UNIVERSITY OF CALIFORNIA

Santa Barbara

Emergency Medical Service Ambulance System Planning: History and Models

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Geography

by

Carlos Alain Baez Tapia

Committee in charge:

Professor Richard L. Church, Chair

Professor Stuart H. Sweeney

Professor Alan T. Murray

December 2017

The thesis of Carlos Alain Baez Tapia is approved.

_____

Alan T. Murray

_____

Stuart H. Sweeney

_____

Richard L. Church, Committee Chair

December 2017

Emergency Medical Service Ambulance System Planning: History and Models

ACKNOWLEDGEMENTS

First, I would like to dedicate this thesis to my grandpa Carlos Baez who is unfortunately not here to see me finish this thesis. Que en paz descanse.

Second, I want to thank my parents Carlos and Elva Baez for the incredible sacrifices they had to make beginning with our journey from Michoacán, Mexico all the way through UCSB and today.

Third, I want to thank my family: my grandmas, uncles (including Gordo), aunts, my brother Erwin and even my sister Amber. In one way or another I would not be here without your help.

Fourth, I want to thank my colleagues at UCSB and elsewhere that helped in more way or another. I especially want to thank Grant Brokenzie, Dan Ervin, Mike Alonzo, Kevin Mwenda, Olaf Mezner, Lumari Pardo, Jacky Banks(!), Matt and Tim Niblett, Yiting Ju, Dylan Parenti, the fifth floor, and everyone else (not that I mean to exclude anyone but I have a filing deadline!). I also want to thank Jose Saleta who was invaluable to every graduate student at UCSB Geography and just a wonderful person who is dearly missed.

Fifth, I want to thank all my friends (not mentioned above) that helped me get here in one way or another. I don't think I could've done it without your support (that includes letting me crash on your couch). So thank you Jesse Vasquez, Erin Corrigan, Sonya, Michelle Himden, Saba Dowlatshahi, Ryan Darby, Svetlin "Foolio" Bostandjiev, Don "I love Mayo" Buyers, Maciek Baranski, Atif Khan, Taylor Horgan, Laurel Patterson, all my friends from the Merton, Fernando Castorena, Sergio Herrera and Lenny, Yuri N., Andrew "Ding Dong" Johnson (unfortunately), and that's again just at the top of my head.

Sixth, shout out to my bros/the gang Nick, Yasin, Cody, Gaykers, Rel (I guess), Colleen, Vicky, Debi, Andrew, Quade, JeBee, Arman Bro, Jen and Dorian, and also, Elaine and Joe, who unfortunately can't be here with us. For lack of better words, your friendship means so much to me.

Seventh, I really want to thank my adviser Rick "The Doc" Church. I don't know how I'm still here or where I would be it wasn't for his brilliance, compassion, and patience. We're still not done but thanks keeping me along!

Lastly, I would like to say, THANKS OBAMA. In 2012, I didn't know if I was going to be able to stay at UCSB but out of nowhere DACA came out. So thank you, President Barrack Obama.

#RESIST

ABSTRACT

Emergency Medical Service Ambulance System Planning: History and Models

by

Carlos Alain Baez Tapia

Integer linear programming models that incorporate probabilistic and stochastic components represent one approach for capturing the stochastic nature of emergency medical service ambulance systems. This includes modeling non-deterministic call arrival and servicing rates and congestion in the ambulance network (i.e., ambulance unavailability). These models focus on maximizing the total population that can find an available ambulance within a set service time standard (s) with a probability of at least α%. In MALP the concept of *local vehicle busyness* estimates is introduced to estimate the availability of service in a neighborhood given the neighborhood's level of demand and the number of ambulance vehicles located in the neighborhood. QMALP is an extension of MALP where queue-theory derived parameters are implemented in the MALP model framework in order to relax the assumption that the probability of different ambulances being busy are independent. Despite this considerable development, several concerns remained about MALP and QMALP, namely the *districting assumption* where its assumed that a neighborhood's calls for service are served only by an ambulance in the area, that ambulances in a neighborhood only serve calls for service originating within the neighborhood, or that at least the flow of ambulance service to and from external

neighborhoods was roughly equal. Questions have been raised about the validity of MALP and QMALP's reliability estimates, that is, whether a neighborhood actually received α-reliable service.

To address these issues, we developed the Resource-Constrained Queue-based Maximum Availability Location Problem (RC-QMALP). This model is based on a location-allocation framework that (1) assigns workload from neighborhoods to ambulances located within *s* and ambulance idle capacity to neighborhoods and (2) includes additional constraints designed to help ensure the validity of the original MALP and QMALP constraints used to establish whether a neighborhood can find an available ambulance with α-reliability. We also implemented a secondary *minsum* objective that minimizes the average travel distance between ambulances and the neighborhoods they service while maintaining the priority of the MALP and QMALP coverage objective.

In this thesis, we validated RC-QMALP by comparing the reliable coverage levels predicted by the RC-QMALP to the ambulance system simulations that used the locational configurations suggested by the RC-QMALP. We found that MALP 2 and QMALP provided higher levels of reliable coverage and that RC-QMALP's secondary objective has a negligible impact on system performance. However, RC-QMALP-based models provide more accurate estimates of reliable coverage and location solutions whose simulated reliable coverage performance was always within 5% of the optimal solution with the same system parameters (we tested 1,080 different model configurations). Our work suggests that (1) more work is needed on developing simulation models that can accommodate the modeling assumptions that underlie location optimization models and that (2) service

reliability location models should consider additional factors such as ambulance workloads (and their distribution).

TABLE OF CONTENTS

1. INTRODUCTION

Emergency medical service (EMS) involves the organized provision of *pre-hospital care* to sick or injured individuals with the ultimate goal of reducing patient mortality and morbidity. An *EMS system* encompasses three general activities - *response*, *treatment to stabilize the patient*, and *transport*. This service entails, respectively, (1) responding to calls for urgent medical assistance, (2) providing medical treatment on-scene, and, if necessary, (3) transporting the sick or injured from the scene to a hospital for care. As such, the objective of an EMS system planner is to develop procedures, policies, and a resource allocation plan that effectively address each of the three outlined tasks.

Ideally, an EMSS (emergency medical service system) responds to calls for service immediately after a request for service is made, always assists patients with the most effective equipment and treatment, and delivers patients promptly and efficiently to the appropriate medical treatment facilities. The reality is, however, that EMSSs are hampered by a variety of issues. For instance, in some EMSSs ambulances located closest to a medical emergency are often not dispatched to that emergency, thus resulting in slower system response times (Dean, 2008; Williams, 2007).[1,2] In terms of the provision of pre-hospital care, Wang et al. (2005) raised concerns about advanced EMS responders losing proficiency in certain types of medical

[1] Interestingly enough, Carter et al. (1972) developed an ambulance response model that showed how a nearest-ambulance dispatching strategy could be suboptimal with respect to minimizing the average system response time. However, the EMS system surveys of Williams (2007) and Dean (2008) do not suggest that this is the case.

[2] Williams (2007) cites a variety of operational practices (*e.g.*, jurisdictional issues, agency protocols, technological difficulties) as the primary explanation for this dispatching strategy, while Dean (2008) identifies the combination of a lack of ambulance location information and fixed-deployment models, as well as, ambulance crew shift change procedures in impacting response times.

treatment due to a lack of opportunities to practice or use such treatment protocols. Lastly, patient care after transport is often complicated with delayed hospital treatment due to emergency room overcrowding. In response to over congested emergency departments, some hospitals actually close their emergency department during periods of high demand and refuse treatment to new patients which results in ambulances having to transport patients to a different, often further away, hospital (Hoot & Aronsky, 2008).[3]

EMS agencies and providers of all sizes face complex challenges of every size and type that concern both internal and external factors. Economic challenges are one of the most common issues whereby EMS agencies are required to justify their operations/financial decisions, improve their efficiency, or adapt to budget cuts or downsizing. For example, in 2005 the South Ogden Fire Department measured the cost-efficiency of its operations in response to lower than expected ambulance revenues (Powers, 2005). Likewise, the County Commissioners of Pinellas County, Florida expressed concerns about the fiscal sustainability of the County's EMSS and fire response operations and commissioned a report to examine the current state of these systems and to analyze several models and proposals (Fitch & Associates, 2013). Other economic factors include the high cost of ambulance equipment (McIntire, 2003) and emergency care and transport (Rosenthal, 2013).

At the organizational level, two ongoing debates in EMSS management include the privatization and insourcing/outsourcing of EMS response and transport responsibilities. Privatizing public services is not new (Greene, 1996) but it remains a controversial matter in many communities where it is being considered (Balskovitz, 2011; Laverty, 2013). Another

[3] The emergency medicine literature typically refers to this as an *ambulance diversion problem*.

equally controversial proposal involves shifting EMS responsibilities to fire departments in what proponents see as a cost saving move while opponents claim that these changes could endanger lives (O'Toole, 2011; Welsh, Linthicum, & Lopez, 2013).

Another highly controversial issue (and the focus of this thesis) is the performance of EMSS, namely in terms of response times. A quick internet search reveals that in the last six years citizens have complained about slow service in San Francisco (T. Goldberg, 2016); Akron (Molnar, 2011); Los Angeles (Linthicum & Lopez, 2012); Sacramento (Chabria, 2016); San Jose (Colgan, 2014); and New York City (Short, 2015).

The case of Los Angeles Fire Department (LAFD) is rather notable in that numerous issues plagued the organization which ultimately led to the Los Angeles Mayor Eric Garcetti to ask for the resignation of then LAFD Chief Brian L. Cummings (Welsh et. al, 2014). In 2012, the Los Angeles Times reported on several issues with the LAFD's performance including response times of over 45 minutes for some incidents and delays in dispatch due to malfunctioning equipment (Linthicum & Lopez, 2012). LAFD firefighters expressed concerns with the organization's abilities, however, both Chief Cummings and then Mayor Antonio Villaraigosa assured the public that the city was safe. Nonetheless, a series of LAFD's system failures prompted the Mayor to call for a review of LAFD operations. In addition, issues about misleading LAFD statistics (an issue raised by the LA Times) prompted several LA City Council members to call for an audit of the LAFD.[4]

[4] The validity of these figures was particularly important as they were used to make decisions about deep cuts to EMS spending in the previous year

In late 2012, the LA Times released two additional reports about the response times in Los Angeles. The first report covered delays in response resulting from geographic and jurisdictional issues between LAFD and Los Angeles County Fire Department (LACFD) (Lopez, Welsh, & Linthicum, 2012). In this investigation, the LA Times analyzed of over 1 million LAFD responses over the previous five years and LACFD dispatch records. Their analysis revealed that LAFD rarely reached out to LACFD and that LAFD dispatchers contacted LACFD dispatchers less than 10% of the time in cases where the nearest County facility was closer to the caller than nearest City facility. They add that 70,000 of these calls were medical calls and that 1,300 of these calls concerned cases of cardiac arrest where fast response times can reduce morbidity and mortality. Also, they reported that callers located within a quarter mile from city boarders were 50% more likely to wait more than 10 minutes for first-responders to arrive. Finally, the report notes that the two agencies worked on a mutual-aid agreement in 1979 to assist LAFD with calls originating near the boundary of the City of Los Angeles and the two agencies eventually signed a formal automatic-aid agreement. This included provisions to connect their dispatching systems, however, this was never implemented and without that the process in moving a call to the LACFD takes too much time (as LAFD dispatchers have to contact LACFD dispatchers via telephone). Other major fire agencies in California are involved in mutual-aid agreements and have "automatic aid" dispatching systems including agencies from Orange County, San Diego and San Jose. Notably, eight agencies in San Diego County entered into an automatic-aid agreement as a cost-saving measure.

In the second report, the LA Times analyzed EMS response times at the block level for the City of Los Angeles (Linthicum, Welsh, & Lopez, 2012). Here they reported the LAFD

regularly failed to respond to many affluent communities within six minutes (the national time standard adopted by LAFD). The LAFD reportedly failed to respond to calls from the "affluent hillside communities stretching from Griffith Park to Pacific Palisades" within six minutes nearly 85% of the time and nearly 90% of the time to calls from the Bel-Air neighborhood. They also report that the average response time to cardiac arrest calls from Bel-Air were twice as long as the average response times of nearby communities (11 ½ minutes). Moreover, the LA Times reported that system congestion contributed to longer response times as the nearest stations could not respond to about 15% of all calls. They added that areas with a high concentration of fire stations were less prone to this issue while areas with fewer fire stations were more prone to this problem including "east San Fernando Valley, the southern edge of Playa del Rey and some neighborhoods in the Santa Monica Mountains as well as Bel-Air. LAFD officials cited difficult driving conditions in mountainous areas. LAFD Chief Brian Cummings argued that his department would need to almost double the number of fire stations to meet the six-minute respond standard. However, the LA Times notes that the LAFD's budget reduction resulted in the closing of 20% of the city's fire stations.

1.1 Thesis Scope and Motivation

In this thesis, the main focus is on the *response* component of EMSSs associated with congested EMS ambulance systems, i.e. EMSSs that frequently exhibit significantly low levels of ambulance availability. The overall goal is to improve the performance of an EMSS in terms of the availability of ambulances via the strategic management of ambulance posting/dispatch locations. To support this goal the main objective of this thesis is to develop an ambulance location planning model that prescribes effective station location/posting configurations. Here, the effectiveness of the configuration is determined by the ability of ambulances positioned in

such configurations to respond to emergency calls within some time standard a relatively high percentage of the time.

The main motivating factor in this work is that for many patients, significant delays in the provision of medical care can increase patient morbidity or reduce patient survival (Wilde, 2013). In cases of cardiac arrest, the benefits of early intervention have been consistently documented (e.g. De Maio, Stiell, Wells, & Spaite, 2003; Eisenberg, Bergner, & Hallstrom, 1980; Wik *et al*., 2003; Vukmir, 2006). Weaver *et al*. (1986) estimated the impact of a minute delay in the application of a defibrillator to as 4% decrease in chances of survival while Larsen *et al*. (1993) found a 3.2% decrease in chances of survival. Moreover, addressing issues of congestion is important in order to avoid a suboptimal or inequitable provision of service as with the case of the LAFD.

A secondary focus of this thesis is the development of a historical account and general overview of EMSSs planning, management, and analysis from a *location science* perspective. This review is mostly centered on EMSSs in the United States beginning with the early EMSSs of the 1800s before moving to early efforts from the mid-1900s to systematically plan, manage, and analyze EMSSs. Finally, we present an overview of early EMSS location models as well as the theoretical foundations of such models as *public facility location models*.

While researching the literature for this thesis, it soon became apparent that overviews of the *technical*/*operational* aspects of EMSSs mostly focus on location models and their technical properties. This is problematic because the highly selective nature of these reviews (namely with respect to covering a specific modeling paradigm) results in discussions that preclude the context or environment in which the EMSS models were developed. Of course, one cannot reasonably expect even reviews to cover everything but the lack of a comprehensive

account (as a single document or a collection of them) for a subfield that is over 50 years old is terribly concerning and not simply because of the absence of such account. The concern here is that this is indicative of a hyper-focus on EMSS subcomponents at the expense of more comprehensive EMSS research. This is not to say that there isn't a need for specialized work or that it shouldn't be a high priority, but rather that more work is needed that improves our understanding of EMSSs as a whole, rather than solely in terms of their parts.[5]

1.2 Thesis Organization

This thesis consists of three parts. Chapters 2 and 3 comprise the first section which covers the development of EMSSs in the United States. Chapter 2 begins with a historical overview of EMSS planning in the United States starting with early EMSSs from the mid-to-late 1800s and ending with the EMS revolution of the 1960s and 1970s. A discussion of the theoretical, methodological, organizational, technologic, and scientific advances related to EMSSs is also included.

The second section, comprised of Chapters 4 through 6, presents a new ambulance location problem, the *Resource Constrained Queueing Maximal Availability Location Problem* (RC-QMALP). Chapter 4 provides the background relating to the ambulance location models used to develop RC-QMALP. RC-QMALP is formally presented and discussed in Chapter 5 while Chapter 6 discusses computational results in solving RC-QMALP along with a comparison to its predecessor, QMALP. We conclude with an overview of this thesis, a discussion of future

[5] Spaite et al. (1995) raise this concern about EMS research while comparing *systems research* and *component research*. The essence of the former approach is that it addresses complex and interrelated problems that usually require complex models and high-levels of collaboration between different types of experts to address them.

work and the state of ambulance location modeling, and a reflection about EMSSs in the United States.

2. History of Emergency Medical Service System Planning

2.1 Early EMS Systems in the United States

Prior to the mid-1800s, there was a general sense of neglect and apathy towards emergency medical transport and treatment in the United States (Haller, 1990). The provision of any pre-hospital emergency care relied on volunteered efforts from nearby individuals or establishments and until the introduction of the ambulance, patients were required to walk to and from their destination or had to find some type of vehicle or apparatus that could accommodate them (Hart, 1978). With the industrialization of the United States however, came the increasing need for more suitable ways of transporting the injured (Willard, 1883).

In 1865, the first civilian EMS system appeared in the United States when the Commercial Hospital of Cincinnati began the first civilian-run and hospital based ambulance service (Pozner, et al., 2004). Shortly thereafter, another municipal based EMS system formed in New York City under the direction of Bellevue Hospital with guidance from the New York City Metropolitan Board of Health (Barkley, 1974) and by the 1880s, the number of EMS systems increased dramatically (although they were mostly confined to large urban centers). Philadelphia began the development of an EMS system (based on the New York City system) in the early 1880s (Willard, 1883) as did the District of Columbia (Barkley, 1974), the City of Cleveland (Metzenbaum, 1908), and New Orleans (Barkley, 1978).

While EMS systems proliferated throughout the United States, the quality of EMS systems also increased overall due to the work of various organizations and individuals. Advances in

medical transportation and communication technology helped EMS systems become more responsive and allowed responders to be better equipped. In terms of prehospital medical treatment, however, emergency medicine remained relatively unchanged between the American Civil War and World War 1 (Robbins, 2005; Trunkey, 2000). Moreover, in the 19th century emergency medicine educational resources or references were very limited or non-existent (Haller, 1990). Surgical textbooks of the time described how to perform various treatments but rarely discussed topics important to emergency responders such how to move patients or how to stabilize their condition or injury.[6] As for the rare emergency medical text books that existed, their intended audience were military surgeons. As such, the texts were not ideal for urban emergency responders, but did cover many topics relevant to the challenges faced by them.

In the area of transportation, by 1868, the Bellevue Hospital in New York City, via the efforts of Dr. Edward Dalton, developed the prototypical civilian ambulance by altering military ambulances (Barkley, 1974). These ambulances would prove to be more comfortable for patients, easier for drivers to manage in urban environments given their tighter turning radius, faster due to their lighter construction, and better equipped as the ambulances were reconfigured to carry less individuals in exchange for being able to carry more medical supplies (Leonard, 1885). The commercial production of civilian ambulances began in 1890 (Robbins, 2005) and in 1894, St. Louis adopted electric streetcar ambulances (Haller, 1990). The first motorized ambulance appeared in 1899 when five Chicago businessmen donated a battery-

[6] Haller (1990) notes one exception- an article in *Wood's Medical and Surgical Monograph* (1890) that describes how to move sick or injured individuals.

driven ambulance wagon to the Michael Reese Hospital in Chicago (Haller, 1990; Robbins, 2005).

As for the impact of new communication technologies, one example includes New York City's early ambulance system that utilized telegraph and telephone services in order to notify more quickly hospitals of medical emergencies (Schroeder, 1902; Leonard, 1989). In this system, police officers communicated requests for emergency medical treatment to police headquarters that would then forward the message to the nearest hospital. Alternatively (and less common), phone calls for medical service were placed from signal boxes dispersed throughout the city that would notify hospitals about a request for service with an alarm.[7] A similar system was developed in Philadelphia, known as the "Gamerel System", whereby every city square would have a telephone connection between it and police headquarters (Willard, 1883; Evatt, 1886). It differed from New York City's EMS system, however, in that in Philadelphia the police department maintained some medical response equipment at its stations and would often respond to emergencies themselves rather than always forwarding calls to a hospital.

Following the 19th century, the provision of emergency medical services would become increasingly prevalent as a result of the many individual efforts by local governments, hospitals, and non-hospital civilian organizations throughout the United States to create or develop their own EMS systems (Robbins, 2005). Most notable are the efforts of the American Red Cross (ARC). Beginning in 1910, the ARC began providing standard courses about basic

[7] Each signal box was connected to a specific fire station and hospitals were connected to the systems of certain fire stations. Thus, when a call was placed at a signal box, it would notify the corresponding fire station and in turn, *all* the hospitals connected to that fire station (Haller, 1990).

first aid (Robbins, 2005). In addition, in response to the increasing number of automobile accidents, the ARC worked on increasing access to EMS along highways in the United States. Beginning in 1936, the American Red Cross established hundreds of "Emergency First Aid Stations" at various locations including fire stations, stores, inns, and gas stations. Other important institutions included various military organizations that created or advanced various medically related technologies that improved the treatment of sick or injured patients. During World War I, this included the further development of motorized ambulances, techniques for treating contaminated wounds and the practice of blood banking (Trunkey, 2000). As for World War II, the Vietnam War, and the Korean Conflict, the state of trauma care medicine advanced, as did some aspects of EMS organization and operations due to research efforts and experience gained by EMS practitioners involved in these conflicts (Robbins, 2005; Trunkey, 2000). Many of these advances were adopted by civilian EMS systems,[8] although some notable developments pertaining to pre-hospital medicine were not such as advanced on-scene medical treatment (*e.g.*, the application of intravenous fluids to a patient by a non-physician) (Robbins, 2005).

2.2 Prelude to the Quantitative Revolution in EMS System Planning and Management

Despite the technological progress in the decades following the development of civilian EMS systems and the overall rise in the number of EMS systems across the United States, by

[8] According to Pozner et al. (2004), military EMS advances from the first and second world wars were not readily replicated in a civilian setting until the 1950s when two civilian physicians, JD "Deke" Farrington and Sam Banks, developed a first-aid training program that incorporated many of those EMS advances. This program, which was developed for the Chicago Fire Department, is considered the prototype of the first basic emergency medical technician (EMT) program in the United States.

the late 1960s, most EMS systems were inadequate by modern standards (Rockwood, et al., 1976). Beyond the efforts of the Red Cross and the developments in motorized ambulance transport from World War I, the provision of EMS remained largely unchanged (Robbins, 2005). With respect to the lack of progress in the area of EMS transport, Briggs and Palmer (1963), Pozner et al. (2004), Rockwood et al., (1976), and Bass (2015) highlight the lack of the adequate vehicles and personnel for transporting the injured. Pozner et al. (2004), citing Blackwell (1993), notes how in the first half of the 20th century, the majority of vehicles used to transport patients to the hospital were hearses that belonged to local funeral homes. Similarly, to underscore the lack of progress in terms of providing adequate vehicles for emergency medical transport, Briggs and Palmer (1963) cited the survey results presented in Hampton (1960). This survey of EMS systems[9] in the United States, conducted in 1958, revealed that, among other things, only 54% of all the vehicles that used as an ambulance were adequate for transporting the injured.

With respect to the training of personnel, Rockwood et al. (1976) and Bass (2015) both discussed the lack of training among ambulance attendants during the era. Rockwood et al. (1976) noted that prior to the passage of the Highway Safety Act of 1966, only 46% of the estimated 200,000 ambulance and rescue personnel received training that was comparable to the advanced level training offered by the Red Cross and that often personnel had no training whatsoever. Likewise, Bass (2015) refers to the work of Barkley (1974) which claimed that in

[9] The survey only considered cities with a population over 10,000.

the post-war era, half of all ambulances were operated by mortuary attendants and that most of these attendants lacked basic first aid training.[10]

Many suggestions have been provided to explain the lack of progress in the development of EMS services. These include: the lack of innovation in the field of emergency medicine (Waller, 1965; Shah, 2006); the lack of knowledge about patients (Mitchell, 1965; Waller, 1965); financial issues related to the collection of payments from patients using EMS (Mitchell, 1965; Stevenson, 1971; Waller, 1965); the lack of adequately equipped ambulances (Briggs & Palmer, 1963); the lack of qualified EMS staff (Mitchell, 1965; Stevenson, 1971; Waller, 1965); the lack of adequate facilities to provide emergency treatment (Skudder & Wade, 1964); the disorganized nature of EMS operations (Stevenson, 1971); and the lack of regulations governing EMS operations (Briggs & Palmer, 1963). In addition, prior to the mid-1960s, the EMS system planning literature was limited in both quantity and in its scope (Stevenson, 1971; Waller, 1965). Regarding the lack of journal articles, Waller (1965) observed that the literature on medical care was expanding, but that it was "uniformly silent on the subject of ambulance services." Taubenhaus and Kirkpatrick (1967) echoed similar claims about a lack of studies about hospital ambulance services and added that when EMS articles did exist, they focused mostly on the issues of equipment and training of EMS personnel. Likewise, Stevenson (1971) noted that prior to 1966, the scope of the EMS literature was very narrow as it was primarily limited to private research conducted by concerned doctors

[10] Bass (2015) also provides a partial explanation for the lack of adequately trained ambulance personnel. According to Bass, when America entered World War II, the military's demand for physicians resulted in many ambulance interns being pulled from their positions. These interns did not return and thus, after the war, ambulance systems were left with poorly trained staff.

or focused on the operations of individual cities. One example of the latter is the work of Lehman and Hollingsworth (1960).[11]

Concerns about the state of the EMS literature would remain through at least the mid-1970s. Gibson (1974) reached the conclusion that "with few but notable exceptions, presently available EMS research papers are not in fact research products and do not satisfy even minimal standards for health services research."[12] Some specific aspects or types of EMS articles/presentations that Gibson found troublesome were "uncritical advocacy descriptions of some intended or completed EMS activity." In this respect, Gibson took issue with (1) the gross exaggeration of claims related to the lives that were lost because of inadequate EMS and the number of lives that could be saved with improved EMS and (2) the "reaction of isolationism [of EMS research] in response to a hostile or apathetic environment." Focusing on the latter issue, according to Gibson, one consequence of this overreaction[13] was the

[11] In Lehman and Hollingsworth (1960) analyzed the results of a nationwide survey of EMS systems in other US cities to compare the ambulance service of Seattle to that of other cities. The analysis was both brief and simple – it was predominantly a collection of descriptive statistics (about the other EMS systems) although it included an attempt to establish a relationship between the cost of ambulance service and other variables including: population, number of vehicles, calls per year, and type of service used. Alongside this analysis, various descriptions and an analysis about Seattle's EMS system were presented. This included statistics about the emergency calls that were made in Seattle, a rudimentary financial assessment of the cost of the ambulance service (in terms of costs per call), and a description of the Seattle's EMS system operations regulations.

[12] Within the publication, Gibson cites the generally "low technical quality" of the EMS papers presented at the 1974 American Public Health Association's (APHA) meeting in New Orleans as his motivation for writing the article. However, the greater importance of these comments (and why this work is notable within the context of the development of EMS in the United States) was the question of whether EMS research should be approached within the context of general health services research or as a separate research area. As noted by Gibson, the Emergency Medical Services Act of 1973 required the establishment of an EMS research program and the agency assigned the responsibility to manage the program, the Bureau of Health Services Research, had to decide which approach to adopt. As such, for Gibson the poor quality of the research presented at the 1974 APHA meeting and that of EMS research in general really indicated, among other things, the need to integrate EMS research with general health services research.

[13] To describe the nature of this overreaction we quote Gibson (1974a) directly:

"In EMS it is embodied as a general proposition, uncritically advanced as the revealed truth, that programming and research in EMS are qualitatively different from programming and research in other health

tendency for research questions to be developed based on pre-selecting a potential EMS intervention and then only considering the problems that such intervention can address. The drawback of this approach, Gibson added, was that "with this method of problem selection is that the problem selected is restated to fit the available solution" and that "[i]n addition, the available solution is not compared with other potential solutions (internal or external to EMS) in terms of relative effectiveness." Combining this with a tendency of EMS research to exaggerate results, Gibson argued that this resulted in a situation where for limited one-case studies, such studies reveal nothing about their generalizability of their results, or rather, the applicability of the EMS intervention being considered *in general*.[14]

Nonetheless, efforts to develop systematic guidelines for planning and managing EMS systems began to appear and develop in the 1960s. At first, journal articles contained general guidelines that consisted of rather simple and descriptive recommendations. However, as time progressed, EMS researchers would develop more specific and sophisticated guidelines.

2.2.1 Early EMS System Management and Planning Efforts

The first guidelines for planning or managing civilian EMS systems are have existed for almost as long as EMS systems have. For military systems, Larrey's memoirs (Larrey, 1814),

service sectors, that these (EMS) activities are based upon a unique set of knowledge and skills not otherwise available, and that categorically unique strategies are necessary for funding, manpower, research, health planning, etc. This proposition, if pragmatically argued, is not inherently unreasonable; indeed, it parallels the professionalizing strategies so successful in the emergence of the medical profession and its subspecialties. Within EMS, however, the difficulty is that this proposition is not pragmatically argued but ideologically asserted: it is used not so much to derive solutions to problems within EMS as to justify the existence of an EMS "social movement."

[14] Gibson also questioned the validity of the reports themselves.

as noted and cited by Robbins (2005), describe many of the principles that were developed and used to create one of the first, if not the very first, EMS *system.*[15] A little over a century later, Watt (1916) also published a set of guidelines concerning the organization of field ambulances.[16] For civilian EMS systems, Edward Dalton prepared a set of guidelines for the Bellevue Hospital (Haller, 1990). Dalton's rules addressed issues related to the governance of the ambulance services, ambulance dispatching policies, ambulance staffing requirements, the command structure of the ambulance crew, a protocol for hiring ambulances, the financing of the ambulance system, and the duties of the medical attendant (including when patients were to be treated) (Haller, 1990).[17]

Dalton's pragmatic and experienced-based approach for managing EMS systems represented the dominant approach for developing *general* EMS operational policies and rules until the late 1950s.[18] For instance, Benjamin Howard's evaluation of the New York ambulance system (Howard, 1881) describes much of the system's operations and focuses on the aspects of the ambulance system that seem to greatly support or improve the provision of EMS or the

[15] Larrey's system did not constitute the first attempt to provide prehospital services (for examples of the provision of some emergency medical services pre-dating Larrey's efforts see Robbins, 2005), however, his system was exceptional due to its comprehensive, planned, and integrated nature. To support this assertion we quote Robbins (2005):

"[Larrey] conceptualized and implemented a cogent, comprehensive pre-hospital care system that, for the first time, triaged the injured, provided immediate, temporary medical care and transported the injured from the battle field to strategically placed medical aid stations in a formal, regulated way using special apparatus."

[16] Given that a large portion of the plans of Larrey's, and later those of Watt (1916), are applicable primarily to military EMS, for the purpose of brevity we forgo outlining/describing their plans here.

[17] This summary of the guidelines presented in Haller (1990) concern guidelines appearing in Miles (1885) that were used to govern an ambulance service in New Orleans. As such, this summary is not immediately about Dalton's guidelines *per se*. However, according to Haller (1990), the differences between these Miles's (1885) and Dalton's guidelines "not measurably different."

[18] Granted, the number of publications that contain or develop guidelines between the 1860s and 1960s were very few.

system's efficiency. Similarly, the recommendations of Watt (1916) were based on his personal and his staff's experiences working in a military field ambulance unit.

Beginning in the mid-to-late 1950s however, EMS operational guidelines would evolve in three important ways. First, guidelines began addressing increasingly specific facets of EMS operations via thorough discussions, simple investigations, or detailed descriptions. For instance, Magelaner and McElroy (1955) discussed the role and impact of ambulance sirens, Curry (1956) discussed the issue of providing medical training to ambulance attendants, and Curry and Lyttle (1958) began an investigation regarding the impact of speeding ambulances. Likewise, early trauma care research was mostly descriptive although some insights and concerns arose from these investigations (Cales & Trunkey, 1985). Zollinger (1955) examined the quality of trauma care afforded to motor vehicle accident victims and suggested that "[t]he problem of trauma deserves the consideration commensurate with its frequency of occurrence". Root & Christensen (1957) examined traffic accident victims that received surgical care and suggested that quality of care may influence mortality. Similarly, Perry & McClellan (1964) studied traffic accident fatalities and suggested a relationship between patient mortality and the patient's arrival condition.

As for medical training for EMS, regular courses for both physicians and ambulance attendants began appearing in the 1950s (Hampton, 1972) while in the literature, several publications, like those of Carl Young (Young, 1954, 1958), began addressing the issue of providing first aid training to ambulance attendants and other emergency response professionals. Young's first book was written for a wide audience that included law enforcement officials and hospital workers while his second book 'Transportation of the Injured' (Young, 1958) focused more on training ambulance attendants and discussed other

aspects of ambulance operations such as the duties of ambulance dispatchers (i.e., beyond simply dispatching calls) and the proper use of sirens. With respect to proper ambulance equipment, both Young (1958) and Curry and Lyttle (1959) provided early, yet partial, lists of proper equipment for modern ambulances.

Second, as information about EMSSs became easier to collect, statistical descriptions of EMSS operations were increasingly being used to evaluate EMSSs and, in some respects, to develop EMS operational policies. In particular, EMS research began placing an emphasis on statistics about EMSS operational costs, the characteristics of EMS providers, EMSS resources (e.g. the number and type of ambulances), and the characteristics of demand (including potential demand). The value and need for statistics and analyses concerning EMSSs and operations was clear by the mid-1960s (Mitchell, 1965), however, according to Waller, et al. (1966), by early 1966, only a few papers had investigated patterns of ambulance care (i.e., ambulance service and patients) and the problems associated with EMS.

Initially, the use of statistics in EMS research was mainly in studies about the safe operation of ambulances and/or in surveys about EMS systems. One early EMS survey included a study commissioned by Kansas City, Missouri (Bureau of Municipal Research, 1955) that examined laws and ordinances regarding speed limits, siren use, and right-of-way privileges in 54 US cities in 29 states. Other early survey studies included the works of Krieger (1958) and Hampton (1960) that, respectively, examined the ownership and some operational characteristics of EMS systems throughout the United States. Lehman and Hollingsworth (1960) also presented results from a 1958 survey prepared by the Seattle-King County Health Department that was developed to allow the EMS service in Seattle to be compared with EMS service in other cities. With respect to the operation of ambulances, Magelaner and McElroy

(1955) studied the relationship between the use of ambulance sirens, ambulance right-of-way privileges, and ambulance accidents during various periods between 1949 and 1954.

Another area of EMS research where the use of statistics began to emerge was in analyzing the demand for ambulance service. Lehman and Hollingsworth (1960) presented numerous statistics pertaining to the volume of calls for service they received (as well as who received the call), the outcomes associated with each call for service, the frequency of calls for ambulance service in 3-hour intervals, the age of the patients that received ambulance service, and the cause/reason behind the call for service. It is also worth noting that the work of Lehman and Hollingsworth (1960) was in part motivated by the frequency at which ambulance service had to be provided to individuals injured in traffic accidents and the number of casualties that resulted from traffic accidents. In Seattle, Lehman and Hollingsworth (1960) reported that traffic accidents were the most common cause given for ambulance service and noted that a statistical analysis performed by Anderson (1957) revealed that traffic casualties comprised more than two-fifths of the total accidental deaths in the US - a far greater amount than any other type of accident. The study by Waller et al. (1966) focused on the demand for ambulance use in rural communities and reported ambulance use statistics similar to those presented by Lehman and Hollingsworth (1960). However, Waller et al. (1966) also included statistics concerning the type of ambulance services that were provided (e.g., simple transports, emergency transports, etc.), the fatality rates associated with different patient diagnoses, and the ambulance utilization rates of individuals with and without prepaid ambulance service.

With respect to how EMS system statistics were reported, most of the studies listed above simply described EMS systems although some when beyond a simple description. Some of this earlier work has served as a starting point for analyses presented in later publications. For

instance, Waller (1965) and Briggs and Palmer (1963) used the works of Krieger (1958) and Hampton (1960), respectively, to highlight the poor quality of ambulance transport throughout the United States. The more advanced work at this time began including quasi-experimental research or analyzed the implications of their statistical findings more thoroughly. Despite this progress however, most work of this era focused on only a few issues, namely the significance of faster ambulance response or total travel times and the economics of ambulance service.[19]

In the first category, Magelaner and McElroy (1955) examined the ratio of ambulance accidents to the number of emergency calls that were serviced under various conditions (i.e., periods when ambulances had different sets of rights-of-way and siren use privileges) and concluded that this ratio was most favorable when ambulances were denied the right-of-way. They also found that the recommended policy did not reduce the efficiency of the ambulance system. Curry and Lyttle (1958) examined 2,500 ambulance runs and estimated that speeding was unnecessary in 98.2% of the cases as there was only a single case where it was decided that faster travel-times would have made a difference. As such, Curry and Lyttle (1958) concluded that ambulances should operate within the speed limit and that they could use sirens. However, unlike Magelaner and McElroy (1955), Curry and Lyttle (1958) recommended that ambulances have the right-of-way. These conclusions about the relationship between the efficiency of an ambulance system and ambulance response or travel times, however, would be indirectly challenged by Waller et al. (1964). In their study that compared urban and rural fatalities resulting from traffic accidents, they observed that in comparison to their individuals that were injured in urban traffic accidents, individuals that died in rural accidents tended to

[19] Other topics included calls for increasing ambulance personnel training in order to better handle certain medical cases and complication (Waller et al., 1966; West et al., 1964).

die more frequently at the scene of the accident, sooner after injury, and from less serious injuries. After ruling out some extraneous factors (such as road or driving conditions, the role of urban and out-of-state drivers, and injuries) they hypothesized that longer response times and longer travel-times to medical care facilities contributed to higher traffic-related fatality rates (in comparison to urban rates).[20] Waller, et al. (1966) would later challenge the second hypothesis however, as their study did not find any substantial relationship between longer travel times and increased patient mortality except for patients with cardio-vascular-respiratory problems.[21]

In the second category, statistics were used in various topics related to the economics of ambulance systems. As previously mentioned, Lehman and Hollingsworth (1960) used statistics to evaluate their EMS system and one part of this analysis used survey data that they collected to investigate the relationship between ambulance service costs and various factors including population, number of ambulances, total yearly call volume, and the type of organization that managed the EMS system. They reported finding no significant correlations. Caldwell (1961) analyzed the payment of ambulance service bills and found a negative relationship between bill repayment and the distance that a patient was transported, as well as (separately), instances where accident care was provided (in comparison to non-accident care). Waller, et al. (1966) made similar observations with respect to rural EMS systems in the United

[20] To support this, Waller et al. (1964) noted, respectively, that rural traffic accidents occurred at times when they were less likely to be discovered (early hours in the morning) and that individuals in rural traffic accidents died from less severe injuries than those involved in urban accidents.

[21] Despite this conclusion, Waller, et al. (1966) expressed caution over its significance noting that several potentially important factors were unaccounted or not well understood. This included concerns over data sample sizes, the geography of roads, and the possibility that certain emergencies were dealt with and responded to in different ways (thereby making it difficult to evaluate the role of travel-times to medical care facilities).

States and concluded that the resulting financial uncertainty was likely to hinder the development of EMS systems in terms of producing and retaining qualified ambulance personnel. Moreover, some accidents involving non-residents of rural communities further complicated the financial health of rural EMS systems. Waller, et al. (1964) observed that a significant amount of service was provided to non-residents of rural communities. They did not consider the financial impact of this trend, however, as Waller, et al. (1966) argued that this represented a substantial economic burden on rural EMS systems considering that a larger proportion of ambulance calls in rural areas were accident cases, that accident cases required longer trips, and that these two factors were related to lower rates of repayment.

In all, during this time the view that statistics was a necessary tool to understand and improve EMS began taking prominence and eventually became accepted by the end of the 1960s. The beginning of this shift can be seen in the works of Lehman & Hollingsworth (1960), Howard (1965), and Waller (1965). Lehman & Hollingsworth (1960) noted that "[n]o statistically accurate or valid appraisal of traffic laws regulating emergency ambulance service is possible from an evaluation of only 30 local ordinances." In an study about emergency care and medical transportation in the Eastern United States, Howard (1965) pondered that "there might be some field of investigation that could dispel this fog of specialized subjective opinions by collecting statistical objective facts on the subject of emergency care." Likewise (although more assertively), Waller (1965) challenged "[t]he assumption that the usual procedures for providing emergency care in an accident or illness are known" adding that "[a]ctually, little has been documented about patient characteristics and who does what at the scene of an accident or *en route* to the hospital." Finally, King & Sox (1967) captures the transition to a complete acceptance of the necessity of EMS statistics in the introduction of

their analysis of the San Francisco EMS system: "[k]nowledge of the population, nature and distribution of emergencies, and geography and is a basic requirement for setting up an emergency medical system and can be used to evaluate existing or proposed systems and facilities. But there have been no such data with which to work. The San Francisco study was undertaken to accumulate samples of these data."

During this time-period EMS researchers suggested or developed increasingly more general and structured EMS operational guidelines. That is, these guidelines began including new issues that had been previously overlooked and they became more generic or nonrepresentational so that their applicability was broad and not limited to a single or limited number of EMS systems. Moreover, despite the broad and comprehensive nature of these guidelines they did not amount to an unstructured collection of related facts or suggestions. Rather, EMS researchers began deliberately developing these guidelines in a cohesive manner that acknowledged the relationships between different components of EMS operations. In other words, they began looking at EMS as a *system* rather than as collection of tasks and obligations.

Again, the notion of EMS as a system is indeed at least as old as the first modern ambulances (see Jean Dominique Larrey's "ambulance volantes"). Larrey did not just invent the first modern ambulance but also devised a system that jointly considered how and where to transport and treat injured soldiers (Bass, 2015). Nonetheless, it was not until the late 1950s that a significantly general and comprehensive set of EMS guidelines appeared. As previously discussed, by the mid- to late-1950s courses and textbooks regarding the proper *transportation* of injured people and the provision of *emergency* medical treatment were available yet little was said about EMS as a system.

In this investigation, the earliest work that was found that alludes to EMS as a system is an article by Curry & Lyttle (1959) where a model statute is proposed for the purpose of "[improving] the quality of transportation of the sick and injured." This article begins with a criticism of the state of EMS transportation services, specifically with the claim that patients were not being transported properly despite the availability of instructional materials regarding the proper transport of injured individuals. Then, it describes a potential solution involving the city of Flint, Michigan and its ordinance for regulating ambulance systems[22] as well as the successful "mutual cooperation between the morticians, independent ambulance companies, the city health officer and the local Committee on Trauma of the American College of Surgeons." Consequently, Curry & Lyttle (1959) proposed a model statue that addressed: (1) what qualified as ambulance services, (2) qualification and training requirements for ambulance attendants, (3) ambulance equipment, (4) regulating ambulance maintenance, (5) the proper operation of ambulances (e.g., with respect to traffic laws and patient welfare), and (6) punishments violating this statue. Their proposal contained complete legal statements (i.e., an operational ordinances) addressing each issue but notably, Curry & Lyttle (1959) included a concise discussion of most issues but also very specific guidelines regarding ambulance operator training and qualifications as well as ambulance inventory requirements (for ambulance companies in Flint, Michigan). Thus, although Curry & Lyttle (1959) don't use the word "system" in their article, they effectively discussed the management of an EMS system by deliberately detailing a sufficiently broad and cohesive set of EMS guidelines.

[22] Curry & Lyttle (1959) noted the existence of ambulance ordinances in Louisiana and Massachusetts but emphasized that the ordinances did not really regulate the transport of patients or set qualification requirements for ambulance operators.

After Curry & Lyttle (1959), a cluster of four journal articles appeared between 1963 and 1965 that suggested a move towards analyzing and managing EMS as a system. These publications were primarily motivated by automobile accidents and the needed response to them, but they all recommended a complete reevaluation and modernization of EMS operations.

First, Briggs & Palmer (1963), like Curry & Lyttle (1959), expressed concerns about the quality and the lack of regulation of emergency transportation. Specifically, they highlighted the dismal results of Hampton's (1960) survey of American EMS systems that reported that: (1) almost half of EMS transportation vehicles were inadequate for transporting injured individuals, (2) the uncertainty about the proportion of ambulance operators with some first aid training, and (3) the lack EMS regulations at the state and city level. In response, they outlined suggestions about the "basic elements of good service" This discussion included the nature of the EMS agency (public, private, volunteer, etc.), the dispatching system, equipment, and the training and selection of ambulance attendants. They also discussed the regulation, inspection, and licensure of EMS (and their staff) and encouraged collaboration among local organizations or agencies that were concerned with EMS.

Skudder & Wade's (1964) brief set of emergency transport guidelines also focus on having properly trained ambulance attendants. However, their work is notable in two respects. First, their focus on EMS is from the standpoint of the hospital emergency room. In their overview of emergency care, they recognized the changing nature of emergency services including the higher demand for hospital facilities (and inadequate space, equipment and staff to handle this change), the necessity to operate 24 hours a day, and the lack of standards and guidance concerning the provision of emergency care. Subsequently, they discussed several topics in

great detail including the organization and staffing of hospital emergency facilities (including ambulance staff and service), the planning or modernization of emergency medical facilities (including the construction of emergency department facilities), and properly equipping and supplying a hospital EMS department. Second, Skudder & Wade (1964) also argue, albeit briefly, that emergency departments should assume responsibility over the treatment of patients before they arrive at the hospital and their transportation and went as far as clamming that these tasks were "an integral part of [the patients] over-all management and may have a direct relationship to morbidity and mortality after admission." With this claim, Skudder & Wade (1964) effectively "elevated" the status of emergency transport from just a transportation service to a medical service.

To understand the significance of this statement, it is important to understand some of the changes occurring in emergency medicine between the 1950s and 1960s. Before 1960, many emergency rooms (ERs) were mostly "accident rooms" staffed with nurses or physicians (staffed on an as-needed basis or as part of a rotation) that provided basic care to patients (Merritt, 2014). The first full-time ER physicians were not hired until 1961 when the Alexandria Hospital in Virginia hired several physicians dedicated to running its ER. Then in the 1960s, ERs expanded their duties to treating accident victims and patients with urgent medical needs. At the same time, patients also began relying more on ERs and less on their general practitioners (GPs) for both urgent and non-urgent medical issues (previously patients relied on general practitioners and established close relationships with them). Merritt (2014) attributes this change to four factors: (1) an increasingly mobile population, (2) an increase in physician use of ERs, (3) the emergence of group practices, and (4) a shift towards medical specialization. Merritt (2014) also notes that many of the first ER doctors were GPs.

Considering that ER medicine was beginning to establish itself in the 1960s, it's no surprise that medical transportation was not equated with medical care. Despite the urging for increased training for ambulance attendants, Curry & Lyttle (1959) were primarily concerned with improper transportation further harming patients.[23] Briggs & Palmer (1963) went further suggesting that ambulance attendants should not be considered "laymen to be trained in the mere rudiments of first aid" and that they were "paramedical personnel with an important and often crucial role in patient care." They also suggested advanced training to handle a variety of situations besides vehicle accidents but their focus remained on medical qualifications and regulations rather than further integrating emergency transport and medical care.

Eventually, Waller (1965) directly addressed the issue, concluding that "[a]mbulance service frequently is the first phase of the medical care sequence and therefore must be considered as a bona fide area of medical care" and called for ambulance services to be considered in comprehensive medical care planning. Moreover, in Waller's (1965) overview of ambulance care he explicitly identifies "several procedures and systems" related to ambulance care indicating a view of EMS as a system. These procedures and systems included: (1) the ownership and organization of ambulance service providers, (2) ambulance personnel, (3) ambulance equipment, (4) the finance and economics of ambulance operations, (5) the characteristics of patients (or the lack of knowledge about them), and (6) the regulation of ambulances. The fourth point is rather notable as previous works *at most* noted this aspect of EMS operations. For this issue, Waller (1965) discussed various approaches for financing EMS

[23] This position in clearly stated in their text: "In many cases, poor transportation of the injured can do as much or more harm than the original accident. It can also influence the type and definitive management and the ultimate result of treatment of a specific injury. It, therefore, behooves the medical profession as well as the general public to insist that those engaged in transportation of the injured be properly qualified."

systems as well as the work by Caldwell (1961) which documented the economic struggles of many EMS agencies (notably those in rural areas) and the non-payment for ambulance services by some patients. Lastly, Waller (1965) also expressed serious concerns about the lack of knowledge about medical emergencies and the assumption that EMS operations in rural and urban areas should be managed similarly.

Mitchell (1965) presented an overview similar to Waller's (1965) in which he discussed: (1) the need to gather more information about the patients served by ambulances and the care being provided, (2) the different types of EMS organizations/providers, (3) evaluating ambulance staff and equipment, (4) the logistical problems faced by ambulance operations, (5) ambulance economics, and (6) the role of public health agencies. Mitchell (1965) notably highlighted situations where geography complicated or dictated response efforts such as, respectively, accidents in remote areas or extreme conditions (e.g., accidents in the desert) and the higher incidence of vehicle accident related deaths and incidents in rural areas than in urban areas.[24]

2.3 Developments in Emergency Medical Service Policy

As emergency medical care garnered an increasing amount of attention from medical professionals during the 1950s and 1960s, policy makers also became more interested in medical care (Shah, 2006). This included a great concern for the growing number of traffic accident fatalities, a problem that would attract the attention of several US presidents

[24] They referenced Waller et al.'s (1964) investigation of traffic fatalities for the latter issue.

(including Presidents Dwight E. Eisenhower, John F. Kennedy, and Lyndon B. Johnson) (Robbins, 2005).

President Eisenhower responded to this crisis by establishing the *President's Committee on Traffic Safety* through an executive order.[25] The order required the committee to synthesize and develop plans to reduce deaths and injuries involving motor vehicle; work with government agencies (at all levels) and interested national organizations to "study traffic-safety needs, adopt uniform traffic laws and ordinances, and conduct balanced traffic-safety programs". This order also called for the creation of advisory groups to "aid citizen leaders in developing effective support organizations, assist public officials in determining specific needs and applying remedial measures, plan and guide nationwide traffic safety educational efforts, and advance all areas of highway safety."

In 1960, President Kennedy confirmed the importance of this issue declaring that "[t]raffic accidents constitute one of the greatest, perhaps the greatest, of the nation's public health problems" (USDHEW, 1968). Despite President Kennedy's assassination, President Johnson maintained his predecessor's interest in traffic accidents (Shah, 2006) and in 1965, the *President's Commission of Highway Safety* (established in 1946) published a report, *Health, Medical Care and Transportation of the Injured* (President's Commission on Highway Safety, 1965). Here the Commission recommended the establishment of a national highway safety program to reduce death and injuries and also, suggested a need for the adequate and timely care of injured patients (Bass, 2015; Rockwood et al., 1976). In the following year, President Johnson discussed highway safety in his State of the Union speech (Shah, 2006).

[25] Executive Order No. 10858 – The President's Committee on Traffic Safety (January 13, 1960).

Besides traffic accidents, an interest in heart disease and strokes also fueled an interest in medical care that led to the advancement of EMSSs in the United States. This push was led by social and medical activist as well as President Johnson who announced his interest in heart disease, cancer, and strokes in a 1964 Health Message and later commissioned a report, *Report to the President: A Program to Conquer Heart Disease, Cancer, and Stroke* (President's Commission on Heart Disease, Cancer, and Stroke, 1965), that outlined a plan and several recommendations to advance the state of medical science and emergency services in the United States (Shah, 2006). Notably, the report contributed to the establishment of Regional Medical Programs (RMP) through the *Heart Disease, Cancer and Stroke Amendment*[26] (Sanazaro, 1967). The purpose of RMPs was "to encourage and assist in the establishment of regional cooperative arrangements among medical schools, research institutions, and hospitals for research and training, including continuing education, and for related demonstration of patient care". According to Shah (2006), RMPs were critical for the development of EMS in that they: (1) helped organize EMSSs and train EMTs, (2) served as a critical source of funding for EMSS, (3) impressed its medical priorities (heart disease, cancer and strokes) on EMS, and (4) promoted regionalized health care. Without RMPs, Shah (2006) argues that it was "unlikely that sufficient funds would have been available in an organized manner to advance EMS".

2.3.1 "Accidental Deaths: The Neglected Disease of Modern Society", National Highway Safety Act of 1966, and Heartmobiles

In 1966, the US National Academy of Science (NAS) and the National Research Council (NRC) marked the beginning of the modern era of pre-hospital care with the publication of

[26] Public Law 89-239.

