EVALUATING THE PERFORMANCE OF EMERGENCY MEDICAL SYSTEM IN THE US

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Zhila Dehdari Ebrahimi

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Transportation and Logistics

November 2022

Fargo, North Dakota

North Dakota State University

Graduate School

Title

EVALUATING THE PERFORMANCE OF EMERGENCY MEDICAL

SYSTEM IN THE US

By

Zhila Dehdari Ebrahimi

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Jeremy Mattson

Chair

Joseph Szmerekovsky

Mohammad Delasay

Kimberly Vachal

Linda Langley

Approved:

11/16/2022

Date

Tim Peterson

Department Chair

ABSTRACT

New and exciting opportunities are emerging for operational researchers to create and use models that provide managers with solutions to enhance the quality of their services as the importance of the service sector grows in industrialized countries. The key to this process is the creation of time-dependent models that analyze complicated service systems and produce efficient staff schedules, allowing organizations to strike a balance between delivering high-quality services and avoiding unnecessary personnel costs. There is a need, particularly in the healthcare sector, to encourage effective management of an EMS, where the likelihood of survival is strongly correlated with the response time.

Motivated by case studies investigating the operation of the Emergency Medical System (EMS), this dissertation aims to examine how operations research (OR) techniques can be developed to determine staff scheduling and maximize the ambulance to decrease service system delays. A capacity planning tool is developed that integrates a combination of queueing theory and optimization techniques to reduce the delay in the service system and maximize ambulance coverage.

The research presented in this dissertation is novel in several ways. Primarily, the first section considers the Markovian models with sinusoidal arrival rates and state-dependency of service rate and uses a numerical method known as Stationary Independent Period by Period (SIPP) to determine the staff requirement of the service system. The final section considers the time dependency in locating an ambulance station across the network and allocating the ambulance to the patients to cover more 911 calls.

ACKNOWLEDGEMENTS

I would like to thank the following people, without whom I would not have been able to complete this research: Dr. Jeremy Mattson, my academic advisor, whose insight and knowledge of the subject matter steered me through this research. I would like to convey my great appreciation to my committee members, Dr. Mohammad Delasay, Dr. Joseph Szmerekovsky, Dr. Kimberly Vachal, and Dr. Linda Langley. They have served as both professional and personal advisors to me. Their guidance has made my graduate studies both productive and enjoyable.

Last but not least, I want to express my humble gratitude to my parents, Fatemeh and Mansour, my brother Ebrahim and his wife, Neda, for all the support you have shown me through this research. Most significantly, I would like to acknowledge the love of my life, Mohsen Momenitabar, for his never-ending understanding, support, and love. This journey would be impossible without their unconditional support, compassion, and love.

DEDICATION

To my beloved family who patiently encouraged me from 6,400 miles away during my four and half year absent while working on this dissertation.

To my friend, husband and love of my life, Mohsen Momenitabar for his never-ending support, faith, and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Emergency Medical Service (EMS) refers to providing acute medical services outside of a hospital and transferring patients to hospitals by ambulance. EMS has evolved and expanded into a significant component of modern health care systems. Indeed, it is a sophisticated network for procuring, distributing, and transferring patients to hospitals in strictly time-limited emergency treatment ambulance units and medical care teams.

The EMS has two goals. First, it aids patients quickly before the condition becomes worse. Second, it ensures that the best service is delivered to the patients. Both goals have measurable impacts on patient outcomes. For the accomplishment of both goals, multiple emergency vehicles are often dispatched with a single service call. Multiple answers allow faster first-aid and guarantee that vehicles are effectively adjusted to the unpredictable requirement of the patient at the possible expense of unavailability of more vehicles compared to only sending one vehicle. If an emergency call is responded to by only one ambulance, then that same unit must still carry the patient to the hospital after on-scene treatment.

On the other hand, all vehicles dispatched may not be needed for the entire service process. After the respondents arrive at the scene, the patient's actual needs are usually apparent, and appropriate EMS unit response becomes more evident. Therefore, modeling the service process in two phases is preferable since certain response vehicles are released during the service cycle rather than a scenario with a single response unit (Yoon & A.Albert, 2020).

1.1. Motivation

The nation's emergency and trauma treatment system highly depends on EMS. Over 16 million medical transports are performed annually by hundreds of thousands of EMS staff. EMS covers the early phases of the emergency care continuum. It comprises dialing 911 in case of an

emergency, dispatching emergency responders to the scene of a crash or illness, and treating and transporting patients by ambulance and air ambulance. The effectiveness and quickness of emergency medical services play a crucial role in the final result for a patient. However, EMS care has some strengths, including automatic crash notification technology and improved medical equipment such as air ambulance service, which is helpful for rural areas where accessibility is so complex or for patients with severe conditions, some systemic problems in the EMS.

The main weakness of the current EMS system is insufficient coordination. In some cases, EMS and other public safety services cannot communicate with one another because they use separate frequencies or incompatible communications technology. So, patients may not be delivered to facilities that are ideal and prepared to receive them because there is insufficient coordination of transport between regions, which leads to inadequate management of the regional flow of patients.

Moreover, disparities in response times are another limitation of the current EMS. The selection of which ambulance needs to respond to emergency calls is highly valuable. An inherent difficulty in EMS is deciding how to distribute scarce resources geographically. The spatial and temporal fluctuation in input elements, such as demand and journey time, must be considered when planning ambulance locations. Demand for ambulance services differs by region and by an hour of the day. Moreover, the time from dispatching the ambulance to arriving at the patient's location also depends on the time. For instance, in crowded areas, ambulance travel lengths are short, but delays may occur due to traffic and other issues, whereas in rural areas, travel distances are more significant, and the terrain can occasionally be challenging.

According to these two main weaknesses of the current EMS system, I investigate to develop Operations Research (OR) techniques to evaluate the service system by determining the

number of required staff in each period, leading to the lowest delay in the system. Also, the limitation of the current service system persuades me to analyze the optimum ambulance locations by using the location-allocation problem to provide services to patients.

1.2. Importance of Research

In the past few decades, the importance of the service sector, including EMS, has expanded in many industrialized countries. As the service sector changes, there is more competition between organizations. Therefore, their managers are trying to evaluate their services to allow them to provide their customers with services as quickly and cheaply as possible. OR approach has dramatically developed in response to the increased evaluation. While the mathematical foundations of stochastic modeling, heretics, and optimization approaches have all been well established, a significant effort has been made to develop theory through the application of models (Ingolfsson, et al., 2010; Izady & Worthington, 2012). In particular, the queueing theory that assists managers in making judgmental decisions based on the resources needed to deliver an efficient service through the disclosure of analytical system quality details in a variety of scenarios has moved from basic server queue analysis to consistent random arrival rates and time-dependent systems (Erlang, 1917). While queueing models can be used to decide on the resources required for the service, their effectiveness depends on the prediction of demand for the system. Thus, an in-depth evaluation of the EMS system involves a mixture of a broad range of statistical and OR techniques ranging from forecasting demands, evaluating the number of staff, and advice on improving service effectiveness by producing staff numbers to ensure the appropriate number of staff at any time.

In this sub-section, I will discuss how applying some methods in these upcoming two chapters can solve the problem in the EMS. Healthcare sector managers are responsible for

promptly delivering services to their patients. Sometimes, they are faced with insufficient staff during peak service times. So, they need to have flexible scheduling to provide better service to their patients to increase the lifetime of the patients. Similarly, in the EMS, assigning ambulances to the patients who call 911 is of high priority to avoid the death of patients. In this way, managing how many ambulances are needed is an important matter that I will address them. The second chapter of my proposal answers this question: how many servers (ambulances) should be assigned to receive different 911 calls?

Applying the method called Stationary Independent Period by Period (SIPP) to analyze the performance measures of queueing systems, including the average number of half-hours with missed service level targets and an average maximum delay service level target, lead to determining the proper system decisions in staffing level and 911 call response.

One of the aims of the EMS is to provide services to the broader area of a received 911 call. The service to the patients will be assigned based on their priorities. So, the third chapter proposes an optimal location of ambulance stations across all demand zones and allocates them to the demand zone to maximize the service coverage to the patients. In this case, the service coverage maximization will result in losing the patients' reach out to the EMS.

1.3. Purpose of Research

This dissertation addresses the urgent needs of patients when 911 calls are received. This proposal primarily aims to investigate how the Operation Research (OR) technique and statistical analysis can be developed to analyze service systems subject to the demand that is urgent and heavily time-dependent.

In the second chapter, I propose a delay probability function in the EMS system to calculate the number of servers (ambulances) in each period. Indeed, the second chapter aims to determine

the number of required staff in each period, leading to the lowest delay probability rate. Also, the third chapter of the dissertation aims to consider the time dependency in locating an ambulance station across the network and allocating the ambulance to the patients to cover more 911 calls. The time dependency in location models is related to variations in travel time, availability of ambulances, and demand. Also, the model of this study is realistic since no study dealt with the maximization of service coverage along with time dependency in travel time, availability of ambulances, and demand.

1.4. Problem Statement

Previous studies investigated EMS in different scopes, but I identified three main gaps. First, most studies focused on performance measures such as delays in the EMS system to decrease the patient's response time. Although most studies evaluated the service rate as constant while the arrival rate is either constant or dependent on time, I assume that the arrival rate is time-dependent, which changes in each period, and the service rate has a state-dependent condition. Therefore, in the second chapter, I focus on capacity planning of service systems with cyclic arrivals and state-dependent service rates to determine staffing requirements to decrease the delay in the service system.

Third, studies worked on the location-allocation problem to find the ambulance's optimum location and provide patient services. Other studies applied the repositioning model to relocate the ambulance location to deliver services in the shortest time. In this study, I assume each area has demand defined by the number of patient calls, which need to be responded to as soon as possible to avoid the loss of the patient by establishing locations for ambulance stations to dispatch them to the scene. So, the problem is determining the optimal location of ambulance stations across all

demand zones and allocating them to the demand zone to maximize the service coverage to the patients.

2. CAPACITY PLANNING OF SERVICE SYSTEMS WITH CYCLIC ARRIVALS AND STATE-DEPENDENT SERVICE RATE

2.1. Introduction

Managers of service establishments where client demand for assistance is cyclic and random frequently change employee levels to ensure a consistent level of service. In an Emergency Department (ED), long waiting times and delays are undesirable because they cause patients unnecessary agony and cause care to be delayed. Overcrowding, extended waiting times, and resultant delays are common in many countries' emergency departments. In China, top-rated hospitals' emergency rooms are constantly congested, and patients face long waiting times (Wang & Xu, 2007). A complex connection between the demand for healthcare services and the availability of healthcare services results in long waiting times (Xie & Or, 2017).

An ED, unlike a clinical department, provides continuous emergency medical care without planned patient appointments, which is one of the key reasons for overcrowding and delay. The ED system is briefly overburdened during peak arrival periods since patient arrivals are frequently stochastic and time-varying throughout the day. On the other hand, the lack of a physician scheduling mechanism to effectively address demand changes exacerbates ED congestion. So, because implementations must take into consideration complex scheduling limitations, developing staffing schedules in such service systems can be problematic. Employees' preferred start times, quit times, and shift durations are all respected, as are policy limits on the number of consecutive hours and days worked, and so on. One of the essential requirements is that there always be enough people on duty to meet the service levels that have been established. In this case, I develop the practice of determining capacity planning to determine the staff in the service system while random

cyclic demands and state dependency service rate by using stationary queueing models. Green et al. (2001) proposed the Stationary Independent Period-by-Period (SIPP) approach to staffing.

In this approach, staffing requirements are determined by dividing the workday into "planning periods" such as shifts, hours, etc. One model for each planning period, a sequence of stationary queueing models is built, which is M/M/s type models. Each model is solved separately to find the smallest number of servers required to satisfy the service target in that time frame. The SIPP technique divides the time of interest into small periods. Each discrete period is subjected to a distinct stationary Markovian queueing model with the average arrival rate as the input parameter. Stationary models can estimate the minimum personnel required based on this approximation.

This chapter aims to determine capacity planning in ED systems with random cyclic demands and state-dependent service rates using the SIPP approach. The model of the SIPP approach is $M(t)/M_{s(t)}/s(t)$ system, while the arrival rate is sinusoidal and the service rate is state-dependent, which refers to the number of servers (staff) at each time. Also, I analyze to measure the performance of the service system during load effect on service rate. The service rate function has two scenarios, including decreasing scenario, where service time decreases with a load, such as social loafing, and the increasing scenario, where service time increase with a load, such as social pressure, has been considered.

The contribution of this chapter is to analyze the performance measure of the SIPP average, which relates to standard SIPP. At the same time, the service rate has state-dependency behavior and ignores the state-dependency of the service rate. Second, develop the state dependency of the SIPP approach with two models, including SIPP Max and SIPP Mix. Third, evaluate the reliability of staffing level and SIPP reliability of models.

2.2. Literature Review

Even though operations research and management have considerably improved vital EDs, staffing and scheduling for stochastic and time-varying patient arrivals remain challenging (Salmon, et al., 2018; Saghafian, et al., 2015). Despite being aware of the day-to-day fluctuation, hospital management allocates staff based on gut instinct and general perceptions, with no methods to assess the impact on critical patient service quality criteria (Green & Soares, 2007). Because different existing systems may experience significant time-dependent changes in their parameters, it is appropriate to enable queuing model parameters like user arrival rates and server count to alter over time. Even though this phenomenon has been seen in various real-world systems, the literature on the issue is relatively young and sparse.

The demand for emergency rooms is influenced by a variety of unpredictable factors, such as the unpredictability of emergencies and the length of surgeries. As a result, some work that addresses physician scheduling issues takes stochastic, time-varying demand into account. Tan et al. (2013) credited Newell with the first studies on time-dependent queues, which were published in a series of papers in 1968 that looked at time-dependent arrival and service rates. According to Schwarz et al. (2016), Kolmogorov's work in 1931 was the first to analyze time-dependent queuing systems. Since then, various studies have looked at time-dependent behavior in a range of queuing systems, including hospitals (Bruin, et al., 2007), emergency medical services (Beojone & Souza, 2020; J.L.Vile, et al., 2016; Singer & Donoso, 2008), call centers (Atlason, et al., 2007; C.Dietz, 2011; Green, et al., 2001). In order to address the time-varying component of physician scheduling issues, Ingolfsson et al. (2002) first developed an integer program model with the stationary independent period-by-period (SIPP) assumption. For the purpose of determining the best schedule with time-varying service levels, Ingolfsson et al. (2010) presented a two-step process that

included both a schedule generator and a schedule evaluator. Green et al. (2006) examined emergency department staffing using the queuing theory.

Moreover, Green et al. (1991) proposed the nonstationary Pointwise Stationary Approximation (PSA) time-varying queues, which assume that the queue reaches a steady state throughout each period and that the steady-state findings can roughly predict the performance of the queue. The PSA is a weighted average of the performance measurements at each point in time (the interval length is set to zero). Whitt (1991) employed the PSA technique for an M(t)/M(t)/c system.

Some approximation methods based on models with piecewise constant input parameters into each time interval within the overall time horizon exist. The SIPP approach divides the entire period into small, independent periods and uses the average arrival rate for each period as the input to a series of stationary studies. Several scholars have used the SIPP to evaluate the behavior of M(t)/M/c(t) queuing systems, such as in the work of Green et al. (2001) and Atlason et al. (2007). The stationary approximation is the most widely used method for assessing performance in a nonstationary queueing system with time-varying demands. The stationary approximation approach converts nonstationary system parameters into stationary counterparts that are then fed into a (series of) stationary models (Defraeye & Nieuwenhuyse, 2016). There have been numerous proposed stationary approximations. Green et al. (2009) worked on the SIPP method to evaluate customer satisfaction and reach out to target customer services. They find that the SIPP method cannot ask about customers' time lag when it will be between peak demand and when the system has peak congestion. So, they test two SIPP approaches to find which achieves the service target with only an increase in staffing. Green et al. (2001) also enhanced the basic SIPP approach by using a time lag between peak demand and peak congestion to improve a basic SIPP's performance.

This change considers that system congestion peaks after arrival rate peaks in nonstationary systems. Green et al. (2006) investigated the performance of the ED queueing model and identified how to modify staffing levels to meet changing needs in an ED of an urban hospital in the United States using the lagged version of the SIPP technique. Green et al. (2006) advocated staffing in EDs, which has reduced the number of patients who leave the ED without being seen. Green et al. (2009) discovered that the basic SIPP approach's suggested staffing in telephone call centers is insufficient to provide customer service levels during critical periods. They designed and evaluated two simple lagged SIPP adjustments that met the service target while requiring only minor staffing additions. With a focus on call centers, Koole et al. (2003) proposed a local search technique for joint staffing and shift schedules with a total service level target, where the SIPP approximation is used to evaluate scheduling performance.

In systems with rapidly fluctuating arrival rates, many demand peaks, lengthy wait times, and constrained operating hours, the SIPP technique is known to perform poorly (Green, et al., 2001; GREEN, et al., 2009; Ingolfsson, et al., 2010). In the extended lag-SIPP approach, the mean service time is used to lag the arrival rate consideration intervals to account for demand carryover (Green, et al., 2001; GREEN, et al., 2009). Both SIPP and lag-SIPP optimize staffing individually for each time interval and do not account for the effects of staffing decisions made for earlier intervals, which can produce a less precise result (Ingolfsson, et al., 2010). Furthermore, steady-state equations are unable to account for customers who are not seen because of the end of the day in systems with limited operating hours and client abandonment.

Another approach to time-dependent small interval staffing is Modified Offered Load (MOL) approximation. In the MOL, a related time-dependent queueing model with an infinite number of servers provides guidance for staffing considerations. Through the use of proxy steady-

state queueing models based on the anticipated number of model customers in the infinite server system at any one time, the MOL approach expands this idea to other performance indicators (Feldman, et al., 2008; Jennings, et al., 1996). The MOL approach does not take into account the effects of staffing decisions made in previous time intervals, but it does take into account the interaction between service time for prior arrivals under a time-dependent arrival rate. Sinreich et al. (2007) calculate small interval staffing for numerous medical resources in an ED care network based on the highest anticipated number of busy model servers in each interval of an infinite server simulation.

Moreover, in telecommunication and service systems, some classical models are utilized for capacity planning (Robbins, et al., 2010; Inman, 1999). They are also widely employed in studies on production and service systems, which assume that the service time distribution parameters are exogenous—unrelated to the state of the system.

Also, some other researchers found that the load, which describes a measure that expresses how occupied or congested a system is at a certain moment, has an effect on service time in the service system (Batt & Terwiesch, 2016; Delasay, et al., 2019). They analyzed some mechanisms, including fatigue, increasing the task of servers, and increasing the server's work contents which push servers to decrease their service speed. Jaeker et al. (2017), and Karaue et al. (1993) analyzed that whenever monitoring the individual server's task was so hard, and servers had less effort, which resulted in decreasing the service rate. Also, researchers investigated the incapacity of servers to work hard for a long time at a high-level causes server to decrease the service speed (Kc & Terwiesch, 2009; KC & Terwiesch, 2012). They found that being overworked and having any task varieties in the system impact decreased server speed (Staats & Gino, 2012; Kc & Terwiesch, 2009). Louriz et al. (2012) and Long et al (2018) investigated a backlog of consumers waiting to

move on to a downstream service, causing congestion in a service that has decreased the service speed. Tan et al. (2014) and Goes et al. (2018) proposed that distributing server resources among numerous customers causes a decrease in service speed due to each customer having full attention from their servers.

