Title

Simultaneous Location of Trauma Centers and Helicopters for Emergency Medical Service Planning

Journal

Operations Research, 62(4)

ISSN

0030-364X

Authors

Cho, Soo-Haeng
Jang, Hoon
Lee, Taesik
et al.

Publication Date

2014-08-01

DOI

10.1287/opre.2014.1287

Peer reviewed

Simultaneous Location of Trauma Centers and Helicopters for Emergency Medical Service Planning

Soo-Haeng Cho · Hoon Jang · Taesik Lee · John Turner*

Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, soohaeng@andrew.cmu.edu
Department of Industrial & Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea, hoon.jang@kaist.ac.kr, taesik.lee@kaist.ac.kr
The Paul Merage School of Business, University of California, Irvine, CA, john.turner@uci.edu

Abstract: This paper studies the problem of simultaneously locating trauma centers and helicopters. The standard approach to locating helicopters involves the use of helicopter busy fractions to model the random availability of helicopters. However, busy fractions cannot be estimated a priori in our problem because the demand for each helicopter cannot be determined until the trauma center locations are selected. To overcome this challenge, we endogenize the computation of busy fractions within an optimization problem. The resulting formulation has non-convex bilinear terms in the objective, for which we develop an integrated method that iteratively solves a sequence of problem relaxations and restrictions. Specifically, we devise a specialized algorithm, called the Shifting Quadratic Envelopes algorithm, that 1) generates tighter outer-approximations than linear McCormick envelopes, and 2) outperforms a Benders-like cut generation scheme. We apply our integrated method to the design of a nationwide trauma care system in Korea. By running a trace-based simulation on a full year of patient data, we find that the solutions generated by our model outperform several benchmark heuristics by up to 20%, as measured by an industry-standard metric: the proportion of patients successfully transported to a care facility within one hour. Our results have helped the Korean government to plan its nationwide trauma care system. More generally, our method can be applied to a class of optimization problems that aim to find the locations of both fixed and mobile servers when service needs to be carried out within a certain time threshold.

Subject Classification: Health care: ambulance service, hospitals. Programming: integer: algorithms: Benders/decomposition. Simulation: applications
Area of Review: Policy Modeling and Public Sector OR

*We thank our collaborators from an earlier study (Kim et al. 2011) commissioned by the Korean Ministry of Health and Welfare, which helped us deepen our understanding of EMS practice. Special thanks go to Dr. Yoon Kim, principal investigator of that study, for allowing us to use the trauma patient data for this paper and for giving us the opportunity to contribute through our work to the establishment of the new trauma care system in Korea. We thank the area editor, Pinar Keskinocak, as well as the anonymous associate editor and three referees, whose comments decidedly improved our manuscript. Many thanks also to those that provided us with useful feedback at INFORMS, MSOM, and POMS conferences, as well as at the Southern California OR/OM Day at UCLA. An earlier version of this paper was awarded the 2012 Lave-Weill Prize for the best unpublished paper on problem solving from Carnegie Mellon University. This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0027093).

1 Introduction

Trauma is a body wound or shock produced by sudden physical injury, as from violence or accident, which may lead to the death of a patient if proper care is not administered in a timely fashion. Trauma is the sixth leading cause of death worldwide and the leading cause of death in the U.S. for those under age 44 (Centers for Disease Control and Prevention (CDC) 2013). Trauma is a serious public health problem with significant social and economic costs.

Providing proper care to trauma patients requires seamless health care delivery operations. When a trauma case occurs, critical care paramedics are quickly dispatched to the scene either by ground ambulance or helicopter. They provide first aid to stabilize the patient at the scene, and then transport him/her to a trauma center. Because any delay in transporting a patient to a trauma center can severely affect his/her survival rate, a general rule of thumb is that an appropriate clinical intervention should be provided *within an hour* from the moment of an injury incident (e.g., CDC 2012). Many countries, including Canada, Germany and Israel, have reported significant improvements in major injury care from designating dedicated trauma care centers (Peleg et al. 2004). A trauma center is a type of hospital that has resources and equipment needed to help care for severely-injured patients. In the U.S., trauma centers are classified as Level I (the highest level of care) to Level IV. The CDC estimates a 25% reduction in deaths for severely-injured patients who receive care at a Level I trauma center rather than at a non-trauma center.

Our paper studies the design of *a nationwide trauma care system*. Specifically, our paper is part of a broader study commissioned by the Korean Ministry of Health and Welfare (KMHW) to make recommendations for establishing a nationwide trauma care system in Korea. In the broader study (Kim et al. 2011), a group of experts conducted research on a variety of issues related to trauma care, including infrastructure, human resources, governance, financing, and quality control. Our paper expands the infrastructure part of this broader study, and provides a quantitative model and analysis using *one-year nationwide data* of 190,193 trauma cases. Our results have helped the Korean government outline its initial plan in 2011. The Korean government is currently refining its initial plan by incorporating feedback from relevant communities. The first trauma center in Korea is scheduled to open in 2014, and the government is planning to open an additional 15 trauma centers across the country. The results developed in this paper can also support future decision-making of the government as it rolls out its final plan. Although we generate numerical results based on the data from Korea, the models and methods developed here could be applied to the design of trauma care systems in other countries or regions. More generally, they could also be applied to a class of optimization problems that aim to find the locations of both fixed servers (e.g., trauma centers, hospitals, warehouses) and mobile servers (e.g., helicopters, ambulances, trucks) when service needs to be carried out within a certain time threshold.

Our objective is to find the optimal locations of trauma centers and helicopters that maximizes the effective coverage of trauma care. Ideally, we would like to always transport every citizen in the country to an available trauma center in under an hour. However, due to limited budgets, this is not practically possible – even in the U.S., nearly 45 million people do not have access to a Level I or II

trauma center within one hour's transportation distance (CDC 2012). At the request of the KMHW, we explore trauma care system designs with various numbers of trauma centers and helicopters, assuming that a sufficiently high number of ground ambulances are deployed.

When there are a limited number of trauma centers, helicopters play the important role of expanding geographic coverage by transporting patients from rural areas to trauma centers. Thus, in designing a nationwide trauma care system, it is important to *simultaneously locate trauma centers and helicopters*. The problems of locating only one type of resource (i.e., either trauma centers or helicopters) belong to well-known classes of optimization problems. The trauma center location problem can be formulated as a capacitated maximal covering location problem (e.g., Pirkul and Schilling 1991), and the helicopter location problem is structurally similar to a probabilistic ambulance location problem (e.g., Daskin 1983). However, our problem poses a unique challenge because the locations of trauma centers affect the demand for helicopters, and vice versa. This dependency is particularly problematic if we explicitly model the probabilistic availability of helicopters. Specifically, in the probabilistic location problem (e.g., ReVelle and Hogan 1989, Borras and Pastor 2002), it is common to model the probabilistic nature of ambulance availability using a "busy fraction" of an ambulance (or a helicopter, in our case). This fraction is usually estimated as a ratio of the workload of an ambulance at a given location (e.g., the daily number of service requests times average service time per patient) to the available service hours of an ambulance. In our model, however, such busy fractions cannot be estimated *a priori* because the demand for helicopters at each given location cannot be determined until *after* the trauma centers are chosen.

To address this challenge, we *endogenize* the computation of busy fractions within an optimization problem, and formulate the problem as a mixed-integer nonlinear program (MINLP) with the objective of maximizing the expected (approximate) number of trauma patients that can be successfully transported within an hour. However, due to the inherent dependency described above, our MINLP formulation has non-convex bilinear terms which present serious computational challenges. Fortunately, we are able to exploit problem-specific structure to develop an integrated method that iteratively solves a sequence of problem relaxations and restrictions, thereby establishing bounds on our model's objective. Specifically, we devise a specialized method, called the *Shifting Quadratic Envelopes (SQE) algorithm,* that creates and shifts quadratic envelopes at each iteration. We show that SQE generates tighter outer-approximations than classical linear McCormick envelopes (McCormick 1976, Floudas and Pardalos 2012), and also outperforms a cut generation scheme based on Generalized Benders Decomposition (GBD) (e.g., Geoffrion 1972). The use of SQE allows us to get within 6% of optimality in most problem instances on the scale required in our problem setting. This is significant, especially considering that the leading global solver BARON (c.f. Tawarmalani and Sahinidis 2005) achieves only 21% of optimality in the same allotted time.

As a point of comparison, we also develop two simple heuristics that are motivated by existing methods in the literature. The first "no-congestion" heuristic is modeled after Branas and ReVelle (2001) who also consider a joint location problem of trauma centers and helicopters as in our paper,

but ignore random availability of helicopters. The second "decoupled" heuristic solves for the locations of trauma centers and helicopters sequentially, ignoring the dependency between these two resources. We have found that the best trauma center locations differ significantly across the three approaches, especially when a large number of helicopters are available. We then compare the performance of the location solutions generated by our integrated approach with those from the two heuristics by carrying out a *trace-based simulation* using one year of patient data. Our simulation treats the locations of trauma centers and helicopters as given, and sequentially processes the times and locations of patient calls from the input stream. We simulate the real-time processes of transporting each patient to a trauma center, and categorize each patient as successful ($\leq$ 60 min transport to an under-capacity trauma center) or unsuccessful. Our results show that our integrated approach outperforms the two heuristics significantly. For example, when using 10 trauma centers and 20 helicopters, we achieve a 12% (resp., 14%) larger number of successful transports using our integrated approach than the no-congestion heuristic (resp., the decoupled heuristic), which translates into a potential 23,000 (resp., 26,000) additional lives saved per year.

The rest of this paper is organized as follows. In §2, we review related literature. In §3, we describe the problem. In §4, we present our solution approaches. In §5, we describe our data and simulation model, and present the results of applying our methods to a trauma center design problem in Korea. We conclude our paper in §6.

2 Related Literature

Our paper is related to the literature on location problems in health care delivery systems. Due to the large volume of literature in this area, we review only the most related papers, and refer readers to Owen and Daskin (1998), Berman and Krass (2002), Brotcorne et al. (2003), Daskin and Dean (2004), ReVelle and Eiselt (2005), and Li et al. (2011) for a more comprehensive review.

Toregas et al. (1971) and Church and ReVelle (1974) are among the first researchers to study Emergency Medical Services (EMS) vehicle location problems. Toregas et al. (1971) study the Location Set Covering Problem (LSCP) that identifies the minimum number of facilities (or ambulances) and their locations which cover all demand points within a certain distance. Church and ReVelle (1974) propose the Maximal Covering Location Problem (MCLP), which locates a fixed number of facilities so as to maximize the amount of demand that is covered by at least one facility. Many variations and extensions have followed these early studies: for example, Schilling et al. (1979) develop a model of locating multiple types of vehicles such as basic and advanced life support ambulances; Hogan and ReVelle (1986) and Gendreau et al. (1997) consider double coverage models in which all demands must be covered by ambulances located at most r_2 minutes away, and in addition, a certain proportion of the demand must also be at most r_1 minutes away.

While these models deal with deterministic location problems, another stream of research takes into account randomness in the availability of ambulances. This randomness is usually modeled as the busy fraction of an ambulance, i.e., the probability that an ambulance is unavailable to respond to a service

request immediately. Assuming a uniform busy fraction for all ambulances, Chapman and White (1974) formulate a probabilistic version of the LSCP, and Daskin (1983) proposes the Maximum Expected Covering Location Problem (MEXCLP) that maximizes the expected value of coverage within a time standard. ReVelle and Hogan (1989) formulate a chance-constrained program, called the Maximum Availability Location Problem (MALP), which positions ambulances so as to maximize the demand covered within a time standard with a given probability. They also propose a method to estimate busy fractions that are specific to each demand region. Several probabilistic location models have extended these models, including Ball and Lin (1993), Marianov and ReVelle (1994, 1996), Borras and Pastor (2002), and Sorensen and Church (2010). All these models require the estimation of busy fractions. However, as ReVelle and Hogan (1989) rightly point out, busy fractions are difficult to estimate because "these values are an *output* of the model and cannot be known a priori," i.e. before knowing the locations of ambulances (Brotcorne et al. 2003).[1] This issue is even more serious in our problem of jointly locating trauma centers and helicopters because, unlike all previous models reviewed above, the busy fraction of a helicopter depends on the locations of *both* trauma centers and helicopters, which are not given a priori.

To the best of our knowledge, Branas and ReVelle (2001) is the only paper that considers a joint location problem of trauma centers and helicopters as in our paper. Branas and ReVelle model this as a *deterministic* location problem, and formulate the problem as a mixed-integer *linear* program. They could not attain solutions within a reasonable amount of time using CPLEX directly, so they developed an iterative heuristic that identifies the best locations of helicopters, holding the locations of trauma centers fixed, and then finds the best locations of trauma centers, holding the locations of helicopters fixed, and so on.

Similar to Branas and ReVelle (2001), this paper considers a joint location problem of trauma centers and helicopters, but takes a fundamentally different approach in both model and solution method. First, our model explicitly models the capacity constraints of trauma centers, which are critical to ensure that trauma centers do not become over-crowded. Second, our model takes into account the randomness in the availability of helicopters as in the second stream of research reviewed above. To address the issues regarding the estimation of busy fractions discussed above, we *endogenize* the computation of helicopter site-specific busy fractions within an optimization problem. According to ReVelle and Hogan (1989),

"It should be noted that the busy fractions used here are not specific for a particular site. The use of [ambulance] site-specific busy fractions, rather than [demand] area-specific busy fractions, would certainly be preferable but such a formulation is not undertaken here for two reasons. First, such site-specific busy fractions cannot be obtained without knowledge of the positions of all other servers, and these positions are only known as an output of the model, not in advance. Second, the constraints that follow from such information are of a form which

[1] Descriptive queueing models such as Larson (1975) and Burwell et al. (1992) can estimate busy fractions under fairly realistic assumptions. However, as Marianov and ReVelle (1994) point out, such descriptive queueing models usually fix the locations of ambulances a priori. There are recent developments in queueing-based location models (see, e.g., Berman and Krass (2002), Aboolian et al. (2008), Zhang et al. (2010) and references therein), but as Berman and Krass (2002) point out, "one invariably has to make simplifying assumptions and approximations to render the model tractable."

requires an integer programming code capable of solving large zero-one problems without special structure."

As ReVelle and Hogan (1989) predicted, the use of site-specific busy fractions within an optimization model requires us to solve a large-scale MINLP with a specialized algorithm. Indeed, our optimization model is a complex mixed-integer *nonlinear* program that determines the allocation of patient demand to trauma centers (using either ground ambulance or helicopter), as well as the locations of trauma centers and helicopters. To solve this program, we develop a novel method that iteratively solves a sequence of problem relaxations and restrictions. Our method exploits the specific structure of the problem formulation to tighten bounds by systematically pushing the solution toward a global optimum. In this way, our method theoretically guarantees convergence to an optimal solution, whereas the heuristic developed by Branas and ReVelle (2001) does not.

In addition, we validate our optimization model's location solution by conducting a trace-based simulation using one year of nationwide patient data from Korea. Our use of simulation for validation is in line with Goldberg et al. (1990), Repede and Bernardo (1994), Borras and Pastor (2002), and Sorensen and Church (2010).

Lastly, we note that our model and solution method may be used in other applications that require the simultaneous location of fixed and mobile servers when service needs to be carried out within a certain time threshold. For example, in the transportation literature, the location-routing problem (see Nagy and Salhi 2007, as well as Prodhon and Prins 2014, for extensive reviews) involves (a) choosing a number of depot locations from which to ship products, (b) assigning one or more trucks to each depot, and (c) finding a short route (tour) for each truck that starts at a depot, makes deliveries to one or more customers in sequence, and returns to the same depot. If we consider the special case where each truck makes a delivery to only one customer, then we can compare this to our problem, and describe how extensions of this problem more closely relate to ours. Notably, location-routing problems usually do not consider transportation delays, since each truck only makes one tour, and unanticipated demands do not occur in the planning horizon. However, if customers order the product at random times and need the product urgently (e.g., pizza delivery from a chain of pizza outlets) then, as in our model, it is appropriate to minimize transportation-delay-induced congestion, and such an objective will naturally have bilinear (or more generally, if multiple trucks are stationed at each depot, nonlinear) terms that can be tackled by our Shifting Quadratic Envelopes algorithm.[2] Although these are the main structural properties that we need to use SQE, we point out that our model and SQE are particularly important for problems where the busy fractions for the mobile servers are hard to estimate a priori, such as when fixed servers also need to be located and the mobile servers must stop at one fixed server along their route. This is because, in this case, route lengths depend on where the fixed servers are located, and so utilization and thus busy fractions depend directly on the route lengths.[3]

[2]Specifically, we need both the arrival rate and workload assigned to each mobile server to be linear functions of the same set of variables. However, as we will see from our model in §4, this occurs quite naturally.

[3]In a location-routing problem, depots are usually facilities that house trucks, and they are analogous to the heliports in our model where the helicopters are stationed; there is no analog to our trauma centers in the canonical location-routing

3 Problem Description

Consider the problem of locating k trauma centers and m helicopters (i.e., air ambulances) to serve $\bar{i}$ demand regions. We call the station where a helicopter is based a "heliport." There are $\bar{j}$ ($\geq k$) candidate sites for trauma centers and $\bar{h}$ candidate heliports. A heliport can either be on the roof of an open trauma center or at a separate location (e.g., an airport) that permits helicopter take-off and landing. We index demand regions by $i \in I = \{1, 2, ..., \bar{i}\}$, heliports by $h \in H = \{1, 2, ..., \bar{h}\}$, and trauma centers by $j \in J = \{1, 2, ..., \bar{j}\}$. More specifically, we assume without loss of generality that heliports 1 through $\bar{j}$ are located on the rooftops of trauma centers 1 through $\bar{j}$, while heliports $\bar{j}+1$ through $\bar{h}$ are not co-located at trauma center sites. Each demand region i has expected demand rate λ_i, and each trauma center j has the fixed capacity of treating up to c_j patients per unit of time. A trauma patient from any demand region can be transported to a trauma center either by a ground ambulance (hereinafter, in short, ambulance) or by a helicopter.

A patient is *geographically* covered if s/he can be transported to an open trauma center within 60 minutes. To define sets of these patients and their transportation modes, we convert travel times (e.g., 60 minutes) into distances between locations as follows (see Figure 1 for illustration). Let d_i^r denote the road distance between the center of demand region i and its nearest ambulance station, d_{ij}^r denote the road distance between the center of demand region i and trauma center j, d_{ij} denote the Euclidean distance between the center of demand region i and trauma center j, and d_{hi} denote the Euclidean distance between heliport h and the center of demand region i. Demand region i is geographically covered by an ambulance if there exists a trauma center j with $d_i^r + d_{ij}^r \leq d_{ground}$, or by a helicopter if there exists a heliport and trauma center pair (h, j) such that $d_{hi} + d_{ij} \leq d_{air}$.[4] In collaboration with practitioners, we have set $d_{ground} = 46\ km$ and $d_{air} = 120\ km$, considering the average time spent in each step of operations (a1-a3) and (h1-h5), respectively, as follows:

• Ambulance: (a1) drive to patient location i from the nearest station ($d_i^r/(50/60)$ minutes), where the average ambulance speed of 50 (km/h) is used; (a2) load the patient into the ambulance (5 minutes); and (a3) drive from patient location i to trauma center j ($d_{ij}^r/(50/60)$ minutes).

• Helicopter: (h1) take off at heliport h (6 minutes); (h2) fly from heliport h to patient location i ($d_{hi}/(180/60)$ minutes), where the average helicopter speed of 180 (km/h) is used; (h3) load the patient into the helicopter (8 minutes); (h4) fly from patient location i to trauma center j ($d_{ij}/(180/60)$ minutes); and (h5) land and hand-off the patient to the trauma center (6 minutes). We assume that a helicopter at heliport h is used to cover demand region i only when $d_i^r + d_{ij}^r > d_{ground}$. Moreover, we assume that each patient is transported (as necessary) to a place where a helicopter can land (e.g., an elementary school, a farming area, etc.) while a helicopter is en route to the patient; thus, maneuvering to the exact pick-up location does not impose any additional delay.

problem. In this sense, although it is tempting to consider depots as fixed servers, they are not. Thus, strictly speaking, the canonical location-routing problem locates only mobile servers, and not fixed servers.

[4] Alternatively, we can define d_i^r as the average road distance between each historical patient in demand region i and his/her nearest ambulance station, and define d_{ij}^r similarly. Geographic coverage is qualitatively unchanged under this

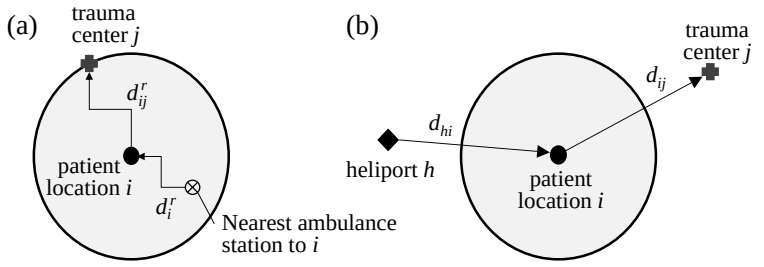


Figure 1: (a) ambulance coverage: $d_i^r + d_{ij}^r \leq d_{ground}$; and (b) helicopter coverage: $d_{hi} + d_{ij} \leq d_{air}$, excluding the area covered by ambulance (note: h and j can be co-located).

We now define the following sets of geographically covered patients and their transportation modes:

• $F^G = \{(i,j) | i \in I,\ j \in J, \text{ and } d_i^r + d_{ij}^r \leq d_{ground}\}$: all feasible ambulance routes (i,j) in which a patient from region i can be transported to trauma center j within 60 minutes;

• $F_i^G = \{j \in J | d_i^r + d_{ij}^r \leq d_{ground} \text{ for fixed } i\}$: the subset of trauma center sites to which a patient from demand region i can be transported by ambulance within 60 minutes;

• $F_j^G = \{i \in I | d_i^r + d_{ij}^r \leq d_{ground} \text{ for fixed } j\}$: the subset of demand regions whose patients can be transported by ambulance to trauma center j within 60 minutes;

• $F = \{(h,i,j) | h \in H, i \in I,\ j \in J, d_i^r + d_{ij}^r > d_{ground} \text{ and } d_{hi} + d_{ij} \leq d_{air}\}$: all feasible helicopter routes (h,i,j) in which a helicopter from heliport h can transport a patient from region i to trauma center j within 60 minutes, excluding any routes that are close enough to be within the ambulance coverage area;

• $F_h = \{(i,j) | d_i^r + d_{ij}^r > d_{ground} \text{ and } d_{hi} + d_{ij} \leq d_{air} \text{ for fixed } h, \text{ where } i \in I,\ j \in J\}$: the subset of pairs of demand region i and trauma center site j to which a helicopter originating from heliport h can transport a patient from demand region i to trauma center j within 60 minutes;

• $F_i = \{(j,h) | d_i^r + d_{ij}^r > d_{ground} \text{ and } d_{hi} + d_{ij} \leq d_{air} \text{ for fixed } i, \text{ where } j \in J,\ h \in H\}$: the subset of pairs of trauma center j and heliport h that can be used to transport a patient from demand region i by air (using the route $h \to i \to j$) within 60 minutes;

• $F_j = \{(h,i) | d_i^r + d_{ij}^r > d_{ground} \text{ and } d_{hi} + d_{ij} \leq d_{air} \text{ for fixed } j, \text{ where } i \in I,\ h \in H\}$: the subset of pairs of heliport h and demand region i that can be used to transport a patient by air to trauma center j (using the route $h \to i \to j$) within 60 minutes.

Geographic coverage is a feasibility criterion, and does not take into account the possibility that patients may be delayed due to congestion. For example, a patient must wait for a helicopter if the nearest helicopter is already transporting another patient. Some proportion of patients who are geographically covered will not receive timely service in expectation. To take into account this congestion effect, we define *expected* covered demand as the expected number of patients that will be transported to an open trauma center in under 60 minutes, without incurring any delays in transportation.

