Abstract

Title of Dissertation: APPLICATIONS OF OPERATIONS RESEARCH MODELS TO PROBLEMS IN HEALTH CARE

Carter Claiborne Price, Doctor of Philosophy, 2009

Dissertation Directed by: Professor Bruce L. Golden

R. H. Smith School of Business

Applied Mathematics and Scientific Computation

This dissertation is divided into two parts. In the first portion, we study inflection points in bi-objective variants of the traveling salesman problem (TSP) related to health care applications. In the second portion, we use a variety of techniques from operations research to improve hospital efficiency.

We used a TSP variant that prioritizes the ability to return to the depot, in addition to the standard distance, to study the collection of blood from remote collection sites. In an application related to emergency response, we looked into behavior of tours generated using the target visitation problem, a TSP variant that also includes node priority in the objective function.

Working with the University of Maryland Medical Center, we did three projects related to hospital efficiency. We used stochastic modeling and simulation to optimize the throughput of a cardiac surgery post-operative unit. We found that altering the mix of post-operative beds could significant increase the effective capacity. After unsuccessfully

attempting to use data mining and survival analysis to predict hospital census, we performed a statistical analysis of patient length of stay patterns and discovered surgeons were improving their chances of having available beds for incoming cases by adjusting their discharge practices by day of week. In the final project, we used integer programming and heuristics to develop schedules that match incoming surgical patients with the discharge of earlier patients. The results indicate that altering the surgical schedule can substantially improve the flow of patients through the hospital.

APPLICATIONS OF OPERATIONS RESEARCH MODELS TO PROBLEMS IN HEALTH CARE

by
Carter Claiborne Price

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Bruce Golden, Chair
Professor David Bigio
Professor Jeffery Herrmann
Professor Wolfgang Jank
Professor Paul Smith
Professor Edward Wasil

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

At over two trillion dollars, health care is the largest single industry in the United States' economy. The US spends more money per capita on health care than any other nation in the world, but does not have the best health care system by several metrics [1, 2]. The US has a population of about 46 million uninsured people. Infant mortality is among the highest in the developed world. Despite the largest per capita expenditures, average life span is in the middle of the industrial world's rankings. Health care expenditures are expected to increase over the next several decades as the baby-boomers retire [3]. In order to maintain or improve upon the current quality of health care in the US in the coming years, efficiency must be increased.

A key component of increasing the efficiency of the overall health care system is improving the delivery of health care services. Hospitals are the primary source of health services in the developed world and, as seen in Figure 1.1, comprise almost one third of US

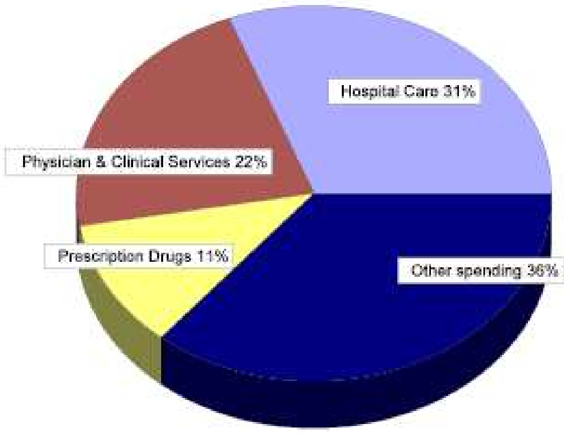

Figure 1.1: A pie chart with the distribution of US health expenditures in 2002 [4].

health expenditures. In this dissertation, we look into issues relating to the improvement of hospital efficiency and problems in the field of efficient emergency response.

1.2 Overview

This dissertation explores the application of mathematical modeling and operations research techniques to improving the efficiency of health care systems. There are two portions to this work. The first (short) portion of the work applies biobjective variants of the traveling salesman problem (TSP) to routing blood collection vehicles and emergency aid distribution vehicles. The second (much larger) portion was done in close consultation with the University of Maryland Medical Center, a large urban hospital, and it seeks to improve various hospital functions related to the surgical services. In both sections, a variety of optimization techniques are used in conjunction with other mathematical methods to tackle real-world problems.

1.2.1 TSP Variants

The traveling salesman problem has been used for a variety of applications and is a very well studied problem. We investigate two biobjective variants of the TSP that have only recently appeared in the literature.

The first variant is referred to as the traveling salesman problem with a center (TSPwC). In this problem, there are two parts to the objective function. The first part is the standard TSP objective which seeks to minimize the Euclidean distance covered by the tour. The second portion seeks to keep the tour close to the center either by minimizing the sum of the distances of the midpoint of each edge or the nearest point of each edge to the center of the graph. The two portions of the objective function are in conflict because minimizing the distance from the center generally results in tours that have intersecting edges (which never happens in an optimal Euclidean TSP tour). Our analysis focuses on the effects of the changes in tour length, average distance from the center, and other factors as the relative emphasis on the two portions of the objective function varies. Essentially, we look into the qualitative and quantitative changes in the tour as each portion's relative importance changes.

This problem has applications to the collection of blood donations for surgical services. Hospitals would like to receive blood supply quickly (Euclidean objective) but may need an emergency delivery at any time (distance from the center). Thus, a good tour must take both of these objectives into account to satisfy the overall goal.

The second variant is the target visitation problem (TVP). This problem was originally formulated to produce tours for unmanned aerial vehicles. Like the TSPwC, this is a biobjective variant of the TSP. The first portion is the Euclidean portion and the second is a priority weighting which is very similar to the linear ordering problem. In the linear ordering problem, each pair of nodes, i and j, has two values, one for visiting i before j and another for visiting i after j. The objective is to maximize the sum of the edge values. The priority weighting used in the TVP is essentially the same as the linear ordering except the objective is to minimize the sum. As with the TSPwC, we focused on the balance between the two portions of the objective function, though in this problem there is no clear conflict between the two components.

This problem can be applied to the delivery of emergency supplies to regions suffering after a natural disaster. In particular, we looked into priority weightings consistent with the destruction caused by earthquakes, tsunami, and hurricanes. We also look at alternative formulations of the problem that cannot be modeled as TSP variants.

1.2.2 Hospital Efficiency

For the second segment of the dissertation, we worked closely with staff and administrators at the University of Maryland Medical Center in Baltimore, Maryland. The people at the hospital suggested areas that were in need of improvement and we determined approaches to investigate these problems. The hospital staff made sure the results were feasible by suggesting additional constraints. There are four parts to this work, each of which involves a different aspect related to improving the flow of post-operative patients through the hospital.

In the first problem, we look at how best to allocate post-operative beds between two cardiac surgery post-operative units. The cardiac surgery service line was not able to meet their desired throughput because of bed capacity. Using queueing theory and simulation, we maximized the throughput by changing the mix of beds. The hospital administrators liked the analysis and the cardiac surgery service line currently uses the mix we suggested.

Our second problem involved predicting post-operative bed capacity. The hospital administrators would like to predict capacity so that they can efficiently allocate nurses and other resources in a timely fashion. Prediction would also benefit the surgeons. Surgeons do not know how many beds will be available in the future and they must schedule their cases independent of the future capacity. This can cause serious problems if surgeons perform without a post-operative bed available for patients. Initially, we focused on cardiac surgery.

We first used data mining and techniques from statistics to attempt to predict the length of stay of individual cardiac surgery patients. The variability was too high to provide a sufficiently accurate estimate and so a different approach was needed. Using clustering techniques and probability theory, we made an estimate of the capacity a few days in advance. We tested the approach using simulation on historical data and the results were quite promising. After the simulation, we implemented the prediction approach in the hospital. These results were much worse than in the simulation and indicated our approach was too simplistic. In particular, it appeared that surgeons' decisions concerning discharges are strongly influenced by the number of scheduled cases and the number of empty beds.

The problems with the prediction led to the third project at the hospital. Using data on several hundred patients for cardiac surgery and general surgery, we performed a detailed statistical analysis of the length of stay of patients. We looked at patient variables like severity of illness or age. We also looked at service line variables such as available operating room time and daily arrivals. We found that factors exogenous to the patient's health state, such as the available operating room hours, had a statistically significant impact on both the length of stay and volume of discharges. We also discuss possible causes for this effect related to the incentives of the physicians.

The final project at the hospital focused on developing approaches to improving the flow of post-operative patients by changing the allocation of operating room time to the different surgical service lines. First, we clustered the service lines into groups with similar post-operative length of stay distributions and case volumes per day in the operating room. Next, we developed an integer program formulation and other scheduling heuristics to reduce the number of patients without post-operative beds following their surgery. We then tested each scheduling approach using a simulation experiment. We found that significant improvement in patient flow could be achieved by altering the scheduling approach.

1.3 Theme

Each of the problems addressed in this dissertation required significant knowledge of both the underlying subject matter and a variety of mathematical disciplines. In order to address these real-world issues, we used techniques from numerical analysis, optimization, and statistics. These projects required detailed simulation models to test the simplified models used for optimization.

One of the primary lessons to be taken from the work as a whole is the importance of comparing disparate models and the available data. The prediction work uncovered a fundamental flaw in the model assumptions because the model results did not work in practice. This was not a wasted effort because it directed us to a deeper understanding of the discharge process. Even hospital administrators were surprised by these findings. Thus, the key lesson to be learned from this work is the importance of using a rigorous model validation process. In each of the following chapters, we will apply these principles to different problems related to health care efficiency.

Chapter 2

Routing Blood Bank Vehicles: An Application of the Traveling Salesman Problem with a Center

Hospitals that perform certain types of surgery are required to maintain a certain level of blood supply on hand. To do this, blood collection vehicles periodically go to donation sites to make pick ups. If there is an emergency, a blood collection vehicle may need to quickly return to the hospital. The traveling salesman problem with a center seeks to minimize the total distance it takes to visit a set of nodes (blood collection cites) while minimizing the distance (using some metric) from the center (hospital). In this paper, we compare different distance metrics, center location, and distributions for the traveling salesman problem with a center determining the qualitative behavior in each case.

2.1 Introduction

Blood is almost as important to a hospital's surgical operations as it is to a human being's operations. Certain surgical cases, such as a cardiac bypass, require several units of blood. To maintain appropriate levels, blood collection vehicles visit blood donation sites to pick up blood throughout the day. If the hospital's stock of blood falls too low, the blood collection vehicle may need to make an emergency visit to the hospital to drop off units of blood. Thus, the driver has two different and competing objectives: the efficiency of the route and the potential need to return to the hospital quickly at any point in time.

The standard formulation of the traveling salesman problem (TSP) attempts to minimize the sum of the Euclidean distances of a Hamiltonian cycle over a set of nodes. Essentially, the goal of the standard TSP is to visit every node while minimizing the total distance traveled [7]. In this paper, we will explore some of the properties of a TSP variant, the TSP with a center (TSPwC). In the TSPwC, there is a bi-objective function: the weighted sum of the tour's Euclidean distance and the distance from the "center" (using some metric). Thus, the general form of the problem on a set of nodes I will be Min $D = (1-\omega)\sum_I d_e + \omega \cdot \sum_I d_c$, where d_e is the Euclidean distance defined in equation (2.1), ω is some number in $[0, 1]$ serving as an affine weight, and d_c is the distance from the center (as defined by some metric) [6]. We will analyze two different ways of measuring the distance from the center: the midpoint distance as defined in equation (2.2) and the nearest point distance as defined in equation (2.3). The aim of the TSPwC is to find the tour that minimizes the bi-objective function D.

