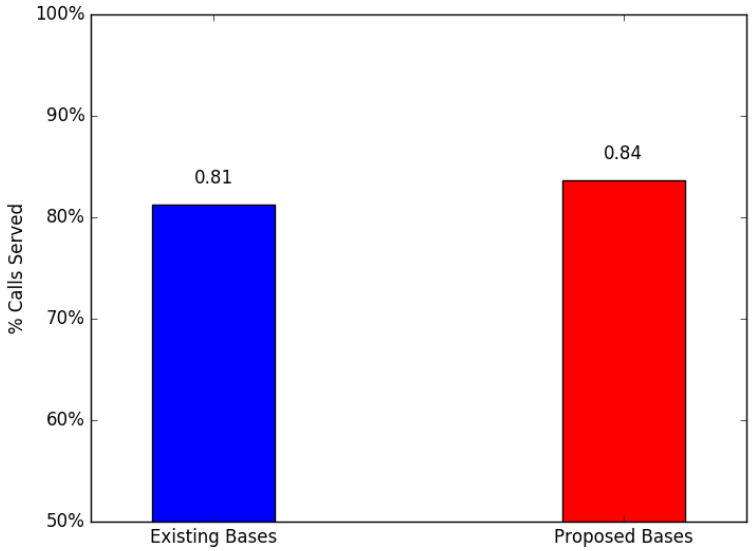

Figure 2.11: Performance with New Base Locations

With the current budget of 58 ambulances, the performance of the greedy algorithm over the new set of bases is marginally higher than the performance with the existing set of bases (see Figure 2.11). The new set of bases in addition to the existing set of bases is shown in Figure 2.12.

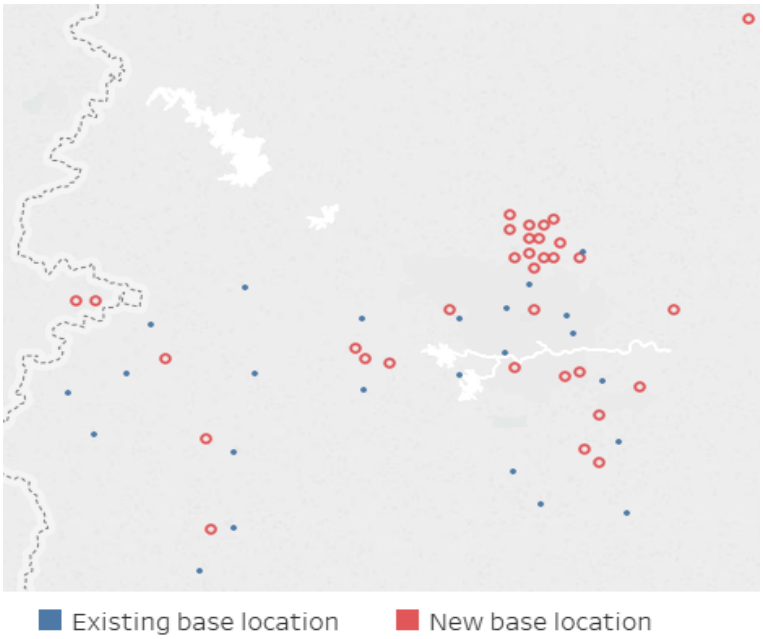

Figure 2.12: New base locations for current budget of ambulances

The spatial distribution of the new set of bases is not very different from the current setup. We see that a lot of the existing bases need to be slightly relocated in the urban areas, while there is no significant difference in the location of bases in the rural areas. Since resource optimization in a redesigned network does not improve service level significantly, we now expand the set of available resources to determine the additional resources we need to match service levels of emerging economies.

2.4.6 Varying the base locations and the ambulance budget

We now study the effect of supply side expansion, both in terms of number of ambulances and the bases. With an increase in ambulance budget we also now have to determine where the corresponding bases need to be situated and how many bases are required. To this end, we repeat the same experiment on the expanded set of bases and also triple the ambulance budget as was the case with 58 bases. The results are summarized in Figure 2.13.

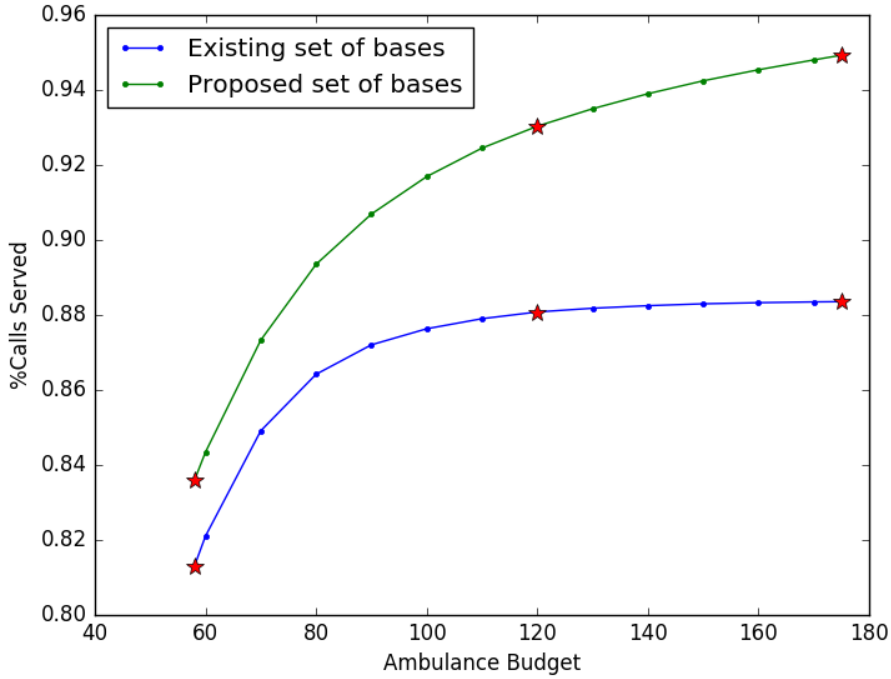

Figure 2.13: Performance with New Base Locations - Calls served

Budget	% Calls Served	
	Current base configuration	Proposed base configuration
58	81.30%	83.59%
120	88.08%	93.03%
175	88.35%	94.92%

Table 2.2: Comparison of fraction of calls served

We see a significant rise in the number of calls served with the expanded set of bases. The gap between this performance and the performance of 58 bases is narrow until a budget of 90 ambulances and deviates significantly after 90 ambulances. When the ambulance budget is tripled, we see a significant rise in the number of calls that were served and not abandoned. In absolute terms, nearly 95% of all calls are served and not abandoned. This is nearly a 14 percentage point increase from the performance of the greedy algorithm with the current ambulance fleet size.