Accidental Deaths: The Neglected Disease of Modern Society (NAS-NRC, 1966). To quote Bass (2015), this seminal report "documented the enormous failure of the United States to provide even minimal care for emergency patient." The problematic issues identified within ESSs in the United States included, among other things: (1) the lack of adequately trained personnel, (2) antiquated communications systems and equipment including a lack of emergency hotlines, (3) slow responses to medical emergencies, (4) the failure on the part of medical and health-oriented organizations to advance the treatment of trauma, (5) the condition of emergency departments in hospitals, (6) a lack emergency treatment protocols, (7) local political authorities failing to provide high quality emergency services, (8) lack of data regarding the impact of inadequate EMSs, (9) the lack of research about the potential of existing Federal programs to assist in the development of EMS; and (10) a lack of prehospital medical treatment (NAS-NRC, 1966). The report outlined a general plan to address these issues including specific recommendations such as improving ambulance communication systems and developing ambulance service standards at the state and local level, as well as, developing pilot programs to evaluate ambulance service in remote sparsely populated areas or in those areas that lack access to proper hospital facilities.

The NAS-NRC's recommendations were consistent with and complemented the *President's Commission of Highway Safety* report and both repots were subsequently used to develop the *National Highway Safety Act of 1966*[27] (Shah, 2006). This act of Congress established the cabinet level Department of Transportation (DOT) and provided the agency with it broad legislative and financial authority to improve EMS. The Act focused on highway

[27] US Public Law 89-564, 80 Stat. 731.

safety programs that included programs and standards for improving EMS planning, equipment, training, and staffing. Moreover, the Act allowed states to be punished for failing to fulfill mandates regarding EMS. Lastly, the Act established a crucial source of funding for EMS projects, studies, equipment, administrative, and personnel costs. In all, between 1968 and 1979 the DOT contributed $142 million to the development of regional ESSs with $10 million going to EMS research and $4.9 million going to EMS demonstration projects (Bass, 2015).

While the Act afforded the Federal government with considerable authority and resources, however, it assigned the tasks of developing EMSs to the states and regional agencies. For instance, the Act provided matching funds for EMS demonstrations and programs, and required states to develop highway safety programs that conformed to DOT regulations and adequate regional EMSSs (Bass, 2015). With this approach, the Act allowed different regions to experiment with different ESSs and policies and avoided expanding the federal government (Shah, 2006).

Besides policy, advances in medical care and technology also brought about changes to EMSs during the 1960s. This included advances related to pharmaceuticals, defibrillation, and trauma care and most notably, mobile cardiac care units that demonstrated immediate and quantifiable benefits (Bass, 2015; Shah, 2006). The latter came about with the work of Pantridge and Geddes (1967) in Belfast, Ireland on the effectiveness of intensive pre-hospital treatment for myocardinal infarction (heart attack) patients using intensive-care ambulances. In the United States, a similar physician-based “Heartmobile” program was established in Columbus, Ohio and the Seattle Fire Department also established “Medic 1” (Shah, 2006). In

all, the success of these and other programs increased the interest in highly advanced and responsive EMSSs.

2.3.2 Emergency Medical Services Act of 1973

Despite the federal government's financial and technical commitments into improving EMS, in 1972 a follow up report published by the NAS and NRC, *Roles and Resources of Federal Agencies in Support of Comprehensive Emergency Systems* (NAS-NRC, 1972), concluded that the federal the government failed to improve EMSs[28] (or match the efforts by other organizations to do so[29]) and that the lack of coordination and planning by federal agencies precluded the optimal use of federal resources.[30] The NAS-NRC report listed several recommendations in the report that urged the Executive branch to develop administrative policies and improve interdepartmental coordination for the implementation of EMS programs.

[28] "The Committee on Emergency Medical Services of the NAS-NRC found little evidence of concern for implementation of recommendations for upgrading emergency medical services by any agency within the Department of Health, Education, and Welfare above the level of the Division of Emergency Health Services. The Division of Medical Sciences of the NAS-NRC in its report, "Accidental Death and Disability: The Neglected Disease of Modern Society," of 1966, along with the American College of Surgeons and the American Academy of Orthopedic Surgeons, in the Airlie Conference report of 1969, recommended new initiatives in this field by the Executive Office of the President. The report of the Department of Health, Education, and Welfare Advisory Committee on Traffic Safety of 1968, under the chairmanship of Dr. Daniel P. Moynihan, recommended that the Department of HEW should assume primary responsibility to establish emergency medical services and consolidate the roles of agencies within the Department for this purpose" (NAS-NRC, 1972).

[29] "Federal agencies have not kept up pace with the efforts of professional and allied health organizations to upgrade emergency medical systems" (NAS-NRC, 1972).

[30] "In its analysis of the ways in which the resources of these agencies might be utilized, the NAS-NRC Committee on Emergency Medical Services finds that while most of the agencies have resources that could and should be used in development of a system of emergency medical services, the most efficient role that each agency may play in an overall program is reduced severely because there are no federal focal points of responsibility for delineation of the essential requirements for communication, transportation }or command and control, which are common to all emergencies, nor is there a federal focal point for overall planning, or for coordination of emergency medical services" (NAS-NRC, 1972).

In response to this report, President Richard Nixon voiced support for improving emergency care but opposed and fought against the passage the several EMS bills including one bill, the *EMS Systems Development Act*, which he vetoed (Shah, 2006). Many prominent national medical organizations and officials testified before Congress about the need for new EMS legislation to address the poor state of EMS in the United States (Shah, 2006),[31] however, opposition to such legislation was based on the idea that EMS was a local, not federal matter and on opposition to non-EMS related clauses such as the continuation of Public Health Service Hospitals (Bass, 2015; Shah, 2006).

Congressional leaders did not relent on reforming EMS and consequently, Congress held additional hearings over EMS which led to the introduction of a new bill in Congress that expanded the federal government's involvement in EMS (Bass, 2015). With this new bill, supporters emphasized the tremendous challenges individual communities faced in establishing regional EMSSs without substantial assistance from the federal government and the bill also discontinued the controversial Public Health Service hospitals (Bass, 2015; Shah, 2006). In November 1973, Congress easily passed this new bill and President Nixon signed into law the *Emergency Services Development Act (ESDA) of 1973*.[32]

The ESDA of 1973 provided wide financial support for developing comprehensive EMSSs throughout the country. It addressed EMSS development, research, and contract grants as well

[31] As noted in Shah (2006), Peter Safar, a key figure in emergency medicine, reiterated the findings of the 1972 NAS-NRC report testifying that the state of EMS as a *". . . disgrace, primarily because of lack of organization, coordination, and clearly defined responsibilities and authorities . . . ,"* and that *"Implementation of national recommendations concerning ambulance services' improvements are still being retarded because of incompetence, bigotry, indifference of the public and governments, and because the interest of providers rather than consumers prevail*." (United States Congress Senate Committee on Labor and Public Welfare, Subcommittee on Health, 1973).

[32] US Public Law 93-154.

as EMS training grants, respectively, through Title XII and a new section in Title VII of the *Public Health Service Act* (Bass, 2015). The grants covered feasibility studies and planning, initial operations, expansions and improvements, and research. The Act was amended and reauthorized to continue spending in 1976, 1978, and 1979 but the Act's underlying expectation was that EMSSs would become financially self-sufficient and not require further federal assistance past 1982 (Shah, 2006). As with the NHSA of 1966, Congress explicitly sought to avoid expanding the federal government. Lastly, the Act emphasized regional ESSs, addressing trauma, and outlined 15 "essential components" of EMSSs to be addressed, including: (1) personnel, (2) training, (3) communications, (4) transportation, (5) emergency facilities, (6) critical-care units, (7) public safety agencies, (8) consumer participation, (9) access to care, (10) patient transfer, (11) standardized patient record-keeping, (12) public education, (13) system review and evaluation, (14) disaster planning, and (15) mutual aid.

2.4 The Systems Approach for Planning and Managing Emergency Medical Services

Beginning in the late 1960s and early 1970s researchers began investigating other facets of EMS with a more systematic approach. Descriptive studies and surveys about EMS systems continued to be published during this time (*e.g.*, Holloway, 1972; West et al., 1972).[33] However, within the broader context of EMS research, the application of a "systems approach" to planning and managing EMS systems, both conceptually and in practice, gained prominence amongst EMS professionals (Boyd & Cowley, 1983). This became a central tenant in US EMS policy. Underpinning this transition was: (1) the notion that EMS was not just a transportation

[33] Examples of more modern surveys include the works of Pozner et al. (2004) and Williams (2007).

service, but also a *medical service*[34];(2) the challenges associated with *implementing* system improvement recommendations; and more generally, (3) a recognition of the need for a systematic way to *evaluate* proposed or existing EMS systems.

West et al. (1972) captured this further shift arguing that advancements in emergency medical care that could reduce patient morbidity should be introduced into the paramedical field. He concluded that ambulance services should no longer be considered a transportation service but rather "an essential component of the emergency medical care system" since "[m]ost of its recommendations were directed toward bringing ambulance service into the medical care field". Notably, two years earlier then California Governor Ronald Regan signed the Wedworth Townsend Act of 1970[35] which allowed paramedics to provide advance medical care under the supervision of a physician but without requiring the physician to be present to directly supervise the paramedic (Pozner et. al, 2004). Similarly, the American Society of Anesthesiologist's (ASA) Committee on Acute Medicine (Committee on Acute Medicine of the American Society of Anesthesiologists, 1968) called for further integration between ambulance services and emergency medical care given that advancements in emergency medical care could improve overall patient care.

EMS professionals and researchers observed that despite the existence of recommendations for improving emergency medical care that they were not being adequately implemented (as noted in the 1972 NAS-NRC report). Boyd & Cowley (1983) commented on *Accidental*

[34] In support of the idea that EMS was viewed primarily as a transportation service Shah (2006) notes, among other things, that the NHSA of 1966 placed EMS under the jurisdiction of the DOT rather than the Department of Health, Education, and Welfare (DHEW). Robbins (2005) also notes that the terms "*emergency medical services*" or "*EMS*" did not appear in the act itself but rather (and sparingly) terms such as "emergency services," "emergency service plans," and "transportation of the injured."

[35] California Health and Safety Code, Sections 1480–1485.

Deaths: The Neglected Disease of Modern Society (NAS-NRC, 1966) that although the report outlined "[t]he basic building blocks and blueprint for an improved trauma care program and most of the developments relevant to EMS and trauma care [at the time]," its major deficiency, in retrospect, was that it did not consider the "methods and approaches" necessary for implementing or effectively integrating the recommendations listed in the reports. Likewise, Hampton (1970) noted that the federal government, via the National Highway Safety Bureau, had already developed standards for the provision of emergency medical service through Standard No. 11[36] and went as far to say that "[*i*]*n urban areas particularly, those hospital emergency departments which cannot meet the standards for emergency departments of the American College of Surgeons or the Joint Commission on Accreditation of Hospitals should be closed as real emergency departments. They should not pretend to be capable of receiving and promptly treating the severely ill or injured. Such casualties should be resuscitated and transferred promptly to a fully equipped, staffed, and ready emergency department at a nearby hospital*."

The EMS community was also heavily critical about the state of EMSS evaluations. King (1968) considers the existing system quality performance measures as "relatively insensitive" in terms of "survival, complications, impairments, and disabilities" but also called for "[establishing] objectives based upon the widely held assumption that the shorter the time between the occurrence of the injury and the administration of an adequate level of medical care, the better will be the outcome for the patient." Likewise, Gibson (1973) was critical of the federal standards from Standard No. 11 given their almost exclusive concern with "in-put

[36] This standard was issued by the DOT secretary in accordance with the NHSA of 1966 (Gibson, 1973).

variables" rather than out-come measures of system performance. Likewise, he criticized data produced by state and local evaluations of EMSSs surveys claiming that "its relevance and usefulness is of dubious value, consisting as it does almost entirely of in-put variables" and also questioned the studies on various methodological grounds including ascertaining or verifying the veracity of data.

2.4.1 *Early EMS System Planning Models and Research*

Soon after EMS researchers first suggested discussing EMSs as systems, several publications appeared that completely embraced the concept. Among the earliest publications found in our literature review that discussed EMSs as such is an article by Richard F. Manegold and Michael Silver from the American Medical Association, *The Emergency Medical Care System* (Manegold& Silver, 1967). Here they presented their conception of an emergency medical care system replete with a schematic relating various EMS functions and factors. Moreover, they identified potential problems within EMSSs (such as delays in treatment) and the causes and impacts of these problems in relation to other system components and functions. Hampton (1970) and Nahum (1971) later authored similar articles about, respectively, a systematic approach to EMSs and emergency medical care systems. Nahum's (1971) article is notable in it outlines a "functional analysis" for EMSSs that relates an EMSS's components (e.g., personnel, equipment), its operations (i.e., notable tasks and events in an EMSS), and the system evaluation. Furthermore, he highlights the potentially complex relationship between these factors and that improving a system along one dimension might require intervention along an adjacent system component or operation.

Despite this paradigm shift, the EMSS planning literature remained reminiscent of earlier EMS planning publications in that they were extensive, yet cohesive discussions about EMS. They notably differed with their emphasis on systems and with increasingly elaborate discussions or detailed guidelines, however, this varied largely by topic with some receiving more attention than others. Examples of these works include: Sigmond (1967), which discussed areawide planning and how to reduce the volume of patients using emergency services and manage EMS-related costs; an extensive set of EMS standard goals by the ASA's Committee on Acute Medicine (ASA, 1968); Huntley's (1970) discussion of organizing community emergency medical care communities; an evaluation of the DOT EMS programs by Lewis (1972), and Hanlon's (1973) presentation on comprehensive emergency medical care systems.

It would be wrong to say that this "transition" period ended given that general system-oriented EMS overviews and surveys are continually published to report on the status of EMSSs in the US and from around the world (often in a highly accessible manner). Nonetheless, around 1973 there appears to be a significant uptick in conceptual and innovative EMSS planning articles. Examples of the first class of articles includes the works of Taubenhaus (1973), Sluyter (1976), Boyd (1976), and Boyd, et al. (1979) that present conceptual frameworks about, respectively, comprehensive EMSSs, EMS communication networks, national EMS systems and programs, and medical control and accountability. For the second class of articles, examples include Vogt's (1976) work with developing EMSS communication subsystems and the work of Boyd, et al. (1973) on the development of state-wide emergency care systems.

2.5 Location Science and EMS Systems

2.5.1 *Theory of Public Facilities*

In the late 1960s, Teitz (1968) proposed an important theoretical development relating to the location of urban public facilities (which can include facilities such as EMS centers and ambulances). Teitz called for the establishment of a theory of urban public facility location noting that location decisions relating to such facilities lacked a sound theoretical basis and instead relied on "mechanical and inadequate" responses based on rules of thumb. He continued that if the government wished to use public facilities to shape urban growth, and social and economic behavior, that such efforts would require the development of evaluation procedures for public services and that many existing quantitative approaches could be used to potentially improve how resources in urban services were utilized in terms of effectiveness and efficiency. For these reasons alone Teitz argued, developing a theoretical structure "might be invaluable."

Teitz (1968) then established some differences between private and public facility location theory, attempted to describe some functions that characterize public facilities as compared to private facilities, outlined a decision-making process as it relates to public facility placement, and lastly, provided a rough example of the application of his proposed theory in various situations. Most, if not all, of the ideas developed by Teitz are applicable (to different degrees) to EMS systems, however, a few points particularly stand out. The first was involved with the structure of public facilities. Teitz noted that the locational nature of a public facility system is strongly influenced by the geometry of facilities (point or network facilities), the services they provide (collection/centralized or distributed services), and how services are provided in terms

location (i.e., the number of public facilities that citizens can use or be served by). In the context of an EMS system, this is exemplified by several ambulances serving a given locale while a single hospital serves people from many locales. As for the importance of these relationships, Teitz noted that the characteristics of the facility system might require that a region be divided for functional reasons but also that it is important to consider how facility systems interact with boundaries established by other systems or organizations. At one extreme, such boundaries are disregarded (e.g., the relationship between public libraries and a city's neighborhood boundaries) and at the other end of the spectrum facilities must operate with strict consideration to such boundaries (e.g., post office delivery regions). Depending on the relationship between the public-sector services and the role of boundaries in the planning of public facilities, the overall effectiveness of a public facility system and/or the quality of the services provided by public facilities (or received by citizens) might be impacted dramatically.

The second key point noted by Teitz about the structure of public facilities is the hierarchy of facilities. According to Teitz, this quality is almost universal in public facility systems and that for point facilities, for instance, hierarchical structures (and their extent, structure, or degree of hierarchy) result from the functions that these facilities perform and the requirements that are necessary to support such facilities. In the context of EMS systems, examples include the high costs of operating advanced ambulance service that limit the number and use of advance ambulances for response.

Teitz also made several important observations related to the nature of public facility decision-making, where he outlined three general challenges that complicate the decision process. The first is that the government has a general resource availability that is established by society as a whole and that society is highly influential in determining how such resources

are allocated. Second, assuming that funds are allocated to "loosely defined" programs, Teitz noted that system location problems are placed under budget restrictions and as such, the budget dictates the provision of a given service rather than societal needs dictating the provision of that service, which Teitz argues, essentially amounts to an inefficient use of resources unless a socially optimal budget is somehow allocated.[37] The third challenge put forth by Teitz involved the general absence of a social welfare function. This poses a major challenge as the lack of quantifiable benefits complicates analyses with respect to the impact of decisions across various places and groups of people. Teitz noted that the extent of this problem varies – decisions that have a clear, positive, and sufficiently large impact are unlikely to be unchallenged as are decisions that have specific or well-defined targets. However, Teitz countered that decisions are highly likely to be challenged when they are made at the local level or in the absence of clearly quantifiable (monetary) impacts. Moreover, concerns about the distribution of impacts arise in both cases, but Teitz argued that when decisions are made at the local level, there are additional challenges stemming from local politics.

To address the problem quantifying benefits, Teitz (1968) argued that understanding the factors that influence a system's cost and efficiency might assist the decision-making process, including issues of scale and location (dispersion). Furthermore, Teitz proposed considering the possibility of formulating a system whose performance is readily measurable. The benefit of this approach, beyond providing a measure of performance, is that the impact of budget changes can be better understood including that of complex systems. In the case of a budget increase, performance measures should improve or at least remain constant and in the case of

[37] Clearly if a budget is below a socially optimal level, social returns could have been increased with a higher budget. If the budget was higher than the socially optimal amount however, Teitz notes that there is pressure to use the complete amount, which would result in inefficiencies in the system (see Parkinson, 1955).

a budget decrease, the decline in system performance (and its extent) can be determined if it occurs. In one of several examples, Teitz (1968) discusses fire stations and response time as a potential measure of system performance. More interestingly, he notes the role of standards and their potential to influence location decisions observing that insurance rates were mostly based on compliance with fire related standards rather than "empirical fire experience" such as response times to fires.

ReVelle et al. (1970) conceptually expanded on the ideas about system performance measurements offered by Teitz (1968) by proposing: 1) identifying and measuring factors that affect social costs; and 2) developing methods of analysis that employ surrogate or substitute measures for social costs. The first option was proposed as an analogue to approaches employed by firms in the private sector - quantifying their interests in terms of monetary value and then developing an objective function that maximizes monetary benefits so as to capture both monetary and non-monetary benefits. According to ReVelle, et al. (1970), efforts in adopting this approach found it difficult to implement and exhibited limited success. As for the second option, the purpose of surrogate measures, ReVelle, et al. (1970) admit, is not necessarily to find a solution to a problem as much as to gain a further understanding of the of the system under study. For potential surrogate measures for a public facility location model, they provide three examples based on: the total average distance traveled by facilities or users in a system (subject to a constraint on the number of facilities to be located); maximizing/minimizing the creation of demand (which is determined as a function of the number, location, and size of facilities); and the maximum distance or time between any facility and a service area/point.

Given such surrogate measures (or possibly others), ReVelle et al. (1970) proposed a framework whereby:

1) Facility location is optimized subject to constraints on investments.
2) The sensitivity of solutions are evaluated with respect to the parameter values assumed in the optimization model.
3) Tradeoffs between investment and utility are compared (when parameters are found to not significantly influence the solution)
4) To make a decision among the various alternatives generated include those with different levels of investment.

They noted that the nature of the surrogate measure be carefully considered, particularly with respect to the process of estimating demand and the length of the planning horizon. Failure to carefully consider both aspects when developing a model can result in solutions that involve sub-optimal locations in the present or near future. In the former, this can be the result of a biased surrogate measure(s) resulting from not correctly capturing the true level demand, while in the latter, this can result from a failure to consider potential changes in demand. ReVelle, et al. (1970) also asserted that the influence of public facilities on future growth should also be considered.

These theories of public facility location would later be expanded (*e.g.* Smolensky, et al., 1970; Austin, 1974; McAllister, 1976; Bigman & ReVelle, 1978; Greenhut & Mai, 1980) and later critiqued (*e.g.*, Dear, 1974; Morrill & Symons, 1977). However, with respect to EMSSs, the works of Teitz (1968) and ReVelle, et al. (1970) proved to be highly influential in the development of many ambulance location models or at least, they presented various elements of a modeling framework that would be applied in many ambulance location models. In

addition to the theoretical contribution from a Location Theory perspective, ambulance systems research produced many theoretical developments as well. The expansion of systems-based conceptualizations of EMS produced many of these advancements; however, numerous important developments also originated from a variety of quantitative EMS facility and system models. With the introduction of numerous mathematical tools and techniques to EMS research, as well as the increasing availability and processing capabilities of computers, ambulance system researchers were able to use models to observe and ask increasingly complicated questions about ambulance systems that were never before possible.

2.5.2 *Early EMS Facility & System Models: 1960s through the 1970s*

The use of mathematics and computers for the purposes of planning or analyzing ambulance systems can be traced to a series of reports, theses, and dissertations published in the late 1960s in both the United States and Europe. In the United States, two key early works include the reports of Dunlap and Associates (1968) and Gordon and Zelin (1968) as both reports developed modeling techniques and approaches that are at the core of various modern ambulance system models.[38]

The contributions of Dunlap and Associates (1968) included the development of methods for determining where to locate ambulances, estimating the demand for ambulance service

[38] Around this time, Hare & Wemple (1969) developed a report for the National Center for Urban and Industrial Health that presented a comprehensive simulation-based model to assist with the planning and development of community EMS systems. The model linked numerous aspects of an EMS system including the detection of emergencies, the process of dispatching, the emergency response, treatment, and the transportation of the patient. Historically, this work is notable as a review of the EMS modeling literature indicates that this work includes the first comprehensive EMS system model to be developed. However, the review also seems to indicate that the impact or further development of this model was extremely limited.

given a certain population size, and determining ambulance availability based on methods employing queue theory. The objective of the latter was to predict the availability of ambulance service as a function of the size of the ambulance fleet size and the demand for ambulance service. As for the work of Gordon and Zelin (1968), they took a different approach for analyzing emergency ambulance systems. They developed a computer-based simulation to study the value of satellite ambulance garages. The motivation here was determining whether a decentralized ambulance system could outperform a centralized ambulance system (that is, a system where ambulances are located at a single central location) in terms of response time, round-trip time, and ambulance utilization.

The long-term impact of these two reports was that their developments and results would end up in two influential journal publications. Dunlap and Associate's work on using queueing theory to determine the number of ambulances needed to provide a certain level of service would be published in Bell and Allen (1969) while Savas (1969) would expand on the work of Gordon and Zelin (1968) and become the first journal article to present a model for analyzing emergency ambulance systems.[39]

Outside of the United States, researchers in Great Britain were also active in the development of the ambulance system models during this period (Gibson, 1973). A model for determining the minimum number of ambulances required to maintain a certain level of service (although for non-emergent cases) was developed by Black (1969) while Dale (1969) considered emergency cases and applied queuing theory in order to determine the appropriate ambulance fleet size. Davidson (1969) synthesized and expanded on the works of both Black

[39] The work of Gordon and Zelin (1968) was published in the Transactions of the New York Academy of Sciences as Gordon & Zelin (1970).

(1969) and Dale (1969) using Markov chains. Other notable research was conducted by the Greater Council of London's Research and Intelligence Unit (July 1967-Janurary 1969) and the Shields (November 1969) both for the London Ambulance Service. These works were similar to those in the Dunlap and Associates (1968) and Gordon and Zelin (1968) reports as they included studies involving the use of simulation, determining optimal fleet size and ambulance location, and predicting demand for ambulance service. Foster (November 1969) would also investigate ambulance demand, optimal fleet size, and changes to the location of ambulance stations although in relation to the development of a new motorway.

After the publications of Savas (1969) and Bell & Allen (1969), the amount of interest and publications in the area of ambulance system modeling expanded dramatically. Within the next decade alone, many articles published within this period would not only significantly expand and develop the core methods and models used in ambulance system modeling, but they would also help transform the practice of analyzing and planning EMS systems from an obtuse, unstructured, and idiosyncratic process towards analyses that were more systematic in nature.

Volz (1971) examined the two versions of the ambulance location problem in the context of a semi-rural area. The first problem considered the location of ambulances that minimized average response time as a function of the number of ambulances that were available. Ambulances were allowed to reposition themselves upon any ambulance becoming busy or available. The second problem was similar to the first except that it required that the average response time to any location served by the ambulance system not exceed some response standard. Such a constraint however would only be in effect when a sufficient number of ambulances were available.

Hall (1971) developed an ambulance location model for a 'dual function' police-ambulance system where select police vehicles would respond to both medical emergency and police calls. In this model, different combinations of ambulance allocation and police call dispatching policies were analyzed in terms of: 1) the probability that at least one ambulance was available in the system, and 2) the proportion of calls that were served by an ambulance located less than a mile from an emergency. The analysis was based on using Markov chains whereby the status and location of each ambulance characterized the system into a set of states. Then, a numerical analysis was used to determine the probability of the system being in any state. A mathematical analysis of this model was presented in Hall (1972).

Fitzsimmons (1971) presented an EMS ambulance system simulation model to aid planners in evaluating existing or proposed EMS systems. Motivating Fitzsimmons's selection of a simulation approach to model EMS systems was the methodological shortcomings of EMS systems being conceptualized as single queue, multi-server models. In particular, Fitzsimmons questioned the typical assumptions in such models about service times noting that service times were dependent on the time of day and that they also were not equal for each ambulance (unless they were located in the same station). Given the limitations of queue based analytical models and a desire to capture the complex nature of EMS systems, Fitzsimmons considered simulation to be the most appropriate tool for modeling EMS systems. The simulation developed in Fitzsimmons (1971) is based on two programs, one to generate incidents and their characteristics (*e.g.*, each incident's location and the type of injury associated with each incident) and the other to simulate the ambulance response process (i.e., the typical sequence of events beginning with EMS system operators receiving a request for service and ending with the ambulance's return to its station). This simulation model was verified and validated

using an approach developed by Naylor & Finger (1967)[40] and with data from various Fire Department Ambulance Companies located in the San Fernando Valley. The study collected information generated by the simulation model about all individual incidents (response time, waiting time, time to hospital) and analyzed such data at the system level. Ambulance system operation and performance statistics were also calculated concerning EMS demand (e.g., call volume and statistics about where these calls originated), ambulance system busyness (e.g., call volume, mean utilization), and ambulance system performance (e.g. response time, mean wait time, time to hospital).

Chaiken (1971) considered the problem of calculating the expected travel times and workloads of emergency response units assigned to defined response areas. The motivation for this work was the problem posed by an imbalance in workloads among firehouses in New York City. Firehouses in some parts of the city responded to a high number of fire alarms which left firefighters working at these stations feeling overworked, while other fire stations in the city responded to far fewer alarms, including some located not too distant from the busy fire stations. One possible solution was to contract the areas that for which busy stations were responsible for (response areas) while expanding those areas of stations that were less busy so as to distribute workload more evenly among all stations. It was noted however that altering

[40] Fitzsimmons (1971) described his entire model assessment process as a "model validation" procedure, however, considering concepts and terminology developed in the simulation literature, this appears to be a misnomer. Based on the terminology presented in Schlesinger et al. (1979), the first step in Naylor & Finger's (1967) "Multi-stage verification" approach coincides with "model verification" as it concerns considering or developing some conceptual model of EMS system and then assessing the EMS system model with respect to the conceptual model. The "Multi-stage verification" process's second and third steps however, are arguably more akin to "model validation" given that they are concerned with the EMS system model's consistency with respect to the intended application of the model. As such the "model validation" process in Fitzsimmons (1971) is arguably a combination of both model verification and model validation procedures. Lastly, the goals and methodology of Fitzsimmons (1971) indicate that the model validation procedure is designed with the intention to establish "model credibility," or the "[concern] with developing in (potential) users the confidence they require in order to use a model and in the information derived from that model" (Sargent, 2005).

response areas of a unit could affect overall response travel-times. As such, Chaiken (1971) wished to calculate the expected travel-times and workloads of units as a function of its response area. To model the emergency response system, and determine these measurements, Chaiken (1971) employed a queue-theory based model to determine the steady-state probabilities of a two ambulance system whereby each ambulance is either busy or available. Then, Chaiken (1971) outlined a procedure to use these probabilities in order to calculate both the workload of each unit and the average response travel-time to each region. Finally, Chaiken (1971) also presented a linear programming model, developed by Edward Ignall, for minimizing the expected generalized travel-time in the special case where demand for service is concentrated at a finite set of points.

Stevenson (1971) presented a very thorough report that discussed the state of EMS systems in the United States, provided a general framework for analyzing EMS systems, and developed a model to evaluate the performance of an EMS system. The model begins with two sub-models that approximate the dispatching delay of an EMS system[41] as a function of the number of the number of ambulances in the system and also, the delays that result from an ambulance's travel from origin location (or station) to the location of the patient. Both models are then combined to develop a facility location model to optimize ambulance location configurations with respect to minimizing response time. The location model is solved with a heuristic based on dynamic programming. Lastly, Stevenson (1971) developed an additional model to determine the minimal number of ambulances that are required to meet a pre-specified level of service in terms of the immediate availability of an ambulance.

[41] This involves determining the probability that a patient experiences a delay in response and the expected length of the delay.

Toregas and ReVelle (1972) expands on the public facility location model developed in Toregas *et al.* (1971) by applying it to emergency services such as fire response and EMS. In Toregas & ReVelle (1972) the location model involves the problem of locating the minimum number of emergency facilities/servers such that the located facilities can cover all demands for service within a time or distance constraint. The problem was formulated within a linear programming framework and solved using a combination of integer linear programming and optimal reduction rules. Within the context of EMS operations, Toregas and ReVelle (1972) is notable because, among other things, it seeks to address the concerns/suggestion of Huntley (1970) regarding providing emergency response within an acceptable amount of time. Church and ReVelle (1974) extended the model of optimizing coverage with respect to a time or distance constraint although rather than trying to establish the minimum number of facilities required to cover all demand, the model in Church and ReVelle (1974) considered the problem of maximizing the amount of demand that could be covered within a time or distance standard with a fixed number of facilities or servers. Like in Toregas and ReVelle (1972), the problem was formulated within a integer linear programming framework, however, solutions were generated by both a heuristic procedure and by using linear programming (in conjunction with a branch-and-bound procedure).

Carter *et al*. (1972) expanded on the work of Chaiken (1971) with respect to establishing response areas that minimized the average response-times although with a slightly different focus. Here, an emphasis was placed on determining shape of the response areas and the objective functions that correspond with response areas that minimize average response-time or that balance workloads. Two important findings are: (1) all 'good' (or undominated) response area candidates 'lie' in between the 'minimum-response-time' response area and the

'equal-alarm-rate' response area, and (2) if alarm rates vary significantly over small distances, a 'closest-unit' division approach to districting response areas does not necessarily produce 'good' candidates.

Keeney (1972) considered a procedure for determining the district boundaries for a naive response area-districting approach. This approach attempts to divide an area (*A*) into *n* areas such that each area is assigned to one of *n* facilities located at fixed points within *A*. To divide *A*, locations are assigned to the nearest facility. The cases that are considered are when a facility is added to the system and when a facility is removed from the system.

Larson and Stevenson (1972) developed a series of analytical models to examine the nature of mean travel times of vehicle responses. This investigation was based on an *area-districting* approach whereby vehicles are positioned at a facility in a district and assigned to respond to calls for service that originate from locations within their districts. Using this framework, two types of models were developed for the analysis of mean travel times of vehicle response whereby vehicles either exclusively serve the district they are assigned to (that is all vehicles operate independently of vehicles outside their assigned district) or are allowed to "cooperate" with other districts by serving some calls for service that originate from outside their assigned district. The former type of model involved a system where multiple vehicles can be located in a district and where no inter-district cooperation is allowed while the latter model type, considered a system where a single vehicle is located in its own district but can respond to calls originating from adjacently located districts. Larson and Stevenson (1972) first analyzed the upper and lower bounds of mean travel time in a system with no inter-district cooperation, when N facilities are located throughout the region. Assumptions about the geography of the region included the use of the Manhattan distance metric and that the demand in the region

was spatially homogeneous. They expanded on this analysis considering the case of an arbitrary distribution of demand, when only two facilities. Larson and Stevenson (1972) also examined a system with inter-district cooperation. This model was based upon dividing a region into two districts, with each district served by a single vehicle, that (they admitted) was "effectively" equivalent to the model developed by Carter *et al.* (1972). However, Larson and Stevenson (1972) extended that work by fixing one of the two vehicles and repositioning the other. When the vehicle is relocated, new district boundary lines are established and the system's mean travel time is recalculated. This process continues until the system's mean travel time is minimized. The procedures used to find district boundaries and the vehicle locations are based on gradient-search.

The ambulance location model proposed in Fitzsimmons (1973), referred to as CALL (Computerized Ambulance Location Logic), combined a stochastic analytical model with a pattern search routine developed by Hooke & Jeeves (1961). The latter routine is used to determine the ambulance locations that would minimize the mean response time for the system - the model's main objective. The CALL model is used by Fitzsimmons (1973) to address the challenges associated with accurately modeling the process of assigning an ambulance to call for service. This is a crucial consideration in EMS systems that experience congestion. In congested systems a specific ambulance may be unavailable (due to having that ambulance respond to or service a different incident) which might require dispatching a more distant ambulance to serve an emergency. Fitzsimmons considered it essential to accurately estimate the probability that a particular number of ambulances are available because mean response time calculations were based on the number of available ambulances in the system. Fitzsimmons (1973) used both a queuing model (based on an $M/G/K$ queue) and a Monte Carlo

simulation procedure to estimate, for each given ambulance location configuration: (1) the mean response time for each possible system state (with each state corresponding to a unique total number of busy ambulances), and (2) the system state probabilities. A $M/G/\infty$ queue based model is used to approximate ambulance availability when the system has 0 or 1 busy ambulances while the Monte Carlo simulation approach is used when 2 or more ambulances are busy. After calculating such quantities, the unconditioned mean service time is estimated iteratively until the difference between the two sequential estimates converge. Then, the pattern search routine is used to nominate a new locational configuration and the process is repeated until a better performing ambulance location configuration cannot be found.

Swoveland *et al*. (1973b) developed a probabilistic ambulance location model that used an enumerative solution procedure (branch-and-bound). The main consideration of the model was to locate ambulance depots so as to minimize the ambulance system's mean response time, where response time is defined as the time between when a call for ambulance service is made and the arrival time of an ambulance at the scene of the accident. This objective is captured in the form of an analytic formula that considers the locations of the k-closest ambulances and the proportion of the total number of calls that are served by the k^{th} closest ambulance at each demand point. Most notably, in this paper, a method is developed to approximate the latter in response to the observation that requests for ambulance service are not always fulfilled by the closest ambulance. The method is based on sampling the results of various ambulance response simulations whereby each simulation instance involved a different ambulance location configuration. To support this approach, a "*stability hypotheses*" conjecture was developed. This basically assumed that, for each demand point, the estimated proportion of the total number of calls that are served by the k^{th} closest ambulance would not differ significantly from

the proportions produced by any other assignment. The details concerning the ambulance system simulation are discussed in Swoveland *et al.* (1973a).

Among the EMS system models presented in this section the one developed by Hamilton (1974) is rather unique. Although it contains a transportation component, transportation was largely an external factor/consideration as the focus of the model was primarily on the potential impact of phasing out certain hospital emergency rooms in terms of how emergency system workloads (e.g., emergency room visits, hospital admissions, inpatient load) would be redistributed amongst the surviving hospitals. The model is based on a simulation that sequentially: generates an emergency occurrence, assigns this emergency to a geographical area, establishes the severity of the emergency, determines the mode of transport for the patient, directs the patient to a specified emergency room, generates a travel time for the trip, computes the arrival time at the hospital, and the patient disposition. The nature of the assignments or decisions at each step are mostly based on historical data or on the nature of an assignment made at a previous state (e.g., the location to which an emergency occurrence is assigned is based on historical data while the patient disposition is based on the severity of the emergency). With respect to the transportation components of the model, travel mode and transit times were mostly exogenous within the simulation as these assignments would be based solely on historical data. Moreover, they were mostly ignored within the development of the projected simulation outcomes of the various proposed scenarios. A potential alternative mode of transportation for serious emergency cases was briefly discussed as a possibility but not seriously considered beyond a remark that the travel times associated with such alternatives would be "within acceptable limits (as defined by physician consultants to the Task Force)." In contrast, one facet of transportation that was highly considered was emergency room

assignments for medical emergencies. In determining the patients' destination, the location of the emergency and travel times were considered but in evaluating simulation outcomes, the relationship between workload distributions and emergency room assignments were carefully considered. Despite the minor role of transportation, the author recognized that in later applications of the proposed model, there should be a greater focus on the EMS transportation system. Some suggestions included explicitly accounting for the size, type, location, and schedule of emergency vehicles in order to ensure an adequate level of service.

In Berlin and Liebman (1974), the ambulance location model developed by Toregas and ReVelle (1972) is combined with an ambulance system simulation model to produce a two-stage ambulance location-allocation model. Within this two-stage model, the model proposed in Toregas and ReVelle (1972) helps address the question of where to establish ambulance depots (which included the task of generating a set of alternative ambulance location configurations) and then the simulation model helps determine the utilization rate of ambulances located at each depot. Motivating the development of this model was the inability of locational models alone (such as that of Toregas and ReVelle, 1972) to consider or describe the impacts of system congestion. In particular, Berlin and Liebman (1974) noted that due to system congestion, the closest ambulance might not respond to an emergency and that response from a more distant ambulance might be necessary. As such, within the modeling framework of Toregas & ReVelle (1972), this would prove problematic if the response time exceeded the maximum response time standard used in the model. Nonetheless, Berlin and Liebman (1974) also noted that static optimization location models were especially suited for systematically determining optimal location configurations. Hence, by combining both models, their two-

stage model was able to generate relatively effective potential solutions and to describe the performance of the system in a more accurate and detailed way.

To assist urban emergency service system administrators in evaluating the performance of emergency response systems, Larson (1973, 1974) developed the "hypercube queuing location model" that attempted to address many of the perceived shortcomings of existing emergency service location and/or districting models.[42] These deficiencies included a lack of consideration for 1) interdisctrict response and the issues associated with it (or resulting from its absence), 2) estimating various system performance measures beyond just mean region-wide response times or other closely related measures, and 3) accounting for the probabilistic nature of EMS systems, namely the stochastic nature of the arrival of calls and the variability in service times. The model is based on the generation of a state transition matrix associated with a finite-state continuous-time Markov process. In this model, the status of each ambulance is tracked and the two possible states, the server is either idle or busy, is represented, respectively, with a 0 and 1. Then, each state in the state transition matrix corresponds to a unique combination of the status of every ambulance in the system, hence the name "hypercube."[43] Server locations (for N servers) are fixed in this model (servers cannot be co-located), and it is assumed that at any moment at most one ambulance can change its state (in either direction). In addition, it is assumed that service times have a negative exponential distribution, are not dependent upon

[42] Here, Larson defines *location problems* as problems closely related to the question of "how should the N response units be located or positioned while not responding to calls for service?" In contrast, Larson defines *districting problems* as those problems closely related to the question of "How should the region be partitioned into areas of primary responsibility (districts) so as to best achieve some level or combination of levels of service?".

[43] Since all ambulances are either idle or busy, in a system with N ambulances, the total number of state spaces is 2^N. To conceptualize the total state-space, each state (a sequence of 0s and 1s) is thought of a vertex in an N-dimensional cube (hence the inclusion of hypercube in the model's name).

location, and that travel times constitute only a small portion of total service time. With respect to demand, the study region is partitioned into individual "atoms of demand" that are each associated with a call arrival rate (which is assumed to have a Poisson distribution). In addition, each demand atom is associated with an immutable ordered server priority list that specifies which ambulance is to respond given the state of the system, that is, if a response unit is requested, the most preferred unit that is available is dispatched to respond.[44] With this, the steady state probabilities of the state transition matrix are calculated by solving a set of 2^N balance equations[45] from which it is possible to calculate performance measures at various levels such as the mean travel time, the workload imbalance, and the fraction of inter-district responses at the regional level; fraction of time spent serving calls (workload), mean travel time, and the fraction of inter-district responses for each response unit; the fraction of responses into each district that are inter-district at the district level; the mean travel time; and the fraction of calls handled by each response unit at the demand atom level.

One significant advantage of the hypercube model proposed by Larson (1973, 1974) is that it does not require an assumption of server independence as all inter-server interactions are captured in the model as each server status is fully tracked. Such tracking however, is computationally expensive as the amount of information that must be maintained grows

[44] In the case that all units are busy, the model can be set up so that that calls are handled by an auxiliary response unit (the system is treated as one with zero capacity) or so that the response is delayed until a response unit becomes available.

[45] Briefly, these equations require that for any state *i*, the sum of the transition rates from all states (except for state *i*) into state *i* is equal to the sum of the transition rates from state *i* to all other states (except for state *i*). The transition rate between two states, say state *i* and state *j*, is strictly positive only, but not necessarily, when such transition is possible, that is, when the system is in state *i*, there is a strictly positive probability that the system changes from state *i* to state *j*. These equations can only be solved if the Markov chain has an equilibrium distribution.

exponentially with the number of servers as the number of balance equations that need to be solved is equal to the total number of state spaces (recall that with N servers the total number of state spaces is 2^N).[46] Consequently, the hypercube queuing model can only consider systems with few servers[47] before the problem becomes computationally intractable. In response to this issue, Larson (1975) developed an approximate version of the hypercube queuing model that reduces the number of balance equations that are needed to be solved from 2^N to N in a system with N servers by not explicitly tracking the status of each server but rather estimating the probability that a given server will respond to a call for service. This probability is estimated by assuming that the probability that a server *j* will respond to a call, P_j, is the product of the probability that server *j* is available and the product that includes the probability that each server preferred to server *j* is busy. Moreover, Larson (1975) completes this calculation by multiplying it with a correction factor, Q, in order to relax the server independence assumption when calculating P_j. Q is a function of a series of queuing factors (called "Q-factors") that are used with an *M*/*M*/*n* queuing model to derive the value of Q. The Q-factors used in Larson (1975) include the number of servers in the system, the response server priority lists, and the system utilization.

Groom (1977) developed a coverage-based stochastic ambulance location model to evaluate the performance of an ambulance system under various scenarios. The prime consideration in the model is service coverage, which is based on the expected proportion of

[46] In turn, an N server system requires a state transition matrix with 2^{2N} elements

[47] At the time, Larson reported it was computationally feasible to model systems with up to 12 units although attempts were made to model up to 15 units. Goldberg (2004) reported computational tractability issues with 20 units.

calls that have a response time below a time standard t[48], although equity of service was also considered in evaluating system performance. To measure coverage, two factors, *range* and *availability*, were considered whereby *range* corresponded to the proportion of emergencies that can be responded to by an ambulance within time t given that r ambulances are available and *availability* corresponded to the proportion of time that r ambulances are available to respond to emergencies. Then, the *range* when r ambulances were available was calculated by summing the proportion of emergencies that was accessible by at least one of r vehicles within time t while *availability* was calculated using queue theory based formulas to determine the probability that r ambulances were available to respond. Two separate scenarios were considered in calculating *availability*, a *single-tier model* and a *double tier operation model*. In the *single-tier model*, ambulances were assigned to respond to emergency calls or to complete non-emergency tasks. An ambulance's task assignment was allowed to vary and was based on the level of standby vehicles available to respond to emergency calls. Also, within this scenario, ambulances were relocated upon the dispatch of an ambulance to an emergency or as an ambulance became available for responding to an emergency (a process assumed to occur instantaneously). In contrast, within the *double tier operation model* ambulances were only assigned to respond to emergency calls. Moreover, no ambulance relocations occurred. Finally, with respect to the equity of service, the level of service provided to each of the various health districts, or sub-regions, in the study area was assessed to ensure that there were no significant disparities in the provision of service.

[48] The response time standards considered by Groom (1977) include a response time of 8 minutes or less for 50% of calls and 20 minutes or less for 95% of calls (with standards of 7 and 15 minutes, respectively, for metropolitan areas).

The model proposed in Achabal (1978) concerns the location of EMS facilities within a multicounty EMS system. Forming the basis of this model is a location-allocation model that is formulated as a mathematical program. The model's objective function is based on minimizing the total costs of the EMS system subject to a number of service constraints. In this model, two types of costs are considered, spatial costs and resource costs. The latter is concerned with the costs associated with providing emergency medical services, but only those that are due to different regionalization plans.[49] Spatial costs are a function of both the *direct costs* of transporting patients, the cost of operating an ambulance on a per-mile basis (using estimates developed by Gibson, 1971), and *indirect costs* of transporting patients, the cost on a per-mile basis associated with the increasing probability of death as it relates to increasing the distance a patient has to be transported to an EMS center. To determine value of the indirect spatial costs, Achabal (1978) relied on the work of Achabal (1975). Here, the implicit social costs from increased travel times were based on a Bernoullian monetary function (Bernoulli, 1954) and the present value of an individual's limetime earnings (a figure that was obtained from Rice, 1966). Then, in consulation with data provided by physicians, Achabal (1975) derived a probability of death function that depended on a patient's travel-time to the EMS center. After accounting for travel time and speed, and substituting the relevant figures into the utility function an estimate for *indirect spatial costs* was derived. In all, the model's objective included both spatial and resource costs and included constraints that requried that the capacity of the system exceed the demand, that all counties were assigned a single facility (from a

[49] Spatial inelastic demand is assumed in this model and as such, the costs associated with treating patients are considered to be constant. In addition, the costs of operating and staffing an ambulance service are not considered as ambulance service operations are assumed to have no influence on the decision of where to locate regional emergency medical service facilities.

selection of different sizes), and that the level-of-service provided to each county operate at or above a service standard based on the minimum proability of survival. The level-of-service parameter used in the latter constraint was selected abritrarily by Achabal (1978) noting that this value was a policy decision while the proability of survival associated with travel between the location of demand and the EMS center was determined with a function developed by Achabal (1978) in consultation with physicians.

Meredith & Shershin (1978) adapted a model developed by the U.S. National Bureau of Standards (Colner, 1973) [50] for determining the optimal locations for fire stations. For establishing the optimal placement of facilities, the National Bureau of Standards (NBS) model used an "exposure index" calculated for each zone in a region that is a function of the response time, desired response time, and alarm frequency (call arrival rate) that was known or determined that corresponds to each zone. Mathematically, the objective value (the Total regional exposure) equaled the sum, over all zones, of the "exposure index" for each zone times the alarm frequency of each zone. Behind this approach was a philosophy that stations should be located such that the total county-wide "delay" in response time[51] was minimized (although desired response times could be normalized by zone to coincide with the priorities of decision makers). This measure was deemed superior to other measures or objectives such as: minimizing average response time, minimizing "delay", balanced workloads among

[50] The reason for this model's adoption was that in 1973, the Dade County Fire Department in Florida was assigned administrative duties for EMS in the Dade County area. Lacking both guidance for managing the system and a dedicated information system for EMS, the department decided to use the existing Fire Departments information system (this was also partly because the system had many attractive data-processing, modeling, and reporting capabilities).

[51] The "delay" in a zone is equal the difference between response time for that zone and desired response time for that zone divided by the desired response time for that zone.

stations, and equal average response times (or "delay") although it is not explained how. One acknowledged disadvantage of the regional "exposure" index approach is that (for management purposes) the measure lacks a physical interpretation. The model's solution procedure relies on an iterative heuristic that locates facilities and allocates zones to the nearest located facilities thereby forming a partition for each facility. Then, facilities are moved to different sites and if the new locations improve the overall total regional exposure, the relocations are made. Otherwise, other relocations were proposed. This process continues until no proposed relocations improve the overall total regional exposure.

Two additional models worth noting due to their novel approach for determining where to position ambulances are the works of Schneider (1971) and Schneider & Symons (1971). Both approached the problem by having people use an interactive computer program (viewed on a CRT monitor) that allowed such participants to locate ambulance dispatch centers from which response districts were created. A network representation was used in these programs - all edges were associated with a travel-time while the set of potential ambulance locations consisted of nodes on a network. Moreover, districts were created by automatically assigning all points to the closest located facility (the modeling framework did not consider/allow for the possibility of ambulances assisting districts besides their own). The model objective was to minimize the mean travel time with the added constraint that the travel-time between every point and its assigned center could not exceed a set maximum travel-time. To assess the performance of the human analyst, the same problems were solved with various heuristics and the quality of each party's solutions were compared. The results of these experiments were that the districts developed by the human analysts outperformed those developed through the heuristic methods all within a limited number of iterations.

2.6 Discussion

The late 1960s and 1970s saw the rise of modeling paradigms and methods such as simulation, mixed-integer programming, heuristic solution procedures, mixed optimization-simulation models, queue-theory based models, and conceptual or theoretical developments in systems modeling. However, research of this era was fraught with many challenges that included limited computational resources, lack of data of about patients, limited technologies, and the budding field of emergency medicine.

In any case, most EMSS location models of today are based on the development of this era. These models have become more sophisticated in many theoretical and technical respects; however, it is not too difficult to connect today's work with projects or ideas from this time. Perhaps the most influential and enduring idea from this era is the use of surrogate performance measures for analyzing public facilities including EMSSs. ReVelle et al. (1970) explored this concept in the context of location models for the public sector while Gibson (1973) explored performance measures for EMSSs. More recently, advances in EMS research have prompted location models that use more direct performance measures (e.g., Zaffar, Rajagopalan, Saydam, Mayorga, & Sharer, 2016), however, their use is still being justified (van Buuren, van der Mei, & Bhulai, 2017).

As for the future of EMSS modeling, Aringhieri, Bruni, Khodaparasti, & van Essen (2017) provide and excellent, extensive overview of the state of EMSS modeling and management. They note how EMSSs have developed in just about every respect but data collection/management issues and the organization of EMSSs are two persistent challenges. Granted, data issues are more complex today as they include developing more sophisticated information and communication systems or collecting new forms of patient data to better

understand outcomes as a function of service metrics (e.g. the time taken to respond to an emergency call). However, organizational issues that include financing, managing, and planning EMSSs remain a challenge because of economic, political, and geographical issues (Pozner et al., 2004).

3. Model Formulation Background

In this section, we present the models considered to be fundamental precursors in the development of the new model presented in this thesis. The first two models, the Location Set Covering Problem (LSCP) of Toregas et al. (1971) and Maximal Covering Location Problem (MCLP) of Church & ReVelle (1974), form the fundamental aspects of our model. A third model is the *p*-Median Problem (PMP), originally defined by Hakimi (1964 & 1965) and formulated as a programming model by ReVelle & Swain (1970). Key elements of the PMP are present in the new model construct as well, however, these components serve as model extensions rather than as core components. Although all three models employ a common mathematical programming modeling framework (explained below), the PMP is based on a different but related class of location models (Church & ReVelle, 1976).

After introducing these models, we discuss the key issues of capacity and congestion when addressing EMSS operations. Both deterministic and non-deterministic location models that attempt to address these issues are briefly presented and discussed. Then, we present several important non-deterministic models such as the Maximum Expected Covering Location Problem (MEXCLP) of Daskin (1982, 1983), the Maximum Availability Location Problems (MALP 1 and 2) of ReVelle & Hogan (1989), and the Queuing Maximum Availability Location Problem (QMALP) of Marianov & ReVelle (1996) which serves as the base model

for our location model – the Resource Constrained Queuing Maximum Availability Location Problem (RC-QMALP). We also present (to different extents) an assortment of location models that contain or develop features present in RC-QMALP.