On the other hand, researchers proposed that some other mechanism, such as a reduction in the server's task and service cancelation, causes increasing the service speed (Batt & Terwiesch, 2016; Kc & Terwiesch, 2009; KC & Terwiesch, 2012). Also, they investigated that decreasing the server's tasks causes servers to increase their service speed (Schultz, 2003; Batt & Terwiesch, 2016; Long & Mathews, 2018). Batt et al. (2016) proposed that whenever servers performed some tasks earlier than usual, the service speed increased.

Although the capacity planning problem has attracted much attention in previous studies, a few studies worked on the SIPP approach. Moreover, most previous studies worked on sinusoidal arrival rate, which did not pay attention to service rate, and kept it constant. While in this chapter, the transient analysis using the SIPP approach has been proposed to evaluate the staff requirement in the service system. Moreover, the SIPP approach has been developed with a cyclic arrival rate, which is time-dependent, and the service rate is state-dependent, which refers to the number of busy servers or staff at the time, into two decreasing and increasing scenarios, where the service rate decreases or increase with load. In this case, I evaluate the performance measure of the SIPP approach for staffing in the service system. Accordingly, the main contributions of this chapter are as follows:

- Transient analysis of delay system in service system with cyclic and time-varying of arrival rate,
- Considering the state dependency on service rate,

- Evaluating the SIPP average, with and without state-dependency of service rate,
- Improving the SIPP average by using the SIPP Max and SIPP Mix, and
- Comparing the performance measures for SIPP- based staffing.

2.3. Methodology

The queueing theory examines how systems allocate their resources to customers who progressively show up at a service facility in order to receive a service. Due to the fact that resources are not always readily available, queues frequently form. Queueing theory tries to predict system performance and assess the level of service that consumers might anticipate receiving in various scenarios. The theory spends a significant amount of time developing performance measurements that assess factors including throughput, delay probability, the number of customers in the line, and predicted queue length (Bhat, 2008). Queueing theory can be used to optimize resource allocation and suggest staffing levels when the goal is to strike a balance between service quality and financial factors. This will prevent queues from building up too much while ensuring that servers are active for a reasonable amount of time.

According to queueing theory, a service system can be thought of as consisting of two components: 1) the service facility itself, which may be staffed by several servers, and 2) a line for service (apart from situations where it may be explicitly stated that queuing is not authorized). Customers arrive at each location and stand in line for a certain activity.

Figure 1. The fundamental process of queuing theory

The two main elements that define every queueing network are the arrival process and the service process. Setting up mathematical models that correspond to Figure 1, analyzing the system,

and assessing various performance indicators are all part of queueing theory. Numerous factors will affect the system's performance, and queueing theory is based on probabilistic analysis because these processes are typically stochastic by nature.

The arrival process specifies the manner in which customers arrive at the service facility (for example, separately or in groups) and the order in which they come. The Poisson process is frequently chosen to represent random client arrivals since it produces uniformly distributed arrival times with an arrival rate.

The resources required for service are specified in the service mechanism. The service time distribution determines how long the service will take, but other factors like the number of servers available and whether they are in series (each server has its own queue) or parallel (one queue for all customers) must be known before a meaningful analysis of the system can be done. In situations where it is assumed that the exponential distribution accurately represents the distribution of service times, one can translate the system to a continuous-time Markov chain that can be solved analytically according to its Markovian (memoryless) feature. The exponential distribution has only one parameter, just like the Poisson distribution. It is conventional to represent the mean service rate by μ, with $\frac{1}{\mu}$ standing in for the mean service time in order to distinguish between the mean arrival and mean service rate.

The additional notation that will be utilized throughout this thesis that is frequently used in the literature to analyze queueing systems is as follows:

p_n: The probability of n customers in the system,

s: The number of servers on duty,

The quantity $p = \frac{\lambda}{s\mu}$, also known as server utilization rate, traffic intensity, or load per server, is a widely used indicator of interest that illustrates the queue's behavior over time, whilst,

$r = \frac{p}{s}$ calculates the system's offered load or the volume of traffic in the queue (Gross, et al., 1998). Given that all servers have the same service rate, the link between these values and the system capacity provides useful insight into the system's performance. In essence, if $p \leq 1$ then the servers are able to process customers faster than the rate at which customers typically arrive, preventing the queue from getting infinitely long. All system characteristics, including the number of users, the length of the queue, and anticipated wait times, will eventually settle down, and the system will operate at a consistent level, which is referred to as "stable" or "stationary" if it runs with a mean service level and inter-arrival rate with $p \leq 1$ for an appropriate amount of time. It is referred to as running in a steady state when the system reaches this point.

A birth-death process can be considered a continuous time stochastic counting process $\{N(t), t \geq 0\}$. Letting $p_n(t) = Prob\{N(t) = n\}$ be the probability that the system is in state n at time t, the transition flow of the birth-death process may be shown in Figure 2. A birth causes the system to move from state n to n + 1, while a death causes it to move from state n to n-1. The process is specified by birth rates $\{\lambda_i\}_{i=0,\dots\infty}$ and death birth rates $\{\mu_i\}_{i=1,\dots\infty}$.

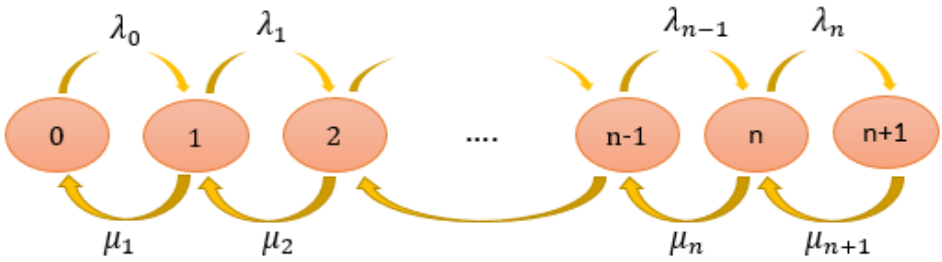

Figure 2. Transition diagram of the birth-death process

In this chapter, I developed the M/M/s model, which is one of the most extensively studied models in the traditional queueing literature because it simultaneously captures unpredictability in arrival and service times, allows for more than one server, and has the appealing advantage of a tractable steady-state solution. It depicts a system with a single queueing location where clients

come and may queue before being served by one of the s same servers. Arrivals follow a Poisson process that is time-homogeneous and has a constant rate, whereas service time has an exponential distribution with a constant mean.

So, I developed the $M(t)/M_{s(t)}/s(t)$ queuing systems, where M(t) represents exponential, independent, and identically distributed arrival rate, which is cyclic and time-dependent; $M_{s(t)}$ represents the distribution of service rate that is dependent on a number of servers on duty, and s(t) represents the number of servers at time t. The function arrival rate has the sinusoidal function, where A>0 is the amplitude and λ is the arrival rate that changes over time (Equation 2.1).

$$\lambda(t) = \lambda + A \times \sin(2\pi t / 24) \tag{2.1}$$

The other model parameters are μ (t), the service rate at time t and s(t), and the number of servers scheduled at time t. The service rate function has two increasing and decreasing functions where the α is a slope for the μ that is negative for decreasing and positive for increasing approaches. In this case, by having decreasing and increasing approach, I can evaluate the performance measure in our model to analyze how the delay decrease or increases and also know in which states they happen. In other words, it is assumed that the service rate has a state-dependent function into increasing scenarios, such as network dispersion, in which the service rate increases with load, and in decreasing scenarios, such as social loafing, in which the service rate decreases with the load. Also, the service rate function in both mentioned scenarios is presented as follows:

$$\mu_{s(t)} = \min\left\{\mu_{\min} + \alpha \times s(t), \mu_{\max}\right\} \tag{2.2}$$

$$\mu_{s(t)} = \max\left\{\mu_{max} + \alpha \times s(t), \mu_{\min}\right\} \tag{2.3}$$

The μ_{min} and μ_{max} are the server's average minimum and maximum service rates, respectively; μ_{min} represents the rate when the system is staffed by one server, and the average service rate increases linearly with the staffing level, at a positive slope α, till it stabilizes at μ_{max} in Eq. (2.2) and in decreasing scenario, the rates μ_{min} and μ_{max} are the servers' average minimum and maximum service rates, respectively; μ_{min} represents the system with one server on duty, and the average service rate decreases linearly with the staffing level, at a negative slope α, till it stabilizes at μ_{max} in Eq. (2.3).

The number of customers in the system is an appropriate state variable in queueing systems that directly model the behavior of individuals who arrive at a service facility in need of a specific service to be provided; as a result, $p_n(t)$ be the periodic steady-state probabilities that there are n customers in the system at time t. The λ_t is average arrival rate that is calculated from (Equation 2.1). Also, I use the randomization method (Gross, et al., 1985) to solve the differential numerical equations system (Eq. (2.4)), which is given in Eq. (2.5) to obtain $p_n(t)$ probabilities.

$$\left\{\begin{array}{ll} P_0^{'}(t) = -\lambda(t).P_0(t) + \mu_{s(t)}.P_1(t) & \\ P_n^{'}(t) = -\lambda(t).P_{n-1}(t) + (n+1)\mu_{s(t)}.P_{n+1}(t) - (\lambda(t) + n.\mu_{s(t)}).P_n(t) & 1 \le n \le s(t) \\ P_n^{'}(t) = \lambda(t).P_{n-1}(t) + s(t).\mu_{s(t)}.P_{n+1}(t) - (\lambda(t) + s(t).\mu_{s(t)}).P_n(t) & n \ge s(t) \end{array}\right\} \quad (2.4)$$

Moreover, $P_D(t)$ is the instantaneous delay probability that a customer arrives at time t. This is also the probability that all servers are busy at epoch t and is given by Eq. (2.5).

$$P_D(t) = 1 - \sum_{n=0}^{s(t)-1} P_n(t) \quad (2.5)$$

The analytic sequence for each scenario is as follows:

- Used the fix scenario's parameters, including λ, the mean arrival rate; μ, the service rate; RA= A/ λ, the relative amplitude; τ, the target probability of delay; and PP, the length of the planning period.
- Divide each workday into planning periods of length PP (e.g., one hour) and compute the average arrival rate λ_{PP} by integrating $\lambda(t)$ in Eq. (2.1) over the interval.
- For each PP, an approximation of the real $M(t)/M_{s(t)}/s(t)$ system by a stationary $M/M_{s(t)}/s(t)$ model with the fixed arrival rate λ_{PP} and state-dependent service rate $\mu_{s(t)}$ to find the minimum staffing level S_{PP} that satisfies the target delay probability τ.
- Simulate the real system at 5-minute intervals by numerically solving Eq. (2.4), keeping the staffing level fixed at S_{PP} during each period.
- Use the $P_n(t)$ probabilities to compute the average delay probability.

Table 1 shows the value of each parameter used in the model. Also, the positive and negative alpha values are for increasing and decreasing approaches. I analyze a complete factorial trial design to test SIPP reliability, with a total of 1,944 possibilities for each approach (parameter combinations).

Table 1. Experimental Setting

Parameter setting	Value
Average arrival rate: λ	{2, 4, 8, 16, 32, 64}
Maximum service rate: μmax	{1.25, 1.5, 1.75}
Target delay probability: τ	{0.05, 0.1, 0.2}
Relative amplitude: RA	{0.1,0.5,1}
Length of planning period: PP	{0.25, 0.5, 1, 2}
Service rate slope: α	{−0.08, −0.04, −0.02, 0.02, 0.04, 0.08}

In the experiments, I set μmin = 1, and the service rate function adheres to Eq. (2.2) and Eq. (2.3). I execute each scenario using the above-described SIPP process, and I then keep track of how many half-hours (denoted by u) occur when the average delay probability is at least 10% higher than the target. I regard SIPP to be reliable for a scenario if the corresponding u = 0 and to be unreliable if u > 0 following (Green, et al., 2001).

2.4. Result and Discussion

2.4.1. SIPP State-dependency

In this section, I investigate the SIPP reliability where the service system has a state-dependent model. To have a state-dependent model for service rate, the queuing model is $M/M_{s(t)}/s(t)$, where the service rate is $M_{s(t)}$, which depends on the number of servers at time t, which is s(t). The state-dependent model is analyzed as two different increasing and decreasing scenarios.

2.4.1.1. Increasing Scenario

Figure 3 illustrates the SIPP approach's state-dependent for increasing approach. The figures show how each parameter affects the system's SIPP reliability when the other parameters

are different, as shown by solid orange lines. As the typical arrival rate rises, SIPP becomes less reliable. When RA = 1, the SIPP reliability is severely reduced by an increase in RA, with the unreliability of 83.95%. The performance of the SIPP is not significantly affected by changes to the service rate-dependency μmax and α. This fact demonstrates that in all experiences, the μmax never achieves its maximal condition. Since personnel levels are set over more extended planning periods, the SIPP is typically less accurate, and the changes in arrival rates are undermined. As expected, SIPP performs better as the target QoS is more relaxed (i.e., higher τ).

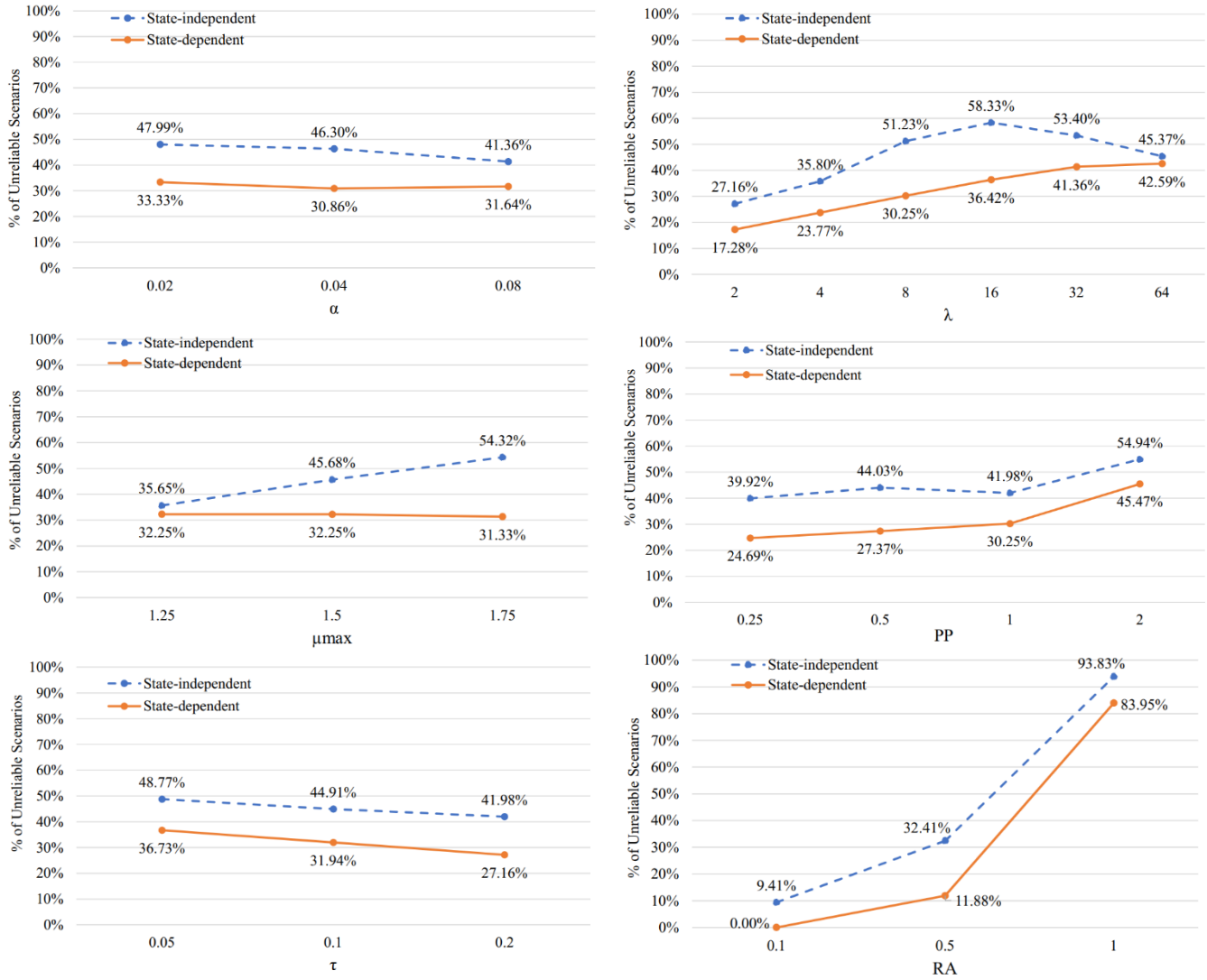

Figure 3. Effects of parameters on increasing SIPP reliability- increasing approach

2.4.1.2. Decreasing Scenario

Figure 4 summarizes the state-dependent SIPP approach's result for decreasing approach. The solid orange lines show the influence of each parameter on the SIPP reliability when other parameters change. When the arrival rate increases, the SIPP reliability decreases. Moreover, for higher RA, the reliability of SIPP decreases, where the unreliability of RA is 81.17% at RA=1. Like SIPP in an increasing approach, μmax and α do not significantly affect SIPP reliability where μ never reaches a maximum value. For higher PP (PP=2), the unreliability of SIPP is near 47%. Moreover, the SIPP reliability is increased for the higher probability of delay.