Figure 2.14 shows the decrease in abandonment with the expanded set of bases. When the budget is tripled, the abandonment decreased by more

than half. While it may not be possible to increase the ambulance budget to thrice the original size, Figure 2.14 also demonstrates that the majority of the abandonment reduction can be achieved with doubling the number of ambulances, i.e., with 120 ambulances.

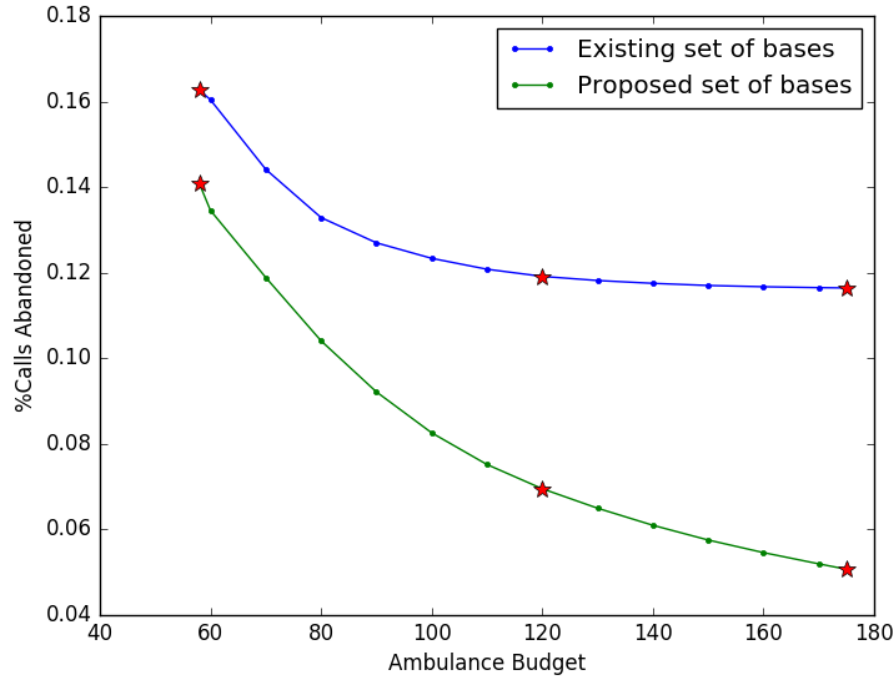

Figure 2.14: Performance with New Base Locations - Calls abandoned

Budget	% Calls Abandoned	
	Current base configuration	Proposed base configuration
58	16.26 %	14.07%
120	11.91 %	6.95%
175	11.64%	5.07%

Table 2.3: Comparison of fraction of calls abandoned

The set of base locations with additional budget of ambulances are shown in figures 2.15 and 2.16. As we increase the budget, we see that more number of bases are first allocated to the densely populated areas or the urban areas followed by addition of new bases to semi-urban areas. In section 2.1, we have described that patients from semi-urban abandon less frequently and wait longer. Since our optimization objective strives to minimize the number

of calls that were served and not abandoned, a bulk of the resources are initially allocated to urban areas which contribute more to abandonment. As we make more resources available to our algorithm, it starts allocating more to semi-urban areas where people wait longer.