3.1 Fundamental Models

3.1.1 The Location Set Covering Problem

At the core of RC-QMALP (as well as that of MCLP, MEXCLP, and MALP) is the LSCP of Toregas et al. (1971). Like the LSCP, RC-QMALP and the other models retain two fundamental modeling constructs concerning how an ambulance system is modeled and analyzed. First, the LSCP is based on a *mathematical programming model framework*. A mathematical program consists of: 1) a set of *decisions* to be made; 2) a set of *constraints* that the decisions must be meet; and 3) an *objective function* that measures the fitness of any decision. Thus, within this framework, the various goals, constraints, and decisions concerning the ambulance system planning process are translated into one of these components and incorporated into a single decision-based mathematical program. Second, to guide the ambulance system planning process the LSCP utilizes a *coverage-oriented modeling* paradigm.[52] Here the focus is centered on determining: 1) when a facility covers a customer or demand node; and 2) what level of coverage is to be provided. Coverage-oriented modeling involves the use of a distance or time standard (or some other metric) and involves serving as many demands as possible or all of the demands within that service standard, although other

[52] For an introduction to alternative paradigms for analyzing ambulance location models see ReVelle et al. (1970), Morrill & Symons (1977), and Savas (1978). ReVelle et al. (1970) discuss the difficulty of developing performance measures for the public sector and proposes some measures, although they are mostly based on efficiency. Morrill & Symons (1977) and Savas (1978) focus on equity-based measures although the concepts of efficiency and effectiveness are also discussed in detail.

factors, such as a facility's capacity or availability, can be considered concurrently. There are two general approaches: *requirement-based* models, where coverage requirements or restrictions are stipulated with constraints, and *goal-based* models, where the objective function promotes or discourages certain forms of coverage or allocations. We note that these two approaches are not mutually exclusive. All models discussed in this section are coverage-oriented except for the PMP.

The classic form of the LSCP is defined on a network of nodes and arcs. Nodes represent places of demand as well as potential facility sites. In the LSCP, the objective is to minimize the number of facilities needed (and locate them) in order to cover each demand node at least once by a facility. Facilities cover a demand node only if they are located within the prescribed distance/time standard, *s*. To capture the decision to locate a facility in the LSCP, for each potential facility location *j* there is a decision variable X_j that takes the value of 1 when a facility is located at site *j* and is 0 otherwise. Thus, the objection function simply involves minimizing the sum of all X_j variables. The coverage requirements are incorporated into an inequality based constraint that stipulates that the sum of the X_j decision variables, corresponding to the set of facilities that can cover node *i*, must be greater than one.

The formulation of the LSCP is as follows:

Model:

$$(LS-O) \quad Minimize\ Z_{LSCP} = \sum_{j \in J} X_j$$

$$(LS-C1) \quad \sum_{\forall j \in N_i} X_j \geq 1\ ;\ \forall i \in I$$

$$(LS-C2) \quad X_j \in \{0,1\};\ \forall j \in J$$

Notation:

Indicies and Sets

$I =$ set of demand nodes.

$J =$ set of potential facility locations.

$i =$ index of demand nodes, i $\in I$.

$j =$ index of potential facility lcoations, $j \in J$.

$N_i = \{j \mid t_{ji} \leq s\}$ - the set of facility locations j in the neighborhood of demand node *i*.

Parameters

$t_{ji} =$ Shortest travel time/distance from facility node j to demand node *i*.

$s =$ Maximal travel time/distance standard.

Decision Variables

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a facility is located at site } j. \\ 0, & \text{otherwise.} \end{cases}$$

The objective of the LSCP (LS-O) is to minimize the number of facilities that are located such that all demand nodes are covered at least once. This goal is formulated as minimizing the sum of the decision variables, X_j as this sum is equivalent to the number of facilities that are needed to provide complete coverage

For every demand node i, there is a corresponding constraint (LS-C1) that specifies that the node must be covered. The left-hand side of (LS-C1) consists of the sum of X_j decision variables are within the coverage standard of i. The right-hand side of the constraint specifies that at least one of these facilities must be selected. Thus, for a solution to be feasible, facilities must be arranged in a way that each demand will have at least one facility in its coverage set, N_i. Constraint (LS-C2) simply stipulates that all X_j location decision variables are 0-1 binary

decision variables. This model is an integer-linear programming problem and is often solved through the use of general purpose optimization software.

3.1.2 The Maximal Covering Location Problem

In terms of planning an ambulance system, the solutions produced by the LSCP are appealing as all demands are covered by at least one facility/ambulance. However, public agencies might not possess the financial resources to provide such a level of coverage. Consequently, an ambulance system planner must inevitably decide how to allocate service when faced with financial constraints

As previously mentioned, to address this issue, Church & ReVelle (1974) developed the MCLP where the goal is to maximize the amount of demand that is covered by a set of facilities given that only a fixed number of facilities can be located.[53] By limiting the number of facilities that are located, the MCLP incorporates the financial constraints of the ambulance service providers into the location model while attempting to achieve the total coverage requirement of the LSCP.

Although the LSCP and MCLP are based on a similar modeling paradigm, there are significant differences between the two models present in all three components of the location model. Even though the MCLP retains the X_j decision variable without any modifications, it is also based upon an additional set of binary 0-1 decision variables, Y_i, which are used to indicate whether specific demand nodes have been covered. The coverage constraints are adapted to allow the model to determine the level of coverage provided to each demand node

[53] White & Case (1974) also defined a similar problem called "the partial cover problem" although it only considered maximizing the total *number* of demand points covered within some standard, that is, the objective function is unweighted. In addition, they did not define this as an integer programming model, w a key feature to the use of the MCLP and its variants.

endogenously (via decision variable Y_i) and a new constraint is added to restrict the number of facilities that can be located to fit within the ambulance system operator's budget. Finally, the objective is changed so that coverage is maximized.

The MCLP is formulated as follows:

Model

$$(MC-O) \quad Maximize\ Z_{MCLP} = \sum_{j \in J} d_i Y_i$$

$$(MC-C1) \quad \sum_{\forall j \in N_i} X_j \geq Y_i;\ \forall i \in I$$

$$(MC-C2) \quad \sum_{\forall j \in N_i} X_j = p$$

$$(MC-C3) \quad X_j \in \{0,1\};\ \forall j \in J$$

$$(MC-C4) \quad Y_i \in \{0,1\};\ \forall i \in I$$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

Parameters

t_{ji} and N_i are as previously defined for the LSCP.

$p =$ total number of facilities to be located.

$d_i =$ call frequency per unit of time at demand node *i*.

Decision Variables

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a facility is located at site } j. \\ 0, & \text{otherwise.} \end{cases}$$

$\forall i \in I$:

$$Y_i = \begin{cases} 1, & \text{if demand node } i \text{ is covered by a facility.} \\ 0, & \text{otherwise.} \end{cases}$$

The objective of the MCLP (MC-O) is to maximize the amount of demand that is covered by at least one located facility within some time/distance standard, *s*. Here the objective value maximizes the sum-product of the demand at location *i*, d_i, and a corresponding decision variable, Y_i. The role of constraint (MC-C1) is to determine whether each demand node *i* is covered at least once. It is similar to constraint (LS-C1) of the LSCP in that the sum on LHS side is equal to the number of located facilities that cover demand node *i* and that demand node *i* is considered to be covered only when the LHS sum is greater than 0. Where (MC-C1) differs from (LS-C1) is that it does not require that at least one facility be established near demand *i*. Instead, (MC-C1) has a 0-1 decision variable Y_i on its RHS that allows for the possibility that no facilities are located within a maximal time/distance *s* of location *i*. To allow for this possibility, Y_i must take on the values 0 as Y_i must take a value less than or equal to the number of located facilities that cover demand node *i*. Constraint (MC-C2) simply requires that exactly *p* facilities are located. The LHS of (MC-C2) is equal to the total number of facilities that are located as this quantity is determined by summing all the X_j decision variables. Then to satisfy constraint (MC-C2), this sum must be equal to *p*. Constraints (MC-C3) and (MC-C4) simply stipulate that the location and coverage decision variables, respectively, X_j and Y_i are binary 0-1 decision variables. When solving the integer restrictions on the y_i variables can be dropped as long as they are restricted to be no greater than 1 in value. This model, like that of the LSCP is an integer-linear programming problem. Reasonable sized problem instances can be solved with general purpose software.

3.1.3 The *p*-Median Problem

The objective in the PMP is to minimize the total weighted travel times/costs, where each demand is assigned to its closest located facility, while locating exactly *p* facilities (Hakimi,

1964, 1965). The problem was first formulated as an integer programming model by Vinod, (1969) and independently by ReVelle & Swain, 1970). The *p*-median problem does not employ a coverage-oriented paradigm but rather a *minsum distance/time* paradigm where the emphasis is on minimizing the average service distance faced by demands (Eiselt & Marianov, 2011). While the model objective is to minimize the sum of travel times/costs between demands and located facilities, the classic PMP does impose restrictions or limits on travel times/costs. [54] In terms of EMS planning, the PMP is naturally appealing because of its focus on reducing the average travel time. Furthermore, it shares some of the appeal of the MCLP as there is a constraint on the maximum number of facilities to be located.

The classic formulation for the PMP relies on a set of assignment or allocation variables. Assignments are captured by X_{ij} binary 0-1 decision variables that take the value 1 when demand node *i* is assigned to a facility at *j*. As such, location decisions are implicitly declared through the X_{jj} decision variables,[55] where self-assignment, $x_{jj} = 1$, represents the fact that demand at j is assigned to itself for service, indicating that site j has been selected a facility.

[54] Church & ReVelle (1976) investigated the theoretical links between the PMP and MCLP and found that the MCLP can be considered a special case of the PMP. To implement a coverage-oriented paradigm into the PMP, they proposed replacing travel times/costs that exceed the standard with the value of 1, and setting all other travel times/costs to zero. This creates an objective of minimizing the amount of demand that is not covered, which is equivalent to maximizing what is covered.

[55] The assumption here is that every demand node is a potential facility site, however, this assumption is easily relaxed.

The formulation is as follows:

Model

$$(PM-O) \quad Minimize\ Z_{PMP} = \sum_{j \in J} \sum_{i \in I} d_i t_{ij} X_{ij}$$

$$(PM-C1) \quad \sum_{\forall j \in J} X_{ij} = 1;\ \forall i \in I$$

$$(PM-C2) \quad X_{jj} \geq X_{ij};\ \forall i \in I,\ j \in J : i \neq j$$

$$(PM-C3) \quad \sum_{\forall j \in J} X_{jj} = p$$

$$(PM-C4) \quad X_{ij} \in \{0,1\};\ \forall i \in I,\ j \in J$$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

Parameters

t_{ji} is as previously defined for the LSCP.

p and d_i are as previously defined for the MCLP.

Decision Variables

$\forall j \in J, i \in I$:

$$X_{ij} = \begin{cases} 1, & \text{if a facility } j \text{ is assigned to demand node } i, \text{ at site } j. \\ 0, & \text{otherwise.} \end{cases}$$

The objective of the PMP (PM-O) is to minimize the sum of the weighted travel times/costs between all demand nodes *i* and their assigned located facility (*j*). The travel times/costs are weighted according to the demand for each node, thus objective function consists of the double sum-product of the demand at node *i* (d_i), the travel time/costs between demand node *i* and a facility location *j* (t_{ij}), and the assignment decision variables X_{ij}. The role of constraint (PM-C1) requires each demand node to assign to a facility. The role of constraint set (PM-C2) is to ensure that demand nodes are assigned only to located facilities, as for any i, X_{ij} can only equal 1 when a facility has been located at that j (i.e. X_{jj}

= 1). Constraint (PM-C3) simply requires that exactly *p* facilities are located. The only difference is that the LHS is the sum of decision variables X_{jj} which implicitly represent the facility location decisions in the PMP. Constraint (PM-C4) simply stipulates that the assignment decision variables X_{ij} are binary 0-1 decision variables.

3.2 Modeling Capacity and Congestion in Location Models

One shortcoming of the LSCP and similar location models (including the MCLP and PMP) as originally formulated is that they do not consider issues of capacity or congestion (Current & Storbeck, 1988; Dearing & Jarvis, 1978). With these models, it's implicitly assumed that the located facility network can handle all demand covered or assigned. In other words, capacity constraints or the possibility of unavailable facilities (due to congestion) are not considered in these models.[56] As previously mentioned, several types of models and strategies have been developed to address these issues. Focusing on developments concerning the three models outlined above, we begin this discussion with models that use *deterministic* approaches to solve these problems of capacity and congestion before moving to models that use *non-deterministic* approaches that are generally more complex.

One additional note is that we distinguish *capacity-based* and *redundancy-based* models in that the former directly specifies some properties about the capacity of the individual

[56] It is important to note that if an EMS system rarely experiences congestion (locally and globally) the assumption that ambulances are always available is likely to be valid as, by definition, it is unlikely for the system to receive an additional call for service from any part of the system (that is covered) while a local ambulance is busy serving another call. Thus, in such systems, using simple, uncapacitated models such as the MCLP can be appropriate. Nonetheless, in EMS systems that experience a significant amount of congestion, it is often the case that the closest ambulance is unable to respond to a call for service as it is busy attending an earlier call for service. Therefore, under such conditions the use of uncapacitated models such as the MCLP would not be an appropriate as they would overestimate the availability of ambulance service.

facilities/servers while the latter are formulated to *implicitly* capture system congestion by encouraging redundant coverage of demand. These properties are not mutually exclusive or limited to deterministic or non-deterministic modeling approaches but it is important to keep this distinction in mind.

3.2.1 Deterministic Location Models

One of the earliest works to consider the problem of both locating facilities and dealing with facility capacity issues in the work of Kuehn & Hamburger (1963). The problem here represented a form of a capacitated warehouse location problem (CWLP) which is similar to the PMP but differs in that the CFLP generally considers a variety of fixed and variable costs (such as costs associated with building and operating a facility) in addition to travel distance/time costs. Moreover, unlike the *p*-median problem, the CWLP does not set a fixed number of facilities to be located. Gough & McCarthy (1975) considered Kuehn & Hamburger's model but it was neither recommended or applied in their investigation.

Two significant developments with capacitated PMPs came with the works of Ross & Soland (1977) and Neebe (1978). Both models retain the PMP's *minsum* approach (that is, the objective function minimizes transportation costs) and share key features but they are based on two different models. Ross & Soland (1977) adapt the generalized assignment problem (GAP) of Ross & Soland (1975) to form the Constrained Capacity PMP (CCPMP) that implements an additional constraint that effectively limits the amount of demand that a located facility can serve. It should be noted that Holmes, Williams, & Brown (1972) formulated an earlier version of the capacitated PMP but they did not attempt to solve it.

Like the PMP, the CCMP of Ross and Soland employs binary 0-1 decision variables ($X_{ij} \in \{0,1\}$) although it uses twice as many decision variables in order to conform to a GAP structure (in addition to needing a second additional constraint).[57] If we let b_j equal the capacity of location j the resulting capacity and decision variable constraints[58] are, respectively, as follows:

$$(CPM-CC) \qquad \sum_{i \in I} d_i X_{ij} \leq b_j \ ; \forall j \in J$$

$$(CPM-DV) \qquad X_{ij} \in \{0,1\}; \forall i \in I, \ j \in J$$

In constraint (CPM-CC), the LHS represents the demand allocation from nodes i to facility j which is limited to the capacity of a facility j (b_j) as noted on the RHS.[59]

Compared to the CCPMP, Neebe (1978) takes a simpler approach that combines the PMP with the CWLP to form the p-Median Transportation Problem (PMTP).[60] Here each facility location carries a limited amount of supplies that must be transported to meet the demand of various locations. Consequently, the PMTP decision variables concern the amount of assignment of supplies from facilities to demand points ($X_{ij} \geq 0$). With b_j as defined for the CCMP and Y_j as defined in the MCLP, the resulting capacity, demand, and decision variable constraints for the PMTP are, respectively, as follows:

[57] In terms of the GAP, the additional "task" variables (X_{ij}) are used to designate located facilities and an extra "agent" (or constraint) is required to keep track of the total number of located facilities.

[58] For clarity, we exclude from the formulation the additional "task" variables used to designate located facilities.

[59] Other parts of the model require that demands are allocated to facilities.

[60] Heller, Cohon, & ReVelle (1989) develop a model similar to that of Neebe (1978) in the context of EMS.

$$(PMT-CC) \qquad \sum_{i \in I} X_{ij} \leq b_j Y_j \; ; \forall j \in J$$

$$(PMT-DC) \qquad \sum_{j \in J} X_{ij} = d_i ; \forall i \in I$$

$$(PMT-DV1) \qquad X_{ij} \geq 0; \forall i \in I, j \in J$$

$$(PMT-DV2) \qquad Y_j \in \{0,1\}; \forall j \in J$$

In constraint (PMT-CC), the LHS represents the demand allocations from nodes *i* to a facility *j* while the RHS represents the capacity of facility *j*. The key to this constraint is that allocations cannot be made to facility *j* unless a facility is located at *j* (i.e. $Y_j = 1$). If no facility is allocated to *j* then the RHS will be zero as $Y_j = 0$. This will then prevent any assignments to that location. The role of constraint (PMT-DC) is to ensure that that all demand from node *i* are assigned across the set of located facilities. Constraint (PMT-DV1) is notable in that the assignment decision variables represent a non-negative flow from a node *i* to a facility *j*. This effectively makes the PMTP more flexible than the CCMP by allowing assignments of a demand to be split among several facilities. In the CCMP, if node *i* is assigned to a facility *j*, all of node *i*'s demand is assigned to that facility *j*, which is quite restrictive when compared to the flexibility of the PMTP.

As with the PMP, there are also capacitated versions of the MCLP and the LSCP. Early formulations of the capacitated MCLP are presented Chung, Schilling, & Carbone (1983), Church & Somogyi (1985), Current & Storbeck (1988), Pirkul & Schilling (1988), and Pirkul & Schilling (1991). With the exception of the models in Current & Storbeck (1988), these capacitated MCLP formulations are based on the original MCLP formulation (with an objective of maximizing covered demand) that is supplemented with a capacity constraints. Furthermore, instead of X_i decision variables there are X_{ij} decision variables that establish the service assignment between nodes (*i*) and facilities (*j*). Chung, Schilling, & Carbone (1983)

use binary 0-1 X_{ij} decision variables to indicate the assignment of node *i* to facility *j*. As such, their capacity constraints resemble the CCMP's capacity constraint (CPM-CC) although the term on the RHS is multiplied by a Y_j (as defined above) so that the facility *j*'s capacity is available only when the facility has been established (i.e., Y_j=1). The relevant capacity constraints are as follows:

$$(CMC1-CC) \qquad \sum_{i\in I} d_i X_{ij} \le b_j \mathrm{Y}_j \;;\forall j \in J$$

$$(CMC1-DV) \qquad X_{ij} \in \{0,1\}; \forall i \in I,\, j \in J$$

Church & Somogyi (1985) opted instead for the use of continuous X_{ij} variables to allow for the possibility of a demand being partially covered and/or served by one facility as well as being totally served by several facilities. But, a decidedly different element was that they allowed for more than one server or facility to be located at a given site. For our purposes, this would mean that it would be possible for several ambulances to be co-located. To add this capability, they expanded the definition of the Y_j to be nonnegative and integer in value, representing the number of servers that are located at node *j*. Thus, $Y_j \in \{0 \cup \mathbb{N}^+\}$). The relevant capacity constraints are the following:

$$(CMC2-CC) \qquad \sum_{i\in I} X_{ij} \le b_j \mathrm{Y}_j \;;\forall j \in J$$

$$(CMC2-DV) \qquad X_{ij} \ge 0; \forall i \in I,\, j \in J$$

Current & Storbeck (1988) present models with both approaches although they're formulated around a version of the MCLP where the objective is to minimize uncovered demand. Pirkul & Schilling (1988) exapnd on the MCLP by considering the workloads for two classes of facilities, a primary service facility (associated with a reponse standard s^p) and a secondary service facility (associated with a reponse standard s^b). Demand from nodes can be

assigned to facilities as primary or secondary service with the only difference being that facilities provide secondary service to more distance nodes (presumably $s^p<s^b$). Likewise, primary and secondary service are treated equally and are added together to determine a facilities total workload. Finally, Pirkul & Schilling (1991) propose a multiobjective model that combines the objective of a capacitated MCLP (with non-binary decision variables) and the PMP. The PMP is included in this model as all demand is required to be covered and so the PMP objective serves to minimize the average travel distance while all demands are covered.

Current & Storbeck (1988) also present formulations for the capacitated LSCP based on the original LSCP with a capacity constraint (similar to CMC1-CC), although here their variables, x_{ij} represent the fraction of demand from node *i* that is assigned to facility *j*. For their first model, the Constrained Set Cover Location Problem 1 (CSCLP1), we have the following constraints:

$$(CLC-CC) \qquad \sum_{i \in I} d_i X_{ij} \leq b_j \mathrm{Y}_j \;; \forall j \in J$$

$$(CLC-DV) \qquad X_{ij} \geq 0; \forall i \in I, j \in J$$

Constraints (CLC-DV) can be modified to have "all-or-nothing" assignment by placing the restriction $X_{ij} \in \{0,1\}$ for all $i \in I$ and $j \in J$. Current & Storbeck (1988) also develop a second model (CSCLP2) based on the Capacitated Plant Location Problem (CPLP). However, the relevant capacity constraints remain unchanged with respect to CSCLP1. Finally, Current & Storbeck (1988) note that a CSCLP2 can be remodeled as a GAP by modifying the GAP version of the CPLP presented in Ross & Soland (1977) .

REDUNDANCY-BASED MODELS

As previously discussed, redundancy-based models attempt to implicitly account for congestion. For coverage-based models, the most common modeling approach is a multi-objective approach that accounts for the number of times that a demand node is covered. These models are generally classified as Hierarchical Objective Location (HOL) models.

Daskin & Stern (1981) developed the Hierarchical Objective Location Set Covering Problem (HOLSCP) that has a primary objective of covering every demand node at least once and a secondary objective of maximizing the sum of the extra number of times demand nodes are covered beyond the initial coverage. In this model, all redundant coverage is considered equally – all nodes are weighted evenly and there are no decreasing returns for every level of additional coverage of a node.[61] Berlin (1972) developed a functionally similar model but his model used a different formulation based on a maximizing objective function rather than a minimizing one as in Daskin & Stern's model (1981). Moreover, Berlin (1972) did not report any computational results for this model (Daskin, Hogan, & ReVelle, 1988). Benedict (1983) further developed Daskin & Stern's (1981) model by allowing non-uniform node weights (equal to the demand at each node) although additional levels of coverage were all counted the same. Thus, both models did not consider a decreasing return or value as the number of times a demand is covered. Hogan & ReVelle (1986) expanded on Benedict (1983) to address some of the problems with the earlier work. This included the use of a relaxed time standard in

[61] For instance, a situation where two nodes are covered redundantly, respectively, 7 and 3 times is equivalent to when two nodes are covered redundantly, respectively, 2 and 8 times.

providing redundant coverage as well as limiting redundant coverage to counted at most once as a second or "backup" facility.

Benedict (1983) and Hogan & ReVelle (1986)[62] also formulate models for the maximal cover version of the HOL problem, the Hierarchical Objective Maximal Covering Location Problem (HOMCLP). The model by Benedict (1983) uses the same formulation as the MCLP with the exception of the objective function, which has two terms. The first term takes the single term in the MCLP's objective function and multiplies it by a large nonnegative weight (W') while the second term counts the number of time redundant coverage is provided for each demand. The counts are weighted according to the amount of demand at each node. As with the HOLSCP, constant returns for additional levels of coverage remain in this model. While Benedict (1983) addressed the HOMCLP alone, Hogan & ReVelle (1986) address a combination of the HOMCLP in addition to the HOLSCP through the introduction of mandatory coverage constraints that correspond to a relaxed time standard (the facilities parameter, *p*, is set such that these mandatory coverage constraints can be satisfied).[63] Moreover, they only allow redundant coverage to be counted at most once for each demand[64]. Benedict (1983) also presents a third model with two coverage standards. The objective is to maximize coverage under either standard and there is a mandatory coverage constraint for the relaxed standard. As with the other models formulated by Benedict (1983), there are no

[62] These models are presented and discussed in Daskin et al. (1988).

[63] This requirement would seem to make this a LSCP-type problem but I defer to the authors as they have chosen what the terms they use mean (the reported motivation for this constraint is that "[it] is often desirable to provide some minimal level of service to all nodes").

[64] Hogan & ReVelle (1986) present an extension to this model that allows for a second level of back up coverage.

decreasing returns for extra levels of coverage under either coverage standard. Lastly, Daskin et al. (1988) present a model with three coverage standard levels where the objective is to maximize the coverage at the most restrictive and relaxed standard. Mandatory coverage at the mid-level standard is required and only an additional level of redundant coverage is counted in their model.

Deterministic coverage-based models that consider congestion are relatively easy to conceptualize given their emphasis on standards. As for deterministic, non-standard focused minsum models, one approach considers "restrictions" on the demand nodes (rather than on the facilities themselves) where there are constraints on how demand from a node can be allocated. This approach appears as early as in the work of Swoveland et al. (1973b) with their stability hypotheses conjecture where its assumed that demand points are served by the k^{th} closest ambulance according to some stable distribution.

Later, Weaver (1979) further developed this approach by formulating the first deterministic, *minsum*-type mathematical program where response by non-closest ambulances was considered - the Vector Assignment *p*-Median Problem (VAPMP). The PMP and the VAPMP share the same fundamental structure but in the latter, it is assumed that each demand will be served a fraction of the time by their k^{th} closest facility. Consequently, every demand node is assigned to multiple facilities in terms of the distance between the demand node and each facility. In a subsequent publication, Weaver & Church (1981) extended this model to consider minimum workload, maximum workload, and workload range (the difference between the busiest and most idle facility) constraints as secondary objectives. Trade-off curves were used to analyze the relationship between the primary and secondary objectives. Soon thereafter, Weaver & Church (1985) formally presented the VAPMP formulation along

with a mathematical proof that an optimal solution exists for any VAPMP (this had only been suggested in Weaver & Church, 1981). This proof however only applied for instances with non-increasing assignment vectors, that is, situations where the k^{th} closest assignment fraction is at least as large as the k^{th} +1 assignment fraction. Lei & Church (2014) relax this restriction with their formulation of the Vector Assignment Ordered Median Problem, however, that model only considers all nodal solutions.

3.2.2 Stochastic and Probabilistic Location Models with Congestion

As noted in our review of EMS response models, the majority of early EMS response models were based on a stochastic or probabilistic modeling approach. Bell & Allen (1969) and Chaiken (1971) employed queueing theory in their models; Swoveland *et al*. (1973b) developed a probabilistic ambulance model; Davidson (1969), Hall (1971), and Larson (1973, 1974) used Markov chains in their models with the latter employing finite-state continuous-time Markov processes; and Savas (1969), Fitzsimmons (1970, 1973), and Siler (1977) used simulation to model system congestion.

Despite these developments, stochastic and probabilistic models were the exception until the 1980s. Many important models were proposed in the late 1980s and throughout the 1990s although the use of such models remained computationally challenging. Location models were often paired with simulation models to model congestion (e.g., Berlin & Liebman, 1974). As Berman & Krass (2015) note, by 2006 the number of publications have been substantially increasing ever since.[65]

[65] Berman & Krass (2015) use rather strict criteria in defining a stochastic location model with congestion and include many models with complex mathematical programs (i.e., highly non-linear programs). Nonetheless, the it remains true that the number of publications in this area has substantially increased.

According to Berman & Krass (2015), stochastic location models that consider congestion are based upon several assumptions: (1) a stochastic stream of demand, (2) facilities with servers that are capacitated or have stochastic service times, and (3) congestion that might result in the formation of queues or lost customers. They also focus on immobile facilities or models where the customers visit the facilities in their review. For this thesis, we expand the scope to include: (1) models with probabilistic measurements (such as server or system busyness fractions) or constraints with probabilistic elements, (2) reliability models that consider when service in unavailable to a demand node, and (3) mobile servers.

All models in this section consider capacity explicitly or implicitly and thus require a different classification scheme. To organize this section, we have divided these models into three categories: (1) reliability-based models, (2) districting models (with and without inter-district cooperation), and (3) other stochastic location models.

RELIABILITY-BASED MODELS

Chapman & White (1974) developed one of the first probabilistic location models, the probabilistic LSCP (PLSCP). In this model, the objective is to minimize the number of located facilities such that all demand nodes can be serviced under some time/distance standard with a minimum level of reliability α, $\alpha \in [0,1]$. Each facility is assumed to be unavailable with a probability q_{ij} which is equal to $1 - p_{ij}$ where p_{ij} equals the probability that customer i is covered by facility j and $p_{ij} = a_{ij}d_j$ where d_j is the probability that facility j is available and $a_{ij} = 1$ if customer i is accessible by a facility j within some distance.

The key aspect of this model was the use of probabilistic "chance-constraints". The LSCP and this PLSCP almost share the exact formulation apart from the mandatory coverage constraint. The LSCP's mandatory coverage constraint (LS-C1) is modified as follows:

$$(PLS-C1A) \qquad 1-\prod_{j\in\{N_i|X_j=1\}} q_{ij} \geq \alpha_i;\ \forall i \in I$$

The RHS is the required level of service reliability for demand node *i* while the LHS is effectively *Prob*[Located facilities in the neighborhood of demand node *i* are available]. A key assumption here is that the facilities' availabilities are independent of each other.

One challenge with this formulation is the non-linear multiplicative term on the LHS of (PLS-C1A). To work around this issue, Chapman & White (1974) replace (PLS-C1A) with the equivalent constraint:

$$(PLS-C1B) \qquad \sum_{j\in N_i} e_{ij} X_j \geq b_i;\ \forall i \in I$$

where e_{ij} = - log q_{ij} and b_i = - log(1-α_i).[66] The equivalence between the two constraints is due to the monotonic and logarithmic nature of the transformations.

Ball & Lin (1993) also developed a Probabilistic Reliability Location Set Covering Problem (PRLSCP) but their model centered around located facilities rather than demand nodes. In Chapman & White (1974), the unavailability of a located server (*j*) is related only to the demand at the location of the facility (d_j) and consequently, there is no consideration for the demand from other nodes in the neighborhood of such a facility (that is $D(j) = \sum_{i\in N_j} d_i$).

[66] If q_{ij} = 0, Chapman & White (1974) recommend setting e_{ij} to an "arbitrarily large positive value." If $\propto_i$ = 1, they recommend assigning b_i "the same arbitrarily large positive value".

Ball & Lin (1993) take a different approach by estimating the probability that a facility *j* has no servers and is not available, $P[D(j) \geq x(j)]$, where $x(j)$ is the number of servers at facility *j*. The underlying assumptions here are that (1) service times are constant ($\overline{T}$), (2) call arrivals are Markovian, and (3) $D(j)$ is a Poisson random variable representing the total call volume of facility *j*'s neighborhood (N_j) during a time-period (t, t+$\overline{T}$). As such, the mandatory coverage constraint in Ball & Lin (1993) is as follows:

$$(PRLS-C1A) \qquad 1-\prod_{j\in N_i}\prod_{1\leq k\leq C_j} P[D(j)\geq k]^{X_{jk}} \geq \alpha_i;\ \forall i \in I$$

$$(PRLS-C2) \qquad \sum_{1\leq k\leq C_j} X_{jk} \leq 1;\ \forall j \in J$$

where C_j is the capacity of facility *j* ($C_j \geq 0$) and X_{jk} is 1 if facility *j* has *k* servers (otherwise X_{jk} = 0). The LHS of Constraint (PRLS-C1A) is the product of the probabilities that a server is not available to demand node *i* and the RHS is the desired level of reliability.

As with PLSCP, the PRLSCP's mandatory coverage constraints are also transformed into a linear form:

$$(PRLS-C1B) \qquad \sum_{j\in N_i}\sum_{1\leq k\leq C_j} a_{jk}X_{jk} \geq b_i;\ \forall i \in I$$

where a_{jk} = -log[$P(D(j) \geq k)$] and b_i = - log(1-$\propto_i$). Also, we have that a_{jk}, $b_i > 0$.

Two important properties or considerations regarding the PRLSCP worth expanding on include how coverage is determined or accounted for and the role of the fixed service time assumption. Regarding coverage in the PRLSCP, every facility in a demand node's neighborhood ($j \in N_i$) is assumed to serve all the demand nodes within its own neighborhood ($i \in N_j$) without regard to whether other facilities (and their servers) outside neighborhood N_i

can serve demand nodes accessible by facilities within neighborhoods N_i (that is, demand nodes $i \in \bigcup_{j \in N_i} N_j$). This assumption can result in overestimating the number of required facilities and servers as relevant facilities outside a demand node's neighborhood are ignored but also because demand within a demand node's neighborhood can be double counted (or more). As for the fixed service time assumption, Baron, Berman, Kim, & Krass (2009) note that assuming fixed service times is rather unrealistic given the high variability in actual service times. More importantly, they demonstrate that it's possible to generate optimal solutions with an unreasonably high number of servers even if the service time parameter serves as an upper bound. Moreover, they also show that using "aggressive" service times (in this case the service times at the 50th percentile)[67] can lead to infeasible solutions where the reliability requirements are not satisfied.

DISTRICTING-BASED MODELS

The primary focus of EMS based districting models is to determine how a region can be divided into smaller districts or subregions that are each served by a facility.[68] There are two general types of districting models, *uncooperative districting models* where facilities cannot provide service across districts and *cooperative districting models* where facilities primarily provide service to their host district but can also provide inter-district service. In this sense, the model of Carter, et al. (1972) is an early stochastic uncooperative districting model. They modeled demand as a collection of Poisson process and servers as queues. The queuing

[67] They assumed an exponential distribution for service times.

[68] Districting models are similar to location-allocation models but they emphasize the boundaries between regions.

elements in their model were essential for estimating the steady-state probabilities of their two-server system (the joint probabilities that each server was busy and/or idle) and for determining workloads.

Berman & Larson (1985) considered a similar two-server uncooperative districting model but in this model: (1) queuing and delays resulting from system congestion were considered, (2) the objective function considered both minimizing travel time and queuing delays, and (3) the facility locations were fixed. Furthermore, servers were modeled as a $M/G/1/\infty$ queue, that is, demand arrival is Markovian (specifically a Poisson process), the service time distribution is General, there is [1] server, and there is an infinite queuing capacity. To solve this problem a "parametric classification" of optimal policies was undertaken for four regions with different demand intensity rates, each representing a continuous interval between 0 and λ (the total demand intensity of the system). Then two heuristics were developed to solve the districting problem and the optimal policies for all λ values.

Berman, Larson, & Chiu (1985) also developed two stochastic districting models, the Stochastic Loss Median Problem (SLMP) on a network where M/G/1/0 queues are used to model servers and the Stochastic Queue Median Problem (SQMP) where M/G/1/∞ queues are used instead.[69] Unlike Berman & Larson (1985)'s model, however, server locations were not fixed in these models and they only located a single server. A location-allocation algorithm was devised to solve the two problems.

Ultimately, Berman & Mandowsky (1986) extended these stochastic districting models to *m* facilities by developing a heuristic that combined the algorithm from the single facility

[69] Batta (1989) considered the SQMP with a finite discrete set of potential facility locations.

location model of Berman, Larson, & Chiu (1985) and the heuristic from the 2-facility model of Berman & Larson (1985). Some interesting observations reported by Berman & Mandowsky (1986) include that: (1) response times are not sensitive to changes in location-allocation policies in situations where there is low demand, (2) slight changes in location or allocation policies in situations where there is high demand can produce substantial (and potentially "disastrous") changes, and (3) when there is high demand, optimal locations are not intuitive (even in simple networks) and "popular median-proximity" location-allocation policies "can cause the system to explode."

As for cooperative districting models, their modeling approach is primarily based on Larson (1974)'s Hypercube model where servers are modeled as a *M*/*M*/*N* system with distinguishable servers. The model is rather powerful as it is possible to calculate the proportion of demand served by each server and the steady-state behavior of the system. Two issues with the Hypercube problem are that approximated travel times are used[70] (unlike the uncooperative models described above) and that it is very computationally expensive. For a system with *m* servers, the Hypercube model involves solving 2^m simultaneous equations. To reduce the problem size, Larson (1975) developed an approximation for the Hypercube problem that requires solving *m* simultaneous nonlinear equations for problems with *m* servers. However, this approximate method requires assuming that service times are identical for all servers, independent of how customers are allocated. Jarvis (1985) later generalized Larson's approximation to allow general service time distributions that may vary by customers and/or servers. Jarvis's approximation only applies to systems with no queues, however.

[70] Using a "Mean Calibration Time" procedure.

Berman & Larson (1982) considered the *p*-Median Problem with Congestion (PMPC), a median problem with the objective of minimizing expected response times and delays due to congestion for random service requests. In this network model, demand originates strictly from demand nodes and occurs as a homogeneous Poisson process. Moreover, customers are serviced by their most preferred available server whereby preferences are fixed, determined beforehand, and can consider factors other than travel times (i.e., specific server or customer characteristics). Customers also enter a queue only when all servers are busy and are served in a first-in, first-out (FIFO) manner. Also, any facility can house multiple servers. To solve this problem, Berman & Larson (1982) extended Berman & Larson's (1985) 1-server model into a multi-server and multi-facility problem based on $M/G/n/\infty$ queues. However, due to the analytical intractability of $M/G/n/\infty$ queues, in Berman & Larson's (1982) model travel times are only implicitly considered in that although their distribution is general, their distribution is not dependent on server location, server location and identity, or the history of the system. Effectively, the assumption is that on-scene travel times are significantly larger than travel times such that the system state probabilities in this model only depend on the intensity of demand at each node, the on-scene service times, and the server preference rankings. Lastly, they prove that there is an all nodal solution for the PMPC given any set of fixed server preferences and that the Hypercube model and Jarvis's algorithm can be used to solve the all nodal PMPC without a loss of generality. Berman, Larson, & Odoni (1981) consider a similar model with some simplifications. In this model, there is no queueing capacity - it is assumed that a back-up system (with a fixed average response time) provides service if all servers are busy. Also, each facility can only house a single server.

Berman, et al. (1987) developed two heuristics based on the Hypercube model to solve the Stochastic Queue *p*-Median Problem (SQPMP). In this model, *p* servers respond to calls for service from the network nodes. Each node is modeled as an independent Poisson generator. Customers are placed in a queue if all servers are busy. Otherwise, they are serviced by the nearest available server on a FIFO basis. Moreover, there is a general service time distribution and thus, the system is modeled as a $M/G/p/\infty$ queue with *distinguishable* servers. As with the PMPC, the SQPMP objective is to locate *p* servers to minimize expected response times to random calls and waiting times. However, because of the lack of a closed form for the expressions or approximations for waiting times in $M/G/p/\infty$ queues an alternative approximation of the objective function is suggested. Notably, the waiting component is the defined as product of the probability that *j* calls are queued times the *expected response time* when *j* calls are queued (summed over $j = 0$ to $j = \infty$). To solve the SQPMP, Berman, Larson, & Parkan (1987) developed a modified version of a heuristic proposed by Jarvis (1976) (Heuristic 1) and a heuristic based on the location-allocation algorithm used in the SQMP (Heuristic 2). Both heuristics performed similarly except for "intermediate" call arrival rates where Heuristic 2 performed better. Nonetheless, the authors "strongly recommend" Heuristic 1 due to its simplicity and lower computational requirements.

MULTIOBJECTIVE AND LOCATION-ALLOCATION MODELS

The models discussed above have objective functions that consider waiting times, travel times, and waiting costs. Other considerations can include the costs associated with locating servers at a facility and costs associated with rejecting a call. Also, these models include some forms of constraints on the number of facilities and servers as well as coverage constraints. Berman & Krass (2001) presents a generalized framework for Location Problems with

Stochastic Demand and Congestion (GLPSDC) that considers all these factors in a generic form where TC_{NC}, TC_{RC}, TC_{WC}, and TC_{LC} are, respectively, the total costs associated with not providing coverage to a customer, rejecting calls from customers that are covered by a facility, waiting times due to travel time and congestion, and locating servers.

Using the notation described above, the formulation provided is:

$$(LSDC-O) \quad Minimize\ Z_{GLPSDC} = TC_{NC} + TC_{RC} + TC_{WC} + TC_{NC}$$

$$(LSDC-C1) \quad \sum_{\forall j \in J} 1\{X_j > 0\} \leq \mathrm{M}$$

$$(LSDC-C2) \quad \sum_{j \in J} X_j \leq p$$

$$(LSDC-C3) \quad \sum_{j \in N_i} X_j \geq b_i Y_i; \forall i \in I$$

$$(LSDC-C4) \quad X_j \in \{0,..., p\}; \forall j \in J$$

$$(LSDC-C5) \quad Y_i \in \{0,1\}; \forall i \in I$$

The objective function (LSDC-O) minimizes the costs associated with the four factors described above, constraint (LSDC-C1) limits the *server* capacity of the facilities, constraint (LSDC-C2) restricts the maximum number of servers that can be located to be less than or equal to *p*, constraint (LSDC-C3) requires that a each demand node is covered by a minimum number of facilities, and constraints (LSDC-C4) constraint (LSDC-C5) define the decision variables for, respectively, the number of facilities at the facility in location *j* and whether a given demand node *i* is covered (Y_i=1) or not (Y_i=0). As this is a generic model, the various factors in the objective function can be weighted (or excluded) accordingly. Likewise, constraints (LSDC-C1) to (LSDC-C4) can also be adjusted accordingly (or excluded).

Numerous models fitting this general framework exist and listing them here is beyond the scope of this work. However, two novel examples with multiple objectives (with respect to the

models discussed above) and the location-allocation elements include the models of Melachrinoudis (1994) and Aboolian, Berman, & Drezner (2008).

Melachrinoudis (1994) developed two versions of the Discrete Location Assignment Problem with Congestion (DLAPC) where the objective is to provide service to all customers (located in a discrete set of locations) so as to minimize total costs. Demand is stochastic (with a general distribution and served on a FIFO basis), service times are distributed exponentially (both types of random variables are independent and identically distributed), and only one facility can be located at most on each site (thus servers are modeled as $G/M/1/\infty$ queues). In their first model (for a "Problem-P"), the objective function includes TC_{WC} and TC_{LC} terms while their second model's objective function (for a "Problem-U"), also includes TC_{WC} and TC_{LC} terms but the TC_{WC} term excludes waiting costs due to congestion. Both models use 0-1 binary decision variables for assignment of customers (i) to facilities (j) (X_{ij}) and for location decisions at every site j (X_j).

Aboolian, Berman, & Drezner (2008) formulated the problem of locating facilities and allocating servers on a congested network (LASCN) where the objective is also to provide service to all customers (located in a discrete set of locations) while minimizing total costs. In LASCN demand is stochastic with a Markovian distribution and multiple servers can be sited at the same facility – servers are thus modeled as $M/M/k_j/\infty$ queues (where k_j is the total number of servers at location j). As for the objective function, the LASCN includes a TC_{WC} term that includes both travel- and congestion-related cost and a TC_{LC} term that includes the fixed costs related to locating a facility at site j and the variable costs associated with number of servers at each facility (servers have a fixed price). The LASCN model also includes a closest-assignment constraint to ensure that customers visit the nearest located facility.

3.3 Essential Probabilistic and Stochastic Location Models

The formulation of RC-QMALP (and its variations) borrow concepts and components from various models. Again, RC-QMALP is an extension of QMALP and its predecessors MALP 1 and 2. These models in turn borrow a key concept from the MEXCLP. Moreover, although RC-QMALP is not strictly a location-allocation problem it borrows elements from and builds on non-deterministic location-allocation models, namely the models presented in Dearing & Jarvis (1978) and Marianov & Serra (1998). Finally, we present and discuss the two-stage modeling framework used by Shariat-Mohaymany, Babaei, Moadi, & Amiripour (2012).

3.3.1 The Queueing *p*-Median Problem

By the late 1960s many research groups were looking into using queues to model ambulance systems, particularly with the intention of capturing congestion. One significant shortcoming for many of these models however, was that they did not consider the location of ambulances (e.g., Bell & Allen, 1969). To address the issue of location and congestion various modeling approaches were adopted including simulation (e.g., Savas, 1969), non-optimizing analytic models (e.g., Larson, 1974), optimizing analytical models (e.g., Carter et al., 1972), mathematical programs with simulation (e.g., Berlin & Liebman, 1974), and analytic models with heuristics (e.g., Fitzsimmons, 1973). The problem with these approaches was that they, respectively, considered a limited set of alternatives, were mostly descriptive (rather than prescriptive), were computationally intractable for reasonably sized problems, produced solutions with models that did not capture the system behavior very well, and produced solutions that might not be optimal. Chapman & White (1974) devised a prescriptive optimization model that captured congestion but it was not implemented due to mathematical

challenges. In this context, the work of Dearing & Jarvis (1978) is notable and rather significant (despite its lack of presence in the literature[71]) in that it implements queues into a mathematical program along and includes an algorithm to find optimal solutions to the Queuing *p*-Median Problem (QPMP).

The QPMP is a network model where calls originate from a finite number of demand points i ($i \in I$, $|I|=n$). Calls for service from each demand node are modeled as independent Poisson processes with at a rate λ_i. At most a single server (and a total of p servers, $p < n$) can be located at any facility site j ($j \in J$, $|J|=m$) and there is a travel time t_{ij} between each demand node i and facility j. Moreover, servers in this system travel to demand points requesting service, provide service, and return to the same facility. Service time is a random variable with an expected on-site service time of τ_{ij} (thus, the total service time between demand node i and facility j is $T_{ij} = 2t_{ij} + \tau_{ij}$). Each server is modeled as an independent $M/G/1/\infty$ queue.

The QPMP's objective is to minimize the average expected travel times such that the expected waiting times at every located facility j is no more than W_j. If the set of demand points (I) is equal to the set of facility sites (J) then QPMP is equivalent to the PMP with an added congestion constraint (and without the added constraint it is exactly equivalent to the PMP). Because the congestion constraint serves as a capacity constraint, the QPMP is a capacitated *minsum* location problem.

[71] A Google Scholar search in August 2017 for works citing Dearing & Jarvis (1978) returned 6 references – 3 journal articles, 2 chapters in a handbook and a book, and 1 PhD dissertation. This article appeared most recently in the latter (published in 2000) and was last citied in a journal article by Melachrinoudis (1994).

The formulation is as follows:

Model

$$(QPM-O) \quad Minimize\ Z_{QPMP} = \sum_{i\in I}\sum_{j\in J}\lambda_i t_{ij} X_{ij}$$

$$(QPM-C1) \quad \sum_{\forall j\in J} X_{ij} = 1;\ \forall i \in I$$

$$(QPM-C2) \quad X_{ij} \le X_i; \forall i \in I,\ j \in J$$

$$(QPM-C3) \quad \sum_{\forall i\in I} X_i \le p$$

$$(QPM-C4A.1) \quad E[W_j] \le W_j;\ \forall j \in J$$

$$(QPM-C4A.2) \quad \rho_j < 1;\ \forall j \in J$$

$$(QPM-C4B) \quad \sum_{\forall i\in I} \lambda_i (T_{ij}^2 + 2W_j T_{ij}) X_{ij} \le 2W_j;\ \forall j \in J$$

$$(QPM-C5) \quad X_{ij} \in \{0,1\}; \forall i \in I,\ j \in J$$

$$(QPM-C6) \quad X_j \in \{0,1\};\ \forall j \in J$$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

Parameters

t_{ij} = is as previously defined for the LSCP.

p = is as previously defined for the MCLP.

τ_{ij} = the expected on-site service time of a server located at facility j for demand node i [time unit].

τ_{ij}^2 = the second moment of service time.

T_{ij} = the total expected time for a server located at facility j for demand node i.

$$T_{ij} = 2t_{ij} + \tau_{ij}$$

λ_i = call intensity per time unit at demand node i.

ρ_j = the utilization rate of the server located at facility j.

$$\rho_j = \sum_{i \in I} T_{ij} \lambda_i X_{ij}$$

W_j = maximum waiting time standard for facility j [time unit].

$E[W_j]$ = the expected waiting time in a queue at facility j.

$$E[W_j] = \sum_{i \in I} \lambda_i T_{ij}^2 X_{ij} \,/\, 2(1 - \sum_{i \in I} \lambda_i T_{ij} X_{ij})$$

Decision Variables

$\forall i \in I, j \in J$:

$$X_{ij} = \begin{cases} 1, & \text{if demand node } i \text{ is assigned to a server at facility } j, \\ 0, & \text{otherwise.} \end{cases}$$

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a server is located at facility j,} \\ 0, & \text{otherwise.} \end{cases}$$

The objective of the QPMP (QPM-O) is to minimize the sum of the weighted travel times/costs between all demand nodes i and their assigned server located at facility (j). The travel times/costs are weighted according to the demand for each node, thus the objective function consists of the double sum-product of the demand intensity at node i (λ_i), the travel time/costs between demand node i and a facility location j (t_{ij}), and the assignment decision variables X_{ij}. The role of constraint (QPM-C1) is like (PM-C1) in that it forces coverage of all demand nodes by requiring the assignment of all demand nodes to a server at a located facility.

Likewise, (QPM-C1) requires that demand nodes be assigned to at most one facility while facilities can be assigned to serve to multiple demand nodes. The role of constraint set (QPM-C2) is similar to that of constraint set (PM-C2) in that they ensure that demand nodes are assigned only to servers at located facilities as for any j, X_{ij} can only equal 1 when a facility is located ($X_j = 1$).

As with constraint (PM-C3), constraint (QPM-C3) simply requires that exactly p facilities are located. They are both formulated with the LHS representing the sum of location decision variables but in the QPMP explicit location decision variables X_i are used. Constraint (QPM-C4A.1) stipulates that the waiting times for a facility j cannot exceed a time standard W_j (on the RHS). The LHS is the expected waiting time formula for a $M/G/1/\infty$ adjusted for the QPM with respect to system utilization rates (ρ_j), the arrival rate of the queue (λ_i), and the second moment of service time (T_{ij}^2). Constraint (QPM-C4A.2) serves as a server capacity constraint as the LHS represents the server's utilization rate. If a server's utilization rate was equal to or greater than 1, the server queue would exhibit unstable behaviors including an ever-increasing queue. Constraint (QPM-C4B) simplifies the model as it is a linear constraint (unlike the QPM-C4A.1 constraint) and it implies both (QPM-C4A.1) and (QPM-C4A.2). Constraint sets (QPM-C5) and (QPM-C6) simply stipulate that the assignment and location decision variables, respectively, X_{ij} and X_j are 0-1 binary decisions variables.

The main difference between the QPMP and RC-QMALP is that there are no waiting related constraints in RC-QMALP. As for similarities, first, the QPMP's objective function (QPM-O) is used in RC-QMALP but as a *second-stage objective function*. RC-QMALP is solved in two stages whereby in the first stage the RC-QMALP objective function is used subject to the RC-QMALP's constraints. Then, for the second-stage, the QPMP objective

function (QPM-O) becomes RC-QMALP's objective function while RC-QMALP's first-stage objective function is transformed into a constraint that is bounded from below by the optimal value associated with the optimal solution of RC-QMALP's first-stage objective. This new constraint is added to the other RC-QMALP constraints. Thus, in the second stage, the RC-QMALP objective minimizes response times subject to a coverage performance constraint and the original RC-QMALP constraints.[72]

The second contribution is the use of continuous, fractional X_{ij} assignment variables ($0 \leq X_{ij} \leq 1$) and their interpretation. Dearing & Jarvis (1978) do not test this approach but suggested that X_{ij} could be assumed to be continuous.[73] As for meaning, continuous X_{ij} values, can be interpreted in two ways: (1) X_{ij} values represent the proportion of demand that is assigned from demand node *i* to the server at facility *j* and (2) X_{ij} values represent the probability that demand node *i* will be served by a server in facility *j*.

3.3.2 The Maximum Expected Covering Location Problem

The MEXCLP of Daskin (1982, 1983) is a derivative of the MCLP that adopts a probabilistic coverage-oriented and goal-based modeling approach. It is structurally similar to the redundancy-based models however, its probabilistic coefficients in the objective function make this a non-deterministic model - the objective function estimates the amount of demand that is expected to be covered (in a probabilistic sense).

[72] We present an extensive discussion of this modeling approach in Section 4.3.7.

[73] They note that Stidham (1971) discussed several models that used continuous, fractional X_{ij} assignment variables and M/M/1 queues to model servers.

The MEXCLP's key contribution is that of server busy fractions and *their use in a mixed-integer linear program*. Prior works such as Volz (1971) included some form of a busy fraction or utilization measure but these modeling approaches/formulations were not compatible with mathematical programming.