Our objective is to find the locations of k trauma centers and m helicopters that maximize the expected covered demand within a time standard of 60 minutes, considering randomness in the avail-

alternative definition.

ability of helicopters. Starting with Daskin (1983), this objective is commonly used in the literature for ambulance location problems (see, e.g., §11.2.1 of Berman and Krass 2002, Sorensen and Church 2010, and references therein). Like ambulances, helicopters in our problem are complimentary to trauma centers since each patient needs both transportation and a care facility. However, unlike ambulances, it is interesting to note that helicopters also act as substitutes for trauma centers by allowing fewer trauma centers to serve a larger coverage area. To model the availability of helicopters, we compute the average service time τ_{hij} for a helicopter to fly the circuit $h \to i \to j \to h$, including steps (h1)-(h5) described above and the following additional steps: (h6) travel from trauma center j to heliport h (where travel time $= d_{jh}/(180/60)$ minutes if $j \neq h$ or 0 otherwise), and (h7) land and refuel at heliport h (5 minutes). During this service time, a helicopter is busy, and thus it is not available to serve any other patients.

As discussed earlier in §1, this problem is challenging due to many reasons – in particular, interdependency in demands for trauma centers and helicopters, and inherent uncertainties. To build a tractable model to support the decision-making of the KMHW, we make a set of assumptions as follows. First, we do not consider randomness in the availability of ambulances.[5] This assumption allows us to ignore the availability and routing details of thousands of ambulances across the country. In fact, the average time it takes for an ambulance to reach a patient location from the moment of a patient call is 10 minutes in major cities in Korea, suggesting that the availability of ambulances is not a serious concern in Korea. We note, however, that we could also consider random availability of ambulances by modeling it in the exact same way as we model random availability of helicopters. Second, in our base model, we assume that each heliport can have at most one helicopter. We give up little with this assumption because, in Korea, 38 candidate hospitals for trauma centers can operate at most one helicopter, whereas 16 separate heliports might be able to operate more than one helicopter. Moreover, helicopters primarily assist rural patients in more sparsely populated areas, and so, assuming helicopters are a scarce resource, this suggests that they should naturally be more spread out rather than clustered together. Indeed, helicopters are scarce in our case: in our numerical study, we have many more candidate heliports (54) than helicopters (between 5 to 25). For completeness, however, Online Appendix A describes how our method can be generalized to the case where more than one helicopter is allowed at each heliport. Moreover, Online Appendix B formulates multi-period and multi-scenario extensions of our model. Third, to make long-term decisions of where to locate trauma centers and helicopters, our optimization model abstracts away from the detailed *real-time* decision-making processes used in practice. For example, in reality, a central operator for EMS (such as 9-1-1 service in the United States) keeps track of each helicopter's availability at each point in time, and dispatches either an ambulance or a helicopter, depending on which option would provide faster service given the system state. The central operator also monitors the availability of beds in each trauma center, and may divert a patient to a farther away, yet less congested, trauma center. To test the performance of the location solutions

[5] Branas and Revelle (2001) also make the same assumption, saying: "This is a consideration that is both realistic and advantageous in analyzing state [Maryland] trauma systems because the number of ambulance depots at the state level is prohibitively large and only a relatively small percentage of ambulance transports are devoted to severe trauma."

our optimization models generate, we use a simulation that captures these real-time decision-making processes. We present details of our simulation model and test results in §5.

4 Optimization Models and Solution Methods

In this section, we model and solve our problem as described in §3. In §4.1, we present an integrated model and outline a general scheme that iteratively solves a sequence of problem relaxations and restrictions to sequentially find tighter bounds for the integrated model. In §4.2, we describe two specific solution methods that tighten the optimality gap in the integrated approach: the Shifting Quadratic Envelopes (SQE) method and a method based on Generalized Benders Decomposition (GBD). Finally, in §4.3, as a point of comparison, we develop heuristic methods that build on existing approaches from the literature.

4.1 Integrated Model and Approach

In our integrated approach, we endogenize the computation of busy fractions within an optimization problem, and formulate the problem as a mixed-integer nonlinear program (MINLP). We explicitly model the allocation of patient demands to trauma centers as well as to ambulances and helicopters. As noted earlier, this allocation represents the long-run average allocation rather than the real-time allocation (which we simulate later in §5). Consequently, our math program has four principal decision variables:

- binary variables y_j that indicate whether or not a trauma center should be opened at site j;
- binary variables x_h that indicate whether or not a helicopter is stationed at heliport h;
- continuous variables s^G_{ij} that represent the (expected) number of patients per unit time to transport from demand region i to trauma center j by ambulance; and
- continuous variables s_{hij} that represent the (expected) number of patients per unit time to transport from demand region i to trauma center j using a helicopter originating from heliport h.

We also define the following four quantities, which are auxiliary decision variables in our math program:

$$\lambda^G = \sum_{i\in I, j\in F^G_i} s^G_{ij}; \tag{1}$$

$$\lambda_h = \sum_{(i,j)\in F_h} s_{hij} \;\forall h \in H; \tag{2}$$

$$\lambda_j = \sum_{i\in F^G_j} s^G_{ij} + \sum_{(h,i)\in F_j} s_{hij} \;\forall j \in J; \tag{3}$$

$$r_h = \sum_{(i,j)\in F_h} \tau_{hij} s_{hij} \;\forall h \in H, \tag{4}$$

where λ^G in (1) represents the total number of patients that we plan to transport by ground ambulance (across all demand regions and all trauma centers); λ_h in (2) represents the total number of patients that we plan to transport using heliport h; λ_j in (3) represents the total number of patients that we plan to transport to trauma center j; and r_h in (4) is the workload assigned to heliport h, which can be explained as follows. Each patient assigned to heliport h uses some helicopter time. Specifically,

a patient flown from demand region i to trauma center j using a helicopter originating from heliport h causes a helicopter to be in service for τ_{hij} units of time, i.e. the time it takes to fly the circuit $h \to i \to j \to h$, plus loading, unloading, and cleanup. The total workload generated by all patients flying from i to j using a helicopter from h is $\tau_{hij} s_{hij}$, which is unitless because τ_{hij} is measured in units of time/patient while s_{hij} is measured in units of patients/time. The workload assigned to heliport h is simply the sum of the workloads from all patients that the plan assigns to heliport h.[6] Table 1 summarizes our notation.

Symbol	Definition
k	Number of trauma centers to be located
m	Number of helicopters to be located
i	Index for demand regions; $i \in I = \{1, 2, ..., \bar{i}\}$
h	Index for heliports; $h \in H = \{1, 2, ..., \bar{h}\}$
j	Index for trauma centers; $j \in J = \{1, 2, ..., \bar{j}\}$
λ_i	Expected demand for region i
c_j	Capacity of trauma center j
d_i^r	Road distance between the center of demand region i and its nearest ambulance station
d_{ij}^r	Road distance between the center of demand region i and trauma center j
d_{ij}	Euclidean distance between the center of demand region i and trauma center j
d_{jh}	Euclidean distance between trauma center j and heliport h
d_{hi}	Euclidean distance between heliport h and the center of demand region i
d_{ground} (d_{air})	Maximum distance that can be covered by an ambulance (a helicopter)
τ_{hij}	Average service time for a helicopter to fly the circuit $h \to i \to j \to h$
F^G, F_i^G, F_j^G	Sets of patients that are covered by ambulances (see §3 for their precise definitions)
F, F_h, F_i, F_j	Sets of patients that are covered by helicopters (see §3 for their precise definitions)
y_j	Variable: Equals 1 if a trauma center is opened at site j, or otherwise equals 0
x_h	Variable: Equals 1 if a helicopter is stationed at heliport h, or otherwise equals 0
s_{ij}^G	Variable: Number of patients per unit time to transport from i to j by ambulance
s_{hij}	Variable: Number of patients per unit time to transport from i to j by helicopter h
λ^G	Variable: Total number of patients per unit time to be transported by ambulance
λ_h	Variable: Total number of patients per unit time to be transported by helicopter h
λ_j	Variable: Total number of patients per unit time to be transported to trauma center j
r_h	Variable: Workload assigned to heliport h

Table 1: Summary of Notation

Ideally, we would like to maximize the expected number of patients that are transported and begin to receive care at a trauma center by the 60-minute threshold. Such an objective would simultaneously incorporate the congestion at both heliports and trauma centers. However, the expression for such an objective involves convolutions of random variables, and is too complex to work with. Instead, we maximize the expected number of patients that are transported without delay, and use a constraint to

[6] In practice, there may be times when a helicopter can fly directly to pick up its next patient without returning to its home heliport. Although such call-to-call travel does not affect the heliports located on the roofs of trauma centers (which comprise 38 out of 54 candidate heliports), it may shorten the service time of a helicopter located in a heliport separate from a trauma center. Berman and Vasudeva (2005) have proposed an approximate approach to model such call-to-call travel. However, if we follow their approach, our objective becomes highly nonlinear and does not yield a tractable solution. When we retain our existing location solutions and add call-to-call travel to our simulation, we find the percentage of successful patients increases by 1%-3%, and that our main results presented in §5.3 remain valid.

ensure sufficient capacity exists at each trauma center, so that patients rarely need to wait for a trauma center bed once they get there. To write down an expression for our objective, we note that when heliport h is open (i.e. $x_h = 1$), heliport h can be considered as a single-server queue with the helicopter as the server, arrival rate λ_h, mean service time $\tau_h = r_h/\lambda_h$, and utilization r_h (where workload and utilization are equivalent in a single-server queue). Then, the probability that an arriving patient finds heliport h busy is equal to the heliport's utilization r_h under the following two assumptions: (i) a patient will wait for a helicopter as needed (the heliport queue backlogs demands; it is not a loss system),[7] and (ii) patient arrivals are Poisson.[8] In other words, r_h is the site-specific busy fraction for heliport h, and it is endogenously-computed, since it depends on the decision variables $\{s_{hij}\}$ (see (4)). Thus, the total number of patients that we expect to be transported (by helicopter or ambulance) without any delay is:

$$\sum_{h \in H} (1 - r_h)\lambda_h + \lambda^G, \tag{5}$$

where, since we assume there are ample ambulances, all λ^G ambulance-transported patients are transported without delay. Our math program maximizes (5), which is a proxy for the expected number of patients transported within 60 minutes.[9] Our objective is consistent with the so-called "expected covered demand" objective commonly used in the literature starting from Daskin (1983) (see, e.g., §11.2.1 of Berman and Krass 2002, Sorensen and Church 2010, and references therein). However, we make the important distinction that we determine heliport-specific busy fractions r_h *endogenously* instead of estimating busy fractions exogenously that are specific to each demand region i (see our earlier discussion in §2).

To ensure that trauma center congestion is kept in check, we pre-compute an appropriate value for the effective capacity c_j of each trauma center j. We briefly describe how we derive an appropriate value for c_j from a probabilistic constraint, while presenting details in Online Appendix C. By following Marianov and Serra (1998) and Berman and Krass (2002) who use an M/M/k queueing model to approximate the flow of patients through a trauma center, we can express the probabilistic constraint Prob[waiting time at trauma center $j \leq \omega$] $\geq \xi$ as $\rho_j \leq \rho_j^{\omega,\xi}$, where ρ_j is the total workload assigned to trauma center j and $\rho_j^{\omega,\xi}$ is a constant that depends on ω and ξ. Moreover, by defining μ_j as the service rate of each server at trauma center j, we can rewrite the constraint $\rho_j \leq \rho_j^{\omega,\xi}$ as $\lambda_j \leq \mu_j \rho_j^{\omega,\xi}$. Finally, we define the effective capacity of trauma center j as $c_j = \mu_j \rho_j^{\omega,\xi}$, and impose a capacity constraint of

[7] Having patients queue for service is in line with Ball and Lin (1993). Others (e.g., see Borras and Pastor 2002) assume that patients do not wait for ambulances and find alternate (private) modes of transportation. Our assumption seems reasonable because in our problem helicopters transport only those patients who are far away from a trauma center and cannot be reached by ambulance within 60 min.

[8] Poisson arrivals is a common assumption in the literature, made for tractability as a first-order approximation of complex systems (e.g., see Berman and Krass 2002, Zhang et al. 2010) even in cases when Markovian assumptions may not hold in a strict sense.

[9] In fact, our objective function is a conservative underestimate for the number of patients transported within 60 minutes. This follows from the fact that the probability a patient experiences no delay is less than or equal to the probability that a patient's delay is small enough that s/he can be transported to a trauma center within 60 minutes.

the form $\lambda_j \leq c_j y_j$, which limits the number of patients served by trauma center j to c_j when it is open, or to zero when it is closed.

Putting all this together, we write our full MINLP model as follows:

$$
\begin{aligned}
(P) \qquad & \max \lambda^G + \sum_{h \in H} (1 - r_h)\lambda_h \\
& \text{s.t. (1)-(4)} \\
& r_h \leq x_h \; \forall h \in H && (6) \\
& \sum_{j \in J} y_j \leq k && (7) \\
& \sum_{h \in H} x_h \leq m && (8) \\
& \sum_{j \in F_i^G} s_{ij}^G + \sum_{(h,j) \in F_i} s_{hij} \leq \lambda_i \; \forall i \in I && (9) \\
& \lambda_j \leq c_j y_j \; \forall j \in J && (10) \\
& x_j \leq y_j \; \forall j \in J && (11) \\
& s_{ij}^G \geq 0 \; \forall (i,j) \in F^G; \qquad s_{hij} \geq 0 \; \forall (h,i,j) \in F && (12) \\
& y_j \in \{0,1\} \; \forall j \in J; \qquad x_h \in \{0,1\} \; \forall h \in H. && (13)
\end{aligned}
$$

Constraint (6) ensures that the busy fraction (or utilization) of heliport h, r_h, should be less than or equal to 1; in other words, heliports should not be overloaded. Constraints (7) and (8) ensure that at most k trauma centers are opened and at most m helicopters are stationed across all heliports. Constraint (9) says we cannot plan to serve more people from region i than the expected demand λ_i from that region, and constraint (10) is our capacity constraint that keeps congestion at trauma center j under control. Constraint (11) makes sure that when a trauma center is closed, so is the heliport on its roof (recall that the set $H(\supseteq J)$ is indexed such that heliport j is on the roof of trauma center j). Finally, constraint (12) makes sure that the number of patients served by all transportation modes must be nonnegative, and constraint (13) makes sure that each trauma center is either open or closed, and each heliport is assigned either one helicopter or no helicopter. Note that, taken together, constraints (2), (4), and (6) enforce the condition that no demands are allocated to closed heliports (i.e., $x_h = 0 \Rightarrow s_{hij} = 0 \; \forall (i,j) \in F_h \Rightarrow \lambda_h = 0$).

There are a few ways that our model differs from much of the existing literature. Instead of pre-grouping demand regions into districts and assuming that each district is served by a pool of helicopters, we make no such assumptions (so called "districting assumptions" in the literature), and allow the math program to determine the assignment of patients to heliports through decision variables s_{hij}. As discussed in Borras and Pastor (2002), demand-area-specific busy fractions within a district can either be server-independent or server-dependent, which boils down to whether servers within a district are modeled as independent single-server queues or one multi-server queue, respectively. In the body of our paper, we assume that each heliport is its own single-server queue. However, if we allow multiple

helicopters per heliport, we can model server dependence, as described in Online Appendix A. Moreover, in our model, because busy fractions are endogenous, we also have another type of dependence, which spans across heliports (analogous to dependence spanning across districts). To illustrate this point, imagine that a demand region can be served by two heliports, h_1 (nearby) and h_2 (further away). Initially, it is optimal to direct patients to h_1. However, as utilization r_{h_1} rises, congestion at h_1 increases, and the math program begins to direct patients to h_2 (which increases r_{h_2}). Consequently, demands get balanced across heliports h_1 and h_2, which is mediated by the fact that the busy fractions r_{h_1} and r_{h_2} are linked through decision variables s_{hij}.

The chief computational difficulty in solving the MINLP problem (P) is the set of non-convex bilinear terms $\lambda_h r_h$ that appear in the objective. As others (c.f. Floudas and Pardalos 2012) have reported, bilinear terms can be notoriously challenging to cope with. In our case, these bilinear terms are embedded in a generalized facility location problem that models both trauma center and heliport locations, as well as the routing of helicopters and ambulances. The resulting problem is significantly more difficult to solve than the canonical facility location problem with linear objective, which itself is hard (c.f. Owen and Daskin 1998). Fortunately, we are able to exploit problem-specific structure to find solutions to (P) that are near-optimal and significantly outperform our benchmark heuristics.

We find solutions of (P) by iteratively solving a sequence of problem relaxations and restrictions, all of which are convex optimization problems and can be solved using a Mixed Integer Quadratic Programming (MIQP) solver such as CPLEX. This turns out to be a more computationally efficient approach than using a general global optimization solver such as BARON, which has the ability to cope with non-convexities but doesn't exploit the problem-specific structure as well as our specialized methods. Our general scheme works as follows. Since we are maximizing the objective, any relaxation yields a valid upper bound, while any restriction produces a lower bound. At each point in time, we can compute an optimality gap by taking the difference between the best (lowest) upper bound and the best (highest) lower bound found thus far, and use this gap to determine whether to continue iterating or stop. In the following, we first introduce some relaxations of (P) in §4.1.1 and then a restriction of (P) in §4.1.2. In §4.2, we describe two methods that we use to reduce the optimality gap.

4.1.1 Relaxations

First, we relax (P) by using McCormick envelopes (McCormick 1976, Floudas and Pardalos 2012) to *linearly* outer-approximate the bilinear $\lambda_h r_h$ terms. A Mixed Integer Linear Program (MILP) relaxation of (P) based on McCormick envelopes is:

$$
\begin{aligned}
(P_{McCormick}) \qquad & \max \lambda^G + \sum_{h \in H} \lambda_h - \sum_{h \in H} w_h \\
& \text{s.t. (1)-(4), (6)-(13)} \\
& \quad w_h \geq \lambda_h^{MAX} r_h + r_h^{MAX} \lambda_h - \lambda_h^{MAX} r_h^{MAX} \; \forall h \in H \qquad (14) \\
& \quad w_h \geq 0 \; \forall h \in H. \qquad (15)
\end{aligned}
$$

Let us explain how we have obtained ($P_{McCormick}$). The McCormick envelopes are derived using known constants λ_h^{MIN}, λ_h^{MAX}, r_h^{MIN} and r_h^{MAX}, which are lower and upper bounds on the allocated demand rate and workload, respectively. Specifically, the McCormick envelope for the bilinear expression $w_h = \lambda_h r_h$ is:

$$w_h \geq \lambda_h^{MAX} r_h + r_h^{MAX} \lambda_h - \lambda_h^{MAX} r_h^{MAX} \tag{16}$$

$$w_h \geq \lambda_h^{MIN} r_h + r_h^{MIN} \lambda_h - \lambda_h^{MIN} r_h^{MIN} \tag{17}$$

$$w_h \leq \lambda_h^{MAX} r_h + r_h^{MIN} \lambda_h - \lambda_h^{MAX} r_h^{MIN} \tag{18}$$

$$w_h \leq \lambda_h^{MIN} r_h + r_h^{MAX} \lambda_h - \lambda_h^{MIN} r_h^{MAX}. \tag{19}$$

Since our objective will try to make w_h as small as possible, it is only the lower bounds for w_h that are needed in our formulation. Thus we include only (14) and (15) in ($P_{McCormick}$), where (14) is the collection of constraints of type (16) for all heliports, and (15) was obtained by substituting $\lambda_h^{MIN} = r_h^{MIN} = 0$ into (17) for all heliports. The values for λ_h^{MAX} and r_h^{MAX} are instance-specific; for example, we can define $r_h^{MAX} = 1$ as the maximum utilization at heliport h and $\lambda_h^{MAX} = \sum_{i \in I:(i,j) \in F_h} \lambda_i$ as the total demand from regions near heliport h. In general, the relaxation ($P_{McCormick}$) can be tightened by using smaller bounds λ_h^{MAX} and r_h^{MAX}, which we derive in Online Appendix D. However, it turns out that the relaxation ($P_{McCormick}$) is quite weak, regardless of the bounds we choose.

To derive a significantly tighter relaxation, we exploit the fact that the variables $\lambda_h = \sum_{i,j} s_{hij}$ and $r_h = \sum_{i,j} \tau_{hij} s_{hij}$ are both defined as linear combinations of s_{hij} variables. In particular, we can interpret $\tau_h = r_h/\lambda_h$ as the *mean service time* of heliport h; that is, the amount of time it takes a helicopter stationed at heliport h to fly $h \to i \to j \to h$, averaged over all pick-up and drop-off points (i, j). Defining $\tau_h^{MAX} = \max_{i,j} \tau_{hij}$ and $\tau_h^{MIN} = \min_{i,j} \tau_{hij}$ as the maximum and minimum mean service times respectively, we derive a *quadratic* outer-approximation by sandwiching the bilinear term $\lambda_h r_h$ as follows:

$$\tau_h^{MIN} \leq r_h/\lambda_h \leq \tau_h^{MAX} \Leftrightarrow \tau_h^{MIN} \lambda_h^2 \leq \lambda_h r_h \leq \tau_h^{MAX} \lambda_h^2.$$

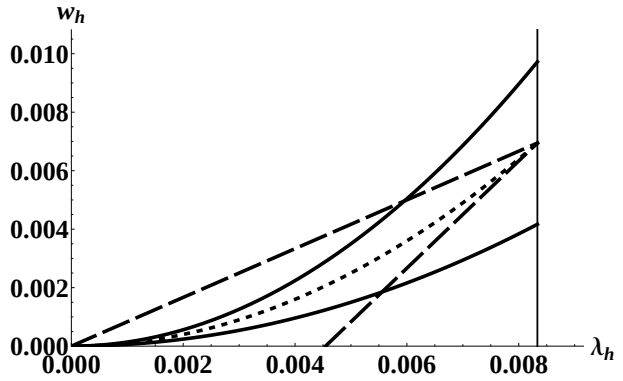
The lower bound for $\lambda_h r_h$ gives us the following Mixed Integer Quadratic Program (MIQP) relaxation of (P):

$$(P_M^{SQE}) \qquad \max \lambda^G + \sum_{h \in H} \lambda_h - \sum_{h \in H} \tau_h^{MIN} \lambda_h^2$$

$$\text{s.t. (1)-(4), (6)-(13).}$$

To compare the quadratic envelope used in (P_M^{SQE}) with the linear McCormick envelope used in ($P_{McCormick}$), we first introduce some notation and then compare both envelopes using an example. Treating the mean service time τ_h as known and fixed, we let $r_h(\lambda_h|\tau_h) = \tau_h \lambda_h$ and $w_h(\lambda_h|\tau_h) = r_h(\lambda_h|\tau_h)\lambda_h = \tau_h \lambda_h^2$ denote the workload r_h and the quantity w_h, respectively, as a function of λ_h. Given a problem instance of (P) with $\tau_h^{MIN} = 60$ minutes and $\tau_h^{MAX} = 140$ minutes, we know that the optimal value for τ_h will be in the range [60, 140]. Let us assume, for purposes of illustration, that the optimal τ_h is midway between its bounds; i.e., $\tau_h = 100$ minutes. Figure 2 plots $w_h(\lambda_h|100)$ (dotted curve), which is sandwiched by the quadratic envelope between $w_h(\lambda_h|\tau_h^{MIN})$ and $w_h(\lambda_h|\tau_h^{MAX})$ (two

solid curves). The linear McCormick envelope for $\tau_h = 100$ minutes (two dashed lines) is computed using (16)-(19) and $r_h(\lambda_h|\tau_h)$ as follows. The bottom line comes from (16), and is the McCormick lower bound for w_h, assuming $\tau_h = 100$ is fixed (which is tighter than $w_h \geq 0$ from (17)). The top line is the McCormick upper bound for w_h, assuming $\tau_h = 100$ is fixed, and is derived from (18) (which in this example is tighter than (19)). Recall that the lower bound for w_h, not the upper bound, is important in our formulation because our objective will try to make w_h as small as possible. As Figure 2 shows, for low values of λ_h the quadratic envelope is tighter (higher), whereas the linear McCormick envelope is tighter for high values of λ_h (the lowest solid curve crosses the lowest dashed line at $\lambda_h = 0.0055$ per minute). Note that we used $\lambda_h^{MAX} = 1/(2\tau_h^{MIN}) = 1/120$, which comes from the tightened bounds described in Online Appendix D.