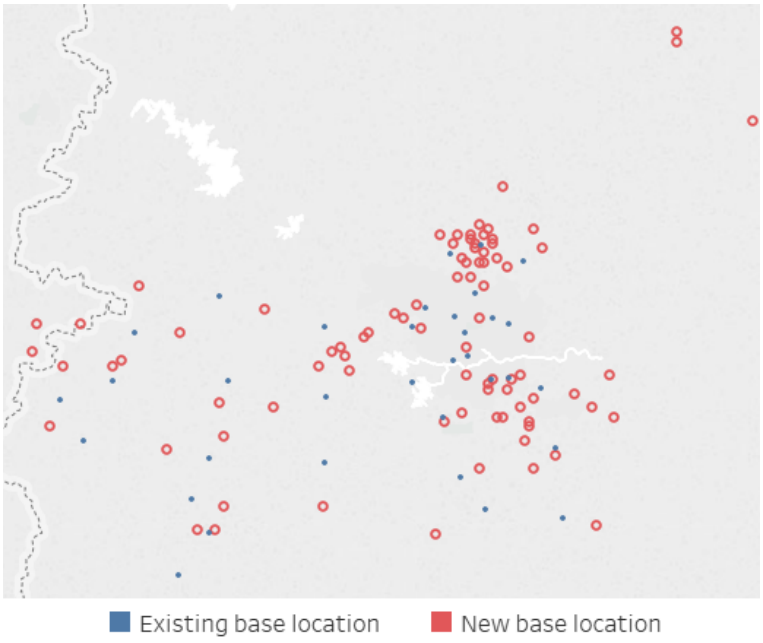

Figure 2.15: New base locations for budget of 120 ambulances

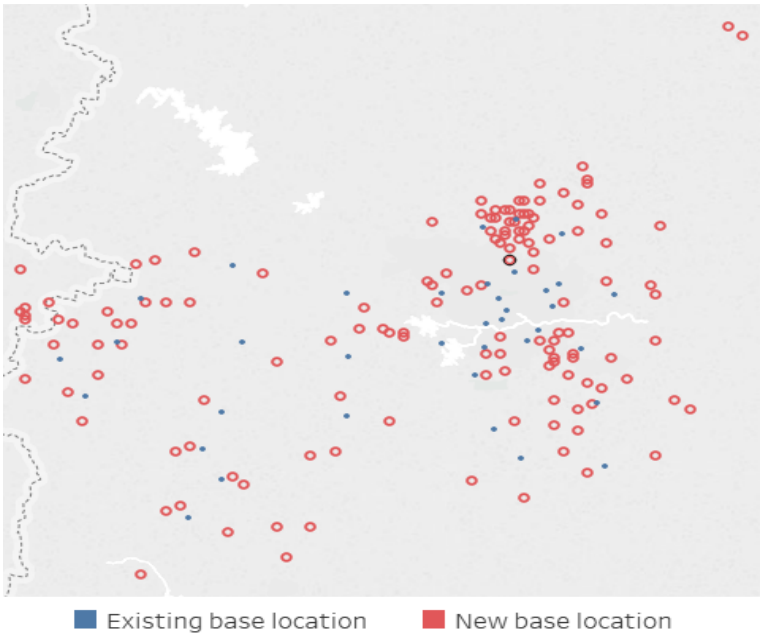

Figure 2.16: New base locations for budget of 175 ambulances

Figures 2.17 through 2.19 represent the improvement in the performance of the EMS system in terms of its timely service. Figure 2.19 depicts the the decrease in the number of calls that were either not served or abandoned. From Figure 2.17 we can see that there is a significant spike in the number of calls that were served within 15 minutes. We also see an improvement in the calls served within 30 minutes. This can be attributed to the enhanced spatial distribution of the new base locations to enable a more distributed system in comparison to the current system, enabling faster service. Hence we not only improve the fraction of calls that were served, but also significantly improve upon the number of calls that were successfully served within 15 and 30 minutes respectively.