The MEXCLP's approach to modeling coverage is implemented through its objective function by using coverage decision variables Y_{ik}. Much like the Y_i coverage decision variables in the MCLP, Y_{ik} variables are used to indicate that demand node i is covered. However, in the MEXCLP the coverage decision variables are extended so that they also indicate the level of coverage provided to a demand node. That is, for each potential level of coverage $k=\{1,\ldots,p\}$,[74] Y_{ik} takes the value 1 if k facilities cover demand node i and 0 otherwise. The mechanism for determining the level of coverage provided is implemented via a coverage constraint for each demand node.

Returning to the issue of determining the amount of demand that is covered in the presence of congestion, the MEXCLP objective function sums the product of the amount of demand (d_i), the decision variable Y_{ik}, and a weight (w_k) over every level of coverage k and each demand node i. The weight w_k is strictly decreasing concave over k to indicate the "diminishing returns" of each additional level of coverage. Moreover, it is calculated so that the sum of the product of w_k, d_i, and Y_{ik} over k is equivalent to the expected amount of demand from node i that can be covered. Hence, the objective function of the MEXCLP determines the overall expected coverage of demand.

[74] The maximum level of coverage a demand node can attain is equal to the number of facilities that are to be located. The total number of facilities to be located is also limited to p in the MEXCLP.

With respect to how w_k is calculated, Daskin (1982) based his calculations on a model parameter q that represents the probability that a randomly selected facility/server is busy. To derive q, first the system-wide workload is estimated with the product-sum of the amount of demand at each node, d_i, and a parameter μ that is equal to the average length of time a server spends servicing a call. Then, this is divided by p and T, which are, respectively, the number of facilities that are to be deployed and the length of the study period[75] while the product of p and T represent the amount of capacity in terms of time. Overall, the entire calculation estimates a *system-wide busy fraction*.

Once the value of q is known, w_k is set to equal the marginal increase in the expected coverage for a demand node i that results from increasing the number of facilities that cover demand node i from k-1 to k for $k \epsilon K$. Consequently, by using these w_k weights in the objective function the objective value is equal to the overall expected coverage of demand. Note that this calculation (provided below) is based upon the assumption that the number of available facilities follows a binomial distribution (and thus that the probability of one server being busy is independent of the state of other servers).

[75] Each facility is presumed in either an idle state or in a busy state serving a call during the entire period. Also, parameters d_i and μ are scaled according to the length and measurement unit of T.

The formulation is as follows:

Model

$(MX-O)$ $\quad Maximize\ Z_{MEXCLP} = \sum_{k\in K}\sum_{i\in I} w_k d_i Y_{ki}$

$(MX-C1)$ $\quad \sum_{\forall j\in N_i} X_j \geq \sum_{k\in K} Y_{ki};\ \forall i \in I$

$(MX-C2)$ $\quad \sum_{\forall j\in J} X_j = p$

$(MX-C3)$ $\quad X_j \in \{0,1\};\forall j \in J$

$(MX-C4)$ $\quad Y_{ki} \in \{0,1\};\ \forall k \in K,\ i \in I$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

$K = \{1,\ldots, p\}$ - the set of possible levels of coverage.

$k =$ index for level of coverage, $k \in K$.

Parameters

t_{ji} and s is as previously defined for the LSCP.

p and d_i are as previously defined for the MCLP.

$T =$ the duration of the study period [time unit].

$\mu =$ the average service time per call [time unit].

$q =$ the system-wide busy fraction. Estimated by:

$$q = \sum_{i\in I} \mu d_i \frac{1}{T}\frac{1}{p}$$

$w_k = (1-q)q^{k-1};\ \forall k \in K$ - the weight associated with covering demand node k times.

Decision Variables

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a facility is located at site } j, \\ 0, & \text{otherwise.} \end{cases}$$

$\forall i \in I, k \in K$:

$$Y_{ki} = \begin{cases} 1, & \text{if a demand node } i \text{ is covered by at least } k \text{ facilities,} \\ 0, & \text{otherwise.} \end{cases}$$

The objective of the MEXCLP (MX-O) is to maximize the expected amount of demand covered within a time/distance standard (s). This is accomplished by maximizing the sum of the product of a weight (w_k) representing the marginal increase in coverage resulting from moving from coverage by k-1 facilities to coverage by k facilities, the demand from node i (d_i), and the binary decision variable Y_{ki} over all demand nodes ($i \in I$) and all levels of coverage ($k \in K$).

Constraints (MX-C1) operate in a similar fashion to the coverage definition constraints of the MCLP (MC-C1) in that the LHS of the constraint counts the number of located facilities that are accessible to a demand node while the RHS determines whether there is a sufficient number of located facilities to consider the demand node to be covered at a given level k. The difference between (MX-C1) and (MC-C1) however, is that (MX-C1) tracks the number of facilities that cover a node rather than only determining whether one or more facilities cover a demand node. To account for multiple levels of coverage, the RHS of (MX-C1) contains a sum of decision variables Y_{ik}. The structure of (MX-C1) requires that the sum of the Y_{ik} decision variables (on the LHS) not exceed the total number of located facilities that cover the demand node i (the RHS sum). Constraint (MX-C1) does not provide any explicit order as to how the Y_{ik} variables are selected.[76] Nonetheless, all (MX-C1) constraints hold to equality (its LHS equals the RHS) and for all demand nodes, their corresponding Y_{ik} variables are equal to 1 for all coverage levels (k) that are less than or equal to the number of located facilities covering a demand node i.[77] This is due to the concave nature of the objective function of the MEXCLP

[76] The here is not explicit requirement, for any i, that Y_{ik} decision variables with higher k value be equal to 0 if there is a Y_{ik} decision variables with lower k value that is equal to 0.

[77] That is if the number of located facilities covering demand node i is equal to n ($n \geq 1$), then $Y_{ik} = 1 \; \forall k$: $1 \leq k \leq n$.

(MX-O).[78] As with constraints (MC-C2) in the MCLP, constraints (MX-C2) simply set the amount of facilities to be located at exactly *p*. The role of constraints sets (MX-C3) and (MX-C4) are to, respectively, define location and coverage decision variables X_j and Y_{ik} as binary 0-1 decision variables.

After the development of MEXCLP, ReVelle & Hogan (1988) attempted to relax the assumption of uniform systemwide server busyness factions as part of an effort to develop a probabilistic MCLP with service reliability constraints (MRCLP). To do so, they created a method to calculate a "local" estimate of server busy fractions. This method (described in more detail in the next section) is local as it involves estimating a busy fraction for each demand node *i* according to the total level of demand in the demand node's neighborhood (the demand in N_i). Sorensen & Church (2010) eventually extended MEXCLP to include local busy fraction calculations in the Maximum Expected Covering Location Problem with local reliability (LR-MEXCLP).

3.3.3 The Maximum Availability Location Problems

Like the MEXCLP, MALP 1 and 2 are probabilistic coverage-based models that utilize busy fractions to determine the availability of servers. More fundamentally, they both adopt a goal-based coverage approach as they do not involve set coverage requirements for each demand node and are based on a redundancy-based framework as they include decision variables that track the level of coverage provided to each demand node.

[78] See Daskin (1983).

The MALP models differ from the MEXCLP in the structure of the objective function and in how coverage is determined. First, whereas the MEXCLP seeks to maximize the expected amount of demand covered within some time/distance standard, the MALP models seek to maximize the amount of demand covered within some time/distance standard such that a minimum level of service *reliability* is provided to the demand node. To meet this service reliability requirement, a demand node must be served by a minimum number of facilities. MALP 1 (like the MEXCLP) use systemwide server busyness fraction to establish this number while MALP 2 uses the local server busyness fraction calculations developed by ReVelle & Hogan (1988) and implemented by ReVelle & Hogan (1989). The latter method estimates the facility requirement by accounting for the total demand in the demand node's neighborhood N_i (rather than the system's total demand).

As such, MALP 1 and 2 are based on a hybrid reliability- and redundancy-based coverage modeling framework. However, in contrast to the MEXCLP, redundancy is not emphasized in MALP 1 and 2 as the coverage-level decision variables only consider the provision of coverage at a single level (i.e. a node is alpha reliable covered or not). Consequently, maximizing the coverage on a local basis is prioritized in MALP 1 and 2 while maximizing the coverage across the entire system is emphasized in the MEXCLPs.

To understand the difference between MALP and MEXCLP, it is useful to consider MALP alongside MEXCLP's formulation. With the MEXCLP, excess coverage is captured with the Y_{ik} decision variables and the w_k parameter represents, respectively, a coverage-level of k facilities for a demand node i and the marginal improvement resulting from an additional facility's coverage (moving from k-1 to k level coverage). In contrast, MALP also adopts Y_{ik} decision variables, however, rather than considering all Y_{ik} variables in the objective function

like MEXCLP, in MALP only one level of coverage is considered, specifically some coverage-level k that ensures that a demand node is served with a minimum level of service reliability (α). In MALP 1, this reliable coverage parameter is b and it is the same for all demand nodes; thus, we have the decision variable Y_{ib}. In MALP 2, the reliable coverage parameter can vary across demand nodes; in that case we use b_i instead and the decision variable becomes Y_{ib_i}. Consequently, in MALP the objective is to maximize the sum of the product of Y_{ib} (or Y_{ib_i}) and d_i or the proportion of demand that is covered with a α-level reliability.

Reliability requirements are determined for the MALP 1 model by first calculating the system-wide busyness (q), just as in MEXCLP. Then q is used to develop chance-constraints similar those used by Chapman & White (1974). However, in MALP 1 α-reliable service is not required coverage, whereas in Chapman and White coverage is required. MALP 2 employs the approach developed by ReVelle & Hogan (1988) that modifies MEXCLP's system-wide busy fraction calculations to consider busyness at a more local level where local-region busy fractions q_i are used instead of system-wide busy fraction q. The key difference here is that rather than considering the total demand in a system with q, in calculating q_i only the total demand in demand node i's neighborhood (M_i) is considered. Then the needed b_i coverage levels are calculated from the corresponding q_i values.

The formulations for MALP1 and MALP 2 are as follows:

Model

MALP 1

$(MA1-O)$ $\quad Maximize\ Z_{MALP1} = \sum_{j \in J} d_i Y_{ib}$

$(MA1-C1)$ $\quad \sum_{\forall j \in N_i} X_j \geq \sum_{k=1}^{b} Y_{ik};\ \forall i \in I$

$(MA1-C2)$ $\quad Y_{ik} \leq Y_{ik-1};\ \forall i \in I,\ k \in \{2,\ldots,b\}$

$(MA1-C3)$ $\quad \sum_{\forall j \in J} X_j = p$

$(MA1-C4)$ $\quad X_j \in \{0,1\}; \forall j \in J$

$(MA1-C5)$ $\quad Y_i \in \{0,1\};\ \forall i \in I$

MALP 2

$(MA2-O)$ $\quad Maximize\ Z_{MALP2} = \sum_{j \in J} d_i Y_{ib_i}$

$(MA2-C1)$ $\quad \sum_{\forall j \in N_i} X_j \geq \sum_{k=1}^{b_i} Y_{ik};\ \forall i \in I$

$(MA2-C2)$ $\quad Y_{ik} \leq Y_{ik-1};\ \forall i \in I,\ k = 2,\ldots,b_i$

$(MA2-C3)$ $\quad \sum_{\forall j \in J} X_j = p$

$(MA2-C4)$ $\quad X_j \in \{0,1\}; \forall j \in J$

$(MA2-C5)$ $\quad Y_i \in \{0,1\};\ \forall i \in I$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

K and k are as previously defined for the MEXCLP.

N_i is as previously defined for the LSCP.

$M_i = \{l \in I \mid t_{li} \leq s\}$ - the set of demand nodes l in the neighborhood of demand node i.

Parameters

t_{ji} and s are as previously defined for the LSCP.

p and d_i are as previously defined for the MCLP.

$q, T,$ and μ are as previously defined for the MEXCLP.

$\alpha =$ the minmium service reliability standard with $\alpha \in (0,1)$.

$q_i =$ the local-region busy fraction. Estimated by:

$$q_i = \frac{\mu \sum_{l \in M_i} d_l}{T \sum_{j \in N_i} X_j}$$

$b = \left\lceil \frac{\log(1-\alpha)}{\log q} \right\rceil -$ number of facilities required for α reliability with uniform system-wide busyness estimates. $; b \in \mathbb{Z}^+$

$b_i = \left\lceil \frac{\log(1-\alpha)}{\log q_i} \right\rceil -$ number of facilities required for α reliability with local-region busyness estimates. $; b_i \in \mathbb{Z}^+$

Decision Variables

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a facility is located at site } j, \\ 0, & \text{otherwise.} \end{cases}$$

$\forall i \in I, k = 1, ..., b_i$:

$$Y_{ik} = \begin{cases} 1, & \text{if a demand node } i \text{ is covered by at least } k \text{ facilities,} \\ 0, & \text{otherwise.} \end{cases}$$

The objective of MALP 1 (MA1-O) and 2 (MA2-O) is to maximize the demand that is covered within a time/distance standard (s) and α-level reliability. This is accomplished by maximizing the sum over all demand nodes ($i \in I$) of the products of the demand at node i (d_i) and the binary 0-1 decision variable Y_{ib} in MALP 1 and Y_{ib_i} in MALP 2. We note that that the second subscripts in Y_{ib} and Y_{ib_i} are equal to the number of facilities that are needed to be located in the neighborhood of demand node i in order to meet the requirements for α-level reliable service. These constraints operate in a similar fashion to the coverage indicator constraints of the MEXCLP (MX-C1) in that the LHS of the constraint counts the number of

located facilities that are accessible to a demand node i while the RHS determines whether there is a sufficient number of located facilities to consider that the demand node has been covered at the α-level. The difference between (MX-C1) and the MALP constraints is that the latter only account for coverage up to a certain level – $k = b$ for MALP 1 and $k = b_i$ for MALP 2. This is reflected in the index of the RHS sums, respectively, in (MA1-C1) and (MA2-C1). These constraints ensure that the decision variables Y_{ik} "behave" properly such that Y_{ik} can only equal 1 if Y_{ik-1} is 1 as well (for $k \geq 2$). The only difference between *(MA1-C2)* and *(MA2-C2)* is that for each demand node *(MA1-C2)* considers the k values from 2 up to b while *(MA2-C2)* considers values from 2 up to b_i. As with constraints (MC-C2) in the MCLP, constraints (MA1-C3) and (MA2-C3) simply set the amount of facilities to be located at exactly p. The role of constraints sets (MA1-C4) and (MA2-C4), is to limit decision variables X_j values to 0 and 1 while the role of constraints sets (MA1-C5) and (MA2-C5) is to limit decision variables Y_{ik} values to 0 and 1.

In MALP 1, there are two assumptions from the MEXCLP that carry over. This includes the assumption that: (1) all servers being equally busy, (2) the probability of a server being available is independent of the state of other servers, and (3) a fixed average service time for all calls. In MALP 2, the first assumption is relaxed with the use of average busy fractions for servers in the region of each demand node i ($j \in N_i$). However, By relaxing assumption (1) ReVelle & Hogan (1989) introduce two additional assumptions (the *Districting Assumption*[79]), (1) that the servers located in a demand node i's neighborhood ($j \in N_i$) only serve demand nodes in the neighborhood ($i \in M_i$) and (2) that all calls originating in M_i are served by facilities in N_i.

[79] We refer to this as the *Districting Assumption* following the terminology of Berman & Krass (2001).

In making this assumption, ReVelle & Hogan (1989) acknowledge a potential issue resulting from some facilities in N_i might serve demand nodes outside N_i and also facilities outside of N_i might serve demand nodes in M_i. The idea is that they assume the net service (from outside to inside and from inside to outside) across the neighborhoods is zero.

In a subsequent publication on a similar model, Marianov & ReVelle (1996) address this issue with a different explanation. Here they claimed that an implicit assumption in MALP is that the call rates in a demand node *i*'s neighborhood do not "differ to a significant extent from the call rate in the neighborhoods that border *i*. They suggested that this established "a rough equivalence between 1) the number of calls originating outside of N_i and requiring servers stationed inside N_i, and 2) the number of calls inside N_i which require servers to come from stations in adjacent, or nearby, neighborhoods."[80] Moreover, they presented an additional assumption - that there was a minimal difference between the response times of servers located outside N_i serving calls in M_i and of servers located inside N_i serving calls outside M_i. With these two assumptions, Marianov & ReVelle (1996) argued that the flows of server in and out of N_i were "not too different" and "approximately cancel each other". This they argued justified their Districting Assumption, that is, treating "each neighborhood as an isolated, independent unit whose demands and servers interact solely with each other."

To test the districting assumption, Murray & Church (1992) assessed MALP-derived locational configurations by comparing their theoretical and simulated MALP objective function values, respectively, Z_{MALP2} and $SIM(Z_{MALP2})$. This experiment involved two data sets (55 and 33 node data sets) with various p and α values. A simple analysis of Z_{MALP2} and

[80] This is also suggested by Marianov & ReVelle (1992) although without the second assumption.

$SIM(Z_{MALP2})$ revealed that for both data sets and all p and α values, $Z_{MALP2} >> SIM(Z_{MALP2})$ apart from six cases. In two instances $Z_{MALP2} = SIM(Z_{MALP2})$ and in four $Z_{MALP2} < SIM(Z_{MALP2})$.

The next step investigated consistency in the differences between Z_{MALP2} and $SIM(Z_{MALP2})$ as consistent differences between the two models would support the robustness of MALP. In this experiment, Murray & Church (1992) generated 100 location configurations and evaluated them using MALP and a simulation for the 55 node data set with six p values (number of servers) with an α value of 0.90. They used a nonparametric statistical test (Sign test) to compare the fitness rankings of the locational configurations produced by evaluating the configurations with MALP and with the simulation model. The null hypothesis was that for any configuration MALP did not consistently produce better solution values than the simulation model (and vice versa). The null hypothesis was rejected for all six p values at a 0.01 significance level. However, an analysis of the variation in objective values differences (as a percentage of demand) for each level of p revealed large standard deviations. Thus, despite MALP and the simulation producing consistent ordering, Murray & Church (1992) concluded that there was a lack of agreement between the objective values produced by MALP and the simulation model.

Given these results, Murray & Church (1992) then investigated potential sources of the large standard deviation in objective values differences. For this analysis, they plotted the MALP and simulation models for each solution (at two p levels) and visually examined the resulting graph. Murray & Church (1992) suggested a tendency for MALP to produce conservative estimates of demand covered with c-reliability particularly for location configurations with mid-level MALP objective values. As a final step, Murray & Church (1992) considered the possibility that these discrepancies were due to the simulation model.

For this, they investigated at a more granular level and examined a single location configuration's results, specifically the reliability levels attained at each demand node. First, they examined the role of α and noted a wide discrepancy in model objective values at $\alpha = 0.90$ but very similar model objectives values with α values of 0.85 and 0.95. Then, they examined an area with two located servers where a large discrepancy in objective values existed. In this instance, they observed a significant difference in demand located in non-overlapping areas (i.e., the set of demand nodes covered exclusively by one of the servers). To test the significance of this difference, they reran both models with an altered data set where the demand in both non-overlapping zones were balanced by increasing/decreasing demand in the set with lower/higher demand. This change reduced the discrepancy in objective values for the server that covered the non-overlapping zone with high demand by *increasing* the MALP objective value estimate. Murray & Church (1992) suggest the local-busyness estimate is problematic because the calculations factor in all demand in a demand node's neighborhood but do not account for the extent to which this demand is served.

Baron *et al.* (2009) also raised similar concerns using an example problem. They used simulations to test the validity of availability measures generated by various location models. With respect to MALP, they generated optimal MALP solutions that are both feasible and infeasible with respect to the α-reliability requirements. These multiple optima concerned the authors because it showed that MALP lacked a mechanism that favored feasible solutions over non-feasible solutions. But more critically, they claimed that "it is not hard to construct a larger problem where all solutions are infeasible" for MALP (and another model).

3.3.4 The Queuing Location Set Covering Problem

Before discussing QMALP, its useful to briefly discuss ReVelle & Hogan's (1989) Probabilistic Location Set Covering Problem (PLSCP-RH)[81] as it represented a significant breakthrough in coverage-based location models that employed a mixed-integer linear programming framework. The 1980s marked the arrival of probabilistic based models with the development of MEXCLP, MALP, and ReVelle & Hogan's (1989) PLSCP-RH, α-Reliability *p*-Center Problem, and Maximum Reliability Location Problem (MRLP).[82] These models represented significant advancement in location modeling by operationalizing the probabilistic optimization modeling paradigm and introducing new concepts that were incompatible or could not be readily implemented with a deterministic modeling framework.

However, despite this significant advancement, coverage-based location modeling remained behind other location modeling approaches when it came to capturing congestion. Coverage-based location models excelled in finding optimal or at least high-quality solutions due to their mixed-integer linear programming framework. However, as these probabilistic coverage-based location models were being developed, models using other modeling frameworks (e.g., Hypercube-based models) and even mixed-integer linear programming-based minsum location models were already incorporating far more sophisticated modeling elements such as queues. Other models even expanded on probabilistic coverage-based location models although at the expense a mixed-integer linear programming friendly framework. For instance, Batta, Dolan, & Krishnamurthy (1989) introduced the Adjusted

[81] The RH is used to distinguish ReVelle & Hogan's (1989) and Chapman & White's (1974) models.

[82] These models applied the local server busyness calculations and chance constraints used in the MRCLP and MALP.

MEXCLP (AMEXCLP) that incorporated elements from Larson's Hypercube model (Larson, 1974,1975) to relax some of assumptions in the MEXCLP, namely the server independence assumption, a particularly important assumption.

The server independence assumption represented a disadvantage to probabilistic mixed-integer linear programming-based location models. This assumption was not just unrealistic for most cases but it results in an overestimate of server availability at the local level.[83] Nonetheless, models that relaxed this assumption also remained at a disadvantage because their of their computationally intensive solution procedures and because they did not always produce better solutions. Saydam, et al. (1994) came to this conclusion for MEXCLP in comparison to the AMEXCLP (and other models) and added that no model was consistently more accurate in estimating expected coverage. In this context, Marianov & ReVelle's (1994) Queueing Location Set Covering Problem (QLSCP) represented a huge breakthrough because it relaxed the server independence assumption in a coverage-based model that maintained a mixed-integer linear programming framework.[84]

The QLSCP and the PLSCP-RH share the same model formulation with the exception of how the α-reliability facility parameter (b_i) is calculated. Unlike in MALP and the PLSCP-RH,

[83] Assume P(A) and P(B) are, respectively, the probability of two ambulances A and B being available. We know that P(AB) = P(A)P(B|A) = P(B)P(A|B). It's reasonable to assume that an ambulance is more likely to be busy when the other ambulance is busy. As such, we have P(B|A) < P(B) and P(A|B) < P(A) which implies that P(A)P(B) > P(AB) = P(A)P(B|A) = P(B)P(A|B) or that the independence assumption over estimates ambulance availability. A more general proof can be easily derived with Bonferroni inequality. We note that it's also possible to underestimate the availability of service by underestimating server availability (i.e., upward biased P(A) and/or P(B) estimates).

[84] We note that Saydam & Aytuğ (2003) later developed a far less computationally intensive model that produced solutions with improved estimates and were often of better quality than the MEXCLP solutions. However, the optimality of their solutions remained in question as they employed a genetic algorithm heuristic.

in the QLSCP facilities are modeled as *M*/*M*/*K*/*K*-loss queues and demand nodes as Poisson processes with "demand intensities" (i.e. call arrival rates). As such, b_i's are calculated using steady-state equations:

(1) $\mu_i P_1 = P_0 \lambda^i; k = 0$

(2) $P_{k-1}\lambda^i + (k+1)\mu_i P_{k+1} = P_k + k\mu_i P_k; k = 1, ..., p_i$

such that:

$\bar{p}_i =$ parameter for number of servers in demand node $i's$ neighborhood, $\bar{p}_i \in \square$.

μ_i = mean rate of service completion per unit of time of a server located a demand node *i*.

λ^i = demand intensity in demand node *i*'s neighborhood:

$$\lambda^i = \sum_{k \in M_i} \lambda_k$$

ρ_i = the utilization rate at demand node *i*.

$$\rho_i = \frac{\lambda^i}{\mu_i}\ ;\ \rho_i \le 1$$

$P_k^i =$ the probability of the system at demand node *i* being in state *k* .

$$P_k^i = \frac{\frac{1}{k!}\rho_i^k}{\sum_{n=0}^{k} \frac{\rho_i^n}{n!}}; k = 1, ..., \bar{p}_i$$

Thus, b_i is calculated as follows:

$$\mathrm{b}_i = \arg\min_{b_i}\ f(b_i) := \{b_i \mid b_i \in \square^+ : P_{b_i}^i \le 1 - \alpha\}$$

That is, we seek to find the smallest integer value b_i such that we satisfy the constraint:

$$\frac{\frac{1}{b_i!}\rho_i^{b_i}}{\sum_{n=0}^{b_i} \frac{\rho_i^n}{n!}} \le 1 - \alpha$$

In formulating this model, some important considerations include that:

- If *s* servers are busy, any calls additional calls are assumed to be lost (hence the *s-loss* designation). Marianov & ReVelle (1994) explain that this loss is from the "point of view of that neighborhood" and that servers located outside the neighborhood fulfill these calls "in practice".
- The system transition rates (λ^i in this case) never change regardless of the state or busyness of a system and does not affect transition rates.
- ρ_i must be less that or equal to 1, otherwise system equilibrium is not possible.
- It can be shown that $P_k^i \leq P_{k-1}^i$ which implies that there is always a b_i such that $P_{b_i}^i \leq 1-\alpha$.

Insofar as modeling assumptions, Marianov & ReVelle (1994) also adopt the districting assumptions previously put forth by ReVelle & Hogan (1988), ReVelle & Hogan (1989), and ReVelle & Hogan (1989b). Moreover, they further justify their use of *M*/*M*/*s*/*s-loss* queues along with the districting assumption as this avoids the need to keep track of the state of each server in the system, in accordance with queue theory-based models (Larson, 1974). They further justify their approach with the claim that α values close to one should be used in order to "obtain useful results". Why useful results require high α values is not explained but they explain that with higher α values, servers in a demand node's neighborhood are more likely to respond and that consequently: (1) "it will only occasionally be necessary for servers outside to cross the boundary and attend calls (unless there is an extreme situation) and (2) "the flow of servers across boundaries should be small". It's also assumed that travel times are significantly smaller than service times as with other models that assume exponentially distributed service rates.

To assess the QLSCP's performance, Marianov & ReVelle (1994) compared the QLSCP and PLSCP-RH's solutions using a 55 node data set (Swain, 1971) and one scenario. They reported that higher b_i values for the PLSCP-RH with low α-reliability values (α = 0.80) and lower b_i values for the PLSCP-RH with higher α-reliability values (α = 0.95,0.99). They interpreted this as the PLSCP-RH overestimating and underestimating congestion, respectively, with low and higher α-reliability values. They also observed more evenly distributed facilities in some cases with high α-reliability values. Computationally, the PLSCP-RH solved faster in every instance (α = 0.80, 0.90, 0.95, 0.99) but the QLSCP times were similar in most cases.

Two important issues concerning QLSCP involve the problem of estimating the parameters for the model and the validity of the QLSCP availability estimates. Beginning with the first issue, Marianov & ReVelle (1994) parameterized the QLSCP with arbitrarily adjusted demand intensity values[85] and developed a mean service rate value that averaged the service times for three possible scenarios.[86] To address this, they briefly discussed some approaches that could estimate these model parameters by observing the system's behavior. Their first suggestion was modeling λ_i and μ_i as doubly stochastic processes, but they disregarded this approach noting that it was unjustifiably complicated unless each random parameter had a simple probabilistic distribution function. As an alternative, they proposed a method for both parameters that establishes confidence intervals using inequalities based on standard formulas

[85] They took the population values associated with each demand node in Swain (1971) and multiplied them by a constant factor (0.7).

[86] They considered the cases where (1) the ambulance arrives, stays on site, and returns to its ambulance depot; (2) the ambulance arrives, transports the patient to the hospital, and returns to the ambulance depot; and (3) the ambulance arrives and returns immediately to its ambulance depot (a false alarm scenario).

about stochastic processes. Then, they used the minimum parameter values that satisfy these inequalities and meet a given confidence coefficient value.

For the second issue, three articles raised the same concern of the validity of the availability estimates used in MALP. Alminana, Borras, & Pastor (1996) first raised this concern reporting that the specified α-reliability was achieved in less than 20% of 36 problems.[87] Next, Borrás & Pastor (2002) conducted an ex-post evaluation of several models (including the QLSCP) with two minimum local reliability level (MLR) measures. One measure assumed server independence (MLR^I) and the other did not (MLR^D). The test included: two different data sets - a modified version of Swain's (1971) 55-node network and Serra's (1989) 79-node network; four call demand configurations (i.e., different times of day); two distance standards; one average service time standard; and nine α-reliability levels (α = 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 0.99). With respect to the QLSCP MLR^I and MLR^D measures, Borrás & Pastor (2002) reported that the stated α-reliability was achieved in, respectively, 63.89% and 47.22% of cases. Moreover, QLSCP solutions meet the stated α-reliability *and* required the fewest number of vehicles (with respect to the other two models in this test) in 50.00% and 38.19% of cases under the MLR^I and MLR^D measures, respectively. Lastly, Baron et al. (2009) used their example problem to analyze the QLSCP and extended its conclusions about the validity of the MALP's availability measures to the QLSCP. In their analysis of the QLSCP (on a *M*/*M*/*K* framework[88]) they also produced both reliability feasible and infeasible solutions for the QLSCP, noted the lack of a guidance mechanism to produce reliability feasible

[87] This experiment was in its preliminary stage and the authors did not report any other statistics regarding such an experiment.

[88] They note that Borrás & Pastor (2002) did not remove this modeling assumption but reached a similar result.

solutions, and that it was possible to generate larger examples where all QLSCP solutions were reliability infeasible.

3.3.5 The Queuing Maximum Availability Location Problem

As with the QLSCP, the QMALP of Marianov & ReVelle (1996) represented a significant breakthrough in location modeling as a coverage- and goal-based model that incorporated queue theory and retained a mixed-integer linear programming friendly structure. Again, we stress the latter property as various models predating QMALP met the first three criteria, including Batta et al. (1989) and Goldberg et al. (1990), but these models relied on heuristic approaches that could produce optimal solutions but not guarantee their production.

Although QMALP and MALP share the same model structure (apart from a few subtle changes and the introduction of additional optional constraints), QMALP is conceptually and technically different as it is based on the queue-theory framework developed by Marianov & ReVelle (1994) for the QLSCP. The most significant difference concerns the calculation of b_i as QMALP is based on the same approach developed in the QLSCP. QMALP also is based on all of the QLSCP assumptions (including the districting assumptions) with a couple of exceptions related to the service times. In the QLSCP, exponentially distributed service times are assumed but this assumption is relaxed to generally distributed service times in QMALP. Nonetheless, the move to *M*/*G*/*s*/*s*-*loss* queues is of minimal operational significance as service times are assumed to include travel times and where both demand node state probabilities (P_i^k) and b_i calculations remain unchanged.[89]

[89] As noted in Berman & Larson (1982), moving to a general distribution service time that includes travel times assumes that service times are not dependent on server location, server location and identity, or the history

Three additional modifications in QMALP include allowing server co-location, the definition of a demand node's neighborhood (N_i), and a workload constraint. Of the three, only the first change is not optional (with the given formulation). Marianov & ReVelle (1996) implement with two changes. They set a server capacity for each location j (C_j) and change the decision variable X_j to X_{kj}. This new decision variable which is still a 0-1 binary decision variable accounts for location (j) and whether it is one of the kth servers at location j (with $k \leq C_j$). The change regarding the neighborhood definition N_i is an attempt to capture the impact of travel times on coverage, that is, N_i is determined such that it only includes other demand nodes such that the probability of reaching those demand nodes from demand node i within a time standard S is greater than or equal to β, a second standard. A normal distribution for travel times is assumed (following Daskin,1987) and an inequality that determines membership is derived. The inequality includes expected travel times between nodes and the variance in terms of a standard deviation.[90] As for the workload constraint, Marianov & ReVelle (1996) propose a constraint that effectively limits workloads by requiring a minimum number of server (g_i servers in a neighborhood N_i) so that the probability that all servers in a neighborhood are busy (P_s^i) remains below some rate $w \in (0,1)$. The value g_i is calculated using the same iterative procedure used to calculate b_i.

The formulation is as follows:

Model

of the system. Thus, this change in QMALP has theoretical and methodological significance but no practical significance. We revisit this issue when discussing RC-QMALP.

[90] We omit presenting this alternative definition and the workload constraint as they are not implemented by the authors.

$$(QMA-O) \quad Maximize\ Z_{QMALP} = \sum_{j \in J} \lambda_i Y_{ib_i}$$

$$(QMA-C1) \quad \sum_{\forall j \in N_i} \sum_{k=1}^{C_j} X_{kj} \geq \sum_{k=1}^{b_i} Y_{ik};\ \forall i \in I$$

$$(QMA-C2) \quad Y_{ik} \leq Y_{ik-1};\ \forall i \in I,\ k = 2,\ldots,b_i$$

$$(QMA-C3) \quad \sum_{\forall j \in J} \sum_{k=1}^{C_j} X_{kj} = p$$

$$(QMA-C4) \quad X_{kj} \in \{0,1\}; \forall j \in J, k = 1,...,C_j$$

$$(QMA-C5) \quad Y_{ik} \in \{0,1\};\ \forall i \in I, k = 1,...,b_i$$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

K and k are as previously defined for the MEXCLP.

N_i is as previously defined for the LSCP.

Parameters

p is as previously defined for the MCLP.

$C_j =$ the server capacity at location j.

λ_i is as previously defined for the QPMP.

α is as previously defined for the PLSCP.

μ_i is as previously defined for the QLSCP.

λ^i is as previously defined for the QLSCP.

P_k^i is as previously defined for the QLSCP.

b_i is as previously defined for the QLSCP.

Decision Variables

$\forall j \in J, k = 1,...,C_j$:

$$X_{kj} = \begin{cases} 1, & \text{if a } k\text{th facility is located at site } j, \\ 0, & \text{otherwise.} \end{cases}$$

$\forall i \in I, k = 1,...,b_i$:

$$Y_{ik} = \begin{cases} 1, & \text{if a demand node } i \text{ is covered by at least } k \text{ facilities,} \\ 0, & \text{otherwise.} \end{cases}$$

As with the MALP 2 objective (MA2-O), the QMALP objective (QMA-O) maximizes the amount of demand that is covered with α-reliability. Constraints (QMA-C1) and (QMA-C2) establish the coverage-level in each demand node *i*'s neighborhood. Note that the objective improves only when $Y_{b_i} = 1$ and so without (QMA-C2) we would have $Y_{b_i} = 1$ when then there is at least one server located in N_i regardless of the value b_i. To prevent this, (QMA-C2) requires the presence of $k - 1$ facilities in N_i to allow the possibility of Y_k being 1 and thus the Y_k values are properly set in (QMA-C1) beginning with Y_1. Constraint (QMA-C3) limits the total number of facilities to *p* and the capacity of locations (*j*) are established by the limiting the highest index value of the second summation to C_j.

Marianov & ReVelle (1996) assessed QMALP (and its performance relative to MALP) with Swain's (1971) 55-node network with a 45 minute average service time, five α-reliability levels (α = 0.85, 0.90, 0.90, and 0.97), and a 1.5 mile service standard. Interestingly, the investigation used a single level of demand intensity where demand nodes generated an average of 0.4 calls *per day*. In comparing MALP and QMALP, Marianov & ReVelle (1996) reported a similar distribution of b_i values in MALP and QMALP across most α-reliability levels although they observed an upward skewed distribution of b_i values in QMALP when α = 0.99. However, they also reported lower estimates of server availability for demand nodes with MALP than QMALP when a *single* server was located in the demand node's neighborhood. Finally, they also noted that the marginal decrease in the percentage of demand covered with α-reliability increased with higher α values (i.e., the drop in α-reliable coverage between α = 0.85 and 0.90 was significantly smaller than the drop between α = 0.90 and 0.95).

With respect to the validity of QMALP's availability measures, the QLSCP results and critiques of Alminana et al. (1996), Borrás & Pastor (2002), and Baron et al. (2009) also apply

to QMALP. Erkut, Ingolfsson, & Budge (2008) present another, broader critique of QMALP (and MALP). First, they present a general critique of set covering based models (probabilistic, stochastic, or deterministic) in that they produce uneconomically sound solutions as they include an excessive number of vehicles. Moreover, they argue that these type of objectives (minimum coverage or reliability levels) do not coincide with the priorities of EMS system practitioners. Second, they construct a pathological example to show that a model emphasizing systemwide reliability rather than local reliability (with equal α values[91]) can produce more desirable solutions (i.e., solutions with fewer vehicles).

As for specific issues concerning QMALP, Erkut et al. (2008) remark on the lack of guidance in setting both the α and β parameters and that, to their knowledge, EMS practitioners do not measure, track, or discuss such measures. Likewise, they also highlight a lack of guidance for setting the average service time parameter and note the difficulty in obtaining an accurate estimate a priori as these values are contingent on the server locations. As for the nature of QMALP solutions, they present four observations highlighting the challenges and disadvantages with using QMALP. First, they reported that QMALP solutions are sensitive to both α and β parameters after observing some large changes in coverage (20%+) after changing some parameter values. Second, they reported that the "best" α and β values varied with the number of servers although they observed a consistent relationship with high β values and high expected coverage. Lastly, they reported that QMALP compared unfavorably to the Hypercube-based model of Ingolfsson et al. (2008) in that (1) its solutions always outperformed the best QMALP solutions (given a fixed number of facilities over various α and

[91] Erkut et al. (2008) propose a systemwide reliability constraint of the form that requires that α fraction of the total system demand be covered reliably.

β values for QMALP) by covering 0.1-0.6% more expected demand and (2) QMALP took 2-6 times longer to solve.

3.3.6 The Queueing Maximal Covering Location-Allocation Problem

Marianov & Serra (1998) present another stochastic coverage-based location model called the Queueing Maximal Covering Location-Allocation Problem (QMCLAP). It is most similar to QMALP given its queue theory-based framework and implicit focus on system performance where the primary focus is to maximize coverage within a time/distance standard given *p* facilities/servers to locate). However, it differs in that the model uses different queue related performance measures/standards (waiting times and queue lengths) and uses a location-allocation framework. Moreover, while Marianov & ReVelle (1996) developed QMALP to model a system with mobile facilities that visit customers, Marianov & Serra (1998) model a system with immobile facilities that customers visit (such as banks, healthcare service centers, and distribution centers).

Although RC-QMALP's focuses on mobile servers that visit customers and return to their base, QMCALP is of interest because of how the location-allocation framework is used to capture congestion explicitly. Whereas models such as QMALP capture congestion implicitly (through service reliability constraints), the QMCALP assignment decisions help capture congestion explicitly by determining server allocations endogenously. The meaning of these assignments do not translate perfectly to a mobile server problem like RC-QMALP but they're useful for explicitly capturing congestion, a problem with QMALP. We examine this issue in the discussion section immediately below.

Marianov & Serra's (1998) QMCLAP formulation effectively combine the MCLP and QPMP formulations. As in both models, they denote location decisions with 0-1 binary X_j decision variables and like the QPMP, 0-1 binary X_{ij} decision variables are used to indicate an assignment of demand node i to server j when $X_{ij} = 1$. In a second formulation that is not presented here, they relax the implicit facility co-location restriction and allow the co-location of m_j servers at location j (also subject to a maximum total number of facilities, p). They present two performance standard constraints concerning queue length and waiting times. Using $M/M/1/\infty$ queues to model each facility, they develop two chance-constraints that set a lower bound for performance (α) for the probability that an arriving customer will (1) encounter b customers in the queue and (2) wait at the facility longer than some time W. Marianov & Serra (1998) use $M/M/K/\infty$ queues in for their second QMCLAP model and later Moghadas & Kakhki (2011) and Moghadas, et al. (2013) extend QMCLAP with $M/G/1/\infty$ and $M/G/K/\infty$ queues, respectively. Finally, we note that it is assumed that customers visit the nearest facility although QMCLAP does not contain constraints that enforce such customer behavior.

The formulation is as follows:

Model

$$(QMLA-O) \quad Maximize\ Z_{QMCLAP} = \sum_{i \in I} \sum_{j \in J} d_i X_{ij}$$

$$(QMLA-C1) \quad X_{ij} \le X_j; \forall i \in I,\ j \in N_i$$

$$(QMLA-C2) \quad \sum_{j \in N_i} X_{ij} \le 1; \forall i \in I$$

$$(QMLA-C3A) \quad \sum_{i \in N_j} \lambda_i x_{ij} \le \mu_j \sqrt[b+2]{1-\alpha}; \forall j \in J$$

$$(QMLA-C3B) \quad \sum_{i \in N_j} \lambda_i x_{ij} \le \mu_j + \frac{1}{W} \ln(1-\alpha); \forall j \in J$$

$$(QMLA-C4) \quad \sum_{j \in J} X_j = p$$

$$(QMLA-C5) \quad X_j \in \{0,1\}; \forall j \in J$$

$$(QMLA-C6) \quad X_{ij} \in \{0,1\}; \forall i \in I,\ j \in N_i$$

Notation

Indicies and Sets

I and J as well as i and j are as previously defined for the LSCP.

N_i is as previously defined for the LSCP.

Parameters

p is as previously defined for the MCLP.

s is as previously designed for the MCLP.

d_i is as previously defined for the MCLP.

α = a probabilistic service quality standard, $\alpha \in (0,1)$.

b = a maximum queue length standard, $b \in$ □ .

W = a maximum waiting standard [time unit].

Decision Variables

$\forall i \in I,\ j \in J$:

$$X_{ij} = \begin{cases} 1, & \text{if demand node } i \text{ is assigned to a server at facility } j, \\ 0, & \text{otherwise.} \end{cases}$$

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a server is located at facility j,} \\ 0, & \text{otherwise.} \end{cases}$$

The QMCLAP objective function (QMLA-O) maximizes the total amount of demand that is assigned to facilities located within some time/distance standard (*s*). Constraint (QMLA-1) allows assignments only between a demand node *i* and a facility *j* if facility *j* is located (X_j = 1); otherwise, when $X_j = 0$ then the assignments X_{ij} on the LHS must also be 0. Constraint (QMLA-2) restricts the assignment of demand nodes to at most a single server and only to servers located in their neighborhood. Constraints (QMLA-3A) and (QMLA-3B) represent two options for controlling system performance. Constraint (QMLA-3A) requires that the sum of the demand intensities assigned to facility *j* (the LHS) remain below a limit so that the probability that a facility at *j* has at most *b* customers is greater than or equal to $\propto$. The second constraint option (QMLA-3B) also requires that the sum of the demand intensities assigned to facility *j* (the LHS) remain below a limit so that the probability that the waiting times at facility *j* are at most *W* is greater than or equal to $\propto$. Marianov & Serra (1998) derived the RHS using known formulas about $M/M/1/\infty$ queues. Also, it is important to note that the summation of demand intensities on the LHS is valid for both equations because the sum of several independent Poisson processes is equivalent to a single Poisson process. QMALP relies on this property to calculate the arrival rate in a demand node *i*'s neighborhood (λ^i) but Marianov & ReVelle (1996) only assumed that λ^i was a Poisson process. Constraint (QMLA-4) simply limits the total number of located facilities to *p*.

After QMCLAP, Marianov & Serra (2002) presented a set covering version of QMCALP the Probabilistic Location–Allocation Set Covering Model with co-location of a pre-specified number *m* of servers per center (PLASC*m*). In this publication, the authors explicitly distinguish their fixed-server location model, from emergency service models such as QMALP and MEXCLP. They explain that with the PLASC*m* they model server capacity statistically

and that demand is assumed to arrive instantaneous. Furthermore, PLASC*m* constraints explicitly model closest-assignment constraints.

In this explanation, Marianov & Serra (2002) raise the important issue of how models of mobile facilities such as ambulances differ from immobile facilities such as a bank or hospital. In a chapter about location models with stochastic demand and congestion, Berman & Krass (2001) attempt to distinguish between the two facility types based on operational characteristics of each type of system as well as on methodological grounds. They begin with an $M/M/K/1$ model where all servers are co-located and explain that such models are not entirely appropriate for mobile server systems due to travel times that are typically not distributed exponentially. As such, they argue that $M/G/K/1$ models are more appropriate for such systems but that these systems are difficult to use because of a lack of some key analytical formulas for such systems. Then they move on to an M/G/K/1 system where the k and K-k servers are located at two distinct locations and the nearest available location responds to an emergency call. This case they note, is a system with *distinguishable servers* which pose the additional challenges of (1) an absence of approximate analytical results and (2) that the service times for consecutive calls are not independent (because the servers from the further location might need to respond). Lastly, they note that mobile systems typically operate under a directed choice policy where a central authority determines how customers are served rather than customers selecting which facility or server to use.

Berman & Krass (2015) revisit this issue in a chapter on stochastic models with congestion and make a similar argument that mobile system models need to consider *distinguishable servers* as these systems cannot be readily decoupled as a set of independent queueing systems. However, they emphasize and clarify that the need to distinguish servers arises from the

dynamic (or state dependent) nature of server assignments. Interestingly, they also note that the tractability of immobile facility location models depends on static server assignments even if they can be decoupled into a set of independent queues. With these observations, Berman & Krass (2015) effectively highlighted the nature of server assignments and identified it as a more fundamental factor than facility type. The argument here is that some immobile facility models operate more like a mobile facility (and vice-versa) due to the nature of assignments. For instance, they note that mobile facility models are more appropriate for immobile facility location systems where customer-facility assignments depend on the system state or those that have dynamic customer allocation systems. Likewise, they suggest immobile facility location models that can model mobile server systems with static and non-intersecting service regions for all facilities. Consequently, Berman & Krass (2015) suggest that perhaps it is more useful to differentiate between systems with static and dynamic assignments than between immobile and mobile server systems.

3.3.7 Ranked Multiobjective Location Models

In this review, we presented both *minsum* and coverage location models but have yet to discuss how these two modeling approaches can be unified within a single model (as is the case with RC-QMALP). RC-QMALP is predominantly a coverage-based model but it also includes considerations for travel times. Motivating this decision is a second hypothesis that reducing average travel times can improve the overall system performance.

Strict coverage models include little guidance with respect to a server's relative location to the demand nodes its serves. Consequently, the intuition behind this is that accounting for travel times should produce more central server location configurations at the neighborhood

level (i.e., server locations in a neighborhood's busier areas). Moreover, at an operational level, ambulances become more available through lower overall service times as they include travel times.

One approach for capturing both coverage and travel times is the *multiobjective approach* alluded to in the GLPSDC (Berman & Krass, 2001) where coverage (TC_{NC}) and a travel time (TC_{WC}) objectives are jointly considered in the objective function. Each objective is usually assigned a weight of θ and $(1- \theta)$, where $\theta \in [0,1]$, although some models include arbitrary weights if any at all. Daskin (1995) presents a simple, context-free version of such a model and more recently, Hosseini & Jabal Ameli (2011) developed a multiobjective EMS model with coverage and travel time objectives.

Multiobjective models are appealing because they allow an analyst to consider the tradeoffs between two objectives. That is, if Z^1 and Z^2 are the objective values for two objectives, we can use the weight θ (using the setup described above) to set the relative importance of Z^1 to Z^2. With a composite objective function $Z^M = \theta * Z^1 + (1 - \theta) * Z^2$, note that if $\theta = 1$ (or $\theta = 0$) then only objective Z^1 (or Z^2) is considered and with all θ values in between zero and one both objectives are considered.

Technically, multiobjective models are used to generate the non-inferior tradeoffs between two objectives or more objectives. However, in some cases, the interest in not in tradeoffs between objectives but rather, in solving problems with a hierarchy of objectives. Here objectives are first ranked in terms of their importance ($Z^1 \succsim Z^2 \succsim \ldots \succsim Z^N$).[92] Then the overall problem solved first with only the highest ranked objective in the objective function ($Z = Z^1$).

[92] $\succsim$ is a preference/ranking operator. If A is weakly preferred to B then $A \succsim B$ and vice-versa.

For subsequent objectives, the same problem is solved with only the lesser ranked objective ($Z = Z^n$, $1 < n \leq N$) but the problem is amended so that all higher ranked objectives (Z^i, $\forall i < n$) are included as constraints (individually) where the objective functions must be as good as or equal to its corresponding optimal solution value (if the objective maximizes or minimizes the objective then the constraint should be, respectively greater or less than the corresponding optimal solution value). If Z^{i*} represents the optimal objective function solution value to the problem with objective function Z^i then the problem is formulated as follows:

$$\textit{Minimize } Z = Z^n$$

such that:

$$Z^i \leq Z^{i*}; \forall i = 1, ..., n-1$$
$$b_j \geq 0; \forall j \in J$$
$$c_k = 0; \forall k \in K$$

where b_j and c_k represent generic constraints. This problem is completely formulated when $n = N$. Note that if the objective is maximized (minimized) when it's a constraint, its value must be greater (less) than or equal to the corresponding optimal value.

This approach to solving problems where solutions from one problem stage are used as parameters in a subsequent problem stage in known as the *ε-constraint method* (Ehrgott, 2005) developed by Haimes, Lasdon, & Wismer (1971) (as noted by Chanta, Mayorga, Mclay, & Wiecek, 2009). In formulating and solving these types of problems (for two objectives), Haimes et al. (1971) proved that this method is appropriate only when every problem stage $n < N$ generates a feasible solution (with respect to the subsequent problem stage) that is unique whether in terms of the objective value or the specific solution values. Also, the generic constraints do not have to be the same across all problem stages with the ε-constraint method.

However, EMS models (including RC-QMALP) mostly utilize the objective values (Z^{i*}) from each stage to parametrize subsequent stages as in the model formulated above.

Chanta et al. (2009) developed an EMS location model using the ε-constraint method that minimized the maximum distance between uncovered demand areas and located facilities subject to maximizing the total expected coverage of demand (within some acceptable bound). Shariat-Mohaymany, et al. (2012) also developed an ε-constraint EMS model where they first sought to minimize the costs associated with locating ambulances and stations and then searched for the solution that minimizes ambulance arrival times. As such, the first-stage objective minimized the weighted number of ambulances and ambulance stations where ambulance stations were capacitated. Ambulances were allocated to these stations. This was subject to MALP-like constraints with additional constraints to limit the server workloads at the neighborhood level. In the second stage, the model's objective function was to minimize the total response times subject to a constraint on the number of ambulances and ambulance depots. The objective function amounted to the sum of the travel times between the located ambulances and each demand node weighted by the corresponding proportion of demand assigned to the located ambulances from each demand node. The second stage problem included the previous stage's constraints in addition to constraints that limited and balanced the server workloads for each ambulance.

If both conditions required for the ε-constraint method are not met, then an integrated model that simultaneously considers multiple objectives is required. However, in some situations it's preferable to forego the ε-constraint method despite satisfying both of its prerequisites. In his dissertation, Church (1974) presented a simple, context-free model combining the MCLP and PMP while prioritizing the MCLP's objective function. This

model's objective function included a composite PMP objective function that represents both covering and median objectives, which preemptively maximizes coverage over weighted distance. This model formulation required less computational resources in comparison to using the *ε-constraint* method. Another challenge was that solution times with the *ε-constraint* method can increase dramatically if it is difficult to find an initial feasible solution. This was our experience when testing RC-QMALP prototypes where many hours passed without the MIP solver finding any feasible solutions. To address this issue, we reformulated the model into a multiobjective model with both objectives included in the objective function but with the weight biased for the primary coverage objective. This drastically reduced computation times from hours to seconds in some cases.

In this section, we have reviewed most of the models used to develop RC-QMALP and several important results including their problems and limitations. We began with the foundational deterministic model before moving on to more sophisticated stochastic and probabilistic models that were developed to capture system congestion. This review concluded with a discussion of various multiobjective modeling approaches.

After presenting the fundamental location models, we discussed the issue of capacity at length because the underlying motivation for modern EMS system models is making the best use of limited resources. The capacitated deterministic location models we reviewed represented a natural extension of the fundamental uncapacitated location models. However, in the stochastic and probabilistic location model review, we showed that in addition to concerns of capacity, there are questions and concerns about system congestion that are related to system capacity but cannot be readily addressed with deterministic models, namely the availability of servers. Then, we delved into four classes of stochastic and probabilistic models

(reliability-, districting-, multiobjective- and location-allocation-based models), their general structures, and limitations. Here, we discussed two critical issues associated with the first two model classes, respectively, the challenges with determining and satisfying reliability constraints and modeling intradistrict cooperation or non-cooperation (as well as the implications or limitations of each approach).