In order, reading from top to bottom along the right-hand-side of λ_h's domain

- $w_h(\lambda_h|\tau_h^{MAX})$
- $w_h = \lambda_h^{MAX} r_h(\lambda_h|\tau_h)$
- $w_h(\lambda_h|\tau_h)$
- $w_h = \lambda_h^{MAX} r_h(\lambda_h|\tau_h) + r_h^{MAX}\lambda_h - \lambda_h^{MAX} r_h^{MAX}$
- $w_h(\lambda_h|\tau_h^{MIN})$

Figure 2: Comparison of the quadratic envelope used in (P_M^{SQE}) with the linear McCormick envelope used in $(P_{McCormick})$, plotted over the range $\lambda_h \in [0, \lambda_h^{MAX} = 1/120]$.

In our computational experiments, we have observed that formulation (P_M^{SQE}) solves faster than $(P_{McCormick})$, suggesting that the quadratic envelope is usually tighter than the McCormick envelope, possibly owing to the fact that the continuous relaxation of our problem spreads demands across many heliports, causing λ_h to be on the low side, where the quadratic envelope is tighter as depicted in Figure 2. We conjectured that we could even do better by enforcing both the quadratic and McCormick envelopes, as in the following (P_M^{GBD}) formulation (where $w_h \geq 0$ is redundant):

$$
\begin{aligned}
(P_M^{GBD}) \qquad & \max \lambda^G + \sum_{h\in H} \lambda_h - \sum_{h\in H} w_h \\
& \text{s.t. (1)-(4), (6)-(13)} \\
& \quad w_h \geq \tau_h^{MIN} \lambda_h^2 \ \forall h \in H \\
& \quad w_h \geq \lambda_h^{MAX} r_h + r_h^{MAX} \lambda_h - \lambda_h^{MAX} r_h^{MAX} \ \forall h \in H \\
& \quad w_h \geq 0 \ \forall h \in H.
\end{aligned}
$$

However, this turns out to be a bad idea because CPLEX is better at handling the quadratic objective

of (P_M^{SQE}) than the quadratic constraints in (P_M^{GBD}); see Online Appendix E for computational details. Therefore, in §4.1.2, we will use (P_M^{SQE}) as our master problem.

4.1.2 Restriction

Solving the master problem (P_M^{SQE}) gives us a feasible solution to (P), since all of (P)'s constraints are present in (P_M^{SQE}). We evaluate the quality of this solution using the true objective from (P), i.e., $\lambda^G + \sum_h \lambda_h - \sum_h \lambda_h r_h$. The optimality gap of this solution is the difference between the optimal value of (P_M^{SQE}) and the true value of this solution; i.e., $(\lambda^G + \sum_h \lambda_h - \sum_h \tau_h^{MIN} \lambda_h^2) - (\lambda^G + \sum_h \lambda_h - \sum_h \lambda_h r_h) = \sum_h (\lambda_h r_h - \tau_h^{MIN} \lambda_h^2)$. Notice that since $r_h/\lambda_h \geq \tau_h^{MIN}$, the optimality gap will always be nonnegative.

Although we could use the feasible solution from (P_M^{SQE}) directly, we can often find better feasible solutions to (P) by *re-optimizing* over a subset of the decision variables. Specifically, from a feasible solution from (P_M^{SQE}), we fix the set of open trauma centers $\{y_j\}$, the locations of helicopters $\{x_h\}$, the demand for helicopters $\{\lambda_h\}$, and the demand for ambulances $(\{s_{ij}^G\}, \lambda^G)$. Then, we ignore the helicopter routing pattern $\{s_{hij}\}$ suggested by the master problem and re-assign helicopter-transported patients across heliports and trauma centers, with the goal of shifting the workload $\{r_h\}$ to the heliports that are under-utilized. That is, given the fixed values for the set of master problem variables $\Theta = \left\{\{y_j\}, \{x_h\}, \lambda^G, \{\lambda_h\}, \{s_{ij}^G\}\right\}$, we solve a *restriction* of (P) to optimize over the remaining variables $\{\{s_{hij}\}, \{r_h\}\}$.[10] With the variables Θ fixed, constraints (1), (7), (8), (11), and (13) in (P) can be ignored because any feasible master problem solution already satisfied these constraints. For brevity, define the constants $a_j = c_j y_j - \sum_i s_{ij}^G \; \forall j$ and $b_i = \lambda_i - \sum_j s_{ij}^G \; \forall i$, which depend only on problem data and the fixed variables Θ. Our restriction of (P), which we call our *subproblem*, optimizes over the variables $\{\{r_h\}, \{s_{hij}\}\}$ and is defined as the following Linear Program (LP):

$$
\begin{aligned}
(P_S^{\Theta}) \qquad \min \; & \sum_{h \in H} \lambda_h r_h \\
\text{s.t.} \; & \sum_{(h,i) \in F_j} s_{hij} \leq a_j \; \forall j \in J \\
& \sum_{(j,h) \in F_i} s_{hij} \leq b_i \; \forall i \in I \\
& \sum_{(i,j) \in F_h} s_{hij} = \lambda_h \; \forall h \in H \\
& \sum_{(i,j) \in F_h} \tau_{hij} s_{hij} - r_h = 0 \; \forall h \in H \\
& r_h \leq x_h \; \forall h \in H \\
& s_{hij} \geq 0 \; \forall (h,i,j) \in F.
\end{aligned}
$$

The optimality gap between the master problem and subproblem is still measured as

$$
(\lambda^G + \sum_{h \in H} \lambda_h - \sum_{h \in H} \tau_h^{MIN} \lambda_h^2) - (\lambda^G + \sum_{h \in H} \lambda_h - \sum_{h \in H} \lambda_h r_h) = \sum_{h \in H} (\lambda_h r_h - \tau_h^{MIN} \lambda_h^2), \tag{20}
$$

[10] Technically, since λ_j depends on s_{hij}, it is also re-optimized when we solve (P_S^{Θ}). However, since λ_j does not appear directly in (P_S^{Θ}), it is clearer to omit λ_j from the set of remaining variables.

but now the (smaller) $\{r_h\}$ values come from the subproblem instead of the master problem, and as a result the gap is reduced.

4.2 Solution Methods for Tightening the Relaxation

Once we have a feasible solution to (P) and a corresponding optimality gap, the next question is: Can we make any inferences from the incumbent solution or its dual that allow us to tighten the master problem's relaxation and thereby reduce the optimality gap? A tighter relaxation of (P_M^{SQE}) yields not only a tighter (i.e., lower) upper bound for (P), but also when Θ is fixed to a solution that is closer to the true optimum, the subproblem (P_S^{Θ}) finds better (i.e., higher) lower bounds for (P). We have studied two methods to tighten the master problem's relaxation: a method that relies on shifting quadratic envelope boundaries (§4.2.1), and a Benders-like cut generation scheme (§4.2.2).

4.2.1 Shifting Quadratic Envelopes (SQE) Algorithm

The first method that we use to tighten the master problem uses one or more quadratic envelopes for each heliport. We use binary variables to control which envelope is active at each candidate solution, and re-define the envelope boundaries at each iteration in an attempt to tighten the relaxation. We first describe a naïve implementation of our approach, and then discuss modifications that we have made for computational tractability.

From (P_S^{Θ}), we have a feasible solution to (P), which we denote $\Psi = \Theta \cup \{\{r_h\}, \{s_{hij}\}\} = \left\{\{y_j\}, \{x_h\}, \lambda^G, \{\lambda_h\}, \{\lambda_j\}, \{s_{ij}^G\}, \{r_h\}, \{s_{hij}\}\right\}$. The two crucial variables that describe the performance of heliport h are its allocated demand λ_h and its workload r_h. Therefore, for each heliport h, we use the point $(\lambda_h, r_h) \in \mathbb{R}^2$ from the feasible solution Ψ to update the quadratic envelope boundaries for heliport h. Let $\tau_h = r_h/\lambda_h$ be the mean service time of candidate solution (λ_h, r_h) and $\tau_h^* = r_h^*/\lambda_h^*$ be the mean service time of the optimal solution (λ_h^*, r_h^*) of (P). For each heliport h, our goal is to pick a mean service time τ_h^{FIX} that underestimates the optimal mean service time τ_h^* as closely as possible. This will allow us to closely underestimate the bilinear term $w_h = \lambda_h r_h$ using a convex quadratic function of λ_h. Define two functions of λ_h; namely, $r_h(\lambda_h|\tau_h^{FIX}) = \tau_h^{FIX}\lambda_h$ and $w_h(\lambda_h|\tau_h^{FIX}) = \lambda_h r_h(\lambda_h|\tau_h^{FIX}) = \tau_h^{FIX}\lambda_h^2$. As long as $\tau_h \geq \tau_h^{FIX}$, it follows by definition that $r_h(\lambda_h|\tau_h^{FIX}) \leq r_h$ and $w_h(\lambda_h|\tau_h^{FIX}) \leq w_h$. That is, for points (λ_h, r_h) where $r_h \geq r_h(\lambda_h|\tau_h^{FIX}) = \lambda_h\tau_h^{FIX}$, the convex quadratic function $w_h(\lambda_h|\tau_h^{FIX}) = \tau_h^{FIX}\lambda_h^2$ provides a valid underestimate for w_h. Since, in general, we know only that $\tau_h^* \in [\tau_h^{MIN}, \tau_h^{MAX}]$ before solving our problem (P), the best choice for τ_h^{FIX} that is always guaranteed to underestimate the optimal mean service time is τ_h^{MIN}. This is why, in §4.1.1, we used the underestimate $\tau_h^{FIX} = \tau_h^{MIN}$ to construct (P_M^{SQE}) by replacing the bilinear terms $w_h = \lambda_h r_h$ in the objective of (P) with their quadratic relaxations $w_h(\lambda_h|\tau_h^{MIN}) = \tau_h^{MIN}\lambda_h^2$. We will now describe how we can use other estimates for τ_h^{FIX} that lead to tighter relaxations.

Notice that since $r_h = \tau_h\lambda_h$ and $\tau_h \in [\tau_h^{MIN}, \tau_h^{MAX}]$, all feasible (λ_h, r_h)-points must lie in the cone defined by $\tau_h^{MIN}\lambda_h \leq r_h \leq \tau_h^{MAX}\lambda_h$, as shown in Figure 3 (e.g., consider $r_h(\lambda_h|\tau_{h,1}) = \tau_h^{MIN}\lambda_h$ and $r_h(\lambda_h|\tau_{h,2}) = \tau_h^{MAX}\lambda_h$ in Figure 3(a)). Moreover, we can subdivide this cone into m_h slices of equal size by splitting the domain of τ_h into the m_h subdomains $[\tau_{h,1}, \tau_{h,2}], [\tau_{h,2}, \tau_{h,3}], \ldots, [\tau_{h,m_h}, \tau_{h,m_h+1}]$, where $\tau_{h,n} = \tau_h^{MIN} + ((n-1)/m_h)(\tau_h^{MAX} - \tau_h^{MIN})$ for $n = 1, 2, ..., m_h + 1$. When the solution (λ_h, r_h) is in

the nth subdomain (i.e., in the slice defined by $\tau_{h,n}\lambda_h \le r_h \le \tau_{h,n+1}\lambda_h$), we can use $\tau_h^{FIX} = \tau_{h,n}$ as an underestimate for the true mean service time τ_h at the point (λ_h, r_h). To keep track of which subdomain each heliport's mean service time is in, we use binary variables; i.e. we let $x_{hn} = 1$ if heliport h's mean service time is in the subdomain $[\tau_{h,n}, \tau_{h,n+1}]$, or $x_{hn} = 0$ otherwise. Moreover, we use the constraint $\sum_n x_{hn} = 1$ to ensure that each heliport's mean service time falls in exactly one subdomain (and when τ_h is on the boundary of two subdomains, we count heliport h's mean service time as being in only one of the neighboring subdomains). Such a setup allows us to activate the lower bound $r_h \ge r_h(\lambda_h|\tau_{h,n})$ whenever the solution (λ_h, r_h) to (P_M^{SQE}) is in the n^{th} subdomain, thereby replacing the objective term $w_h = \lambda_h r_h$ of (P) with the quadratic underestimate $w_h(\lambda_h|\tau_{h,n}) = \tau_{h,n}\lambda_h^2$. The full formulation of (P_M^{SQE}), which has slices indexed by the sets $N_h = \{1, ..., m_h\}$, is as follows:

$$
\begin{aligned}
(P_{M2}^{SQE}) \qquad & \max \lambda^G + \sum_{h\in H} \lambda_h - \sum_{h\in H, n\in N_h} \tau_{h,n}\lambda_{hn}^2 \\
& \text{s.t. (1)-(4), (6)-(13)} \\
& \quad \lambda_h = \sum_{n\in N_h} \lambda_{hn}, \quad r_h = \sum_{n\in N_h} r_{hn} && \forall h \in H \\
& \quad 0 \le \lambda_{hn} \le \lambda_h^{MAX} x_{hn}, \quad 0 \le r_{hn} \le r_h^{MAX} x_{hn} && \forall h \in H,\ \forall n \in N_h \\
& \quad \tau_{h,n}\lambda_{hn} \le r_{hn} \le \tau_{h,n+1}\lambda_{hn} && \forall h \in H,\ \forall n \in N_h \\
& \quad \sum_{n\in N_h} x_{hn} = 1 && \forall h \in H \\
& \quad x_{hn} \in \{0,1\} && \forall h \in H,\ \forall n \in N_h.
\end{aligned}
$$

In addition to the variables and constraints of (P_M^{SQE}), the formulation (P_{M2}^{SQE}) includes the binary variables $\{x_{hn}\}$ that define subdomain membership, new continuous variables $\{\lambda_{hn}\}$ and $\{r_{hn}\}$, and several logical constraints that link these quantities. The added constraints make sure that when $x_{hn} = 1$, then λ_{hn} and r_{hn} are equal to the allocated demand and workload of heliport h (i.e., $\lambda_{hn} = \lambda_h$ and $r_{hn} = r_h$), whereas when $x_{hn} = 0$, then $\lambda_{hn} = r_{hn} = 0$. Therefore, for each heliport h, only one conic slice $\tau_{h,n'}\lambda_{hn'} \le r_{hn'} \le \tau_{h,n'+1}\lambda_{hn'}$ is ever active at a time (corresponding to the n' with $x_{hn'} = 1$).

As we subdivide the domain of each τ_h into finer slices, our relaxation (P_{M2}^{SQE}) becomes tighter. Moreover, as the number of slices $m_h \to \infty$ for all heliports h, the area of each slice collapses to zero and the optimal value of (P_{M2}^{SQE}) converges to the optimal value of (P); that is, in theory we can approximate (P) to any arbitrary precision by simply slicing the subdomains of τ_h finely enough. However, as the number of slices m_h increases, the problem (P_{M2}^{SQE}) becomes much harder to solve due to the larger number of binary variables $\{x_{hn}\}$. As a result, we abandon the idea of using equally-spaced slices, and instead use a more efficient method to decide where to slice each cone $\tau_h^{MIN}\lambda_h \le r_h \le \tau_h^{MAX}\lambda_h$, thereby producing a tight relaxation of (P) using only a few slices.

Our Shifting Quadratic Envelopes (SQE) algorithm begins with only one slice defined for each heliport; i.e., $m_h = 1$, $\tau_{h,1} = \tau_h^{MIN}$, and $\tau_{h,2} = \tau_h^{MAX}$ $\forall h$. At each iteration and for each heliport, the algorithm subdivides one (judiciously chosen) slice into two. The general idea is that we should focus

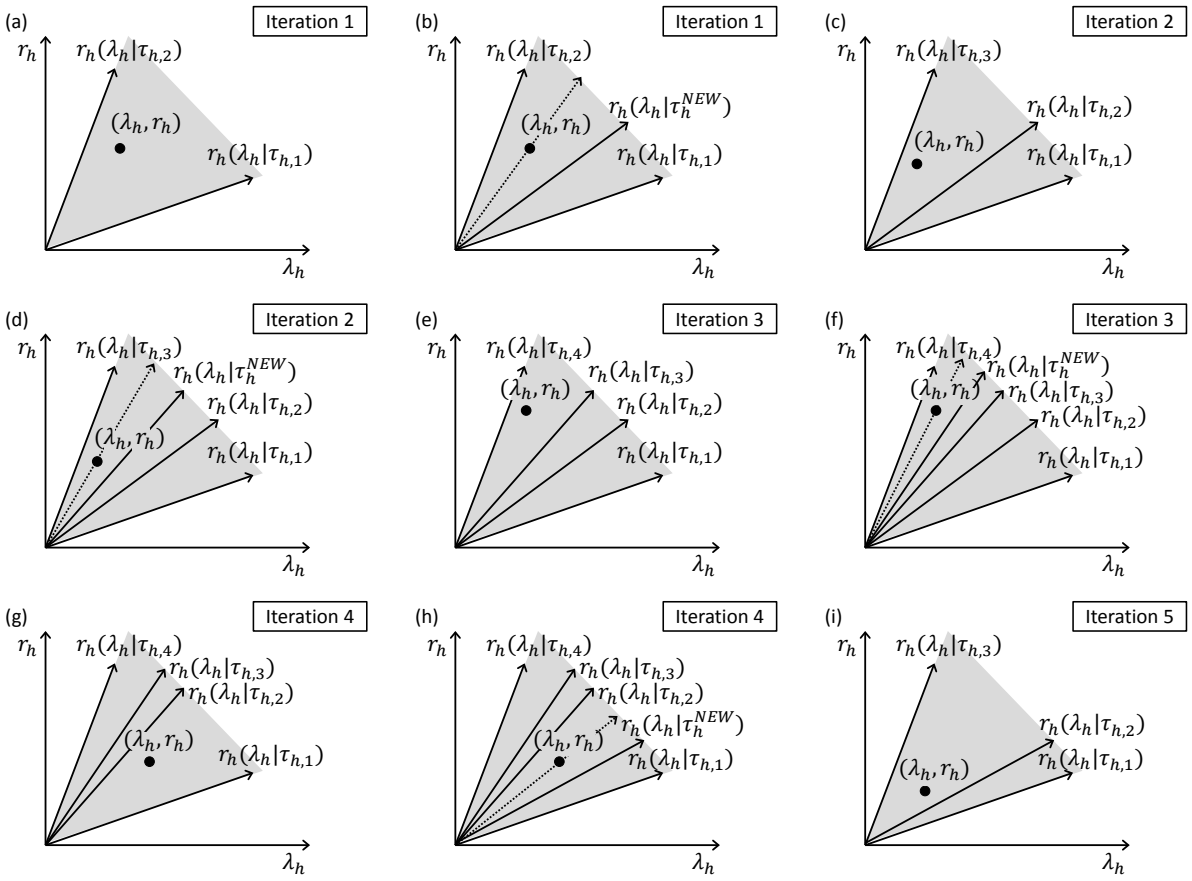


Figure 3: Illustration of the Shifting Quadratic Envelopes algorithm

our attention on narrowly refining the partition of τ_h's domain in areas that are likely to be close to the (P)-optimal τ_h^*. Therefore, at each iteration, we solve (P_{M2}^{SQE}) and (P_S^{Θ}) to get the (P)-feasible solution (λ_h, r_h), which lies in a particular slice, say the slice defined by $\tau_{h,n_h^*}\lambda_h \leq \tau_h \leq \tau_{h,n_h^*+1}\lambda_h$ for some n_h^*. Next, we subdivide this slice into two unevenly-sized slices. Because we care about producing underestimates for τ_h rather than overestimates, we cut the slice in two along the ray $r_h(\lambda_h|\tau_h^{NEW}) = \tau_h^{NEW}\lambda_h$, where $\tau_h^{NEW} = (\tau_h + \tau_{h,n_h^*})/2$ is the average of the true mean service time $\tau_h = r_h/\lambda_h$ at the point (λ_h, r_h) and the slope τ_{h,n_h^*} of the lower boundary of that slice (which can be interpreted as the previous underestimate of the mean service time at the point (λ_h, r_h)). Note that, although we choose where to split each slice heuristically, our choices are motivated by efficiency and have no bearing on the correctness of our approach, or on our ability to generate valid quadratic envelopes. Moreover, to keep the total number of slices small, whenever we suspect that the (P)-optimal (λ_h^*, r_h^*) has a low likelihood of being in a particular slice, we merge that slice with the slice below it. From one iteration to the next, we may add or delete slices, but we never use more than a small number of slices per heliport, which keeps the computational complexity of (P_{M2}^{SQE}) in check. It turns out that in practice, keeping just 3 slice boundaries below and 1 slice boundary above the current solution (λ_h, r_h) results in a good trade-off between precision and tractability. Because this procedure causes the boundaries of the subdomains of τ_h to "shift" at each iteration, which in turn define the quadratic envelope boundaries

$w_h(\lambda_h|\tau_{h,n}) = \tau_{h,n}\lambda_h^2, n = 1, 2, \ldots, m_h$, we call this method the Shifting Quadratic Envelopes (SQE) algorithm.