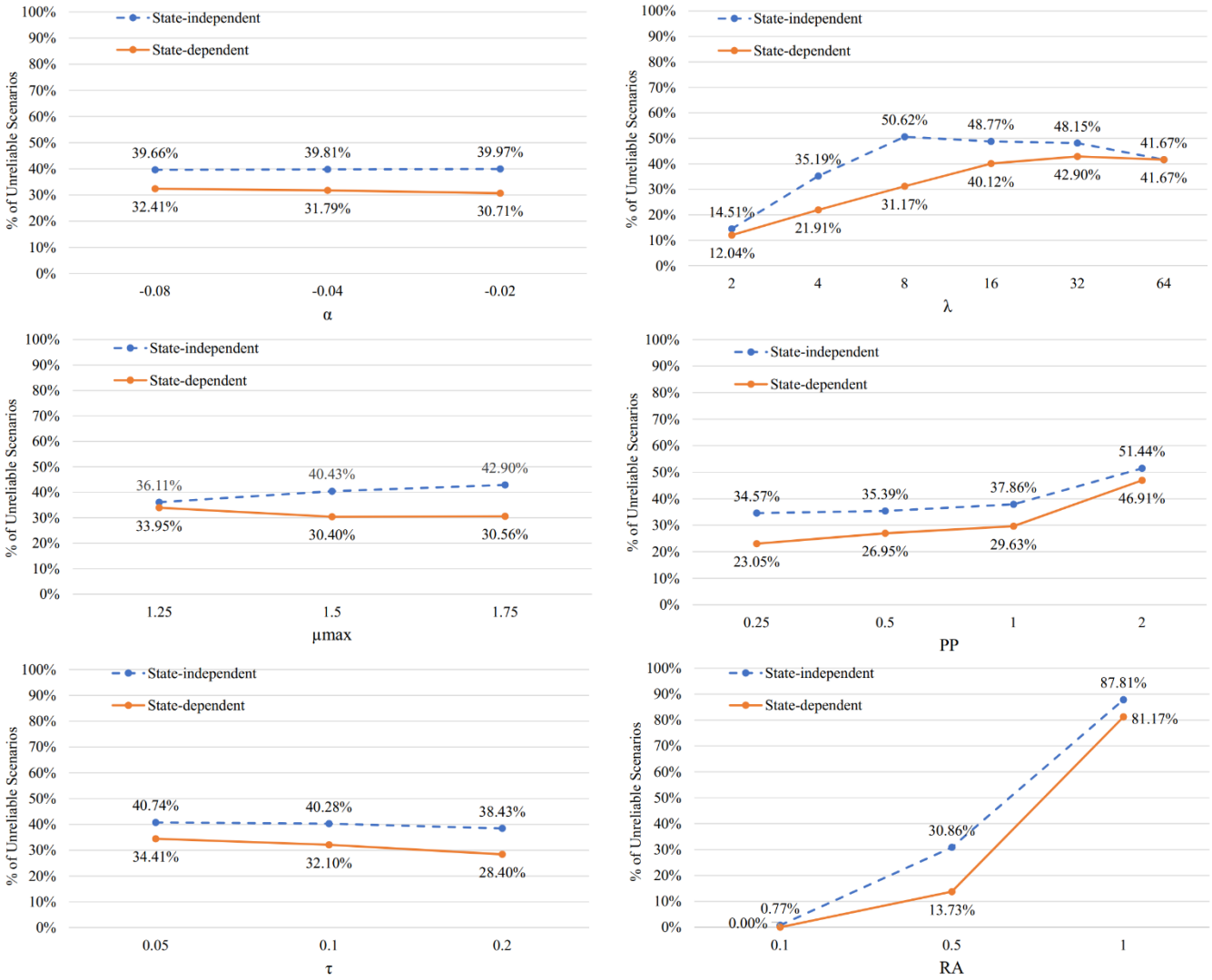

Figure 4. Effects of parameters on increasing SIPP reliability- decreasing approach

2.4.1.3. Classification Tree

Figure 5 and Figure 6 represent the SIPP unreliability for increasing and decreasing approaches. According to Figure 5, the SIPP is unreliable in increasing approach while the scenario has RA = 0.1, 0.5, PP = 0.25, 0.5, 1, and lambda = 2, 4 decreased. Also, the scenarios with RA = 1, lambda= 2, 4, and target probability delay = 0.1, 0.2, and PP = 0.25, 0.5 has a high SIPP reliability. On the other hand, Figure 6 shows the SIPP unreliability for decreasing scenarios; for RA = 1 and lambda> 4, SIPP has unreliable. Also, the scenarios with PP= 0.25, 0.5, and 1 have a low SIPP reliability.

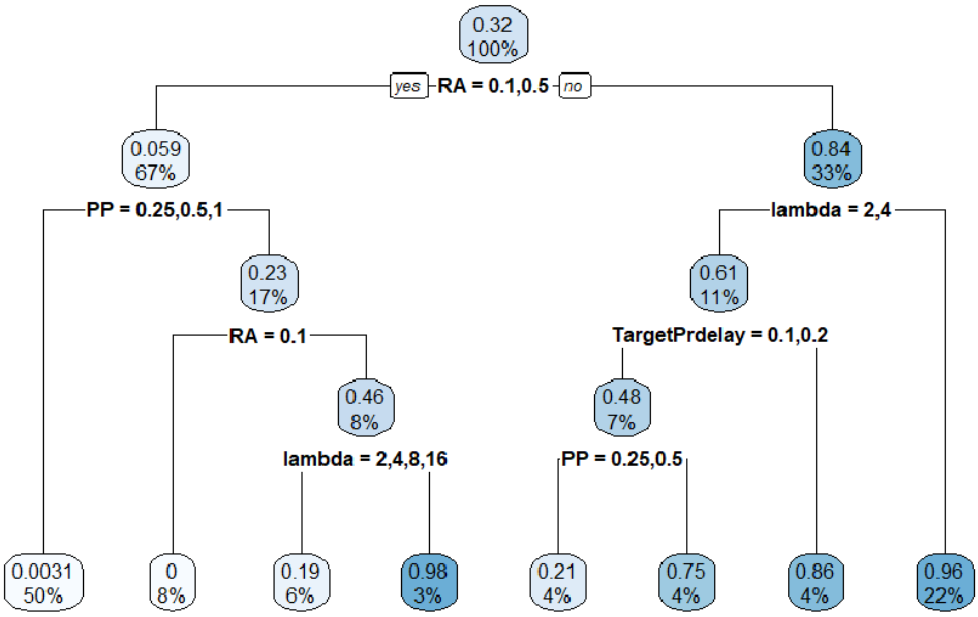

Figure 5. Classification tree- Unreliability of state-dependent model-increasing approach

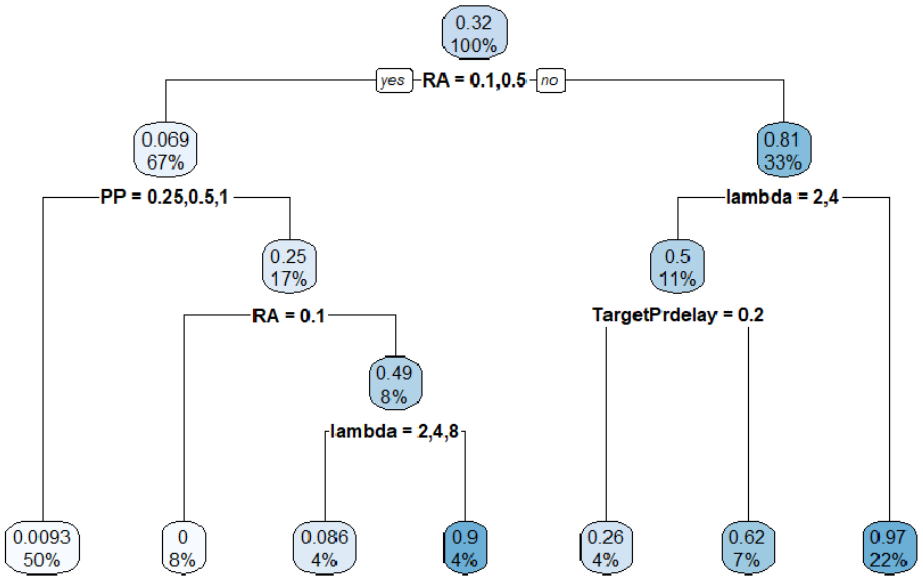

Figure 6. Classification tree- Unreliability of state-dependent model-decreasing approach

2.4.2. Ignoring the SIPP State-dependency

In this section, I try to examine the SIPP reliability while I ignore the state-dependent service rate model. I ignore the state-dependency service rate model and use an M/M/s(t) (instead of an $M/M_{s(t)}/s(t)$) system in step (ii) of the SIPP approach, considering the average system service rate for all states (I refer to this approximate model as state-independency).

2.4.2.1. Increasing Scenario

The dashed blue lines in Figure 3 illustrate how the system performance would decline if step (ii) of the SIPP approach used an M/M/s(t) system rather than an $M/M_{s(t)}/s(t)$ system, taking into account the average system service rate for all states (I refer to this approximate model as state-independency). As the arrival rate (and consequently, system load) rises, the effects become more severe, resulting in 45.37% unreliability at $\lambda = 64$. When RA = 1, the state-dependent SIPP is already so unstable that ignoring state-dependencies does not significantly worsen the performance. To complement my findings, I look into system parameter combinations where

neglecting state-dependent service rates significantly lower QoS. To improve the SIPP approach's reliability for the system, I will change it in addition to my analyses.

2.4.2.2. Decreasing Scenario

According to Figure 4, neglecting the state-dependency of service rate has no positive effect on reliability. The dashed blue line in the figure indicates ignoring the state-dependencies model of service rate. Like the increasing scenario, I used an M/M/s(t) (instead of an $M/M_{s(t)}/s(t)$) system in step (ii) of the SIPP approach, considering the average system service rate for all states. The result shows that while λ increases, the reliability is not superior to the state-dependent service rate strategy. The SIPP reliability is unaffected by changing the μ_{max} and when the unreliability of μ_{max} is increasing continuously. Because the state-dependent technique is already highly unreliable for RA = 1, the state-dependency has no impact on SIPP reliability. The dependability of SIPP remained unchanged for longer PP and higher delay probabilities.

2.4.3. Reliability with/without SIPP State-dependency

Table 2 and Table 3 show the average amount of system delays for two increasing and decreasing scenarios. According to these tables, the bold font indicates that the average number of half-hour target probability delays by at least 10% for the state-independent model is higher than for state-dependent ones. So, it is proof that ignoring the state-dependency of service rates will make SIPP unreliable. Indeed, ignoring the state-dependency of service rate increases the probability of delay.

Table 2. Reliability/unreliability of state-dependency and independency (based on a half-hour target delay of 10%- increasing approach)

State-dependent	State-independent		
	Reliable	Unreliable	Total
reliable	1057	266	1323
Unreliable	8	613	**621**
Total	1065	**879**	1944

Table 3. Reliability/unreliability of state-dependency and independency (based on a half-hour target delay of 10%- decreasing approach)

State-dependent	State-independent		
	Reliable	Unreliable	Total
reliable	1164	165	1329
Unreliable	6	609	**615**
Total	1170	**774**	1944

Furthermore, Table 4 indicates the total staffing requirement in the SIPP average with and without the state-dependent model. The number of staffing for the state-dependent model is 0.6% more than staffing when I ignore the state-dependent of service rate. Indeed, by comparing the result from Table 2 and Table 3 with Table 4, the state-dependent model is more reliable with good staffing for services.

Table 4. Staffing requirement in SIPP Average method

Scenario	Method	
	State-dependent	State independent
Increasing	1773678	1793878
Decreasing	2343428	2298160
Total	4117106	4092038

2.4.4. SIPP Improvement

Results in the preceding section showed that the traditional SIPP technique has drawbacks. The issues are severe enough to invalidate SIPP's use in many service systems. The following logical query is, "Are there straightforward substitutes that perform better for systems for which SIPP is unreliable?" In this part, I investigate the validity of three different SIPP-based staffing requirement techniques.

2.4.4.1. SIPP Max

According to Green et al. (2001), the usage of the planning period average to represent the arrival rate over the whole planning period is the cause of many of SIPP's reliability issues. Using the maximum value of lambda for the planning period instead, as this frequently results in understaffing, is one potential solution. The SIPP Max technique is the name of this modification. I ran 3,888 scenarios described in sections 2.4.1 and 2.4.2 for increasing and decreasing approaches using the SIPP Max method. The result of increasing and decreasing approaches are explained as follows.

2.4.4.1.1. Increasing

The dashed blue lines in Figure 7 illustrate the reliability of the SIPP Max- increasing approach. In this approach, I used the state dependency model and used $M/M_{s(t)}/s(t)$ system. The figure shows that SIPP Max is always more reliable than standard SIPP, which is referred to as the SIPP average. While the SIPP average is unreliable for 621 of the 1,944 experiences run, SIPP Max is unreliable for 121 experiments. Moreover, the SIPP max is safe whenever $\alpha > 0.04$. Also, for a low arrival rate, $\lambda = 2$, the reliability of SIPP Max is high, and for higher λ, SIPP Max is not safe. Moreover, like the SIPP average, the performance of the SIPP is not significantly affected by changes in μmax because μmax never achieves maximum conditions. Also, the SIPP Max

reliability for low relative amplitude (RA≤ 0.5) is high. Since personnel levels are set over more extended planning periods, the SIPP is typically high accurate. As expected, SIPP performs better as the target QoS is more relaxed (i.e., higher τ).

On the other hand, from Figure 8, which gives the SIPP Max- increasing approach results, I see that none of any experiences works for the case with μmax and α. This is likely the result of Figure 7 describing each parameter's behavior patterns in the SIPP max-increasing approach. According to Figure 8, the unreliability of SIPP Max is high when RA=1, PP=0.25, τ= 0.05. On the other hand, while RA=0.1,0.5 and PP=0.5, 1, 2, the reliability rate of the SIPP Max-increasing approach is high.

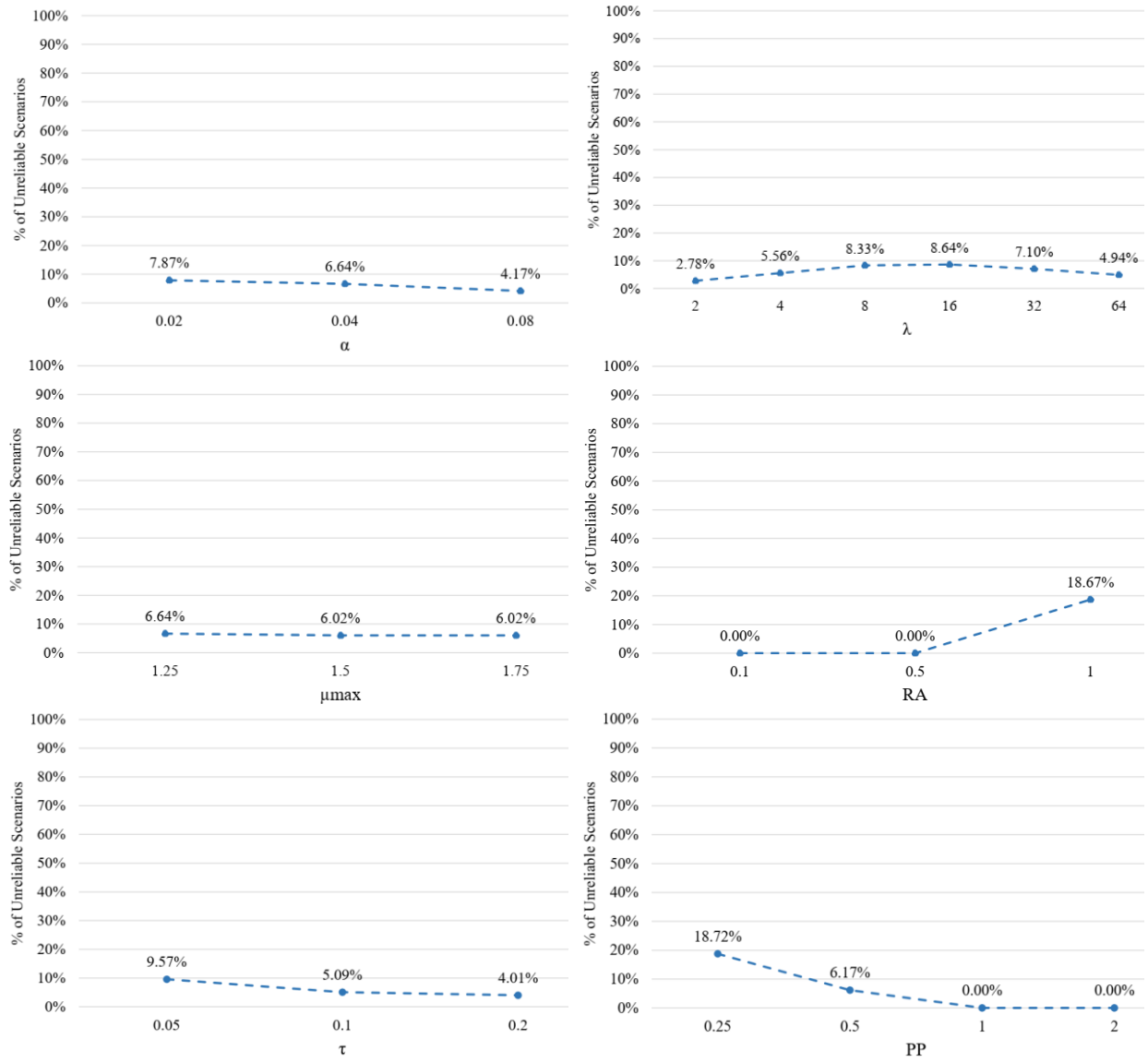

Figure 7. Effects of parameters on SIPP Max reliability- increasing approach

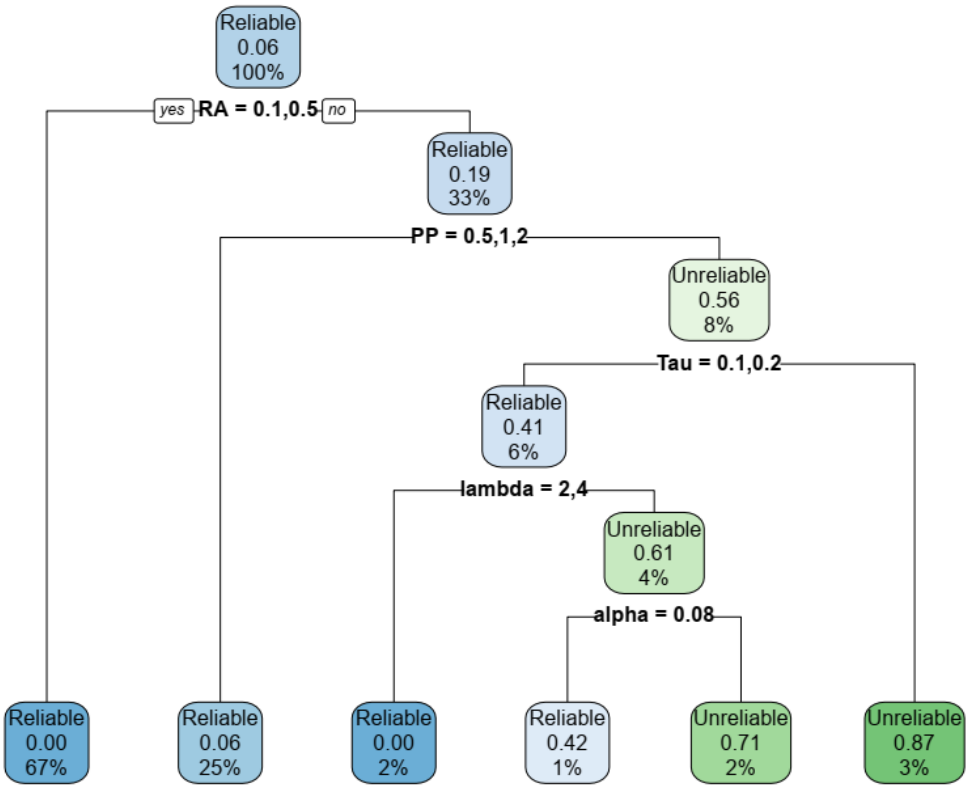

Figure 8. Reliability of SIPP Max- Increasing approach

2.4.4.1.2. Decreasing

Figure 9 summarizes the result of the SIPP Max-decreasing approach. The dashed blue lines show the influence of each parameter on the SIPP reliability when other parameters change. While the arrival rate increase, the SIPP Max reliability decrease. Also, whenever RA= 1, the reliability of SIPP Max decrease. Moreover, for higher μmax and α, the reliability of SIPP Max is improved, where for α = -0.02 and μmax = 1.75, the SIPP Max is 4.17% and 3.55 respectively. For longer planning periods, whenever PP ≥ 1, the reliability of SIPP Max is high. Moreover, the SIPP reliability is increased for the higher probability of delay.

According to Figure 10, which shows the reliability of the SIPP Max- decreasing approach, 102 unreliable experiences happened while RA=1, PP=0.25, and λ=16, 32, 64. On the other hand,

most reliability of the SIPP Max-decreasing approach occurred while RA=0.1, 0.5 (67% reliable), RA=0.1, 0.5, and PP= 0.5, 1, 2 (25% reliability).