In this paper, we characterize the behavior of the TSPwC in several different cases. We provide qualitative results in the third section by comparing the two metrics for the biobjective function: the midpoint and the nearest point distances. We look at cases with different shaped distributions of the nodes and various locations for the center. In the fourth section, we discuss how the weight ω can be used to change the qualitative behavior of the tour to fit various problems. We provide a concrete example in Section 2.5 of an application of the TSPwC to blood donation collection. We also extend the formulation to multiple centers in Section 2.6 and discuss additional uses for this structure in Section 2.7.

2.2 Methodology

Lipowski and Lipowska [5] looked at the TSPwC using the distance from the depot to the midpoint of the line segment between two nodes as the second portion of the objective function. We will denote the coordinates of the i^{th} node as (x_i, y_i). c will refer to the index of the center with coordinates (x_c, y_c). d_e, d_m, and d_n are the Euclidean (equation (2.1)), midpoint (equation (2.2)), and nearest point distances (equation (2.3)) respectively. We have

$$d_e(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad \text{and} \tag{2.1}$$

$$d_m(i, j, c) = \sqrt{\left(\frac{x_i + x_j}{2} - x_c\right)^2 + \left(\frac{y_i + y_j}{2} - y_c\right)^2}. \tag{2.2}$$

We compare the results of this midpoint distance to the nearest point distance on several

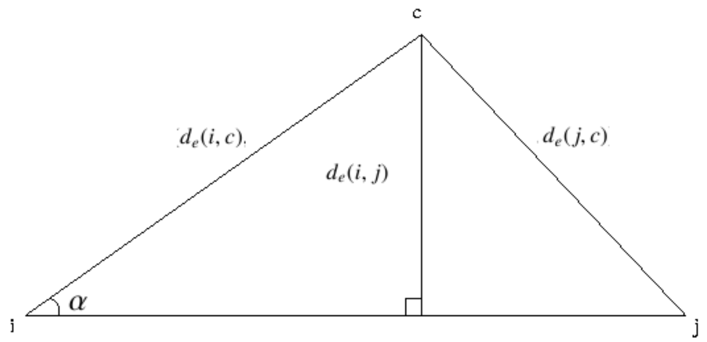

Figure 2.1: The nearest point distance between the center, c, and the line segment between points i and j is $d_n(i, j, c)$.

distributions with a variety of centers. The nearest point distance metric measures the minimum distance of any point on the line segment connecting two nodes from the center.

$$d_n(i,j,c) = \begin{cases} \min(d_e(i,c), d_e(j,c)) & \text{if } \max\left(d_e(i,c), d_e(j,c)\right) > d_e(i,j) \\ d_e(i,c)sin\left(cos^{-1}\left(\frac{d_e(i,c)^2 + d_e(i,j)^2 - d_e(j,c)^2}{2d_e(i,c)d_e(i,j)}\right)\right) & \text{otherwise.} \end{cases} \tag{2.3}$$

We derived the nearest point distance using the Law of Cosines and some basic trigonometry. Looking at Figure 2.1 and applying the Law of Cosines, we have:

$$cos(\alpha) = \frac{d_e(i,c)^2 + d_e(i,j)^2 - d_e(j,c)^2}{2d_e(i,c)d_e(i,j)} \tag{2.4}$$

From the right angle in Figure 2.1, we know:

$$d_n(i,j,c) = d_e(i,c)sin(\alpha) = d_e(i,c)sin\left(cos^{-1}\left(\frac{d_e(i,c)^2 + d_e(i,j)^2 - d_e(j,c)^2}{2d_e(i,c)d_e(i,j)}\right)\right) \tag{2.5}$$

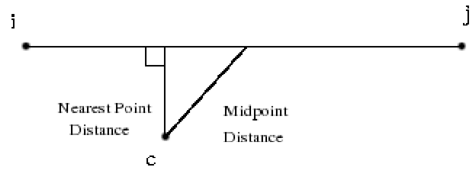

Figure 2.2: A comparison of the nearest point and midpoint distances.

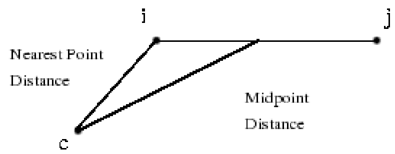

Figure 2.3: A comparison of the nearest point and midpoint distances in a different configuration.

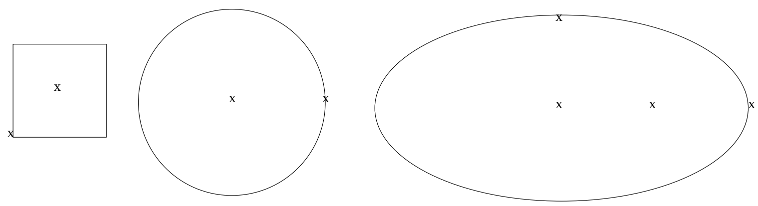

Figure 2.4: These are the different shapes used for the distributions. In each of these, x marks the locations of the different centers used in the tests.

Looking at Figures 2.2 and 2.3, it should be clear that in all situations $d_n \leq d_m$.

In order to explore the behavior of the TSPwC and compare the two distance metrics, 100 points were randomly distributed and a center was specified. It is important to note that the term "center" does not necessarily mean the geometric center. Instead, we use "center" to refer to the point (x_c, y_c) from the equations above. It should be thought of as a home base. As seen in Figure 2.4, the points were distributed in regions of different shapes with different centers. We tested points distributed in the unit square with centers located at $(.5, .5)$ or $(0, 0)$, the unit circle with centers $(0, 0)$ or $(0, 1)$, and an ellipse (with a semiminor axis of length one and a semimajor axis of length two) with centers at $(0, 0)$, $(1, 0)$, $(0, 1)$, or $(0, 2)$. For each shaped region, ten distributions of 100 points were generated. On each set of points, the compound objective function was tested using both the midpoint distance and the nearest point distance and values of ω ranging from zero to one. LKH-1.3, the Lin-Kernighan solver by Keld Helsgaun, was used to find a near optimal solution to the problem [8]. The results of these tests are discussed in sections 2.3 and 2.4.

2.3 Distance Results

In some sense, minimizing the Euclidean distance is antithetical to minimizing the distance from the center. Both center distance metrics essentially drive the solution to oscillate across the center (edges that cross the center have much lower values for d_m and d_n than those that do not come near the center), while an optimal Euclidean tour never crosses itself. Thus, there is a push and pull at play with the weight, ω, serving to determine the qualitative behavior of the tour. The graphs of the tours resulting from each of the distance metrics with respect to the weight ω (Figures 2.5 and 2.6) indicate a phase shift or inflection point around $\omega = \frac{2}{3}$. The change in the qualitative nature of the tours produced as ω varies can be seen in Figures 2.7 through 2.13. Further examination of the inflection point is discussed in the next section. Some of the characteristics of the inflection point for the midpoint distance TSPwC were discussed in the Lipowski paper [5]. Though it is important to note that their work only looked at the midpoint distance with points randomly distributed uniformly on the unit square with the center located at the geometric center $(.5, .5)$.

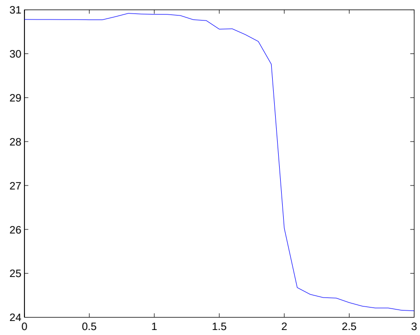

Figure 2.5: The value of the nearest point distance as ω varies. It is important to note the apparent inflection point around $\omega = \frac{2}{3}$.

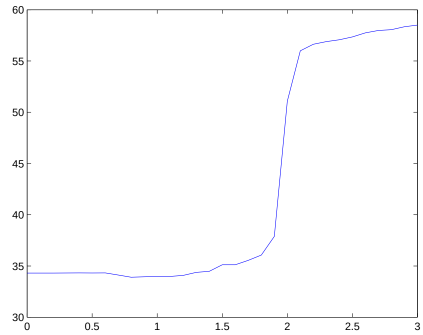

Figure 2.6: The value of the Euclidean distance as ω varies. Here too, there appears to be an inflection point around $\omega = \frac{2}{3}$.

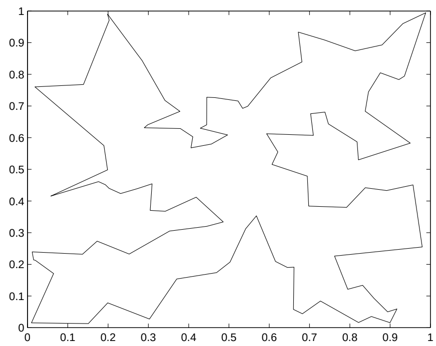

Figure 2.7: Euclidean tour ($\omega = 0$)

2.4 Inflection Points

A qualitative analysis of the graph of tour length versus ω (see Figures 2.5 and 2.6) indicates what appears to be an inflection point located near $\omega = \frac{2}{3}$. As shown in Figures (2.5) and (2.6), the qualitative nature of the tour plots change as ω varies. For values of ω