At each iteration of the SQE algorithm, we proceed as follows: (1) solve the master problem (P_{M2}^{SQE}) to get a "good" solution to (P); (2) solve the subproblem (P_S^{Θ}) to improve this solution; (3) evaluate the quality of the improved solution by computing its optimality gap using Equation (20); and (4) terminate if a time limit or optimality gap threshold is reached, otherwise shift the quadratic envelope boundaries in (P_{M2}^{SQE}) and continue. Figure 3 illustrates the progression of the SQE algorithm on a hypothetical example. For each iteration, we plot the subdomains of $[\tau_h^{MIN}, \tau_h^{MAX}]$ for a single heliport h as conic slices in the (λ_h, r_h)-plane. (Although we only describe what happens at a single heliport h, the conic slices are adjusted for all heliports at each iteration of our algorithm.) Graphically, we make use of the fact that, for any fixed value of τ_h, the function $r_h(\lambda_h|\tau_h) = \tau_h\lambda_h$ defines a ray from the origin with slope τ_h. Thus, we begin with a single slice bounded by $r_h(\lambda_h|\tau_{h,1})$ and $r_h(\lambda_h|\tau_{h,2})$, where $\tau_{h,1} = \tau_h^{MIN}$ and $\tau_{h,2} = \tau_h^{MAX}$ (Figure 3(a)). At *Iteration 1*, we solve (P_{M2}^{SQE}) and (P_S^{Θ}), and use the solution (λ_h, r_h) to compute the mean service time $\tau_h = r_h/\lambda_h$ (this τ_h is the slope of the dotted ray in Figure 3(b)). Next, we split the subdomain of $[\tau_{h,1}, \tau_{h,2}]$ into two slices by introducing a new slice boundary $r_h(\lambda_h|\tau_h^{NEW})$, where $\tau_h^{NEW} = (\tau_h + \tau_{h,1})/2$ is the average of the actual mean service time τ_h at the point (λ_h, r_h) and the previous underestimate $\tau_{h,1}$ (Figure 3(b)). At *Iteration 2*, after re-labelling the slice boundaries τ_h^{NEW} and $\tau_{h,2}$ as $\tau_{h,2}$ and $\tau_{h,3}$ respectively, we then re-solve (P_{M2}^{SQE}) and (P_S^{Θ}) to get a new point (λ_h, r_h) and its associated mean service time $\tau_h = r_h/\lambda_h$. Assuming the point (λ_h, r_h) is in the top slice (Figure 3(c)), we proceed by splitting the top slice in two. We do this by creating a new slice boundary with slope $\tau_h^{NEW} = (\tau_h + \tau_{h,2})/2$, i.e. a slope that is midway between the actual mean service time τ_h and the previous underestimate $\tau_{h,2}$ (Figure 3(d)). Once again, we re-label the slice boundaries and re-solve (P_{M2}^{SQE}) and (P_S^{Θ}). At *Iteration 3*, assume the point (λ_h, r_h) also falls into the topmost slice (Figure 3(e)). As in the previous iteration, we split the top slice in two by defining a new region boundary with slope $\tau_h^{NEW} = (\tau_h + \tau_{h,3})/2$ that is midway between $\tau_h = r_h/\lambda_h$ and the underestimate defined by the lower boundary of that slice, $\tau_{h,3}$ (Figure 3(f)). But before the next iteration, we also delete one region boundary to keep the problem size manageable. Motivated by our desire to generate underestimates, we keep up to three region boundaries below the incumbent point (λ_h, r_h), and only one region boundary above. Specifically, we delete the region boundary defined by $r_h(\lambda_h|\tau_{h,2})$, i.e. the lowest region boundary that can be deleted. Note that we must keep the original region boundaries $r_h(\lambda_h|\tau_h^{MIN})$ and $r_h(\lambda_h|\tau_h^{MAX})$, since the solution (λ_h, r_h) may have a mean service time $\tau_h = r_h/\lambda_h$ that falls anywhere in the full domain $[\tau_h^{MIN}, \tau_h^{MAX}]$. At *Iteration 4*, with the slice boundaries re-labelled, we re-solve (P_{M2}^{SQE}) and (P_S^{Θ}) once again. Assuming the new solution (λ_h, r_h) now falls into the bottommost slice (Figure 3(g)), we split the bottommost slice in two by introducing the slice boundary $\tau_h^{NEW} = (\tau_h + \tau_{h,1})/2$ and delete all but one slice boundary that lies above the point (λ_h, r_h); i.e. we delete $r_h(\lambda_h|\tau_{h,2})$ and $r_h(\lambda_h|\tau_{h,3})$ (Figure 3(h)). Finally, *Iteration 5* begins with two slices for heliport h, as shown (Figure 3(i)). The algorithm continues until either the optimality gap is reduced below a desired threshold or a time limit has been reached. Pseudocode for the SQE algorithm

can be found in Online Appendix F.

4.2.2 Algorithm Based on Generalized Benders Decomposition

The second method that we use to tighten the master problem is based on Generalized Benders Decomposition (GBD) (e.g., Benders 1962, Geoffrion 1972). GBD is a technique that can be used to solve a complex math program by structurally decomposing it into a master problem and one or more subproblems. The subproblems' dual solutions are used to infer one or more cuts that are then added to the master problem to make its formulation tighter. GBD iterates back and forth between solving the master problem and subproblems until a provably near-optimal solution to the full problem is found. It is worth pointing out that when the math program being decomposed is nonlinear, special care must be taken to implement GBD to make sure that Benders cuts do not inadvertently cut off the optimal solution; see, e.g., Geromel and Belloni (1986) and Sahinidis and Grossman (1991). In our case, this "special care" requires us to add binary variables to our formulation for each cut generated.

At each iteration t, we augment the master problem (P_M^{GBD}) with a Benders cut of the form $z \leq B_t(\cdot)$, where z is the objective of the master problem. For completeness, the full master problem used in the Benders decomposition takes the form:

$$
\begin{aligned}
(P_{M2}^{GBD}) \qquad & \max\ z \\
& \text{s.t. (1)-(4), (6)-(13)} \\
& \quad z \leq \lambda^G + \sum_{h \in H} \lambda_h - \sum_{h \in H} w_h \\
& \quad z \leq B_t(\cdot)\ \forall t = 1..nCuts \\
& \quad w_h \geq \sum_h \tau_h^{MIN} \lambda_h^2\ \forall h \in H \\
& \quad w_h \geq \lambda_h^{MAX} r_h + r_h^{MAX} \lambda_h - \lambda_h^{MAX} r_h^{MAX}\ \forall h \in H \\
& \quad w_h \geq 0\ \forall h \in H.
\end{aligned}
$$

The Benders subproblem is the restriction of (P) with the variables in the set $\Theta = \left\{\{y_j\}, \{x_h\}, \lambda^G, \{\lambda_h\}, \{s_{ij}^G\}\right\}$ fixed to the master problem solution. This is exactly the previously-introduced linear program (P_S^Θ) from §4.1.2. Its dual at each iteration t is the linear program (D_S^Θ) that optimizes over the variables $\left\{\{\alpha_j^t\}, \{\beta_i^t\}, \{\Delta_h^t\}, \{\gamma_h^t\}\right\}$ (derived in Online Appendix G). Define:

$$
B_t(\Theta) = \lambda^G + \sum_j \alpha_j^t a_j + \sum_i \beta_i^t b_i + \sum_h \Delta_h^t \lambda_h + \sum_h \max(\lambda_h, \gamma_h^t) x_h. \tag{21}
$$

Lemma 1 *At each iteration t, $z \leq B_t(\Theta)$ is a Benders optimality cut.*

To implement the $\max(\lambda_h, \gamma_h^t)x_h$ expressions in the Benders cut $z \leq B_t(\Theta)$, we introduce binary variables. For details of this implementation and the proof of Lemma 1, see Online Appendix G.

4.2.3 Computational Results

To test the SQE and GBD-based algorithms described in §4.2.1 and §4.2.2, respectively, we ran various instances while varying the number of trauma centers k and the number of helicopters m. We conducted

all of our computational experiments on a Dell Precision T5500 workstation with Intel Xeon CPU X5680 @ 3.33GHz (6 cores) and 12.0 GB RAM, running Windows 7, 64 bit. We used CPLEX 12.3 and AMPL 2011.07.25 to solve the MIQP and MILP formulations of our problem instances. We have found that: (1) when using SQE, the (P_M^{SQE}) formulation outperforms the (P_M^{GBD}) formulation, and (2) SQE outperforms the GBD-based algorithm. Performance varied by instance (k, m), with the optimality gap of the SQE method after 18 hours being in the range of 1.64% to 9.34%. For example, for $(k, m) = (10, 15)$, we reached a gap of 7.61%, whereas for $(k, m) = (14, 25)$, we reached a gap of 2.8%. Our SQE method is substantially faster than the global solver BARON, which, at the 18-hour mark has a gap of 21.9% and 13.6% for $(k, m) = (10, 15)$ and $(14, 25)$, respectively. Further details of our computational experiments are in Online Appendix E.

Although striving for a low optimality gap and solution time is theoretically justified, when it comes to measuring the real-world performance of a particular solution, one of the best ways is to run the location solutions through a simulation model. For this reason, in §5 we use simulation to compare the performance of the solutions that we get using SQE with the solutions from two heuristics that we present next.

4.3 Benchmark Heuristics

We now introduce two benchmark heuristics that were inspired by existing methods in the literature. We call the first the "no-congestion" heuristic, and the second the "decoupled" heuristic.

The *no-congestion heuristic* is inspired by the model of Branas and ReVelle (2001). As described in §2, Branas and ReVelle (2001) also study a joint location problem of trauma centers and helicopters, but model the problem as a deterministic location problem without taking into account the random availability of helicopters. To isolate the effect of helicopter congestion, we construct a Mixed Integer Linear Program (MILP) relaxation from (P) by simply dropping the bilinear $\lambda_h r_h$ terms from its objective:

$$(P_{NC}) \qquad \max \ \lambda^G + \sum_{h \in H} \lambda_h$$
$$\text{s.t. (1)-(4), (6)-(13).}$$

Problem (P_{NC}) maximizes the number of patients served, assuming heliports do not get congested. The no-congestion heuristic is simpler to implement (in terms of computation) than our original problem (P) because the non-convex bilinear terms $\lambda_h r_h$ do not appear in the objective. As we shall see in §5.3, ignoring helicopter congestion has a detrimental impact on overall performance. (Note that (P_{NC}) is not exactly the same as the formulation of Branas and ReVelle (2001) because, most notably, (P_{NC}) additionally models the capacity constraints of trauma centers.)

In contrast, the *decoupled heuristic* models congestion, but locates trauma centers and helicopters in sequence, allowing us to test how important it is for trauma centers and helicopters to be located *simultaneously*. This heuristic has practical relevance, since one of the planning methods the KMHW considered was to determine the locations of trauma centers first, while postponing the decision of where to locate helicopters. Recall from our discussion in §1 that both of the single-resource allocation problems

(i.e., either trauma centers or helicopters) belong to well-known classes of optimization problems: trauma centers can be located using a capacitated version of the maximal covering location problem (e.g., Pirkul and Schilling 1991), and helicopters can be located using a probabilistic ambulance location formulation (e.g., Daskin 1983). The latter method estimates a helicopter busy fraction, which can only be computed after the trauma center locations are fixed. Thus, our decoupled heuristic solves the problem sequentially as follows. First, we solve a capacitated variant of the maximal covering location problem, assuming that a helicopter is stationed at every heliport. The solution from this problem yields the trauma center locations as well as an allocation of patient demands to trauma centers. Second, after we estimate a helicopter busy fraction from the solution of the first step, we solve a variant of the maximum expected covering location problem (Daskin 1983) to establish the heliport locations. Further details of the decoupled heuristic are presented in Online Appendix H.

5 Application

In this section, we apply our model and solution methods to the design of a nationwide trauma care system in Korea. In §5.1, we briefly describe the data used for our analysis. Then, in §5.2, we present a trace-based simulation model, which takes the location solution of an optimization model as input, and simulates the arrival and service processes of trauma patients. Finally, in §5.3, we present the location solutions from the different approaches described in §4 on the map of Korea, and compare their performance, as measured by our simulation.

5.1 Data

We use the following two data sets that were produced as part of the broader study (Kim et al. 2011) commissioned by the KMHW: (1) demand-side data: one year's worth of nationwide trauma patient calls, including the times and locations of incidents; and (2) supply-side data: the number and location of candidate trauma centers and heliports.

[1] Demand-Side Data: Estimating the demand-side data involves many practical challenges such as fragmented data sources and a lack of clinical information to measure injury severity scores. Below we briefly describe the main data issues, and refer the reader to Kim et al. (2011) for further details.

Estimates of the total *annual number* of trauma patients come from two data sources: the National Emergency Department Information System (NEDIS), and the National Health Insurance (NHI). Injury-related Emergency Department (ED) visits were identified by their diagnosis code, yielding a total of 1,223,750 cases. Among these, only those cases with an Excess Mortality Ratio-Based Injury Severity Score (EMR-ISS) higher than 15 were classified as trauma cases, yielding a total of 190,193 trauma cases.

The NEDIS and NHI data sets, while useful for computing accurate total patient volumes, lack the fine granularity that we need to model patient arrivals. For the specific *locations and times* of trauma incidents, we consulted the nationwide data of emergency telephone calls. This data set includes field-triage records of which 80,300 were classified as trauma cases. Additionally, from the NEDIS data we identified 32,630 trauma patients who were self-transported or transferred from other local hospitals;

for these patients, the locations and times of their incidents were assumed to be those of their ED visits. For the remaining 77,263 ($= 190,193 - 80,300 - 32,630$) trauma patients, we assigned their locations and times by subsampling from the 112,930 ($= 80,300 + 32,630$) trauma patients, while taking care to match the regional demand rates in the original data.[11] See Figure 4(a) for the geographical distribution of trauma patients in Korea.

For our tests, we split the year into two halves: January-June (90,265 trauma cases) and July-December (99,928 trauma cases). This allowed us to test our methods both in-sample (e.g., by optimizing the trauma center and heliport locations using January-June data and then evaluating the performance of this solution with a simulation using January-June data) and out-of-sample (e.g., optimizing with January-June data and evaluating with July-December data). We report only the in-sample results from the January-June data set here, and include the out-of-sample results, which are qualitatively similar but serve as a robustness check, in Online Appendix I.

Finally, we aggregated patient demand by geographic area to keep the size of our optimization models manageable. After consultation with practitioners, we used a $25km \times 25km$ grid to subdivide Korea into 204 ($= \bar{\imath}$) demand regions. The optimization models use the aggregate demand rates λ_i for each region $i \in I = \{1, 2, ..., \bar{\imath}\}$, where all patient demands in a region are assumed to come from its center. As a precautionary measure, we also solved our optimization models with other grid sizes (e.g., $15km \times 15km$ and $20km \times 20km$) and found that solutions were similar – i.e., a reasonable amount of data aggregation seems to have only a marginal impact on solution quality.[12]

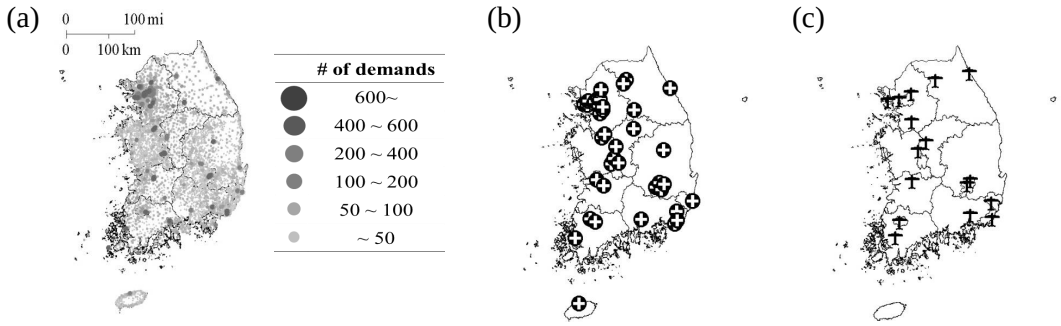


Figure 4: Geographical distribution of: (a) trauma cases in Korea, (b) candidate trauma centers, and (c) candidate heliports (excluding the heliports co-located at trauma centers)

[2] Supply-Side Data: The KMHW provided us with a list of 38 ($= \bar{j}$) candidate trauma center sites, all of which are existing hospitals that expressed interest in dedicating resources specifically for trauma patients. These hospitals are general hospitals with operating emergency departments, most of which

[11] Alternatively, we could have used the locations and times of the ED visits of these 77,263 patients. After much debate among medical and field professionals, Kim et al. (2011) concluded that the sampling approach described here captured reality better than this alterative approach.

[12] We have also conducted simulation experiments with both aggregated and unaggregated data, and tested solutions by increasing or decreasing regional demands by x% (where $x = \pm 5$, ± 10, and ± 15). Our results are robust against these changes.

are associated with medical schools. Figure 4(b) shows the locations of these candidate trauma centers. To limit congestion at trauma centers, the effective capacity of each center, c_j, is set to 50 patients per day. More details on this derivation can be found in Online Appendix C.

For candidate heliports, we use 16 that the Korean National Emergency Management Agency (NEMA) is currently operating; see Figure 4(c) for their locations. Recall that helicopters can also be stationed at open trauma centers. So, the total number of candidate heliport locations, $\bar{h}$, is 54 ($= 38 + 16$). Currently, the NEMA operates helicopters for fire-fighting and rescue as well as for EMS missions, so its helicopters are not specifically designed for transporting trauma patients. Thus, the KMHW was interested in exploring the optimal deployment of new EMS-dedicated helicopters. For this purpose, we vary the total number of available helicopters, m, in our study.

5.2 Simulation Model

To benchmark the performance of the location solutions generated by our integrated (SQE) method from §4.1 and §4.2 with those generated by the heuristics from §4.3, we use a trace-based simulation. Taking the locations of trauma centers and helicopters as given, we sequentially process each patient call from the historical data set. We simulate the real-time processes of serving each patient according to the flowchart shown in Figure 5, and categorize each patient as successful or unsuccessful. Successful patients are those that get transported to an available (under-capacity) trauma center within 60 minutes. After processing all patient calls, we measure the proportion of the nation's trauma patients that are served successfully.

The following features of our simulation, which are abstracted away in our optimization model and heuristics, capture the real-time decision-making processes in practice:

• Helicopter Assignment: We keep track of each helicopter's availability at each point in time. While a helicopter is transporting a patient to a trauma center or returning to its home heliport, it is temporarily unavailable for serving another patient. If more than one patient must wait for a helicopter, patients get served according to the first-in-first-out rule.

• Helicopter Diversion: We monitor the availability of beds at each trauma center. If a trauma center does not have an available bed to admit a new patient, we divert helicopters to the nearest trauma center that has an available bed.[13]

• Multiple Resources: When multiple resources are available to serve a patient (e.g., multiple trauma centers that operate under their capacities, multiple helicopters/ambulances), we always transport a patient to the nearest available (under-capacity) trauma center in the fastest way possible.

For further details of our simulation, most notably the decision points marked A, B, C, and D in Figure 5, see Online Appendix J.

[13] We have obtained similar results when implementing alternative diversion policies based on the number of waiting patients or boarding patients.

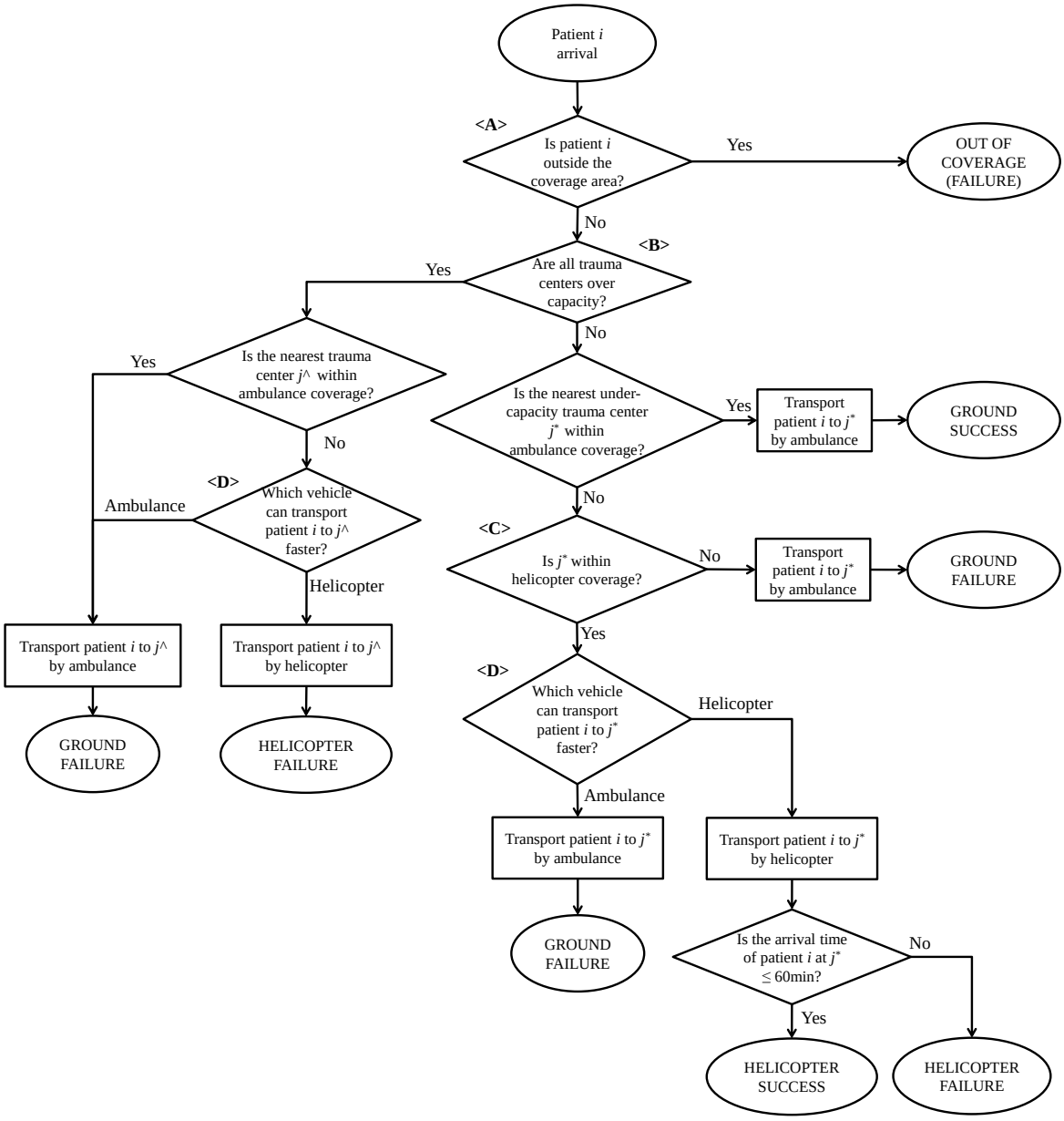


Figure 5: Simulation flowchart

5.3 Results

We tested the following three approaches: the integrated approach (presented in §4.1 and §4.2), and the no-congestion and decoupled heuristics (presented in §4.3). The KMHW was interested in exploring different numbers of trauma centers (k) and helicopters (m), since the budget for the trauma care system had not yet been determined. Below, we report test results for $k = 10$, 12, 14, and $m = 5$, 10, 15, 20, 25. Let TkHm denote the test case with k trauma centers and m helicopters.

First, we examine the location solutions of the trauma centers computed under the three different approaches. Figure 6 plots the test cases T10H5, T10H15, and T10H25 as we increase the number of helicopters from 5 to 25 holding the number of trauma centers fixed at 10. From this figure, we can make the following two important observations. First, the trauma center locations depend on the

number of available helicopters, implying that it is important to consider the locations of trauma centers in conjunction with helicopter transportation. Second, the trauma center locations differ significantly across the three approaches, especially when a large number of helicopters are used. For example, in the case of T10H25, both the no-congestion heuristic and the decoupled heuristic place five trauma centers at different locations, as compared to the integrated approach. The integrated approach tends to locate trauma centers in high demand regions (darker areas in the figure), whereas the two heuristics tend to spread trauma centers more broadly across the country. This is because the two heuristics ignore helicopter congestion when locating trauma centers, and simply maximize the total *geographic* coverage. This tends to result in trauma centers that are placed in areas that make heavy use of helicopters and light use of ambulances. On the other hand, the integrated approach overcomes this shortcoming by accounting for helicopter congestion when choosing trauma center sites.