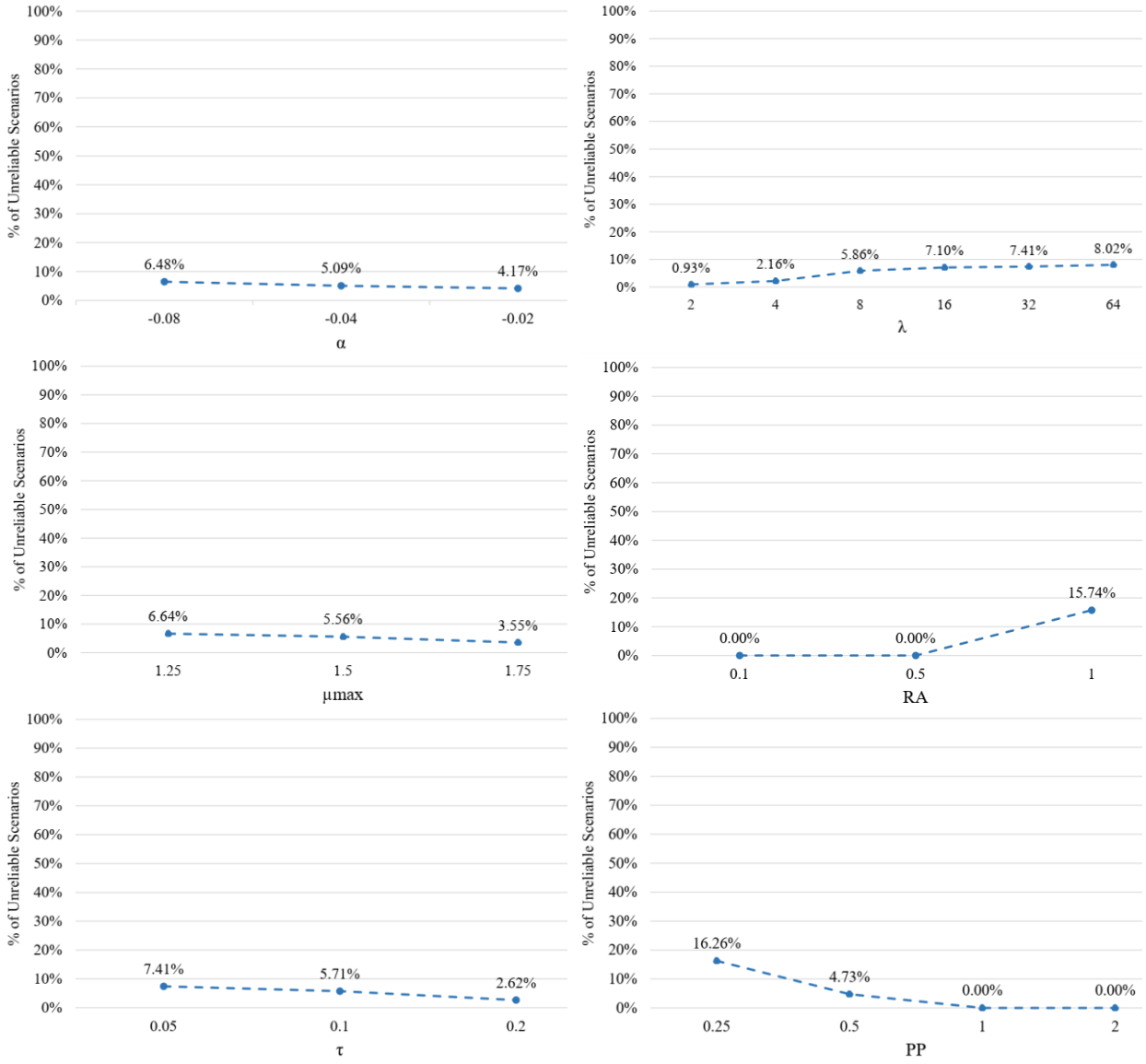

Figure 9. Effects of parameters on SIPP Max reliability- decreasing approach

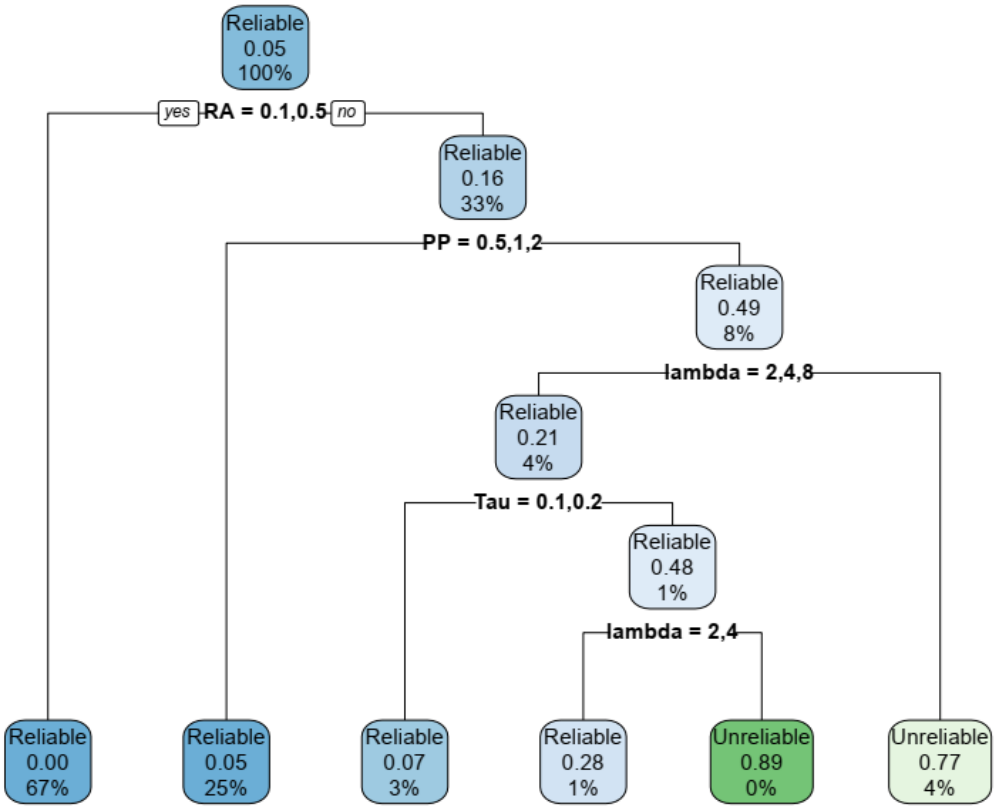

Figure 10. Reliability of SIPP Max-decreasing approach

2.4.4.2. SIPP Mix

When the arrival rate decreases, the SIPP average is more likely to be understaffed since there is a lag between the arrival rate and the ensuing delay. Therefore, employing the maximum arrival rate for the planning period should be the most cost-effective option for those planning periods during which the arrival rate strictly decreases or achieves its maximum value. Therefore, the SIPP Mix technique is the name of modification that employs the average planning period arrival rate of periods in which the arrival rate is strictly increasing and the maximum planning period arrival rate otherwise. I ran 3,888 scenarios for increasing and decreasing approaches using the SIPP Mix method. The result of increasing and decreasing approaches are explained as follows.

2.4.4.2.1. Increasing Scenario

The solid green lines in Figure 11 illustrate the reliability of the SIPP Mix- increasing approach. The figure shows that SIPP Mix is always as or more reliable than standard SIPP, which is referred SIPP average. While the SIPP average is unreliable for 621 of the 1,944 experiences run, SIPP Mix is unreliable for 213 experiments. The experiments show that the reliability of SIPP Mix is high for a low arrival rate, $\lambda = 2$. Moreover, unlike the SIPP average, the reliability of the SIPP Mix for $\alpha > 0.04$ is better. Also, the SIPP Mix reliability for the lower RA, $RA \leq 0.5$, is high. Since personnel levels are set over more extended planning periods, the SIPP is typically high accurate, specifically for PP=1, 2. As expected, SIPP performs better as the target QoS is more relaxed (i.e., higher τ). According to Figure 12, which illustrates the reliability of the SIPP Mix-increasing approach, the dark green box is related to unreliability experiences among 1944 experiences, and the blue box shows the reliability ones. From Figure 12, the most unreliable experiences happened while RA=1 and PP=0.25, 0.5. Moreover, when RA=1, PP=0.25, 0.5, and $\lambda \neq 2, 4$, the reliability is decreased.

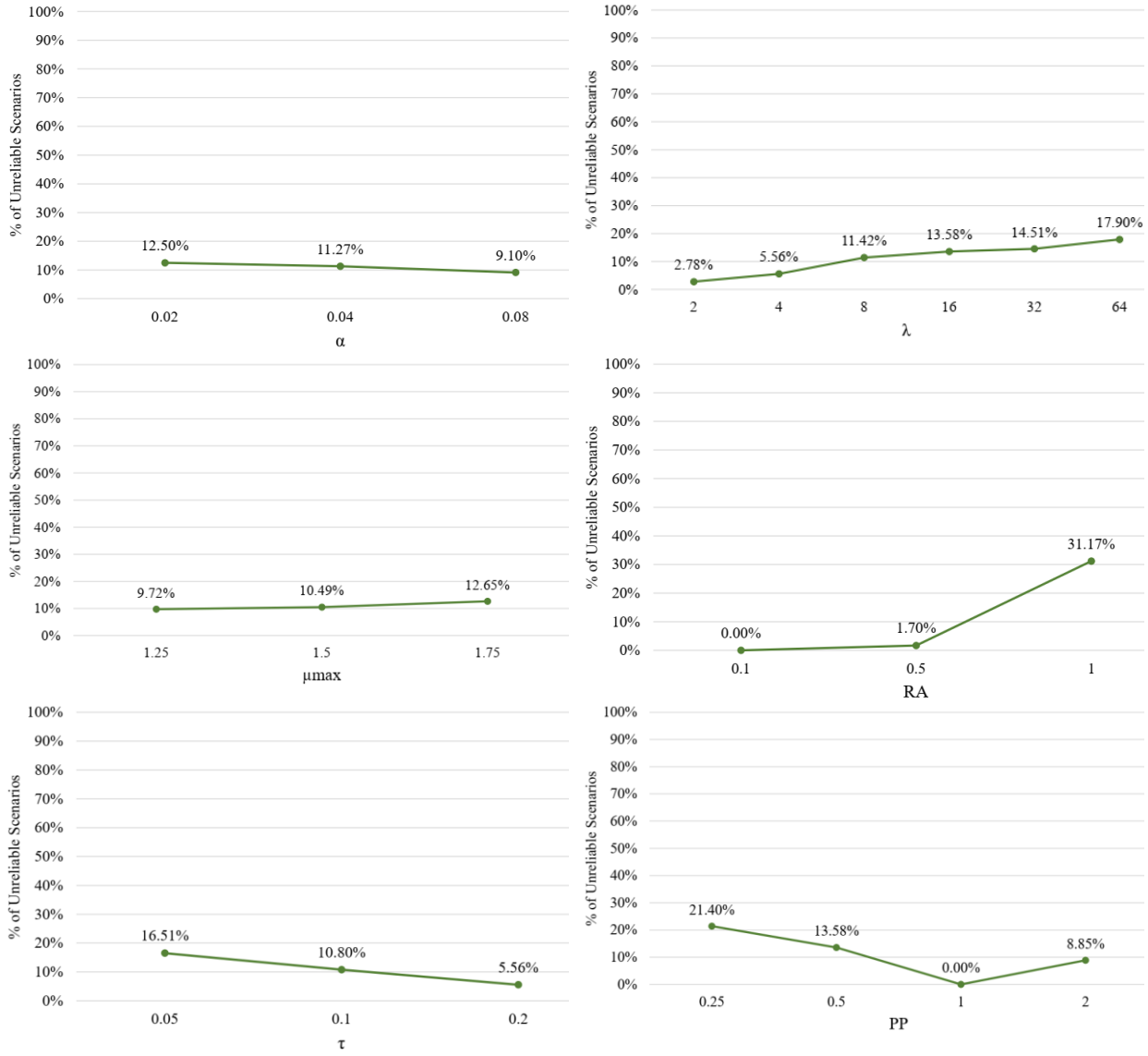

Figure 11. Effects of parameters on SIPP Mix reliability- Increasing approach

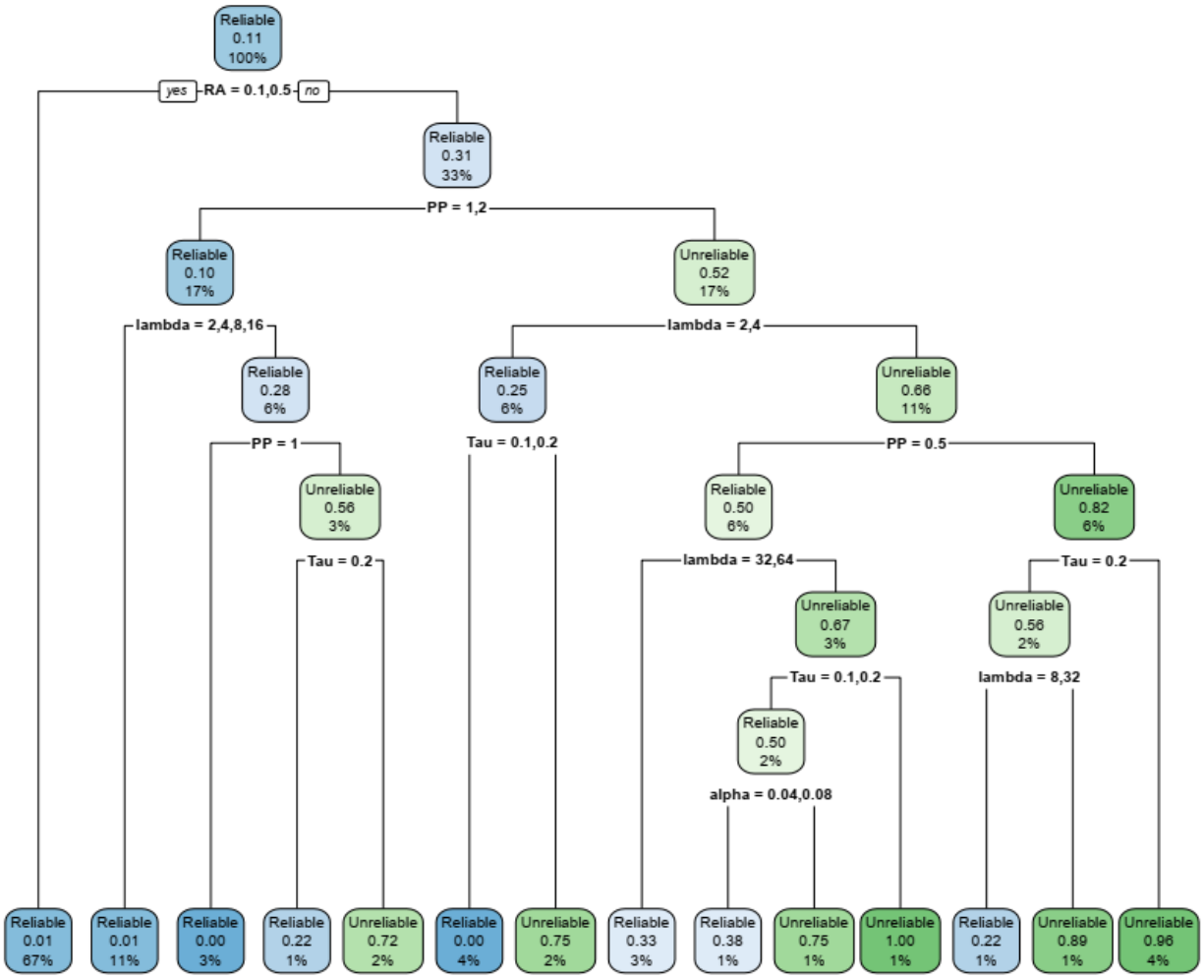

Figure 12. Reliability experiences SIPP MIX- Increasing approach

2.4.4.2.2. Decreasing Scenario

Figure 13 illustrates the reliability of the SIPP Mix for decreasing approach. The figure shows that SIPP Mix is always as or more reliable than standard SIPP, which is referred SIPP average. While the SIPP average is unreliable for 615 of the 1,944 experiences run, SIPP Mix is unreliable for 145 experiments. The experiments show that the reliability of SIPP Mix is high for a lower arrival rate. Also, the SIPP Mix reliability for RA$\leq$ 0.5 is high. Since personnel levels are set over more extended planning periods, the SIPP is typically high accurate.

Moreover, like the SIPP average, the SIPP Mix reliability is higher for higher target planning delays. Figure 13 shows the reliability of the SIPP Mix-decreasing approach, the dark

green box indicates the unreliability experiences percentage, and the blue one offers reliable experiences. As we can see, the most unreliable experiences, nearly 8% among 1,944 decreasing experiences, happened when RA = 1 and PP ≠ 0.25. Figure 14 describes the behavior pattern of parameters, especially for α and μmax. As we can see, the unreliability is high for lower μmax, where μmax ≤ 1.5.