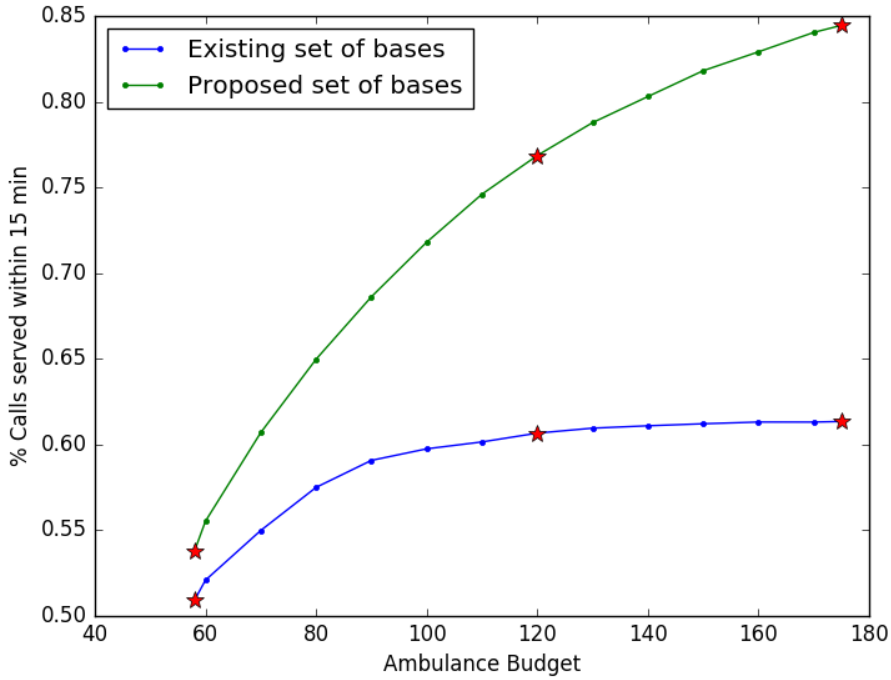

Figure 2.17: Fraction of calls served within 15 minutes

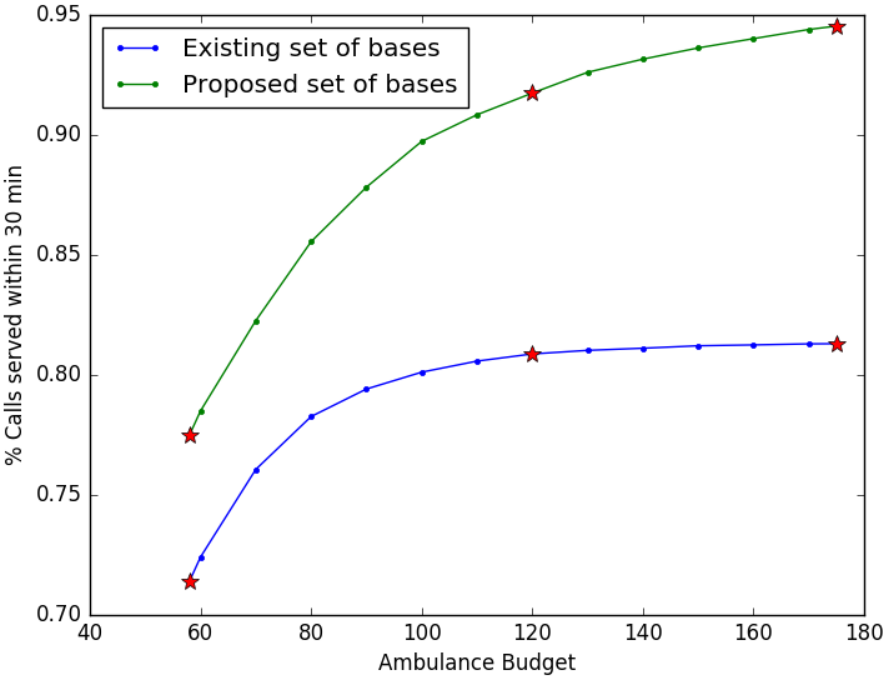

Figure 2.18: Fraction of calls served within 30 minutes

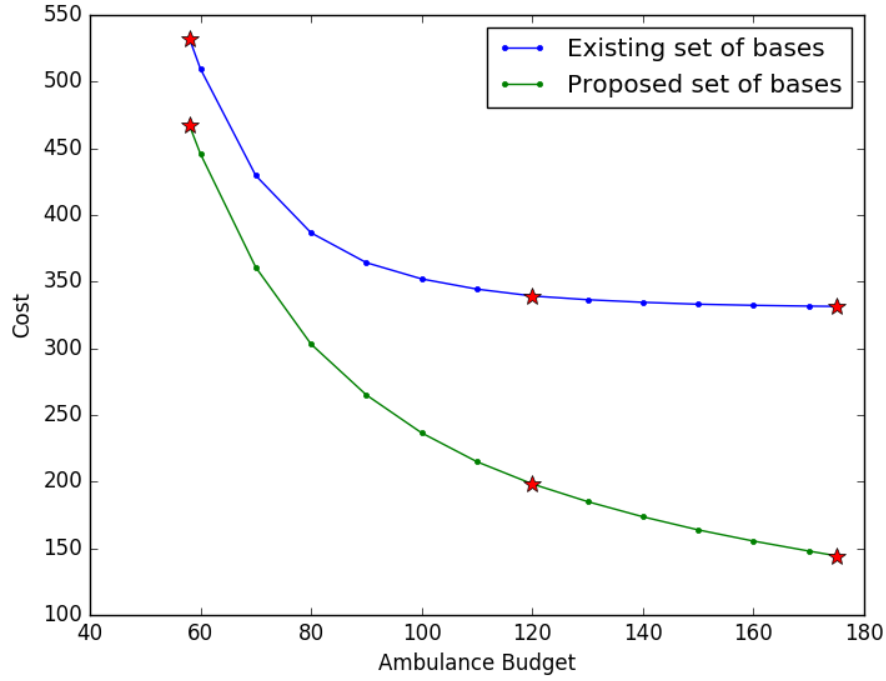

Figure 2.19: Cost when varying ambulance budget

2.4.7 Varying the base locations, dispatch policy and ambulance budget

With the existing set of bases, we determined earlier that the current dispatch policy of sending the nearest free ambulance to all the calls was the best. We now try to understand if the same policy holds good for the new set of bases as well. The results of this experiment are shown in Figure 2.20. We observe that the policy of selective non-dispatch when $\alpha = 0.9$ performs well, as we had seen with the current set of bases.

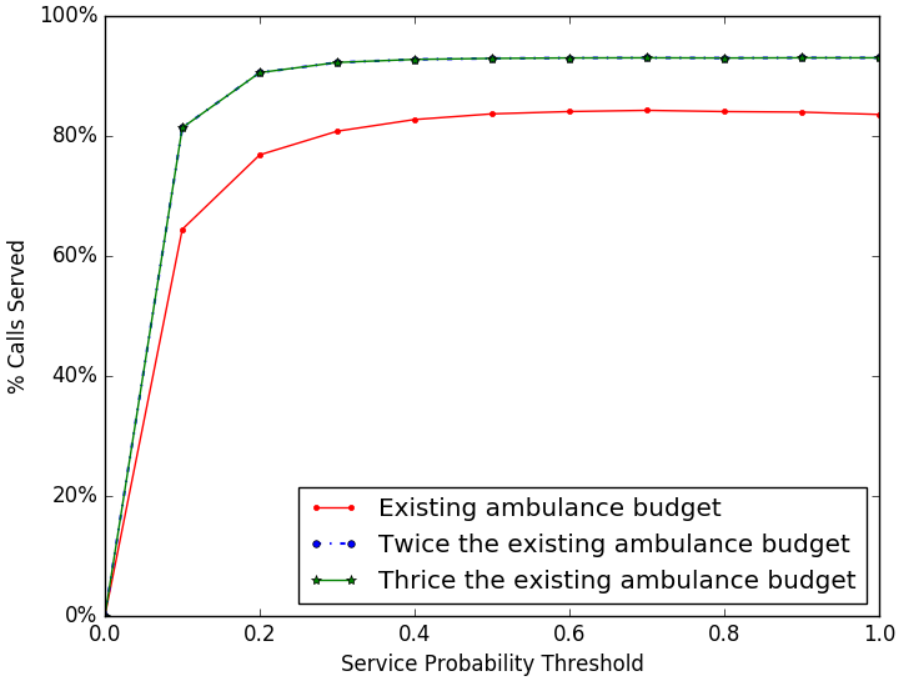

Figure 2.20: Performance with New Base Locations - Varying dispatch policy

However, because selective dispatch can be very difficult to implement in practice, and moreover does not provide significant gains; we recommend the policy of dispatching ambulances to all callers irrespective of the tendency to abandon.

Thus we have identified network redesign and operational strategies that can be used by the operator to operate with a clearer awareness of abandonment, resulting in improved service to users across the system.

Chapter 3

Robust Ambulance Allocation using Risk-Based Metrics

This chapter includes previously published material from (Krishnan et al., 2016) [1] ©2016 IEEE. We do not use the same data set that was used for chapter 2. Instead, we use a modified version of the data from (Yue et al., 2012a) for our analysis here.

3.1 Objective

The primary question addressed in this chapter is not only how to allocate a fleet of ambulances to maximize its service level but also to design the network in such a way that it is robust to unexpected demand patterns.