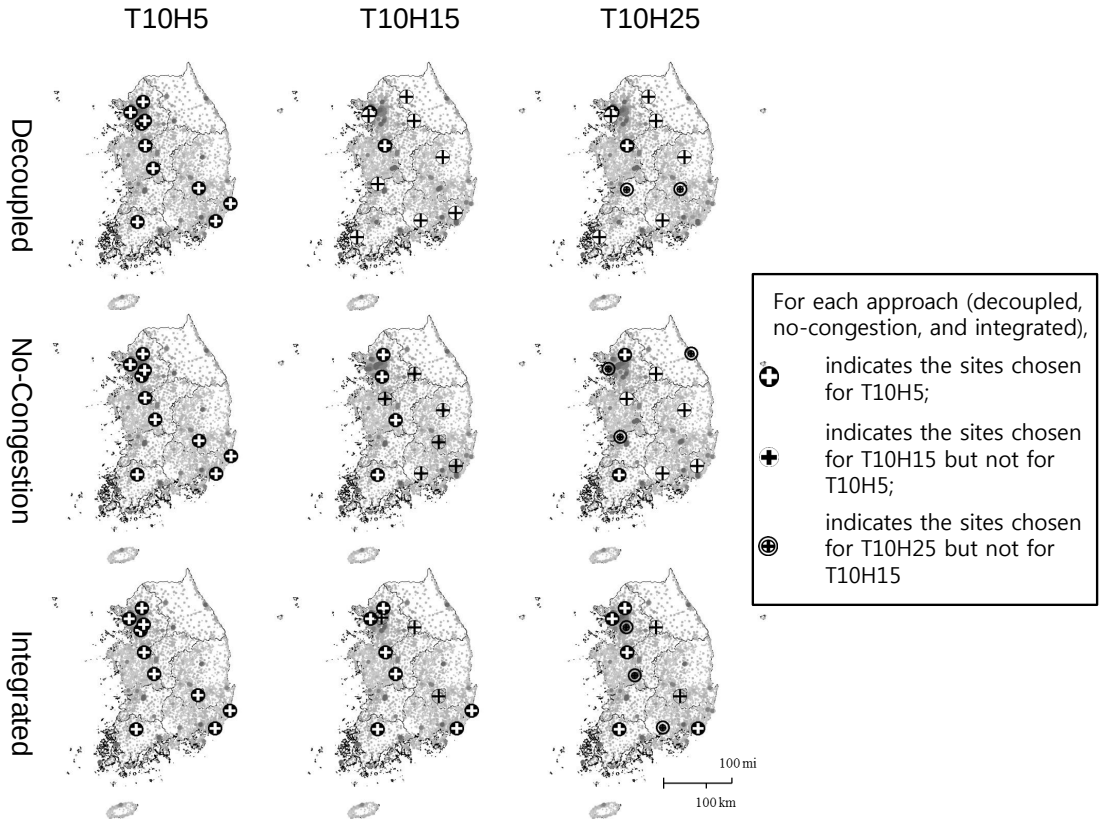


Figure 6: Trauma center locations determined from the three approaches for T10H5, T10H15 and T10H25 (Note: Helicopter locations are omitted to avoid clutter.)

Next, given the locations of trauma centers and helicopters determined from each of the three approaches, we examine the percentage of successful transports as measured by our simulation. As seen in Figure 7, when the number of helicopters is very low (e.g., TkH5), the performance is about the same across all three approaches. This is expected because with a small number of helicopters most patients are transported by ambulances, which are modeled roughly in the same manner in all three approaches. As the number of helicopters increases (i.e., m increases), we observe that *the integrated*

approach outperforms the two heuristics significantly. Moreover, as more helicopters become available, the percentage of successfully transported patients increases under the integrated approach (which is intuitive), whereas it decreases in some cases under the two heuristics (which is counter-intuitive).[14]

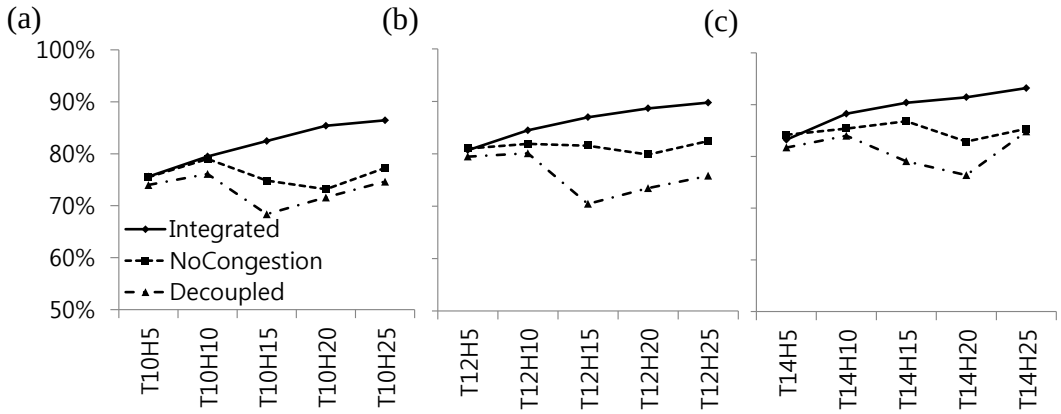


Figure 7: The proportion of successful patients when: (a) $k = 10$, (b) $k = 12$, and (c) $k = 14$.

To explain these intuitive and counter-intuitive results, we examine the role of helicopter transportation and how helicopter congestion affects the performance of the trauma care system. Figure 8 shows the total number of patients transported by helicopters when different numbers of helicopters are available. When there are enough helicopters (e.g. $m = 10,\ 15,\ 20$ or 25), the two heuristics assign significantly more patients to helicopters than the integrated approach, and a larger number of these patients under the heuristic solutions fail to get to a trauma center within 60 minutes. The reason for the larger number of failures in the heuristic cases is due to the delay caused by helicopter congestion, as is evident from Figure 9: A higher proportion of the patients transported by helicopter experience delay under the heuristic solutions than under the integrated solutions. In addition, patients experience fewer delays as more helicopters become available under the integrated approach, whereas using more helicopters does not always lead to fewer delays under the heuristics. While Figures 8 and 9 show the results for the 10 trauma center test cases (i.e., $k = 10$), we also observe the same pattern with different numbers of trauma centers.

We explain the underlying reasons for the results presented in Figures 7, 8 and 9 as follows. When the number of helicopters is very low ($m = 5$), the location solutions from the three approaches do not differ much (as shown in Figure 6), hence we see similar performance. As the number of helicopters increases (e.g., $m = 10$ or 15), helicopter transport becomes more crucial. Compared to the two heuristics, the integrated approach makes more judicious use of helicopters (Figure 8), keeping helicopter utilization

[14]We have also evaluted the value of the objective given in equation (5) *analytically* by substituting the solutions from the integrated approach and the no-congestion heuristic into (5). We have observed the same general pattern as shown in Figure 7: as m increases, the objective value increases under the integrated approach, whereas the objective value decreases in some cases under the no-congestion heuristic.

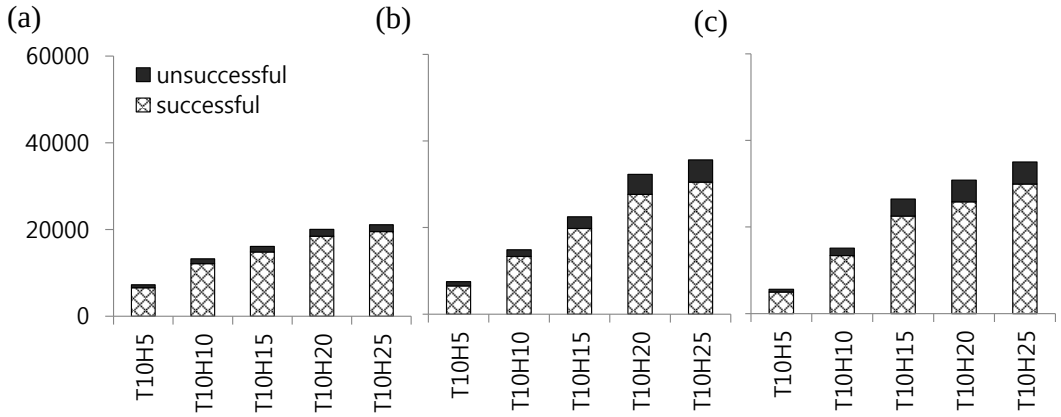


Figure 8: The total number of successful and unsuccessful helicopter transports under the solutions determined by: (a) integrated approach, (b) no-congestion heuristic, and (c) decoupled heuristic.

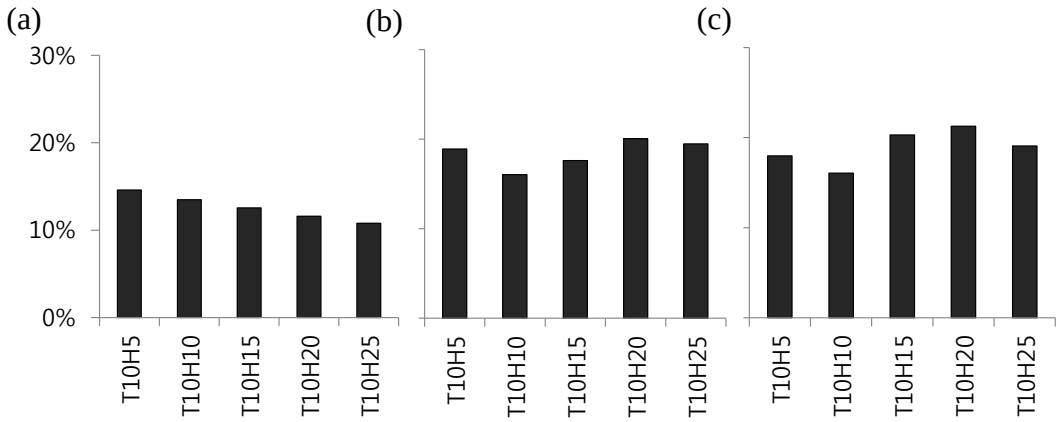


Figure 9: The proportion of helicopter transports that were delayed under the location solutions determined by: (a) integrated approach, (b) no-congestion heuristic, and (c) decoupled heuristic.

and the frequency of helicopter delays lower (Figure 9). As the number of helicopters increases further (e.g., $m = 25$), helicopters become a less-constrained resource: utilization drops even for the heuristics, although the integrated method still outperforms by a significant margin.

The simulation demonstrates the advantages of our integrated approach. First, by locating trauma centers and helicopters *simultaneously*, our integrated approach significantly outperforms the decoupled heuristic, which sites trauma centers and helicopters sequentially. Second, by taking helicopter *congestion* into account, our integrated approach significantly outperforms the no-congestion heuristic. In the case of T10H20, Figure 7 shows that we achieve a 12% (resp., 14%) larger number of successful transports using our integrated approach than the no-congestion (resp., decoupled) heuristic, which translates into a potential 23,000 (resp., 26,000) additional lives saved per year!

6 Concluding Remarks

This paper studies the problem of simultaneously locating trauma centers and helicopters. Our problem poses a unique challenge because the locations of trauma centers affect the demand for helicopters, and vice versa. To overcome this challenge, we endogenize the computation of a busy fraction within an

optimization problem, and develop the Shifting Quadratic Envelopes algorithm that outperforms a Benders-like cut generation scheme. We apply our method to the design of a nationwide trauma care system in Korea. By running a trace-based simulation on a full year of patient data, we find that the solutions generated by our model outperform several benchmark heuristics significantly.

Our main methodological contribution, the Shifting Quadratic Envelopes method, can be applied more generally to network design problems cast as mathematical programs where the probability of a server being busy is a function of the workload sent to that server, and both the arrival rate and workload are linear functions of the same set of variables (e.g., in our case, the routing variables s_{hij}). It would be interesting to know what classes of problems that fit this general description can be solved using the SQE method more efficiently than other solution methods. For example, the problems of locating two classes of fixed servers (c.f. Marianov and Serra 2001, Grmez et al. 2011) or two classes of mobile servers (c.f. Mandell 1998) share several features with our model. In these problems, demand for one type of server depends on the locations of a second type of server, in which case modeling congestion requires both types of servers to be simultaneously located, making our model relevant. For these problems, we expect SQE to be particularly helpful when service times are a function of where the demands originate. We leave a more exhaustive computational study of the SQE method, including a comparison of different heuristics for slice selection and subdivision, to future work.

To date, the Korean government has taken several steps toward implementing a nationwide trauma care system. The earlier, broader study by Kim et al. (2011) was used in the government's initial feasibility study, and the KMHW has continued to refine its plan, taking into account recommendations we prepared in an earlier version of this paper. The first trauma center in Korea is scheduled to open in spring 2014. In addition, the KMHW designated five trauma centers in fall 2012 and another four centers in 2013, and it plans to establish six more trauma centers in the near future. We are pleased to report that the ten hospitals selected by the KMHW to become trauma centers are generally consistent with our recommended solutions, with minor variations owing to qualitative factors such as hospital reputation. Moreover, with our completed paper in hand, the KMHW now has more accurate sensitivity analyses for selecting additional trauma centers and deploying helicopters moving forward. In addition to the current project of designing a trauma care system, the KMHW has been working to create or redesign many parts of the country's healthcare service infrastructure such as restructuring the current tiered EMS system and deploying two different types of ambulances (advanced and basic life support ambulances), and helicopters (specialized air ambulances and multi-purpose helicopters). Our methods would likely benefit these other projects, since they also involve integrated planning of fixed and mobile servers where congestion is a concern, and busy fractions cannot be reasonably estimated a priori.

References

Aboolian, R., O. Berman, Z. Drezner. 2008. Location and Allocation of Service Units on a Congested Network. *IIE Transactions* **40**(4), 422-433.

Ball, M.O., L.F. Lin. 1993. A Reliability Model Applied to Emergency Service Vehicle Location. *Operations Research* **41**(1) 18-36.

Benders, J.F. 1962. Partitioning Procedures for Solving Mixed-Variables Programming Problems. *Numerische Mathematik* **4**(1) 238-252.

Berman, O., D. Krass. 2002. Facility Location Problems with Stochastic Demands and Congestion. Z. Drezner and H. W. Hamacher, eds. *Facility Location: Applications and Theory*, Springer, Berlin.

Berman, O., S. Vasudeva. 2005. Approximating Performance Measures for Public Services. *IEEE Transactions on Systems, Man, and Cybernetics* **35**(4) 583-591.

Borras, F., J.T. Pastor. 2002. The Ex-Post Evaluation of the Minimum Local Reliability Level: An Enhanced Probabilistic Location Set Covering Model. *Annals of Operations Research* **111** 51-74.

Branas, C.C., C. ReVelle. 2001. An Iterative Switching Heuristics to Locate Hospitals and Helicopters. *Socio-Economic Planning Sciences* **35**(1) 11-30.

Brotcorne, L., G. Laporte, and F. Semet. 2003. Ambulance Location and Relocation Models. *European Journal of Operational Research* **147**(3) 451-463.

Burwell, T.H., M.A. McKnew, J.P. Jarvis. 1992. An Application of a Spatially Distributed Queueing Model to an Ambulance System. *Socio-Economic Planning Sciences* **26** 289-300.

Centers for Disease Control and Prevention. 2013. http://www.cdc.gov/TraumaCare/. Last accessed on September 9, 2013.

Chapman, S., J. White. 1974. Probabilistic Formulations of Emergency Service Facilities Location Problems. ORSA/TIMS paper, San Juan, Puerto Rico.

Church, R., C. ReVelle. 1974. The Maximal Covering Location Problem. *Papers in Regional Science* **32**(1) 101-118.

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

Daskin, M.S., L. K. Dean. 2004. *Location of Health Care Facilities*, Chapter 3 in the Handbook of OR/MS in Health Care: A Handbook of Methods and Applications, F. Sainfort, M. Brandeau and W. Pierskalla, editors, Kluwer.

Erkut, E., A. Ingolfsoon, T. Sim, and G. Erdogan. 2009. Computational Comparison of Five Maximal Covering Models for Locating Ambulances. *Geographical Analysis* **41**(1) 43-65.

Floudas, C.A., P.M. Pardalos. 2012. *State of the Art in Global Optimization: Nonconvex Optimization and Its Applications*, Springer.

Gendreau, M., G. Larporte, and F. Semet. 1997. Solving an Ambulance Location Model by Tabu Search. *Location Science* **5**(2) 75-88.

Geoffrion, A.M. 1972. Generalized Benders Decomposition. *Journal of Optimization Theory and Applications* **10**(4) 237-260.

Geromel, J.C., M.R. Belloni. 1986. Nonlinear Programs with Complicating Variables: Theoretical Analysis and Numerical Experience. *IEEE Transactions on Systems, Man and Cybernetics* **16**(2) 231-239.

Goldberg, J., R. Dietrich, J. Chen, G. Mitwasi, T. Valenzuela, L. Criss. 1990. A Simulation Model for Evaluating a Set of Emergency Vehicle Base Locations: Development, Validation, and Usage. *Socio-Economic Planning Sciences* **24** 125-141.

Görmez, N., M. Kksalan, F.S. Salman. 2011. Locating Disaster Response Facilities in Istanbul. *Journal of the Operational Research Society* **62** 1239-1252.

Hogan, K., C. ReVelle. 1986. Concepts and Applications of Backup Coverage. *Management Science* **32**(11) 1434-1444.

Kim, Y., D.W. Kang, Y.S. Rho, S.B. Park, J.H. Park, C.B. Park, S.D. Shin, K.O. Ahn, S.R. Yeom, J.H. Oh, E.J. Lee, T, Lee, J.S. Cho, and W.C. Cha. 2011. *Study on an Implementation Plan for Trauma Care System Element.* Report to the Korean Ministry of Health and Welfare. Seoul National University.

Larson, R.C. 1975. Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23**(5) 845-868.

Li, X., Z. Zhao, X. Zhu, and T. Wyatt. 2011. Covering Models and Optimization Techniques for Emergency Response Facility Location and Planning: a Review. *Mathematical Methods of Operations Research* **74**(3) 281-310.

Mandell, M. 1998. Covering Models for Two-tiered Emergency Medical Services Systems. *Location Science* **6** 355-368.

Marianov, V., C. ReVelle. 1994. The Queueing Probabilistic Location Set Covering Problem and Some Extensions. *Socio-Economic Planning Sciences* **28**(3) 167-178.

Marianov, V., C. ReVelle. 1996. The Queueing Maximal Availability Location Problem: a Model for the Siting of Emergency Vehicles. *European Journal of Operations Research* **93** 110-120.

Marianov, V., D. Serra. 1998. Probabilistic Maximal Covering Location-Allocation for Congested System. *Journal of Regional Science* **38**(3) 401-424.

Marianov, V., D. Serra. 2001. Hierarchical Location-Allocation Models for Congested Systems. *European Journal of Operational Research* **135** 195-208.

McCormick, G.P. 1976. Computability of Global Solutions to Factorable Nonconvex Programs: Part I – Convex Underestimating Problems. *Math Programming* **10** 147-175.

Nagy, G., S. Salhi. 2007. Location-Routing: Issues, Models and Methods. *European Journal of Operational Research* **177** 649-672.

Owen, S.H., M.S. Daskin. 1998. Strategic Facility Location: A Review. *European Journal of Operational Research* **111**(3) 423-447.

Peleg, K., L. Aharonson-Daniel, M. Stein, Y. Kluger, M. Michaelson, A. Rivkind, V. Boyko, the Israel Trauma Group. 2004. Increased Survival among Severe Trauma Patients. *ARCH SURG* 139, November 1231-1236.

Pirkul, H., D.A. Schilling. 1991. The Maximal Covering Location Problem with Capacities on Total Workload. *Management Science* **37**(2) 233-248.

Prodhon, C., C. Prins. 2014. A Survey of Recent Research on Location-Routing Problems. *European Journal of Operational Research*, Forthcoming.

Repede, J. J. Bernardo. 1994. Developing and Validating a Decision Support System for Locating Emergency Medical Vehicles in Louisville, Kentucky. *European Journal of Operational Research* **75** 567-581.

ReVelle, C., H.A. Eiselt. 2005. Location Analysis: A Synthesis and Survey. *European Journal of Operational Research* **165**(1) 1-19.

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

Sahinidis, N.V., I.E. Grossman. 1991. Convergence Properties of Generalized Benders Decomposition. *Computers & Chemical Engineering* **15**(7) 481-491.

Schilling, D.A., D.J. Elzinga, J. Cohon, R. Church, and C. ReVelle. 1979. The Team/Fleet Models for Simultaneous Facility and Equipment Siting. *Transportation Science* **13**(2) 163-175.

Sorensen, P., R. Church. 2010. Integrating Expected Coverage and Local Reliability for Emergency Medical Services Location Problems. *Socio-Economic Planning Sciences* **44** 8-18.

Tawarmalani, M. and N. V. Sahinidis. 2005. A Polyhedral Branch-and-Cut Approach to Global Optimization. *Mathematical Programming* **103**(2) 225-249.

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

Zhang, Y., O. Berman, P. Marcotte, V. Verter. 2010. A Bilevel Model for Preventive Healthcare Facility Network Design with Congestion. *IIE Transactions* **42**:12, 865-880.

Online Appendix

A. Multiple Helicopters per Heliport

The Shifting Quadratic Envelopes (SQE) method can be extended to handle multiple helicopters per heliport. This section illustrates the theory, and provides some computational results. The multi-helicopter (per heliport) instances are more difficult to solve computationally than single-helicopter instances. As a result, we were not able to solve multi-helicopter instances on the scale required in our problem setting. However, the multi-helicopter SQE extension is theoretically interesting in its own right, and may be useful for smaller instances (e.g., trauma center planning at the scale of a metropolitan area, for example).

Recall that, in the single-helicopter case, we used the variable r_h to mean either (i) the probability a patient routed via heliport h experiences transportation delay, (ii) the utilization of heliport h, or (iii) the workload assigned to heliport h (measured in Erlangs). Throughout the paper, we have used all three interpretations, depending on which definition was more convenient for the context at hand. In the multi-helicopter case, however, we need to be more careful, since the expressions for (i), (ii), and (iii) are no longer equal. We begin this section by defining analytical formulas for (i), (ii), and (iii), that apply more broadly to the multi-helicopter case. Then, we show how our model and the SQE method can be generalized to the multi-helicopter case.

We begin by modeling heliport h as an M/M/k queue[15], where each server is a helicopter, we have exactly k helicopters stationed at heliport h, and the time a helicopter is in service is the time it takes to fly the circuit $h \to i \to j \to h$, plus loading, unloading, and cleanup times. Therefore, the workload assigned to heliport h remains defined as $r_h = \sum_{i,j} \tau_{hij} s_{hij}$, whereas utilization is now r_h/k, since the workload is now distributed across k helicopters. Finally, the probability that a patient routed via heliport h experiences transportation delay is $f_k(r_h)$, where

$$f_k(r) = \frac{r^k}{(k-1)!(k-r)} \times \left[\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{(k-1)!(k-r)} \right]^{-1} \tag{22}$$

is known as the Erlang-C formula. As a check, notice that if $k = 1$, both utilization r_h/k and the delay probability $f_k(r_h)$ simplify to r_h, which is what we expect for the M/M/1 case.[16] We now state three useful lemmas.

Lemma 2 *The Erlang-C function $f_k(r)$ is a non-negative, non-decreasing, convex function in r.*

Proof. See Lee, H. L. and M. A. Cohen 1983. A note on the convexity of performance measures of M/M/c queueing systems. *Journal of Applied Probability.* 20(4) 920–923. ■

Lemma 3 *Let $a(x)$ and $b(x)$ be two non-negative, non-decreasing convex functions. Then, their product $c(x) = a(x)b(x)$ is also non-negative, non-decreasing, and convex.*

[15] Although the M/G/k queue would technically be a more accurate model, the M/M/k queue has a closed-form analytical expression for the delay probability (the well-known Erlang-C formula), whereas the M/G/k queue does not. Moreover, it is well-known that the Erlang-C formula provides a good approximation to the delay probability of an M/G/k queue; see, for example: Kimura, T., 2010. The M/G/s Queue. *Wiley Encyclopedia of Operations Research and Management Science.* Cochran, J. et al., Eds. John Wiley & Sons, Inc.

[16] Writing $f_k(r) = \frac{r^k \pi_0}{(k-1)!(k-r)}$, where $\pi_0 = \left[\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{(k-1)!(k-r)} \right]^{-1}$, we substitute $k = 1$ to get $\pi_0 = \left[\frac{r^0}{0!} + \frac{r^1}{(1-1)!(1-r)} \right]^{-1} = \left[1 + \frac{r}{(1-r)} \right]^{-1} = \left[\frac{1}{1-r} \right]^{-1} = 1 - r$; and $f_k(r) = \frac{r^1 \pi_0}{(1-1)!(1-r)} = \frac{r}{1-r} \cdot \pi_0 = \frac{r}{1-r} \cdot (1-r) = r$.