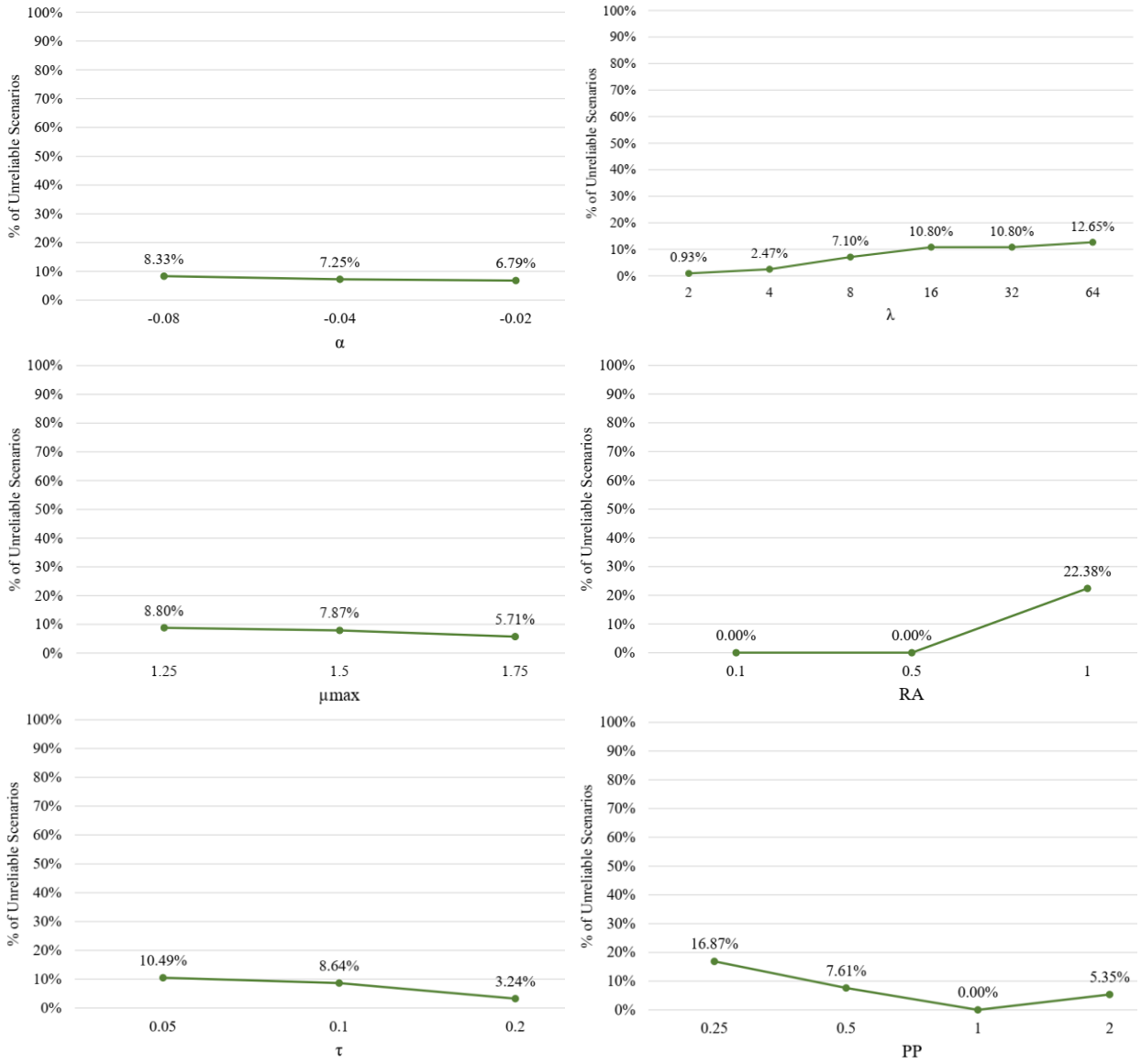

Figure 13. Effects of parameters on SIPP Mix reliability- decreasing approach

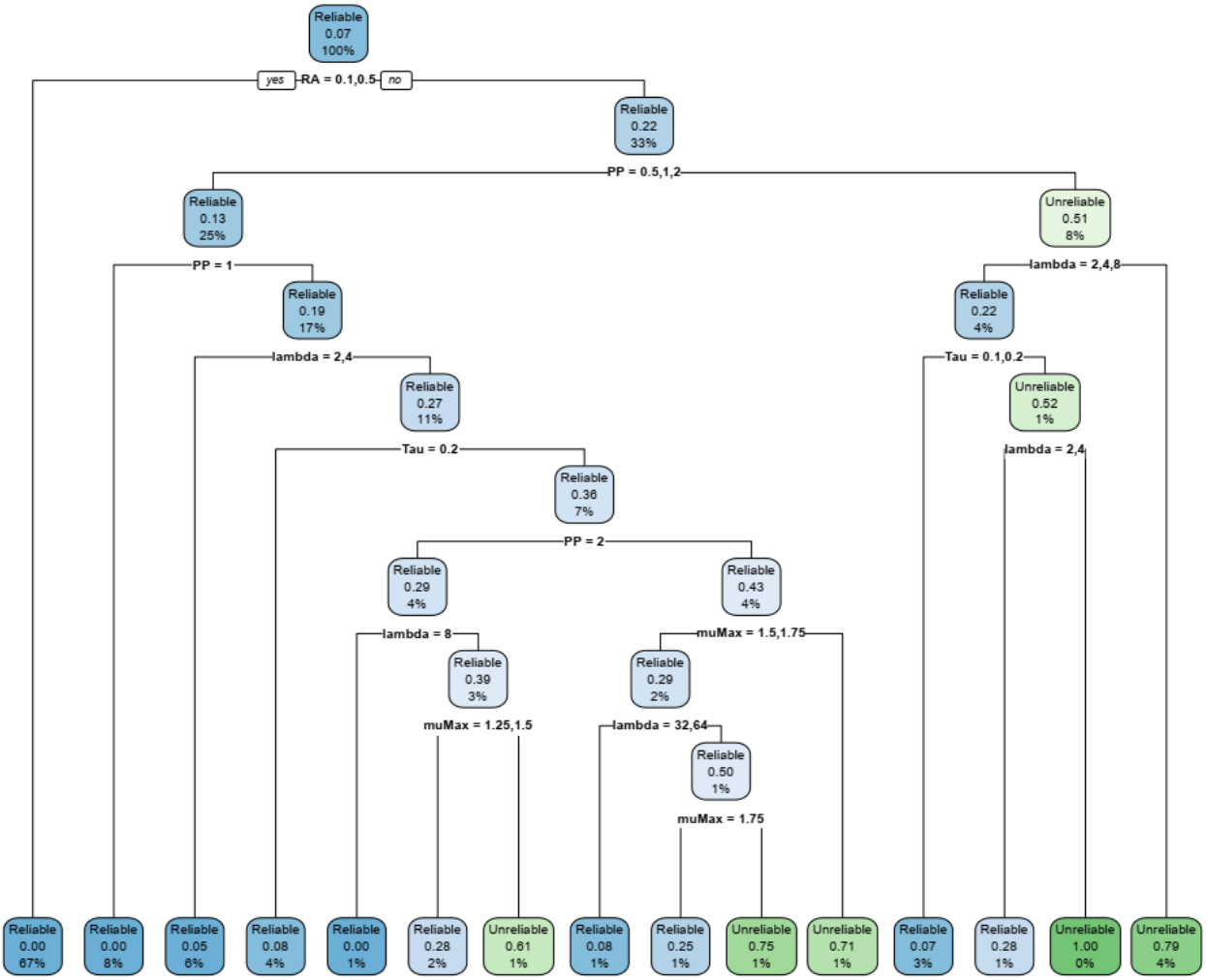

Figure 14. Reliability experiences SIPP Mix- Decreasing approach

2.4.5. Summary

Table 5 shows the state-dependent service rate model with the sinusoidal arrival rate model correctly identifying that SIPP Max is more reliable than SIPP Mix. Although the amount of staffing in SIPP Max is higher than in SIPP Mix, the reliability is high, and the SIPP Max method didn't meet the target. Indeed, SIPP Max has high reliability without low delay with high staffing during the service time to deliver the services in the system.

Table 5. Staffing requirement SIPP methods

Measure	Scenarios	SIPP average	SIPP Max	SIPP Mix
Reliability	Increase	1,329	1,823	1,731
	Decrease	1,323	1,842	1,799
Staffing	Increase	1,773,678	1,808,836	1,791,257
	Decrease	2,343,428	2,393,128	2,368,278

2.5. Conclusion

This chapter proposed the staff requirements in a service system with random cyclic demand and state-dependency service rate. Also, I used the Markovian model and SIPP approach to determine staff requirements in the EMS. By comparing the number of delays in the service system, SIPP Max can be more reliable than SIPP Mix and SIP average. In all parameter values, SIPP Max has a good performance and a low probability of delay.

The finding described in this chapter is based on sinusoidal arrival rate and state-dependent service rate in the service system. Moreover, the chapter is based on the probability of delayed performance targets for some reasons. First, we are familiar with most real-world implementations, and the research on these issues used the likelihood of delay measures. Because, given the method of numerically solving the differential equations of the system, computation of probability delay was theoretically possible. Finally, adopting a performance metric with a single parameter makes the analysis more straightforward and clearer conclusions.

This chapter addressed the staff requirement and capacity planning using the SIPP technique. Indeed, by using the sinusoidal arrival rate model and state dependency of service rate, I developed the SIPP approach to have highly reliable performance with a low level of delay in the service system. This approach is more broadly applicable in setting outside of service systems that provide interdisciplinary services, for instance, in emergency departments. This approach can

inform the successful implementation of capacity planning in the healthcare system and improve access to community-based healthcare.

For future studies, researchers can apply some scenarios such as SIPP Max and SIPP Mix for service rate. Indeed, in the SIPP Max approach, the system selects the maximum level of service rate. Also, in the SIPP Mix approach, whenever the service rate is increasing level, the system uses the average level of service rate. Otherwise, the system uses the increased level of service rate.

3. TIME-DEPENDENT MAXIMAL COVERING LOCATION PROBLEM CONSIDERING REPOSITIONING AND AMBULANCE RANKING

3.1. Introduction

The EMS systems aim to provide prompt medical attention to preserve lives and lower morbidity. The ability to guarantee a rapid response time is crucial since EMS systems sometimes assist patients who need emergency requests. To meet the patient demand, the arrangement of all facilities in EMS systems, including the location of emergency facilities, ambulance stations, emergency vehicles, crew schedules, and ambulance routing, needs to be planned.

An essential long-term planning issue in developing an EMS system that enables emergency response to calls is the placement of ambulance stations and the allocation of ambulances to these stations. The spatial and temporal variability of input parameters, such as demand and travel time, must be considered when determining the locations of ambulances. The need for ambulance requests varies by region and day (Cantwell, et al., 2013). The time it takes to go from the ambulance's position to the patient's location also depends on the time of day (Schmid & F.Doerner, 2010). In urban areas, the variance in travel times throughout different day hours can substantially impact the actual service level and cause delays while transferring the patient to the hospital. The urban areas of many developing nations have dense populations and congested roadways, which makes this difference more challenging and sometimes does not let ambulances pass (Boutilier & Chan, 2020).

This chapter of my dissertation has utilized the percentage of emergency calls that should be answered within a predetermined time frame to assess the quality of an EMS system. When there is a strong probability that an ambulance will be available at the location closest to a call, a high level of service quality is attained. Therefore, a predetermined number of ambulances must

be distributed throughout the sites to ensure the highest possible level of coverage. Also, in mathematical modeling, maximizing the coverage and minimizing the response time are among the popular objective measures to select the best locations for ambulance stations. Erkut et al. (2008) demonstrated how these measurements are ineffective in differentiating between the effects of two different response times. These policies are predicated on the idea that territory is covered if it can be reached within a predetermined window. For instance, if the required coverage time is 12 minutes, a patient's place will be regarded as covered, 5 or 11.8 minutes away from an ambulance station. On the other hand, whether a patient's place is 12.2 or 20 minutes from the ambulance station will be regarded as uncovered. This is a significant problem because the difference between 11.8 and 12.2 minutes may not be a problem, but the difference between 5 and 11.8 minutes will be more problematic as the patients may lose their lives. In this manner, Erkut et al. (2008) presented a non-linear survival function that monotonically declines with response time and considered patients' chances of survival to solve the problem. Also, Raviarun et al. (2021) considered a continuous survival function-based objective by including station-level service rate, arrival rate, and the busy probability of ambulances to propose a mixed-integer non-linear program.

This chapter also investigates finding the location of ambulances and their relocations between those ambulance stations. The location of ambulances has been selected among the demand zone as candidate locations should be placed. On the other hand, the concept of relocation has been utilized, which means the ambulance is placed at various base locations in consecutive time periods. Therefore, the ambulance should be relocated between two time periods. Also, to reduce the number of opened locations and their relocations, penalty cost has been given to them

to control the total costs of the network (Berg & Aardal, 2015). Indeed, considering penalties in the model allows the decision-makers to manage their budget better to avoid unnecessary costs.

This chapter of my dissertation aims to consider the time dependency of ambulance location stations across the network and allocate the ambulance to the patients to cover more 911 calls. The time dependency in location models is related to variations in travel time, availability of ambulances, and demand. Also, the model of this chapter is realistic since few studies dealt with the maximization of service coverage along with time dependency on travel time, availability of ambulances, and demand. The overall contribution of this chapter is as follows. First, I propose a Mixed-Integer Programming (MIP) model by considering the time dependency in the location model. Second, to maximize network coverage, I penalized the objective function to fully reflect the effect of opening the location and relocating the ambulance during the allocation to avoid high costs imposed on the network. Third, the proposed model is tested on a commercial CPLEX solver to find an optimal solution. Finally, the sensitivity analysis on crucial parameters has been done to show the applicability of the model.

3.2. Literature Review

Due to the importance of EMS planning, the location-allocation problem of ambulances has attracted much attention in the literature studies. In addition to context-free location models, (ReVelle, et al., 1977) provided an early assessment of ambulance location models. A summary of the evolution of operations research challenges for the planning of EMS and fire departments is provided (Goldberg, 2004). A review of ambulance localization problems using static, probabilistic, and dynamic models is presented by Brotcorne et al. (2003). A study of EMS planning issues, including issues with ambulance location, is presented by Aringhieri et al. (2017) throughout the entire emergency treatment route. In a review of relevant research that focuses on

the interdependencies across various challenges, Reuter-Oppermann et al. (2017) provided an overview of the logistical issues in EMS. A recent analysis of EMS fleet management challenges is presented by Bélanger et al. (2019), with a particular emphasis on the location, relocation, and dispatch-related issues and the interactions between these issues. The comprehensive literature on ambulance location issues is more thoroughly surveyed in these review studies. Next, I provide an overview of a few papers that are pertinent to this chapter in this research.

The Location Set Covering Model (LSCM) and Maximum Covering Location Problem (MCLP) are two early location models for placing ambulances that have been documented in the literature. To reduce the number of ambulances needed to cover all demand zones, Toregas et al. (1971) developed the LSCM. The goal of the MCLP, developed by Church et al. (1974), was to provide the greatest amount of coverage with a predetermined number of ambulances. In a seminal work employing queueing theory, Larson (1974) introduced the hypercube queuing model, which considers server congestion and coverage. The maximal expected covering location problem (MEXCLP), created by Daskin et al. (1983), maximizes the expected coverage while explicitly accounting for the busy likelihood of ambulances. The Maximum Availability Location Problem (MALP), developed by ReVelle et al. (1989), extended the MCLP to take server availability probabilities into account. To maximize the amount of demand that two ambulances can cover at one location, Gendreau et al. (1997) suggested the Double Standard Model (DSM).

Emergency call demand follows a circadian rhythm, which implies that it changes continually throughout 24 hours. Low demand occurs at night, then rises to a peak at noon and another in the evening. This pattern describes how demand fluctuates (Cantwell, et al., 2015; Cantwell, et al., 2013;Momenitabar, et al., 2022; Momenitabar, et al., 2022). To consider temporal fluctuation in demand, Repede et al. (1994) proposed the maximal expected coverage location

model with time variation (TIMEXCLP), an extension to the MEXCLP. A dynamic available coverage location (DACL) model was proposed by Rajagopalan et al. (2008) that considers several times with fluctuating demands and minimizes the number of ambulances while still maintaining a predetermined level of ambulance availability. Saydam et al. (2013) proposed an extension of DACL that considers two goals: reducing the number of ambulances and reducing the frequency of ambulance shift relocations.

The time it takes ambulances to respond to patients is another significant input that changes throughout the day. The DSM is expanded by Schmid et al. (2010) to account for several periods with variable travel times throughout the day. An ambulance location model that considers response time uncertainty by considering a short random dispatch delay and the random journey time is presented by Ingolfsson et al. (2008). To account for temporal fluctuation in demand, travel duration, and ambulance availability, Berg et al. (2015) extended the MEXCLP. To reduce the number of stations and relocations while increasing predicted coverage, they penalize the target function by including the cost of opening new stations and moving ambulances. In a lower-middle-income country setting of Dhaka, Bangladesh, Boutilier et al. (2020) explored the issue of identifying and routing ambulances while emphasizing the variability in geographical demand and uncertainty in trip times. To deal with data uncertainty, they combine field data and prediction models and employ a simulation-based methodology to address problems with the policy.

Most models used to locate ambulances in the literature either employ coverage or response time restriction-based objective functions. Through their analysis of EMS performance metrics, McLay et al. (2010) concluded that varied response time criteria could be effectively employed as a proxy for survival probability and fairness. By extending the MSLP to incorporate heterogeneous patients with additional outcome metrics, such as survival function and coverage based on patient

category, Knight et al. (2012) developed the maximal expected survival location model for heterogeneous patients (MESLMHP). A more realistic ambulance location model is presented by Leknes et al. (2017) formulation of the maximum expected performance location problem for heterogeneous regions (MEPLP-HR), which computes the busy probability for each station separately while considering a heterogeneous performance measure for various patient categories. Andersson et al. (2020) proposed an adaptation to the MEPLP-HR that considers different periods and evaluates many scenarios, including closing an emergency department, designating cars for non-urgent calls, and time-dependent fluctuation in demand.

Based on what has been discussed so far, a few studies dealt with finding simultaneously optimal locations of ambulance stations along with relocating ambulances when returning from serving demand zone. Also, no studies considered different parameters, including the availability of ambulances, demand, and travel time of ambulance arriving at the scene. Repede et al. (1994) proposed the TIMEXCLP, which introduced several time frames and increased the predicted coverage throughout the entire day. Also, Schmid et al. (2010) introduced the time dependency model, which is the extension of the Double Standard model that assumes that the travel time is time-dependent. Moreover, a penalty has been added for the number of relocations to deal with time-dependency in travel time. Therefore, this chapter aims to maximize the coverage along the network by minimizing the number of ambulance stations and penalizing the opened locations and their relocations to fully address EMS planning. Also, the model with penalty costs for relocation and locations between different time periods has been proposed. Moreover, the model of this chapter is realistic since few studies dealt with the maximization of service coverage along with time dependency on travel time, availability of ambulances, and demand, simultaneously. Furthermore, the ambulance stations have been ranked to prioritize the stations based on their

received calls. Indeed, the proposed model has applicability to be implemented in reality due to including the time independency and coverage maximization that are concerns of decision-makers.

3.3. Problem Formulation

In this section, firstly, the problem statement of this chapter has been proposed. Second, the mathematical formulation of the model has been discussed in detail.

3.3.1. Problem Statement

Assume that a geographical area has been divided into different zones to cover the patient's needs promptly. Indeed, the demand of each zone is defined by the number of calls received from the patients. Accordingly, these calls need to be responded to as soon as possible to avoid the loss of the patient. For that purpose, some locations need to be established to place the ambulance. Figure 15 shows the schematic view of the covering location model that has been discussed in this chapter.

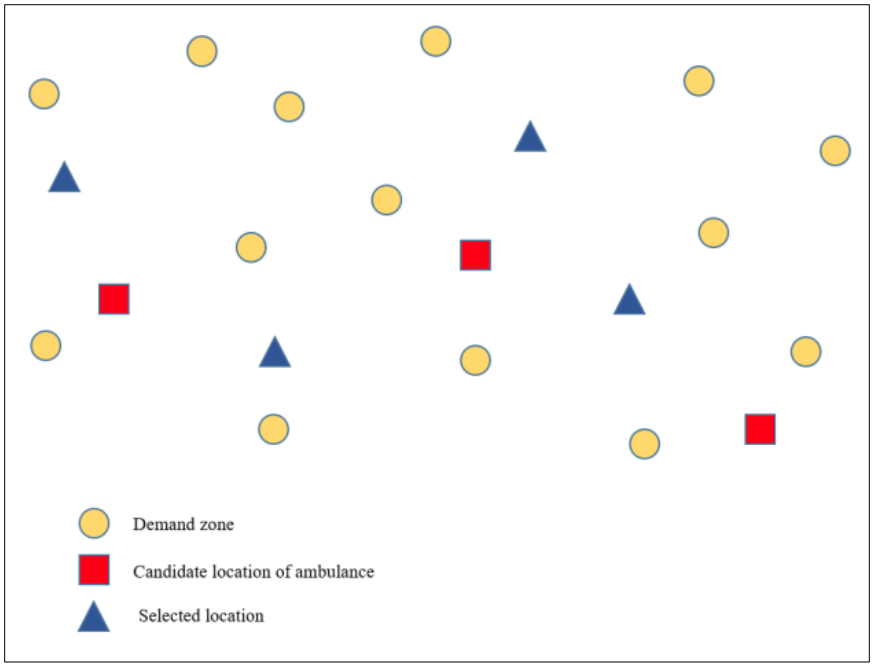

Figure 15. Network of covering location problem

Also, each station is ranked 1 to r in providing patient service. For instance, rank 1 means that the ambulance station can be dispatched to the scene after receiving the call to respond to the patient. Rank 2 means the ambulance can be dispatched to the scene if the first ambulance station is not available. Also, the model of this chapter is a multi-period model, meaning that the time is divided into various times to comprehensively cover all 24 hours in a day. The model has the penalty cost for relocation and locations. This penalty works for the number of locations and relocation between time periods to decrease the cost in the time-dependent travel times model. Also, the model has the busy fraction, which is the percentage of time a resource—such as an ambulance, a location, or a system—is in use and thus unavailable to respond to incoming emergency calls.