In the third subsection, we presented and discussed several essential models in a detailed manner along with the previously raised issues although in a more thorough manner and in a more specific context. With the QPMP, we discussed how this model could be integrated into RC-QMALP framework and how the server-demand node assignment decision variables can be interpreted. The presentation of the MEXCLP served to introduce the concept of systemwide busyness fractions and how they are calculated as this information is critical to understanding MALP. MALP effectively relaxed some of the limitations of systemwide busyness fractions with local-reliability calculations but also shifted to a service reliability-oriented version of the MCLP. For MALP, the discussion focused on its assumptions as many researchers have questioned their validity as well as that of the MALP solutions with respect to the service reliability that they promise to provide. Likewise, although QMALP relaxes some of MALP's assumptions with the uses of queueing theory, many works have challenged QMALP on similar grounds with respect to the validity of its assumptions and solutions. With QMCLAP, the focus of this review was to expand the discussion on server-demand node assignment variables that have been used in stochastic locations models. QMCLAP and the PLASC*m* are designed to model immobile facilities rather than mobile facilities. However, making this distinction raises the question as to how such a distinction is justified. Ultimately, this leads to dynamic and static server-demand node assignments and how this might be a more

significant or proper system property to consider when modeling an EMS system. Lastly, we discussed multiobjective models, particularly problems with ranked objectives. We presented the ε-constraint method for formulating and solving these problems but also discussed alternative methods that have desirable computational requirements.

Moving forward the two most significant challenges are: (1) adequately capturing the system network interactions, that is intra- and inter-district/neighborhood server cooperation, and (2) ensuring the validity of RC-QMALP with respect to the reliability constraints. Capturing system network interactions is persistently a challenge for location modelers because it requires making *a priori* estimates about the system that might not be consistent with the *ex post* system behavior. Using server assignments with a QMALP location model is a promising avenue for addressing this issue but the value of this approach needs to be assessed particularly in terms of how to interpret assignment decisions. The second challenge is an extension of the first, however, it is important because RC-QMALP is a reliability-based model and thus, it's integral that the reliability constraints hold. However, complex system network interactions and interdependencies complicate efforts to mathematically validate location models (without at least making unrealistic assumptions or decisions) and so with RC-QMALP the focus is on whether it can generate consistent predictions about system performance.

4. The Resource Constrained Queuing Maximum Availability Location Problem

RC-QMALP is an effort to improve QMALP by addressing issues related to intra- and inter-neighborhood server interactions, the validity of reliability constraints, and the location of servers within a neighborhood. RC-QMALP maintains much of the QMALP model, however, there are three key additions that address these issues.

First, RCQMALP includes assignment decision variables that track server workload assignments as well as *server idle capacity* (a server's unassigned capacity). As discussed above, many location models explicitly include workload assignment variables (e.g., Heller, et al., 1989) in order to account for server capacity as well as to determine whether a demand node can be served. In contrast, server idle capacity is mostly addressed indirectly through objectives and constraints on waiting times and queue lengths (Berman & Krass, 2001; Marianov & Serra, 1998), workload constraints (e.g., Neebe, 1978), or workload balance objectives (e.g., Weaver & Church, 1981). Models typically include these constraints or objectives as a way to improve system performance, encourage equitable or well distributed workloads, or to simply to conform to system requirements or goals. With RC-QMALP these goals are at most secondary. Rather, these assignment variables are included to help determine the aggregate availability of service or server capacity within a demand node's neighborhood. In QMALP, b_i servers must be located in a demand node i's neighborhood (N_i) in order to satisfy the α-reliability service constraints. However, as noted above, it's possible that the b_i servers have sufficient commitments to demand nodes outside N_i such that more than b_i servers are required to provide α-reliable service in N_i. Conversely, it's possible that servers outside of N_i serve demand nodes inside N_i such that less than b_i servers are required to provide α-reliable service in N_i. Thus, we address the inter-neighborhood cooperation issue and relax QMALP's districting assumption by tracking both server workload and idle capacity allocations as server workload assignments alone do not indicate whether enough server capacity is present in a demand node's neighborhood.

Second, RCQMALP includes several new constraints to accommodate the location-allocation model approach and to bolster the QMALP reliability constraint. The former

includes supply and demand constraints for, respectively, servers and demand nodes while the latter accounts for allocations of capacity at the neighborhood and inter-neighborhood level.

Third, RC-QMALP includes a second QPMP-like objective function (QPM-O). We hypothesize that this second objective function will (1) result in superior solutions (in terms of reliability) and (2) help reduce computation times. Regarding the first, as previously discussed, QMALP offers no guidance as to where to locate servers within neighborhoods and thus QMALP solutions might fail to include attractive locations or site configurations. For example, a QMALP solution might include server locations in less congested areas of a neighborhood despite the availability of sites in more congested areas (of course assuming that the alternative solution is equally fit). Thus, we expect to generate improved solutions by including this second objective function. Regarding the second hypothesis, we expect faster computation times by including this objective function (rather than omitting it). Again, we expect to produce better solutions with the second objective function but also that the second objective function helps to reduce the solution space and helps eliminate alternative solutions more quickly when involving a branch and bound algorithm.

Finally, before introducing the RC-QMALP formulation it's important to address a common criticism of MALP-type models. As noted above, some publications have raised doubt about the usefulness of this this modeling paradigm. For instance, Erkut et al. (2008) argue that local-reliability objectives in MALP-type models do not coincide with the goals of EMS practitioners. This is a fair criticism although a recent publication by van Buuren, van der Mei, & Bhulai (2017) indicates that there is value in this modeling paradigm. They explain that local governmental figures or organizations frequently demand that their sub-regional districts receive adequate levels of service. Consequently, they report a shift in interest away

from aggregate coverage models and an increasing interest in maximum availability and minimum reliability models due to local intraregional politics (e.g., a mayor demanding better coverage for his/her city). Their study is based out of the Netherlands; however, it raises the point that the widespread adoption of certain EMS goals or objectives should not preclude consideration of other approaches.

4.1 Model Formulation

As with QMALP, RC-QMALP maintains the 0-1 binary decision variable Y_i that indicates whether b_i facilities (required for α-reliable service) are located in demand node i's neighborhood. We note that this set of decision variables only track a single level of coverage (b_i) whereas QMALP's formulation included decision variable for all levels of coverage up to b_i. Likewise, the total number of facilities that can be located are limited to p where X_j decision variables track locational decisions for each site j . In RC-QMALP server co-location is not allowed and hence every X_j is a 0-1 binary decision variable.

RC-QMALP's allocation framework includes two sets of decisions variables for server workload and idle capacity assignments. Workload is assigned from each demand node to located servers and this is tracked with continuous, non-negative Γ_{ij} decision variables. In contrast, server idle capacity is assigned from located servers to demand nodes and these assignments are tracked with continuous, non-negative Φ_{ji} decision variables. Servers cannot accept more work than they can handle and demand nodes cannot be assigned more demand than they generate (on average). Thus, the total server workload from a server is limited to the server's capacity (μ_j) while on the demand node side they are limited to the demand node's intensity (λ_i). Likewise, a server's idle capacity assignments are limited by its capacity slack

(μ_j - $\sum_{\forall i \in I} \Gamma_{ij}$). However, in RC-QMALP there is no limit as to how much server idle capacity can be assigned to any demand node. Lastly, the travel times associated with both server workload and idle capacity assignments are included in RC-QMALP's second *p*-median oriented objective function.

As for determining the reliability of service, the local-reliability and b_i calculations from QMALP are applied in RC-QMALP. Thus, the assumption that each demand node's neighborhood N_i operates as a *M*/*G*/s/*s*-*loss* system. Moreover, it's assumed that the server's capacity parameters (μ_j) include travel time.[93]

Model

[93] We discuss the implications of this and other assumptions in *Section 5.3*.

$$(RCQ-O1) \quad Maximize\ Z^1_{RCQMALP} = \sum_{i \in I} d_i Y_i$$

$$(RCQ-O2) \quad Minimize\ Z^2_{RCQMALP} = \sum_{i \in I} \sum_{j \in J} t_{ij} \Gamma_{ij} + t_{ji} \Phi_{ji}$$

$$(RCQ-C1) \quad \sum_{\forall j \in N_i} \Gamma_{ij} \leq \lambda_i ; \ \forall i \in I$$

$$(RCQ-C2) \quad \sum_{\forall i \in N_j} \Gamma_{ij} + \Phi_{ji} \leq \mu_j X_j ; \forall j \in J$$

$$(RCQ-C3) \quad \sum_{\forall k \in M_i} \sum_{\forall j \in N_k} \Gamma_{kj} \geq \lambda^i Y_i ; \forall i \in I$$

$$(RCQ-C4) \quad \sum_{\forall j \in N_i} X_j \geq b_i Y_i ; \forall i \in I$$

$$(RCQ-C5) \quad \sum_{\forall k \in M_i} \sum_{\forall j \in N_k} \Gamma_{kj} + \sum_{\forall k \in M_i} \sum_{\forall j \in N_i : k \neq j} \Phi_{jk} + \Phi_{ii} \geq b_i Y_i ; \forall i \in I$$

$$(RCQ-C6) \quad \sum_{j \in J} X_j = p$$

$$(RCQ-C7) \quad X_j \in \{0,1\} ; \forall j \in J$$

$$(RCQ-C8) \quad Y_i \in \{0,1\} ; \ \forall i \in I$$

$$(RCQ-C9) \quad \Gamma_{ji} \geq 0 ; \forall j \in J, i \in N_j$$

$$(RCQ-C10) \quad \Phi_{ji} \geq 0 ; \forall j \in J, i \in N_j$$

Notation

Indicies and Sets

$I =$ set of demand nodes.

$J =$ set of potential facility locations.

$i =$ index of demand nodes, $i \in I$.

$j =$ index of potential facility locations, $j \in J$.

$N_i = \{ j \mid t_{ji} \leq s \}$ - the set of facility locations j in the neighborhood of demand node i.

$M_i = \{ l \in I \mid t_{li} \leq s \}$ - the set of demand nodes l in the neighborhood of demand node i.

Parameters

t_{ji} = shortest travel time from site j to site i.

s = maximal travel time standard [time unit].

p = total number of facilities to be located.

$\bar{p}_i$ = parameter for number of servers in demand node i's neighborhood, $\bar{p}_i \in \square$.

μ_i = mean rate of service completion per unit of time of a server located a demand node i.

d_i = poupulation at demand node i.

λ_i = call intensity per unit of time at demand node i.

λ^i = call intensity per unit of time in demand node i's neighborhood (N_i).

$$\lambda^i = \sum_{k \in M_i} \lambda_k$$

ρ_i = the utilization rate at demand node i.

$$\rho_i = \frac{\lambda^i}{\mu_i} \; ; \rho_i \leq 1$$

P_k^i = the probability of the system at demand node i being in state k.

$$P_k^i = \frac{\frac{1}{k!}\rho_i^k}{\sum_{n=0}^{k} \frac{\rho_i^n}{n!}}; k = 1, ..., \bar{p}_i$$

$$\mathrm{b}_i = \arg\min_{b_i} \; f(b_i) := \{b_i \mid b_i \in \square^+ : P_{b_i}^i \leq 1 - \alpha\}$$

Decision Variables

$\forall j \in J$:

$$X_j = \begin{cases} 1, & \text{if a server is located at facility j,} \\ 0, & \text{otherwise.} \end{cases}$$

$\forall j \in J, i \in N_j$:

Γ_{ji} = workload from demand node i is assigned to a server at location j, $\Gamma_{ji} \geq 0$,

Φ_{ji} = service capacity from assigned

4.2 Model Components

As with the QMALP objective (QMA-O), the first objective of RC-QMALP (RCQ-O1) maximizes the population that is covered with α-reliability. The second RC-QMALP objective (RCQ-O2) is minimizes the travel times between servers and the assigned locations of their

workload along with the assigned idle capacity assignments. To implement both objectives, we planned to use the ε-constraint method with (RCQ-O2) as our objective function and (RCQ-O1) as a constraint bounded below by the optimal solution to RC-MALP. This approach turned out to be computationally intensive as most problem instances required hours to solve. To understand this issue, we analyzed the model statistics as the model solved and observed that the solver struggled to find a feasible solution to the combined problem. Consequently, we abandoned the ε-constraint method in favor of a multi-objective minimization model with both (RCQ-O1) and (RCQ-O2) in the objective function. The resulting new objective was:

$$(RCQ-O3) \qquad Minimize\ Z^{3}_{RCQMALP} = \sum_{i\in I}\sum_{j\in J} t_{ij}\Gamma_{ij} + t_{ji}\Phi_{ji} - M * \sum_{i\in I} d_i Y_i$$

Recall that (RCQ-O1) is maximized and thus in (RCQ-O3) we make this objective negative so that it corresponds with the minimize objective. Moreover, to make certain that the optimal (RCQ-O1) value is generated with (RCQ-O3), we scaled (RCQ-O1) with a very large value $M \in \mathbb{R}^{+}$ such that (1) minimizing (RCQ-O1) is absolutely prioritized over (RCQ-O2) and (2) the solution is Pareto optimal over both objectives.

Constraints (RCQ-1) and (RCQ-2) constrain server workload and idle capacity assignments. Constraints (RCQ-1) limit the maximum workload assignments to every demand node to its demand intensity and constraints (RCQ-2) limit the total workload and idle capacity assignment for every located server to its service capacity.

Constraints (RCQ-C3) through (RCQ-C5) set the requirements for establishing that a demand node has access to α-reliable service. Note that the RHS of these constraints contain the Y_i decision variable multiplied by a parameter. These parameters represent requirements that must be met to establish α-reliability at demand node i as $Y_i = 1$ only when the LHS is

larger than the RHS's parameter value. For constraints (RCQ-C3) the LHS represents the total server workload assigned to demand node *i*'s neighborhood (N_i), or the amount of fulfilled demand in N_i. Thus, one condition to establish the α-reliability service for a demand node is that all demand in the demand node's neighborhood is fulfilled. Note that the LHS sum accounts for workload assignments from servers inside and outside N_i. Constraints (RCQ-C4) are similar to the α-reliability constraints of MALP2 (MA2-C1 and C2) and QMALP (QMA-C1 and 2), although in RC-QMALP these constraints are reduced to a single set of constraints. These "physical facility" constraints require that the total number of facilities located in N_i (the LHS) exceed b_i to establish α-reliability service at each demand node *i*. Constraints (RCQ-C5) represent the server capacity constraints that require that the total server capacity available in N_i (the LHS) needs to equal or exceed the capacity of b_i facilities to establish α-reliability service at each demand node *i*. However, we note that while source of server workload is not restricted (so long as the assignments conform to the specified service time/accessibility standards), there are two restrictions in accounting for server idle capacity for each demand node. First, for server idle capacity assignment to count for a demand node they must originate from a located server and be assigned to demand nodes that are both accessible to that demand node (that is only assignments Φ_{jk} where $j \in N_i$ and $k \in M_i$). Second, if server idle capacity is "self-assigned" that is assigned from a server at site *j* to demand node *k* where $j = k$, then this assignment Φ_{jj} (or Φ_{kk}) is only factored towards establishing the α-reliability constraint of demand node *k*. The first restriction attempts to ensure that server idle capacity is available to a demand node rather than only to its neighbors[94] while the second restriction attempts to

[94] This prevents, for example, situations where server idle capacity is assigned by a server located outside a demand node *i*'s neighborhood to a demand node near the border of demand node *i*'s neighborhood.

address double counting of server idle time when servers are located at the intersection of two demand node neighborhoods (that is, facilities j such that $j \in \{N_i \cap N_k\}$ for demand nodes i and k, $i \neq k$). Finally, constraint (RCQ-C6) limits the total number of facilities to p and constraints (RCQ-C7) to (RCQ-C10) define the domain of our decision variables.

4.3 Discussion

In the previous chapter, we identified two main challenges for modeling ambulance systems: adequately capture interactions and ensure the validity of the reliability constraints. With RC-QMALP, we address both issues with a location-allocation framework that tracks the allocation of server workloads and idle capacity. Given this fundamental change, it's necessary to address new concerns and revisit the assumptions of the essential models used to develop RC-QMALP.

4.2.1 Workload Assignments

The first substantial issue is the significance of the assignments. In RC-QMALP, servers are assigned demand and demand nodes are assigned server idle capacity. In establishing these assignments, the intention isn't to require that these assignments actually manifest themselves through a districting or dispatching policy. Rather, they serve primarily as an accounting method of sorts for server capacity and thus, one should proceed with caution when analyzing or interpreting any assignment values. It's possible to use these assignments to develop districting or dispatching policies, however, no dispatching or districting policies are explicitly assumed in RC-MALP. Admittingly, there is an implicit assumption that servers are more likely to serve demand nodes that are closer (and busier) than those that are father away (and less busy). In location modeling, such a limitation is not unusual but rather the norm even when

considering alternative approaches. As previously discussed, Hypercube models and different approximations provide highly descriptive output measures but the models require computationally intensive MSC calculations to relax the unrealistic assumption that service times are exponentially distributed. Likewise, Jarvis's (1975) simplified Hypercube model reduces the computational burden associated with MSC calculations but it requires balanced workloads among servers, which may or may not be a reasonable condition. With more prescriptive models, works such as Swoveland et al. (1973b) and Weaver & Church, (1985) are based upon a "stability hypothesis" regarding the distribution of service performed by the k^{th}-closest facility. Goldberg et al. (1990) computes "optimal" fixed preference schemes but found substantial differences between the dispatching predicted by the optimization model and the "actual" dispatching from their validation model. Heller et al. (1989) also encountered issues when validating a deterministic PMTP location model's workload capacity constraints using simulation. They noted that the dispatching policy in their simulation model did not consider their model's workload balancing objective/constraints and that situations with binding workload constraints would prove problematic given the stochastic nature of their simulation model. In all, they generated superior approximations of server workload with their model solutions (compared to PMP solutions) but reported that their model underestimated the simulated maximum server workload values.

Despite such underwhelming results from this previous work, there are several promising results and insights. For instance, although the fixed preference schemes generate by Goldberg et al.'s (1990) model were not practically useful, the generated model solutions improved system performance in terms of balancing workloads and improving on-time response rates. Moreover, they found that the discrepancy between the predicted and actual dispatching

operations is partly attributable to server preference ties resulting from equidistantly located servers (that is to a demand zone). They did not observe this issue in situations with low vehicle utilization rates but when the problem appeared they altered the zone sizes to prevent ties. Likewise, Heller et al. (1989) reported a consistent relationship between their PMTP's workload-oriented objective function and the simulated workloads and suggested that the PMTP showed great promise within a "multiobjective context." Moreover, Heller et al. (1989) suggest that optimal PMTP *location* configurations are robust with respect allocations and availability in the presence of alternative optima for allocation decisions.[95] In other words, while their PMTP model might not produce optimal workload allocations, it does not preclude the generation of an optimal location solution that can accommodate an optimal workload allocation. In all, the implication for RC-QMALP is that the assignment decisions are important for capturing the system interaction and although the actual assignment values can be important, they are not as integral to the solution as the locational component.

4.3.2 *Server Idle Capacity Assignments and the Queue Systems*

For RC-QMALP, we previously explained that both server workloads and idle capacities are considered in order to determine that there is at least enough server capacity equivalent to the capacity of the b_i servers that are required to α-reliability serve each demand node *i*. Such workload and idle level assignments allow server cooperation by allowing a neighborhood's workload to be handled by an outside server. In support of this proposition (beyond an intuitive explanation), we note that by requiring that the total/all demand in a neighborhood to be served

[95] This is within the context of Heller's (1985) analysis of capacitated location-allocation systems where she showed that optimal location configurations do not necessarily require unique allocations.

plus the total service idle capacity exceed b_i we are effectively limiting the server utilization rate of the neighborhood. As such, even if a server in the local neighborhood assists an outside demand node, there must be enough server idle capacity within the neighborhood to compensate for this external transfer.

To see this, we note that b_i calculations consider the total demand in a neighborhood (λ^i) and assume that k servers are completely available. Thus, we use the following utilization rate (ρ_i) to calculate b_i (where k is determined by α):

$$\rho_i = \frac{\lambda^i}{k\mu_i} \; ; \rho_i \leq 1$$

Because λ^i is fixed and we required that all λ^i be assigned to establish α-reliability, we only need to concern ourselves with the denominator or server capacity (server workload and idle capacity). Clearly, if all server workload and idle capacity remains within a neighborhood there is no issue in terms of assigned capacity, but if a server helps an outside demand node, then there must remain enough server idle capacity to meet the b_i capacity requirement. However, if there is an outside server assisting with a neighborhood's demand, *ceteris paribus*, it is not clear if this situation is equivalent to that of an interior server handling the workload. Unfortunately, the answer to this question depends on various factors, namely the service times distribution, the system queueing capacity, and the independence of service times.[96] To show this, let $L(k, \rho)$ be the system loss (e.g., dropped calls) that is a function of the number of servers ($k_i \in \mathbb{N}^+$), the demand arrival rate (λ_i), the service rate(μ), and utilization rate ($\rho = \lambda / \mu$)

[96] These results are as presented in (Smith & Whitt, 1981).

, It's well known that the function $L(k,\rho) = \lambda * B(k,\rho)$ where B is the Erlang blocking function as previously defined[97]:

$$B(k,\rho) = \frac{\frac{1}{k!}\rho^k}{\sum_{n=0}^{k}\frac{\rho^n}{n!}}$$

If we assume a *M*/*M*/*K*/*K*-*loss* system that can be divided into two M/M/k/k-loss systems ($K = k_1 + k_2$ with demands λ_1 and λ_2),[98] then we have the following inequality:

$$L(k_1 + k_2, \lambda_1 + \lambda_2, \mu) \leq L(k_1, \lambda_1, \mu) + L(k_2, \lambda_2, \mu)$$

This implies that a combined system of servers loses *at most* the same amount of calls as compared to two systems operating in parallel. Moreover, it can be shown that $B(t*k, t*\rho)$ strictly decreases with t. However, if we consider a *M*/*G*/*K*/*K*-loss system we have $L(k_1 + k_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(\rho_1 + \rho_2))$ where a similar inequality does not apply as it can be shown that in some cases:

$$L(k_1 + k_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(\rho_1 + \rho_2)) \geq L(k_1, \lambda_1, \mu) + L(k_2, \lambda_2, \mu)$$

In other words, sometimes two parallel sever *M*/*G*/k_i/k_i-*loss* systems lose less calls than a combined *M*/*G*/*K*/*K*-loss system. Smith & Whitt (1981) suggest that this is likely when the server systems have substantially different service times.

---

[97] It was previously defined as P_i^k for QMALP, etc.

[98] Here we assume that each system can only serve arrivals from their own system (i.e., no server cooperation). In the context of neighborhoods, the servers inside the neighborhood comprise one system and the external servers comprise another.

With QMALP, there is an assumption that there is a balance of intra-district server cooperation and thus, these results further cast doubt on the suitability of this assumption. This is also the case with RC-QMALP, however, it is far less concerning because of the additional server capacity constraint.

In any case, these inequalities are useful in analyzing local interactions. Unfortunately, questions remain about the behavior of the system as a whole, that is, whether the queue system is stable. With queueing systems, the most critical consideration is that the utilization rate (ρ) does not exceed 1. This is particularly important for buffered systems (e.g., *M*/*G*/*K*/∞ queues) as with $\rho > 1$ queues become *unstable* or they almost surely grow. In contrast, with unbuffered systems (e.g., *M*/*G*/*K*/*K*-*loss* queues) as the utilization rate grows it is less likely to find the system in a state where one or more servers are available.

Baron et al. (2009) present an interesting perspective on queue stability in their analysis that involves restricted inter-district cooperation (i.e., calls can only be served by servers in demand node's neighborhood). These types of problems are a form of a Multi-Class Multi-Server Queueing (MCMSQ) System with partially accessible queues (PAQ)[99] whereas systems where all servers are accessible to all customers are fully accessible queue (FAQ) systems. In this paper, Baron et al. (2009) propose two location set covering problems with stochastic demand and congestion and PAQs. Although they limit their analysis to *M*/*M*/*K*/∞ systems and explore decoupling systems into a set of PAQ systems, there remains the question of stability. Caldentey and Kaplan (2007) have proved that an MCMSQ system with *M*/*M*/*K*/∞ queues is stable if and only if:

[99] We refer to this article for references about MCMSQ systems.

$$\sum_{i \in V} \lambda_i < \sum_{s \in S(V)} \mu_s; \forall V \subset N$$

where N is the set of all customers, λ_i is the arrival rate of customer type i, S is the set of servers ($s \in S$), μ_s is the service rate of server s, and $S(V)$ is the set of all servers accessible to the customers in subset $V \subset N$. This property holds under any *work conserving discipline* (servers cannot be idle if there is an unserved customer in the system and servers cannot terminate a job with a customer before completing the job) which include the FIFO discipline. This approach is limited however, in that it assumes an exponential service time distribution but more importantly $2^{|N|}$ subsets need verification to establish queue stability. Nonetheless, this is a serious concern with all MALP-based models, This is especially true when dealing with uncovered demand, that is, demand nodes without a server in their neighborhood. In QMALP, this is handled by what amounts to a PAQ system as uncovered demands are ignored since they are not factored in any part of the model. Admittingly, this is also an issue with RC-QMALP although the server capacity constraints represent an attempt to promote local queue stability by considering the demand node neighborhood subsets. Another option enabled by RC-MALP's location-allocation framework is to factor all demand directly in the model but this goes beyond the scope of this thesis (although we revisit this issue in the discussion).

The final issue of importance involves the independence of service times. It is a well-known result that the blocking probabilities in *M*/*G*/*K*/*K-loss* depend only on the mean service time (Burman, 1981). Models such as QMALP and RC-QMALP use these queue systems due to their flexibility in this respect. However, Singer & Donoso (2008) and others have observed that the spatial distribution of servers and demand play a critical role in the suitability of using queues in location models. Despite the insensitivity of *M*/*G*/K to service time distributions in

affecting the mean (e.g., blocking probabilities) they are sensitive to any system delays or other issues that significantly alter the mean service times. One notorious cause is the unavailability of the nearest facility (to an incident) which might require dispatching a more distant unit (and hence increasing average travel times). This issue of *geographical dispersion* (Delasay et al., 2015) is most pronounced in congested buffered queue systems (modeled or otherwise), but can also be an issue with unbuffered FAQ systems. Aside from affecting mean service times, another assumption with *M*/*G*/K queues is that the service times are identically and independently distributed. Geographical dispersion can certainly conflict with this assumption but demand side issues (e.g., emergencies requiring multiple ambulances) can also pose some problems. On the supply side, one "solution" is to employ PAQs to limit geographical dispersion while others have proposed "adjustment" or correction factors (e.g., Batta, Dolan, & Krishnamurthy, 1989) to account for server dependence. The first option is interesting but might conflict with EMS response policies while the later introduces non-linearities that are not compatible with the mathematical programming approach employed here. With RC-QMALP this is admittingly an open issue as it is assumed that travel times are included in the total service time, however this worthy problem is beyond the scope of this work.

4.3.3 *Impact of Median Objectives*

Another facet of RC-QMALP's location-allocation component is the PMP-like objective function. As previously discussed, by adding this objective: (1) we expect to improved computational performance; (2) we hope to generate "reasonable" server workload and idle capacity assignments between servers and demand nodes that tend to be close assignments as compared to something farther away; and (3) we expect that chosen locations will be close to

areas with high demand. It remains to be seen, however, whether these elements will improve system performance with respect to both total and reliable coverage.

Encouraging close or closer assignments may help in creating realistic estimates of workload and whether reliable coverage levels can be met. But, some have questioned such assignment policies. For example, Carter *et al.* (1972) has challenged nearest server dispatch policies by showing that these policies are suboptimal in some cases, notably where there are large variations in demand over short distances. Likewise, Berman & Mandowsky (1986) also report that system performance becomes increasingly more sensitive to location and allocation decisions as congestion increases and that optimal facility locations in cases of high demand are not intuitive with respect to "popular median-proximity" location-allocation policies. However, the non-cooperative districting approach used in both models limits the general applicability of these results. Larson & Odoni (1981) note that cooperative server systems have more balanced workloads than districting systems. This is important as without thresholds on blocking probabilities it can be shown that a system of M/G/*K*/*K*-*loss* queues [100] (with homogenous service rates) optimize throughput (i.e., minimized blocked calls) when workload is evenly distributed among the systems (Yao & Shanthikumar, 1987).

Berman, et al. (2007) provide another interesting perspective in the context of unreliable facilities where they note that facilities tend to become more centralized in order to accommodate disruptions while *p*-median models tend to "spread" out facilities in order to minimize travel costs. Likewise, Church, et al. (2004) also provide some insight with their

[100] Note that the queue systems must all have the same number of servers.

PMP (and MCLP) based model where r of p facilities are expected to be interdicted, however, the $p - r$ remaining facilities must continue to serve the system. Thus, the objective is to locate p facilities such that, respectively, the average travel distance is minimized (or the total coverage is maximized) upon the removal of the r facilities that result in the maximum increase in average travel distance (or the maximum decrease in coverage). For both models, they showed that robust solutions for the interdiction models had significantly lower objective values than their non-interdiction model counterparts which raises the possibility that of the p-median objective alone produces inferior solutions.

In any case, these results and insights are highly interesting but difficult to apply to RC-QMALP. First, the MALP-family of models is rather unique in that despite being a coverage-based model, it promotes centralization (due to the b_i requirements). As the number of facilities increase, the extent of coverage should grow to cover more demand, however, these additions to coverage can be expected when there are sufficient servers to support the expansion. Second, while p-median objective is included in RC-QMALP and should encourage a "spread" in locational configurations, the model is subject to the QMALP objective which will tend to cluster facilities in order to meet alpha reliability constraints. Third, workload balancing constraints are not implemented within RC-QMALP and so it is not clear how the p-median objective might affect workload. Again, the b_i requirements should assist in balancing workloads but this is not clear to what extent any of these factors might affect system performance.

5. Results and Analysis

With RC-QMALP we expect to improve the reliability estimates predicted by the optimization model as compared to reliability estimates of MALP and QMALP. We expect RC-QMALP to outperform both models with respect to aggregate on-time response coverage.

These hypotheses effectively amount to an *operational validation* of the location optimization models which involves determining how the optimization location model outputs (the location configuration and its associated α-reliability objective value) correspond to the system they represent[101] (Sargent, 2005). This process can be subjective or objective where the former type relies on exploring the model behavior (e.g., using parameter variability-sensitivity analysis) and graphical instruments (e.g., graphs and charts) while the latter involves statistical tests and procedures. In any case, operational model validity is consistent with providing a high degree of confidence in the model's output within its domain and with respect to range of accuracy required by the model's purpose or application (Sargent, 2005; Schlesinger et al., 1979)

In this thesis, we use the optimization-simulation approach employed by Sorensen & Church (2010) to compare MEXCLP-LR with MEXCLP and MALP along with objective and subjective approaches. They generated optimal location solutions using each model and then tested each solution using an ambulance simulation program. They assessed the three models by tabulating the instances where each model uniquely (and jointly) produced the best solution according to the simulation model and compared the deviations between the α-reliable

[101] The system and its behavior can described by empirically collected data or generated by another model (Aboueljinane et al., 2013).

coverage of the best location configuration and the corresponding solution of each model for each problem instance. Moreover, they performed a *t*-test on a comparison of the aggregate on-time response coverage of MEXCLP and MEXCLP-LR.

Validating location models with simulations is not a new idea (Ignall, Kolesar, & Walker, 1978) and several works have validated optimization location models using simulation (see Aboueljinane et al., 2013) and but as Sorensen & Church (2010) note, these analyses "[appear] to be the exception rather than the rule."[102] Comparing models strictly through their objective function values and over several parameters is a great first step, however, this approach does not assess the operational validity of the model particularly with respect to its assumptions (as with QMALP and the districting assumption). Hypercube-based models are often used to validate optimization and heuristic models (e.g., Erkut et al., 2009) given their highly descriptive nature. However, these models are also based on assumptions that can't be readily assessed.[103]

5.1 Experiments

For our experiments, we programmed MALP2, QMALP, and RC-QMALP on FICO's *Xpress-IVE Version* 1.24.06 64 bit using *Xpress Mosel Version* 3.8.0 and solved with *Xpress Optimizer Version* 27.01.02. We used a computer equipped with an Intel i7 3370K with 8 GB of RAM. We established our problem instances for MALP 2, QMALP, and RC-QMALP based

[102] Aboueljinane et al.'s (2013) review of simulation models applied to EMS operations listed less than 10 works related to mathematical programming-based ambulance deployment models. This list appears to be incomplete, however.

[103] For instance, Jarvis's (1985) hypercube approximation assumes exponential services times and does not consider queues. Embedding a hypercube model within a location model can also be problematic (see Chiyoshi, Galvão, & Morabito, 2003).

on Sorensen & Church (2010) with some additional parameter values to test various elements of the models.

Model Parameter Dimensions & Values	
Total system call-volume	2 and 4 calls per hour (CPH)
City diameters (mins)	16, 24, 32
Total number of servers	4, …, 15
Response standards (mins)	6, 8, 10
Total service time (mins)	60 [fixed]
α-Reliability Standards	0.80, 0.85, 0.90, 0.95, 0.99

Table 1 - Location Model Parameter Dimensions & Values

The two call-volume levels represent high and low call-volume scenarios. We used Swain's (1971) 55-node network dataset and adjusted call intensity proportionally according to the demand levels at each demand node.[104] We also scaled this network dataset to three city diameter values. We set a maximum number of servers at 15 as at this point all models generated solutions with complete α-reliable coverage (Z_{Model} = 640). The total service time was fixed at 60 mins due to the software limitations (we discuss this below). Although this is limiting in some respects, generally distributed service times are assumed in RC-QMALP and thus, this does not create a conceptual model validation issue. This is also a standard used in

[104] Our work is admittingly limited by relying exclusively on Swain's (1971) data. However, this dataset is useful as this is a classic dataset in location science modeling and this allows us to estimate the performance of RC-QMALP with other works. Nonetheless, in future works we expect to consider a greater number of datasets.

the industry when a service call includes a patient transport. Five α-reliability standards are considered and thus a total of 1,080 problem instances were solved for each model

To evaluate the model solutions, we used the same simulation program described and used in Sorensen & Church (2010). The general structure of the system is such that (1) all calls have equal priority, (2) service times (travel and on-scene time) are constant, (3) the nearest available server is dispatched to calls, (4) calls are placed in a FIFO queue if all servers are busy, and (5) servers return to their home location before responding to a new call. Each problem instance involved simulating 10,000 calls (the software maximum). The simulation software tracks and reports information about the total demand served with the specified time standard and the reliability of service at each demand nodes.

5.1.1. Comparing MALP, QMALP, and RC-QMALP

To compare the three models, we first tabulated the instances that each model outperformed or tied other models, the number of instances where each model produced a solution within three thresholds (1%, 2%, and 5%) of the best solution, and summary statistics about how the simulated results of each solution deviated from predicted solution values (the model objective), how they deviated from the best overall solution, as well as the computational times of each model.

We initially considered follow the Sign Test approach used by Murray & Church (1992) to compare the simulated total and α-reliable coverage of each model as well as to assess the operational validity of RC-QMALP by comparing the predicted and simulated α-reliable coverage. However, we reconsidered this decision upon reviewing the models results as it became clear that our questions require the development of a more proper simulation

experiment (Barton, 2013), particularly in regards to the theoretical components of our questions (Davis, Eusebgardt, & Binghaman, 2007).

5.1.2 Assessing RC-QMALP's Server Cooperation Constraints (RCQ-C5)

Constraints (RCQ-C5) are an integral part of how inter-district cooperation is handled in RC-QMALP. As previously discussed, the constraint is formulated to discourage the double-counting of idle server capacity. However, the efficacy of this formulation is not clear as to whether inter-district cooperation can be effectively handled with a more relaxed constraint. As such, to test these constraints we replaced them with four alternative constraint sets that increasingly relax constraints (RCQ-C5). For each alternative constraint set, we define a new version of RC-QMALP:

Class (C) - RC-QMALP+ILA[105]

$$(RCQ-C2A) \quad \sum_{\forall i \in N_j} \Gamma_{ij} + \Phi_j \leq \mu_j X_j; \forall j \in J$$

$$(RCQ-C5A) \quad \sum_{\forall k \in M_i} \sum_{\forall j \in N_k} \Gamma_{kj} + \sum_{\forall j \in N_k} \Phi_j \geq b_i Y_i; \forall i \in I$$

Class (A) - RC-MALP+FLA

$$(RCQ-C5B) \quad \sum_{\forall k \in M_i} \sum_{\forall j \in N_k} \Gamma_{kj} + \sum_{\forall k \in M_i} \sum_{\forall j \in N_i} \Phi_{jk} \geq b_i Y_i; \forall i \in I$$

Class (B) - RC-QMALP+FLA+OIC

$$(RCQ-C5C) \quad \sum_{\forall k \in M_i} \sum_{\forall j \in N_k} \Gamma_{kj} + \sum_{\forall k \in M_i} \sum_{\forall j \in N_k} \Phi_{jk} \geq b_i Y_i; \forall i \in I$$

[105] In RC-QMALP-ILA constraints (RCQ-C2) are replaced with constraints (RCQ-C2A) as only self-assigned server idle capacity is allowed.

Class (D) - LA-QMALP+DR

($RCQ-C5D$) is removed and not replaced.

With constraints (RCQ-C2A) RCQMALP+ILA (immobile local assignment), self-assignment of server idle capacity is allowed and can count towards the reliability requirement of any demand node in the same neighborhood. However, it can only remain at the server's location. In RC-QMALP+FLA (flexible local assignment), constraints (RCQ-C5B) also allow self-assigned server idle capacity but assignments to other local/neighborhood demand nodes is allowed. The OIC (outside idle capacity) extension in RC-QMALP+FLA+OIC allows idle capacity from outside to demand node's neighborhood to count towards it server capacity requirement. Finally, when (RCQ-C5) is removed from RC-QMALP the resulting model is effectively a location-allocation version of QMALP with the added restriction that establishing α-reliable coverage in a demand node's neighborhood requires accounting for all demand in that neighborhood.

5.1.3 *Assessing the secondary p-Median Objective*

Our key interest in the secondary *p*-Median objective is as to its impact on computation times, its simulated performance, and its predictive power as compared to RC-QMALP without the secondary objective. For this analysis, we shall present some key summary statistics and conduct a Sign test to compare the two models along all three dimensions.

5.2 Results

5.2.1 *Model Predictions*

To visualize these results, Figure 1 and Figure 2 (below) report the predicted α-reliable coverage for each model under varying parameters. Predicted results represent the objective function value of each model, and may not in themselves represent the actual performance of a system. Within each graph the predicted percentage of demand covered with α-reliability (PRC) is reported on the y-axis (PRC. [%]) and the number of facilities is on the x-axis (*P*). Also, within each graph we plot the α-reliable coverage values for all five reliability standards (the color legend is at the bottom of the graph).

In both figures (and other figures Section 5.2) the graphs are sorted according to different parameters. First, Figure 1 corresponds to the low call intensity scenarios (2 CPH) while Figure 2 corresponds to the high call intensity scenarios (4 CPH). Then, along the columns they are sorted by service time standards (6 and 10 minutes) and thus the first column of graphs corresponds to models parameterized with the 6-minute response time standard. Along the rows, four main models MALP 2, QMALP, RCQ (RC-QMALP without the PMP objective), and RCQPM (RC-QMALP) are grouped by the city diameter (16 and 32 minutes). Thus, in the first group the results of the five models are associated with a 16-minute city diameter where columns present results for different coverage standards. For example, the top graph on the far-left concerns MALP 2 (with colored line graphs for all five reliability standards) when there are two calls per hours, a 16-minute city diameter, and a 6-minute service standard. Note that we only report the highest and lowest city diameters and service standards (a total of four dimensions per scenario) as all observed patterns are most pronounced with these

combinations. Also, that we maintain this structure and a model row order of MALP 2, QMALP, RCQ, and RCQPM throughout Section 5.2 unless its noted otherwise.

Beginning with the low call intensity scenario, the two most notable general trends in Figure 1 (below) are that (1) model objective values increase and converge more gradually along *P (number of located units)* with increasing city diameters and (2) objective function values converge faster for all different α-reliability standards as service standards increase. As for model specific trends, MALP 2 generates high objective values the soonest along *P* and this trend is maintained along increasing service time standards although it's less obvious as city diameters increase. The other three models seemingly produce similar solutions along all model parameters and dimensions perhaps reflecting their common queue based framework.

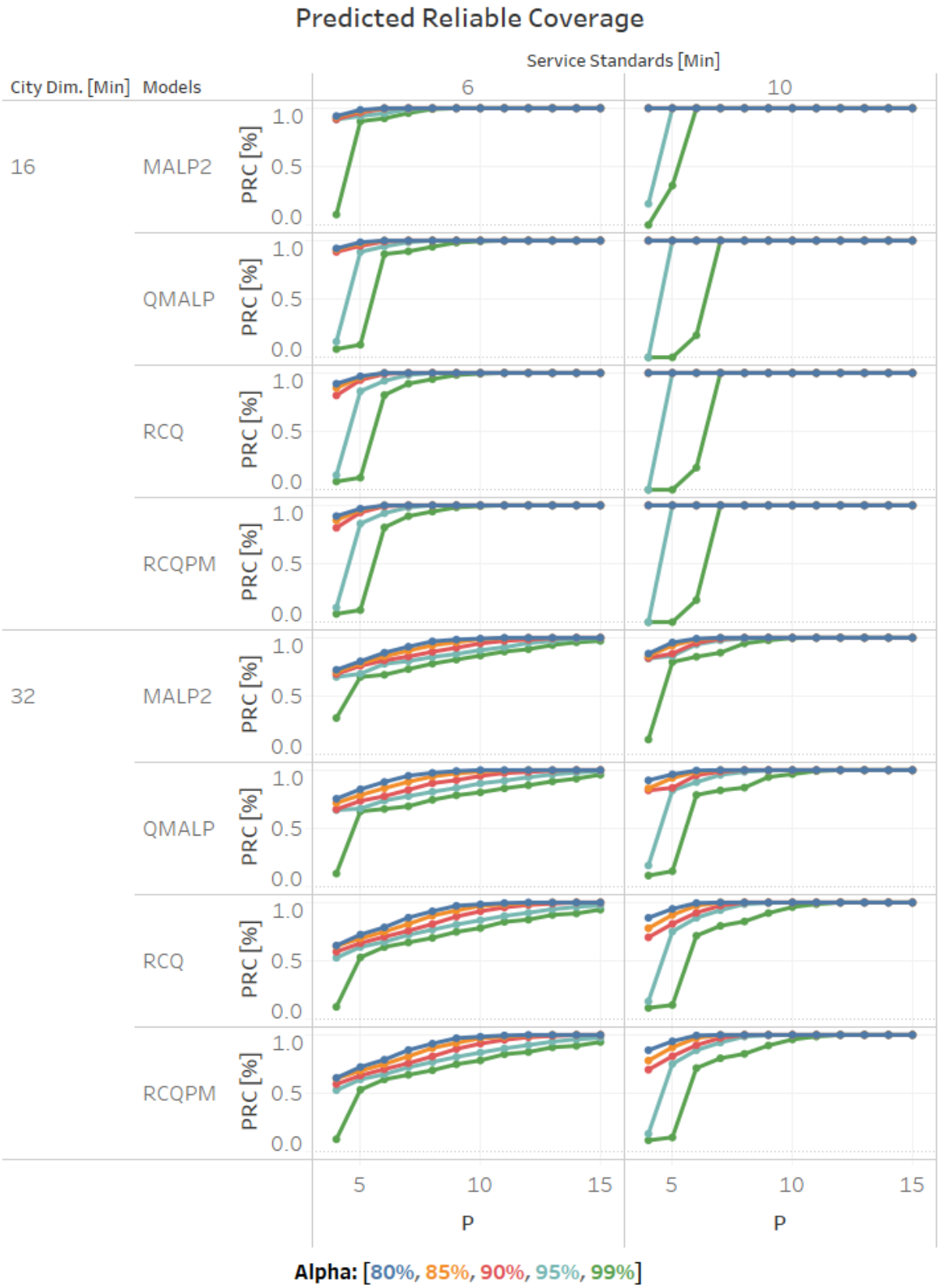

Figure 1 - Predicted α-reliable coverage: Low call intensity scenario (2 CPH)

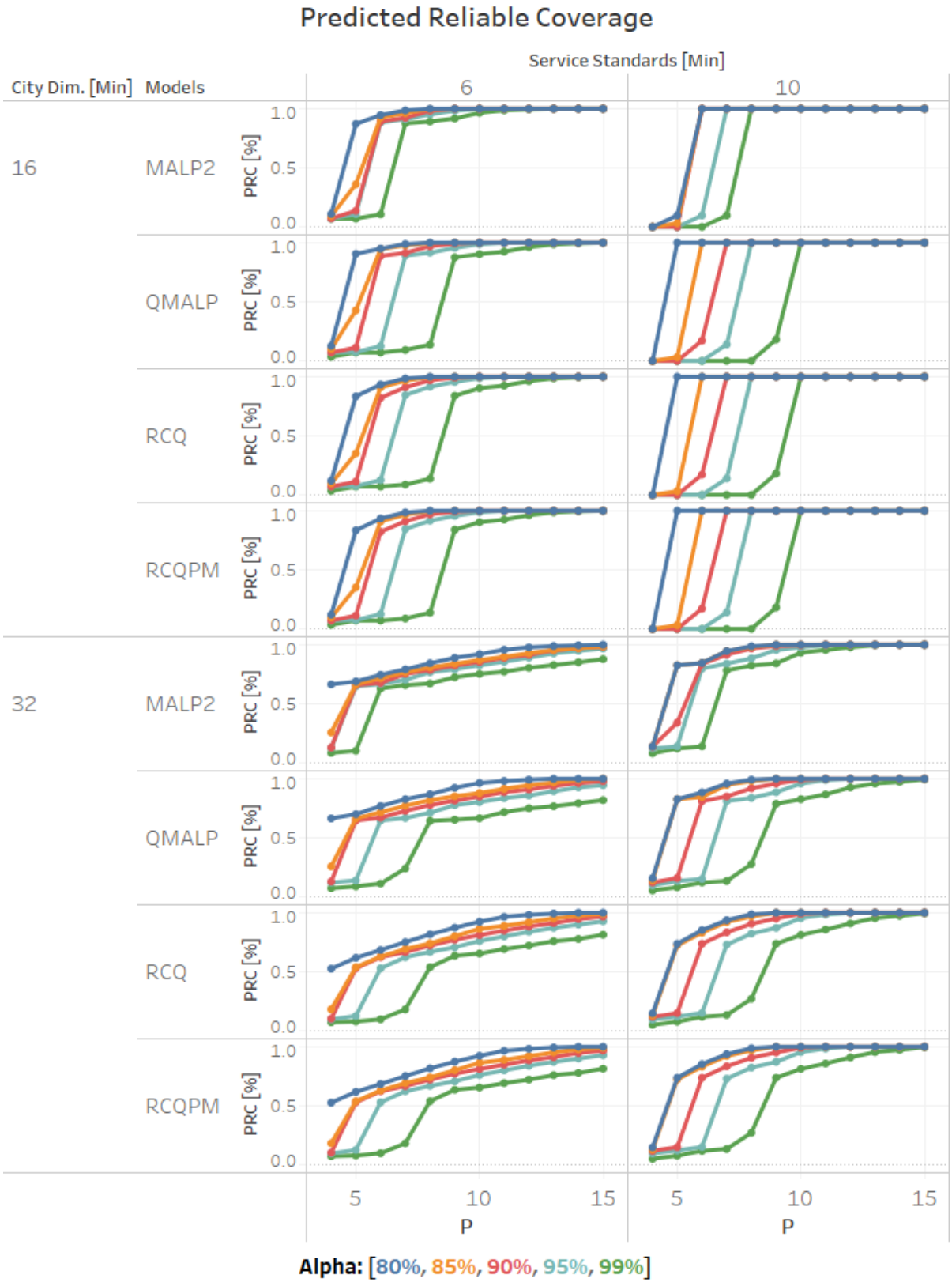

Figure 2 - Predicted α-reliable coverage: High call intensity scenario (4 CPH)

In the high call intensity scenario depicted in Figure 2 (above), we observe similar yet more pronounced patterns. First, the model objective values increase and converge far more gradually along *P* with increasing city diameters. Second, the "base" objective values (i.e., *P* = 4) are far lower than in the low call volume scenario and more servers are required with higher α-reliability standards to achieve similar levels of reliable coverage. Third, the concave objective value functions also appear to increase less gradually with higher service standards and α-reliability standards.

To understand the differences between models, we turn to our tabulation analysis where we compared the objective values generated by each model to the maximum objective value (Z^{PM}) generated by these models for every problem instance. In Table 2 the instances where each model matched Z^{PM} or produced objective values within 1%, 2%, and 5% of Z^{PM} as well as unique objective values. The counts are aggregated across all model dimensions and parameters.

	MALP 2		**QMALP**		**RCQ**		**RCQPM**	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	**978**	**90.56%**	729	67.50%	571	52.87%	571	52.87%
0.01	**1011**	**93.61%**	798	73.89%	624	57.78%	624	57.78%
0.02	**1040**	**96.30%**	836	77.41%	667	61.76%	667	61.76%
0.05	**1065**	**98.61%**	929	86.02%	756	70.00%	756	70.00%
Unique	**351**	**32.50%**	90	8.33%	0	0.00%	0	0.00%

Table 2 – Highest predicted reliable coverage: Aggregated across all scenarios

From Table 2 it's clear that MALP 2 has the highest proportion of maximum predicted objective values at all four levels and also produced the highest number of unique solutions. The other three models produced considerably fewer solutions with lower objective values of

α-reliability, although QMALP generated a few unique solutions whereas RCQ and RCQPM did not produce any.

The next dimensions we consider are both different *P* values and α-reliability standards.

		P											
		4	5	6	7	8	9	10	11	12	13	14	15
α: 80%	**MALP 2** - Max	33.33%	27.78%	38.89%	55.56%	55.56%	77.78%	77.78%	83.33%	94.44%	94.44%	94.44%	**100.00%**
	0.01	33.33%	50.00%	55.56%	61.11%	83.33%	83.33%	83.33%	94.44%	94.44%	94.44%	**100.00%**	**100.00%**
	0.02	38.89%	72.22%	66.67%	77.78%	83.33%	83.33%	94.44%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	50.00%	83.33%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**0.00%**	**0.00%**
	QMALP - Max	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.01	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**61.11%**	**61.11%**	**55.56%**	**44.44%**	**27.78%**	**22.22%**	**22.22%**	**5.56%**	**5.56%**	**5.56%**	**0.00%**	**0.00%**
	RCQ - Max	22.22%	27.78%	44.44%	50.00%	72.22%	77.78%	77.78%	88.89%	94.44%	94.44%	**100.00%**	**100.00%**
	0.01	27.78%	27.78%	50.00%	66.67%	77.78%	77.78%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	33.33%	38.89%	55.56%	72.22%	77.78%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	50.00%	50.00%	77.78%	77.78%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**0.00%**	**0.00%**
	RCQMALP - Max	22.22%	27.78%	44.44%	50.00%	72.22%	77.78%	77.78%	88.89%	94.44%	94.44%	**100.00%**	**100.00%**
	0.01	27.78%	27.78%	50.00%	66.67%	77.78%	77.78%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	33.33%	38.89%	55.56%	72.22%	77.78%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	50.00%	50.00%	77.78%	77.78%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**0.00%**	**0.00%**

Table 3 - Highest predicted reliable coverage: Aggregated for all *P* and α, α = 80%

With α at 80%, QMALP produced the highest proportion of maximum predicted objective values at all four tiers as well as the highest proportion of unique solution values for all *P* values. Moreover, only QMALP produced unique solution values at this level of reliability although the number of solutions decreased as *P* increased. Note that both RCQ and RCQPM tie QMALP for the highest proportion of maximum predicted objective values beginning at *P* = 14 while MALP 2 ties QMALP beginning at *P* = 15.