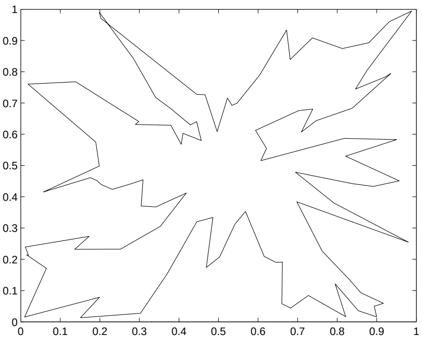

Figure 2.8: nearest point; ω = .6

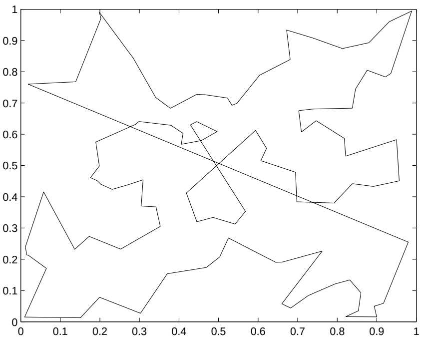

Figure 2.9: midpoint; ω = .6

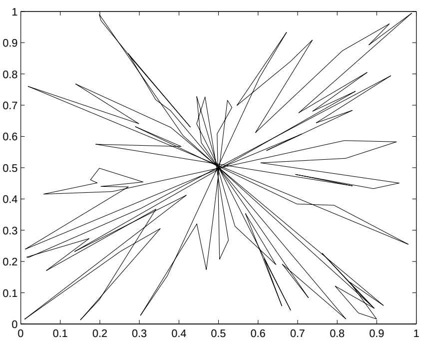

Figure 2.10: nearest point; $\omega = \frac{2}{3}$

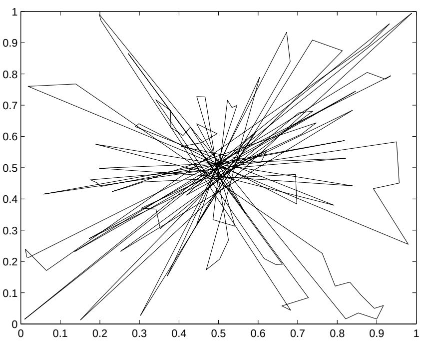

Figure 2.11: midpoint; $\omega = \frac{2}{3}$

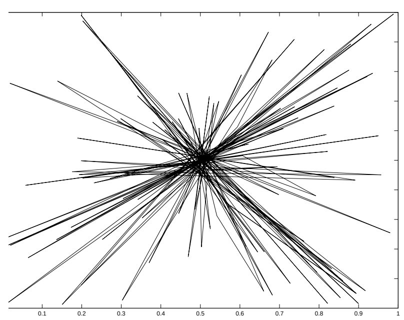

Figure 2.12: nearest point; ω = .75

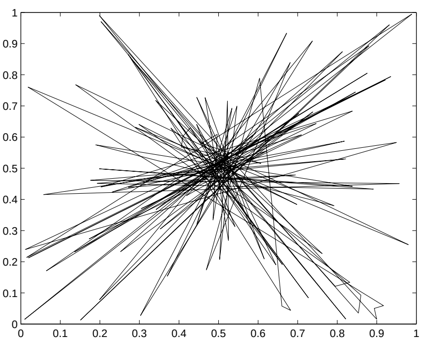

Figure 2.13: midpoint; ω = .75

less than two thirds, the paths rarely cross (an optimal Euclidean TSP tour will never have intersecting edges). As ω grows larger, an increasing number of intersections occur. For very high values of ω (i.e., $\omega \geq 3/4$), the only edges that do not intersect near the center are those that are nearly collinear with the center.

The inflection point in tour length as a function of ω can be thought of as dividing the set of solutions into two categories. To the left of the inflection point (roughly the interval [0,0.66)), minimizing the Euclidean distance portion of the objective function dominates the aim of minimizing the distance to the center. To the right of the inflection point (approximately the interval (0.66,1]), the objective function is dominated by the goal of minimizing the center distance. This can be seen in the qualitative similarities among the graphs of the tours on either side of the inflection point. Knowing the qualitative behavior of the solution can accelerate the process of finding a solution by suggesting an initial solution to seed the solver. Thus, knowledge of the location and the nature of the inflection point can be particularly useful when working with iterative heuristics.

To more accurately determine the location of this inflection point, we estimated the second derivative of the function with a Taylor series approximation at two points in equation (2.6). Here, $D_e(\omega)$ is the total Euclidean distance of the TSPwC tour using weight ω and h is the step size. For the tests, we used $h = .01$, $\omega_1 = .65$, and $\omega_2 = .67$

$$D_e''(\omega_i) = \frac{D_e(\omega_i - h) - 2D_e(\omega_i) + D_e(\omega_i + h)}{h^2} + O(h^2) \tag{2.6}$$

We estimated the inflection point using the linear interpolation formula in equation (2.7). We will use ω^* for the estimated inflection point.

$$\omega^* = \omega_2 - D_e''(\omega_2)\frac{\omega_2 - \omega_1}{D_e''(\omega_2) - D_e''(\omega_1)} + O((\omega_2 - \omega_1)^2) \tag{2.7}$$

Shape	center	nearest point average	midpoint average
Square	(0,0)	.6619	.6729
Square	(.5,.5)	.6617	.6634
Circle	(0,0)	.6608	.6694
Circle	(0,1)	.6608	.6780
Ellipse	(0,0)	.6607	.6549
Ellipse	(1,0)	.6608	.6403
Ellipse	(0,1)	.6599	.6584
Ellipse	(0,2)	.6601	.6495

Table 2.1: The estimated value of ω^*, the inflection point, for the various cases.

Table 2.1 indicates the inflection point is consistently around .66 for the nearest point distance metric. However, the location of the inflection point was slightly more varied for the midpoint metric but it was generally near .66. A value greater than .5 should be expected for the inflection because the mean midpoint and nearest point distances are generally smaller than the mean Euclidean distance for a collection of points.

2.5 Blood Collection Example

Now we will use the TSPwC framework to route a blood collection vehicle. The driver begins at the hospital, visits ten donation sites, and then returns to the hospital. In this example, the hospital will be located at (.5,.5). If there is an emergency demand at the hospital, the driver would like to be able to deviate from the tour and get to the hospital as quickly as possible. In an emergency, the response time is a key factor. We will model this as a TSPwC using the hospital as the center. The results in Tables 2.2, 2.3, and 2.4 were obtained by randomly generating ten points in a unit square and raising ω from 0 to 1. 100 different random sets were tested and the results reported are the mean values.

Table 2.2 shows the total distance traveled and the mean emergency response time for each objective function with a variety of ω values. Clearly, the maximum response time is fixed at the distance of the node farthest from the center. If an emergency occurs when the driver is at point (x^*, y^*) along the edge between i and j, the response time will be the distance from the center, $\sqrt{(x^* - x_c)^2 + (y^* - y_c)^2}$. Because any point on the line can be represented as an affine weight of the end points, we can write the response time as $\sqrt{(\gamma x_i + (1-\gamma)x_j - x_c)^2 + (\gamma y_i + (1-\gamma)y_j - y_c)^2}$. Thus, for a tour S, we have a mean response time (MRT) of:

$$\text{MRT} = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \sqrt{(\gamma x_{s(i)} + (1-\gamma)x_{s(i+1)} - x_c)^2 + (\gamma y_{s(i)} + (1-\gamma)y_{s(i+1)} - y_c)^2}d\gamma \quad (2.8)$$

The results found in Table 2.2 indicate that the response time can be reduced by about 25%, but at the cost of nearly doubling the tour length. The evaluation of the integral in equation (2.9) can be found in Appendix A.

The data in Table 2.2 do not tell the whole story. When the vehicle is forced to return to the hospital, trip length increases. Tables 2.3 and 2.4 show the quartiles for the response times and the average additional cost of travel for a trip to the center as ω varies. The column labeled return trip cost is the average additional distance traveled per trip to the center.

ω	tour length (MD)	response time (MD)	tour length (NPT)	response time (NPT)
0.0	3.12	0.35	3.12	0.35
0.1	3.13	0.35	3.13	0.35
0.2	3.14	0.35	3.13	0.35
0.3	3.18	0.35	3.15	0.35
0.4	3.21	0.35	3.18	0.35
0.5	3.34	0.34	3.27	0.35
0.6	3.80	0.32	3.63	0.33
0.7	5.24	0.28	5.05	0.29
0.8	5.52	0.27	5.66	0.27
0.9	5.65	0.27	5.69	0.28
1.0	5.70	0.27	5.74	0.27

Table 2.2: The mean tour length in Euclidean distance and mean response time for different values of ω for the midpoint distance (MD) and the nearest point distance (NPT).

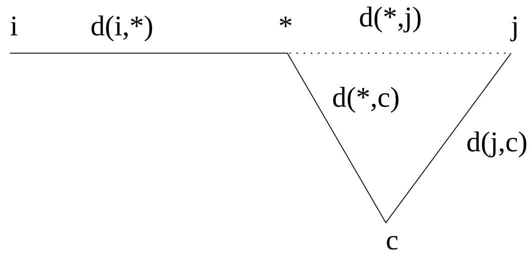

Figure 2.14: If the driver is notified of an emergency at point $*$, the additional distance traveled will be $d_e(*, c) + d_e(j, c) - d_e(*, j)$.

The return trip cost for a point (x^*, y^*) along the edge (i, j) is the response time distance plus the distance from the center to the next node in the trip minus the remaining distance along the edge (i, j) (that is to say, on average, the additional time added to the tour per return trip will be $d_e(*, c) + d_e(j, c) - d_e(*, j)$). The quartiles of the distribution of response time provide a better look at the advantage higher values of ω have when seeking to minimize distance from the center.

If there are two or more emergencies on an average day, the total trip length is less than 50% longer and the response time is generally between $20 - 30\%$ faster using an $\omega = 1$

ω	Tour Length	Return Trip Cost	25% RT	Mean RT	75% RT
0.0	3.0187	0.5905	0.3273	0.3568	0.3708
0.1	3.0404	0.5867	0.3257	0.3541	0.3694
0.2	3.0828	0.5841	0.3239	0.3537	0.3692
0.3	3.1023	0.5832	0.3226	0.3536	0.3692
0.4	3.1886	0.5781	0.3199	0.3529	0.3688
0.5	3.2510	0.5747	0.3181	0.3526	0.3687
0.6	3.6784	0.5384	0.2941	0.3377	0.3612
0.7	5.1930	0.4178	0.2368	0.2927	0.3387
0.8	5.2841	0.4086	0.2317	0.2887	0.3367
0.9	5.3915	0.4038	0.2302	0.2881	0.3364
1.0	5.6127	0.3884	0.2264	0.2844	0.3346

Table 2.3: Tour length and the average cost in distance per return trip to the center as a function of omega for the midpoint distance metric. RT is the response time. 25% RT and 75% are the first and third quartiles for the response time, respectively.

instead of $\omega = 0$. When selecting the "best" value for ω (hence, the best route), a driver will need to carefully determine the relative value of trip efficiency over the ability to quickly return. Tables 2.2, 2.3, and 2.4 provide an impression of the tradeoffs involved in adjusting the value of ω. If the mean response time must be below a certain threshold, these tables provide an indication of the trade-offs needed to achieve this.

2.6 Conclusion

We have presented an analysis and comparison of both the midpoint and nearest point objective functions for the TSPwC. As seen in the Figures 2.7 - 2.13, for high values of ω, the nearest point distance function provides a "tighter" tour than the midpoint method. Intuitively, the nearest point comes to mind when minimizing the distance from the cen-

ω	Tour Length	Return Trip Cost	25% RT	Mean RT	75% RT
0.0	3.0187	0.5905	0.3273	0.3568	0.3708
0.1	3.0825	0.5847	0.3255	0.3541	0.3694
0.2	3.1173	0.5811	0.3229	0.3523	0.3685
0.3	3.1688	0.5784	0.3221	0.3522	0.3685
0.4	3.1713	0.5777	0.3208	0.3514	0.3681
0.5	3.2061	0.5743	0.3197	0.3499	0.3673
0.6	3.7547	0.5272	0.2929	0.3308	0.3575
0.7	5.0145	0.4248	0.2426	0.2908	0.3378
0.8	5.4846	0.3904	0.2288	0.2800	0.3324
0.9	5.7023	0.3765	0.2262	0.2780	0.3314
1.0	5.7234	0.3760	0.2248	0.2765	0.3306

Table 2.4: Tour length and the average cost in distance per return trip to the center as a function of omega for the nearest point distance metric. RT is the response time. 25% RT and 75% are the first and third quartiles for the response time, respectively.

ter is a component of the objective function. The selection of the objective function and the value of the weight depends on the application desired. Because there is an inflection point in the Euclidean distance as a function of the weight, the qualitative behavior can vary dramatically as the weight changes. Some applications might require a tour with a different qualitative behavior. Understanding the inflection points is key to manipulating the qualitative behavior and determining the cost involved with these changes. The blood collection application discussed in Section 2.6 shows the qualitative structure's impact on other important factors such as response time.