Due to the spatio-temporal nature of emergencies occurring in EMS systems, the resource constraints on ambulances and the network structure of the city, there are cascading effects on the calls causing each call to be dependent on the previous calls and the ambulances assigned to them. These cascading dependencies are captured by assuming that the call arrivals follow a Poisson process and optimizing for the average-case metrics. However in emerging economies where resources are heavily constrained, the notion of Poisson call arrivals may not always hold and the above approach may not be robust. Hence we need to account for the tail behavior in addition to average-case metrics. We achieve this by incorporating a risk metric, Conditional Value at Risk (CVaR) in the objective function.

[1] K. Krishnan, L. Marla, and Y. Yue, Robust ambulance allocation using risk-based metrics, in Communication Systems and Networks (COMSNETS), 2016 8th International Conference on. IEEE, 2016, pp. 16
I was provided with the base code by the co-authors of this paper which I had used and built upon
The IEEE grants authors the license to re-use their paper in their thesis

3.2 Motivation

When some parts of the network incur heavy-tailed call arrivals, we observe that the entire resource-constrained network behaves in a heavy tailed manner.We illustrate this statement considering data from an Indian metropolitan city. The city contains 83 sub-city districts. When as few as six sub-districts begin to follow a heavy-tailed distribution, the entire call stream follows a heavy-tailed distribution (see Table 3.1).

	Call Log 1	Call Log 2
Sub-city districts (Light-tailed)	83	77
Sub-city districts (Heavy-tailed)	0	6
Distribution Parameters	Poisson Rate = 0.28	Weibull Shape = 0.97, Scale = 3.9

Table 3.1: Comparing call streams from light-tailed and heavy-tailed distributions ©2016 IEEE

The performance of these allocations are evaluated using the data-driven discrete-event simulator described in (Yue et al., 2012b). The procedure followed in the simulator is described in Algorithm 5. The dispatch policy is the same as what we had described in Algorithm 1 in chapter 2. The nearest available ambulance is dispatched to service a request. When the nearest ambulance is not available to serve the next call, it creates a *dependency* between two requests.

Informally, r depends on r' if the assignment of $y_{r'}$ to r' causes r to be assigned y_r such that $y_{r'} \succ_{p_r} y_r$ (Yue et al., 2012b). The formal definitions follow.

Definition 1 *There exists an active dependency $\gamma_{r,r',y_{r'}}$ from request r to request r' with label $y_{r'}$ if*

1. $t_{r'} < t_r$ *(r' arrives before r)*
2. $\bar{t}_{r'}(y_{r'}) > t_r$ *(r' completes after r arrives – this indicates that the two requests "overlap" in time)*

3. $y_{r'} \succ_{p_r} y_r$ *(r' is assigned an ambulance from a higher priority base, w.r.t. r's priority queue, than the ambulance ultimately assigned to r)*

The dependency structure in the network is dependent on the call arrivals, the ambulance allocation and dispatch policy. An efficient allocation is defined as one which allows for more calls to be served by the nearest ambulance, thereby maximizing the service level. When such dependencies occur with Poisson call arrivals, it will be more pronounced with heavy-tailed call arrivals.

To illustrate this behavior, we consider the naive allocation that the operator uses and evaluate the two call logs (see Table 3.2)

	Call Log 1	Call Log 2
Mean number of calls not served	8.87%	10.23 %
90^{th} quantile of calls not served	10.11%	12.58%

Table 3.2: Naive allocation performance on light-tailed and heavy-tailed call logs

Hence it is important that we allocate ambulances considering the performance at the tail of the distribution as well.

The data structure for input to the discrete event simulator described below is the same as what is described in chapter 2, but we do not consider patient waiting time and emergency classification.

Algorithm 5 SIMULATOR: Data-driven Simulator Method

input: (R, A), DISPATCH
$W \leftarrow A$ *//keeps track of which ambulances are free*
$\hat{R} \leftarrow \emptyset$ *//keeps track of active requests*
initialize $Y = \{y_r\}_{r \in R}$ such that $y_r \leftarrow \perp$
initialize events $\mathcal{E} \leftarrow R$ sorted in arrival order
while $|\mathcal{E}| > 0$ **do**
  remove next arriving event e from $\mathcal{E}$
  if $e =$ new request r **then**
    $y_r \leftarrow \text{DISPATCH}(r, W, R)$ *//dispatch policy*
    if $y_r \neq \perp$ **then**
      $\hat{R} \leftarrow \hat{R} + r(y_r)$ *//updating active requests*
      $W \leftarrow W - y_r$ *//updating free ambulances*
      insert job completion event at time $\bar{t}_r(y_r)$ into $\mathcal{E}$
    end if
  else if $e =$ job completion event $\bar{t}_r(y_r)$ **then**
    $\hat{R} \leftarrow \hat{R} - r(y_r)$ *//updating active requests*
    $W \leftarrow W + y_r$ *//updating free ambulances*
  end if
end while
return: Processed assignments of ambulances to requests Y

3.3 Modeling Approach

Our objective function uses a linear combination of expected value metrics and risk metrics. As a risk metric, we consider Conditional-Value-at-Risk (CVaR) due to its properties of coherence (Rockafellar and Uryasev, 2002),(Rockafellar and Uryasev, 2000).

For a general loss function described by random variable X and $0 < \alpha < 1$, $CVaR$ is defined as $CVaR_\alpha = \frac{1}{\alpha}\int_{1-\alpha}^{1} VaR_\alpha(X)d\alpha$ where VaR_α is the Value-at-risk (VaR). This can be equivalently written as:

$$CVaR_\alpha = -\frac{1}{\alpha}\left(E[X\ 1_{\{X \leq x_\alpha\}}] + x_\alpha(\alpha - P[X \leq x_\alpha])\right) \tag{3.1}$$

where $x_\alpha = \inf\{x \in \mathbb{R} : P(X \leq x) \geq \alpha\}$ is the upper α-quantile and

$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$ is the indicator function.

Let A denote an allocation of ambulances to a set of bases $\mathcal{A}$ (there can be more than one ambulance at a base). We represent A as a multiset of elements in $\mathcal{A}$. Let $M(\mathcal{A})$ denote the multi-powerset of $\mathcal{A}$ and $L(A)$ as the cost of allocation A. $L(A)$ may correspond to the fraction of requests not served, fraction of requests whose service time is above some target threshold, fraction of requests served at each service level; and is a tunable component of the framework.