The problem here is to determine an optimal location of ambulance stations across all demand zones and allocate them to the demand zone to maximize the service coverage to the patients. Due to the high costs of establishing an ambulance location, the penalty has been given to the opened locations and the relocations of an ambulance in different periods to avoid the high cost imposed on the network.

3.3.2. Mathematical Formulation

Before proposing the equations of the model, the notation of the mathematical model has been defined which are as follows in Table 6 :

Table 6. Notation of the mathematical model

Notation	**Definition**
Set	
i	Index of the demand zone $i = 1, 2, ..., I$
j	Index of candidate locations of ambulances $j = 1, 2, ..., J$
k	Index of ambulances $k = 1, 2, ..., K$
t	Index of time, $t = 1, 2, ..., T$
W_i^t	Index of base locations that can cover demand point i in period t
r, q	Index of ambulance station's rank $r = 1, 2, ..., R$, $q = 1, 2, ..., Q$
Parameters:	
q_t	Busy fraction in period t
β	A penalty for each ambulance location
γ	A penalty for each ambulance relocation
P_t	Number of available ambulances during period t
M	Large number
L_t	Maximum load which is assigned to each ambulance in each period
λ_{it}	Number of calls per unit times received from demand zone i in period t
S_{ijt}	Service time of received call from demand zone i responded by ambulance station j
Variables:	
X_{jt}	The number of ambulances located at station j during period t
Y_{ikt}	If demand point i is covered by at least k ambulances during period t
Z_j	If the based location j is used at a one-time period
R_{ijt}	If the ambulance is relocated from location i to location j between period t and t+1
dd_{ijrt}	The proportion of demand from zone i covered by station j with rank r during period
$Delta_{jrit}$	If station j is assigned rank r for demand zone i during period t

The proposed model of this chapter can be formulated as follows:

$$MaxZ_1 = \sum_i \sum_k \sum_t \lambda_{it} \times (1-q_t) \times q_t^{k-1} \times Y_{ikt} + \sum_i \sum_j \sum_r \sum_t \lambda_{it} \times dd_{ijrt} \quad (3.1)$$

$$MinZ_2 = \beta \times \sum_j Z_j + \gamma \times \sum_i \sum_j \sum_t R_{ijt} \quad (3.2)$$

Subject to:

$$\sum_{j \in W_i^t} X_{jt} \geq \sum_k^{p_t} Y_{ikt}, \quad \forall i = 1,2,...,I, \forall t = 1,2,...,T \quad (3.3)$$

$$\sum_{j \in W} X_{jt} \leq P_t, \quad \forall t = 1,2,...,T \quad (3.4)$$

$$\sum_{t \in T} X_{jt} \leq M \times Z_j, \quad \forall j = 1,2...,J \quad (3.5)$$

$$X_{jt} + \sum_i R_{ijt} + \sum_i R_{jit} = X_{j(t+1)}, \quad \forall j = 1,2...,J, \forall t = 1,2...,T-1 \quad (3.6)$$

$$X_{jT} + \sum_{i \in W} R_{ijT} + \sum_{i \in W} R_{jiT} = X_{j1}, \quad \forall j = 1,2...,J \quad (3.7)$$

$$\sum_j Delta_{jrit} = 1, \quad \forall r = 1,2,...,R, \forall i = 1,2,...,I, \forall t = 1,2,...,T \quad (3.8)$$

$$Delta_{jrit} \geq dd_{ijrt}, \quad \forall i = 1,2,...,I, \forall j = 1,2,...,J, \forall r = 1,2,...,R, \forall t = 1,2,...,T \quad (3.9)$$

$$Delta_{jrit} \geq dd_{ijqt}, \quad \forall i = 1,2,...,I, \forall J = 1,2,...,J, q \neq r \in R, \forall t = 1,2,...,T \quad (3.10)$$

$$\sum_j dd_{ijrt} \leq \sum_j dd_{ij(r-1)t}, \quad \forall i = 1,2,...,I, \forall j = 1,2,...,J, r \in R \mid r > 1, t = 1,2,...,T \quad (3.11)$$

$$\sum_{j}\sum_{r} dd_{ijrt} = 1, \quad \forall i = 1,2,...,I, \forall t = 1,2,...,T \tag{3.12}$$

$$\sum_{i}\sum_{r} \lambda_{it} \times S_{ijt} \times dd_{ijrt} \le L_t \times X_{jt}, \quad \forall j = 1,2,...,J, \forall t = 1,2,...,T \tag{3.13}$$

$$\sum_{k}^{P_t} Y_{ikt} \le \sum_{j}\sum_{r} Delta_{jrit}, \quad \forall i = 1,2,...,I, \forall t = 1,2,...,T \tag{3.14}$$

$$Z_j \le \sum_{i}\sum_{t} Delta_{jrit}, \quad \forall j = 1,2,...,J, \forall r = 1,2,...,R \tag{3.15}$$

$$\begin{aligned} &Y_{ikt}, Z_j, R_{ijt}, Delta_{jrit} \in \{0,1\}, \forall i, j, k, r, t \\ &0 \le dd_{ijrt} \le 1, \forall i \in I, \forall i, j, r, t \\ &X_{jt} \in Z, \forall j, t \end{aligned} \tag{3.16}$$

The first objective function (Eq. (3.1)) aims to maximize the coverage of all demands across the network. The second objective function plans to penalize the number of locations that have been established. Also, the second term of Eq. (3.2) aimed at penalizing the total number of ambulance relocations between different periods. Eq. (3.3) ensures that all demand should be covered by at least k ambulances. Also, Eq. (3.4) confirms that the available number of ambulances in each period is limited. Eq. (3.5) indicates that the ambulance can be assigned to each location only if the ambulance stations are opened. Also, the balance between the number of ambulances in different periods has been ensured by Eq. (3.6) and Eq. (3.7).

Eq. (3.8) through Eq. (3.12) have been written to show the coverage constraints for the model of this chapter. For instance, Eq. (3.8) shows that only one ambulance station can obtain one rank based on the model. While Eq. (3.9) assures that if the station is opened, the demand can be covered by that station. Eq. (3.10) tells us that one rank can be assigned to one station and

cannot be given to multiple stations. The more significant portion of demand has been given to a station with a higher rank, which is shown by Eq. (3.11). Eq. (3.12) guarantees that each demand zone in each period can only be covered by one station. The maximum service time availability at each station is shown by Eq. (3.13). Eq. (3.14) states that if one rank has been assigned to one station, then the demand can be covered by at least k ambulances. Eq. (3.15) represents the relationship between the location of the ambulance station and the rank of each ambulance station for covering the demand zone in each period. Lastly, the nature of all defined variables has been shown in Eq. (3.16).

3.3.3. Solution Methodology

To solve the propsoed multi-objetcive optimization model in section 3.3.1, the Epsilon constraint method has been utilized to obtain the result. This method continually optimizes one of the objective functions if the decision maker defines the highest acceptable constraint for other objectives in the form of constraints. The structure of this method has been proposed as follows:

- Select one of the objective functions as desired.
- Solve the problem each time according to one of the objective functions and extract the optimal values.
- Get a table of values for $\varepsilon_2,...,\varepsilon_n$.
- Each time solve the problem with the primary objective function with any of the values $\varepsilon_2,...,\varepsilon_n$.
- Report the Pareto frontier.
- By changing the values to the right-hand side of the constraints (ε_i), efficient solutions to the problem are obtained.

Also, Eq. (3.17) shows how this technique can be formulated and adjusted to the proposed model of this chapter.

$$
\begin{aligned}
&M \text{ in } \quad \mathrm{f}_1(x), \quad x \in X \\
&\mathrm{f}_1(x) \le \varepsilon_2 \\
&. \\
&. \\
&. \\
&\mathrm{f}_n(x) \le \varepsilon_n
\end{aligned}
\tag{3.17}
$$

3.4. Computational Results

The proposed model of this chapter is a Mixed Integer Programming (MIP) model. Therefore, it can be solved using commercial solvers like CPLEX, Gurobi, and BARAN. This chapter utilizes the CPLEX solver to obtain the result. Also, all computations were executed on a 2.9 GHz Intel® Core i7-3520M PC with 16 GB of RAM.

3.4.1. Input Parameters

The input data for sunning the model has been provided in Table 7. It is noteworthy to mention that the input data of the model has been obtained from Berg et al. (2015).

Table 7. Input data of the model

Parameters	Definition	Value
q_t	The busy fraction in period t	Uniform (0.15, 0.45)
β	A penalty for each ambulance location	360.00
γ	A penalty for each ambulance relocation	3.60
P_t	Number of available ambulances during period t	Uniform (5, 10)
L_t	Maximum load is assigned to each ambulance in each period	Uniform (190, 200)
d_{it}	Demand for point i in period t	Uniform (2000, 2500)
λ_{it}	Number of calls per unit times received from demand zone i in period t	Uniform (9, 10)
S_{ijt}	Service time of received call from demand zone i responded by ambulance station j in period t	Uniform (0.20, 0.25)

To run the model, I have generated 15 demand points and six candidate locations for placing ambulances in 6 different periods. As can be seen in Table 7, the busy fraction is uniformly generated between 0.15 and 0.45. The penalty for each ambulance location and relocation is considered to be 0.5 and 0.005 percent of the total number of demands in all periods. I divide the time between 6 seasons, which is shown in Table 8 (see reference (Berg & Aardal, 2015)). The deviation from the average demand during a period is referred to as call intensity. As a result, the average number of calls per hour during period 6 is only 88% of the daily average. Travel time displays the deviation from the average travel time over a given period. For example, traveling from point A to point B takes 15% less time in period 1. Also, the demand, number of ambulances, and call duration are used to calculate the busy fractions. The call duration is the average amount of time an ambulance is unavailable after responding to a call. This time does not include the time it takes to return to the base because ambulances are dispatched during that time.

Table 8. Input data for different periods

Interval	Call intensity	Travel time	Number of ambulances	Busy fraction
00:00-04:00	0.72	0.85	5	0.212
04:00-08:00	0.58	1.22	7	0.283
08:00-12:00	1.13	1.35	10	0.429
12:00-16:00	1.84	0.97	10	0.455
16:00-20:00	1.31	1.24	10	0.363
20:00-24:00	0.88	0.93	7	0.257

3.4.2. Basic Results

After running the model of this chapter, the result is shown in Table 9. Based on Table 9, I can see that different coverage levels in each time period have been obtained. Also, the total number of calls is between 89% and 96% in each interval. Also, the third and fourth columns show the number of locations and relocations of ambulance sites. The computational time of running this model is about 25 seconds. Also, I have set the limitation time for the CPLEX solver to 300 seconds which could obtain a result less than this threshold time.

Table 9. Result of the model

Interval	Percentage of coverage	Number of locations	Number of relocations
00:00-04:00	0.962	5	1
04:00-08:00	0.953	6	0
08:00-12:00	0.898	4	0
12:00-16:00	0.935	6	2
16:00-20:00	0.917	5	1
20:00-24:00	0.942	6	0
00:00-24:00	0.934	6	4

As can be seen, the coverage is different in different periods. The final result shows that a total number of 6 locations and four relocations is needed to reach the coverage of 93.4%.

3.4.3. Sensitivity Analysis

3.4.3.1. Response Time

The result of the basic model is obtained according to the maximum travel time of 12 minutes, which is the response time of 15 minutes minus a pre-trip delay of 3 minutes. Indeed, the pre-trip delay is the time between the arrival call and when the ambulance leaves the center. This sub-section aims to know changes in response time and how much they affect the coverage of the model. In this way, the result of variations in response time is shown in

Figure 16. As is visible, the coverage value increased by increasing the response time. Also, it can be noted that increasing the response time after 16 minutes has not resulted in changes significant in the service coverage.

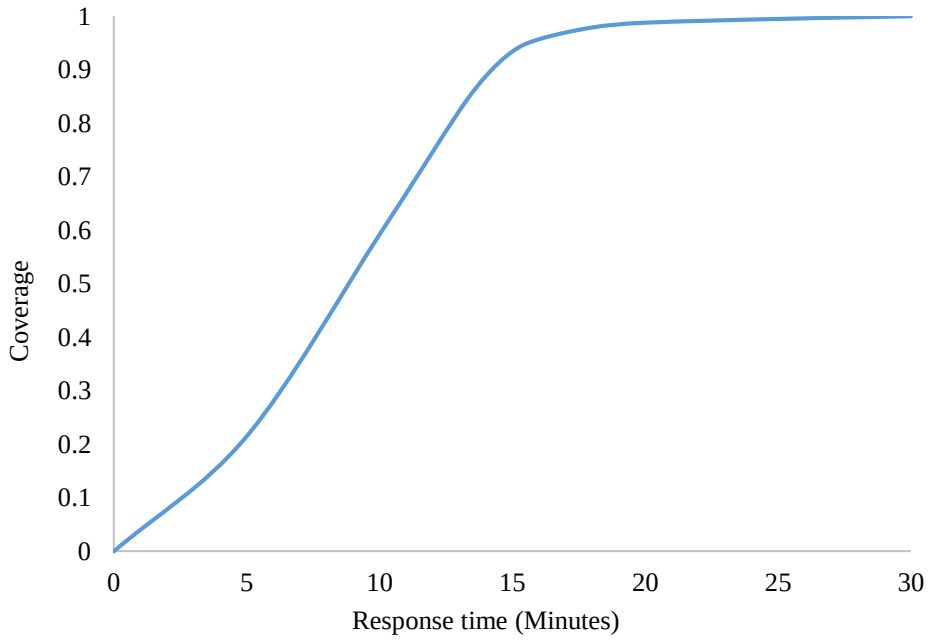

Figure 16. Coverage of ambulances based on various response times

3.4.3.2. Penalty Coefficient

Another important parameter that I have included in this chapter is the effect of changing the two penalties for opened locations and their relocations between different periods. The effect

of these two parameters on objective function has been shown in Table 10. As can be seen, ignoring the penalties in the objective function resulted in improving the expected coverage considerably but led to an increase significantly in the total number of locations and relocations in the model with 11 and 18, respectively. Indeed, considering penalties minimized the total costs of establishing the ambulance locations. Interestingly, the trade-off between the costs of opening ambulance locations and service coverage can help the decision-makers behave properly.

Table 10. Effects of penalties

	Objective function	
	With Penalties	Without penalties
Expected coverage	0.934	0.978
Number of locations	6	11
Number of relocations	4	18

3.4.3.3. Availability of Ambulances

The other parameter is the availability of ambulances in different periods of time. In Table 11, this parameter has been generated using uniform distribution between 5 and 10, which is considered a basic scenario. But, some other scenarios have been included, which show the availability of ambulance lower and upper than of the basic scenarios, which are shown in Table 11.

Table 11. Availability of an ambulance in different scenarios

Time interval	Scenario				
	1	2	3 (basic)	4	5
00:00-04:00	1	3	5	7	9
04:00-08:00	3	5	7	9	11
08:00-12:00	6	8	10	12	14
12:00-16:00	6	8	10	12	14
16:00-20:00	6	8	10	12	14
20:00-24:00	3	5	7	9	11

The result of different scenarios on various time intervals for the first objective function has been reported in Table 12. As can be seen, increasing the availability of an ambulance resulted in improving the coverage of ambulances in the network. This is expected since increasing the number of ambulances helps the network to cover more demand, which leads to improving the first objective function. Also, it can be understood that depending on how much the decision maker wants to invest, the coverage can be improved.

Table 12. Impact of availability of an ambulance on the first objective function

Time interval	Percentage of coverage				
	1	2	3 (basic)	4	5
00:00-04:00	0.928	0.945	0.962	0.971	0.983
04:00-08:00	0.917	0.930	0.953	0.966	0.975
08:00-12:00	0.856	0.873	0.898	0.912	0.929
12:00-16:00	0.895	0.911	0.935	0.945	0.958
16:00-20:00	0.872	0.899	0.917	0.924	0.934
20:00-24:00	0.905	0.928	0.942	0.951	0.967
00:00-24:00	0.895	0.914	0.934	0.945	0.957

3.4.4. Time Dependency Versus Time-independency in the Parameter of the Model

As I discussed previously, the proposed model of this chapter is time-dependent. It means that the busy fraction, travel time, demand, and number of ambulances are constant during the day. Therefore, I solve the model in a time-independent condition and compare it with a time-dependent condition to see how much coverage and number of locations alter. The result, which is brought in Table 13, shows that running the model in a time-independent condition resulted in the loss of 1.7 % of the total received calls.

Table 13. Effects of penalties

	Time independent	Time-dependent
Expected coverage	0.917	0.934
Number of locations	6	6
Number of relocations	0	4

3.4.5. Compared with Published Studies

The result of this chapter has been compared with published studies to show how much the proposed model could bring benefit to the decision-makers (Berg & Aardal, 2015) and (Raviarun A. Nadar, 2021). By proposing the model of this chapter, I could obtain better results in terms of expected coverage and minimize the total costs of opened locations and ambulances' relocations between those opened locations. For example, Berg et al. (2015) reported the expected coverage of their time-dependent MEXCLP with the value of 91.5%, while this value for the model of this chapter is 93.4%. Also, the result of time-independent MEXCLP has been compared, and it found that the model of this chapter resulted in better coverage with a value of 91.7% versus 90.15%. This comparison shows that considering time dependency on travel time, availability of ambulances, and demand, simultaneously resulted in reaching better ambulance coverage in the network, which led to delivery services efficiently to the patient's needs.

3.4.6. Scalability

To show the scalability of the model, the proposed model has been tested on large-size networks to ensure that the model has capability in real conditions. The network considered in this chapter consists of 161 demand zone, and 9 of these demand zone are indicated as candidate base locations for ambulance locations. Figure 17 shows the demand points across the real network. Also, the input parameters of the model have been provided in Table 14.