Chapter 3

The Target Visitation Problem and Distribution of Emergency Supplies

3.1 Background

In this chapter, we will discuss the target visitation problem (TVP), a variant of the traveling salesman problem (TSP) which prioritizes some nodes or targets. The TVP has applications to the delivery of emergency supplies. When delivering emergency supplies, not all the destinations are equally urgent. For example, after an earthquake, the area around the epicenter will be much more damaged and in need of quick delivery of emergency supplies. The periphery of the effected area will be substantially less effected and will not require emergency supplies immediately. When practical, more urgent locations should be visited before the less urgent locations. The TVP seeks to balance the need to visit the emergency sites quickly with the desire for efficiency.

In the next section, we will discuss the formulation of the TVP. The third section contains an investigation into the balance between the two portions of the objective function.

The fourth section covers some real world type applications to emergency response. An alternate formulation for routing based on priorities is discussed in the fifth section.

3.2 Formulation

The TVP is a hybrid of the TSP and the linear ordering problem (LO). Grundel and Jeffcoat [9] first introduced the TVP as an approach to ordering targets for an unmanned aerial vehicle (UAV). They used a greedy randomized adaptive search procedure (GRASP) to find tours. Arulselvan et al. [10] also investigated the TVP but they used a genetic algorithm to construct tours.

The TSP seeks to find a tour on a set of nodes that minimizes the total distance traveled [7]. Assume there are n locations that are in need of emergency supplies. Let (x_i, y_i) be the location of the i^{th} emergency delivery point. For our simplified case, the distance traveled between points i and j will be the Euclidean distance:

$$d_e(i, j) = d_e(j, i) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

This is a symmetric distance metric.

The LO problem seeks to order a set of nodes based on preferences. For any two nodes i and j, there is a preference weighting p_{ij} for visiting i before j and p_{ji} for visiting j before i. The LO problem seeks to order the nodes in such a way that the sum of the preferences is maximized. There is no requirement that the nodes be well ordered. That is to say, $p_{ij} > p_{ji}$ and $p_{jk} > p_{kj}$, do not necessarily imply $p_{ik} > p_{ki}$. To reconcile this, the earlier papers assumed there is a set of experts that can vote to determine if $p_{ij} > p_{ji}$. Essentially p_{ij} is the number of experts that believe i is more important than j. The analytical hierarchy process (AHP) [11] could be used to create an ordering on the system. AHP constructs a well ordered set with relative weights from a set of pairwise comparisons. Thus, the expert opinions can be used in AHP to create a set of values u_i for each node i. Because we are looking at emergency supply delivery not UAV routing, we will use a different approach to order the nodes.

Each delivery point i can be assigned a level of urgency u_i. This urgency could be a measure of the remaining supplies, the number of people relying on the supplies, the population acuity, etc. The set of weights could also be developed using AHP. Without loss of generality, we will assume u_i is a number in $[0, 1]$, where lower values correspond to higher priority nodes (this may seem counter-intuitive given the examples above, but it will become clear later). Instead of a normal concept of distance, we will use a priority metric between nodes i and j. We would like the priority of any ordering of the nodes to depend on the relative urgency of the nodes. That is to say, we want the value associated with a tour to be a function of the order in which the nodes are visited. With this in mind, each pair of nodes (i, j) will have two edges (one from i to j the other from j to i) each with different values. Essentially, if i is more urgent than j there should be a penalty for visiting j before i. This penalty should also be in some way proportional to the relative values of i and j.

That is to say, the penalty for visiting j before i should be larger if $u_i \ll u_j$ than if $u_i \simeq u_j$.

Taking these goals into account, we suggest the following priority metric:

$$d_p(i, j) = 1 - d_p(j, i) = \frac{u_i}{u_i + u_j}.$$

Clearly this is not symmetric. If the distance traveled was irrelevant, the optimal strategy would be to visit the nodes in order of priority.

We will use a small example to illustrate the use of this metric. Assume there are four nodes 1-4, such that $u_1 = .1$, $u_2 = .2$, $u_3 = .4$, and $u_4 = .8$. The priority distance matrix, D_p, would be:

$$D_p = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{5} & \frac{1}{9} & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} & \frac{1}{5} & 0 \\ \frac{4}{5} & \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ \frac{8}{9} & \frac{4}{5} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where the fifth column and row represent the depot. If our only consideration is the priority distance, the optimal tour would be $depot \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow depot$, with a cost of $0 + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + 0 = 1$. Any other tour would use an edge below the main diagonal which would result in a tour with a value greater than one (implying the suggested tour is optimal).

If the location of each delivery point is within the unit square, $\sum_{i,j} d_p(i, j)$ will be roughly the same order of magnitude of $\sum_{i,j} d_e(i, j)$. These two metrics can be put together to form the TVP objective function:

$$L(\omega) = \text{Min}_S \sum_{i=1}^{n} \Big((1-\omega)d_e(s_i, s_{i+1}) + \omega d_p(s_i, s_{i+1})\Big). \tag{3.1}$$

In this equation, S is a tour over the nodes such that s_i is the i^{th} node along the tour.

3.3 Inflection Point

Using the same methodology as the work on the traveling salesman problem with a center (TSPwC), we looked for an inflection points in the target visitation problem. As in the earlier work, an inflection point, ω^*, of a bi-objective problem can be thought of as the point where one portion of the objective function is no longer dominated by the other. Unlike with the TSPwC, the two portions of the objective function are not in clear opposition (as long as the position and urgency are uncorrelated). Because the priority and Euclidean metrics are not in direct competition, it seems reasonable to expect that there may not be a well-defined inflection point. To test if there was an inflection point, we took 100 randomly distributed points with random priorities, solved the resulting TVP using the asymmetric TSP solver in LKH [8], and estimated the inflection point. We repeated this process ten times. The results of these experiments can be found in Table 3.1.

Unlike for the TSPwC, the TVP does not appear to have a consistent inflection point. All of the values are in the higher range of possible ω ($\omega > .8$). This could indicate that the priority weight does not significantly alter the tour until it is weighted substantially more

Run	Estimated Inflection Point (ω^*)
1	.9895
2	.9754
3	.8162
4	.9677
5	.9713
6	.9496
7	.9515
8	.9316
9	.8874
10	.9072
Average	.9347

Table 3.1: This table contains estimations of the inflection point for ten different distributions of 100 points.

than the Euclidean portion.

One possible reason for the size of ω^* is the relative gain from a swap. If two points, i and j, have very similar urgency, $d_p(i, j) \approx d_p(j, i) \approx \frac{1}{2}$. On the other hand, if $u_i \gg u_j$, $d_p(i, j) \approx 0$ and $d_p(j, i) \approx 1$. Thus the difference between a "good" edge ($u_i \approx u_j$) and a "bad" edge ($u_i \gg u_j$) is about .5. On the other hand, the difference between a "good" Euclidean distance edge ($d_e(i, j) \approx 0$) and a "bad" Euclidean distance edge ($d_e(i, j) > 1$) is more significant. Essentially, it is not the range of values that creates the imbalance (the ranges are both similar) it is the distribution of the values within this range.

3.4 Sample Applications

While there are certainly cases where the position and urgency would be uncorrelated, it may also be useful to correlate a node's location with its urgency. The distribution of the urgency could be correlated with the position to mimic the effects of certain types of natural disasters. For instance, earthquakes have an epicenter and the urgency could be a function of the distance from the epicenter. Hurricanes move across a path and the urgency could be a function of the distance from that path. A tsunami might start at the coast causing great devastation, but the damage decreases farther inland. Each of these disasters has a different pattern of destruction and the urgency generally correlates with the destruction.

3.4.1 Earthquake

When an earthquake occurs, the damage expands radially from the epicenter and decreases as a function of the radius. If the urgency is directly proportional to the damage and the epicenter is taken to be (.5, .5), then the urgency could be defined as:

$$u_i = \sqrt{(x_i - .5)^2 + (y_i - .5)^2}.$$

Figure 3.1 shows the distribution of u_i. Table 3.2 contains the value of the inflection point. The inflection point for the earthquake scenario is less variable than the uncorrelated case and, on average, lower than the uncorrelated case. This is likely because “good” edges in terms of the priority metrics are more likely to be relatively close together. Thus, the competition between the two portions of the objective function does not significantly alter the tour until ω is large. Furthermore, because the two parts of the biobjective function are correlated the transition happens in a tighter range than the uncorrelated case. Figure 3.2 shows a sample tour for $\omega = 1$. In this case, the construction of the optimal tour orders the

Run	Estimated Inflection Point (ω^*)
1	.9144
2	.8615
3	.8956
4	.8724
5	.8702
6	.8691
7	.8693
8	.8706
9	.8715
10	.8919
Average	.8787

Table 3.2: This table contains estimations of the inflection point for ten different distributions of the earthquake scenario with 100 points.

nodes by the distance from the origin.

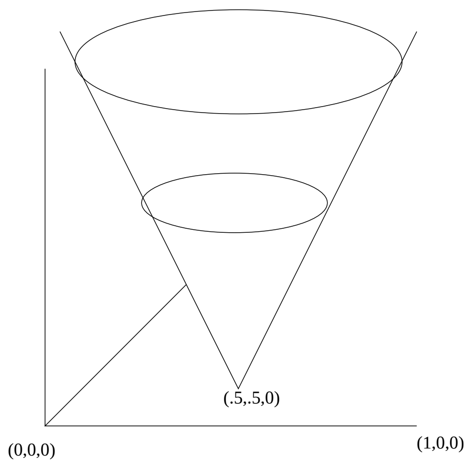

Figure 3.1: A representation of the earthquake urgency function as a function of x and y.