		P											
		4	5	6	7	8	9	10	11	12	13	14	15
α: 85%	**MALP 2** - Max	72.22%	66.67%	77.78%	72.22%	77.78%	77.78%	94.44%	94.44%	94.44%	94.44%	94.44%	94.44%
	0.01	88.89%	77.78%	83.33%	88.89%	77.78%	94.44%	**100.00%**	94.44%	94.44%	94.44%	94.44%	94.44%
	0.02	94.44%	77.78%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	94.44%	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	QMALP - Max	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.01	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**27.78%**	**33.33%**	**22.22%**	**27.78%**	**22.22%**	**11.11%**	**5.56%**	**5.56%**	**5.56%**	**5.56%**	**5.56%**	**5.56%**
	RCQ - Max	27.78%	22.22%	44.44%	50.00%	61.11%	77.78%	77.78%	77.78%	88.89%	94.44%	94.44%	94.44%
	0.01	27.78%	27.78%	44.44%	50.00%	72.22%	77.78%	77.78%	88.89%	94.44%	94.44%	94.44%	94.44%
	0.02	27.78%	33.33%	55.56%	55.56%	77.78%	77.78%	**100.00%**	94.44%	94.44%	**100.00%**	**100.00%**	**100.00%**
	0.05	38.89%	44.44%	66.67%	77.78%	77.78%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	RCQMALP - Max	27.78%	22.22%	44.44%	50.00%	61.11%	77.78%	77.78%	77.78%	88.89%	94.44%	94.44%	94.44%
	0.01	27.78%	27.78%	44.44%	50.00%	72.22%	77.78%	77.78%	88.89%	94.44%	94.44%	94.44%	94.44%
	0.02	27.78%	33.33%	55.56%	55.56%	77.78%	77.78%	**100.00%**	94.44%	94.44%	**100.00%**	**100.00%**	**100.00%**
	0.05	38.89%	44.44%	66.67%	77.78%	77.78%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4 - Highest model predicted coverage: Aggregated for all *P* and α, α = 85%

With α at 85%, again QMALP produced the highest proportion of maximum predicted objective values at all four tiers and the highest proportion of unique solution values. Likewise, only QMALP generated unique solutions and their number decreased as *P* increased. No model tied QMALP for the highest proportion of maximum predicted objective values at any level of *P* although MALP 2's proportion of solutions matching Z^{PM} surpassed 90% at *P* = 10 while RCQ and RCQPM surpassed 90% at *P* = 13.

		4	5	6	7	8	9	10	11	12	13	14	15
α: 90%	**MALP 2** - Max	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**94.44%**	**100.00%**	**100.00%**	**94.44%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.01	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**66.67%**	**83.33%**	**83.33%**	**66.67%**	**38.89%**	**33.33%**	**33.33%**	**16.67%**	**5.56%**	**16.67%**	**5.56%**	**5.56%**
	QMALP - Max	33.33%	16.67%	16.67%	33.33%	61.11%	66.67%	66.67%	83.33%	94.44%	83.33%	94.44%	94.44%
	0.01	55.56%	22.22%	66.67%	61.11%	61.11%	77.78%	**100.00%**	**100.00%**	**100.00%**	94.44%	**100.00%**	**100.00%**
	0.02	61.11%	44.44%	66.67%	72.22%	77.78%	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	83.33%	61.11%	83.33%	83.33%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	5.56%	0.00%	0.00%	5.56%	0.00%	0.00%	0.00%	0.00%
	RCQ - Max	22.22%	16.67%	11.11%	33.33%	44.44%	50.00%	66.67%	77.78%	77.78%	77.78%	94.44%	94.44%
	0.01	22.22%	16.67%	22.22%	38.89%	44.44%	55.56%	77.78%	77.78%	94.44%	94.44%	94.44%	94.44%
	0.02	22.22%	22.22%	22.22%	55.56%	55.56%	66.67%	77.78%	83.33%	94.44%	94.44%	94.44%	**100.00%**
	0.05	27.78%	33.33%	27.78%	55.56%	66.67%	83.33%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	RCQMALP - Max	22.22%	16.67%	11.11%	33.33%	44.44%	50.00%	66.67%	77.78%	77.78%	77.78%	94.44%	94.44%
	0.01	22.22%	16.67%	22.22%	38.89%	44.44%	55.56%	77.78%	77.78%	94.44%	94.44%	94.44%	94.44%
	0.02	22.22%	22.22%	22.22%	55.56%	55.56%	66.67%	77.78%	83.33%	94.44%	94.44%	94.44%	**100.00%**
	0.05	27.78%	33.33%	27.78%	55.56%	66.67%	83.33%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 5 - Highest predicted reliable coverage: Aggregated for all *P* and α, α = 90%

With α at 90%, MALP 2 produced the highest proportion of maximum predicted objective values at all four tiers as well as the highest proportion of unique solution values. The number of unique solutions also decreased as *P* increased, however, QMALP also produced unique solutions when *P* = 8 and 11. No model tied QMALP for the highest proportion of maximum predicted objective values at any level of *P* although QMALP's proportion of solutions matching Z^{PM} surpassed 90% at *P* = 12, 14, and 15 while RCQ and RCQPM surpassed 90% at *P* = 14.

		P											
		4	5	6	7	8	9	10	11	12	13	14	15
α: 95%	**MALP 2** - Max	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.01	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**72.22%**	**77.78%**	**88.89%**	**88.89%**	**55.56%**	**55.56%**	**44.44%**	**33.33%**	**22.22%**	**22.22%**	**22.22%**	**22.22%**
	QMALP - Max	27.78%	22.22%	11.11%	11.11%	44.44%	44.44%	55.56%	66.67%	77.78%	77.78%	77.78%	77.78%
	0.01	33.33%	38.89%	27.78%	27.78%	55.56%	55.56%	66.67%	77.78%	77.78%	77.78%	83.33%	94.44%
	0.02	33.33%	50.00%	27.78%	44.44%	55.56%	66.67%	83.33%	83.33%	83.33%	94.44%	94.44%	94.44%
	0.05	44.44%	72.22%	50.00%	83.33%	77.78%	77.78%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	RCQ - Max	16.67%	16.67%	11.11%	11.11%	33.33%	33.33%	44.44%	55.56%	77.78%	77.78%	77.78%	77.78%
	0.01	16.67%	22.22%	11.11%	22.22%	44.44%	44.44%	50.00%	72.22%	77.78%	77.78%	77.78%	88.89%
	0.02	16.67%	22.22%	11.11%	22.22%	44.44%	44.44%	55.56%	77.78%	77.78%	77.78%	88.89%	88.89%
	0.05	16.67%	27.78%	22.22%	33.33%	66.67%	66.67%	77.78%	83.33%	94.44%	94.44%	94.44%	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	RCQMALP -Max	16.67%	16.67%	11.11%	11.11%	33.33%	33.33%	44.44%	55.56%	77.78%	77.78%	77.78%	77.78%
	0.01	16.67%	22.22%	11.11%	22.22%	44.44%	44.44%	50.00%	72.22%	77.78%	77.78%	77.78%	88.89%
	0.02	16.67%	22.22%	11.11%	22.22%	44.44%	44.44%	55.56%	77.78%	77.78%	77.78%	88.89%	88.89%
	0.05	16.67%	27.78%	22.22%	33.33%	66.67%	66.67%	77.78%	83.33%	94.44%	94.44%	94.44%	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 6 - Highest predicted reliable coverage: Aggregated for all *P* and α, α = 95%

With α at 95%, MALP 2 again produced the highest proportion of maximum predicted objective values at all four tiers well as the highest proportion of unique maximum predicted objective values. Again, the number of unique maximum objective values also decreased as *P* increased, however, no other model produced any unique maximum objective values. No model tied QMALP for the highest proportion of maximum objective values at any level of *P* or had their proportion of maximum objective values matching Z^{PM} exceeded 80% at any level of *P*.

		P											
		4	5	6	7	8	9	10	11	12	13	14	15
α :99%	**MALP 2** - Max	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.01	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**83.33%**	**88.89%**	**94.44%**	**88.89%**	**94.44%**	**88.89%**	**77.78%**	**66.67%**	**61.11%**	**55.56%**	**50.00%**	**38.89%**
	QMALP - Max	16.67%	11.11%	5.56%	11.11%	5.56%	11.11%	22.22%	33.33%	38.89%	44.44%	50.00%	61.11%
	0.01	16.67%	11.11%	5.56%	11.11%	11.11%	16.67%	33.33%	44.44%	44.44%	50.00%	55.56%	77.78%
	0.02	16.67%	11.11%	5.56%	11.11%	11.11%	22.22%	38.89%	44.44%	50.00%	55.56%	72.22%	83.33%
	0.05	16.67%	16.67%	38.89%	22.22%	33.33%	50.00%	66.67%	61.11%	72.22%	83.33%	94.44%	94.44%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	RCQ - Max	16.67%	11.11%	5.56%	11.11%	5.56%	11.11%	22.22%	33.33%	38.89%	44.44%	50.00%	61.11%
	0.01	16.67%	11.11%	5.56%	11.11%	11.11%	16.67%	33.33%	38.89%	44.44%	50.00%	55.56%	77.78%
	0.02	16.67%	11.11%	5.56%	11.11%	11.11%	22.22%	33.33%	44.44%	50.00%	55.56%	72.22%	77.78%
	0.05	16.67%	11.11%	5.56%	16.67%	22.22%	27.78%	50.00%	44.44%	55.56%	83.33%	77.78%	83.33%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	RCQMALP - Max	16.67%	11.11%	5.56%	11.11%	5.56%	11.11%	22.22%	33.33%	38.89%	44.44%	50.00%	61.11%
	0.01	16.67%	11.11%	5.56%	11.11%	11.11%	16.67%	33.33%	38.89%	44.44%	50.00%	55.56%	77.78%
	0.02	16.67%	11.11%	5.56%	11.11%	11.11%	22.22%	33.33%	44.44%	50.00%	55.56%	72.22%	77.78%
	0.05	16.67%	11.11%	5.56%	16.67%	22.22%	27.78%	50.00%	44.44%	55.56%	83.33%	77.78%	83.33%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 7 - Highest predicted reliable coverage: Aggregated for all *P* and α, α = 99%

Lastly, with α at 99%, MALP 2 again produced the highest proportion of maximum predicted objective values at all four tiers as well as the highest proportion of unique maximum predicted objective values. Likewise, the number of unique maximum objective values also decreased as *P* increased and no other model produced any unique maximum objective values. Again, no model tied QMALP for the highest proportion of maximum objective values at any level of *P* or had their proportion of maximum objective values matching Z^{PM} exceeded 65% at any level of *P*.

In all, the tabulations at this level indicated that QMALP begins as the dominant model but MALP 2 begins predicting high levels of reliable coverage beginning at α = 90%. Moreover, at this point the proportion of maximum objective values matching Z^{PM} continually drops for all other models and interestingly, at α = 99% the proportions for these three models converged at every tier and all P values.

To explore other factors, in Table 8,

Table 9, and

	City Diameter: 32 [Min]					
Demand [Call/Hr.]:	2			4		
Service Std. [Min]:	6	8	10	6	8	10
MALP 2 - Max	**75.00%**	**81.67%**	**95.00%**	**68.33%**	**86.67%**	**93.33%**
0.01	**88.33%**	**90.00%**	**98.33%**	**73.33%**	**88.33%**	**95.00%**
0.02	**93.33%**	**96.67%**	**98.33%**	**90.00%**	**91.67%**	**96.67%**
0.05	**100.00%**	**98.33%**	**98.33%**	**100.00%**	**98.33%**	**98.33%**
Unique	**48.33%**	**31.67%**	**26.67%**	**58.33%**	**55.00%**	**45.00%**
QMALP - Max	51.67%	68.33%	73.33%	41.67%	45.00%	55.00%
0.01	61.67%	78.33%	80.00%	53.33%	53.33%	60.00%
0.02	70.00%	81.67%	83.33%	60.00%	58.33%	61.67%
0.05	96.67%	91.67%	90.00%	76.67%	71.67%	71.67%
Unique	25.00%	15.00%	5.00%	30.00%	11.67%	6.67%
RCQ - Max	15.00%	41.67%	60.00%	3.33%	18.33%	38.33%
0.01	21.67%	48.33%	65.00%	6.67%	26.67%	43.33%
0.02	28.33%	51.67%	71.67%	16.67%	31.67%	48.33%
0.05	46.67%	63.33%	80.00%	30.00%	41.67%	60.00%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQPM -Max	15.00%	41.67%	60.00%	3.33%	18.33%	38.33%
0.01	21.67%	48.33%	65.00%	6.67%	26.67%	43.33%
0.02	28.33%	51.67%	71.67%	16.67%	31.67%	48.33%
0.05	46.67%	63.33%	80.00%	30.00%	41.67%	60.00%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 10 (below) we disaggregate the model results along all other dimension while aggregating along p and α.

	City Diameter:16 [Min]					
Demand [Call/Hr.]:	2			4		
Service Std. [Min]:	6	8	10	6	8	10
MALP 2 - Max	**100.00%**	**100.00%**	**100.00%**	**86.67%**	**96.67%**	**98.33%**
0.01	**100.00%**	**100.00%**	**100.00%**	**90.00%**	**96.67%**	**98.33%**
0.02	**100.00%**	**100.00%**	**100.00%**	**91.67%**	**96.67%**	**98.33%**
0.05	**100.00%**	**100.00%**	**100.00%**	**95.00%**	**96.67%**	**98.33%**
Unique	**23.33%**	**10.00%**	**5.00%**	**36.67%**	**23.33%**	**10.00%**
QMALP - Max	76.67%	90.00%	95.00%	63.33%	76.67%	90.00%
0.01	86.67%	93.33%	95.00%	68.33%	83.33%	90.00%
0.02	88.33%	93.33%	95.00%	75.00%	83.33%	90.00%
0.05	93.33%	93.33%	95.00%	85.00%	83.33%	90.00%
Unique	0.00%	0.00%	0.00%	13.33%	0.00%	0.00%
RCQ - Max	70.00%	90.00%	95.00%	50.00%	76.67%	90.00%
0.01	75.00%	93.33%	95.00%	55.00%	83.33%	90.00%
0.02	80.00%	93.33%	95.00%	65.00%	83.33%	90.00%
0.05	86.67%	93.33%	95.00%	76.67%	83.33%	90.00%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQPM - Max	70.00%	90.00%	95.00%	50.00%	76.67%	90.00%
0.01	75.00%	93.33%	95.00%	55.00%	83.33%	90.00%
0.02	80.00%	93.33%	95.00%	65.00%	83.33%	90.00%
0.05	86.67%	93.33%	95.00%	76.67%	83.33%	90.00%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 8 – Highest predicted reliable coverage: City diameter 16 [Min], aggregated across p and α.

In Table 8, we observe that models generally generate an increasing proportion of objective values equal to Z^{PM} and a decreasing proportion of unique maximum objective values as the service distance standards increase. Across different demand intensities, there appears to be a decrease in the proportion of solutions equal to Z^{PM} except this effect is less pronounced with MALP 2. In the lower call intensity scenario, the top tiers of QMALP, RCQ, and RCQPM exceed 90% beginning with the 8-minute service standard. For the higher demand intensities, the top tiers of QMALP, RCQ, and RCQPM exceed 90% only at the 10-minute service standard but they converge at the 8-minute service standard. Lastly, only MALP 2 generated unique maximum objective values in the low demand intensity scenario and both MALP 2 and

QMALP produced the unique maximum objective values in the high demand intensity scenario although MALP 2 produced more at every service standard level.

	City Diameter: 24 [Min]					
Demand [Call/Hr.]:	2			4		
Service Std. [Min]:	6	8	10	6	8	10
MALP 2 - Max	**81.67%**	**93.33%**	**100.00%**	**86.67%**	**93.33%**	**93.33%**
0.01	**90.00%**	**98.33%**	**100.00%**	**88.33%**	**95.00%**	**95.00%**
0.02	**96.67%**	**98.33%**	**100.00%**	**91.67%**	**96.67%**	**96.67%**
0.05	**98.33%**	**100.00%**	**100.00%**	**98.33%**	**98.33%**	**96.67%**
Unique	**31.67%**	**26.67%**	**21.67%**	**55.00%**	**41.67%**	**35.00%**
QMALP - Max	68.33%	73.33%	78.33%	45.00%	58.33%	65.00%
0.01	78.33%	80.00%	85.00%	53.33%	61.67%	68.33%
0.02	81.67%	86.67%	88.33%	58.33%	65.00%	73.33%
0.05	91.67%	91.67%	95.00%	71.67%	76.67%	83.33%
Unique	15.00%	6.67%	0.00%	11.67%	5.00%	5.00%
RCQ - Max	41.67%	66.67%	73.33%	18.33%	45.00%	58.33%
0.01	48.33%	70.00%	81.67%	26.67%	48.33%	61.67%
0.02	51.67%	70.00%	83.33%	31.67%	51.67%	68.33%
0.05	63.33%	78.33%	91.67%	41.67%	61.67%	76.67%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQPM - Max	41.67%	66.67%	73.33%	18.33%	45.00%	58.33%
0.01	48.33%	70.00%	81.67%	26.67%	48.33%	61.67%
0.02	51.67%	70.00%	83.33%	31.67%	51.67%	68.33%
0.05	63.33%	78.33%	91.67%	41.67%	61.67%	76.67%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 9 - Highest predicted reliable coverage: City diameter 24 [Min], aggregated across p and α.

In Table 9, we observe that models tend to generate an increasing proportion of objective values equal to Z^{PM}, a decreasing number of unique maximum objective values as the service time standards increase, and a decreasing number of objective values equal to Z^{PM} with increasing demand. The latter trend is less pronounced with MALP 2, however. In the lower call intensity scenario, only the 5% tier of QMALP, RCQ, and RCQPM exceeded 90% and at the 10-minute service standard QMALP does not converge with the RC-QMALP models. For

the higher demand intensities, the top tiers of QMALP, RCQ, and RCQPM do not exceed 90% at any service time standard and again, QMALP and the RC-QMALP models do not converge. Lastly, both MALP 2 and QMALP generated unique maximum objective values in both call intensity scenarios although, again, MALP 2 produced more unique solutions at every service standard level in both call intensity scenarios.

	City Diameter: 32 [Min]					
Demand [Call/Hr.]:	2			4		
Service Std. [Min]:	6	8	10	6	8	10
MALP 2 - Max	**75.00%**	**81.67%**	**95.00%**	**68.33%**	**86.67%**	**93.33%**
0.01	**88.33%**	**90.00%**	**98.33%**	**73.33%**	**88.33%**	**95.00%**
0.02	**93.33%**	**96.67%**	**98.33%**	**90.00%**	**91.67%**	**96.67%**
0.05	**100.00%**	**98.33%**	**98.33%**	**100.00%**	**98.33%**	**98.33%**
Unique	**48.33%**	**31.67%**	**26.67%**	**58.33%**	**55.00%**	**45.00%**
QMALP - Max	51.67%	68.33%	73.33%	41.67%	45.00%	55.00%
0.01	61.67%	78.33%	80.00%	53.33%	53.33%	60.00%
0.02	70.00%	81.67%	83.33%	60.00%	58.33%	61.67%
0.05	96.67%	91.67%	90.00%	76.67%	71.67%	71.67%
Unique	25.00%	15.00%	5.00%	30.00%	11.67%	6.67%
RCQ - Max	15.00%	41.67%	60.00%	3.33%	18.33%	38.33%
0.01	21.67%	48.33%	65.00%	6.67%	26.67%	43.33%
0.02	28.33%	51.67%	71.67%	16.67%	31.67%	48.33%
0.05	46.67%	63.33%	80.00%	30.00%	41.67%	60.00%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQPM -Max	15.00%	41.67%	60.00%	3.33%	18.33%	38.33%
0.01	21.67%	48.33%	65.00%	6.67%	26.67%	43.33%
0.02	28.33%	51.67%	71.67%	16.67%	31.67%	48.33%
0.05	46.67%	63.33%	80.00%	30.00%	41.67%	60.00%
Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 10 - Highest predicted reliable coverage: City diameter 32 [Min], aggregated across *p* and α.

In Table 10, we observe many of the same patterns as in the previous table but with this additional scenario it becomes apparent that the number of unique solutions generated by MALP 2 and QMALP increase along with the city diameter. Moreover, the RC-QMALP objective values are drastically lower along this dimension, particularly in the high call

intensity scenario. Together, this supports the idea that RC-QMALP models are very conservative with their predictions (likely due to the presence of additional constraints) or at least that MALP 2 and QMALP are more flexible although for different reasons.

As reported above, QMALP generates higher maximum objective values with lower α-reliability values while MALP 2 generates higher values with higher α-reliability values. This is not a coincidence but rather an arguably overlooked point that is based on the calculations of the b_i requirements to establish α-reliable service. To discuss this it's useful to consider a subtle conundrum in Marianov & ReVelle's (1996) presentation of QMALP. Here they clearly reported the later point by comparing *MALP* and QMALP in that with higher α-reliability values, more demand nodes required higher b_i values with QMALP. They also emphasized their findings of achieving higher levels of availability for demand nodes under QMALP for the case where one server was located in each neighborhood. This seemingly suggests a relationship between higher service availability and QMALP but it also appears to contradict their findings about QMALP and MALP's b_i requirements. The answer (which we shall discuss in more depth later) is that the way that MALP calacualtes b_i leads to significant overestimates, particualrly with low α-reliability values. With MALP, a utilization rate of 1 (i.e., 1 call per hour) requires 2 servers to surpass 50% reliability. In contrast, a utilization rate of 1 with an *M*/*G*/1/1-loss queue results in a server availability of 50%!

5.2.2 *Simulation Results*

To report our simulation results we follow the same reporting style as in the earlier section. We begin with the simulated reliable coverage results before reporting the simulated total coverage results.

5.2.2.1 Simulated Reliable Coverage

The graphs in this section report the simulation performance for all four models in terms of simulated reliable coverage for, respectively, a low call intensity scenario (2 CPH) and then a high call intensity scenario (4 CPH) under the same city diameters, service time standards, and α-reliability values. Within each graph the simulated percentage of demand covered with α-reliability (SRC) is reported on the y-axis (SRC [%]) and the number of facilities is on the x-axis (*P*).

Beginning with general trends, perhaps the most striking feature about the simulated reliability coverage values is that the function they formed was at times not monotonic. With the predicted reliability value functions, adding an ambulance improved the level of coverage or at least maintained the current level. In contrast, in our simulation study in some instances *adding* an ambulance resulted in lower reliable coverage levels. The problem appeared with all models although the size and frequencies of these drops varied widely although they emerged more often with shorter service standards and larger city diameters. Another related issue included simulated reliability value lines for different α-reliability values actually crossing each other. In generating optimal solutions, solutions to problem instances with lower α-reliability values must be at least as good as solutions for similar problem instances with higher α-reliability values. However, in our simulation study we observed both low α-reliability value lines surpass high α-reliability value lines and high α-reliability value lines sink below low α-reliability value lines.[106]

[106] This issue is well known in the literature (e.g., Erkut et al., 2008) and presents a serious challenge to solving and formulating MALP-like problems.

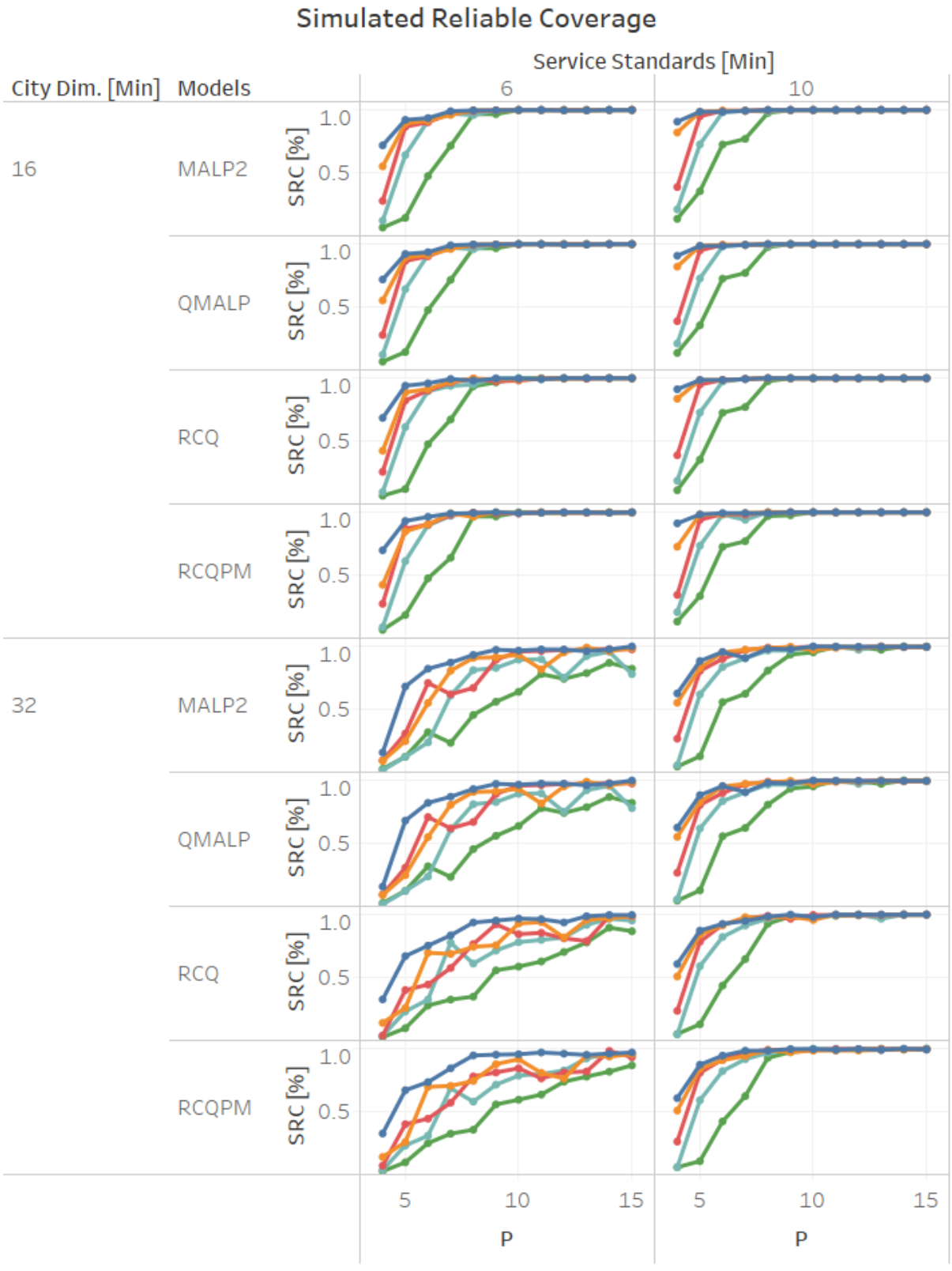

Figure 3 - Simulated α-reliable coverage: High call intensity scenario (2 CPH)

In the low call intensity scenario depicted in Figure 3 (above), the general characteristics of predicted reliability value lines from the previous section also apply here for the most part, although they are not as pronounced due to the issues described above. For instance, the model objective values increase and converge more gradually along the *P* axis as the city diameter increases. There are also sudden and often drastic non-monotonic changes (i.e., over some interval of *P* values). Likewise, the objective function values converge faster as the service standards increase. But, again, some value lines do not absolutely converge due to non-monotonic changes that occur for large city diameters and low service standards. Unlike the predicted reliability value graphs, no model stood out as all four models produced similar objective value functions for the most part although two model groups (one with MALP and QMALP and the other with RCQ and RCQPM) each generated objective value functions with distinct features, namely similar kinks or changes at similar positions in the graph.

In the high call intensity scenario depicted in **Error! Reference source not found.** (below), overall trends patterns become less clear. First, the most obvious pattern is the zero or near zero objective values for the low values of *P*. Second, after a "lag", the objective value lines form concave like function and converge to high objective values the fastest with higher service standards and lower city diameters. Notably, these increases are relatively steep with lower α-reliability values. But with the highest city diameters and smallest service standards, the lines of coverage tend to rise much more slowly and slowly converge to the highest value. No model stands out above the rest but the two model groups produce objective value lines with distinctive issues, like lines crossing for different alpha values and sometimes even dipping with additional units.

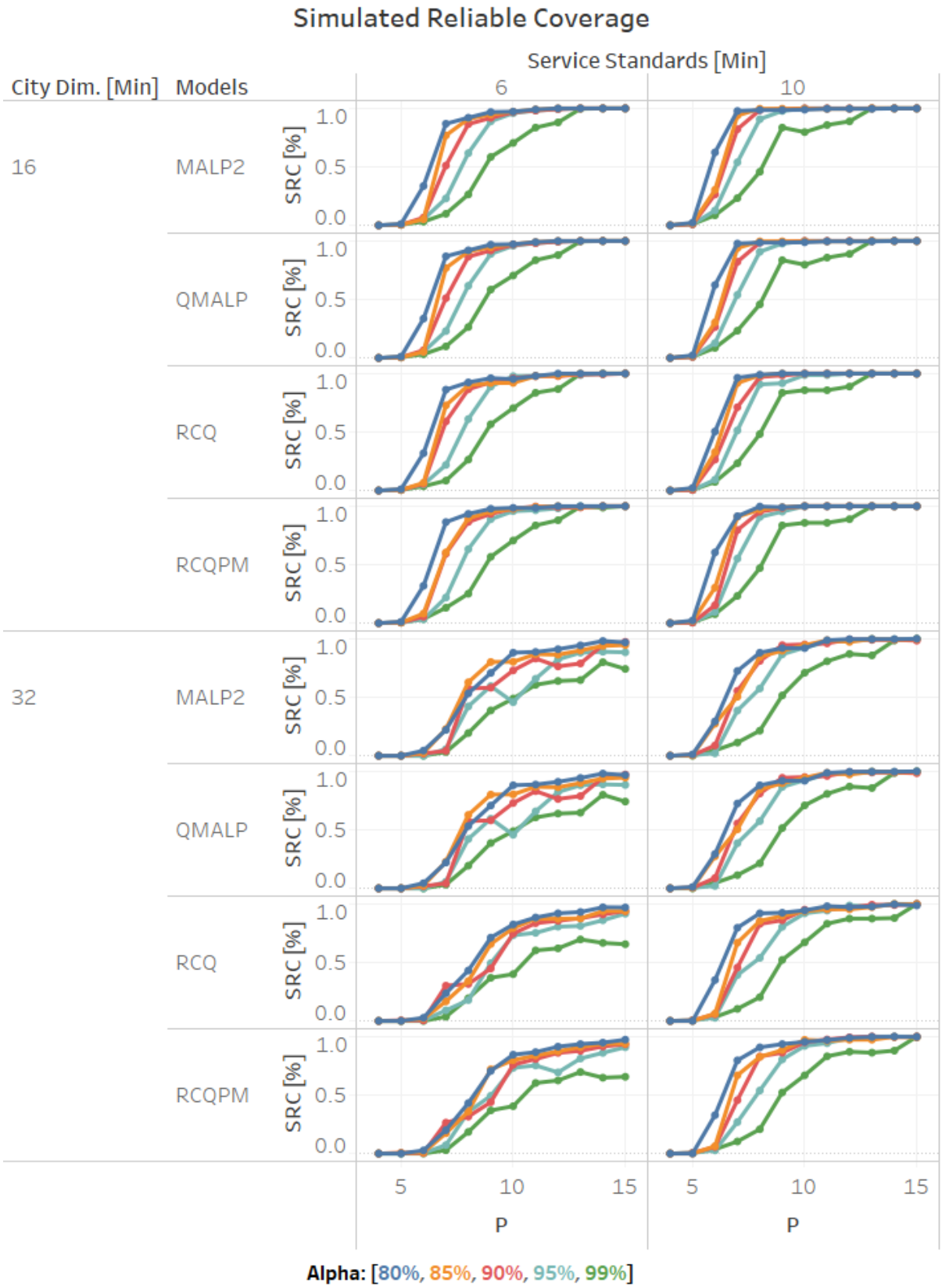

Figure 4 - Simulated α-reliable coverage: High call intensity scenario (4 CPH)

Table 11 (below) reports the relative performance between the four models with respect to simulated α-reliable coverage. MALP 2 and QMALP produce the highest proportion of solutions with objective values greater than or equal to the maximum simulated reliability value for each problem instance (Z^{SR}). For the top three tiers (Max, 0.01, and 0.02), the difference between MALP 2 and QMALP and the RC-QMALP models is between ~11 and 17%. However, RQPM outperformed all models in the fourth tier with all objective values being within 5% Z^{SR}. The other models generated objective values that were lower by at least ~7% (MALP 2) and 18% (RCQ) at most. Lastly, only RCQPM and RCQ produced unique maximum objective values and at a similar rate with a slight edge to RCQPM.

	MALP 2		**QMALP**		**RCQ**		**RCQPM**	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	**739**	**68.43%**	**739**	**68.43%**	586	54.26%	556	51.48%
0.01	**865**	**80.09%**	**865**	**80.09%**	716	66.30%	711	65.83%
0.02	**921**	**85.28%**	**921**	**85.28%**	804	74.44%	775	71.76%
0.05	1001	92.69%	892	82.59%	880	81.48%	**1080**	**100.00%**
Unique	0	0.00%	0	0.00%	119	11.02%	**125**	**11.57%**

Table 11 - Highest predicted reliable coverage: Aggregated across all scenarios

Upon disaggregating these results along *P* and α, these general patterns persisted although and both RCQPM and RCQ outperformed or tied MALP 2 and QMALP on several occasions. However, we did not find any distinct or consistent patterns that would suggest a significant trend. With QMALP and MALP, the only discernable trend we identified along *P* and/or α was that they generated higher proportions across all models along *P* values below $P = 13$ although at that these proportions mostly decreased until about $P = 8$ or 9.

6.2.2.2 Simulated Total Coverage

The graphs in this section report the simulation performance for all four models in terms of total expected coverage for, respectively, a low call intensity scenario (2 CPH) and then a high call intensity scenario (4 CPH) under the same city diameters, service time standards, and all *p*. Within each graph the simulated percentage of demand covered with the service time standard (TC) is reported on the y-axis (TC [%]) and the number of facilities is on the x-axis (*P*).

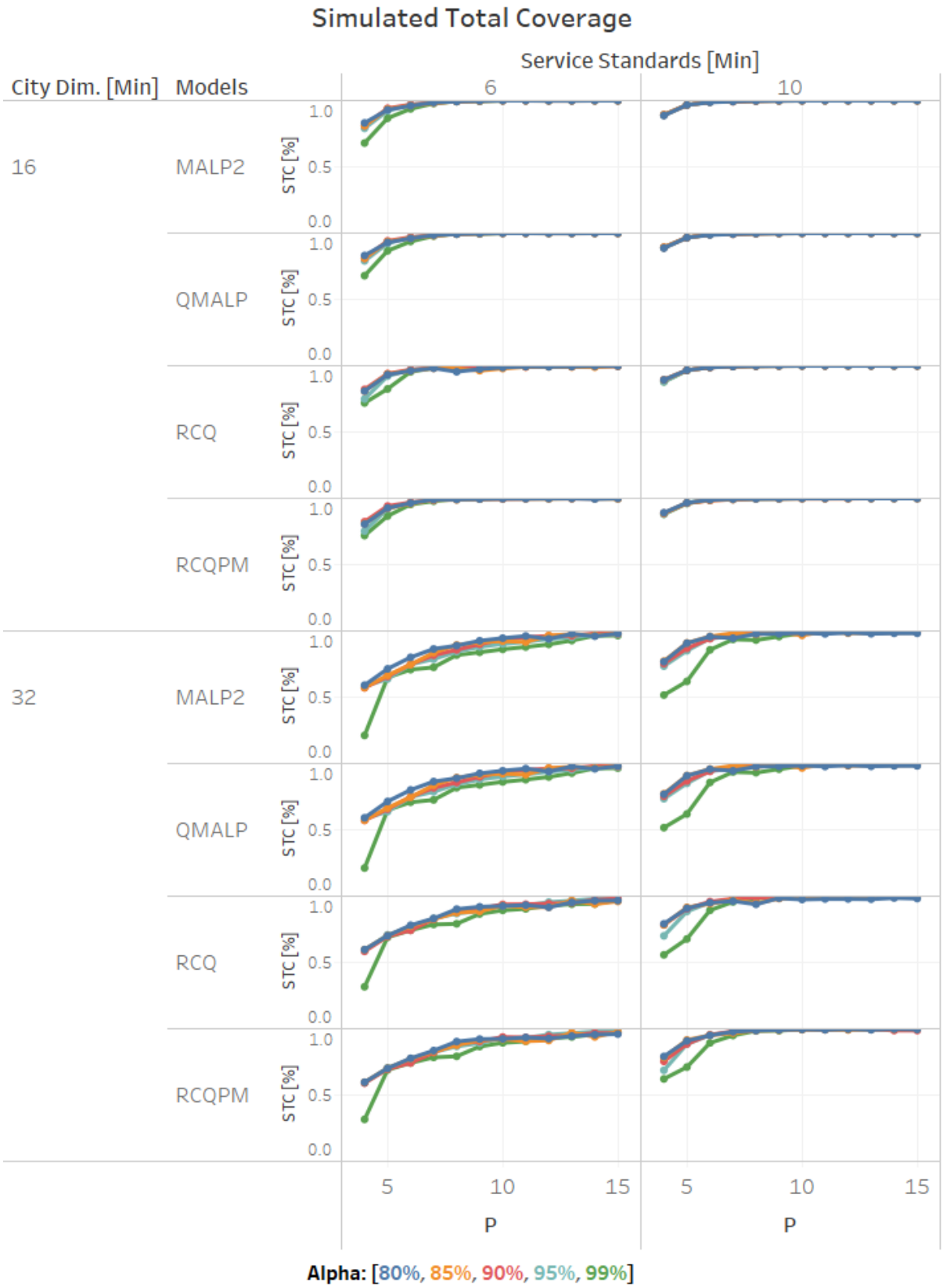

Figure 5 – Simulated total coverage: Low call intensity scenario (2 CPH)

For the low call intensity scenario depicted in Figure 5 (above), we note that all models provide similar levels of coverage throughout all values of *P*. For the smallest city dimeter and shorter service time standards, there is a slight difference between the lower and higher α-reliability values with the lower *P* values (which are slightly larger with the RC-QMALP models) but all models mostly converge along all α-reliability levels by *P* = 7. With the higher service time standard, all models at all α-reliability levels are effectively similar throughout all values of *P*. For the larger city diameter and shorter service time standards, the initial coverage levels (*P* = 4) are substantially lower as compared to the smaller city diameter (some models drop more than 40%) and the difference between the lower and higher α-reliability values are also larger (between 25% and 35%). The gaps are smaller with the larger service time standard, however. Lastly, all models converge quite fast with larger city diameters but the rates of increase are smaller, particularly with the lower service time standards.

For the high call intensity scenario depicted in Figure 6 (below), all models provide similar levels of coverage throughout all *P* values. The difference between the high and low call intensity scenarios are most apparent in terms of the initial coverage levels where all models begin with total coverage levels near 0 for all city diameter and service time standard parameters. However, there was drastically smaller differences within models along different α-reliability values. With smaller city diameters, all models converge faster than with larger city diameters. Here all models achieved at least 90% total coverage at *P* = 8 with 6-minute service time standards and at *P* = 7 with 10-minute service time standards while with larger city diameters we observed convergence at *P* = 15+ and P = 10 with, respectively, smaller and higher service time standards. Also, at *P* = 6 we observed the largest differences within models

under different α-reliability values for almost all models although occasionally we observed smaller differences at higher *P* values.

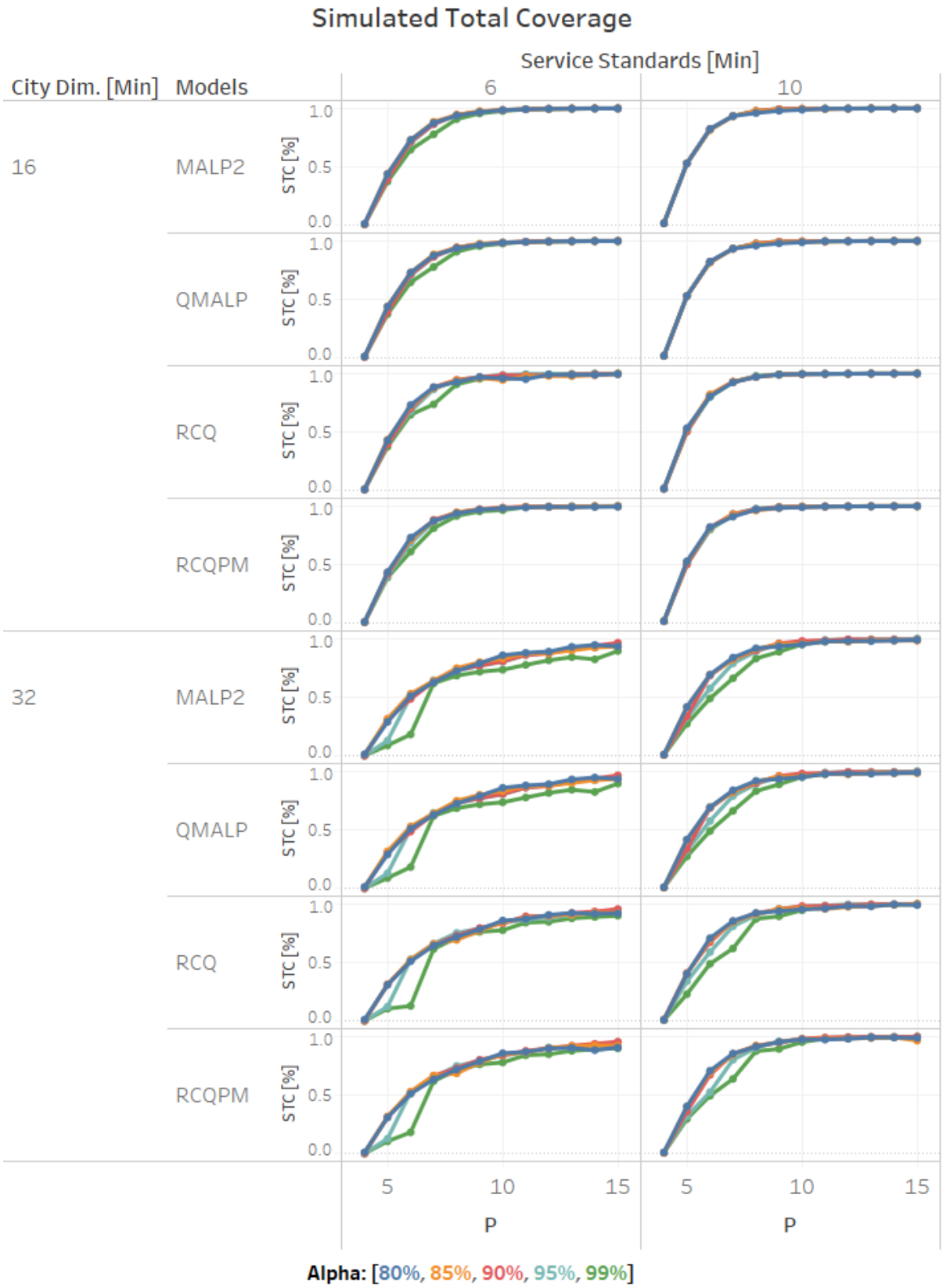

Figure 6 - Simulated total coverage: High call intensity scenario (4 CPH)

Insofar as performance across models, in Table 12 we observed similar coverage values in all tiers except within the unique maximum value tier. In the top tier, MALP 2 and QMALP were tied again and produced the highest proportion of simulated coverage values greater than or equal to the maximum total coverage value for each problem instance (Z^{TC}). However, the RC-QMALP models generated similar proportions that were within ~2% of MALP 2 and QMALP. In the other four tiers, RCQPM outperformed all other models by producing between ~4% and 7% more objective values within 1% of Z^{TC}, between ~1 and cgv4% more solutions within 2% of Z^{TC}, and between ~5 and 6% more objective values within 5% of Z^{TC} (in this case 100% of RCQPM's objective function values were within 5%).

	MALP 2		**QMALP**		**RCQ**		**RCQPM**	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	**528**	**48.89%**	**528**	**48.89%**	505	46.76%	504	46.67%
0.01	814	75.37%	814	75.37%	846	78.33%	**886**	**82.04%**
0.02	918	85.00%	918	85.00%	954	88.33%	**965**	**89.35%**
0.05	1017	94.17%	1025	94.91%	1030	95.37%	**1080**	**100.00%**
Unique	0	0.00%	0	0.00%	193	17.87%	**222**	**20.56%**

Table 12 - Highest total simulated coverage: Aggregated across all scenarios

Upon disaggregating by P and α, we observe that MALP 2 and QMALP generate better solutions with low α-reliability values (α = 80% and 85%). In Table 13 and Table 14 (below) it is evident that MALP 2 and QMALP generate more objective values equal to or near Z^{TC} for a wide range of P values. The RC-QMALP models only tied MALP 2 and QMALP at the highest tiers in very specific problem instances (at α = 80% with P = 7 and 8), however, only the RC-MALP models generated the unique maximum objective values. Furthermore, all of RCQPM's objective values remained within 5% of Z^{TC} although the other three models remained within this range for most problem instances.

		P										
	α: 80%	5	6	7	8	9	10	11	12	13	14	15
MALP 2	Max	**55.56%**	**50.00%**	44.44%	**55.56%**	**55.56%**	**66.67%**	**72.22%**	**55.56%**	**66.67%**	**72.22%**	**66.67%**
	0.01	**83.33%**	**83.33%**	**77.78%**	**88.89%**	77.78%	**83.33%**	**88.89%**	**83.33%**	**88.89%**	**88.89%**	**83.33%**
	0.02	**88.89%**	**94.44%**	**88.89%**	**100.00%**	**94.44%**	88.89%	**100.00%**	**94.44%**	**100.00%**	**100.00%**	**100.00%**
	0.05	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	**55.56%**	**50.00%**	44.44%	**55.56%**	**55.56%**	**66.67%**	**72.22%**	**55.56%**	**66.67%**	**72.22%**	**66.67%**
	0.01	**83.33%**	**83.33%**	**77.78%**	**88.89%**	77.78%	**83.33%**	**88.89%**	**83.33%**	**88.89%**	**88.89%**	**83.33%**
	0.02	**88.89%**	**94.44%**	**88.89%**	**100.00%**	**94.44%**	88.89%	**100.00%**	**94.44%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	83.33%	**100.00%**	**100.00%**	**100.00%**	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	50.00%	11.11%	**50.00%**	22.22%	22.22%	22.22%	22.22%	33.33%	27.78%	50.00%	44.44%
	0.01	61.11%	55.56%	**77.78%**	55.56%	55.56%	44.44%	61.11%	66.67%	83.33%	83.33%	66.67%
	0.02	**88.89%**	66.67%	**88.89%**	88.89%	88.89%	61.11%	77.78%	88.89%	94.44%	94.44%	**100.00%**
	0.05	**100.00%**	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	94.44%	**100.00%**
	Unique	**22.22%**	0.00%	**16.67%**	16.67%	11.11%	11.11%	11.11%	16.67%	11.11%	**22.22%**	11.11%
RCQPM	Max	22.22%	**50.00%**	**50.00%**	27.78%	33.33%	27.78%	27.78%	38.89%	33.33%	27.78%	55.56%
	0.01	61.11%	72.22%	72.22%	83.33%	**83.33%**	72.22%	77.78%	72.22%	77.78%	72.22%	72.22%
	0.02	**88.89%**	77.78%	**88.89%**	**100.00%**	**94.44%**	**94.44%**	83.33%	77.78%	88.89%	94.44%	77.78%
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	**38.89%**	**16.67%**	**27.78%**	**27.78%**	**16.67%**	**16.67%**	**22.22%**	**22.22%**	5.56%	**16.67%**

Table 13 - Highest simulated total coverage: Aggregated for all *P* and α, α = 80%

	α: 85%	5	6	7	8	9	10	11	12	13	14	15
MALP 2	Max	**61.11%**	**61.11%**	**50.00%**	**61.11%**	**61.11%**	**55.56%**	**44.44%**	33.33%	**61.11%**	50.00%	**55.56%**
	0.01	**88.89%**	**83.33%**	**66.67%**	**88.89%**	**88.89%**	**83.33%**	**88.89%**	72.22%	**88.89%**	**88.89%**	72.22%
	0.02	88.89%	**83.33%**	**88.89%**	**100.00%**	**100.00%**	83.33%	**100.00%**	88.89%	**100.00%**	**100.00%**	**100.00%**
	0.05	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	**61.11%**	**61.11%**	**50.00%**	**61.11%**	**61.11%**	**55.56%**	**44.44%**	33.33%	**61.11%**	50.00%	**55.56%**
	0.01	**88.89%**	**83.33%**	**66.67%**	**88.89%**	**88.89%**	**83.33%**	**88.89%**	72.22%	**88.89%**	**88.89%**	72.22%
	0.02	88.89%	**83.33%**	**88.89%**	**100.00%**	**100.00%**	83.33%	**100.00%**	88.89%	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	88.89%	**100.00%**	94.44%	**100.00%**	**100.00%**	**100.00%**	94.44%	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	44.44%	38.89%	27.78%	16.67%	38.89%	22.22%	**44.44%**	**61.11%**	27.78%	**61.11%**	50.00%
	0.01	55.56%	72.22%	61.11%	66.67%	66.67%	61.11%	77.78%	77.78%	83.33%	77.78%	72.22%
	0.02	88.89%	**83.33%**	83.33%	77.78%	83.33%	**88.89%**	83.33%	88.89%	**100.00%**	94.44%	**100.00%**
	0.05	**100.00%**	88.89%	**100.00%**	94.44%	**100.00%**	**100.00%**	**100.00%**	94.44%	**100.00%**	**100.00%**	**100.00%**
	Unique	**16.67%**	5.56%	16.67%	5.56%	**16.67%**	16.67%	**22.22%**	**38.89%**	11.11%	**27.78%**	**22.22%**
RCQPM	Max	27.78%	38.89%	33.33%	38.89%	22.22%	27.78%	33.33%	27.78%	44.44%	55.56%	50.00%
	0.01	61.11%	77.78%	**66.67%**	72.22%	77.78%	72.22%	77.78%	**88.89%**	**88.89%**	72.22%	**83.33%**
	0.02	**94.44%**	**83.33%**	**88.89%**	83.33%	83.33%	**88.89%**	**100.00%**	**94.44%**	88.89%	77.78%	88.89%
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	5.56%	**11.11%**	**22.22%**	**27.78%**	11.11%	**27.78%**	**22.22%**	16.67%	**27.78%**	16.67%	16.67%

Table 14 - Highest simulated total coverage: Aggregated for all *P* and α, α = 85%

With a higher α-reliability values (α = 90%), the trend begins reversing with RCQ and RCQPM producing a greater proportion of objective values greater than or equal to Z^{TC}. In Table 15 both groups performed similarly at the highest tier although the RC-QMALP models (mainly RCQPM) mostly performed better along lower *P* values while MALP 2 and QMALP

performed better at higher *P* values. Again, all of RCQPM's objective values remained within 5% of Z^{TC} while the other three models remained within this range for most problem instances.