Proof. Differentiating twice, we see that $c'(x) = a'(x)b(x) + a(x)b'(x)$ and $c''(x) = a''(x)b(x) + 2a'(x)b'(x) + a(x)b''(x)$. Since a and b are convex, we have $a''(x) \geq 0$ and $b''(x) \geq 0$. Since a and b are non-decreasing, we have $a'(x) \geq 0$ and $b'(x) \geq 0$. Finally, since a and b are non-negative, we have $a(x) \geq 0$ and $b(x) \geq 0$. Therefore, both terms of $c'(x)$ are non-negative, which implies c is non-decreasing. Moreover, all three terms of $c''(x)$ are non-negative, which implies c is convex. ■

Lemma 4 *The expression $\lambda_h f_k(\tau_h^{FIX}\lambda_h)$ is convex in λ_h on the domain $\lambda_h \geq 0$, for any non-negative constant τ_h^{FIX}.*

Proof. Since both λ_h and $f_k(\tau_h^{FIX}\lambda_h)$ are non-negative, non-decreasing, convex functions of λ_h on the domain $\lambda_h \geq 0$, the result follows by Lemma 3. ■

In the multi-helicopter case, we would like the number of helicopters at heliport h to be modeled by the variable $x_h \in \{0, 1, \ldots, \overline{k}\}$, where $\overline{k}$ is the maximum number of helicopters per heliport. However, it is instructive to first describe the special case where we fix the number of helicopters at heliport h to the constant k_h. For this special case, our objective, which maximizes the number of patients that don't experience a transportation delay, is:

$$\max \lambda^G + \sum_{h \in H} \lambda_h (1 - f_{k_h}(r_h)),$$

which is the natural generalization of our previous objective (5). The nonlinear terms $\lambda_h f_{k_h}(r_h)$ present a computational challenge in much the same way that the bilinear terms $\lambda_h r_h$ did in our original model. We will first describe how a generalization to our SQE method can tackle these nonlinear terms. Afterward, we will return to the issue of generalizing our results to the case where the number of helicopters at each heliport is variable.

A.1 Description of Multi-Helicopter SQE Method

Let us now describe how SQE works in the multi-helicopter case with a fixed number of k helicopters at heliport h. Generally speaking, all that is required is to replace the (convex) quadratic underestimate $w_h(\lambda_h|\tau_h^{FIX}) = \tau_h^{FIX}\lambda_h^2$ with the more general convex (non-quadratic) underestimate $w_h(\lambda_h|\tau_h^{FIX}) = \lambda_h f_k(\tau_h^{FIX}\lambda_h)$. As a check, we note that for $k = 1$, we recover $\lambda_h f_1(\tau_h^{FIX}\lambda_h) = \tau_h^{FIX}\lambda_h^2$, as before. The main parts of the SQE algorithm described in Section 4.2.1 remain unchanged, except for the second paragraph of Section 4.2.1, which we generalize as follows (the complete paragraph is copied, with changed parts in bold):

> From (P_S^Θ), we have a feasible solution to (P), which we denote $\Psi = \Theta \cup \{\{r_h\}, \{s_{hij}\}\} = \Big\{\{y_j\}, \{x_h\}, \lambda^G, \{\lambda_h\}, \{\lambda_j\}, \{s_{ij}^G\}, \{r_h\}, \{s_{hij}\}\Big\}$. The two crucial variables that describe the performance of heliport h are its allocated demand λ_h and its workload r_h. Therefore, for each heliport h, we use the point $(\lambda_h, r_h) \in \mathbb{R}^2$ from the feasible solution Ψ to update the quadratic envelope boundaries for heliport h. Let $\tau_h = r_h/\lambda_h$ be the mean service time of candidate solution (λ_h, r_h) and $\tau_h^* = r_h^*/\lambda_h^*$ be the mean service time of the optimal solution (λ_h^*, r_h^*) of (P). For each heliport h, our goal is to pick a mean service time τ_h^{FIX} that underestimates the optimal mean service time τ_h^* as closely as possible. This will allow us to closely underestimate the **nonlinear** term $w_h = \boldsymbol{\lambda_h f_k(r_h)}$ using a **convex** function of λ_h. Define two functions of λ_h; namely, $r_h(\lambda_h|\tau_h^{FIX}) = \tau_h^{FIX}\lambda_h$ and

$w_h(\lambda_h|\tau_h^{FIX}) = \boldsymbol{\lambda_h f_k(r_h(\lambda_h|\tau_h^{FIX})) = \lambda_h f_k(\tau_h^{FIX}\lambda_h)}$. As long as $\tau_h \geq \tau_h^{FIX}$, it follows by definition that $r_h(\lambda_h|\tau_h^{FIX}) \leq r_h$ and $w_h(\lambda_h|\tau_h^{FIX}) \leq w_h$. That is, for points (λ_h, r_h) where $r_h \geq r_h(\lambda_h|\tau_h^{FIX}) = \lambda_h\tau_h^{FIX}$, the **convex** function $w_h(\lambda_h|\tau_h^{FIX}) = \boldsymbol{\lambda_h f_k(\tau_h^{FIX}\lambda_h)}$ provides a valid underestimate for w_h. Since, in general, we know only that $\tau_h^* \in [\tau_h^{MIN}, \tau_h^{MAX}]$ before solving our problem (P), the best choice for τ_h^{FIX} that is always guaranteed to underestimate the optimal mean service time is τ_h^{MIN}. This is why, in §4.1.1, we used the underestimate $\tau_h^{FIX} = \tau_h^{MIN}$ to construct (P_M^{SQE}) by replacing the **nonlinear** terms $w_h = \boldsymbol{\lambda_h f_k(r_h)}$ in the objective of (P) with their **convex** relaxations $w_h(\lambda_h|\tau_h^{MIN}) = \boldsymbol{\lambda_h f_k(\tau_h^{FIX}\lambda_h)}$. We will now describe how we can use other estimates for τ_h^{FIX} that lead to tighter relaxations.

A.2 Variable Number of Helicopters per Heliport

Of course, we do not know a priori how many helicopters to assign to each heliport. To address this problem, we can make the following changes to our model. First, we introduce binary variables $x_{hk} = 1$ when k helicopters are stationed at heliport h, and $x_{hk} = 0$ otherwise. Next, we add constraints of the form $\sum_{k=0..n} x_{hk} = 1$ for all heliports $h \in H$. Finally, we introduce continuous variables λ_{hk}, where we force $\lambda_{hk} = \lambda_h$ when $x_h = k$, and $\lambda_{hk} = 0$ otherwise. This is accomplished using constraints $\lambda_h = \sum_{k=1..n} \lambda_{hk}\ \forall h \in H$ in conjunction with Big-M constriants $\lambda_{hk} \leq \lambda_h^{MAX} x_{hk}$. This allows us to write our objective function as:

$$\max \lambda^G + \sum_{h\in H, k=1..n} \lambda_{hk}(1 - f_k(r_h)).$$

The chief computational difficulty is dealing with the nonlinear products of the form $\lambda_{hk} f_k(r_h)$, which we have already seen can be addressed using an extension of our Shifting Quadratic Envelopes (SQE) approach (simply substitute λ_{hk} for λ_h in the preceding analysis).

A.3 Details of Multi-Helicopter SQE

We now describe the necessary changes to the master problem (P_{M2}^{SQE}) and subproblem (P_S^{Θ}). We will refer to the multiple-helicopter versions of these problems as $(P_{M2-MULT}^{SQE})$ and (P_{S-MULT}^{Θ}), respectively. In the master problem, we replace (11) with its generalization

$$x_j \leq \overline{k} y_j\ \forall j \in J, \tag{23}$$

where $\overline{k}$ is the maximum number of helicopters allowed at a single heliport. Let $K_0 = \{0, 1, 2, \ldots, \overline{k}\}$ be a set that lists the possible number of helicopters that can be stationed at a heliport, and $K = \{1, 2, \ldots, \overline{k}\}$ be the same set without zero. We also replace (13) with

$$y_j \in \{0, 1\}\ \forall j \in J; \qquad x_h \in K_0\ \forall h \in H, \tag{24}$$

since now x_h, the number of helicopters at heliport h, can be more than 1. The binary variable x_{hn} is replaced by the binary variable x_{hkn}, which we set to 1 if the solution (λ_h, r_h) is in slice n *and* there are k helicopters placed at heliport h, or 0 otherwise. Analogously, λ_{hn} and r_{hn} are replaced by λ_{hkn} and r_{hkn}, and we allow the bound

r_{hk}^{MAX} to depend on k (a reasonable value for this bound is $r_{hk}^{MAX} = k$). Thus, the master problem is now:

$$
\begin{aligned}
(P_{M2-MULT}^{SQE}) \quad & \max \lambda^G + \sum_{h \in H} \lambda_h - \sum_{h \in H, k \in K, n \in N_h} \lambda_{hkn} f_k(\tau_{hn} \lambda_{hkn}) \\
& \text{s.t. (1)-(4), (6)-(10), (12), (23), (24)} \\
& \lambda_h = \sum_{k \in K, n \in N_h} \lambda_{hkn}, \quad r_h = \sum_{k \in K, n \in N_h} r_{hkn} && \forall h \in H \\
& x_h = \sum_{k \in K, n \in N_h} k x_{hkn} && \forall h \in H \\
& 0 \le \lambda_{hkn} \le \lambda_h^{MAX} x_{hkn}, \quad 0 \le r_{hkn} \le r_{hk}^{MAX} x_{hkn} && \forall h \in H,\ \forall k \in K,\ \forall n \in N_h \\
& \tau_{h,n} \lambda_{hkn} \le r_{hkn} \le \tau_{h,n+1} \lambda_{hkn} && \forall h \in H,\ \forall k \in K,\ \forall n \in N_h \\
& \sum_{k \in K_0, n \in N_h} x_{hkn} = 1 && \forall h \in H \\
& x_{hkn} \in \{0, 1\} && \forall h \in H,\ \forall k \in K_0,\ \forall n \in N_h.
\end{aligned}
$$

The subproblem (P_{S-MULT}^{Θ}) is nearly identical to the single-helicopter version (P_S^{Θ}). The constraint set is unchanged, and the objective becomes $\min \sum_{h \in H} \lambda_h f_k(r_h)$, where λ_h is a fixed parameter, $k = x_h$ is fixed, and $f_0(r_h) = 0$ is defined for completeness[17].

At each iteration of the SQE algorithm, we proceed exactly as in the single-helicopter case: (1) solve the master problem to get a "good" solution to (P); (2) solve the subproblem to improve this solution; (3) evaluate the quality of the improved solution by computing its optimality gap; and (4) terminate if a time limit or optimality gap threshold is reached, otherwise shift the quadratic envelope boundaries (in exactly the same way as in the single-helicopter case) and continue.

A.4 Computational Results

We solved some small multi-helicopter instances, using KNITRO rather than CPLEX to solve both the master problem and subproblem.[18] The first instance we tested is from the southwest corner of Korea and has 38 demand regions, 10 candidate trauma centers, and 4 candidate heliports. Allowing for up to 3 helicopters per heliport, and assuming we can open up to 4 trauma centers and use up to 8 helicopters, we ran the multi-helicopter version of SQE and found a near-optimal solution very quickly: After 1 iteration (3 seconds), we had a solution within 1.64% of optimality; after 3 iterations (37 seconds) we had a solution within 0.54% of optimality; after 6 iterations (2 minutes) we got to within 0.27%; and after 10 iterations (5 minutes) we had reduced the gap to 0.07%. The second instance we tested is from the northwest corner of Korea, which includes Seoul and is more populous. In this instance there are 52 demand regions, 25 candidate trauma centers, and 9 candidate heliports. With 6 trauma centers and 8 helicopters, and allowing up to 3 helicopters per heliport, we get within 2.1% from optimality after 3 hours, and with 8 trauma centers and 10 helicopters, we get within 0.7% of optimality after 3

[17] When $x_h = 0$, then we must have $\lambda_h = 0$, so the whole term $\lambda_h f_k(r_h)$ reduces to 0 regardless of how f_0 is defined.

[18] The master problem is now a convex MINLP (Mixed Integer Nonlinear Program), and the subproblem is a convex NLP (Nonlinear Program). KNITRO can solve both convex MINLP's and convex NLP's, whereas the only nonlinear functions that CPLEX can handle are quadratic. Thus, while in the single-helicopter case, we can use CPLEX to solve both the master problem MIQP and subproblem LP, we now need to use KNITRO, which supports a broader problem class.

hours, with additional time necessary to further close the gap. Some heliports were assigned 2 or 3 helicopters, which we expect since there are more helicopters than heliports. We also attempted to run the multi-helicopter SQE method on our full Korea-wide instance, but this proved to be too large for KNITRO to find good solutions in a reasonable amount of time. Thus, we conclude that the single-helicopter SQE is better suited for nation-scale trauma center design problems, while multi-helicopter SQE is viable for planning smaller, e.g. metropolitan, areas.

B. Robust Model

The Shifting Quadratic Envelopes (SQE) method that we introduce in this paper is quite general, and can be applied more broadly to other, more complex, variants of our trauma care system design problem. For instance, we may be interested in knowing how to locate trauma centers and helicopters when, over some time horizon, one or more of the following is expected to change in a predictable way: (i) the demand for care in each region, (ii) the effective capacity of each trauma center, (iii) the number of helicopters, or (iv) the number of trauma centers. The problem in this case would be to produce an optimal rollout plan, which would decide where both trauma centers and helicopters would be placed at each stage in a multi-period horizon. Generally speaking, multi-period location problems are known as *dynamic facility location problems*, and many specialized algorithms have been developed to tackle different such variants[19].

It is up to the modeler to decide what model variant is most appropriate for a given setting. In the main body of our paper, our focus was on the single-period problem, which is appropriate if either (i) problem parameters such as demands, capacities, and the number of helicopters and trauma centers do not appreciably change over time; (ii) some problem parameters may change, but not in a predictable way (when they change, we re-solve the single-period model); or (iii) we are most interested in knowing what the best location solution is given the current parameter values, without regard to possible future values. In practice, single-period models are often periodically run to check how far from optimal the current system is. The magnitude of the optimality gap allows policymakers to assess how important re-configuring the system is, and differences between the optimal solution and the current system suggest how the current system should be modified. In our case, the KMHW is planning substantial changes to Korea's trauma care system which involve many logistical details beyond the strategic decision of where to locate trauma centers and helicopters. A single-period model (i) allowed the KMHW to focus on improving the current state of care versus what previously existed, and (ii) was reasonable, given that some of the parameters (e.g., trauma center capacities) may change in the future, but in ways that cannot be accurately predicted at present.

On the other hand, when it is known how demand and resource availability will change over time, a multi-period model may be more appropriate. However, multi-period models make more assumptions about logistical details, and as a result their solutions specify not only where to locate trauma centers and heliports, but when to open and close them over time. Constraints in such models can be used to restrict how the solution changes over time; e.g., since there are significant fixed costs for establishing a trauma center, it makes sense to assume that once a trauma center is designated, it remains open for an extended period of time. Helicopters, on the other hand, may be more easily re-located, especially if the heliports themselves are already built. We view multi-

[19]For a comprehensive review, see chapter 15 of Farahani, R. Z., and M. Hekmatfar (Eds.) 2009. *Facility location: concepts, models, algorithms and case studies.* Springer-Verlag, Berlin.

period models as being complementary to single-period models, and recommend that they be used to supplement single-period strategic analysis when additional logistical details are available. For example, if the population in certain areas is growing or aging, we can predict the impact on future trauma rates. Governments may also have visibility into future resource availability, e.g., if multi-year budgets have been approved to purchase 2 new helicopters per year for the next 5 years, then it may be useful to run a multi-period model as a robustness check to confirm that the single-period model does not select trauma centers that are in completely different locations than will be needed in 5-years' time.

For an example of a multi-period model, assume the KMHW wants to designate all k trauma centers at the beginning of the horizon, and initially procure m_1 helicopters. Then, over the time periods $t = 1..T$, the KMHW could add additional helicopters so that m_t, the total number of helicopters available in period t, is nondecreasing. A multi-period variant of our problem (P) can be used to decide (i) where to place the trauma centers, (ii) where to initially locate the m_1 helicopters, and (iii) where to locate the m_t helicopters in each of the subsequent periods $t = 2..T$. Moreover, we can also account for growing and shrinking populations by using forecasts for λ_i^t, the number of people who will need trauma care in each region i in each time period t. The extension to (P) that we can use in this case is as follows, where x_h^t, $s_{ij}^{G,t}$, s_{hij}^t, $\lambda^{G,t}$, λ_h^t, and r_h^t are the analogs of the original decision variables that apply to each period t:

$$
\begin{aligned}
(P) \qquad & \max \sum_{t=1..T} w_t \left(\lambda^{G,t} + \sum_{h \in H} (1 - r_h^t) \lambda_h^t \right) \\
& \text{s.t. } \lambda^{G,t} = \sum_{i \in I} \sum_{j \in F_i^G} s_{ij}^{G,t} \; \forall t = 1..T \\
& \lambda_h^t = \sum_{(i,j) \in F_h} s_{hij}^t \; \forall h \in H, \; \forall t = 1..T \\
& r_h^t = \sum_{(i,j) \in F_h} \tau_{hij} s_{hij}^t \; \forall h \in H, \; \forall t = 1..T \\
& r_h^t \leq x_h^t \; \forall h \in H, \; \forall t = 1..T \\
& \sum_{j \in J} y_j \leq k \qquad (25) \\
& \sum_{h \in H} x_h^t \leq m_t \; \forall t = 1..T \\
& \sum_{j \in F_i^G} s_{ij}^{G,t} + \sum_{(h,j) \in F_i} s_{hij}^t \leq \lambda_i^t \; \forall i \in I, \; \forall t = 1..T \\
& \sum_{i \in F_j^G} s_{ij}^{G,t} + \sum_{(h,i) \in F_j} s_{hij}^t \leq c_j y_j \; \forall j \in J, \; \forall t = 1..T \qquad (26) \\
& x_j^t \leq y_j \; \forall j \in J, \; \forall t = 1..T \qquad (27) \\
& s_{ij}^{G,t} \geq 0 \; \forall (i,j) \in F^G; \qquad s_{hij}^t \geq 0 \; \forall (h,i,j) \in F, \; \forall t = 1..T \\
& y_j \in \{0,1\} \; \forall j \in J; \qquad x_h^t \in \{0,1\} \; \forall h \in H, \; \forall t = 1..T \qquad (28)
\end{aligned}
$$

The weights w_t in the objective function determine the relative importance of optimizing the solution for each period t. If all periods are the same length, then we could choose $w_t = w_1 \delta^{(t-1)} \; \forall t$, where $\delta \in (0,1)$ is a discount rate. This would ensure that $w_1 > w_2 > \cdots > w_T$; i.e., that we place more weight on optimizing the present than the future. Or, if periods differ in length, then each w_t could be chosen proportionally to the length of period

t. For example, if period 1 is a 5-year initial ramp-up period, and period 2 is the next 20 years, then we could choose $w_1 = 1$ and $w_2 = 4$ if we do not discount the future, or $w_2 < 4$ if we choose to apply some discounting.

If heliports should only be added but never moved or shut down, then we'd also impose the following constraint, to make sure that the set of heliports in service at period t is always a superset of the set of heliports in service at period $t-1$:

$$x_h^{t-1} \leq x_h^t \; \forall h \in H, \; \forall t = 2..T.$$

Moreover, we can also model the introduction and/or closure of a number of trauma centers over time by using the variables y_j^t to model whether or not trauma center j is open in period t, using k_t as a parameter that denotes the total number of trauma centers to operate in period t, and replacing (25), (27), and (28), respectively, with:

$$\sum_{j \in J} y_j^t \leq k_t;$$
$$x_j^t \leq y_j^t \; \forall j \in J, \; \forall t = 1..T; \text{ and}$$
$$y_j^t \in \{0,1\} \; \forall j \in J, \; \forall t = 1..T; \qquad x_h^t \in \{0,1\} \; \forall h \in H, \; \forall t = 1..T.$$

As well, if the effective capacity of each trauma center is expected to change over time, we can use the parameter c_j^t to denote the effective capacity of trauma center j at period t, and replace (26) with:

$$\sum_{i \in F_j^G} s_{ij}^{G,t} + \sum_{(h,i) \in F_j} s_{hij}^t \leq c_j^t y_j^t \; \forall j \in J, \; \forall t = 1..T.$$

Finally, if new trauma centers can be established over time, but old ones cannot be moved or shut down, then we'd additionally want to enforce the following constraint:

$$y_j^{t-1} \leq y_j^t \; \forall j \in J, \; \forall t = 2..T.$$

Since the functional forms of the objective and the constraints are preserved in all of these model adaptations, the Shifting Quadratic Envelopes (SQE) algorithm developed for our original model can handle all of these model variants. However, it is perhaps no surprise that the introduction of multiple periods increases the number of binary variables in the formulation, and hence makes the problem instances more computationally challenging. While it should be possible to exploit the structure of these problems using decomposition techniques tailored for multi-period problems whose periods are weakly linked, this is beyond the scope of the current paper, and so we leave this question for future research.

C. Modeling Congestion at Trauma Centers

This appendix describes how our model limits congestion at trauma centers. Specifically, we present our approach of deriving the effective capacity c_j for constraint (10) that limits the number of patients served by trauma center j.

Let us first describe (potential) patient flows within a trauma center in practice. As mentioned in the main body, Korea does not yet have any open trauma centers, and thus many of the operational details still need to be determined. Consequently, the broader study of Kim et al. (2011) uses data from large hospitals that currently

operate EDs, and describes patient flows among major resources as follows. A trauma center operates as an independent facility with all its resources exclusively dedicated to treating trauma patients arriving at the center. At a high level, a patient moves among three major departments: Emergency Room (ER), Intensive Care Unit (ICU), and Inpatient Ward (IW); see Figure 10.[20] All patients enter a trauma center through the ER. There are three different types of trauma patients:

(i) ER Type: This type of patient is treated in the ER, and then s/he is discharged;

(ii) ICU Type: This type of patient requires intensive care. If an ICU has an available bed, then s/he is admitted immediately to the ICU; otherwise, s/he stays in the ER until an ICU bed becomes available;

(iii) IW Type: This type of patient is first treated in the ER, and then s/he is transferred to the IW.

The data from large hospitals in Korea shows that 63.5% of trauma patients are of ER type, 7.9% are of ICU type, and 28.6% are of IW type. Moreover, the average Length Of Stay (LOS) of an ER-type or IW-type patient in the ER is 9.24 hours[21]; the average LOS in the ICU is 10 days; and the average LOS in the IW is 10 days for ICU-type patients and 10.6 days for IW-type patients, respectively. For the current demand of trauma patients, the study estimates that each trauma center will need 30 beds at the ER, 50 beds at the ICU, and 220 beds at the IW.

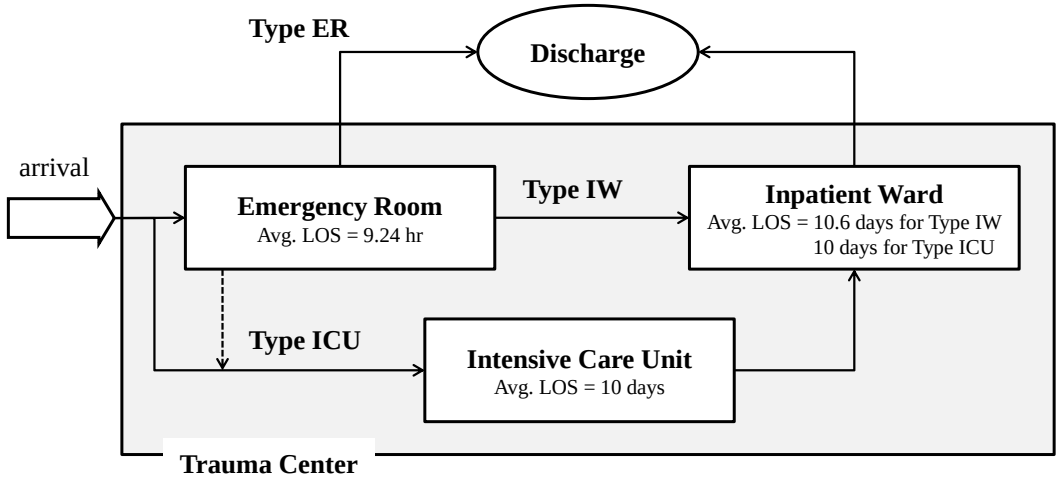


Figure 10: Patient flows within a trauma center

As discussed in the main body of the paper, when locating trauma centers and helicopters, one might wish to maximize the expected number of patients that are transported and begin to receive care at a trauma center by the 60-minute threshold. Such an objective would simultaneously incorporate the congestion at both heliports and trauma centers. However, the expression for such an objective involves convolutions of random variables, and it is too complex to work with in our optimization model. Instead, we maximize the expected number of patients that are transported without delay, and use a constraint to ensure sufficient capacity exists at each trauma center,

[20]Note that there are a number of other resources in a trauma center, including physicians, nurses, various testing equipment, and operating rooms. Kim et al. (2011) state that beds at the ER, the ICU and the IW are most critical resources at major EDs in Korea, and that other resources rarely become bottlenecks.