3.4.2 Hurricane

When hurricanes carve a path of destruction through an area, the devastation is worst near the path of the eye and decreases farther from the path. If the eye travels across the

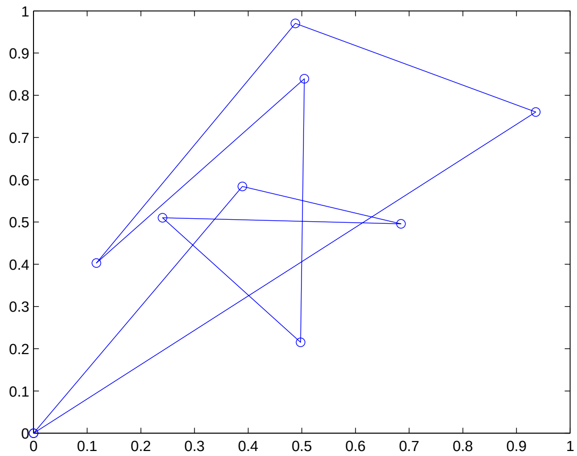

Figure 3.2: A sample tour of region devastated by an earthquake. Here the origin is the depot and $\omega = 1$. The urgency is proportional to the distance from the center. Thus, the optimal tour with $\omega = 1$ visits those nodes closest to the center first and then moves outward.

unit square through $x = .5$, the urgency could be:

$$u_i = |x_i - .5|.$$

Figure 3.3 shows the distribution of u_i. Table 3.3 contains the value of the inflection point. The inflection point for the hurricane scenario is less variable than the uncorrelated case and, on average, lower than the uncorrelated case. In this way, it is similar to the earthquake. Though, the value for the inflection point in the hurricane scenario does appear to be lower than the earthquake case. Figure 3.4 shows a sample tour for $\omega = 1$. In the optimal tour, the nodes are ordered by the distance from the line $x = .5$. This results in the tour oscillating across $x = .5$.

Run	Estimated Inflection Point (ω^*)
1	.8429
2	.8539
3	.8741
4	.8601
5	.8759
6	.8784
7	.8718
8	.8939
9	.8900
10	.8828
Average	.8724

Table 3.3: This table contains estimations of the inflection point for ten different distributions of the hurricane scenario with 100 points.

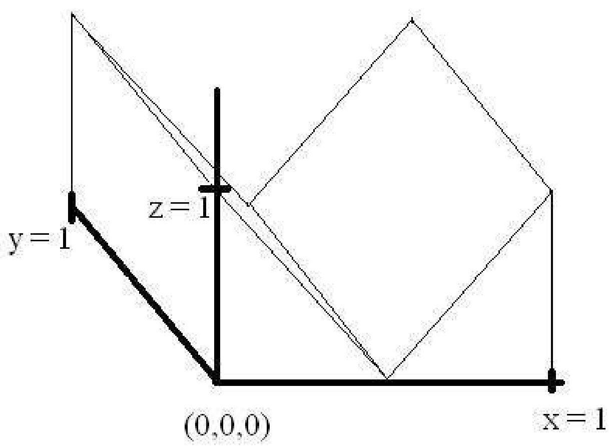

Figure 3.3: A representation of the hurricane urgency function as a function of x and y.

3.4.3 Tsunami

A Tsunami devastates the coastal areas but the impact decreases the farther from the shore. Thus, if the coast is the edge $x = 0$, the urgency could be:

$$u_i = x_i.$$

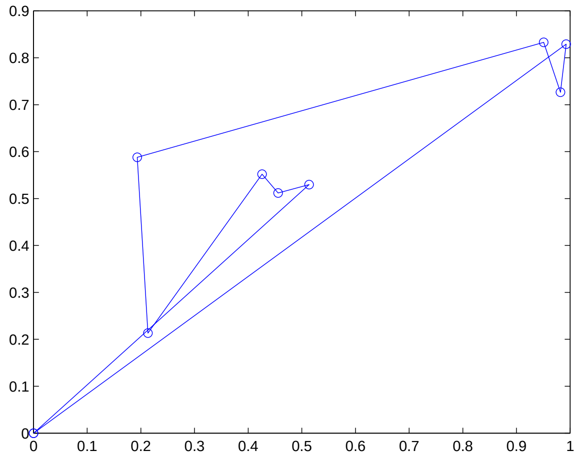

Figure 3.4: A sample tour of region devastated by a hurricane. Here the origin is the depot $\omega = 1$. The urgency is proportional to the distance from the line $x = .5$. Thus, the optimal tour with $\omega = 1$ visits those nodes closest to the line first and then moves farther from the line.

Figure 3.5 shows the distribution of u_i. Table 3.4 contains the value of the inflection point.

The inflection point for the tsunami scenario is less variable than the uncorrelated case and, on average, lower than the uncorrelated case. Figure 3.6 shows a sample tour for $\omega = 1$. The nodes in the optimal tour are ordered by the distance from the edge $x = 0$.

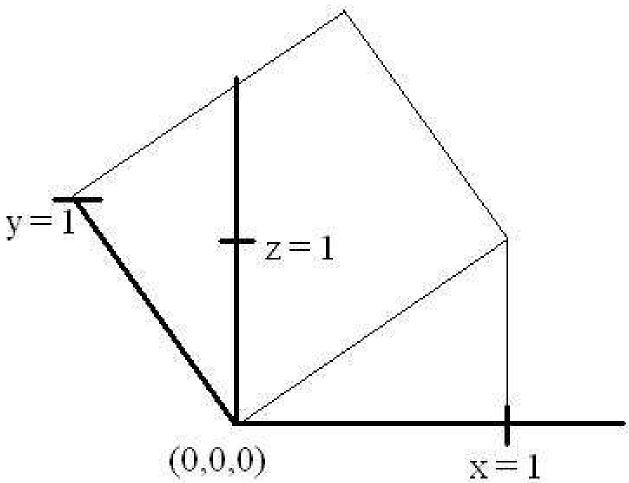

Figure 3.5: A representation of the tsunami urgency function as a function of x and y.

Run	Estimated Inflection Point (ω^*)
1	.9800
2	.8953
3	.9252
4	.9308
5	.9325
6	.9299
7	.9299
8	.9303
9	.9304
10	.9260
Average	.9310

Table 3.4: This table contains estimations of the inflection point for ten different distributions of the tsunami scenario with 100 points.

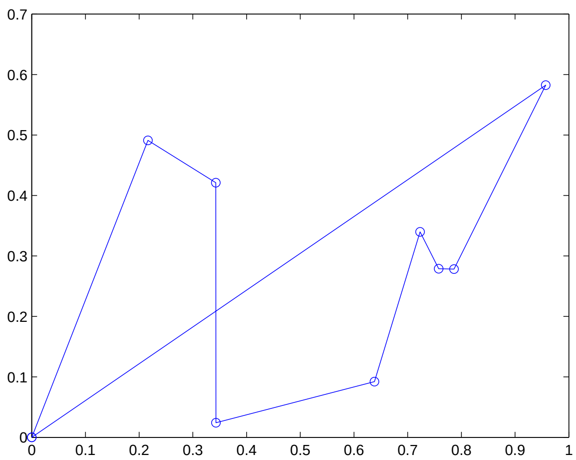

Figure 3.6: A sample tour of region devastated by a tsunami. Here the origin is the depot and $\omega = 1$. The urgency is proportional to the distance from the x-axis. Thus, the optimal tour with $\omega = 1$ visits those nodes with the smallest x-coordinate first.

3.5 Alternate Formulation

The formulations of the TVP discussed earlier might be too simplistic for practical use. When delivering emergency supplies, it is not the order of the deliveries that matter so much as the time at which the delivery is made. Essentially, if a driver is passing a low priority node on his way to a high priority node, in the real world, the delivery should be made to the low priority on the way instead of backtracking after the high priority node is visited. This is, of course, contingent upon the time required to make a delivery being negligible relative to the total travel time. With this critique in mind, an alternate way to formulate the objective function would be:

$$L(\omega) = \text{Min}_S \sum_{i=1}^{n} \Big(u_i \sum_{j=1}^{i} d_e(s_j, s_{j+1}) \Big), \tag{3.2}$$

where u_i is the urgency and $d_e(i, j)$ is the Euclidean distance between i and j. This formulation has the advantage that it factors in the time taken to reach a node instead of the order in which the nodes are visited. However, this formulation has the distinct disadvantage that ATSP software cannot be used to find tours.

Because this formulation cannot be solved using standard TSP solvers, we will test the solutions from the original TVP formulation on this alternate formulation. For these tests, we will randomly distribute twenty points in the unit square and randomly (uniformly) assign each point an urgency between zero and one. We will then produce a tour using the original formulation for different values of ω (0, .5, .8, .9 and 1). We repeated this ten times

ω	0.0	0.5	0.8	0.9	1.0
1	44.14	50.87	45.21	52.26	61.48
2	49.91	43.29	46.06	52.96	52.23
3	43.37	56.27	56.33	49.12	59.62
4	32.94	41.73	42.30	45.35	49.14
5	27.03	40.31	32.95	36.52	47.86
6	41.14	34.65	43.21	43.21	57.63
7	41.58	33.30	33.30	33.30	48.99
8	46.60	44.51	43.08	48.01	45.59
9	20.18	45.98	31.64	29.57	49.46
10	48.13	55.68	55.46	39.06	65.32
Average	39.50	44.66	42.96	42.94	53.73

Table 3.5: This table contains the value of the objective function from equation (3.2) using the tours generated using the objective function in equation (3.1).

and report the results below.

Table 3.5 has the value of equation (3.2) for ten different tours generated using different approaches. Clearly, for the objective function in equation (3.2), the tours generated using the standard TSP formulation ($\omega = 0$) performed better, on average, than the tours generated using equation (3.1) with $\omega > 0$. This might not be the case if the points were correlated or if the time spent at each node was not zero (ie there was a time penalty for visiting each node).

3.6 Conclusion

In this paper, we have investigated the TVP both on a theoretical and a practical level. The balance between the two parts of the objective function, the TSP portion and the LO portion, does not have a tight equilibrium like in the TSPwC. This equilibrium occurs when the LO portion of the objective function is weighted at least four times the TSP portion. Practically, this means the LO portion only dominates the TSP portion when the order is much more important than the distance.

We also looked into possible real-world applications for disaster response to regions struck by earthquakes, hurricanes, and tsunamis. In each of these cases the inflection point was more stable (the range of values was tighter), but still fairly high.

Next, we developed an alternate formulation. In a very practical sense, the goal in disaster response is to visit urgent targets quickly but that does not necessarily imply the most urgent site must be first. If a lower priority node is on the way to a high priority node, it might make sense to visit the lower priority node first. The TVP, as formulated in the literature, would penalize this behavior unnecessarily. Problems such as these highlight the need for decision makers to carefully evaluate their priorities. The approach taken or formulation used should reflect the objectives of the decision makers. To that effect, inflection points and alternative formulations should always be considered when approaching any problem in the real world.

Chapter 4

Maximizing Cardiac Surgery Throughput

4.1 Introduction

Efficient bed management and nurse resource allocation are challenges confronting every major medical center in the country. Hospitals are bursting at the seams as they attempt to manage increased patient demands in an era of nursing shortages and bed scarcities. Several of the nation's experts have tied the crisis in emergency room diversions to the downstream bottlenecks seen in intensive care units, step-down units, and medical surgical units (see [12] and [13]). However, building new inpatient capacity is an expensive option and the nursing shortage is predicted to become even more severe in the next decade. Research conducted in non-healthcare related industries (e.g., manufacturing, aviation, distribution) has demonstrated that throughput can be maximized by using mathematical modeling techniques. Historically, however, healthcare has not embraced these methodologies. One example is that bed allocation decisions are often made without scientific evidence and frequently rely on historical data and subjective anecdotes.