	α: 90%	5	6	7	8	9	10	11	12	13	14	15
MALP 2	Max	27.78%	**50.00%**	33.33%	33.33%	38.89%	**44.44%**	**55.56%**	44.44%	**61.11%**	**77.78%**	**77.78%**
	0.01	44.44%	77.78%	55.56%	77.78%	88.89%	83.33%	**94.44%**	88.89%	**94.44%**	**94.44%**	**94.44%**
	0.02	50.00%	**94.44%**	61.11%	88.89%	94.44%	94.44%	94.44%	**94.44%**	**100.00%**	**100.00%**	**100.00%**
	0.05	77.78%	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	27.78%	**50.00%**	33.33%	33.33%	38.89%	**44.44%**	**55.56%**	44.44%	**61.11%**	**77.78%**	**77.78%**
	0.01	44.44%	77.78%	55.56%	77.78%	88.89%	83.33%	**94.44%**	88.89%	**94.44%**	**94.44%**	**94.44%**
	0.02	50.00%	**94.44%**	61.11%	88.89%	94.44%	94.44%	94.44%	**94.44%**	**100.00%**	**100.00%**	**100.00%**
	0.05	77.78%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	38.89%	**50.00%**	44.44%	33.33%	44.44%	38.89%	38.89%	**55.56%**	38.89%	38.89%	50.00%
	0.01	66.67%	**83.33%**	77.78%	**100.00%**	83.33%	94.44%	83.33%	88.89%	**94.44%**	**94.44%**	88.89%
	0.02	**72.22%**	**94.44%**	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**94.44%**	**100.00%**	**100.00%**	**100.00%**
	0.05	77.78%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	16.67%	**27.78%**	16.67%	11.11%	16.67%	22.22%	16.67%	**22.22%**	5.56%	5.56%	**16.67%**
RCQPM	Max	**55.56%**	27.78%	**50.00%**	**66.67%**	**50.00%**	38.89%	38.89%	38.89%	50.00%	33.33%	44.44%
	0.01	**72.22%**	77.78%	**94.44%**	**100.00%**	**94.44%**	**100.00%**	88.89%	**94.44%**	83.33%	**94.44%**	77.78%
	0.02	**72.22%**	**94.44%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**94.44%**	**100.00%**	**100.00%**	88.89%
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**38.89%**	11.11%	**27.78%**	**44.44%**	**27.78%**	**27.78%**	**27.78%**	**22.22%**	**22.22%**	**16.67%**	5.56%

Table 15 - Highest simulated total coverage: Aggregated for all *P* and α, α = 90%

At the highest α-reliability values, both RCQ and RCQPM outperform MALP 2 and QMALP by generating more objective values equal to or near Z^{TC} for a wide range of *P* values. Whereas MALP 2 and QMALP generate similar results, RCQ and RCQPM performed differently along the α and *P* dimensions. In Table 16 where α = 95%, RCQ outperformed or tied all other models along most *P* values at almost every tier. RCQPM outperformed and tied RCQ in the highest tiers of performance (Max, 0.01, etc.) in two instances with high *P* values (respectively *P* = 13 and 14) and mostly tied RCQ in between the 2nd and 4th tiers across all *P* values (notably RCQPM remained consistent the 4th tier with all its objective values remained within 5% of Z^{TC}). Nonetheless, in instances where RCQ produced more maximum objective value solutions, RCQ generated between 11 and 22% more unique solutions than RCQPM.

	α: 95%	5	6	7	8	9	10	11	12	13	14	15
MALP 2	Max	22.22%	33.33%	33.33%	27.78%	22.22%	22.22%	**66.67%**	**55.56%**	55.56%	55.56%	61.11%
	0.01	44.44%	50.00%	50.00%	66.67%	77.78%	66.67%	83.33%	77.78%	**94.44%**	88.89%	**100.00%**
	0.02	44.44%	61.11%	66.67%	77.78%	94.44%	88.89%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**
	0.05	88.89%	88.89%	94.44%	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	22.22%	33.33%	33.33%	27.78%	22.22%	22.22%	**66.67%**	**55.56%**	55.56%	55.56%	61.11%
	0.01	44.44%	50.00%	50.00%	66.67%	77.78%	66.67%	83.33%	77.78%	**94.44%**	88.89%	**100.00%**
	0.02	44.44%	61.11%	66.67%	77.78%	94.44%	88.89%	88.89%	94.44%	**100.00%**	**100.00%**	**100.00%**
	0.05	94.44%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	**50.00%**	**50.00%**	**61.11%**	**61.11%**	**77.78%**	**61.11%**	44.44%	44.44%	38.89%	**61.11%**	**66.67%**
	0.01	**77.78%**	**77.78%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	94.44%	**94.44%**	**100.00%**	**100.00%**
	0.02	**83.33%**	**83.33%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	94.44%	**100.00%**	**100.00%**	**100.00%**
	0.05	72.22%	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**27.78%**	**27.78%**	**33.33%**	**27.78%**	**33.33%**	**27.78%**	5.56%	11.11%	5.56%	**16.67%**	**27.78%**
RCQPM	Max	**50.00%**	38.89%	38.89%	50.00%	55.56%	50.00%	38.89%	50.00%	**61.11%**	**61.11%**	50.00%
	0.01	66.67%	**77.78%**	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**94.44%**	88.89%	94.44%
	0.02	66.67%	77.78%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**27.78%**	22.22%	16.67%	22.22%	11.11%	22.22%	**16.67%**	**16.67%**	**27.78%**	11.11%	5.56%

Table 16 - Highest simulated total coverage: Aggregated for all *P* and α, α = 95%

Overall at α = 99% (Table 17) RCQ and RCQPM outperformed MALP 2 and QMALP in most problem instances (QMALP's objective values were all within 5% of Z^{TC} for all $P \geq 7$) and perform similarly at all tiers although their relative performance varied along *P*. For $P \leq 7$, RCQPM topped RCQ in all tiers and problem instances (save for one tie in the 4th tier) by considerable margins that at times exceeded 50%. Beginning with $P = 8$, RCQ began to outperform RCQPM at the highest and lowest tiers in five of the eight problem instances and at least tying RCQ in six of them. However, both models tied along most of these *P* values in the 2nd through 4th tiers with a slight edge to RCQPM. Also, RCQ's improvements were not as large and just exceed 20% twice for $P = 15$ at the highest and 5th tier.

	α: 99%	5	6	7	8	9	10	11	12	13	14	15
MALP 2	Max	16.67%	33.33%	27.78%	22.22%	33.33%	33.33%	50.00%	50.00%	33.33%	**66.67%**	66.67%
	0.01	27.78%	44.44%	50.00%	55.56%	61.11%	61.11%	66.67%	88.89%	66.67%	83.33%	88.89%
	0.02	33.33%	66.67%	50.00%	66.67%	72.22%	66.67%	**88.89%**	88.89%	83.33%	83.33%	**100.00%**
	0.05	44.44%	94.44%	72.22%	77.78%	83.33%	83.33%	94.44%	**100.00%**	**100.00%**	94.44%	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	16.67%	33.33%	27.78%	22.22%	33.33%	33.33%	50.00%	50.00%	33.33%	**66.67%**	66.67%
	0.01	27.78%	44.44%	50.00%	55.56%	61.11%	61.11%	66.67%	88.89%	66.67%	83.33%	88.89%
	0.02	33.33%	66.67%	50.00%	66.67%	72.22%	66.67%	**88.89%**	88.89%	83.33%	83.33%	**100.00%**
	0.05	66.67%	72.22%	77.78%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	33.33%	44.44%	38.89%	**55.56%**	**61.11%**	61.11%	**77.78%**	**72.22%**	**77.78%**	61.11%	**100.00%**
	0.01	38.89%	55.56%	66.67%	**94.44%**	**100.00%**	**94.44%**	**88.89%**	**100.00%**	**100.00%**	94.44%	**100.00%**
	0.02	44.44%	66.67%	72.22%	**94.44%**	**100.00%**	**100.00%**	**88.89%**	**100.00%**	**100.00%**	94.44%	**100.00%**
	0.05	88.89%	83.33%	**100.00%**	94.44%	**100.00%**	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	0.00%	22.22%	22.22%	**33.33%**	**27.78%**	16.67%	**22.22%**	**22.22%**	**33.33%**	5.56%	**22.22%**
RCQPM	Max	**83.33%**	**55.56%**	**61.11%**	**55.56%**	55.56%	**77.78%**	66.67%	66.67%	66.67%	**66.67%**	77.78%
	0.01	**88.89%**	**77.78%**	**88.89%**	72.22%	**100.00%**	83.33%	**88.89%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.02	**88.89%**	**77.78%**	**88.89%**	83.33%	**100.00%**	88.89%	**88.89%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**50.00%**	**33.33%**	**50.00%**	**33.33%**	22.22%	**33.33%**	5.56%	16.67%	22.22%	**16.67%**	0.00%

Table 17 - Highest simulated total coverage: Aggregated for all *P* and α, α = 99%

To further explore total coverage performance along other parameter values we further disaggregated our simulation results along call intensity, city diameter and service time standards (as in Table 8). The most interesting results concerned the performance at the highest tier (the proportion of objective values matching Z^{TC}). Consequently, we summarize these results in Table 19 (below) and tabulate the results in Table 19.[107] In both tables we shorten the combination of MALP 2 and QMALP to MQ (both models generated similar results), RCQ to R, and RCQPM to RPM.

Demand [Call/Hr.]:		2			4	
Service Time Std.[Min]:	6	8	10	6	8	10
16 [Min]	MQ	R	R	MQ	R	MQ
24 [Min]	R	RPM	MQ	MQ	RPM	MQ
32 [Min]	R	R, RPM	RPM	RPM	MQ	RPM

Table 18 – Model(s) with most objective values matching Z^{TC}**:** Aggregated for all city diameters, call intensities, and service time standards

[107] We refer the reader to Appendix for more detailed tables.

	Totals		
	MQ	R	RPM
16 [Min]	3	3	0
24 [Min]	3	1	2
32 [Min]	1	2	4
Subtotal	**7**	**6**	**6**
2 [CPH]	2	5	3
4 [CPH]	5	1	3
Subtotal	**7**	**6**	**6**
6 [Min]	3	2	1
8 [Min]	1	3	3
10 [Min]	3	1	2
Subtotal	**7**	**6**	**6**

Table 19 – Instances where models generated the highest proportion of solutions matching Z^{TC} [Count]: Aggregated for all city diameters, call intensities, and service time standards

In Table 19 (above) we observe that the subtotals for each model are about the same along the city diameter, call intensity, and service time standard parameter but disaggregating along these dimensions reveals that they don't perform equally. As the city diameter increases, MALP 2 and QMALP's performance at the highest tier decreases, RCQ's performance dips significantly for the middle value of city diameter but remains about the same, while RCQPM's performance increases. With increasing call intensity, MALP 2 and QMALP's performance increases significantly, RCQ's performance decreases significantly, and RCQPM's performances remains unchanged. Lastly, with increasing service time standards, the patterns are not as clear. For MALP 2 and QMALP they appear to stay relatively high except for the

drop along the middle service time standard and for RCQ and RCQPM they appear to decrease and increase, respectively.

5.2.3 *Relative Model Accuracy Assessment*

The graphs in this section report the simulation performance for all four models in terms of how accurately they predicted reliable coverage for, respectively, a low call intensity scenario (2 CPH) and then a high call intensity scenario (4 CPH) under the same city diameters, service time standards, and all α-reliability values. Within each graph the difference between the predicted and simulated percentage of demand covered with α-reliability (PRC [%] – SRC [%]) is reported on the y-axis (Deviation [%]) and the number of facilities is on the x-axis (*P*).

We present the results for the low call intensity scenario in Figure 7 (below). With the small city diameter and both service time standards, we notice that that the model deviations generally decrease with *P*. At lower levels of *P* (i.e., $P = 4$) the higher α-reliability standards result in higher deviations. For the lower service time standard, the queue-based model deviations appear to be lower than the MALP 2 deviations at least until the models begin to mostly converge as the number of units being located reaches $P = 8$. Also, the queue-based models underestimate the simulated reliable coverage at $P = 5$ by between 2 and 8%. For the higher service time standard, the model deviations become more erratic although they generally decrease for $\alpha = 80\%$ and 85% across all models. With the 90% α-reliability standard the model deviations remain at the same level although they increase slightly with RCQ and RCQPM and decrease slightly with MALP 2 and QMALP. With the 95% α-reliability standard, all models further underestimate reliable coverage at $P = 4$ and then at $P = 5$, the queue-based models further overestimate reliable coverage by a couple of percentage points

(MALP 2). Lastly, at the 95% α-reliability standard the MALP 2 deviations decreased overall while the queue-based model deviations substantially increased between $P = 4$ and 6 by further underestimating simulated reliable coverage.

Figure 7 – Model deviation (*Predicted - Simulated reliability*) [%]: Low call intensity scenario (2 CPH)

We present the results for the high call intensity scenario in Figure 9. Compared with the low call intensity scenario and with the smaller city diameters, the initial deviations at $P = 4$ were substantially smaller among all models with the 80%, 85%, and 90% α-reliability standards and slightly higher with the 95% and 99% α-reliability standards. For larger values of *P*, there is a sharp increase with in the 80%, 85%, and 90% α-reliability standard lines that peak at $P = 6$ with deviations that exceed the maximums for the same lines in the low call intensity scenario. These lines converged to a deviation of about 0 at a later point (between $P = 8$ and 10). For the higher α-reliability standard, MALP 2 and the queue-based models produced different patterns. At the 95% α-reliability standard, MALP 2 deviations peaked at $P = 6$ at slightly higher level and converged with the other models at $P = 10$. The queue-based models peaked at similar levels at $P = 7$ and converged at $P = 9$. For a 95% α-reliability standard we can observe significant differences between MALP 2 and the queue-based models. With a higher call intensity, MALP 2's deviation peaked at a higher value and with a higher number of facilities ($P = 9$). In contrast, the deviations of all three queued-based models gradually declined (one with a negative deviation) to a slightly lower level as the number of units were increased to $P = 8$. All four models converged at P = 13 however.

With the larger city diameter, all four models were also rather similar overall with the smaller service standard although the MALP 2 model deviations appear to be higher than the queue-based model deviations particularly with the 99% α-reliability standard line. In comparison to the lower call intensity scenario, the various α-reliability standard lines were smoother and more gradual in their rise and decline. Moreover, the 80% α-reliability standard line shows that deviations were substantially higher with every model. With the higher service

time standard, the α-reliability standard lines are quite variable, sometimes with dramatic swings between negative and positive deviations. The models were also rather similar except MALP 2 generated higher model deviations with the 99% α-reliability standard between P = 7 and P = 9 while RCQ and RCQPM produced higher peaks with the 85% α-reliability standard at P = 7.

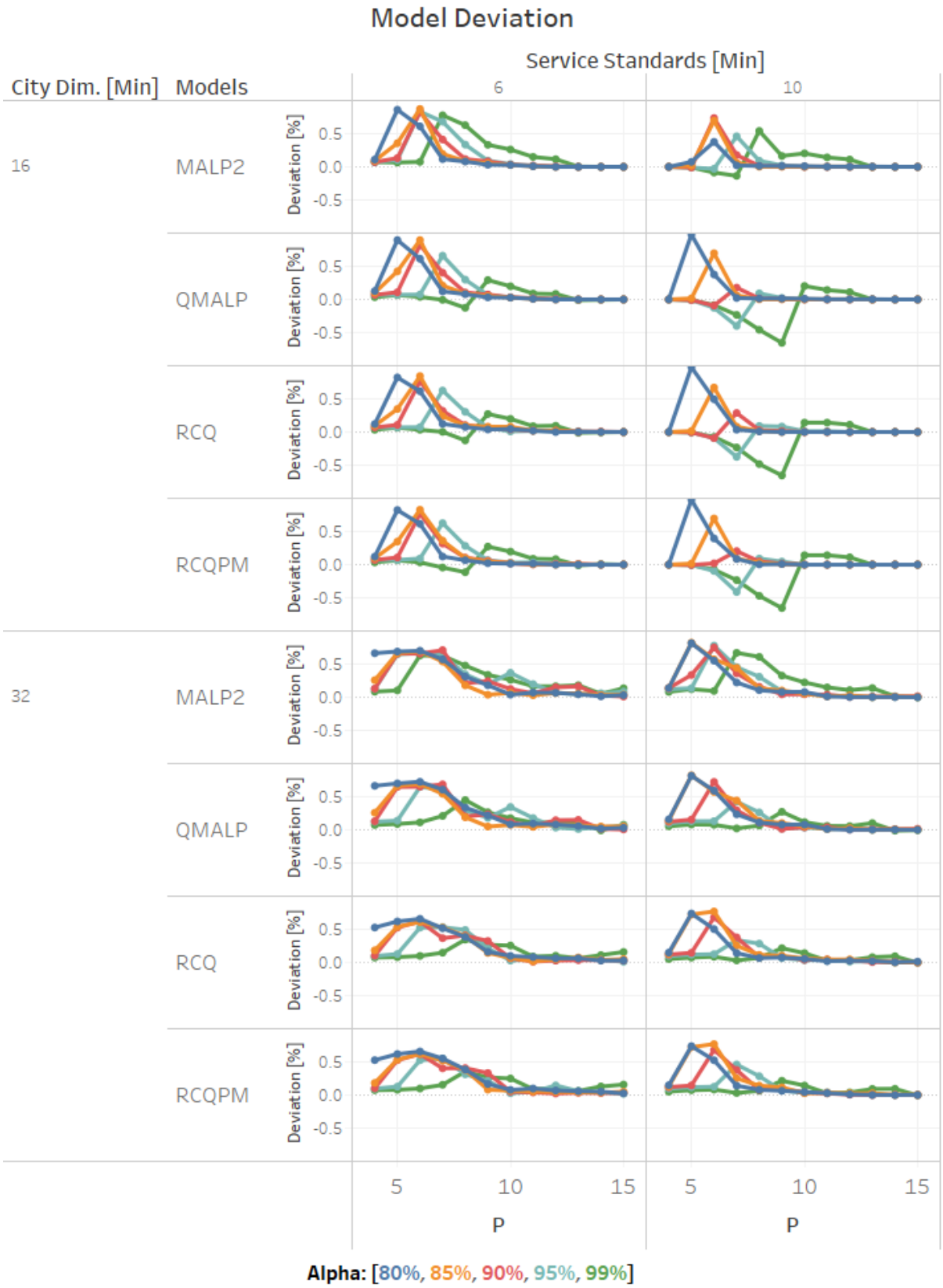

Figure 8 - Model deviation (*Predicted - Simulated reliability*) [%]: High call intensity scenario (4 CPH)

Table 20 provides summary statistics about all models aggregated across all scenarios/ parameter variations. These results suggest that RCQ outperforms the other models with smaller average deviations, variance, and a lower median for deviations. For the exception of MALP 2's minimum deviation and maximum deviation measures, the differences across models were relatively small. However, upon further disaggregating the model deviation results we found that the aggregate results held up only partly and were not consistent across other dimensions/parameters.

	MALP2	QMALP	RCQ	RCQPM
Average Dev.:	12.66%	9.55%	**8.90%**	9.07%
Population Std. Dev.:	20.56%	18.75%	**17.22%**	17.20%
Max Dev.:	**87.34%**	97.81%	97.81%	97.81%
Median Dev.:	2.73%	2.03%	**2.19%**	2.50%
Min. Dev.:	**-13.91%**	-65.47%	-65.47%	-65.47%

Table 20 – Model deviation summary statistics (*Predicted* - *Simulated reliability*) [%]: Aggregated across all scenarios

		Minimum Max. Error														
α	*P*:	4	5	6	7	8	9	10	11	12	13	14	15	*Subtotal* (P)	*Average* (P)
80%	MALP 2	0	1	0	0	1	0	0	1	0	1	0	0	4	30.27%
	QMALP	0	0	0	0	0	0	0	0	0	0	0	0	0	32.64%
	RCQ	1	0	1	1	0	1	0	0	1	0	1	1	**7**	**29.90%**
	RCQPM	1	0	1	0	0	0	0	0	0	0	0	0	2	30.47%
85%	MALP 2	0	0	0	0	1	1	1	0	1	0	0	1	**5**	**30.29%**
	QMALP	0	0	0	0	1	1	1	0	0	0	0	0	3	31.00%
	RCQ	1	1	0	0	0	0	0	1	0	0	1	0	4	31.08%
	RCQPM	1	1	1	1	0	0	0	0	0	1	0	0	**5**	31.93%
90%	MALP 2	1	0	0	0	0	0	0	0	0	0	0	1	2	34.15%
	QMALP	1	0	0	0	1	1	0	1	1	1	1	1	**8**	32.45%
	RCQ	1	1	1	0	0	0	1	0	0	0	0	1	5	**31.64%**
	RCQPM	1	1	1	1	0	0	0	0	0	0	0	0	4	32.46%
95%	MALP 2	0	0	0	0	0	0	0	0	0	0	0	0	0	40.20%
	QMALP	0	0	0	0	0	1	0	0	0	1	0	0	2	34.44%
	RCQ	1	1	1	1	0	0	1	1	1	0	1	1	**9**	**26.51%**
	RCQPM	1	1	1	1	1	0	1	0	0	0	0	0	6	27.64%
99%	MALP 2	0	0	0	0	0	0	0	0	0	0	0	0	0	40.20%
	QMALP	0	0	0	0	0	0	1	0	0	0	1	0	2	34.44%
	RCQ	1	1	1	1	1	1	0	0	0	1	1	0	**8**	**26.51%**
	RCQPM	1	1	0	1	0	1	0	1	1	0	0	0	6	27.64%

Table 21 - Instances where models generated the minimum maximum error [Count] and average minimum maximum error across all *P* values [%]: Aggregated over all city diameters, call intensities, and service time standards.

Table 21 (above) presents model deviations across α and *P*. These results are similar to what was presented in Table 20. In four of the five α-reliability levels, RCQ either produced or matched the minimal maximum error (MMXE) across all values of *P* or produced the lowest average MMXE across all values of *P*. Moreover, in three of the five α-reliability levels (for both low and high α-reliability levels), RCQ uniquely produced the most MMXEs and the lowest average MMXE value. These results are summarized below in Table 22. Note that RCQ produced almost 1.5 times more MMXEs that the second highest MMXE producer (RCQPM).

Minimum Max. Error						
α:	80	85	90	95	99	*Subtotal*
MALP 2	4	**5**	2	0	0	11
QMALP	0	3	**8**	2	2	15
RCQ	**7**	4	5	**9**	**8**	33
RCQPM	2	**5**	4	6	6	23

Table 22 - Instances where models generated the minimum maximum error in each and for all α [Count]: Aggregated over all city diameters, call intensities, service time standards, and *P*.

To address model performance over *P*, we aggregated the results given in Table 21 over α in Table 22. RCQ again frequently produced the most MXMEs, but note that this is also over most values of *P* (8 of 12) and over the widest range, that is, from *P* = 4 to *P* = 15. This range was also twice as large as the second ranked model in this category (QMALP) and 6 and 12 times as large as the third and fourth ranked models, respectively.[108]

Minimum Max. Error														
	P												Count (*P*)	Range (*P*)
MALP 2	1	1	0	0	**2**	1	1	1	1	1	0	2	1	1
QMALP	1	0	0	0	**2**	**3**	**2**	1	1	**2**	2	1	4	6
RCQ	**5**	**4**	**4**	3	1	2	**2**	**2**	**2**	1	**4**	**3**	**8**	**12**
RCQPM	**5**	**4**	**4**	**4**	1	1	1	1	1	1	0	0	4	*2*

Table 23 – Range and Instances where models generated the minimum maximum error in each and across *P* [Count]: Aggregated over all city diameters, call intensities, service time standards, and α.

We calculated the medians of deviation errors across α and *P*. In this analysis the results, reported in Table 24, only partially agree with the aggregated results. Here RCQ only produces

[108] This appears to call for some ordered statistics tests. However, the observations over *P* are not *both* independently and identically distributed (IID) - they are independent but not identically distributed. To my limited knowledge, most order statistic require IID random variables with the exception of methods based on something like the Bapat–Beg theorem (Bapat & Beg, 1989) that can consider independent but not necessarily identically distributed random variables. Unfortunately, it appears that this specific approach is not easily implementable to due inordinately high computational requirements (Glueck et al., 2008).

the minimum median of deviation errors (MMDEs) twice (although for the highest and lowest α-reliability levels) and the MMDE average once. Furthermore, unlike in the previous analysis, these measurements did not occur simultaneously. We do note, however, that no model established a majority or plurality in terms of MMDE counts (including ties), however, QMALP produced the lowest average MMDE values in the three highest α-reliability levels.

Minimum Median Error

α	*P*:	4	5	6	7	8	9	10	11	12	13	14	15	*Subtotal* (P)	*Average* (P)
80%	MALP 2	1	1	1	0	0	0	0	1	0	0	1	1	6	**5.56%**
	QMALP	0	0	0	0	0	0	0	1	0	0	1	1	3	8.24%
	RCQ	0	0	0	1	0	1	0	1	1	1	1	1	**7**	8.20%
	RCQPM	0	0	0	0	1	0	1	1	0	0	1	0	4	7.85%
85%	MALP 2	0	0	1	1	1	0	1	0	1	1	0	1	**7**	9.41%
	QMALP	0	0	0	1	0	0	0	0	1	1	0	1	4	9.79%
	RCQ	1	1	0	0	0	1	0	0	0	0	1	1	5	**9.40%**
	RCQPM	0	1	0	0	0	0	0	1	0	0	0	0	2	9.92%
90%	MALP 2	0	0	0	0	0	0	1	0	1	1	1	1	5	9.04%
	QMALP	0	1	1	1	0	1	1	0	1	1	1	1	**9**	**6.30%**
	RCQ	1	0	0	0	1	0	0	0	0	0	0	1	3	7.50%
	RCQPM	1	0	0	0	0	0	0	1	1	1	0	0	4	6.94%
95%	MALP 2	0	0	0	0	1	0	1	0	1	1	1	1	6	9.04%
	QMALP	1	0	0	1	1	1	1	1	1	1	1	1	**10**	**6.30%**
	RCQ	0	1	1	0	0	0	0	0	0	0	1	1	4	7.50%
	RCQPM	1	1	0	0	0	0	0	0	0	0	1	1	4	6.94%
99%	MALP 2	0	0	0	0	0	0	0	0	0	0	1	1	2	9.04%
	QMALP	0	0	0	0	0	1	1	0	0	0	1	1	4	**6.30%**
	RCQ	0	1	1	1	0	0	0	1	1	0	1	1	**7**	7.50%
	RCQPM	1	1	1	0	1	0	0	0	0	1	0	1	6	6.94%

Table 24 - Instances where models generated the minimum median error [Count] and average median maximum error across all *P* values [%]: Aggregated over all city diameters, call intensities, and service time standards.

To further this MDE analysis, we disaggregated the information given in Table 24 in terms of both α and *P* in Table 25 and27**.** Table 25 elucidates our observations about overall model performance along α-reliability levels, but perhaps more importantly, it shows that the difference between the top ranked model and the two second-ranked models (in terms of

counts) is not as large as in the previous analysis. Here QMALP only produces the lowest MMDEs 1.15 times more often than RCQ and MALP 2. QMALP produced the lowest MMDE 1.5 times more often than bottom-ranked RCQPM but RCQ produced 3 and 2.2 times more MMXEs than the third and fourth ranked models.

Minimum Median Error

α:	80	85	90	95	99	*Subtotal*
MALP 2	6	**7**	5	6	2	26
QMALP	3	4	**9**	**10**	4	**30**
RCQ	**7**	5	3	4	**7**	26
RCQPM	4	2	4	4	6	20

Table 25 - Instances where models generated the minimum median error in each and for all α [Count]: Aggregated over all city diameters, call intensities, service time standards, and *P*.

Likewise, QMALP also produces the MMDE over the largest number of values of *P* over a wide range, but its relative performance is questionable. RCQ produced MMXEs for twice as many values of *P* than the second and third ranked models and for eight times as many values of *P* than the fourth ranked model. In contrast, the subtotals in Table 26 indicate that QMALP produced MMDE's for just as many values of *P* as the second ranked model (MALP 2) and 1.75 times as the third and fourth ranked models. Moreover, MMDEs of RCQ appear over a wider range of *P* (11) than second ranked QMALP (10 – tied for second) and the MMDEs QMALP are 1.25 times more frequent than the fourth ranked model (RCQPM).

Median Max. Error

	P												Count (*P*)	Range
MALP 2	1	1	**2**	1	**2**	0	**3**	1	**3**	**3**	**4**	**5**	**7**	10
QMALP	1	1	1	**3**	1	**3**	**3**	2	**3**	**3**	**4**	**5**	**7**	10
RCQ	2	**3**	**2**	2	1	2	0	2	2	1	**4**	**5**	4	**11**
RCQPM	**3**	**3**	1	0	**2**	0	1	**3**	1	2	2	2	4	8

Table 26 – Range and instances where models generated the minimum maximum error in each and across all *P* [Count]: Aggregated over all city diameters, call intensities, service time standards, and α.

To further explore model performance (particularly given the discrepancy between the two analyses), we disaggregated the model deviation results along city diameters, call intensities, and service time standards while aggregating along P and α. We summarize the details of these tables grouped by call intensity, city diameter, and service time standards.

Beginning with the low call intensity scenario (Table 27), with a 16-minute city diameter, RCQ generates the highest number of MMXEs and MMDE as well as the lowest average MMXE and MMDE values. QMALP also tied RCQ's counts but only matched the average MMDE value and RCQPM only matched the MMXE count. With the 24-minute city diameter, RCQ again generated the highest MMXE and MMDE counts along with the lowest average for both measures although QMALP matched the MMDE count while RCQPM matched the MMXE count and average. With the 32-minute city diameter, RCQ generated the highest MMXE and MMDE counts while QMALP matched the MMDE count but with the lowest average while RCQPM matched the MMXE count also with the lowest average.

			Demand [Call/Hr.]:		2		
Service Time Std. [Min]:			6	8	10	*Subtotal*	*Average*
16 [Min.]	**MALP 2**	Max	0	0	1	1	68.07%
		Median	0	1	1	2	0.10%
	QMALP	Max	0	1	1	**2**	62.81%
		Median	1	1	1	**3**	**0.00%**
	RCQ	Max	0	1	1	**2**	**60.36%**
		Median	1	1	1	**3**	**0.00%**
	RCQPM	Max	1	1	0	**2**	61.25%
		Median	0	1	1	2	0.10%
24 [Min.]	**MALP 2**	Max	0	0	0	0	73.07%
		Median	0	0	1	1	0.99%
	QMALP	Max	0	0	0	0	58.75%
		Median	1	1	1	**3**	**0.10%**
	RCQ	Max	1	1	1	**3**	**52.60%**
		Median	1	1	1	**3**	**0.10%**
	RCQPM	Max	1	1	1	**3**	**52.60%**
		Median	0	0	0	0	1.98%
32 [Min.]	**MALP 2**	Max	0	0	0	0	68.54%
		Median	0	0	0	0	3.46%
	QMALP	Max	0	0	0	0	60.89%
		Median	1	1	0	**2**	**0.91%**
	RCQ	Max	1	1	1	**3**	51.72%
		Median	0	1	1	**2**	0.94%
	RCQPM	Max	1	1	1	**3**	**49.79%**
		Median	0	0	0	0	4.53%

Table 27 - Instances where models generated the minimum maximum and median error in each and across all city diameters and service standards (2 CPH) [Count] and the average values [%]: Aggregated over all *P* and α values.

In Table 28 the results of Table 28 are aggregated city diameters. Here, we observe that RCQ generates the highest MMXE and MMDE counts although QMALP and RCQPM match the MMDE and MMXE counts. RCQ generates the longest range for both MMXE and MMDE values from the high end of service time standards while QMALP and RCQPM, respectively, match the MMDE and MMXE range but from the lower range of service time standards.

		Demand [Call/Hr.]:	2			
Service Time Std. [Min]:		6	8	10	*Subtotal*	*Range*
MALP 2	Max	0	0	1	1	0
	Median	0	1	2	3	0
QMALP	Max	0	1	1	2	0
	Median	**3**	**3**	2	**8**	**2**
RCQ	Max	2	**3**	**3**	**8**	**2**
	Median	2	**3**	**3**	**8**	**2**
RCQPM	Max	**3**	**3**	2	**8**	**2**
	Median	0	1	1	2	0

Table 28 - Instances where models generated the minimum maximum and median error in each and across all service standards (2 CPH) [Count] and the average values [%]: Aggregated over all city diameters, *P*, and α values.

With the high call intensity scenario (Table 29), with a 16-minute city diameter, MALP 2 generates the highest MMXE count with the lowest average value. Both QMALP and RCQ generated the highest MMDE counts along with the lowest average value. With the 24-minute city diameter, RCQ generated both the highest MMXE and MMDE counts along with the lowest average values for both measures. QMALP and RCQPM matched RCQ on counts and the lowest average value but only for the MMDE and MMXE measures, respectively. With the 32-minute city diameter, RCQ generated both the highest MMDE and MMXE counts along with the lowest average values for both measures although RCQPM matched the MMXE count and MMDE values.

			Demand [Call/Hr.]:		4		
Service Time Std. [Min]:			6	8	10	*Subtotal*	*Average*
16 [Min.]	MALP 2	Max	0	1	1	**2**	**82.50%**
		Median	0	0	1	1	2.89%
	QMALP 2	Max	0	0	0	0	94.95%
		Median	1	1	1	**3**	**0.00%**
	RCQ	Max	0	0	0	0	93.23%
		Median	1	1	1	**3**	**0.00%**
	RCQPM	Max	1	0	0	1	92.71%
		Median	0	0	1	1	1.51%
24 [Min.]	MALP 2	Max	0	0	1	1	80.78%
		Median	0	0	0	0	9.06%
	QMALP 2	Max	0	0	0	0	83.49%
		Median	0	1	1	**2**	1.04%
	RCQ	Max	1	1	0	**2**	**76.20%**
		Median	1	0	1	**2**	**0.99%**
	RCQPM	Max	1	1	0	**2**	**76.20%**
		Median	0	0	0	0	6.67%
32 [Min.]	MALP 2	Max	0	0	0	0	75.94%
		Median	0	0	0	0	13.26%
	QMALP 2	Max	0	0	0	0	76.67%
		Median	0	0	1	1	3.52%
	RCQ	Max	1	1	1	**3**	**69.11%**
		Median	1	1	1	**3**	**2.79%**
	RCQPM	Max	1	1	1	**3**	**69.11%**
		Median	0	0	0	0	9.17%

Table 29 - Instances where models generated the minimum maximum and median error in each and across all city diameters and service standards (4 CPH) [Count] and the average values [%]: Aggregated over all *P* and α values.

In Table 30 we further summarize results of Table 29 by aggregating across city diameters. Here, we observe that RCQ generates the highest MMDE counts and the longest MMDE range while RCQPM generates the highest MMXE counts and range from the lower service time standard.

		Demand [Call/Hr.]: 4				
		6	8	10	*Subtotal*	*Range*
MALP 2	Max	0	1	2	3	0
	Median	0	0	1	1	0
QMALP	Max	0	0	0	0	0
	Median	1	**2**	**3**	6	2
RCQ	Max	2	**2**	1	5	1
	Median	**3**	**2**	**3**	**8**	**3**
RCQPM	Max	**3**	**2**	1	**6**	**2**
	Median	0	0	1	1	0

Table 30 - Instances where models generated the minimum maximum and median error in each and across all service standards (4 CPH) [Count] and the average values [%]: Aggregated over all city diameters, *P*, and α values.

To conclude this section, we offer two final tables where we present the results where we have aggregated across call intensity (Table 31) and then along service time standards (Table 32) to return to a different but fully high-level view. In our analysis, these tables summarize the findings about model deviation in this section and support the results from Table 20, although in a more nuanced manner.

	Service Time Std. [Min]:		6	8	10	*Subtotal*	*Range*
16 [Min]	**MALP 2**	Max	0	**1**	**2**	**3**	**2**
		Median	0	1	**2**	3	1
	QMALP	Max	0	**1**	1	2	1
		Median	**2**	**2**	**2**	**6**	**3**
	RCQ	Max	0	**1**	1	2	1
		Median	**2**	**2**	**2**	**6**	**3**
	RCQPM	Max	**2**	**1**	0	**3**	**2**
		Median	0	1	**2**	3	1
24 [Min]	**MALP 2**	Max	0	0	**1**	1	1
		Median	0	0	1	1	0
	QMALP	Max	0	0	0	0	0
		Median	1	**2**	**2**	**5**	2
	RCQ	Max	**2**	**2**	**1**	**5**	**3**
		Median	**2**	1	**2**	**5**	**3**
	RCQPM	Max	**2**	**2**	**1**	**5**	**3**
		Median	0	0	0	0	0
32 [Min]	**MALP 2**	Max	0	0	0	0	0
		Median	0	0	0	0	0
	QMALP	Max	0	0	0	0	0
		Median	1	1	1	3	0
	RCQ	Max	**2**	**2**	**2**	**6**	**3**
		Median	1	**2**	**2**	**5**	**2**
	RCQPM	Max	**2**	**2**	**2**	**6**	**3**
		Median	0	0	0	0	0

Table 31 - Instances where models generated the minimum maximum and median error in each and across all city diameters and service standards [Count] and the average values [%]: Aggregated over all *P* and α values, and call intensities.

From Table 31, one can see that for the 16-minute city diameter both MALP 2 and RCQPM generated the highest MMXE count and longest/complete MMXE range while both QMALP and RCQPM generated the highest MMDE count and MMDE range. MALP 2 performed well with longer service time standard while RCQPM was better at the lower end. For larger city sizes MALP 2 is virtually absent in generating the best values for MMXE and MMDE. In contrast, the other three models appeared to generate higher MMDE counts, RCQPM with MMXE counts and range, and RCQ with both MMDE and MMXE counts and ranges. Notably, MALP

and QMALP failed to match that the performance of RCQ and RCQPM for larger city diameters.

Service Time Std. [Min]:		6	8	10	*Subtotal*	*Range*
MALP 2	Max	0	1	3	4	0
	Median	0	1	3	4	0
QMALP	Max	0	1	1	2	0
	Median	4	**5**	5	14	1
RCQ	Max	4	**5**	**4**	13	**2**
	Median	**5**	**5**	**6**	**16**	**3**
RCQPM	Max	**6**	**5**	3	**14**	**2**
	Median	0	1	2	3	0

Table 32 - Instances where models generated the minimum maximum and median error in each and across all service standards [Count] and the average values [%]: Aggregated over all *P* and α values, city diameters, and call intensities.

Finally, in Table 32 we tabulated the total MMXE and MMDE counts over every dimension. Again, we observe the same relative strengths of QMALP, RCQ, and RCQPM, however, we note that only marginal differences for both the subtotals and ranges. This again would suggest that the models are rather similar but this is misleading because as noted about they perform differently under different conditions.

5.2.4 *Interdisctrict Reliability Constraints*

In this section, we examine the performance of RCQ and RCQPM. In Table 32 and Table 33 we compared the predicted reliable coverage values across all RCQMALP variations (the version corresponds to the four different constraint classes outlined in Section 5.1.2). What we found was that the models predicted similar objective values with the exception of the Class C variation where idle capacity is restricted to the server location. Also, all model classes (except

C) generated objective values that were all within 5% of the best solution of a similarly parameterized model.

	RCQMALP		RCQMALPA		RCQMALPB		RCQMALPC		RCQMALPD	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	1,037	96.02%	1,039	96.20%	1,039	96.20%	505	46.76%	1,039	96.20%
0.01	1,058	97.96%	1,060	98.15%	1,060	98.15%	579	53.61%	1,060	98.15%
0.02	1,073	99.35%	1,073	99.35%	1,073	99.35%	679	62.87%	1,073	99.35%
0.05	**1,080**	**100.00%**	**1,080**	**100.00%**	**1,080**	**100.00%**	842	77.96%	**1,080**	**100.00%**
Unique	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%

Table 33 - Highest predicted reliable coverage (Non-PMP models): Aggregated across all scenarios

	RCQMALPPM		RCQMALPPMA		RCQMALPPMB		RCQMALPPMC		RCQMALPPMD	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	1,037	96.02%	1,037	96.02%	**1,080**	**100.00%**	505	46.76%	**1,080**	**100.00%**
0.01	1,058	97.96%	1,058	97.96%	**1,080**	**100.00%**	579	53.61%	**1,080**	**100.00%**
0.02	1,073	99.35%	1,073	99.35%	**1,080**	**100.00%**	679	62.87%	**1,080**	**100.00%**
0.05	**1,080**	**100.00%**	**1,080**	**100.00%**	**1,080**	**100.00%**	842	77.96%	**1,080**	**100.00%**
Unique	0	0.00%	0.00%	0.00%	0	0.00%	0	0.00%	0	0.00%

Table 34 - Highest predicted reliable coverage (PMP models): Aggregated across all scenarios

Likewise, the model simulations suggest that there are no pronounced differences between the PMP and the non-PMP versions of the model. In any case, RC-QMALPB and RC-QMALPD appear to perform the best with a slight edge to RC-QMALPB in the fourth tier and RC-QMALPD in the second and third tiers.

	RCQMALP		RCQMALPA		RCQMALPB		RCQMALPC		RCQMALPD	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	512	47.41%	441	40.83%	**545**	**50.46%**	501	46.39%	**545**	**50.46%**
0.01	659	61.02%	608	56.30%	683	63.24%	663	61.39%	**686**	**63.52%**
0.02	761	70.46%	719	66.57%	782	72.41%	749	69.35%	**785**	**72.69%**
0.05	894	82.78%	879	81.39%	**912**	**84.44%**	893	82.69%	910	84.26%
Unique	26	2.41%	33	3.06%	2	0.19%	57	5.28%	2	0.19%

Table 35 - Highest simulated reliable coverage (Non-PMP models): Aggregated across all scenarios

	RCQMALPPM		RCQMALPPMA		RCQMALPPMB		RCQMALPPMC		RCQMALPPMD	
	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]	[Count]	[%]
Max	446	41.30%	470	43.52%	550	50.93%	540	50.00%	541	50.09%
0.01	659	61.02%	663	61.39%	689	63.80%	664	61.48%	682	63.15%
0.02	736	68.15%	744	68.89%	770	71.30%	755	69.91%	777	71.94%
0.05	883	81.76%	881	81.57%	886	82.04%	884	81.85%	883	81.76%
Unique	8	0.74%	10	0.93%	46	4.26%	**59**	**5.46%**	31	2.87%

Table 36 - Highest simulated reliable coverage (PMP models): Aggregated across all scenarios

In **Error! Reference source not found.**, **Error! Reference source not found.**, Figure 12, an d **Error! Reference source not found.** (below) we provide the SRC [%] graphs for all model variations. We note that the models are mostly similar along all dimensions although slight artifacts appear in some cases but they do not appear to establish any consistent pattern. As such, we conclude that the PMP models have little to offer in terms of increasing reliability considering that the differences between RCQ or RCQPM and QMALP or MALP were significantly higher than the differences among these RC-QMALP variations.

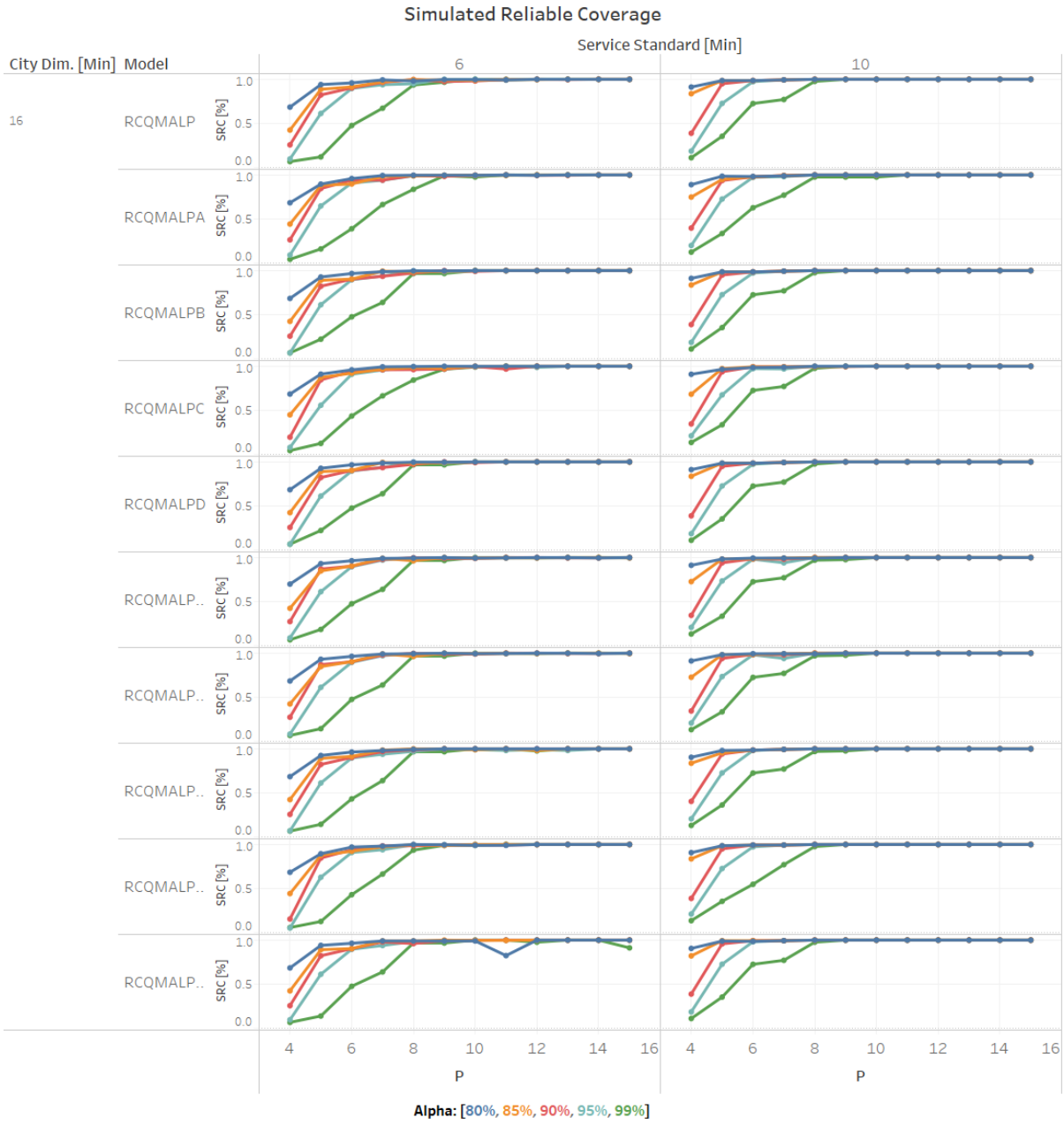

Figure 9 - Simulated α-reliable coverage: Low call intensity scenario (2 CPH) and small city diameter (16 [Min])

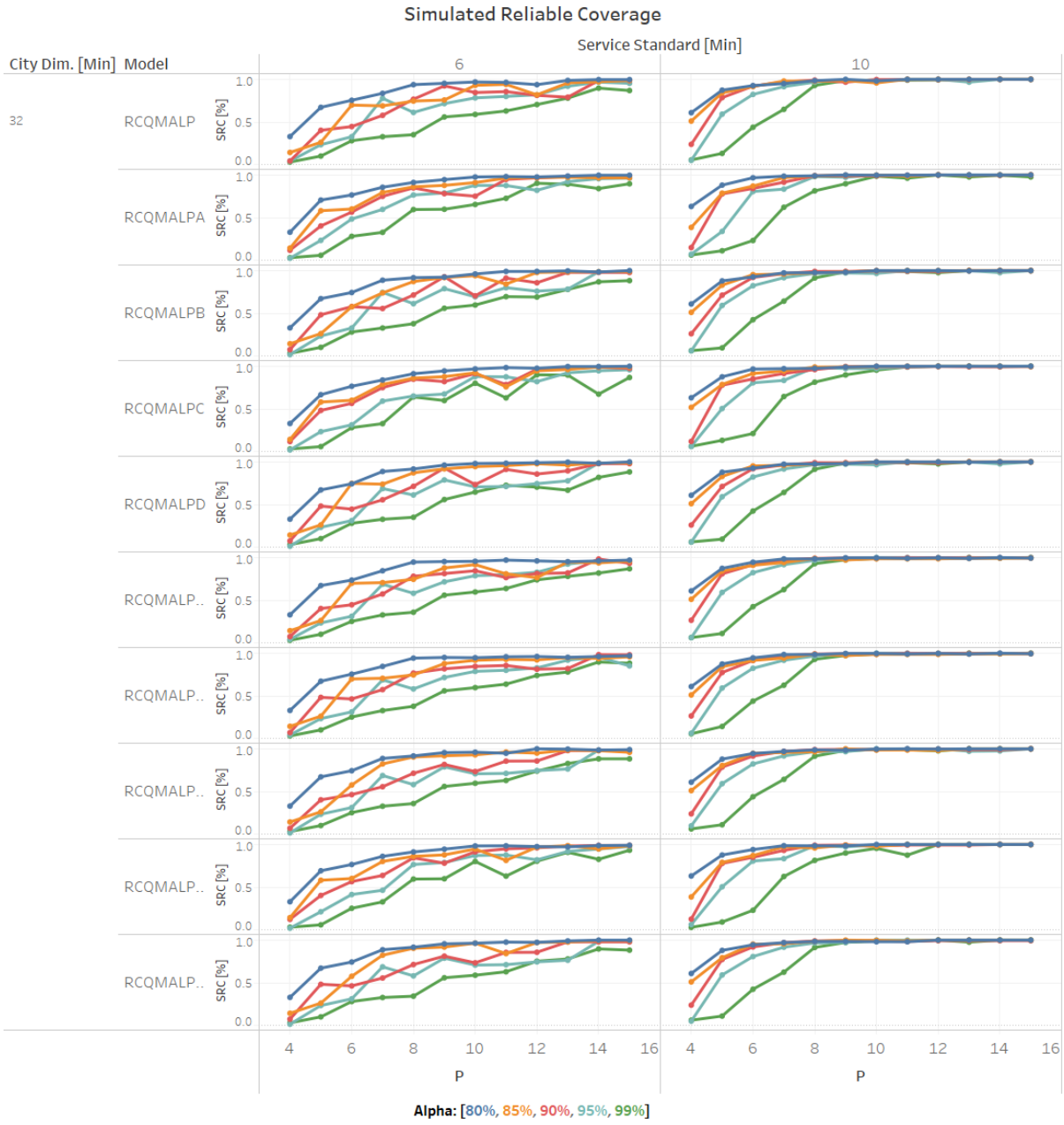

Figure 10 - Simulated α-reliable coverage: Low call intensity scenario (2 CPH) and large city diameter (32 [Min])

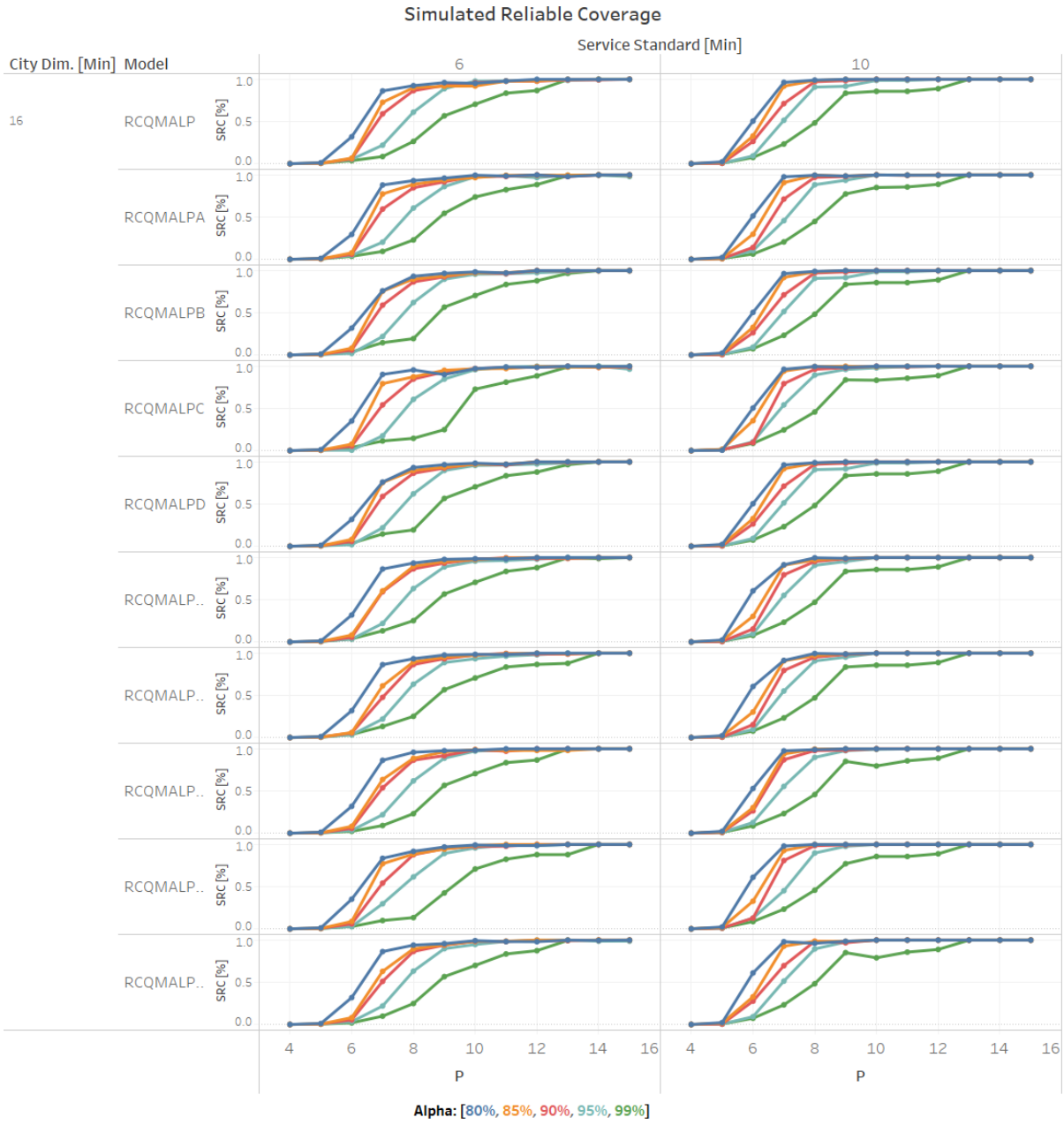

Figure 11 - Simulated α-reliable coverage: High call intensity scenario (4 CPH) and small city diameter (16 [Min])

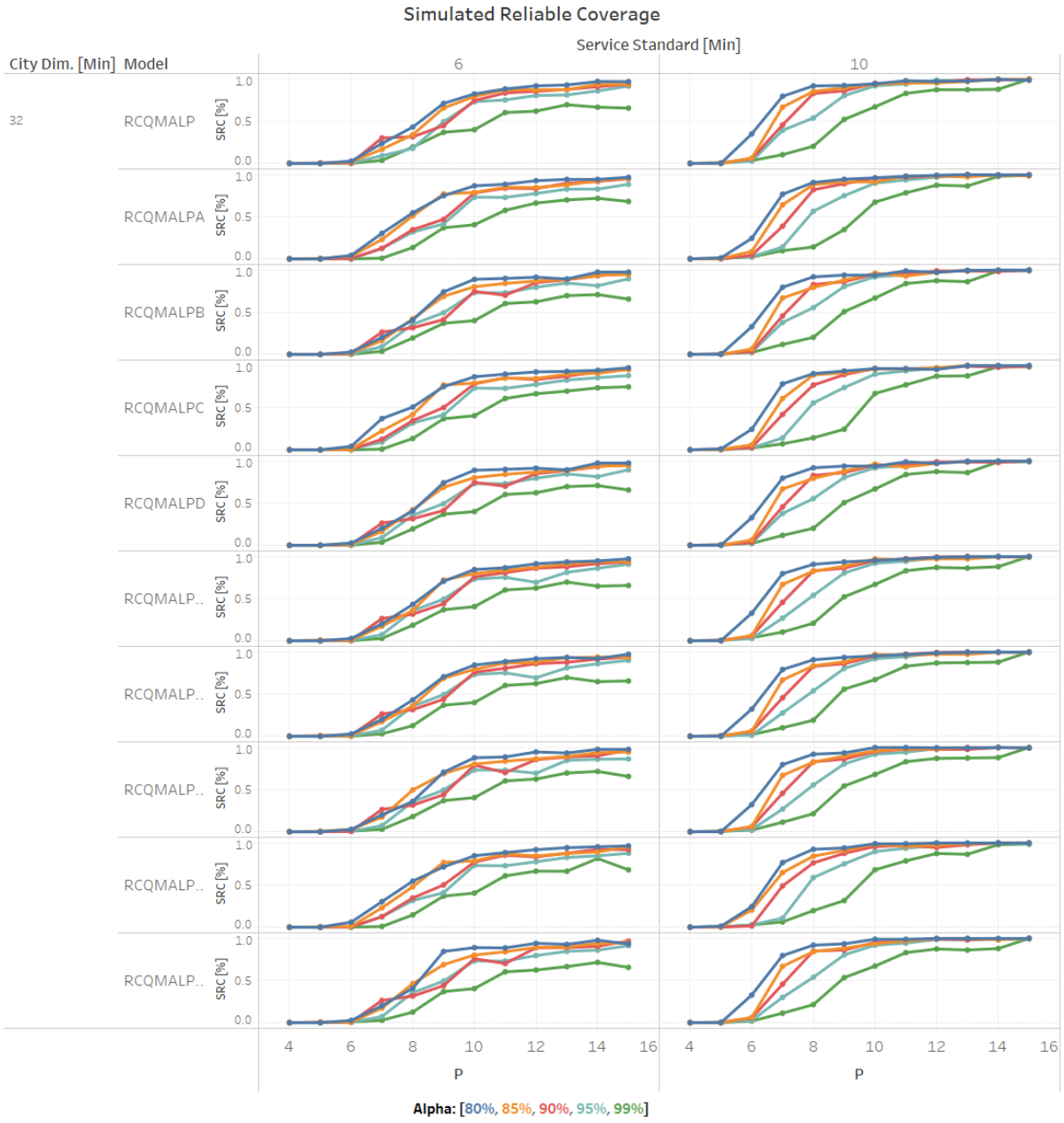

Figure 12 - Simulated α-reliable coverage: High call intensity scenario (4 CPH) and large city diameter (32 [Min])

6. CONCLUSION

The objectives of this thesis were twofold: (1) review the history of EMSS location models and EMSSs in general; and (2) develop a new resource constrained model building upon the components of the MALP modeling paradigm of ReVelle and Hogan (ReVelle and Hogan, 1989). The history of EMSS speaks volumes about the complexity of EMSS planning because of problems that include a lack of information about patients, financing issues, technological limits, the state of emergency medical science research, and a patchwork of varying legal and regulatory frameworks. Notably, most of these problems have been identified and discussed for over 50 years in the Unites States and yet they persist to this day. The review has also included a review of the fundamental modeling approaches that have been developed to analyze and plan EMS systems, ranging from the Hypercube queuing approach to the Maximal Availability Location problem (MALP). This review also included the current concerns within the medical community and the "modeling" community.