[21]The average LOS at the ER is from Jang, H., Kim, Y., and Lee, T. 2012. A framework for building an ED simulation model to investigate ED overcrowding. *KIIE Fall Conference*, Ansan, Korea.

so that patients rarely need to wait for a trauma center bed once they get there. For this constraint, we follow the approach developed by Marianov and Serra (1998) and Berman and Krass (2002), which we will now describe briefly.

Marianov and Serra (1998) develop a location model that considers the congestion of fixed servers. In their model, each customer is allocated to a particular facility (a trauma center in our case), and each facility is assigned one or more servers. They model the congestion at each facility by using the the probabilistic constraint Prob[waiting time at facility $j \leq \omega] \geq \xi$; i.e., the probability that a customer arriving to facility j waits at most ω units of time must be greater than or equal to ξ. Berman and Krass (2002) show how the M/M/1 queueing model used by Marianov and Serra (1998) can be naturally extended to an M/M/k model. The M/M/k queueing model is well-suited for our purpose, since we can employ a closed-form analytical formula to measure congestion that does not require estimates of higher moments of the service time distribution. Moreover, the M/M/k model is often used as a first-order approximation of complex systems (e.g., Zhang et al. 2010) even in cases when Markovian assumptions may not hold in a strict sense.

In such models, this constraint has the effect of capping the capacity of each facility, turning the formulation into a capacitated facility location problem. Using the analysis from Berman and Krass (2002) (page 354), we can rewrite the probabilistic constraint Prob[waiting time at trauma center $j \leq \omega] \geq \xi$ as $\rho_j \leq \rho_j^{\omega,\xi}$, where ρ_j is the total workload assigned to trauma center j and $\rho_j^{\omega,\xi}$ is a constant that depends on ω and ξ. Moreover, by defining $\lambda_j = \sum_{i \in F_j^G} s_{ij}^G + \sum_{(h,i) \in F_j} s_{hij}$ (patients/day) as the total patient arrival rate into trauma center j from all sources, and μ_j (patients/day) as the service rate of each server at trauma center j, we can rewrite the constraint $\rho_j \leq \rho_j^{\omega,\xi}$ as $\lambda_j \leq \mu_j \rho_j^{\omega,\xi}$. Finally, we define the effective capacity of trauma center j as $c_j = \mu_j \rho_j^{\omega,\xi}$, so that we can write our capacity constraint in the form $\lambda_j \leq c_j y_j$, which limits the number of patients served by trauma center j to c_j when it is open, or to zero when it is closed.

To find an appropriate value for the effective capacity c_j of trauma center j, we first compute $\rho_j^{\omega,\xi}$. This requires the estimated values for parameters ω and ξ. To this end, we have consulted the EMS physicians who participated in the broader study of Kim et al. (2011). They argued that delays ought to be avoided entirely to the extent possible, and hence suggested that the congestion target should be specified in terms of the fraction of patients who experience no delay. Based on their suggestions, we fix $\omega = 0$, and consider a target value for ξ between 0.9 and 0.95. Having estimated the values of ω and ξ, we can then compute $\rho_j^{\omega,\xi}$ as follows. For an M/M/k queue, Prob[waiting time > 0] is modeled by the Erlang-C probability $f_k(\rho_j)$ defined by (22) in Appendix A, where k is the number of servers and $\rho_j = \lambda_j/\mu_j$ is the total workload assigned to trauma center j. By Lemma 2 in Appendix A, $f_k(\rho_j)$ is an increasing function of ρ_j. Therefore, as ρ_j increases, Prob[waiting time at trauma center $j \leq 0$] decreases. It follows that there exists some threshold value $\rho_j = \rho_j^{\omega,\xi}$ such that Prob[waiting time at trauma center $j \leq 0$] equals ξ. By taking $\rho_j^{\omega,\xi}$ to be the value of ρ_j that satisfies $1 - f_k(\rho_j) = \xi$, we find the largest value of ρ_j that satisfies Prob[waiting time at trauma center $j \leq 0] \geq \xi$. (Note that $\rho_j^{\omega,\xi}$ can be computed for any $\omega \geq 0$ by following a similar method; c.f. Berman and Krass 2002).

A remaining issue is what constitutes the servers when we approximate the patient flow at a trauma center using an M/M/k queue. As shown in Figure 10, a trauma center consists of different types of resources, and patients follow various paths through the trauma center. If we could identify the bottleneck resource, then we could approximate the entire trauma center with only the bottleneck resource. However, as mentioned earlier, no trauma centers are open yet, so we are working with limited information. Thus, we examine each scenario in

which one of the ER, the ICU, and the IW is the bottleneck. The service rate μ_j in each scenario is obtained as the inverse of the respective average LOS, and k is taken to be the number of beds.[22] The resulting values of effective capacity c_j are shown in the following table, where we consider two cases in which $\xi = 0.9$ and $\xi = 0.95$:

Bottleneck Resource	k	μ_j	c_j when $\xi = 0.9$	c_j when $\xi = 0.95$
ER	30	2.597	58.92	55.18
ICU	50	0.100	51.16	48.69
IW	220	0.096	52.17	51.01

Units for μ_j and c_j are patients/day.

From the above results, we conclude that a reasonable value for effective capacity c_j is 50 patients/day. Constraint (10) then has the following interpretation: Given the number of beds in each trauma center department, we cap the rate at which patients are sent to each trauma center j to 50 patients/day. This ensures that 90-95% of patients receive care immediately without delay.

D. Bounds on Allocated Demand and Workload

Our relaxations of the math program (P) rely on upper bounds for λ_h and r_h, which we have denoted λ_h^{MAX} and r_h^{MAX}, respectively. In this section, we describe how the values for these upper bounds were computed.

We define λ_h^{MAX} as the minimum of three upper bounds:

$$\lambda_h^{MAX} = \min\Big(\sum_{i \in I:(i,j) \in F_h} \lambda_i,\ \sum_{j \in J:(i,j) \in F_h} c_j,\ 1/(2\tau_h^{MIN})\Big) \tag{29}$$

The first bound, $\sum_{i \in I:(i,j) \in F_h} \lambda_i$, is the total patient demand surrounding heliport h, and the second bound, $\sum_{j \in F_h} c_j$, is the total trauma center capacity surrounding heliport h. The third bound comes directly from a first order condition that we derive as follows. Recall that the contribution of heliport h to (P)'s objective function, i.e., the number of patients transported without delay by heliport h, is $\lambda_h - w_h$, where $w_h = r_h \lambda_h$. Treating the mean service time τ_h as known and fixed, we can define the objective contribution of heliport h given τ_h as $f(\lambda_h|\tau_h) = \lambda_h - w_h(\lambda_h|\tau_h) = \lambda_h(1 - r_h(\lambda_h|\tau_h)) = \lambda_h - \tau_h \lambda_h^2$. The function $f(\lambda_h|\tau_h)$ is concave quadratic in λ_h, which means that beyond a certain point, increasing the number of patients served by that heliport causes enough congestion that the number of patients served without delay (our performance metric) begins to decrease. Since the point $\lambda_h = 1/(2\tau_h)$ maximizes the function $f(\lambda_h|\tau_h)$, we know that $\lambda_h \leq 1/(2\tau_h)$ must hold for any optimal solution to the problem (P). Of course, τ_h is not known before solving (P), but we do know that $\tau_h \in [\tau_h^{MIN}, \tau_h^{MAX}]$. Therefore, we can impose the upper bound $\lambda_h \leq 1/(2\tau_h^{MIN})$ on λ_h, since $\tau_h^{MIN} \leq \tau_h \implies 1/(2\tau_h) \leq 1/(2\tau_h^{MIN})$. Note that this congestion-based bound is significant, since in many of the instances we tested, it ends up being the tightest for most of the heliports.

Next, we define r_h^{MAX} as the minimum of two upper bounds:

$$r_h^{MAX} = \min(1/2,\ \lambda_h^{MAX} \tau_h^{MAX}) \tag{30}$$

[22] Defining service rate as the inverse of average LOS implies that we treat LOS as equivalent to service time. Since LOS tends to be larger than service time, it gives an underestimate for the service rate μ_j, which in turn leads to a conservative estimate of $c_j(\omega, \xi)$.

The first bound, 1/2, follows from applying the congestion-based bound once again. Notice that $\lambda_h \leq 1/(2\tau_h)$ can be rewritten as $\lambda_h \tau_h \leq 1/2$, and from here we just need to make the substitution $r_h = \lambda_h \tau_h$ to get $r_h \leq 1/2$. Finally, the second bound in (30) holds because $r_h = \lambda_h \tau_h \leq \lambda_h^{MAX} \tau_h^{MAX}$, where λ_h^{MAX} and τ_h^{MAX} are upper bounds for λ_h and τ_h, respectively. In most instances that we tested, the first bound tends to be tighter than the second. However, sometimes a few heliports have small λ_h^{MAX} values and in these cases the second bound is tighter.

E. Computational Results: SQE vs. GBD-based Algorithm

In this appendix, we report the computational results for the test case $(k, m) = (10, 15)$ when using four different techniques that we developed: (1) Tightening (P_M^{SQE}) by Shifting Quadratic Envelopes, (2) Tightening (P_M^{GBD}) by Shifting Quadratic Envelopes, (3) Tightening (P_M^{GBD}) without the linear McCormick envelope constraints using the GBD-based algorithm, and (4) Tightening (P_M^{GBD}) using the GBD-based algorithm. We compare the performance of our specialized methods with the global solver BARON 11.8 run on the original problem (P). Recall that the formulation of (P_M^{SQE}) includes only quadratic envelopes, while (P_M^{GBD}) includes both quadratic and linear McCormick envelopes. In Figure 11, we show how the optimality gap decreases over time as each method progresses. We plot the performance of the T10H15 test case with 16 candidate heliports (i.e., trauma centers do not have their own heliports), as well as with 54 candidate heliports (i.e. trauma centers have their own heliports). As shown in Figure 11(a), the 16-heliport case yields initial solutions very quickly for technique 1 and BARON (within 10 minutes, which appears at hour 0 on the plot), and we continue to run the solver for an additional 4 iterations of 3 hours each. For the 54-heliport case shown in Figure 11(b), we ran the solver for 6 iterations of 3 hours each. In each iteration, if the master problem (P_M^{SQE}) or (P_M^{GBD}) is not solved to optimality in the allotted 3 hours, the solution found thus far is used to compute the optimality gap.

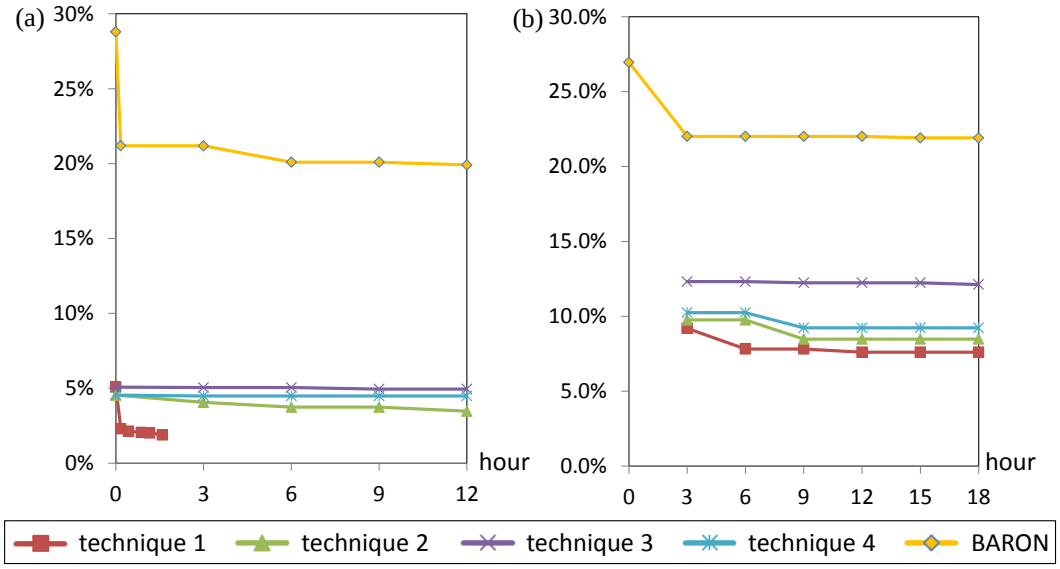


Figure 11: Optimality gaps when using four different methods of tightening the relaxations or the global solver BARON: (a) 16 candidate heliports (none of which are trauma centers), and (b) 54 candidate heliports. (Parameters: $k = 10$ trauma centers and $m = 15$ helicopters)

As you can see from Figure 11, technique (1) outperforms techniques (2), (3), and (4). The difference is most pronounced for the 16-heliport case shown in Figure 11(a): when using technique (1), the optimality gap is reduced to 5.09% in 9 seconds after the first iteration, to 2.30% in 10 minutes after the second iteration, and to 1.89% in 1.6 hours after 6 iterations (at which point we terminated the algorithm). In comparison, the second-best method, technique (2), reaches an optimality gap of 4.54% in 5 minutes after the first iteration, then slowly narrows the gap to 3.47% by the 12-hour mark after 5 iterations. All of our methods significantly outperform the global solver BARON.

F. Formal Specification of the SQE Algorithm

In this appendix we present a formal description of our Shifting Quadratic Envelopes Algorithm. The main algorithm is called `SQESolve`, which makes use of the subroutine `ShiftEnvelopes`.

F.1 Main Algorithm: SQESolve

Description: The main Shifting Quadratic Envelopes algorithm produces a near-optimal solution to the trauma system design problem (P).

1. **Required Parameters.**
   tol = the optimality gap target; once we get below this threshold, we terminate.

2. **Initialization.**
   Set $Iter \leftarrow 1$; $LB \leftarrow -\infty$; $UB \leftarrow \infty$; $gap \leftarrow \infty$; and $bestSol \leftarrow \emptyset$.
   Begin with a single slice for each heliport, with boundaries τ_h^{MIN} and τ_h^{MAX}: Set $m_h \leftarrow 1\ \forall h$; $\tau_{h1} \leftarrow \tau_h^{MIN}\ \forall h$; and $\tau_{h2} \leftarrow \tau_h^{MAX}\ \forall h$.

3. **Main Loop.**
   Repeat {
   - Solve the master problem (P_{M2}^{SQE}). Store the optimal value as z^M, and the optimal solution as Θ.
   - Solve the subproblem (P_S^{Θ}). Store the optimal value as z^S, and the optimal solution as Ψ.
   - If $z^S < LB$, then update the best solution found thus far: $bestSol \leftarrow \Psi$.
   - Update the upper bound: $UB \leftarrow \min(UB,\ z^M)$.
   - Update the lower bound: $LB \leftarrow \max(LB,\ z^S)$.
   - Update the optimality gap: $gap \leftarrow (UB - LB)/UB$.
   - If $gap < tol$, then terminate, returning $bestSol$ with value LB. Otherwise, for each heliport h call the subroutine `ShiftEnvelopes` to update the quadratic envelopes. Increment the iteration counter $Iter \leftarrow Iter + 1$, and repeat this main loop.

   }

F.2 Subroutine: ShiftEnvelopes

Description: Re-defines the quadratic envelope boundaries for a single heliport h. We keep at most ψ^A boundaries above and ψ^B boundaries below the point $(\lambda_h,\ r_h)$. Note that τ_h^{MIN} and τ_h^{MAX} always remain the lowest and highest-indexed boundaries, respectively.

1. **Required Parameters.**
   ψ^A = maximum number of boundaries to keep above the current solution $(\lambda_h,\ r_h)$
   ψ^B= maximum number of boundaries to keep below the current solution $(\lambda_h,\ r_h)$

   Note: From experience, we have chosen $\psi^A = 1$ and $\psi^B = 3$.

2. **Algorithm.**
   - Compute the actual mean service time τ_h^{ACTUAL} for the solution $(\lambda_h,\ r_h)$: If $\lambda_h = 0$, then set $\tau_h^{ACTUAL} \leftarrow 0$. Otherwise, set $\tau_h^{ACTUAL} \leftarrow r_h/\lambda_h$.
   - Store the index n^* of the slice that the solution $(\lambda_h,\ r_h)$ is in: $n^* \leftarrow \{n \mid x_{hn} = 1\}$.
   - Create a new slice boundary τ_h^{NEW} midway between the current lower bound for the mean service time τ_{h,n^*} and the actual mean service time τ_h^{ACTUAL} used at point $(\lambda_h,\ r_h)$: $\tau_h^{NEW} \leftarrow \left(\tau_{h,n^*} + \tau_h^{ACTUAL}\right)/2$.
   - Compute the index $n' \leftarrow \min(n^*+1, \psi^B)$, which will become the index for the newly-derived boundary τ_h^{NEW}.
   - Compute the index $n'' \leftarrow \min(n'+\psi^A, m_h+1)$, which will become the index for the highest boundary τ_h^{MAX}.
   - Update the slice boundaries: $\tau_{h,1} \leftarrow \tau_h^{MIN}$;
     $\tau_{h,n} \leftarrow \tau_{h,n^*-n'+n+1}\ \forall n = 2..n'-1$; $\tau_{h,n'} \leftarrow \tau_h^{NEW}$; $\tau_{h,n} \leftarrow \tau_{h,n^*-n'+n}\ \forall n = n'+1..n''-1$; and $\tau_{h,n''} \leftarrow \tau_h^{MAX}$.
   - Update the slice count: $m_h \leftarrow n''-1$.

G. Derivations for the GBD-Based Algorithm

In this appendix, we complete some derivations for the GBD-based algorithm. In particular, we (1) derive the subproblem dual (D_S^Θ) and prove Lemma 1; and (2) describe how the $\max(\lambda_h, \gamma_h^t)x_h$ expressions in the Bender's cut $z \le B_t(\Theta)$ are implemented using binary variables.

G.1 Derivation of the GBD-based Algorithm

The subproblem primal (P_S^{Θ}) is:

$$
\begin{aligned}
u^*(\Theta) = \min \sum_h \lambda_h r_h & \qquad \underline{\text{Dual Variables}} \\
\text{s.t.} -\sum_{h,i} s_{hij} \geq -a_j \; \forall j \in J & \qquad \ldots \alpha_j \geq 0 \\
-\sum_{h,j} s_{hij} \geq -b_i \; \forall i \in I & \qquad \ldots \beta_i \geq 0 \\
-\sum_{i,j} s_{hij} = -\lambda_h \; \forall h \in H & \qquad \ldots \Delta_h \text{ free} \\
-\sum_{i,j} \tau_{hij} s_{hij} + r_h = 0 \; \forall h \in H & \qquad \ldots \gamma_h \text{ free} \\
-r_h \geq -x_h \; \forall h \in H & \qquad \ldots \eta_h \geq 0 \\
s_{hij} \geq 0 \; \forall (h,i,j) \in F. &
\end{aligned}
$$

Its dual (D_S^{Θ}) is:

$$
\begin{aligned}
d^*(\Theta) = \max -\sum_{j\in J} a_j \alpha_j - \sum_{i \in I} b_i \beta_i - \sum_{h \in H} \lambda_h \Delta_h - \sum_{h \in H} x_h \eta_h & \\
\text{s.t.} -\alpha_j - \beta_i - \Delta_h - \tau_{hij} \gamma_h \leq 0 \; \forall (h,i,j) \in F & \\
\gamma_h - \eta_h = \lambda_h \; \forall h \in H & \\
\alpha_j \geq 0 \; \forall j \in J \qquad \beta_i \geq 0 \; \forall i \in I \qquad \eta_h \geq 0 \; \forall h \in H. &
\end{aligned}
$$

By eliminating $\{\eta_h\}$ and substituting $\eta_h = \gamma_h - \lambda_h$ into the objective, we get:

$$
\begin{aligned}
d^*(\Theta) = \max -\sum_{j\in J} a_j \alpha_j - \sum_{i \in I} b_i \beta_i - \sum_{h \in H} \lambda_h \Delta_h - \sum_{h \in H} x_h \gamma_h + \sum_{h \in H} x_h \lambda_h & \\
\text{s.t. } \alpha_j + \beta_i + \Delta_h + \tau_{hij} \gamma_h \geq 0 \; \forall (h,i,j) \in F & \\
\lambda_h \leq \gamma_h \; \forall h \in H & \\
\alpha_j \geq 0 \; \forall j \in J \qquad \beta_i \geq 0 \; \forall i \in I. &
\end{aligned}
$$

Finally, we note that since $x_h \in \{0,1\}$ and $x_h = 0 \implies \lambda_h = 0$ in the master problem (P_M^{GBD}), we can replace the objective term $\sum_h x_h \lambda_h$ with $\sum_h \lambda_h$. The result is the following dual problem, which optimizes over the variables $\{\{\alpha_j\}, \{\beta_i\}, \{\Delta_h\}, \{\gamma_h\}\}$:

$$
\begin{aligned}
d^*(\Theta) = \sum_{h \in H} \lambda_h + \max -\sum_{j\in J} a_j \alpha_j - \sum_{i \in I} b_i \beta_i - \sum_{h \in H} \lambda_h \Delta_h - \sum_{h \in H} x_h \gamma_h & \qquad (31) \\
\text{s.t. } \alpha_j + \beta_i + \Delta_h + \tau_{hij} \gamma_h \geq 0 \; \forall (h,i,j) \in F & \qquad (32) \\
\lambda_h \leq \gamma_h \; \forall h \in H & \qquad (33) \\
\alpha_j \geq 0 \; \forall j \in J \qquad \beta_i \geq 0 \; \forall i \in I. & \qquad (34)
\end{aligned}
$$

Since the subproblem is a linear program, by strong duality, we have $d^*(\Theta) = u^*(\Theta)$, where $u^*(\Theta)$ is the optimal value of (P_S^{Θ}). Therefore, a lower bound $v^*(\Theta)$ for full problem (P) using the fixed master problem

solution Θ is:

$$v^*(\Theta) = \lambda^G + \sum_{h\in H} \lambda_h - u^*(\Theta) = \lambda^G + \sum_{h\in H} \lambda_h - d^*(\Theta) = \lambda^G + \sum_{j\in J} a_j\alpha_j + \sum_{i\in I} b_i\beta_i + \sum_{h\in H} \lambda_h\Delta_h + \sum_{h\in H} x_h\gamma_h.$$

It is worth pointing out that all quantities in the above expression depend on the master problem solution Θ. This includes the master problem variables $\left\{\lambda^G, \{\lambda_h\}, \{x_h\}\right\}$, the derived values $\{\{a_j\}, \{b_i\}\}$, and the optimal dual solution $\{\{\alpha_j\}, \{\beta_i\}, \{\Delta_h\}, \{\gamma_h\}\}$.