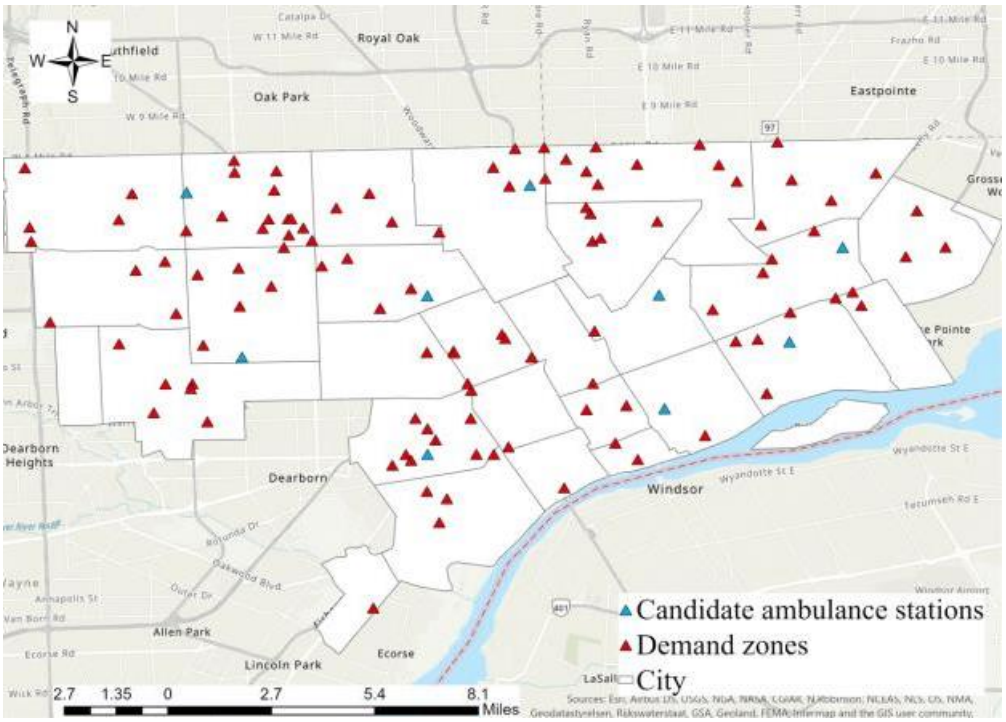

Figure 17. Map of demand points in real condition

Table 14. Data input of the real network

Interval	Total calls	Travel time	Number of ambulances	Busy fraction
00:00–02:00	27,380	1.005	13	0.2678
02:00–04:00	23,212	0.983	13	0.1973
04:00–06:00	18,942	1.036	13	0.1625
06:00–08:00	18,139	1.081	13	0.2275
08:00–10:00	39,430	1.008	18	0.3811
10:00–12:00	44,518	0.970	18	0.4912
12:00–14:00	47,526	0.979	18	0.5483
14:00–16:00	49,527	0.981	18	0.5151
16:00–18:00	35,342	1.003	17	0.4866
18:00–20:00	41,808	0.987	17	0.4200
20:00–22:00	42,505	0.962	17	0.3639
22:00–24:00	33,114	0.999	17	0.2837

Based on the input data into the model of this chapter, the result has been reported in Table 15. Again, the CPLEX solver was utilized to solve the model in a large size in just 10 minutes and 38 seconds to provide the result.

As can be seen, in Table 15, the expected coverage is around 97%, which shows the expected coverage that has been reported by the solver. Also, a total number of 5 locations has been chosen for the ambulance locations among the nine available candidate locations, and a total number of 3 relocations have been selected to occur between different ambulance stations in different time periods of the network.

On the other hand, the result of the model has been compared with existing conditions, which shows that the model could find better results in terms of expected coverage and the total number of locations and their relocations. Also, by comparing the result of the model with the current situation, the number of locations that need to be established has decreased from 9 to 5, which shows a reduction of almost 44.5 %. Also, the same situation has been obtained for the total number of relocations in the network, with a 50 % decrease.

Table 15. Result of the real network

Title	Value	
	Optimal	Current
Expected coverage	0.9751	0.9328
Number of locations	5	9
Number of relocations	3	6

3.5. Conclusion

This chapter proposed a time-dependent probabilistic location model for emergency medical services. By presenting a mathematical model, I could find optimal locations of ambulance stations across the network to serve received calls. Also, my model has considered relocations in

different periods to avoid the high costs of opening ambulance stations. After running the model, I understood that the model was able to obtain an optimal solution by covering more demand compared to the model in the literature. Indeed, my proposed model, compared with the time-independent model in the literature, could find better coverage. More interestingly, I found that my model was sensitive to some critical parameters, including penalties for opening ambulance stations and the relocations of ambulances between different periods, which resulted in more coverage (see Table 10) with more number of opening ambulance stations.

Considering temporal variance in ambulance allocation enables better planning in estimating ambulance needs and coverage. This improvement comes at the expense of a more computationally challenging problem that requires longer to solve and more ambulance moves between intervals. These factors imply that, despite the minor benefits, dividing a day into more periods may not be as beneficial as increasing the number of periods. Effective location-allocation decisions are essential since they impact short-term EMS planning decisions such as crew planning, ambulance relocation, and real-time route planning for ambulances. For instance, the number of ambulances needed at various times of the day can be used to decide the number of crews to assign and the scheduling of shifts for these crews.

The sensitivity analysis has been done on some crucial parameters of the proposed model. First, the response time has been involved in evaluating its effect on the expected coverage of the model. It found that the coverage value improved by increasing the response time. The second parameter was to involve the effects of two penalties for opened locations and their relocations between different periods. The results confirmed that ignoring the penalties in the second objective function resulted in improving the expected coverage considerably but led to an increase significantly in the total number of locations and relocations in the model with 11 and 18,

respectively. Therefore, it can be concluded that the trade-off between the costs of opening ambulance locations and service coverage helps the decision-makers to decide properly.

Also, the model has been solved in two conditions, in which the busy fraction, travel time, demand, and the number of ambulances were constant during the day (time-independent), compared to other conditions where the model was time-dependent. The result represented that running the model in a time-independent condition led to the loss of 1.7 % of the total received calls. The last parameter of the model was the impact of ambulance availability on the first objective function. The result confirmed that increasing the availability of an ambulance resulted in improving the coverage of ambulances in the network, which is expected.

For future studies, researchers can focus more on factors including disruption, weekends, and seasonal factors that might affect the model's result. Also, some solution procedures can be applied for solving the model in a large size by utilizing a meta-heuristics algorithm in the fastest time.

4. CONCLUSION

This research was carried out to determine performance measures of the EMS system in the US. The second chapter of this dissertation determined the staff requirement in the service system using the Markovian chain process and SIPP approach with random cyclic demand and state-dependent service rates. In this chapter, three approaches have been proposed, including SIPP Average, SIPP Max, and SIPP Mix, with two increasing and decreasing scenarios by applying some reliable methods to evaluate the staff scheduling in the service system.

In the third chapter of this dissertation, the mathematical programming model has been proposed to determine an optimal location of the ambulance station by providing services to the patients. Also, it is assumed that each area has demand that has been defined by the number of patient calls, which need to be responded to as soon as possible to avoid the patient's loss.

Moreover, according to the result of the research, the findings are listed as follows:

1. The second chapter indicated that whenever the arrival rate has a sinusoidal function, and the service rate is a state-dependent function, the SIPP Max is more reliable for determining staff requirements. Also, this method is reliable because the probability of delays in the service system is lower than the SIPP average and SIPP Mix approaches. The family of SIPP-based techniques examined here aimed at generating the bare minimum staff numbers that were practicable.
2. In the third chapter, the time-dependent probabilistic location model has been utilized in the service system to determine the optimal location of ambulances across the network to maximize coverage. Also, the coverage rate has been compared, which resulted in better results compared with past studies. Moreover, the sensitivity analysis showed that the penalties for opening ambulance stations and the relocations of ambulances between

different periods, with more opening ambulance stations, significantly impacted the model of this chapter.

This dissertation demonstrated that the EMS could be more efficient in management. It introduced a method for staff requirements by analyzing the factors that have an effect on ambulance response time and ambulance coverage. In addition, this study has some implementations as follows:

1. According to the second chapter, the method can be used in various healthcare settings, including community health centers, walk-in clinics, and urgent care facilities. Also, implementing the SIPP approach can inform the successful implementation of capacity planning in the healthcare system and improve access to community-based healthcare to have the lowest delay in the service systems.
2. Because ambulance crews are the EMS system's first point of physical contact with patients, a faster response time can conceivably enhance the likelihood that patients would receive early medical interventions. Implementing strategies, such as providing more in-depth information to analyze the best approach, such as the best route and location for EMS responders to use quick response vehicles, may help shorten the response time of ambulances and improve the morbidity and mortality results for patients.
3. Better planning in estimating ambulance requirements and coverage is made possible by considering temporal variation in ambulance allocation. This improvement comes at the expense of a more computationally difficult problem that takes longer to solve and more ambulance relocations between intervals. These elements suggest that splitting a day up into more periods may not be as effective as increasing the number of periods, even with the marginal benefits. Additionally, selecting the number of periods that strike a balance

between the improvement in solution and the rise in computational demands and ambulance relocations is essential. Effective location-allocation choices are crucial because they influence short-term EMS planning choices like crew planning, ambulance relocation, and real-time route planning for ambulances.

For future research, the two proposed chapters can be extended as follows: In the second chapter, researchers can utilize the SIPP approach for staff requirements with available servers at each time. Moreover, researchers can apply some scenarios, such as SIPP Max and SIPP Mix, for service rate. Indeed, in the SIPP Max approach, the system selects the maximum level of service rate. Also, in the SIPP Mix approach, whenever the service rate is increasing level, the system uses the average level of service rate. Otherwise, the system uses the increased level of service rate. For the third chapter, researchers can focus more on factors including disruption, weekends, and seasonal factors that might affect the result of the model. Also, some solution procedures can be applied for solving the model in a large size by utilizing a meta-heuristics algorithm in the fastest time.

REFERENCES

A.H. Nasser, A. et al., 2020. Every minute counts: The impact of pre-hospital response time and scene time on mortality of penetrating trauma patients. *The American Journal of Surgery,* 220(1), pp. 240-244.

A.Waalewijn, R., Vos, R., G.P.Tijssen, J. & W.Koster, R., 2001. Survival models for out-of-hospital cardiopulmonary resuscitation from the perspectives of the bystander, the first responder, and the paramedic. *Resuscitation,* 51(2), pp. 113-122.

Adeyemi, O. J., Arif, A. A. & Paul, R., 2021. Exploring the relationship of rush hour period and fatal and non-fatal crash injuries in the U.S.: A systematic review and meta-analysis. *Accident; Analysis and Prevention,* Volume 163.

Adeyemi, O. J., Paul, R. & Arif, A., 2021. An assessment of the rural-urban differences in the crash response time and county-level crash fatalities in the United States. *the Journal of Rural Health.*

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

Alnowibet, K. A. & Perros, H., 2006. The Nonstationary Loss Queue: A Survey. *Communication Networks and Computer Systems,* pp. 105-125.

Andersson, H. et al., 2020. Using optimization to provide decision support for strategic emergency medical service planning – Three case studies. *International Journal of Medical Informatics,* Volume 133.

Aringhieri, R., Bruni, M., S.Khodaparasti & Essen, J., 2017. Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research,* Volume 78, pp. 349-368.

Atar, R., Keslassy, I. & Mendelson, G., 2019. Subdiffusive Load Balancing in Time-Varying Queueing Systems. *Operations Research,* 67(6), pp. 1678-1698.

Atlason, J., Epelman, M. A. & Henderson, S. G., 2007. Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. *Management Science,* 54(2), pp. 295-309.

Bakke, H. K. et al., 2015. Bystander first aid in trauma - prevalence and quality: a prospective observational study. *Acta Anaesthesiologica Scandinavica,* 59(9), pp. 1187-1193.

Batt, R. J. & Terwiesch, C., 2016. Early Task Initiation and Other Load-Adaptive Mechanisms in the Emergency Department. *Management Science,* 63(11), pp. 3531-3551.

Bélanger, V., Ruizb, A. & Soriano, P., 2019. Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research,* 272(1), pp. 1-23.

Beojone, C. V. & Souza, R. M. d., 2020. Improving the shift-scheduling problem using non-stationary queueing models with local heuristic and genetic algorithm. *Pesquisa Operacional,* Volume 40.

Berg, P. L. d. & Aardal, K., 2015. Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research,* 242(2), pp. 383-389.

Bhat, U. N., 2008. *An Introduction to Queueing Theory: Modeling and Analysis in.* s.l.:Springer.

Boutilier, J. J. & Chan, T. C. Y., 2020. Ambulance Emergency Response Optimization in Developing Countries. *Operations Research,* 68(5), pp. 1315-1334.

Brotcorne, L., Laporte, G. & Semet, F., 2003. Ambulance location and relocation models. *European Journal of Operational Research,* 147(3), pp. 451-463.

Brown, J., Sajankila, N. & Claridge, J. A., 2017. Prehospital Assessment of Trauma. *Surgical Clinics of North America,* 97(5), pp. 961-983.

Bruin, A. M. d., Rossum, A. C. v., Visser, M. C. & Koole, G. M., 2007. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science volume* , Volume 10, p. 125–137.

Budge, S., Ingolfsson, A. & Zerom, D., 2010. Empirical analysis of ambulance travel times: the case of Calgary emergency medical services. *Management Science*, 56(4), pp. 716-723.

Bürger, A. et al., 2018. The Effect of Ambulance Response Time on Survival Following Out-of-Hospital Cardiac Arrest. *Deutsches Arzteblatt International,* pp. 541-548.

Buuren, M., Kommer, G. J., Mei, R. d. & Bhulai, S., 2017. EMS call center models with and without function differentiation: A comparison. *Operations Research for Health Care*, Volume 12, pp. 16-28.

Byrne, J. P. et al., 2019. Association Between Emergency Medical Service Response Time and Motor Vehicle Crash Mortality in the United States. *JAMA Surgery,* 154(4), pp. 286-293.

Byrne, J. P. et al., 2015. Redefining “dead on arrival”: Identifying the unsalvageable patient for the purpose of performance improvement. *Journal of Trauma and Acute Care Surgery*, 79(5), pp. 850-857.

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

C.Dietz, D., 2011. Practical scheduling for call center operations. *Omega,* 39(5), pp. 550-557.

Calland, J. F. et al., 2012. The effect of dead-on-arrival and emergency department death classification on risk-adjusted performance in the American College of Surgeons Trauma Quality Improvement Program. *The Journal of Trauma and Acute Care Surgery,* 73(5), pp. 1086-1091.

Call, D. A., Medina, R. M. & Black, A. W., 2019. Causes of weather-related crashes in Salt Lake county, Utah. *The Professional Geographer,* 71(2), pp. 253-264.

Cantwell, K., Dietze, P., Morgans, A. E. & Smith, K., 2013. Ambulance demand: random events or predicable patterns?. *Emergency Medicine Journal,* 30(11), pp. 883-887.

Cantwell, K. et al., 2015. Time of Day and Day of Week Trends in EMS Demand. *Prehospital Emergency Care,* 19(3), pp. 425-431.

Carlson, L. C., Reynolds, T. A., Wallis, L. A. & Hynes, E. J. C., 2019. Reconceptualizing the role of emergency care in the context of global healthcare delivery. *Health Policy and Planning,* 34(1), pp. 78-82.

Chen, B., Maio, R. F., Green, P. E. & Burney, R. E., 1995. Geographic variation in preventable deaths from motor vehicle crashes. *The Journal of Trauma,* 38(2), pp. 228-232.

Church, R. & ReVelle, C., 1974. The maximal covering location problem. *Papers of the Regional Science Association,* Volume 32, pp. 101-118.

Cruz, M. C. & Ferenchak, N. N., 2020. Emergency Response Times for Fatal Motor Vehicle Crashes, 1975–2017. *Transportation Research Board,* 2674(8).

Dai, J. G. & Shi, P., 2017. A Two-Time-Scale Approach to Time-Varying Queues in Hospital Inpatient Flow Management. *Operations Research,* 65(2), pp. 514-536.

Daskin, M. S., 1983. A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution. *Transportation Science,* 17(1), pp. 48-70.

Defraeye, M. & Nieuwenhuyse, I. V., 2016. Staffing and scheduling under nonstationary demand for service: A literature review. *Omega,* Volume 58, pp. 4-25.

Delasay, M., Ingolfsson, A. & Kolfal, B., 2016. Modeling Load and Overwork Effects in Queueing Systems with Adaptive Service Rates. *Operations Research,* 64(4), pp. 867-885.

Delasay, M., Ingolfsson, A., Kolfal, B. & Schultz, K., 2019. Load effect on service times. *European Journal of Operational Research,* Volume 279, pp. 673-686.

Detroit's Open Data Portal, 2020. *City of Detroit Open Data Portal.* [Online] Available at: https://data.detroitmi.gov/datasets/detroitmi::911-calls-for-service/about [Accessed 8 September 2020].

Dibene, J. C. et al., 2017. Optimizing the location of ambulances in Tijuana, Mexico. *Computers in Biology and Medicine,* Volume 80, pp. 107-115.

Di, T. S. et al., 2020. Factors associated with delayed ambulance response time in hospital university sains malaysia, Kubang kerian, Kelantan. *Malaysian Journal of Public Health Medicine,* 20(1).

Erkut, E., Ingolfsson, A. & Erdoğan, G., 2008. Ambulance location for maximum survival. *Naval Research Logistics,* 55(1), pp. 42-58.

Erlang, A., 1917. Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges. *Post Office Electrical Engineer's Journal,* Volume 10, pp. 189-197.

Estochen, B., Strauss, T. & Souleyrette, R. R., 1998. An assessment of emergency response vehicle pre-deployment using GIS identification of high-accident density locations.

Feero, S., R Hedges, J., Simmons, E. & Irwin, L., 1995. Does out-of-hospital EMS time affect trauma survival?. *The American Journal of Emergency Medicine,* 13(2), pp. 133-135.

Feldman, Z., Mandelbaum, A., Massey, W. A. & Whitt, W., 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science,* 54(2), pp. 324-338.

G.Bakalos, et al., 2011. Advanced life support versus basic life support in the pre-hospital setting: A meta-analysis. *Resuscitation,* 82(9), pp. 1130-1137.

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

Goldberg, J., 2004. Operations research models for the deployment of emergency services vehicles. *EMS Management Journal,* 1(1), pp. 20-39.

Gonzalez, R. P. et al., 2009. Does increased emergency medical services prehospital time affect patient mortality in rural motor vehicle crashes? A statewide analysis. *The American Journal of Surgery,* 197(1), pp. 30-34.

Gopalakrishnan, S., 2012. A Public Health Perspective of Road Traffic Accidents. *Journal of Family Medicine and Primary Care,* 1(2), pp. 144-150.

Green, L. & Kolesar, P., 1991. The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals. *Management Science,* 37(1), pp. 84-97.

Green, L. V. & Kolesar, P. J., 1997. The Lagged PSA for Estimating Peak Congestion in Multiserver Markovian Queues with Periodic Arrival Rates. *Management Science,* 43(1), pp. 80-87.

Green, L. V. & Kolesar, P. J., 1998. A Note on Approximating Peak Congestion in Mt/G/∞ Queues with Sinusoidal Arrivals. *Management Science,* 44(11-part-2), pp. S137-S144.

Green, L. V., Kolesar, P. J. & Soares, J., 2001. Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research,* 49(4), pp. 549-564.

Green, L. V., Kolesar, P. J. & Soares, J., 2009. An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management,* 12(1), pp. 46-61.

Green, L. V. & Soares, J., 2007. Note—Computing Time-Dependent Waiting Time Probabilities in M(t)/M/s(t) Queuing Systems. *Manufacturing & Service Operations Management,* 9(1), pp. 54-61.

Green, L. V., Soares, J., Giglio, J. F. & Green, R. A., 2006. Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing. *Academic Emergency Medicine,* 13(1), pp. 61-68.