The MALP modeling paradigm and the related Queuing Based MALP (QMALP) (Marianov and ReVelle, 1989) are fundamentally important approaches to modeling probabilistic and stochastic elements found in EMSS. They have been applied in real-world situations and have influenced the development of many new models including variations of the models themselves. Nonetheless, an increasing number of publications have questioned both the applicability and foundations of these two models. Some publications have questioned the validity of these models' assumptions (Baron et al., 2009; Murray & Church, 1992) and others have questioned the usefulness of the modeling approach in the area of EMSS ambulance deployment (Erkut et al., 2008).

The development of the Resourced Constrained QMALP (RC-QMALP) model represented an attempt to answer these critiques by implementing within QMALP a location-allocation framework. We hoped that adding a resource constrained framework would address some of biggest flaws of MALP and QMALP, namely, that of relaxing the districting assumption that had very little theoretical support, and demonstrating the validity of reliability constraints.

To test RC-QMALP we used a simulation method and a subjective comparison approach to validate our new model and test MALP2 and QMALP. We stress that the latter represented a response to the limits with statistical approaches that require much more thought that is beyond the scope of this thesis (in terms of establishing more sophisticated experiments) but also technical limits given that the observed solutions (i.e., simulations under varying parameters) are not identically distributed. With these limited efforts, we argue that even though RC-QMALP was not the best model in terms of producing the locational solutions with the highest reliable coverage, it produced solutions that were desirable in other ways including with respect to their accuracy and the total coverage that they produced.

RCQ was the more balanced of all models, producing better solutions as measured by the median of deviation errors (MMDE) and Minimum Maximum Errors (MMXE) across most parameter values, while the RCQPM tended to produce better solutions as measured by the average of the Minimum Maximum Errors (MMXE) under lower service time standards and larger city diameters. Also, RCQPM always produced solutions within 5% of the best-found configurations as estimated by simulation. MALP 2 and QMALP produced the highest proportion of solutions that had the highest reliable performance. Overall, RCQ and RCQPM generated the highest proportion of unique optimal solutions. Consequently, the overall results are somewhat mixed, in that there was no clear winner over all categories of comparison.

There remain two promising versions of RC-QMALP (RC-QMALPB and RC-QMALPD) that are based upon relaxing the idle capacity constraint. These models were compared to the basic versions of RC-QMALP (RCQ and RCQPM) and although they did not significantly improve upon RCQ and RCQPM in terms of reliable coverage these improvements suggest that they might perform well against MALP2 and QMALP in other respects, a task left for future research. There are also issues that should be addressed with respect to the simulation model that was used to test the validity of all model solutions. This simulation model maintains queues of calls whereas the underlying assumption of most EMS models is that if a queue occurs, calls will be either dropped or handled by a different service. Because queues do form in some of the simulations, the current results may be overly conservative in estimating expected and reliable coverage.

REFERENCES

Aboolian, R., Berman, O., & Drezner, Z. (2008). Location and allocation of service units on a congested network. *IIE Transactions, 40*(4), 422–433.

Aboueljinane, L., Sahin, E., & Jemai, Z. (2013). A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering*, *66*(4), 734–750.

Achabal, D. D. (1975). *The Development of spatial delivery system for emergency medical services*. Univesity of Texas.

Achabal, D. D. (1978). The development of a spatial delivery system for emergency medical services. *Geographical Analysis, 10*(1), 47–63.

Alminana, M., Borras, F., & Pastor, J. T. (1996). The Centralized Probabilsitic Location Set

Covering Problem. *Studies in Locational Analysis*, (9), 5–8.

Anderson, M. (1957). Injuries caused by motor vehicles a statistical analysis. *California Medicine*, *86*(Feburary), 115–118.

Aringhieri, R., Bruni, M. E., Khodaparasti, S., & van Essen, J. T. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers and Operations Research*, *78*(September 2016), 349–368.

Austin, C. M. (1974). The Evaluation of Urban Public Facility Location: An Alternative to Benefit-Cost Analysis. *Geographical Analysis*, *6*(2), 135–145.

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

Balskovitz, A. (2011, April 6). The ins and outs of EMS. *Lansing City Pulse*. Lansing, Michigan.

Barkley, K. T. (1974). The history of the ambulance. In *Proceedings, International Congress of the History of Medicine* (Vol. 23, pp. 456–466).

Barkley, K. T. (1978). *The ambulance: the story of emergency transportation of sick and wounded through the centuries*. Exposition Press.

Baron, O., Berman, O., Kim, S., & Krass, D. (2009). Ensuring feasibility in location problems with stochastic demands and congestion. *IIE Transactions*, *41*(5), 467–481.

Barton, R. R. (2013). Designing simulation experiments. In *Winter Simulation Confrence* (pp. 2223–2231).

Bass, R. R. (2015). History of EMS. In D. Cone, J. H. Brice, T. R. Delbridge, & B. Myers

(Eds.), *Emergency Medical Services: Clinical Practice and Systems Oversight* (Second, Vol. 1, pp. 1–16). Wiley.

Batta, R. (1989). The Stochastic Queue Median Over a Finite Discrete Set. *Operations Research*, *37*(4), 648–652.

Batta, R., Dolan, J. M., & Krishnamurthy, N. N. (1989). The Maximal Expected Covering Problem: Revisited. *Transportation Science*, *23*(4), 277–287.

Bell, C. E., & Allen, D. (1969). Optimal Planning of an Emergency Ambulance Service. *Socio-Economic Planning Sciences*, *3*(2), 95–101.

Benedict, J. M. (1983). *Three hierarchical objective models which incorporate the concept of excess coverage to locate EMS vehicles or hospitals*. Northwestern University.

Berlin, G. (1972). *Facility Location and Vehicle Allocation for Provision of an Emergency Service*. Johns Hopkins University, Baltimore.

Berlin, G. N., & Liebman, J. C. (1974). Mathematical analysis of emergency ambulance location. *Socio-Economic Planning Sciences*, *8*(6), 323–328.

Berman, O., & Krass, D. (2001). Facility Location Problems with Stochastic Demands and Congestion. In Z. Drezner & H. W. Hamacher (Eds.), *Facility Location: Applications and Theory* (pp. 329–371). Springer-Verlag Berlin.

Berman, O., & Krass, D. (2015). Stochastic Location Models with Congestion. In G. Laporte, S. Nickel, & F. S. da Gama (Eds.), *Location Science* (pp. 443–486). Springer.

Berman, O., Krass, D., & Menezes, M. B. C. (2007). Facility Reliability Issues in Network p-Median Problems: Strategic Centralization and Co-Location Effects. *Operations*

Research, *55*(2), 332–350.

Berman, O., & Larson, R. C. (1982). The median problem with congestion. *Computers and Operations Research*, *9*(2), 119–126.

Berman, O., & Larson, R. C. (1985). Optimal 2-Facility Network Districting in the Presence of Queuing. *Transportation Science*, *19*(3), 261–277.

Berman, O., Larson, R. C., & Odoni, A. R. (1981). Developments in network location with mobile and congested facilities. *European Journal of Operational Research*, *6*(2), 104–116.

Berman, O., Larson, R. C., & Parkan, C. (1987). The Stochastic Queue p-Median Problem. *Transportation Science*, *21*(3), 207–216.

Berman, O., Larson, R., & Chiu, S. (1985). Optimal server location on a network operating as an M/G/1 queue. *Operations Research*, *33*(4), 746–771.

Berman, O., & Mandowsky, R. R. (1986). Location-allocation on Congested Networks. *European Journal of Operational Research*, *26*(2), 238–250.

Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica: Journal of the Econometric Society*, 23–36.

Bigman, D., & ReVelle, C. (1978). The theory of welfare considerations in public facility location problems. *Geographical Analysis*, *10*(3), 229–240.

Black, R. P. (1969). *Fleet Size Requirements of a General Stretch Ambulance Service for an Urban Area* (D. R. C. Thesis). University of Strathclyde, Scotland.

Blackwell, T. H. (1993). Prehospital care. *Emergency Medical Clinics of North America*, *11*(1), 1–14.

Borrás, F., & Pastor, J. T. (2002). The ex-post evaluation of the minimum local reliability level: an enhanced probabilistic location set covering model. *Annals of Operations Research*, *111*(1), 51–74.

Boyd, D. R. (1976). Emergency medical services systems development: A national initiative. *IEEE Transactions on Vehicular Technology*, *25*(4), 104–115.

Boyd, D. R., & Cowley, R. A. (1983). Comprehensive Regional Trauma/Emergency Medical Services (EMS) Delivery Systems: The United States Experience. *World Journal of Surgery*, *7*(1), 149–157.

Boyd, D. R., Mains, K. D., & Flashner, B. A. (1973). A Systems approach to statewide emergency medical care. *The Journal of Trauma and Acute Care Surgery*, *13*(4), 276–284.

Boyd, D. R., Micik, S. H., Lambrew, C. T., & Romano, T. L. (1979). Medical control and accountability of emergency medical services (EMS) systems. *IEEE Transactions on Vehicular Technology*, *28*(4), 249–262.

Briggs, A. E., & Palmer, F. C. (1963). Regulation of Emergency Services. *Public Health Reports*, *78*(1), 41.

Bureau of Municipal Research. (1955). *Operation of emergency vehicles*. Syracuse, New York.

Burman, D. Y. (1981). Insensitivity in Queueing Systems. *Advances in Applied Probability*, *13*(4), 846–859.

Caldentey, R. A., & Kaplan, E. H. (2007). *A Heavy Traffic Approximation for Queues with Restricted Customer-Server Matchings* (SSRN Scholarly Paper No. ID 1293130). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=1293130

Caldwell, L. A. (1961). Ambulance services and traffic casualties. *Ontario Medical Review*, *28*, 172–182.

Cales, R. H., & Trunkey, D. D. (1985). Preventable Trauma Deaths: A Review of Trauma Care Systems Development. *JAMA*, *254*(8), 1059–1063.

Carter, G. M. M., Chaiken, J. M. M., & Ignall, E. (1972). Response areas for two emergency units. *Operations Research*, *20*(3), 571–594.

Chabria, A. (2016, August). Sacramento fire crews arrive nearly 2 minutes later than national standard. *The Sacramento Bee*. Retrieved from http://www.sacbee.com/news/local/article94439132.html

Chaiken, J. M. (1971). *Allocation of Emergency Units: Response Areas*. *P-4745*. New York.

Chanta, S., Mayorga, M., Mclay, L., & Wiecek, M. (2009). A Bi-Objective Covering Location Model for EMS Systems. *Proceedings of the 2009 Industrial Engineering Research Conference*, (February 2015), 1868–1873.

Chapman, S., & White, J. (1974). Probabilistic formulations of emergency service facilities location problems. In *ORSA/TIMS Conference, San Juan, Puerto Rico*.

Chiyoshi, F. Y., Galvão, R. D., & Morabito, R. (2003). A note on solutions to the maximal expected covering location problem. *Computers and Operations Research*, *30*(1), 87–96.

Chung, C. H., Schilling, D. A., & Carbone, R. (1983). The capacitated maximal covering problem: A heuristic. In *Proceedings of Fourteenth Annual Pittsburgh Conference on Modeling and Simulation* (pp. 1423–1428).

Church, R. L. (1974). *Synthesis of a Class of Public Facilities Location Models*. The John Hopkins University.

Church, R. L., & ReVelle, C. S. (1976). Theoretical and Computational Links between the p-Median, Location Set-covering, and the Maximal Covering Location Problem. *Geographical Analysis, 8*(October), 406–415.

Church, R. L., Scaparra, M. P., & Middleton, R. S. (2004). Identifying Critical Infrastructure: The Median and Covering Facility Interdiction Problems. *Annals of the Association of American Geographers, 94*(3), 491–502.

Church, R. L., & Somogyi, C. (1985). Optimizing service and access coverage. Presented at the North American Meetings of the Regional Science Association, Philadelphia, PA.

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers in Regional Science, 32*(1), 101–118.

Colgan, M. (2014, February 4). Slow Response Times Put San Jose Fire Department In Hot Seat. *KCBS*. Retrieved from http://sanfrancisco.cbslocal.com/2014/02/04/slow-response-times-put-san-jose-fire-department-in-hot-seat/

Colner, D. R. (1973). User's Instructions for a Computer Program to Compute and Display Distributions Geographically and Tabularly. U.S. Department of Commerce, National Bureau of Standards, U.S. Government.

Commitee on Trauma and Commitee on Shock. (1966). *Accidental death and disability: the neglected disease of modern society*. Washington D.C.: National Academy of Science-National Research Council.

Committee on Acute Medicine of the American Society of Anesthesiologists. (1968). Community-Wide Emergency Medical Services. *JAMA*, *204*(7), 595–602.

Current, J. R., & Storbeck, J. E. (1988). Capacitated covering models. *Environment & Planning B: Planning & Design*, *15*(2), 153–163.

Curry, G. J. (1956). The ambulance attendant: his qualifications and training. *Bulletin of the American College of Surgeons*, *41*(6), 465.

Curry, G. J., & Lyttle, S. N. (1958). The speeding ambulance. *American Journal of Surgery*, *95*(4), 507–511.

Curry, G. J., & Lyttle, S. N. (1959). The ambulance. *American Journal of Surgery*, *98*(4), 530–533.

Dale, W. A. (1969). *Fleet Size Requirements for an Urban Emergency Ambulance Service* (D. R. C. Thesis). University of Strathclyde, Scotland.

Daskin, M. S. (1982). Application of an Expected Covering Model To Emergency Medical Service System Design. *Decision Sciences*, *13*(3), 416–439.

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

Daskin, M. S. (1987). Location, dispatching and routing models for emergency services with stochastic travel times. In A. Ghosh & G. Rushton (Eds.), *Spatial Analysis and Location-*

Allocation Models (pp. 224–265).

Daskin, M. S. (1995). Extensions of Location Models. In *Network and Discrete Location: Models, Algorithms, and Applications* (1st ed., pp. 309–382). John Wiley & Sons.

Daskin, M. S., Hogan, K., & Revelle, C. (1988). Integration of multiple, excess, backup, and expected covering models. *Environment & Planning B: Planning & Design*, *15*(1), 15–35.

Daskin, M. S., & Stern, E. H. (1981). A Hierarchical Objective Set Covering Model for Emergency Medical Service Vehicle Deployment. *Transportation Science*, *15*(2), 137–153.

Davidson, D. (1969). *Analysis of the Transit Behavior of the Emergency Ambulance Fleet Size Problem Using Markov Chains, Appendix 1* (Operational Research Project: Scottish Ambulance Service). Scotland: Department of Operational Research, University of Strathclyde.

Davis, J. P., Eusebgardt, K. M., & Binghaman, C. B. (2007). Develop Theory Through Simulation Methods. *Academy of Management Review*, *32*(2), 480–499.

De Maio, V. J., Stiell, I. G., Wells, G. a., & Spaite, D. W. (2003). Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine*, *42*(2), 242–250.

Dean, S. F. (2008). Why the closest ambulance cannot be dispatched in an urban emergency medical services system. *Prehospital and Disaster Medicine*, *23*(2), 161–165.

Dear, M. J. (1974). A paradigm for public facility location theory. *Antipode*, *6*(1), 46–50.

Dearing, P. M., & Jarvis, J. P. (1978). A Location Model with Queueing Constraints. *Computers & Operations Research*, *5*, 273–277.

Delasay, M., Ingolfsson, A., Kolfal, B., & Schultz, K. (2015). *Load Effect on Service Times* (SSRN Scholarly Paper No. ID 2647201). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2647201

Ehrgott, M. (2005). *Multicriteria Optimization* (2nd ed.). Springer, Berlin.

Eiselt, H. A., & Marianov, V. (2011). Pioneering Developments in Location Analysis. In H. A. Eiselt & V. Marianov (Eds.), *Foundations of Location Analysis* (pp. 3–22).

Eisenberg, M., Bergner, L., & Hallstrom, A. (1980). Out-of-hospital cardiac arrest: improved survival with paramedic services. *The Lancet*.

Erkut, E., Ingolfsson, A., & Budge, S. (2008). Maximum availability/reliability models for selecting ambulance station and vehicle locations: a critique. *Natural Sciences and Engineering Research Council of Canada*, 1–11.

Erkut, E., Ingolfsson, A., Sim, T., & Erdoğan, G. (2009). Computational Comparison of Five Maximal Covering Models for Locating Ambulances. *Geographical Analysis*, *41*(1), 43–65.

Fitch & Associates. (2013). *Operational Analysis of EMS & Fire Deployment/Response: Pinellas County, Florida*.

Fitzsimmons, J. A. (1970). *Emergency Medical Systems: A Simulation Study and Computerized Method for Deployment of Ambulances*. University of Califronia, Los Angeles.

Fitzsimmons, J. A. (1971). An emergency medical system simulation model. *Proceedings of the 5th Conference on Winter Simulation - WSC '71*.

Fitzsimmons, J. A. (1973). A Methodology for Emergency Ambulance Deployment. *Management Science*, *19*(6), 627–636.

Foster, F. A. (November 1969). *A Study of the Organization of the Ambulance Services in North Berkshire*. Nuffield Operational Research (Health Services) Unit, Department of Applied Statistics, University of Reading.

Gibson, G. (1971). Status of Urban Services. *Hospitals*, *45*(23), 49–54.

Gibson, G. (1973). Evaluative criteria for emergency ambulance systems. *Social Science & Medicine*, *7*(6), 425–454.

Gibson, G. (1973). Evaluative criteria for emergency ambulance systems. *Social Science & Medicine*, *7*(6), 425–454.

Gibson, G. (1974). Emergency medical services research: integration or isolation? *Health Services Research*, *9*(4), 255–269.

Goldberg, J. B. (2004). Operations Research Models for the Deployment of Emergency Services Vehicles. *EMS Management Journal*, *1*(1), 20–39.

Goldberg, J., Dietrich, R., Ming Chen, J., Mitwasi, M. G., Valenzuela, T., & Criss, E. (1990). Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ. *European Journal of Operational Research*, *49*(3), 308–324.

Goldberg, T. (2016, April 14). S.F. Firefighter Leaders Say Morale Is a Problem — and the Chief Should Go. *KQED News*. Retrieved from https://ww2.kqed.org/news/2016/04/14/s-

f-firefighter-leaders-say-morale-is-a-problem-and-the-chief-should-go/

Gordon, G., & Zelin, K. (1968). *A Simulation Study of Emergency Ambulance Service in New York.*

Gough, J., & McCarthy, W. O. (1975). *The Ambulance Facility Location Problem - A Survey of Methods and a Simple Solution* (Vol. 73). Retrieved from http://researcharchive.lincoln.ac.nz/handle/10182/1232

Greater London Council: Research and Intelligence Unit (1967). *The Control System for London Emerging Ambulance* (Papers 1-16).

Greene, J. D. (1996). How much privatization? A research note examining the use of privatization by cities in 1982 and 1992. *Policy Studies Journal*, *24*(4), 632–640.

Greenhut, M. L., & Mai, C. C. (1980). Towards a general theory of public and private facility location. *The Annals of Regional Science*, *14*(2), 1–11.

Groom, K. N. (1977). Planning Emergency Services. *Operational Research Quarterly*, *28*(3), 641–651.

Haimes, Y. Y., Lasdon, L. S., & Wismer, D. A. (1971). On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization. *IEEE Journals & Magazines*, *47*(JULy), 296–297.

Haj Mohammad Hosseini, M., & Jabal Ameli, M. S. (2011). A bi-objective model for emergency services location-allocation problem with maximum distance constraint. *Management Science Letters*, *1*(2), 115–126.

Hakimi, S. L. (1964). Optimum Locations of Switching Centers and the Absolute Centers and

Medians of a Graph. *Operations Research*, *12*(3), 450–459.

Hakimi, S. L. (1965). Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems. *Operations Research*, *13*(3), 462–475.

Hall, W. (1971). Management science approaches to the determination of urban ambulance requirements. *Socio-Economic Planning Sciences*, *5*(5), 491–499.

Hall, W. K. (1972). The Application of Multifunction Stochastic Service Systems in Allocating Ambulances to an Urban Area. *Operations Research*, *20*(3), 558–570.

Haller, J. S. (1990). The beginnings of urban ambulance service in the United States and England. *The Journal of Emergency Medicine*, *8*(6), 743–755.

Hamilton, W. F. (1974). Systems Analysis in Emergency Care Planning. *Medical Care*, *12*(2), 152–162.

Hampton, O. (1960). Transportation of the Injured - a report. *Bulletin of the American College of Surgeons*, *45*, 55–59.

Hampton, O. P. (1970). A Systematic Approach to Emergency Medical Services. *Archives of Environmental Health: An International Journal*, *21*(2), 214–217.

Hampton, O. P. (1972). The committee on trauma of the American College of Surgeons. *Bulletin of the American College of Surgeons*, *57*, 7–13.

Hanlon, J. J. (1973). Emergency medical care as a comprehensive system. *Health Services Reports*, *88*(7), 579–87.

Hart, H. (1978). The conveyance of patients to and from hospital, 1720—1850. *Medical*

History, *22*(4), 397–407.

Heller, M. (1985). *Location Optimization and Simulation for the Analysis of Emergency Medical Service Systems* (PhD thesis). Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimote, MD.

Heller, M., Cohon, J. L., & Revelle, C. S. (1989). The use of simulation in validating a multiobjective EMS location model. *Annals of Operations Research*, *18*, 303–322.

Heller, M., Cohon, J., & Revelle, C. (1989). The use of simulation in validating a multiobjective EMS location model. *Annals of Operations Research*, *18*, 303–322.

Hogan, K., & Revelle, C. (1986). Concepts and applications of backup coverage. *Management Science*, *32*(11), 1434–1444.

Holloway, R. M. (1972). New York City's experience in improving ambulance service. *Health Services Reports*, *87*(5), 445–450.

Holmes, J., Williams, F. B., & Brown, L. a. (1972). Facility Location under a Maximum Travel Restriction: An Example Using Day Care Facilities. *Geographical Analysis*, *4*(3), 258–266.

Hooke, R., & Jeeves, T. A. (1961). "Direct Search'' Solution of Numerical and Statistical Problems. *Journal of the ACM*, *8*(2), 212–229.

Hoot, N. R., & Aronsky, D. (2008). Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions. *Annals of Emergency Medicine*, *52*(2).

Howard, B. (1881). The New York Ambulance System. *British Medical Journal*.

Howard, C. R. G. (1965). Emergency care and medical transportation in the Eastern quarter of

the United States of America, 1963. *The Journal of the College of General Practitioners*, *9*(1), 48–54.

Huntley, H. C. (1970). Emergency health services for the nation. *Public Health Reports*, *85*(6), 517–522.

Ignall, E. J., Kolesar, P., & Walker, W. E. (1978). Using Simulation to Develop and Validate Analytic Models: Some Case Studies. *Operations Research*, *26*(2), 237–253.

Ingolfsson, A., Budge, S., & Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science*, *11*(3), 262–274.

Jarvis, J. P. (1976). A location model for spatially distributed queueing systems. In *Proceedings of the International Conference on Cybernetics and Society* (pp. 32–35).

Jarvis, J. P. (1985). Approximating the Equilibrium Behavior of Multi-Server Loss Systems. *Management Science*, *31*(2), 235–239.

Keeney, R. L. (1972). A Method for Districting Among Facilities. *Operations Research*, *20*(3), 613–618.

King, B. G. (1968). Estimating community requirements for the emergency care of highway accident victims. *American Journal of Public Health and the Nation's Health*, *58*(8), 1422–1430.

Krieger, W. M. (1958). Ambulance Operations. *Mortuary Management*, (October), 16.

Kuehn, A. A., & Hamburger, M. J. (1963). A Heuristic Program for Locating Warehouses. *Management Science*, *9*(4), 643–666.

Larrey, D. J. (1814). *Memoirs of Military Surgery and Campaigns of the French Armies (Vol.*

1).

Larson, R. (1975). Approximating the performance of urban emergency service systems. *Operations Research*, *23*(5), 845–868.

Larson, R. C. (1973). *A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services*. *R-1238-HUD* (Vol. 1). New York City.

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

Larson, R. C. (1975). Approximating the Performance of Urban Emergency Service Systems. *Operations Research*, *23*(5), 845–868.

Larson, R. C., & Odoni, A. R. (1981). *Urban Operations Research*. Retrieved from https://trid.trb.org/view.aspx?id=205030

Larson, R. C., & Stevenson, K. A. (1972). On Insensitivities in Urban Redistricting and Facility Location. *Operations Research*, *20*(3), 595–612.

Laverty, D. (2013, March 10). Private or fire-based ambulance services? It's a hot issue. *The Times of Northwest Indiana*. Retrieved from http://www.nwitimes.com/news/local/lake/hobart/private-or-fire-based-ambulance-services-it-s-a-hot/article_17fdbe33-b0fb-5c24-a79a-b9aa44311bb7.html

Lehman, S., & Hollingsworth, K. (1960). Ambulance service in Seattle. *Public Health Reports*, *75*(4), 343–351.

Lei, T. L., & Church, R. L. (2014). Vector assignment ordered median problem: a unified median problem. *International Regional Science Review*, *37*(2), 194-224.

Leonard, G. (1885). Ambulances and the Ambulance Service in the Larger Cities. In *A reference handbook of the medical sciences: embracing the entire range of scientific and practical medical and allied sciences*. New York.

Lewis, F. J. (1972). DOT and emergency medical care. *Journal of the American College of Emergency Physicians*, *1*(3), 29–33.

Linthicum, K., & Lopez, R. J. (2012, March 17). Injured and ailing people wait as dispatch problems slow LAFD. *Los Angeles Times*. Retrieved from http://articles.latimes.com/2012/mar/17/local/la-me-fire-dispatch-20120318

Linthicum, K., Welsh, B., & Lopez, R. J. (2012, November 15). Medical response time lags in many pricey L.A. neighborhoods. *Los Angeles Times*. Retrieved from http://articles.latimes.com/2012/nov/15/local/la-me-lafd-response-disparities-20121115

Lopez, R. J., Welsh, B., & Linthicum, K. (2012, October 20). Delayed 911 responses a matter of Geography and jurisdictions. *Los Angeles Times*.

Magelaner, I., & McElroy, M. (1955). Are ambulance sirens necessary? *Hospitals*, *29*(4), 89–90.

Manegold, R. F., & Silver, M. H. (1967). The Emergency Medical Care System. *JAMA*, *200*(4), 124–128.

Marianov, V., & Revelle, C. (1994). The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences*, *28*(3), 167–178.

Marianov, V., & ReVelle, C. (1992). The capacitated standard response fire protection siting problem: deterministic and probabilistic models. *Annals of Operations Research*, *40*,

303–322.

Marianov, V., & ReVelle, C. (1996). The queueing maximal availability location problem: A model for the siting of emergency vehicles. *European Journal of Operational Research*, *93*(1), 110–120.

Marianov, V., & Serra, D. (1998). Probabilistic, Maximal Covering Location—Allocation Models for Congested Systems. *Journal of Regional Science*, *38*(3), 401–424.

Marianov, V., & Serra, D. (2002). Location–Allocation of Multiple-Server Service Centers with Constrained Queues or Waiting Times. *Annals of Operations Research*, *111*, 35–50.

McAllister, D. M. (1976). Equity and efficiency in public facility location. *Geographical Analysis*, *8*(1), 47–63.

McIntire, M. (2003, October 21). Anatomy of a $133,000 Ambulance; City Pays Premium, but Its Tough Specs Draw Few Bidders. *The New York Times*. Retrieved from http://www.nytimes.com/2003/10/21/nyregion/anatomy-133000-ambulance-city-pays-premium-but-its-tough-specs-draw-few-bidders.html

Melachrinoudis, E. (1994). A Discrete Location Assignment Problem With Congestion. *IIE Transactions*, *26*(6), 83–89.

Meredith, J., & Shershin, A. (1978). Locating emergency medical rescue vehicles under conditions of urgency. *Computers & Industrial Engineering*, *2*(1), 31–39.

Merritt, A. K. (2014). The rise of emergency medicine in the sixties: Paving a new entrance to the house of medicine. *Journal of the History of Medicine and Allied Sciences*, *69*(2), 251–293.

Metzenbaum, M. (1908). Cleveland's present ambulance system. *Cleveland Medical Journal*, *7*, 7–12.

Miles, A. (1885). The Charity Hospital ambulance service. *New Orleans Medical and Surgical Journal*, *13*, 51–56.

Mitchell, H. (1965). Ambulances and emergency medical care. *American Journal of Public Health and the Nations Health*, *55*(11), 1717–1724.

Moghadas, F. M., & Kakhki, H. T. (2011). Queueing Maximal Covering Location-Allocation Problem: An Extension with M/G/1 Queueing Systems. *Advances in Decision Sciences*, *2011*, 1–13.

Moghadas, F. M., Monabbati, E., & Kakhki, H. T. (2013). Emergency Location Problems with an M/G/k Queueing System. *Iranian Journal of Operations Research*, *4*(1), 1–13.

Molnar, N. (2011, March 24). EMS response time called slow. *Akron Beacon Journal*. Akron, Ohio.

Morrill, R., & Symons, J. (1977). Efficiency and equity aspects of optimum location. *Geographical Analysis*, *IX*(July).

Murray, A. T., & Church, R. L. (1992). *Estimating α-Reliable coverage using Local-busyness: An assessment*.

Nahum, A. M. (1971). Emergency Medical Care Systems. *JAMA*, *217*(11), 1530–1532.

Naylor, T. H., & Finger, J. M. (1967). Verification of Computer Simulation Models. *Management Science*, *14*(2), B-92-B-101.

Neebe, A. W. (1978). A Branch and Bound Algorithm for the p-Median Transportation

Problem. *Journal of Operations Research Society*, 989–995.

O'Toole, M. (2011, July 16). We're in a war with the fire department. *National Post*. Toronto.

Pantridge, J. F., & Geddes, J. S. (1967). A mobile intensive-care unit in the management of myocardial infarction. *Lancet*, (August), 5–7.

Parkinson, C. N. (1955, November). Parkinson's Law. *The Economist*. Retrieved from http://www.economist.com/node/14116121

Perry, J. F., & McClellan, R. J. (1964). Autopsy findings in 127 patients following fatal traffic accidents. *Surgery, Gynecology & Obstetrics*, *119*, 586–590.

Pirkul, H., & Schilling, D. (1988). The siting of emergency service facilities with workload capacities and backup service. *Management Science*, *34*(7), 896–908.

Pirkul, H., & Schilling, D. A. (1991). The maximal covering location problem with capacities on total workload. *Management Science*, *37*(2), 233–248.

Powers, D. K. (2005). *Measuring Cost Effectiveness of South Ogden Fire Department Ambulance Service*. South Ogden City, Utah.

Pozner, C. N., Zane, R., Nelson, S. J., & Levine, M. (2004). International EMS Systems: The United States: past, present, and future. *Resuscitation*, *60*(3), 239–244.

President's Commission on Highway Safety. (1965). *Health, mdeical care, and transportation*. Washington, D.C.

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

ReVelle, C., & Hogan, K. (1988). A reliability-constrained siting model with local estimates of busy fractions. *Environment and Planning B: Planning and Design*, *15*(2), 143–152.

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

ReVelle, C., Marks, D., & Liebman, J. C. (1970). An Analysis of Private and Public Sector Location Models. *Management Science*, *16*(11), 692–707.

ReVelle, C. S., & Swain, R. W. (1970). Central facilities location. *Geographical Analysis*, *2*(1), 30–42.

Rice, D. (1966). *Estimating the cost of illness*. Washington D.C.

Robbins, V. D. (2005). A History of Emergency Medical Services Medical Transportation Systems in America.

Rockwood, C. C., Mann, C. C., Farrington, J., Hampton, O. O., & Motley, R. R. (1976). History of emergency medical services in the United States. *The Journal of Trauma*, *16*(4), 299–308.

Root, G. T., & Christensen, B. H. (1957). Early surgical treatment of abdominal injuries in the traffic victim. *Surgery, Gynecology & Obstetrics*, *105*(3), 264–267.

Rosenthal, E. (2013, December 4). Think the E.R. Is Expensive? Look at How Much It Costs to Get There. *The New York Times*. Retrieved from http://www.nytimes.com/2013/12/05/health/think-the-er-was-expensive-look-at-the-ambulance-bill.html

Ross, G. T., & Soland, R. M. (1975). A branch and bound algorithm for the generalized

assignment problem. *Mathematical Programming*, *8*(1), 91–103.

Ross, G. T., & Soland, R. M. (1977). Modeling Facility Location Problems as Generalized Assignment Problems. *Management Science*, *24*(3), 345–357.

Sanazaro, P. J. (1967). The Evaluation of Medical Care under Public Law 89-239. *Medical Care*, *5*(3), 162–168.

Sargent, R. G. (2005). Verification and Validation of Simulation Models. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, & J. A. Joines (Eds.), *Proceedings of the 2005 Winter Simulation Conference* (Vol. 1, pp. 130–143). Orlando, Florida: IEEE.

Savas, E. (1969). Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service. *Management Science*, *15*(12), 608–627.

Savas, E. S. (1978). On Equity in Providing Public Services. *Management Science*, *24*(8), 800–808.

Saydam, C., & Aytuğ, H. (2003). Accurate estimation of expected coverage: revisited. *Socio-Economic Planning Sciences*, *37*(1), 69–80.

Saydam, C., Repede, J., & Burwell, T. (1994). Accurate estimation of expected coverage: A comparative study. *Socio-Economic Planning Sciences*, *28*(2), 113–120.

Schlesinger, S., Crosbie, R. E., Gagné, R. E., Innis, G. S., Lalwani, C. S., Loch, J., … Bartos, D. (1979). Terminology for model credibility. *Simulation*, *32*(3), 103–104.

Serra, D. (1989). *The pq-median problem: Location and districting of hierarchical facilities*. John Hopkins University.

Shah, M. N. (2006). The formation of the emergency medical services system. *American*

Journal of Public Health, *96*(3), 414–423.

Shariat-Mohaymany, A., Babaei, M., Moadi, S., & Amiripour, S. M. (2012a). Linear upper-bound unavailability set covering models for locating ambulances: Application to Tehran rural roads. *European Journal of Operational Research*, *221*(1), 263–272.

Shariat-Mohaymany, A., Babaei, M., Moadi, S., & Amiripour, S. M. (2012b). Linear upper-bound unavailability set covering models for locating ambulances: Application to Tehran rural roads. *European Journal of Operational Research*, *221*(1), 263–272.

Shields, L. R. (1969). *Ambulance and Control Manning in the London Ambulance Service*. Greater London Council: Operational Research Unit.

Short, A. (2015, October 13). Slow FDNY EMS Response Times in Bronx. *The New York Post*. Retrieved from http://www.jems.com/articles/2015/10/slow-fdny-ems-response-times-in-bronx.html

Sigmond, R. M. (1967). Areawide Planning for Emergency Services. *JAMA*, *200*(4), 308–12.

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

Skudder, P. A., & Wade, P. A. (1964). The Organization of Emergency Medical Facilities and Services. *Journal of Trauma and Acute Care Surgery*.

Sluyter, A. J. J. (1976). The Role of Communication Systems in Emergency Medical Services. *IEEE Transactions on Vehicular Technology*, *25*(4), 175–186.

Smith, D. R., & Whitt, W. (1981). Resource Sharing for Efficiency in Traffic Systems. *The Bell System Technical Journal*, *60*(1), 39–55.

Smolensky, E., Burton, R., & Tideman, N. (1970). The Efficient Provision of a Local Non-Private Good. *Geographical Analysis*, *2*(4), 330–342.

Sorensen, P., & Church, R. L. (2010). Integrating expected coverage and local reliability for emergency medical services location problems. *Socio-Economic Planning Sciences*, *44*(1), 8–18.

Spaite, D. W., Criss, E. A., Valenzuela, T. D., & Guisto, J. (1995). Emergency Medical Service Systems Research: Problems of the Past, Challenges of the Future. *Annals of Emergency Medicine*, *26*(2), 146–152.

Stevenson, K. (1971). *Operational aspects of emergency ambulance services*. Cambridge. Retrieved from http://trid.trb.org/view.aspx?id=115138

Swain, R. (1971). *A decomposition algorithm for a class of facility location problems*. Cornell University.

Swoveland, C., Uyeno, D., Vertinsky, I., & Vickson, R. (1973a). A simulation-based methodology for optimization of ambulance service policies. *Socio-Economic Planning Sciences*, *7*(6), 697–703.

Swoveland, C., Uyeno, D., Vertinsky, I., & Vickson, R. (1973b). Ambulance Location: A Probabilistic Enumeration Approach. *Management Science*, *20*(4–Part–II), 686–698.

Taubenhaus, L. J. (1973). The emergency service spectrum. *Journal of the American College of Emergency Physicians*, *2*(5), 327–330.

Taubenhaus, L. J., & Kirkpatrick, J. R. (1967). Analysis of a Hospital Ambulance Service. *Public Health Reports (1896-1970)*, *82*(9), 823–827.

Teitz, M. B. (1968). Toward A Theory of Urban Public Facility Location. *Papers in Regional Science*, *21*(1), 35–51.

Toregas, C., & ReVelle, C. (1972). Optimal location under time or distance constraints. *Papers in Regional Science*, *28*(1), 133–144.

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

Trunkey, D. D. (2000). History and development of trauma care in the United States. *Clinical Orthopaedics and Related Research*, (374), 36–46.

van Buuren, M., van der Mei, R., & Bhulai, S. (2017). Demand-point constrained EMS vehicle allocation problems for regions with both urban and rural areas. *Operations Research for Health Care*.

Vinod, H. D. (1969). Integer programming and the theory of grouping. *Journal of the American Statistical Association*, *64*(326), 506–519.

Vogt, F. B. (1976). Problems in EMS Planning and Communications. *IEEE Transactions on Vehicular Technology*, *25*(4), 122–128.

Volz, R. A. (1971). Optimum Ambulance Location in Semi-Rural Areas. *Transportation Science*, *5*(2), 193–203.

Vukmir, R. B. (2006). Survival from prehospital cardiac arrest is critically dependent upon response time. *Resuscitation*, *69*(2), 229–234.

Waller, J. A. (1965). Ambulance Service: Transportation or Medical Care. *Public Health Reports (1896-1970)*, *80*(10), 847–853.

Waller, J. A., Curran, R., & York, N. (1964). Traffic Deaths: A Preliminary Study of Urban and Rural Fatalites in California. *California Medicine*, *101*(4), 272–276.

Waller, J. A., Garner, R., & Lawrence, R. (1966). Utilization of ambulance services in a rural community. *American Journal of Public Health and the Nations Health*, *56*(3), 513–520.

Wang, H. E., Kupas, D. F., Hostler, D., Cooney, R., Yealy, D. M., & Lave, J. R. (2005). Procedural experience with out-of-hospital endotracheal intubation. *Critical Care Medicine*, *33*(8), 1718–1721.

Watt, W. L. (1916). The Field Ambulance and its Organization. *The Canadian Medical*.

Weaver, J. R. (1979). *Context-Free Vector Assignment Location Problems*. The University of Tennessee.

Weaver, J. R., & Church, R. L. (1981). Average response time and workload balance: two criteria for ambulance station location. *Systems Science in Health Care*, 975–983.

Weaver, J. R., & Church, R. L. (1985). A Median Location Model with Nonclosest Facility Service. *Transportation Science*, *19*(1), 58–74.

Weaver, W. D., Cobb, L. a., Hallstrom, A. P., Fahrenbruch, C., Copass, M. K., & Ray, R. (1986). Factors influencing survival after out-of-hospital cardiac arrest. *Journal of the American College of Cardiology*, *7*(4), 752–757.

Welsh, B., Linthicum, K., & Lopez, R. J. (2013, April 26). LAFD chief to shift firefighters from trucks to ambulances. *Los Angeles Times*. Los Angeles, California.

West, I., Kleinman, G., Taylor, E., Majors, A., & Mitchell, H. W. (1964). Speeding ambulance survey: a preliminary report. *Aid (Journal of the Ambulance Association of America)*, 8.

West, I. M. I., Gettinger Jr, C. E., Meyer, D., Rosenthal, M., Snow, R., Weiner, F. F. R., … Hoaglin, L. M. W. L. (1972). Emergency Medical Transportation—A Survey of California Ambulance Operations. *California Medicine, 116*(2), 35–43.

White, J. A., & Case, K. E. (1974). On Covering Problems and the Central Facilities Location Problem. *Geographical Analysis*, *6*(3), 281–294.

Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes? *Health Economics*, *22*(7), 790–806.

Willard, D. (1883). *Ambulance Service in Philadephia*.

Williams, D. (2007). 2006 JEMS 200-city survey. EMS from all angles. *JEMS*, *32*(2), 38–42.

Williams, D. M. (2007). 2006 JEMS 200-city survey. EMS from all Angles. *JEMS*, *32*(2), 38–54.

Yao, D. D., & Shanthikumar, J. G. (1987). The Optimal Input Rates to a System of Manufacturing Cells. *Infor*, *25*(1), 57–65.

Young, C. B. (1954). *First aid and resuscitation; emergency procedures for rescue squads, firemen, policemen, ambulance crews, interns and industrial nurses.* Springfield, Ill.: Thomas.

Young, C. B. (1958). *Transportation of the Injured*. Springfield, Ill.: Charles C. Thomas.

Zaffar, M. A., Rajagopalan, H. K., Saydam, C., Mayorga, M., & Sharer, E. (2016). Coverage, survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation–optimization. *Operations Research for Health Care*, *11*, 1–12.

Zollinger, R. W. (1955). Traffic Injuries: A surgical problem. *AMA Archives of Surgery*, *70*(5), 694–700.

APPENDIX A

		City Diameter:16 [Min]					
Demand [Call/Hr.]:		2			4		
Service Std. [Min]:		6	8	10	6	8	10
MALP 2	Max	**65.00%**	71.67%	80.00%	**60.00%**	43.33%	**76.67%**
	0.01	**95.00%**	**98.33%**	**100.00%**	**90.00%**	88.33%	**96.67%**
	0.02	**98.33%**	**100.00%**	**100.00%**	90.00%	91.67%	**100.00%**
	0.05	98.33%	**100.00%**	**100.00%**	96.67%	96.67%	**100.00%**
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	**65.00%**	71.67%	80.00%	**60.00%**	43.33%	**76.67%**
	0.01	**95.00%**	**98.33%**	**100.00%**	**90.00%**	88.33%	**96.67%**
	0.02	**98.33%**	**100.00%**	**100.00%**	90.00%	91.67%	**100.00%**
	0.05	98.33%	**100.00%**	**100.00%**	90.00%	**100.00%**	91.67%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	26.67%	**75.00%**	**86.67%**	25.00%	**73.33%**	56.67%
	0.01	83.33%	**98.33%**	96.67%	58.33%	**96.67%**	80.00%
	0.02	88.33%	**100.00%**	**100.00%**	80.00%	**98.33%**	81.67%
	0.05	**100.00%**	**100.00%**	**100.00%**	93.33%	98.33%	95.00%
	Unique	8.33%	**13.33%**	**10.00%**	11.67%	**31.67%**	**21.67%**
RCQPM	Max	41.67%	51.67%	50.00%	46.67%	50.00%	35.00%
	0.01	**95.00%**	96.67%	96.67%	83.33%	91.67%	80.00%
	0.02	96.67%	**100.00%**	**100.00%**	**91.67%**	95.00%	81.67%
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	**21.67%**	11.67%	3.33%	**25.00%**	13.33%	0.00%

Table 37 - Highest simulated total coverage: City diameter 16 [Min], aggregated across *p* and α.

Demand [Call/Hr.]:			2			4	
Service Std. [Min]:		6	8	10	6	8	10
MALP 2	Max	40.00%	36.67%	**73.33%**	**43.33%**	28.33%	**56.67%**
	0.01	66.67%	83.33%	90.00%	61.67%	56.67%	80.00%
	0.02	80.00%	93.33%	96.67%	75.00%	78.33%	88.33%
	0.05	91.67%	98.33%	**100.00%**	88.33%	91.67%	91.67%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	40.00%	36.67%	**73.33%**	**43.33%**	28.33%	**56.67%**
	0.01	66.67%	83.33%	90.00%	61.67%	56.67%	80.00%
	0.02	80.00%	93.33%	96.67%	75.00%	78.33%	88.33%
	0.05	**100.00%**	96.67%	98.33%	83.33%	90.00%	91.67%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	**46.67%**	36.67%	45.00%	41.67%	35.00%	28.33%
	0.01	**86.67%**	75.00%	83.33%	**75.00%**	60.00%	70.00%
	0.02	**98.33%**	95.00%	91.67%	**78.33%**	75.00%	85.00%
	0.05	96.67%	**100.00%**	**100.00%**	85.00%	88.33%	96.67%
	Unique	**20.00%**	16.67%	5.00%	**21.67%**	26.67%	10.00%
RCQPM	Max	43.33%	**58.33%**	50.00%	35.00%	**48.33%**	45.00%
	0.01	73.33%	**90.00%**	**91.67%**	63.33%	**75.00%**	**91.67%**
	0.02	85.00%	**98.33%**	**98.33%**	73.33%	**81.67%**	**96.67%**
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	15.00%	**38.33%**	**11.67%**	16.67%	**41.67%**	**25.00%**

Table 38 - Highest simulated total coverage: City diameter 24 [Min], aggregated across *p* and α.

		City Diameter: 32 [Min]					
Demand [Call/Hr.]:		2			4		
Service Std. [Min]:		6	8	10	6	8	10
MALP 2	Max	38.33%	40.00%	26.67%	38.33%	**41.67%**	20.00%
	0.01	55.00%	66.67%	56.67%	50.00%	61.67%	60.00%
	0.02	68.33%	81.67%	80.00%	56.67%	**78.33%**	73.33%
	0.05	88.33%	93.33%	95.00%	86.67%	88.33%	90.00%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
QMALP	Max	38.33%	40.00%	26.67%	38.33%	**41.67%**	20.00%
	0.01	55.00%	66.67%	56.67%	50.00%	61.67%	60.00%
	0.02	68.33%	81.67%	80.00%	56.67%	**78.33%**	73.33%
	0.05	98.33%	**100.00%**	98.33%	93.33%	83.33%	95.00%
	Unique	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RCQ	Max	**45.00%**	**45.00%**	43.33%	38.33%	40.00%	53.33%
	0.01	**65.00%**	**86.67%**	75.00%	71.67%	**73.33%**	75.00%
	0.02	78.33%	**98.33%**	90.00%	**86.67%**	**78.33%**	**86.67%**
	0.05	98.33%	**100.00%**	98.33%	93.33%	83.33%	90.00%
	Unique	**25.00%**	**16.67%**	20.00%	11.67%	**23.33%**	28.33%
RCQPM	Max	36.67%	**45.00%**	**58.33%**	**55.00%**	35.00%	**55.00%**
	0.01	**65.00%**	73.33%	**93.33%**	**73.33%**	63.33%	**80.00%**
	0.02	**81.67%**	90.00%	**95.00%**	**86.67%**	73.33%	83.33%
	0.05	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Unique	16.67%	**16.67%**	**36.67%**	**26.67%**	18.33%	**31.67%**

Table 39 - Highest simulated total coverage: City diameter 32 [Min], aggregated across *p* and α.