At each iteration t, given master problem solution Θ_t, we generate a Benders cut of the form $z \leq B_t(\Theta)$ that holds for *all* feasible master problem solutions Θ. To generate this cut, we use the optimal solution $SD_t = \left\{\{\alpha_j^t\}, \{\beta_i^t\}, \{\Delta_h^t\}, \{\gamma_h^t\}\right\}$ from the subproblem dual $(D_S^{\Theta_t})$. As required for a Benders optimality cut, each cut t satisfies $B_t(\Theta) \geq v^*(\Theta)$ for all feasible master problem solutions Θ, and this inequality holds at equality when evaluated at Θ_t; i.e., $B_t(\Theta_t) = v^*(\Theta_t)$.

To construct $B_t(\Theta)$, we first need a (preferably near-optimal) feasible solution to (D_S^{Θ}) that we can express as a closed-form expression of the master problem variables Θ and the optimal subproblem dual solution SD_t. Let $\alpha_j(\Theta, SD_t) := \alpha_j^t \ \forall j$, $\beta_i(\Theta, SD_t) := \beta_i^t \ \forall i$, $\Delta_h(\Theta, SD_t) := \Delta_h^t \ \forall h$, and $\gamma_h(\Theta, SD_t) := \max(\lambda_h, \gamma_h^t) \ \forall h$.

Lemma 5 *The vector* $\{\{\alpha_j(\Theta, SD_t)\}, \{\beta_i(\Theta, SD_t)\}, \{\Delta_h(\Theta, SD_t)\}, \{\gamma_h(\Theta, SD_t)\}\}$ *as defined above is feasible in* (D_S^{Θ}).

Proof. Checking constraint (32) for each tuple (h, i, j), we find that:

$$\begin{aligned}&\alpha_j(\Theta, SD_t) + \beta_i(\Theta, SD_t) + \Delta_h(\Theta, SD_t) + \tau_{hij}\gamma_h(\Theta, SD_t)\\ &= \alpha_j^t + \beta_i^t + \Delta_h^t + \tau_{hij}\max(\lambda_h, \gamma_h^t) \geq \alpha_j^t + \beta_i^t + \Delta_h^t + \tau_{hij}\gamma_h^t \geq 0,\end{aligned}$$

where the first inequality follows since $\tau_{hij} > 0 \ \forall (h, i, j) \in F$, and the second inequality follows since $\left\{\{\alpha_j^t\}, \{\beta_i^t\}, \{\Delta_h^t\}, \{\gamma_h^t\}\right\}$ was optimal (and hence feasible) in $(D_S^{\Theta_t})$. Furthermore, checking constraint (33) for each heliport h, we find that $\lambda_h \leq \gamma_h(\Theta, SD_t) = \max(\lambda_h, \gamma_h^t)$ holds trivially by definition. ■

Evaluating $\{\{\alpha_j(\Theta, SD_t)\}, \{\beta_i(\Theta, SD_t)\}, \{\Delta_h(\Theta, SD_t)\}, \{\gamma_h(\Theta, SD_t)\}\}$ in the objective function of (D_S^{Θ}) gives us:

$$\begin{aligned}d(\Theta, SD_t) &= \sum_h \lambda_h - \sum_j a_j\alpha_j(\Theta, SD_t) - \sum_i b_i\beta_i(\Theta, SD_t) - \sum_h \lambda_h\Delta_h(\Theta, SD_t) - \sum_h x_h\gamma_h(\Theta, SD_t)\\ &= \sum_h \lambda_h - \sum_j \alpha_j^t a_j - \sum_i \beta_i^t b_i - \sum_h \Delta_h^t\lambda_h - \sum_h \max(\lambda_h, \gamma_h^t)x_h.\end{aligned}$$

Finally, we re-state Lemma 1 given in the main body, and provide its proof.

Lemma 1. *The function* $B_t(\Theta)$ *as defined in (21) satisfies* $B_t(\Theta) \geq v^*(\Theta)$ *for all feasible master problem solutions* Θ*, and furthermore* $B_t(\Theta_t) = v^*(\Theta_t)$ *holds at equality for the particular master problem solution* Θ_t*. In other words,* $z \leq B_t(\Theta)$ *is a Bender's optimality cut.*

Proof. Since (D_S^{Θ}) is a maximization problem with optimal value $d^*(\Theta)$, by definition of optimality we have $d(\Theta, SD_t) \leq d^*(\Theta)$. Therefore:

$$v^*(\Theta) = \lambda^G + \sum_h \lambda_h - d^*(\Theta) \leq \lambda^G + \sum_h \lambda_h - d(\Theta, SD_t) = B_t(\Theta).$$

Furthermore, since $d(\Theta_t, SD_t) = d^*(\Theta_t)$, the above inequality becomes an equality for the master problem solution Θ_t.■

G.2 Implementing $\max(\lambda_h, \gamma_h^t)x_h$

To implement the $\max(\lambda_h, \gamma_h^t)x_h$ expressions in the Benders cut $z \leq B_t(\Theta)$ as defined in (21), we introduce binary variables $\{u_{th}\}$ that keep track of which term in the maximum is the largest; i.e., $u_{th} = 1$ if $\lambda_h \geq \gamma_h^t$, and $u_{th} = 0$ otherwise. Because x_h is binary and $x_h = 0 \implies \lambda_h = 0$, we have $\lambda_h x_h = \lambda_h$. Therefore, $\max(\lambda_h, \gamma_h^t)x_h = \lambda_h$ when $u_{th} = 1$, and $\max(\lambda_h, \gamma_h^t)x_h = \gamma_h^t x_h$ otherwise. As well, for each cut t and each heliport h, we introduce a continuous variable λ_{th}^{HIGH} and a binary variable x_{th}^{LOW}. The variable λ_{th}^{HIGH} is forced to be equal to λ_h when $u_{th} = 1$, and is set to zero otherwise. The variable x_{th}^{LOW} is forced to be equal to x_h when $u_{th} = 0$, and is set to zero otherwise. The Benders cut we implement is $z \leq B_t(\Theta)$, where

$$B_t(\Theta) = \lambda^G + \sum_{j \in J} \alpha_j^t a_j + \sum_{i \in I} \beta_i^t b_i + \sum_{h \in H} \Delta_h^t \lambda_h + \sum_{h \in H} \lambda_{th}^{HIGH} + \sum_{h \in H} \gamma_h^t x_{th}^{LOW}.$$

To model the required logic and link together the $\{\lambda_h\}$, $\{\lambda_{th}^{HIGH}\}$, $\{x_h\}$, $\{x_{th}^{LOW}\}$, and $\{u_{th}\}$ variables, we use the following constraints, which we add to the master problem (P_{M2}^{GBD}):

$$\begin{aligned}
&\gamma_h^t u_{th} \leq \lambda_h \leq \gamma_h^t + (\lambda_h^{MAX} - \gamma_h^t)u_{th} \ \forall h \in H, \forall t = 1..nCuts \\
&0 \leq \lambda_h - \lambda_{th}^{HIGH} \leq \lambda_h^{MAX}(1 - u_{th}) \ \forall h \in H, \forall t = 1..nCuts \\
&0 \leq \lambda_{th}^{HIGH} \leq \lambda_h^{MAX} u_{th} \ \forall h \in H, \forall t = 1..nCuts \\
&0 \leq x_h - x_{th}^{LOW} \leq u_{th} \ \forall h \in H, \forall t = 1..nCuts \\
&0 \leq x_h^{LOW} \leq (1 - u_{th}) \ \forall h \in H, \forall t = 1..nCuts.
\end{aligned}$$

H. Decoupled Heuristic

In this appendix, we describe the decoupled heurisitic in detail. Recall that the decoupled heuristic has two stages. In the first stage, we solve for the trauma center locations, and in the second stage, we solve for the heliport locations. In addition to the sets that we defined earlier in §3, we make use of the following sets:

• $F_i^H = \{j \in J | d_{ij} > d_{ground}$ and $d_{hi} + d_{ij} \leq d_{air}$ for fixed i and for some $h \in H\}$: the set of trauma centers that patients in demand region i can reach by helicopter within 60 minutes.

• $F_j^H = \{i \in I | d_{ij} > d_{ground}$ and $d_{hi} + d_{ij} \leq d_{air}$ for fixed j and for some $h \in H\}$: the set of demand regions that are covered by trauma center j using helicopter transportation.

• $H_{ij} = \varnothing$ if $d_{ij} \leq d_{ground}$, or $H_{ij} = \{h \in H | d_{hi} + d_{ij} \leq d_{air}$ for fixed $(i, j)\}$ if $d_{ij} > d_{ground}$: the set of heliports that can be used to fly patients from demand region i to trauma center j within 60 minutes.

The decoupled heuristic's first stage is formulated as a variant of the capacitated maximal covering location problem (cMCLP) (Pirkul and Schilling 1991). This problem has three decision variables: s_{ij}^G, y_j, and s_{ij}^H. The first two variables, defined earlier in §4.1, represent respectively the number of patients per unit time transported by ambulance from demand region i to trauma center j, and an indicator for whether trauma center j is opened. The third variable, s_{ij}^H, is new to the decoupled heuristic, and represents the number of patients per unit time transported by helicopter from demand region i to trauma center j. The problem is formulated as a Mixed-Integer

Linear Program (MILP) as follows:

$$\max \sum_{i \in I} \sum_{j \in F_i^G} s_{ij}^G + \sum_{i \in I} \sum_{j \in F_i^H} s_{ij}^H \tag{35}$$

$$\text{s.t.} \sum_{j \in J} y_j \leq k \tag{36}$$

$$\sum_{j \in F_i^G} s_{ij}^G + \sum_{j \in F_i^H} s_{ij}^H \leq \lambda_i \quad \forall i \in I \tag{37}$$

$$\sum_{i \in F_j^G} s_{ij}^G + \sum_{i \in F_j^H} s_{ij}^H \leq c_j y_j \quad \forall j \in J \tag{38}$$

$$\sum_{i \in I} \sum_{j \in F_i^H} \overline{\tau}_{ij} s_{ij}^H \leq m \tag{39}$$

$$s_{ij}^G \geq 0; \quad \forall i \in I, j \in F_i^G; \qquad s_{ij}^H \geq 0 \quad \forall i \in I, j \in F_i^H \tag{40}$$

$$y_j \in \{0,1\} \quad \forall j \in J, \tag{41}$$

where (35), (36), (37), (38), (39), (40) and (41) can be interpreted similarly to (5), (7), (9), (10), (6), (12) and (13), respectively, in the integrated model presented in §4.1. It is worth noting the following differences between the above formulation and the integrated model. First, program (35)-(41) does not determine the locations of helicopters modeled as x_h in the integrated model. Instead, it assumes that a helicopter is stationed at every heliport; i.e., $x_h = 1 \; \forall h$. Second, since the locations of the helicopters are not determined at this stage, program (35)-(41) solves for s_{ij}^H, i.e., the *total* volume of patients that should be flown between demand region i and trauma center j, instead of the more detailed allocations s_{hij} that, in the integrated model, determine the volume of patients transported from demand region i to trauma center j using each *specific* helicopter (heliport) h. Third, in the integrated model, constraint (6) ($\sum_{(i,j) \in F_h} \tau_{hij} s_{hij} \leq x_h \; \forall h$) ensures that the busy fraction (or utilization) of heliport h is at most 1. Since s_{hij} is not present in program (35)-(41), we introduce the analogous constraint (39), where $\overline{\tau}_{ij}$ is the average service time for a helicopter in the set H_{ij} to fly the circuit $h \to i \to j \to h$.

Once we have solved for the trauma center locations (y_j) and the demands for helicopters (s_{ij}^H) in the first stage, we run the decoupled heuristic's second stage to determine the helicopter locations. We formulate the decoupled heuristic's second stage as a variant of the Maximum Expected Covering Location Problem (MEXCLP) (Daskin 1983). The two primary decision variables are x_h, as defined in §4.1, and s_{ijl}^H, which we set equal to s_{ij}^H when l ($\in \{1, 2, ..., m\}$) helicopters serve the demand s_{ij}^H, or zero otherwise. Binary variables u_{ijl}^H are used to link

s^H_{ij} to s^H_{ijl}. The problem is formulated as the following MILP:

$$\max \sum_{i\in I}\sum_{j\in F^H_i}\sum_{l=1}^{m}(1-\overline{\rho}^l)s^H_{ijl} \tag{42}$$

$$\text{s.t.} \sum_{h\in H} x_h \le m$$

$$x_j \le y_j \quad \forall j \in J$$

$$s^H_{ijl} \le s^H_{ij} u^H_{ijl} \quad \forall i \in I,\ j \in F^H_i,\ \forall l = 1..m$$

$$\sum_{l=1}^{m} u^H_{ijl} \le 1 \quad \forall i \in I,\ j \in F^H_i$$

$$\sum_{l=1}^{m} l\ u^H_{ijl} \le \sum_{h\in H_{ij}} x_h \quad \forall i \in I,\ j \in F^H_i \tag{43}$$

$$s^H_{ijl} \ge 0; \qquad u^H_{ijl} \in \{0,1\} \quad \forall i \in I,\ j \in F^H_i,\ \forall l = 1..m$$

$$x_h \in \{0,1\} \quad \forall h \in H$$

In (42), $\overline{\rho}$ is a *global busy fraction* for helicopters, and is computed from stage one's s^H_{ij} values using the formula $\overline{\rho} = \sum_{i\in I}\sum_{j\in F^H_i} \overline{\tau}_{ij} s^H_{ij}/m$, where $\sum_{i\in I}\sum_{j\in F^H_i} \overline{\tau}_{ij} s^H_{ij}$ is the total workload assigned to helicopters in the first stage; see (39). Thus, $(1-\overline{\rho}^l)$ represents the probability that at least one helicopter is available to transport a patient from demand region i to trauma center j within 60 minutes. This objective, which differs slightly from the original model of Daskin (1983), has been used in the literature (e.g., Sorensen and Church 2010). Constraint (43) specifies that if (i,j) is covered by l heliports, then at least l heliports should be in the set H_{ij} that can serve (i,j). The other constraints are straightforward.

We note that there are different ways that one might model busy fractions in such a two-stage heuristic. For example, we could modify ReVelle and Hogan's (1989) area-specific busy fraction concept to estimate (i,j)-specific busy fractions iteratively, using the locations of trauma centers and helicopters from the current iteration to update the busy fractions for the subsequent iteration.[23] However, such variations of busy fraction estimation schemes have their own weaknesses. Convergence can be an issue, for example, or more stable definitions for ρ_{ij} may rely on somewhat arbitrary assumptions for the estimated workload at each heliport. We find that, in general, the integrated approach outperforms these variations as well, and comes with the added advantage of establishing a theoretical optimality gap.

Note that, in the decoupled heuristic's second stage math program, it is not possible to use heliport-specific busy fractions like the r_h values in our integrated model from §4.1. This is because the first stage only determines the total demand for helicopters, s^H_{ij}, for each (i,j) pair, and does not specify the demands for specific helicopter routes s_{hij} as in our integrated model. That is, the workload at a particular heliport h can, at best, only be coarsely approximated after the first stage.

[23] Details of this approach can be found in: Lee, T., H. Jang, S-H. Cho, and J.G. Turner. 2012. A Simulation-Based Iterative Method for a Trauma Center - Air Ambulance Location Problem. Proceedings of the 2012 Winter Simulation Conference, Berlin, Germany.

I. Out-of-Sample Results

As described in §5.1, we tested our methods both in-sample (e.g., by optimizing the trauma center and heliport locations using January-June data and then evaluating the performance of this solution with a simulation using January-June data) and out-of-sample (e.g., optimizing with January-June data and evaluating with July-December data). We reported the in-sample results in the main body of the paper. The out-of-sample results, which serve to show that our methods are robust even in the face of aggregate-level data uncertainty, are reported here. Note that the out-of-sample results are qualitatively similar to the in-sample results, which further validates our approach.

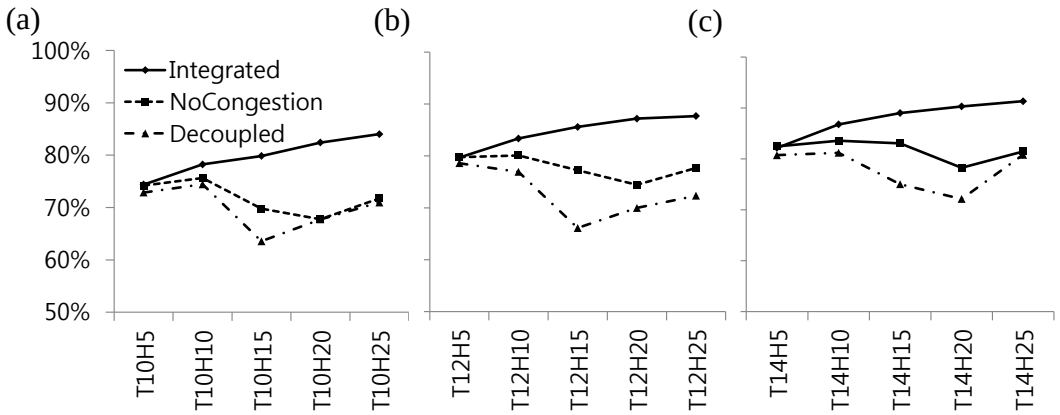


Figure 12: The proportion of successful patients when: (a) $k = 10$, (b) $k = 12$, and (c) $k = 14$.

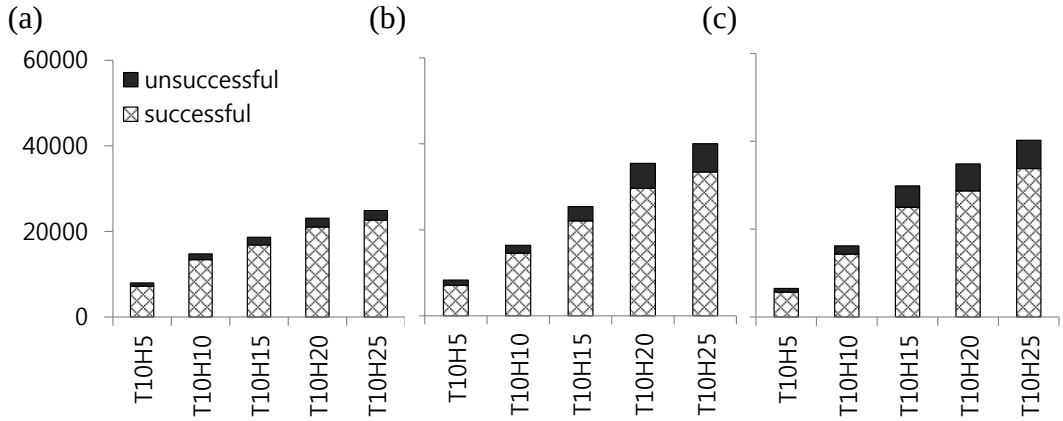


Figure 13: The total number of successful and unsuccessful helicopter transports under the solutions determined by: (a) integrated approach, (b) no-congestion heuristic, and (c) decoupled heuristic.

J. Simulation Details

In this appendix, we describe our simulation model in further detail. In particular, decision points A, B, C, and D in Figure 5 deserve further comment.

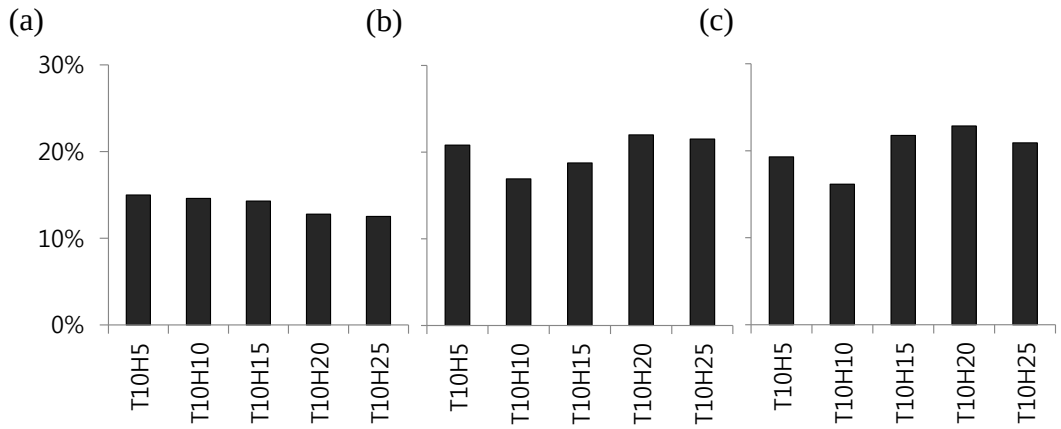


Figure 14: The proportion of helicopter transports that were delayed under the location solutions determined by: (a) integrated approach, (b) no-congestion heuristic, and (c) decoupled heuristic.

As depicted by decision point A, we first determine whether the patient is within the coverage area defined by our location solution. A patient is deemed to be out of the coverage area if no trauma center can ever be reached in under 60 minutes from the patient's location by either ambulance or helicopter. We label out-of-coverage patients as unsuccessful, since these patients are handled by regional hospitals outside of the trauma care system.[24]

At decision point B, if the patient is within the coverage area, we check if there is an available (i.e., under-capacity) trauma center. If all trauma centers are over-capacity, which happened very rarely in our simulation runs, then we send the patient to the nearest trauma center. Otherwise, we send the patient to the nearest *available* (i.e., under-capacity) trauma center. Each patient is either ER type, ICU type, or IW type (upon arrival, we randomly assign these types according to the respective proportions mentioned in Appendix C). For an ER-type or IW-type patient, if a trauma center has an available bed in its ER, then it is under-capacity. For an ICU-type patient, if a trauma center has an available bed in its ER or in its ICU, then it is under-capacity. If no trauma centers within the coverage area are under-capacity, then the patient is sent to the nearest trauma center and waits until a bed becomes available. Once the patient is admitted to a trauma center, s/he follows one of the paths shown in Figure 10 according to his/her type as follows. An ER-type patient occupies an ER bed, and then leaves the trauma center upon the completion of receiving treatment. An IW-type patient receives initial care from the ER and then gets transferred to the IW before being discharged. This type of patient may have to wait in the ER if there is no bed in the IW that is immediately available (such patients are often called boarding patients). An ICU-type patient is admitted directly to the ICU if an ICU bed is available, or otherwise s/he is admitted to the ER, and then stays in the ER until an ICU bed becomes available. Service times in the ER, the ICU, and the IW are sampled from exponential distributions with their means equal to the corresponding LOS values shown in Figure 10 in Appendix C.

Decision point C limits the use of helicopters to trips under 120km in length (i.e., transports within the helicopter coverage area). In other words, if the nearest available trauma center is so far from the patient that even a helicopter cannot fly there within 60 minutes, then we do not dispatch a helicopter. Instead, we transport

[24]In practice, it is possible that some of these patients may get transferred from a regional hospital to one of the trauma centers. However, since they do not receive proper care from a trauma center within 60 min, it seems reasonable to count these patients as unsuccessful for our purpose.

this patient by ambulance, which prevents a helicopter from being used for an excessively long time only to complete an unsuccessful transport.[25]

Finally, when the destination trauma center is outside the ground coverage area but within the helicopter coverage area, we use either an ambulance or a helicopter, depending on which mode will get the patient to the trauma center faster (decision points D). When a nearby helicopter is available, using a helicopter is always faster (because being outside the ground coverage area means that it will take longer than an hour for an ambulance to transport this patient to the nearest trauma center; see §3). On the other hand, when all nearby helicopters are busy, it is sometimes faster to use an ambulance rather than waiting for a helicopter to become available. To make this decision, the model computes the lead time (wait time + transportation time) of using a helicopter from each heliport within the patient's helicopter coverage area based on (i) the location of the patient requiring service, (ii) the location of the patient currently in transit (if any), and (iii) the number and locations of the patients waiting in that heliport's queue.

[25]Our results do not change substantially when we allow helicopters to fly outside of the defined 120km range. This is because only a small number of patients are affected by this rule. Specifically, among all test cases for the integrated method, no more than 0.3% of patients that are within the location solution's coverage area find all nearby trauma centers unavailable and need to be transported farther than 120km.