We define $L(A)$ in terms of balancing reward and risk. In particular, as we want to typically maximize the expected value of an allocation and minimize its CVaR (the expectation of the allocation value being lower than a specified threshold), we consider the function: $\max(\beta * E(gain) - (1-\beta) CVaR_\alpha F(A))$.

We measure the cost of an allocation A using a real-valued objective function $F : M(\mathcal{A}) \to \Re$. We focus on penalty reduction formulations, where we can write $F(A)$ as

$$F(A) = L(\emptyset) - L(A), CVaR(A) = CVaR \tag{3.2}$$

where $L : M(\mathcal{A}) \to \Re$ measures the cost of an ambulance allocation over each request log that spans some period of time (e.g., one week).

We define L using the outcomes of simulated requests over several request logs. Using the simulator, we measure the percentiles of the expected metrics of interest over the set of request logs. Our goal is to maximize the expected gain in performance over some (known) distribution of requests $\mathbf{P}(R)$. Let $Y = \{y_r\}_{r \in R}$ denote the output of Algorithm 5 for request log R. Then we can write the αth quantile of expected penalty as

$$L(A) = \mathbf{E}_{R \sim \mathbf{P}(R)} \left[\sum_{r \in R} L_r(y_r) \right], \tag{3.3}$$

where $L_r(y)$ is the penalty of assigning request r with y_r (e.g., whether or not assigning ambulance y_r to r results in a service time above a target threshold).

In practice, we resort to optimizing over a collection of request logs $\mathcal{R} = \left\{ \{R_{mn}\}_{m=1}^{M} \right\}_{n=1}^{N}$, where each $R_m \in \mathcal{R}$ is sampled i.i.d according to $\mathbf{P}(R)$ and

each $R_m n$ sampled according to ABANDONMENT. In our experiments, we use Sample Average Approximation (Verweij et al., 2003) to bound the difference between our sample average objective and the optimal expected performance. We thus approximate the expectation with the sampled average,

$$L_{\mathcal{R}}(A) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\sum_{r\in R_m n} L_r(y_r) \approx E_R\left[\sum_{r\in R} L_r(y_r)\right]. \tag{3.4}$$

Let $\delta_F(a|A)$ denote the gain of adding a to A,

$$\delta_{E(L)}(a|A) = L(A) - L(A \cup a). \tag{3.5}$$

$$\delta_{CVaR_\alpha(L)} = CVaR_\alpha(L(A)) - CVaR_\alpha(L(A \cup a)). \tag{3.6}$$

$$\delta_F(a|A) = \beta * \delta_{E(L)}(a|A) - (1-\beta)\delta_{CVaR_\alpha(L)} \tag{3.7}$$

$\delta_{E(L)}(A)$ corresponds to the expected value of the dependency chains broken by the allocation A compared to the null allocation, and $\delta_{CVaR_\alpha(L)}$ corresponds to the condistional-value-at-risk of the dependency costs.

Given a budget of K ambulances, the static allocation goal then is to select the ambulance allocation A (with $|A| \leq K$) such that the utility $F(A)$ is maximized. More formally, we can write our optimization problem as

$$\underset{A\in M(\mathcal{A}):|A|\leq K}{\arg\max} F(A). \tag{3.8}$$

The greedy algorithm employed by (Yue et al., 2012b) is used, because the properties of *approximate* submodularity still hold as discussed above. The greedy algorithm has already been described in chapter 2. The algorithm iteratively selects the ambulance a that has maximal incremental gain to the current solution until M ambulances have been allocated. Note that each evaluation of $\delta(a|A)$ requires running the simulator to evaluate $F(A+a)$.

3.4 Computational Results

The data from the Indian metropolitan city contains approximately ten thousand logged emergency requests over the course of one month. Each record in the request log contains the type and location of the request, the ambu-

lance (if any) that was dispatched, and the various travel times (e.g., base to scene, scene to hospital, hospital to base). The request arrivals fit typically into Poisson distributions per sub-city-district and service times fit into log-normal distributions respectively. Request arrivals and service times are independent. However, certain sub-city-districts also have distributions that can be fit to heavy-tailed distributions (Weibull distribution). We will therefore examine if the difference in the assumptions behind the distributions (which makes the sampling consistent with real-world data) results in solutions that are robust to heavy tailed arrival rates.

We run our optimization model described in 3 with training call logs having exponential inter-arrival times, and test call logs sampled according to the following cases:

- Exponential inter-arrival times
- Weibull inter-arrival times in sub-city-districts (call arrivals are heavy-tailed)
- Poisson inter-arrival times with hotspots in some sub-city-districts (chosen such that high arrival areas are simultaneously stressed)
- Weibull inter-arrival times with hotspots in some sub-city-districts (chosen such that high arrival areas are simultaneously stressed)

Our action space contains 58 bases and 58 ambulances. We evaluate our methods over a period of one week. 500 training call logs and 500 test call logs, each spanning one day, and independent of each other, are used.

The following cost function is considered in our experiments.

$$L_r(y) = \begin{cases} 1 & \text{if service time} \geq 30min \\ 0 & \text{otherwise} \end{cases}.$$

Our metrics are the various quantiles of the test calls logs, that evaluate the tail probabilities of failure (non-service), as well as the mean performance. We use $\beta = 0.7$ and for varying values of α (tail CVaR values).

Figures 3.1 and 3.2 present the improvement in the tail metrics (calls not served) for varying values of protection levels α when there are no hotspots observed.