For surgical services in general and cardiac surgical services, in particular, throughput is often tied directly to the availability of downstream beds to care for and recover patients. Surgical schedules in major centers are often adversely affected due to limited downstream capacity which, in turn, constrains efficiency and throughput. Despite this impact, the determination of the appropriate mix of downstream beds is often made based on factors such as space availability, budgetary constraints, and historical practice.

Simulation is a powerful technique that can be used to model a wide range of problems. Making basic assumptions about the nature of the patients (e.g., patients treated in the future will have a similar length of stay distribution as those patients treated in the last two years), simulation allows decision makers to test a variety of scenarios (see [14] and [16]). In this paper, we present a simulation model of the post-operative bed flow for a cardiac surgery unit that maximizes the throughput using the current level of physical resources. Our model uses actual data from the cardiac surgery unit at the University of Maryland Medical Center. The cardiac surgery service line has annual surgical volume of nearly 1,000 patients.

In May, 2006, the University of Maryland Medical Center opened a new 30 bed unit for post-operative cardiac surgery patients. These beds were divided between two units: 15 beds for the intensive care unit (ICU) and 15 beds for the intermediate care unit (IMC, a step-down unit from the ICU). Because of nursing shortages, the minimum staffing requirements could not be met. In general, each ICU bed requires four nurses on staff per week and each IMC bed requires two nurses on staff per week. To conform to the available staffing levels at the time of this study, the bed mix was changed to 11 ICU beds and 18 IMC beds (one bed was closed).

The standard path for a patient is shown in bold in Figure 4.1; other paths are shown as well. After surgery, the patient is sent to the ICU. The patient recovers in the ICU with 24 hour-a-day nurse supervision. When the attending physician (usually the surgeon) determines intensive care is no longer medically necessary, the patient is moved to the IMC. The patient stays in the IMC until the patient is medically determined to be ready to go home or to a non-cardiac surgery ward. The cardiac service line performed 973 cases in fiscal year 2006 (July 1, 2005 to June 30, 2006). According to hospital administrators, the number of cases is expected to grow by about 13% in fiscal year 2007. However, there are serious obstacles to this growth. In particular, the bottlenecks in the beds could lead to greater friction in the system as the growth puts larger strains on scarce resources. Early in the summer of 2006, a cap was placed on the number of surgeries per day because of a lack of bed space in the ICU.

For the 2007 fiscal year, there are $9,840$ patient bed days budgeted. There is a total

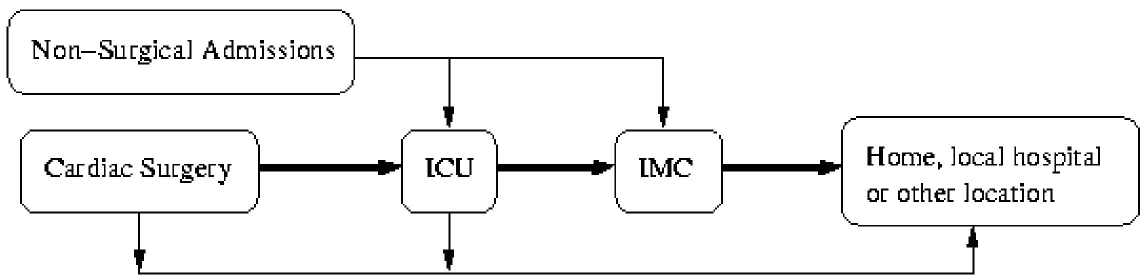

Figure 4.1: The paths that patients take through the system in the simulation model.

capacity of $365 \times 30 = 10,950$ bed days dedicated to cardiac surgery. While a buffer of roughly 1,000 bed days may sound reasonable, it masks two problems. First, the days are not uniformly distributed throughout the year. There are seasonal effects that influence the number of surgeries performed in a particular month. Thus, capacity issues from the high volume may arise (as it did in the summer of 2006 when surgeries had to be limited to three a day because of a lack of beds) and hold down the total number of surgeries. Second, given the cardiac surgery unit's expected annual growth rate of 13% and a constant average length of stay of 9.88 days (according to hospital administrators), fiscal year 2007 will require $9.88 \times 1.13 \times 973 = 10,863$ days. This is essentially at the level of total capacity (i.e., $10,863/10,950$ or 99.2% of capacity). Operating so close to capacity means that small fluctuations in volume or any decrease in capacity will stop the flow of patients into the system. This level of operation requires additional beds or a more efficient use of the current resources (see [17]).

We will investigate the effects of changes in the post-operative bed allocation on patient throughput. An analysis of the available data is provided in the next section. The third section contains a first shot at a model developed using queueing theory. The fourth and fifth sections cover an initial and more complex simulation model, respectively. The sixth section discusses the financial implications of the simulation results.

4.2 Data

The data set used in this study included information from the perioperative services department and the hospitals finance department. These data contained detailed information that included patient length of stay in every unit the patient visited and the number, type, and date of operation for every cardiac surgery patient for fiscal years 2005 and 2006. A total of 1,725 surgical operations were reported in the data.

These 1,725 operations were performed on 1,548 patients (in addition, 127 patients had no operations but spent time in either ICU or IMC; some of these patients had ventricular-assistance devices and needed occasional ICU visits). Thus, there were roughly 11.4% more operations performed than there were patients undergoing surgery. An analysis of this data set produced several useful descriptive statistics. Eighty-three patients ($83/1,548 = 4.96\%$) underwent some type of cardiac surgery, but spent no time in either the cardiac surgery ICU or IMC. Most of these patients either spent time in post-operative units elsewhere in the hospital or had only minor surgery. There were $1,548 - 83 + 127 = 1,592$ patients who used the cardiac surgery post-operative units. Therefore, on average, $1,725/1,592 = 1.085$ operations were performed for each patient that passed through the cardiac surgery ICU or IMC. In other words, there were 8.5% more operations than patients who passed through the cardiac surgery post-operative units since there were 177 multiple surgeries (some patients require multiple operations over the course of their stay). We derived empirical distributions for the length of stay in both ICU and IMC. The correlation between these two distributions was 0.2113. This is a weak positive correlation and we did not take it into account in any of the models.

We calculated the best fit distributions from the empirical distributions. ICU length of stay was lognormally distributed with a mean of 4.04 and a standard deviation of 4.91. IMC length of stay was lognormally distributed with a mean of 5.16 and a standard deviation of 4.31. Neither of these distributions were found to be sufficiently accurate by an Anderson-Darling test (at the $p = .05$ level or even the $p = .10$ level). These distributions had a substantial under-sampling from the tail of the distributions and skewed the estimations. In fact, roughly 5% of patients spend more than three standard deviations from the mean. For these reasons, we elected to use the empirical distributions instead of the fitted distributions for the simulations. In Figure 4.2, we show the distribution of the ICU length of stay; the IMC length of stay is displayed in Figure 4.3.

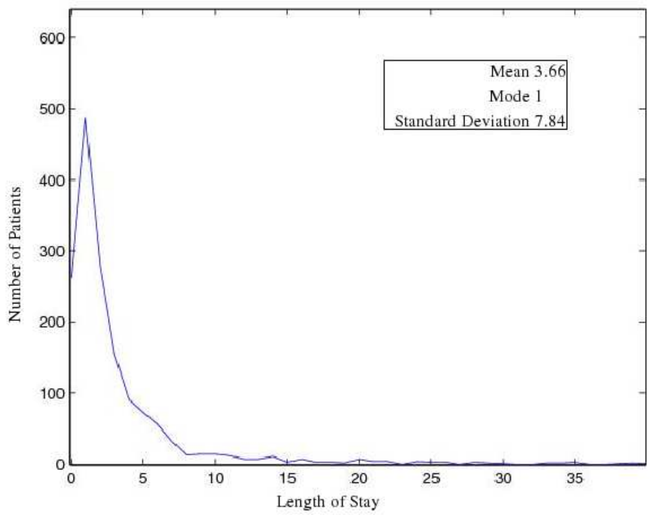

Figure 4.2: The distribution of the ICU length of stay.

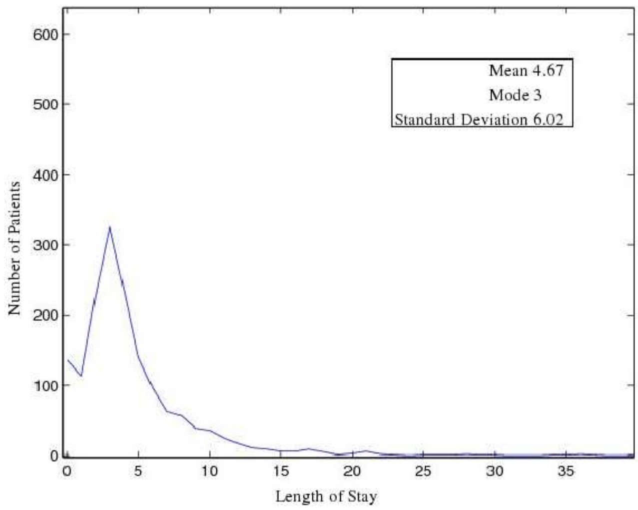

Figure 4.3: The distribution of the IMC length of stay.

4.3 Queueing Model

In order to better understand the nature of the flow through the system, a stochastic model was constructed that had simplifying assumptions to make the solution tractable. Because the model assumes that the ICU is always fully occupied, the ICU length of stay can be thought of as the arrival rate for the IMC, and the IMC length of stay is the service time for the system. It is also assumed that the time in the ICU and the time in the IMC are independent. If we assume that the time spent in the ICU and the IMC are exponentially distributed, we can model the system as an $M/M/c$ system.

Let m be the number of beds in the ICU, n be the number of beds in the IMC, λ be the length of stay for the ICU, and μ be the length of stay for the IMC. The state space for the system will be, $\{i \mid i \in (0, 1, ..., n)\}$, the number of patients in the system. The steady state solution, $\{\pi_i\}_{i=0}^n$, can be found using the flow equations:

$$\begin{aligned}
m\lambda\pi_0 &= \mu\pi_1 \\
m\lambda\pi_1 &= 2\mu\pi_2 \\
&\vdots \\
m\lambda\pi_i &= i\mu\pi_{i+1} \\
&\vdots \\
m\lambda\pi_{m-2} &= (n-1)\pi_{n-1} \\
m\lambda\pi_{m-1} &= n\pi_n.
\end{aligned}$$

Noting that $\sum_{i=0}^n \pi_i = 1$, we can solve the system of equations:

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^n \prod_{j=1}^i \frac{m\lambda(m\lambda+i\mu)}{i\mu(m\lambda+(i-1)\mu)}}$$

$$\pi_i = \pi_0 \prod_{j=1}^i \frac{m\lambda(m\lambda + i\mu)}{i\mu(m\lambda + (i-1)\mu)}.$$

Utilization and throughput are important for comparing bed mixes. If U is the utilization, we know $u = \sum_{i=0}^n \frac{i}{n}\pi_i$. Let $P(t)$ be the number of patients that pass through the system in t days, using Little's theorem we can determine $E(P(t)) = \frac{u \cdot n \cdot t}{\mu}$ $\forall i > 0$.