Griffin, R. & McGwin, G., 2013. Emergency Medical Service Providers' Experiences with Traffic Congestion. *The Journal of Emergency Medicine,* 44(2), pp. 398-405.

Gross, D., Shortle, J. F., Thompson, J. M. & Harris, C. M., 1985. *Fundamentals of Queuing Theory.* 2nd Edition ed. New York: John Wiley & Sons.

Gross, D., Shortle, J. F., Thompson, J. M. & Harris, C. M., 1998. *Fundamentals of Queueing Theory.* Fourth Edition ed. s.l.:Wiley.

Gunduz, M. & Karacan, H. V., 2017. Assessment of abnormally low tenders: a multinomial logistic regression approach. *Technological and Economic Development of Economy* , 23(6), pp. 848-859.

He, Z., Qin, X., Renger, R. & Souvannasacd, E., 2019. Using spatial regression methods to evaluate rural emergency medical services (EMS). *The American Journal of Emergency Medicine,* 37(9), pp. 1633-1642.

Holmén, J. et al., 2020. Shortening Ambulance Response Time Increases Survival in Out-of-Hospital Cardiac Arrest. *Journal of the American Heart Association,* 9(21).

Ingolfsson, A. et al., 2007. A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline. *INFORMS Journal on Computing,* 19(2), pp. 201-214.

Ingolfsson, A., Budge, S. & Erkut, E., 2008. Optimal ambulance location with random delays and travel times. *Health Care Management Science volume,* Volume 11, pp. 262-274.

Ingolfsson, A., Campello, F., Wu, X. & Cabral, E., 2010. Combining integer programming and the randomization method to schedule employees. *European Journal of Operational Research,* 202(1).

Ingolfsson, A., Haque, M. & Umnikov, A., 2002. Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research,* 139(3), pp. 585-597.

Inman, R. R., 1999. Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production and Operations Management,* 8(4), pp. 409-432.

Insurance Information Institute, 2022. *acts + Statistics: Highway Safety.,* s.l.: s.n.

Insurance Institute for Highway Safety, 2015. s.l.: s.n.

Izady, N. & Worthington, D., 2012. Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research,* 219(3), pp. 531-540.

J.L.Vile, J.W.Gillard, P.R.Harper & V.A.Knight, 2016. Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. *Operations Research for Health Care,* Volume 8, pp. 42-52.

J.Trowbridge, M., J.Gurka, M. & E.O'Connor, R., 2009. Urban sprawl and delayed ambulance arrival in the U.S. *American Journal of Preventive Medicine,* 37(5), pp. 428-432.

Jaeker, J. A. B. & Tucker, A. L., 2017. Past the Point of Speeding Up: The Negative Effects of Workload Saturation on Efficiency and Patient Severity. *Management Science,* 63(4), pp. 901-1269.

Jaffe, E., 2021. *Far Beyond Rush Hour: The Incredible Rise of Off-Peak Public Transportation.* s.l.:s.n.

Jennings, O. B., Mandelbaum, A., Massey, W. A. & Whitt, W., 1996. Server Staffing to Meet Time-Varying Demand. *Management Science,* 42(10), pp. 1383-1394.

Jou, R.-C. & Chao, M.-C., 2021. Fail to Yield? An Analysis of Ambulance Crashes in Taiwan. *Sustainability,* 13(3).

Kalyanaraman, R. & Sundaramoorthy, A., 2019. A Markovian single server working vacation queue with server state dependent arrival rate and with randomly varying environment. *AIP Conference Proceedings,* 2177(1).

Karau, S. J. & D.William, K., 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology,* 65(4), pp. 681-706.

Katayama, Y. et al., 2019. Prehospital factors associated with death on hospital arrival after traffic crash in Japan: a national observational study. *BMJ open.*

Kc, D. S. & Terwiesch, C., 2009. Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations. *Management Science,* 55(9), pp. 1486-1498.

KC, D. S. & Terwiesch, C., 2012. An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit. *Manufacturing & Service Operations Management,* 14(1), pp. 50-65.

Khursheed, M. et al., 2015. Dead on arrival in a low-income country: results from a multicenter study in Pakistan. *BMC Emergency Medicine,* 15(S8).

Knight, V., Harper, P. & Smith, L., 2012. Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega,* 40(6), pp. 918-926.

König, D. & Schmidt, V., 1980. Imbedded and Non-Imbedded Stationary Characteristics of Queueing Systems with Varying Service Rate and Point Processes. *Journal of Applied Probability,* 17(3), pp. 753-767.

Koole, G. & Sluis, E. v. d., 2003. Optimal Shift Scheduling with a Global Service Level Constraint. *IIE Transactions,* 35(11), pp. 1049-1055.

Kumar, A. et al., 2017. Trend Analyses of Emergency Medical Services for Motor Vehicle Crashes: Michigan Case Study. *Transportation Research Board,* 2635(1).

Lambert, T. E. & Meyer, P., 2005. Ex-Urban Sprawl as a Factor in Traffic Fatalities and EMS Response Times in the Southeastern United States.

Larson, R. & Odoni, A., 2007. *Urban Operations Research.* s.l.:Transportation Research Board.

Lee, J., Abdel-Atya, M., Cai, Q. & Wang, L., 2018. Effects of emergency medical services times on traffic injury severity: A random effects ordered probit approach. *Traffic Injury Prevention,* 19(6), pp. 577-581.

Lee, J., Abdel-Aty, M., Cai, Q. & Wang, L., 2018. Analysis of Fatal Traffic Crash-Reporting and Reporting-Arrival Time Intervals of Emergency Medical Services. *Transportation Research Board,* 2672(32), pp. 61-71.

Leknes, H. et al., 2017. Strategic ambulance location for heterogeneous regions. *European Journal of Operational Research,* 260(1), pp. 122-133.

Lilley, R. et al., 2019. Geographical and population disparities in timely access to prehospital and advanced level emergency care in New Zealand: a cross-sectional study. 9(7).

Liu, Y. & Whitt, W., 2012. Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research,* 60(6), pp. 1551-1564.

Long, E. F. & Mathews, K. S., 2018. The Boarding Patient: Effects of ICU and Hospital Occupancy Surges on Patient Flow. *Production and Operations Management,* 27(12), pp. 2122-2143.

Louriz, M. et al., 2012. Determinants and outcomes associated with decisions to deny or to delay intensive care unit admission in Morocco. *Intensive Care Medicine,* Volume 38, pp. 830-837.

Lovely, R. et al., 2018. Injury Severity Score alone predicts mortality when compared to EMS scene time and transport time for motor vehicle trauma patients who arrive alive to hospital. *Traffic Injury Prevention,* Volume 19.

Lu, Y. & Davidson, A., 2017. Fatal motor vehicle crashes in Texas: needs for and access to emergency medical services. *Annals of GIS,* 23(1), pp. 41-54.

Mandelbaum, A. & Massey, W. A., 1995. Strong Approximations for Time-Dependent Queues. *Mathematics of Operations Research,* 20(1), pp. 33-64.

Ma, N. & Whitt, W., 2019. Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems,* 9(3), pp. 261-290.

Massey, W. A., 2002. The Analysis of Queues with Time-Varying Rates for Telecommunication Models. *Telecommunication Systems,* Volume 21, pp. 173-204.

Massey, W. A. & Whitt, W., 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems,* Volume 25, pp. 157-172.

McFadden, D., 1987. Regression-based specification tests for the multinomial logit model. *Journal of Econometrics,* 34(1-2), pp. 63-82.

McLay, L. A. & Mayorga, M. E., 2010. Evaluating emergency medical service performance measures. *Health Care Management Science,* Volume 13, pp. 124-136.

Medrano, N. W. et al., 2019. Multi-Institutional Multidisciplinary Injury Mortality Investigation in the Civilian Pre-Hospital Environment (MIMIC): a methodology for reliably measuring prehospital time and distance to definitive care. *Trauma Surgery & Acute Care,* 4(1).

Meng, Q. & Weng, J., 2013. Uncertainty Analysis of Accident Notification Time and Emergency Medical Service Response Time in Work Zone Traffic Accidents. *Traffic Injury Prevention,* 14(2), pp. 150-158.

Mojir, K. Y. & Pilemalm, S., 2016. Actor-centred emergency response systems: a framework for needs analysis and information systems development. *International Journal of Emergency Management (IJEM),* 12(4).

Momenitabar, M. et al., 2022. Designing a sustainable closed-loop supply chain network considering lateral resupply and backup suppliers using fuzzy inference system. *Environment, Development and Sustainability,* pp. 1-34.

Momenitabar, M., Ebrahimi, Z. D., Arani, M. & Mattson, J., 2022. Robust possibilistic programming to design a closed-loop blood supply chain network considering service-level maximization and lateral resupply. *Annals of Operations Research,* pp. 1-43.

National Center for Statistics and Analysis, 2019. *2018 Fatal Motor Vehicle Crashes: Overview. Traffic Safety Fact,* s.l.: National Highway Traffic Safety Administration.

National Center for Statistics and Analysis, 2020. *Preview of Motor Vehicle Traffic Fatalities in 2019. Traffic Safety Fact,* s.l.: National Highway Traffic Safety Administration;.

National Center for Statistics and Analysis, 2022. *Early Estimate of Motor Vehicle Traffic Fatalities for the First 9 Months (January–September) of 2021,* s.l.: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration;, 2017. *MMUCC Guideline: Model Minimum Uniform Crash Criteria,* Washington,DC.: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration, 2017. *FARS annual crash statistics 2017,* s.l.: s.n.

National Highway Traffic Safety Administration, 2018. *Crashes, by Time of Day, Day of Week, and Crash Severity,* s.l.: United States Department of Transportation.

National Highway Traffic Safety Administration, 2019. *Fatality Data Show Continued Annual Decline in Traffic Deaths,* s.l.: United States Department of Transportation.

National Highway Traffic Safety Administration, 2020. *Emergency Medical Services NEMSIS Data Dictionary.* s.l.:National Highway Traffic Safety Administration.

National Safety Council, 2021. *Preliminary Semiannual Estimates. National Security Council,* s.l.: s.n.

Organization, W. H., 2009. *Global status report on road safety time for action* , s.l.: World Health Organization.

Organization, W. H., 2018. *Global status report on road safety 2018,* s.l.: World Health Organization.

Paleti, R. & Naveen Eluru, C. R. B., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis and Prevention,* 42(6), pp. 1839-1854.

Paulo B. Goes, N. I. & Mingfeng Lin, J. L. Z., 2018. When More Is Less: Field Evidence on Unintended Consequences of Multitasking. *Management Science,* 64(7), pp. 3033-3054.

Philip McArthur, D., A.Gregersen, F. & P.Hagen, T., 2014. Modelling the cost of providing ambulance services. *Journal of Transport Geography,* Volume 34, pp. 175-184.

Pirdavani, A., Bellemans, T., Brijs, T. & Wets, G., 2014. Application of Geographically Weighted Regression Technique in Spatial Analysis of Fatal and Injury Crashes. *Journal of Transportation Engineering,* 140(8).

Polley, E. C., Rose, S. & Laan, M. J. v. d., 2011. Super Learning. In: New york, NY: Springer, pp. 43-66.

Pons, P. T. et al., 2005. Paramedic response time: does it affect patient survival?. *Academic Emergency Medicine*, 12(7), pp. 594-600.

R.Aringhieri, M.E.Bruni, S.Khodaparasti & Essen, J., 2017. Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research,* Volume 78, pp. 349-368.

Rajagopalan, H. K., Saydam, C. & Xiao, J., 2008. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research,* 35(3), pp. 814-826.

Raviarun A. Nadar, J. J. J. J. T., 2021. Strategic location of ambulances under temporal variation in demand and travel time using variable neighbourhood search based approach. *Computers & Industrial Engineering,*, Volume 162, p. 107780.

Repede, J. F. & Bernardo, J. J., 1994. Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research,* 75(3), pp. 567-581.

Reuter-Oppermann, M., Berg, P. L. v. d. & Vile, J. L., 2017. Logistics for Emergency Medical Service systems. *Health System,* 6(3), pp. 187-208.

ReVelle, C. et al., 1977. Facility location: a review of context-free and EMS models. *Health Service Research,* 12(2), pp. 129-146.

ReVelle, C. & Hogan, K., 1989. The Maximum Availability Location Problem. *Transportation Science,* 23(3), pp. 192-200.

Robbins, T. R., Medeiros, D. J. & Harrison, T. P., 2010. Does the Erlang C model fit in real call centers?. *Proceedings of the 2010 Winter Simulation Conference,* pp. 2853-2864.

S.Roudsari, B. et al., 2007. Emergency Medical Service (EMS) systems in developed and developing countries. *Injury,* 38(9), pp. 1001-1013.

Saghafian, S., Austin, G. & Traub, S. J., 2015. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering,* 5(2), pp. 101-123.

Salmon, A., Rachuba, S., Briscoe, S. & Pitt, M., 2018. A structured literature review of simulation modelling applied to Emergency Departments: Current patterns and emerging trends. *Operations Research for Health Care,* Volume 19, pp. 1-13.

Sánchez-Mangas, R., García-Ferrrer, A., Juan, A. & Arroyo, A. M., 2010. The probability of death in road traffic accidents. How important is a quick medical response?. *Accident Analysis & Prevention,* 42(4), pp. 1048-1056.

Saydam, C., Rajagopalan, H. K., Sharer, E. & Lawrimore-Belanger, K., 2013. The dynamic redeployment coverage location model. *Health Systems,* 2(2), pp. 103-119.

Schmid, V. & F.Doerner, K., 2010. Ambulance location and relocation problems with time-dependent travel times. 207(3), pp. 1293-1303.

Schultz, K. L., 2003. Overcoming the dark side of worker flexibility. *Journal of Operations Management,* 21(1), pp. 81-92.

Schwarz, J. A., Selinka, G. & Stolletz, R., 2016. Performance analysis of time-dependent queueing systems: Survey and classification. *Omega,* Volume 63, pp. 170-189.

Selinka, G., Franz, A. & Stolletz, R., 2016. Time-dependent performance approximation of truck handling operations at an air cargo terminal. *Computers & Operations Research,* Volume 65, pp. 164-173.

Singer, M. & Donoso, P., 2008. Assessing an ambulance service with queuing theory. *Computers & Operations Research,* 35(8), pp. 2549-2560.

Sinreich, D. & Jabali., O., 2007. Staggered work shifts: A way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science,* 10(3), pp. 293-308.

Staats, B. R. & Gino, F., 2012. Specialization and Variety in Repetitive Tasks: Evidence from a Japanese Bank. *Management Science,* 58(6), pp. 1141-1159.

Stainsby, D., MacLennan, S. & Hamilton, P., 2000. Management of massive blood loss: a template guideline. *British Journal of Anaesthesia,* 85(3), pp. 487-491.

Syahputri, K. et al., 2020. Application of Fuzzy C-Means in Level Clustering of Traffic Accident Vulnerability. *IOP Conference Series: Materials Science and Engineering.*

Tan, T. F. & Netessine, S., 2014. When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity. *Management Science,* 60(6), pp. 1574-1593.

Tan, X., Knessl, C. & Yang, Y. P., 2013. On finite capacity queues with time dependent arrival rates. *Stochastic Processes and their Applications,* 123(6), pp. 2175-2227.

Thind, A. et al., 2015. Prehospital and Emergency Care. pp. 245-262.

Thompson, G. M., 1993. Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. *Journal of Operations Management,* 11(3), pp. 269-287.

Toregas, C., Swain, R., ReVelle, C. & Bergman, L., 1971. The Location of Emergency Service Facilities. *Operations Research,* 19(6), pp. 1363-1373.

Trujillo, L., Álvarez-Hernández, G., Maldonado, Y. & Vera, C., 2020. Comparative analysis of relocation strategies for ambulances in the city of Tijuana, Mexico. Volume 116.

U.S. Environmental Protection Agency, 2021. *United States Environmental Protection Agency.* [Online]
Available at: https://www.epa.gov/smartgrowth/smart-location-database-technical-documentation-and-user-guide
[Accessed 16 June 2021].

V.Bélanger, A.Ruiz & P.Soriano, 2019. Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research,* 272(1), pp. 1-23.

Valenzuela, T. D. et al., 2000. Outcomes of rapid defibrillation by security officers after cardiac arrest in casinos. *Journal of Medicine,* 343(17), pp. 1206-1209.

W.K.Grassmann, 1977. Transient solutions in markovian queueing systems. *Computers & Operations Research,* 4(1), pp. 47-53.

Wang, H. & Xu, T., 2007. Attitudes to triage of Chinese emergency room patients in a Beijing tertiary hospital. *Emergency Medical Journal,* 24(3).

Wang, N. et al., 2018. Study of time-dependent queuing models of the national airspace system. *Computers & Industrial Engineering,* Volume 117, pp. 108-120.

Weather Underground, 2020. *Weather underground.* [Online]
Available at: https://www.wunderground.com/
[Accessed 8 September 2020].

Wei Lam, S. S. et al., 2015. Factors affecting the ambulance response times of trauma incidents in Singapore. *Accident Analysis & Prevention,* Volume 82, pp. 27-35.

Whitt, W., 1991. The Pointwise Stationary Approximation for Mt/Mt/s Queues Is Asymptotically Correct As the Rates Increase. *Management Science,* 37(3), pp. 307-314.

Whitt, W. & You, W., 2019. Time-varying robust queueing. *Operations Research,* 67(6), pp. 1766-1782.

Whitt, W. & You, W., 2020. Heavy-traffic limits for stationary network flows. *Queueing Systems,* Volume 95, pp. 53-68.

Wubben, B. M., Denning, G. M. & Jennissen, C. A., 2019. The Effect of All-Terrain Vehicle Crash Location on Emergency Medical Services Time Intervals. *Safety,* 5(4).

Wu, C. (., Bassamboo, A. & Perry, O., 2018. Service System with Dependent Service and Patience Times. *Management Science,* 65(3), pp. 1151-1172.

Xie, S.-H.et al., 2016. Mortality from road traffic accidents in a rapidly urbanizing Chinese city: A 20-year analysis in Shenzhen, 1994–2013. *Traffic Injury Prevention* , 17(1), pp. 39-43.

Xie, Z. & Or, C., 2017. Associations Between Waiting Times, Service Times, and Patient Satisfaction in an Endocrinology Outpatient Department: A Time Study and Questionnaire Survey. *Inquiry,* Volume 54.

Y.Dimitrakopoulos & A.N.Burnetas, 2016. Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control. *European Journal of Operational Research,* 252(2), pp. 477-486.

Yasmina, S., Eluru, N. & R.Pinjari, A., 2015. Analyzing the continuum of fatal crashes: A generalized ordered approach. *Analytic Methods in Accident Research,* Volume 7, pp. 1-15.

Yasunaga, H. et al., 2011. Population density, call-response interval, and survival of out-of-hospital cardiac arrest. *International Journal of Health Geographics,* 10(26).

Yom-Tov, G. B. & Mandelbaum, A., 2014. Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing. *Manufacturing & Service Operations Management,* 16(2), pp. 283-299.

Yoon, S. & A.Albert, L., 2020. A dynamic ambulance routing model with multiple response. *Transportation Research Part E: Logistics and Transportation Review,* Volume 133.

Zeng, Q. et al., 2019. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. *Accident Analysis & Prevention,* Volume 127, pp. 87-95.