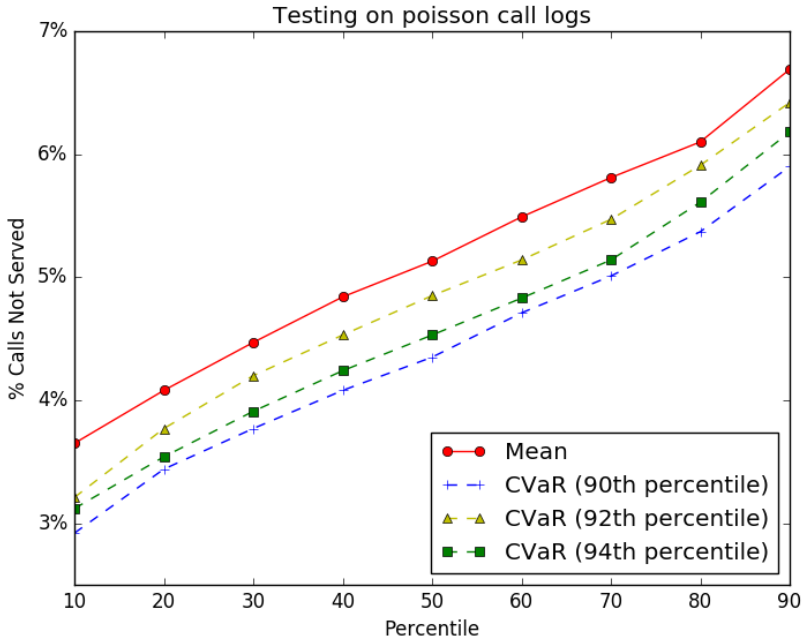

Figure 3.1: Performance of risk-optimized allocation on Poisson call log without hotspots ©2016 IEEE

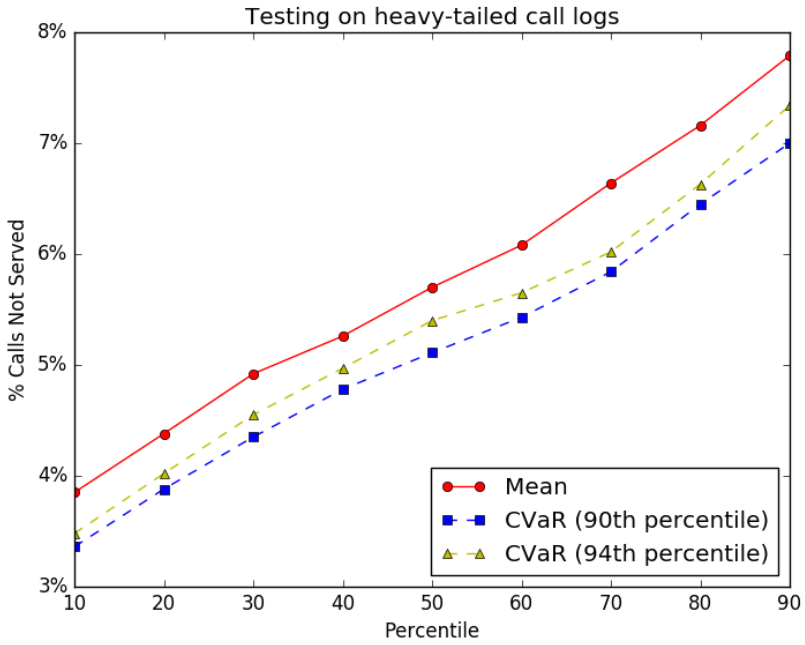

Figure 3.2: Performance of risk-optimized allocation on heavy-tailed call log without hotspots ©2016 IEEE

Figures 3.3 and 3.4 present the improvement in the tail metrics (calls not served) for varying values of protection levels α when there are hotspots.

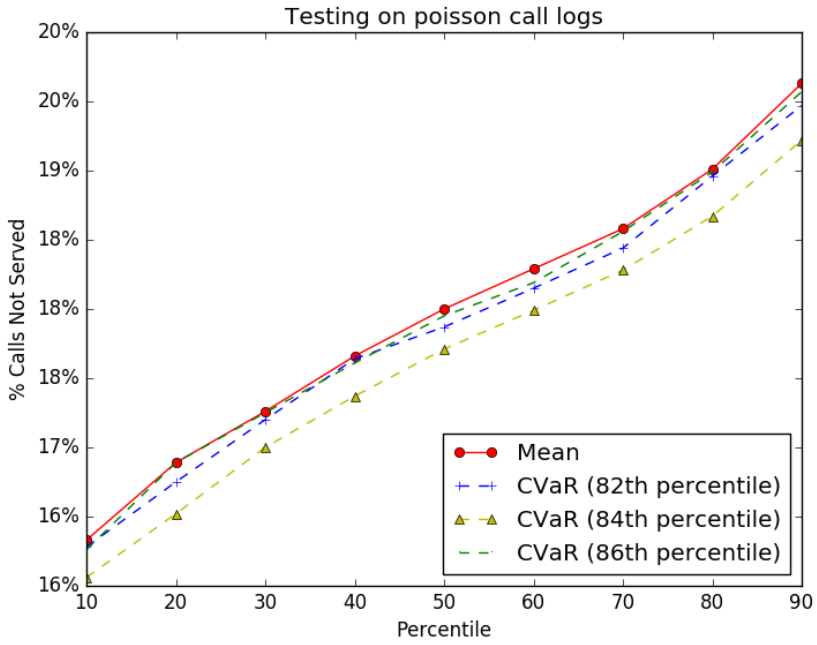

Figure 3.3: Performance of risk-optimized allocation on Poisson call log with hotspots ©2016 IEEE