The results indicate that altering the bed mix can increase the throughput. The 12/18 and 13/17 mixes both increase by about 10 patients per quarter over the 11/18 mix that is currently in use. The length of stay distribution were clearly not exponential so there are

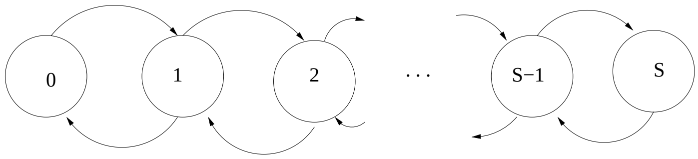

Figure 4.4: The queueing model of the cardiac surgery post-operative unit. Each state represents the number of patients currently in the IMC.

Bed Mix	Utilization	Throughput
11 ICU / 18 IMC	.7707	275.8087
12 ICU / 18 IMC	.8303	287.9818
13 ICU / 17 IMC	.8754	286.7469
14 ICU / 16 IMC	.9087	280.1546
15 ICU / 15 IMC	.9336	269.8393

Table 4.1: These are the results of the stochastic model for 13 weeks. Here throughput is measure in terms of patients per quarter.

some problems using this approach for policy decisions. This does indicate the potential benefits for using a simulation model.

4.4 Preliminary Simulation Model

4.4.1 Model

To determine the best balance between ICU beds and IMC beds, a simulation of a 13 week period (roughly three months) in the post-operative unit was performed. We used the empirical distributions for ICU length of stay and IMC length of stay in our simulation. In order to maximize the allowable utilization levels, an operation was allowed only if there was an empty bed in the ICU. Whenever a bed was open in the ICU, a surgery would be done. Thus, the ICU was completely utilized at all times. For this model, time spent blocked in the ICU did not reduce time spent in the IMC. Thus, if a patient spent four "extra" days in the ICU that patient would not expect a reduced length of stay. We used MedModel [15] to develop our simulation model. To ensure the model was in a steady state, the simulation had a 13-week warm-up period before the 13-week data collection period began. We averaged the results over 999 different simulations for the 13-week period.

4.4.2 Throughput Results

In Table 4.2, we report the number of patients admitted into the ICU. A patient admitted to the ICU was a proxy for an elective cardiac case being performed. Therefore, the number of ICU admissions equals the number of cardiac cases performed. The bed mixes reported are in the following form: number of the ICU beds/number of the IMC beds (e.g., 11/18 denotes 11 beds in the ICU and 18 beds in the IMC). The mean is calculated over the 13-week simulation period. The mean for 11/18 is roughly consistent with the current practice though slightly lower than the 2006 totals. Min 5% (Max 5%) indicates that the smallest (largest) 5% of the simulation results fell below (above) this level. For example, in Table 4.1 for the 11/18 mix, the minimum 5% indicates that 5% of the simulations resulted in fewer than 204 ICU entrances.

There were times when patient flow from the ICU beds to the IMC beds was blocked because the IMC was full. This is an important factor and it plays a role in determining the optimal bed mix. The values for the percentage of time blocked are given in Table 4.3.

The 14/16 case mix and the 15/15 case mix had roughly the same number of allowed cardiac operations (maximum number of admissions). From Table 4.3, the mix of 14/16 had a lower mean blocked percentage (4.44%) than the 15/15 mix (8.43%). In addition, the top 5% and the maximum time blocked for the 15/15 mix were more than 40% higher than the 14/16 bed mix (12.86/7.80 = 1.65 and 16.10/11.23 = 1.43). A large value for % time blocked indicates that patients are receiving ICU care when it is not medically necessary. Because ICU time is relatively expensive, this is inefficient and raises the overall cost of cardiac surgery. The 14/16 bed mix is preferable to both the 15/15 and 11/18 mixes. The 14/16 bed mix allows for a 17.4% increase in the number of operations allowed over the 11/18 bed mix and the 14/16 mix is only blocked 4.44% of the time, on average.

Bed Mix	11/18	12/18	13/17	14/16	15/15
Mean	230.04	248.31	262.88	270.20	268.39
St. Dev.	15.44	14.66	13.17	12.03	11.40
Min	175	202	213	227	227
Min 5%	204	224	240	250	249
Median	230	248	264	271	268
Max 5%	255	272	283	290	288
Max	278	291	299	305	300

Table 4.2: Simulation results for ICU throughput for 13 weeks. We assumed that when a room was available in the ICU, a surgeon would operate. The mean is the number of ICU admissions that were permitted during the simulation timespan. An ICU admission was allowed whenever there was an empty bed in the ICU. The bed mix A/B indicates that there were A ICU beds and B IMC beds.

4.4.3 Problems with the Model

Because in practice almost all patients spend time in the ICU and can spend their IMC recovery time in an ICU bed, it is counter-intuitive that switching an IMC bed to an ICU bed would reduce throughput. The initial simulation's throughput peaks at 14/16 and decreases

Bed Mix	11/18	12/18	13/17	14/16	15/15
Mean	0.33	0.69	2.04	4.44	8.43
St. Dev.	0.37	0.60	1.20	1.88	2.63
Min	0.00	0.00	0.00	0.38	1.51
Min 5%	0.00	0.03	0.45	1.65	4.16
Median	0.19	0.55	1.79	4.21	8.30
Max 5%	1.06	1.86	4.25	7.80	12.86
Max	2.55	3.75	8.76	11.23	16.10

Table 4.3: Simulation results for the percentage of patients blocked. A patient is blocked from moving to the IMC from the ICU when the IMC is full. A blocked patient will advance to the IMC when a bed is available.

at the 15/15 mix. This inconsistency is a direct result of the assumption that extra time spent in the ICU while blocked has no effect on the IMC length of stay. As seen in Table 4.3, more than 8% of the patients see the system blocked.

Additionally, with the 11/18 mix, the hospital has historically seen nearly a thousand patients which is more than the 920 patients per year predicted by the initial simulation. This implies there are some fundamental problems with this preliminary model, but the results are sufficiently promising to warrant further investigation.

4.5 Modified Simulation Model

4.5.1 Model

We corrected the problems from the preliminary simulation by changing the way we account for days spent blocked. In the next simulation model, every additional day spent in the ICU because of a full IMC was subtracted from the time needed to be spent in the IMC. In some cases, patients were discharged directly from the ICU and never passed the IMC because of blocking. This change was difficult to add to the MedModel simulation and so we constructed the simulation from scratch in MatLab. This provided us with much more flexibility and the model ran much faster.

As with the initial simulation, we ran this simulation for 13 weeks with a 13 week warm up period using the empirical distribution. The results presented are from 999 runs.

4.5.2 Results

Bed Mix	11/18	12/18	13/17	14/16	15/15
Mean	285.62	317.85	337.72	367.97	375.77
St. Dev.	35.14	30.26	32.83	35.12	40.42
Min	222	229	254	272	262
Min 25%	265	299	312	346	351
Median	282	318	343	372	382
Max 25%	305	334	360	390	406
Max	360	390	425	446	447

Table 4.4: Simulation throughput results with the assumption that every day spent blocked in the ICU reduces the IMC length of stay by one day.

Bed Mix	11/18	12/18	13/17	14/16	15/15
Mean	8.40	11.73	17.60	25.16	29.96
St. Dev.	5.24	5.20	5.89	6.04	6.90
Min	1.35	1.43	3.85	11.29	14.12
Min 5%	4.61	7.72	13.58	20.76	26.38
Median	7.91	11.39	17.68	25.29	30.46
Max 5%	11.76	15.46	22.13	29.34	35.19
Max	20.28	22.56	32.39	39.01	51.69

Table 4.5: Simulation results for the percentage of patients blocked using the final model. A patient is blocked from moving to the IMC from the ICU when the IMC is full. A blocked patient will advance to the IMC when a bed is available.

The results are reasonable on face. The throughput increases monotonically in ICU beds. The maximum throughput for the 11/18 mix is $4 \times 286 = 1,144$ patients per year, on average. As desired, this throughput is above the 2006 level of just under a thousand patients per year. This indicates that there are no problems with the model on face.

4.5.3 Constant Staffing Level Scenario

One of the primary problems with the current system is the lack of nurses. This makes the 15/15 mix, while the most desirable in terms of throughput, potentially unattainable in the near term. In general, an ICU nurse is responsible for one or two patients (depending on the acuity) and an IMC nurse is takes care of two to four patients at a time. For our purposes, we will assume that two IMC beds must be closed to open one ICU bed. To take the nurses into account, we performed the simulation for a second scenario with a constant level of nurses.

Bed Mix	11/18	12/16	13/14	14/12	15/10
Mean	285.62	317.38	335.77	342.83	325.35
St. Dev.	35.14	30.68	34.85	40.93	59.53
Min	222	232	244	246	237
Min 25%	265	295	310	318	275
Median	282	318	338	348	315
Max 25%	305	334	360	370	368
Max	360	387	409	429	439

Table 4.6: These are the simulation throughput results in the scenario with a fixed number of nurses.

4.6 Financial Results

As mentioned in the introduction, the target growth rate for the cardiac surgery service line was 13%. In Tables 4.8 and 4.9, there are several options that allow this growth rate to

Bed Mix	11/18	12/16	13/14	14/12	15/10
Mean	8.40	17.69	27.89	37.38	42.12
St. Dev.	5.24	6.08	7.07	7.73	9.00
Min	1.35	3.88	9.47	15.04	24.89
Min 5%	4.61	13.40	22.89	30.99	32.60
Median	7.91	17.28	27.95	38.78	40.19
Max 5%	11.76	22.12	32.12	43.17	49.09
Max	20.28	31.66	44.94	55.44	62.71

Table 4.7: The simulation results for percent of time spent blocked in the fixed nurse scenario.

be feasible for at least another year. The 15/15 mix allows for more than two years of 13% growth.

Assuming an ICU nurse costs the hospital a total of $\$100,000$ on average each year and each additional surgery provides the hospital a net revenue of $\$20,000$, switching to the $15/15$ bed mix could increase revenue by roughly $90 \times 4 \times \$20,000 = \$7.2$ million. If there are 8.5% more surgeries than cardiac surgery post-operative patients due to multiple surgeries and patients who do not use cardiac surgery post-operative units, the increase in revenue is $1.085 \times 90 \times 4 \times \$20,000 = \$7.81$ million. We estimate that eight nurses would be required for the new mix (four additional ICU beds require sixteen nurses and we can save six nurses from two fewer IMC beds). The cost is $10 \times \$100,000 = \$1$ million and the overall change in profit is $\$7.81$ million $-\$1$ million $= \$6.81$ million each year. Some portion of this money could be used to attract new nurses and retain other members of the team.