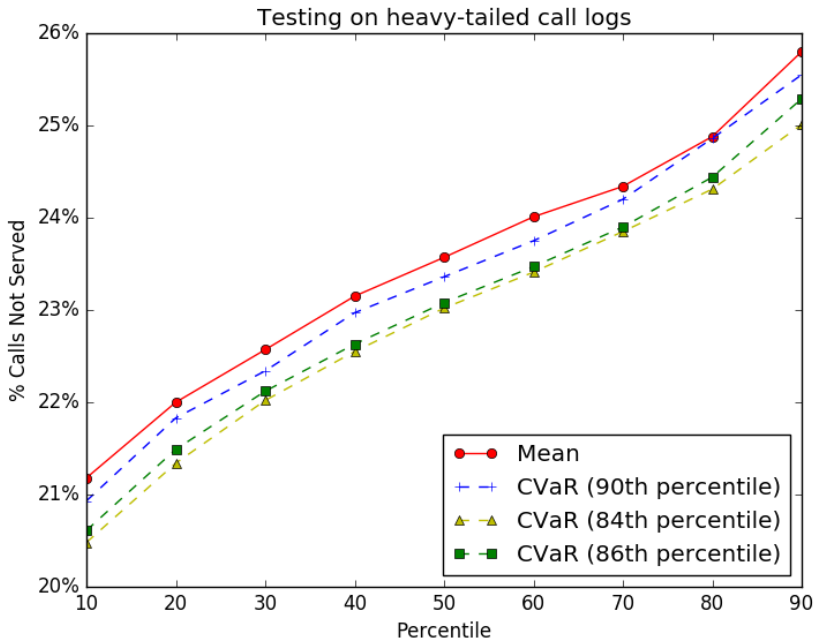

Figure 3.4: Performance of risk-optimized allocation on heavy-tailed call log with hotspots ©2016 IEEE

Our results show that optimizing with a combination of expected-value and risk-metrics provides improved results on tail-related metrics such as service failure probabilities, as compared to optimizing using expected value-based metrics alone. When there is imperfect information such as when the allocation was built assuming Poisson call arrivals but we see Weibull call arrivals in the system, or when there are hotspots, optimizing for the mean performed better. Also we observe that there is no single value of the protection level α that performs well for all cases, indicating the need to conduct further research to fix the right protection level.

Chapter 4

Conclusion

n this work, we have described data-driven models to maximize EMS service levels in emerging economies. We considered two problems specific to emerging economies. the first is that of customer abandonment of ambulances and the second is that of using risk metrics in the resource allocation of such systems.

We observed that due to the resource-constraints in emerging economies, including fewer ambulances and higher traffic, customers can depart to the hospital using other means of transportation without waiting for the dispatched ambulance to arrive at the scene. In this context, we explained the concept of *in-service abandonment* of ambulances and contrasted it abandonment in other settings like a call center. By using a semi-parametric Maximum Likelihood Estimation approach, we empirically estimated waiting time distributions for different classes of patients.

We then built upon the simulation based greedy optimization approach proposed by (Yue et al., 2012a) to design better networks to mitigate abandonment. With the current base configuration, we could not observe a significant decrease in ambulance abandonment even by tripling the fleet size and by varying the ambulance dispatch policy. To overcome this, we considered a problem where potentially every street corner can be considered a candidate base location. This resulted in an extremely large scale resource allocation problem which is two orders of magnitude larger than similar problems discussed in literature. We proposed a new network design consisting of a new set of bases and an allocation of ambulances across these bases, and demonstrated a 6 percentage point increase in calls served and decreased abandonment by 6 percentage points by tripling the fleet size. This can increase the number of calls successfully served to 94%, a considerable improvement over the current system.

Next, we presented an efficient approach to ambulance fleet allocation that

is data-driven and balances expected value-based and risk-based metrics. We show that maximizing with CVaR of gain in addition to expected gain for certain values of the protection level α helps generate more robust allocations than optimizing with the expected value of the gain alone. Further research needs to be done on examining the theoretical properties of the objective function and determine bounds for performance of this algorithm.

References

Aksin, Z., Ata, B., Emadi, S. M., and Su, C.-L. (2013). Structural estimation of callers delay sensitivity in call centers. *Management Science*, 59(12):27272746.

Aksin, Z., Ata, B., Emadi, S. M., and Su, C.-L. (2017). Impact of delay announcements in call centers: An empirical approach. *Operations Research*, 65(1):242–265.

Anderson-Bergman, C. (2017). An efficient implementation of the emicm algorithm for the interval censored npmle. *Journal of Computational and Graphical Statistics*, 26(2):463–467.

Batt, R. J. and Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dong, J., Yom-Tov, E., and Yom-Tov, G. B. (2017). The impact of delay announcements on hospital network coordination and waiting times.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141.

Gilboy, N., Tanabe, P., Travers, D., Rosenau, A. M., et al. (2012). Emergency severity index (esi): a triage tool for emergency department care, version 4. *Implementation handbook*, pages 12–0014.

Groeneboom, P. (1995). Nonparametric estimators for interval censoring problems. *Lecture Notes-Monograph Series*, pages 105–128.

Hathaway, B. A., Emadi, S. M., and Deshpande, V. (2017). Queue now or queue later: An empirical study of callers' redial behaviors.

Huang, J. et al. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169. Springer.

Krishnan, K., Marla, L., and Yue, Y. (2016). Robust ambulance allocation using risk-based metrics. In *Communication Systems and Networks (COMSNETS), 2016 8th International Conference on*, pages 1–6. IEEE.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM.

McCormack, R. and Coates, G. (2015). A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operational Research*, 247(1):294–309.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294.

Restrepo, M. (2008). Computational methods for static allocation and real-time redeployment of ambulances.

Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295.

Verweij, B., Ahmed, S., Kleywegt, A. J., Nemhauser, G., and Shapiro, A. (2003). The sample average approximation method applied to stochastic routing problems: a computational study. *Computational Optimization and Applications*, 24(2):289–333.

Yu, Q., Allon, G., and Bassamboo, A. (2017). How do delay announcements shape customer behavior? *Management Science*, 63(1):1–20.

Yue, Y., Marla, L., and Krishnan, R. (2012a). An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment. In *AAAI*.

Yue, Y., Marla, L., and Krishnan, R. (2012b). An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment. In *AAAI Conference on Artificial Intelligence (AAAI)*.