For the fixed nurse scenario, the 14/12 mix allows for an increase in the maximum throughput possible by about 229 patients per year. This could result in $1.085 \times 57 \times$

$4 \times \$20,000 = \$4.95$ million additional profit because there is no additional expense for personnel.

Bed Mix	12/18	13/17	14/16	15/15
Change in Mean	32.24	52.11	82.35	90.16
% Change	11.29	18.24	28.83	31.57
Change in Profit	$2,178,888.04	$3,568,591.01	$5,788,392.99	$6,212,551.41

Table 4.8: Some financial statistics for the fixed number of beds scenario using the assumptions that each additional surgery profits the hospital $\$20,000$ and each additional nurse costs the hospital $\$100,000$. These totals are for the year and so are four times higher than the quarterly results reported in previous tables. These do not include the 8.5% for multiple surgeries.

Bed Mix	12/16	13/14	14/12	15/10
Change in Mean	31.72	50.16	57.22	39.73
% Change	11.11	17.56	20.03	13.91
Change in Profit	$2,537,699.92	$4,012,551.41	$4,577,303.88	$3,178,492.00

Table 4.9: Financial statistics for the scenario with fixed number of nurses.

4.7 Conclusion

The simulation experiment indicated that by altering the bed mix, the cardiac surgery service line could increase throughput by as much as 32% and increase annual profit by about $6.2 million if additional nursing staff could be acquired. By switching to a more efficient arrangement, the hospital can increase capacity without constructing new ICU or IMC rooms. Without additional nurses throughput could be increased 20% with a profit increase of as much as $4.6 million.

Depending upon the immediate staffing availability and demand for beds, the hospital now uses 13 to 14 ICU beds staffed with 15 to 17 IMC beds.

This work can be applied to other hospitals and surgical service lines as long as representative length of stay is available.

Chapter 5

Cardiac Surgery Capacity Prediction

5.1 Background

Cardiac surgery is a very resource intensive service line. A standard cardiac surgery patient requires several hours in an operating room, a few days in an intensive care unit (ICU) bed, followed by a few additional days in an intermediate care unit (IMC) bed. Each of these stages of treatment requires staff, equipment, supplies, and rooms. Without accurate census estimates, resources can be unexpectedly over-utilized or under-utilized. Anticipating discharges is key to better managing these resources. In this paper, we will discuss approaches to predicting discharges in advance.

At the University of Maryland Medical Center (UMMC), like many hospitals, the lack of sufficient post-operative bed capacity creates a bottleneck that reduces the throughput of the cardiac surgery service line. At the time of this study, the cardiac surgery service line at UMMC had 12 beds devoted to the ICU. The cardiac surgery ICU (CSICU) operates near capacity and has been a bottleneck for the flow of patients. Cardiac surgery patients generally must pass through the CSICU because other ICU units in the hospital do not have nurses with the specialized training required to properly care for cardiac surgery patients. Thus, if the CSICU is full and there are no patients being discharged, new surgical cases may have to be canceled. This is clearly not desirable for the patient. Additionally, because cardiac surgery is a high dollar service line, disruptions in the flow of patients can result in significant reductions in revenue for both the hospital and the affected surgeons. The revenue from cardiac surgery is used to subsidize less profitable but still vital hospital services.

When thinking about bed management, there are two key terms: census and capacity. A unit's census is the number of patients in the unit. A unit's capacity is the number of staffed beds (a staffed bed is a physical bed that has sufficient nursing support to have a patient) in a unit. Discharge and census information is also important for staffing nurses. Improving the management the cardiac surgery post-operative beds is crucial to increasing throughput. Information regarding available capacity is key to improving management's decision making with regards to scheduling patterns. If there is not sufficient capacity, new cases cannot be scheduled.

Staffing patterns are made more than a week in advance and then are adjusted based on perceived need. These decisions are made after the expected discharge and arrival information has become available to the bed management center. Frequently, the expected evening census is not known until the mid-afternoon, because there was no good estimate for the number of discharges until then. When this happens, the nurse managers must scramble to find sufficient staff to cover the patient load. If the discharges could be estimated earlier, the staffing could be adjusted sooner, reducing mid-afternoon census surprises.

Hospital planners would like to accurately predict the number of beds that will be available in the near future (today, tomorrow, or two days from now). This would allow them to adjust staffing patterns in advance. If there is likely to be a capacity problem, cases could be rescheduled a day or two in advance instead of canceled on the day of surgery. In this chapter, we will explore a variety of approaches to predict patient length of stay and use these predictions to estimate CSICU census in the near future.

5.2 Relevant Literature

Farmer and Emami [18] use a time series to predict the future hospital census. They found the time series approach was less prone to model specification error and autocorrelation than simpler trend fitting approaches. The target staffing level is dependent upon the census and acuity of patients. This approach might be useful for the census but does not provide an idea of which patients are ready to be discharged or their acuity. Essentially, a time series approach assumes today's census can be predicted as a function of previous days' censuses. While this may be true, it omits relevant factors from consideration (age, comorbitities, etc) and does not provide substantial insight into the underlying system (at least when compared to more atomistic approaches that focus on the system's constituent parts).

Postma et al. [19] also use a time series approach, but they look at additional pieces of the system. Their model includes many details on the patient such as age and type of disease. They also break the data into age and disease cohorts. This process is essentially arbitrary, though the types of groups they selected seem reasonable. Like Farmer and Emami's findings, their results indicate that a time-series approach provides a good estimate for the census.

An alternative to regression-based estimates of available capacity would be to estimate individual patients' lengths of stay. Essentially, this would be a bottom up atomistic approach. There have been several studies using modern analytical techniques to estimate patient length of stay. In particular, there have been several papers using artificial neural networks (ANN) for this purpose. Zernikow et al. [20] compare an ANN and a multiple linear regression model to predict the length of stay in preterm neonates (babies born prematurely) using information from the first day of life. They find that the results from both methods were highly correlated with the actual length of stay. Because the data used for prediction only contained information available at the beginning of the stay, this work implies that these techniques can be useful in planning. Mobley et al. [21] use a variety of ANN implementations to estimate the length of stay for a post-coronary care unit. While the outputs of the models generally have mean absolute errors of less than two days, the average length of stay for these patients was 3.84 days. This level of accuracy is probably not sufficient to be useful for planning purposes.

5.3 Length of Stay Prediction

Our first approach to estimating future bed capacity relies on an accurate length of stay (LoS) estimate to determine when patients will be discharged. In this section, we will discuss several approaches to LoS prediction. If a patient's length of stay can be accurately determined, the available capacity will be the number of currently available beds plus the number of discharges. The number of discharges would be known because of the LoS predictions.

5.3.1 Data

The data used to estimate individual LoS span two years, covered 1,675 patients and 17,123 patient bed days (one bed day is one patient in a hospital bed for one day). Each patient spent some time in a cardiac surgery post-operative unit. Each patient bed day was labeled as intensive care unit (ICU), non-ICU, or intermediate care unit (IMC). Each of these labels indicates a level of care required by the patient. There were also floor beds, but capacity for these is not as seriously constrained and they are rarely used for cardiac surgery patients. Therefore, we will focus our attention on the ICU, NonICU, and IMC beds. For our analysis, we grouped the NonICU with the IMC.

Included in the data are the age, sex, race, level (emergent, express, or elective), and a list of operations for each patient encounter. There were over three hundred types of operations listed. Each operation type was pre-clustered into groups of related types. For example, mitral valve replacements were grouped with aortic valve replacements in one class. An experienced surgeon determined these coarse operation classes were reasonable. These classes and the current procedural terminology codes (CPT codes) that fall into each class can be found in Appendix B. It is important to note that most patients fell into multiple classes because they had multiple operations. Also, 302 patients did not have operations that fell into one of the 20 classes and 127 patients that had no operations at all. Thus, there are 22 classes (the 20 operation classes, the patients with operations that did not fall into one of these classes and the patients with no operation).

The end result of the preprocessing was a data set that contained the number of days each patient spent in each level of care (ICU or NonICU), each patient's vital statistics, and a list of operation classes for each patient. Certain medical information (e.g., kidney function) was not available and is not always easily accessible to the decision makers that manage the beds. Therefore, this information was omitted from consideration. Further work, could be done to determine what other data are readily available (or could be made available) and relevant to determining a patient's length of stay.

The data had enough extreme outliers (patients with lengths of stay more than three standard deviations away from the mean) that the results could be skewed for some techniques. Extreme outliers amounted to about 2% of the data and outliers (patients with length of stay more than two standard deviations) were more than 4%. We looked into two approaches to handle the outliers: omit the extreme outliers and use a logarithm to transform the data set.

Because the data had the appearance of being lognormally distributed, we looked at

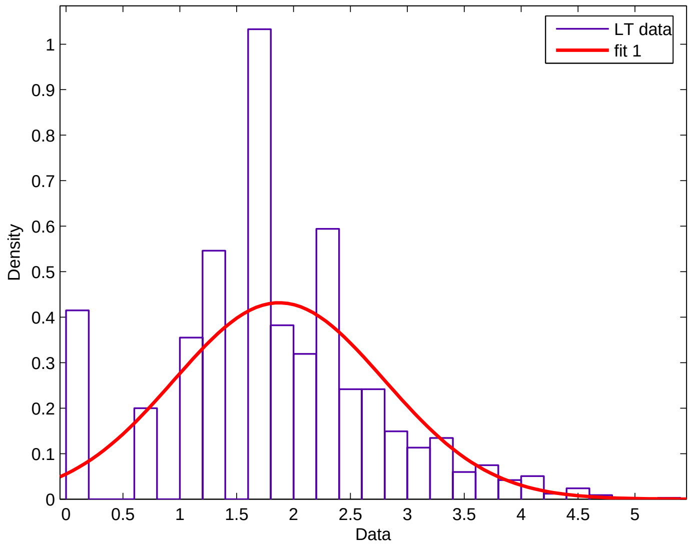

Figure 5.1: A histogram of the lognormal transform of the total length of stay of cardiac surgery patients with the best fit normal curve

the logarithm of total length of stay (some patients skip the ICU and move directly to a NonICU bed, while others are discharged directly from the ICU). This would not work for ICU or NonICU length of stay because they both had patients with zero length of stay. As can be seen in Figure 5.1, the lognormal transform is most likely not normal. We found that only three patients had a transformed length of stay that was longer than three standard deviations away from the mean. For 1,675 standard normally distributed observations, an average of 2.18 are expected to lie beyond three standard deviations of the mean. Thus, the problem with having too large a percentage of the patients lying in the tail has essentially been solved by using the lognormal transformation.

	Total	Transformed	Truncated
Mean	10.22	1.87	8.02
Median	6	1.79	6
Standard Deviation (SD)	13.22	0.92	6.46
% Observations more than 2 SD from Mean	4.30	3.34	2.25
% Observations more than 3 SD from Mean	2.15	0.18	0

Table 5.1: Basic statistics about the data sets

5.3.2 Methods

We tested six different approaches to estimating patient LoS on each of the data sets (Total, Transformed, and Truncated): average by class, regression, chi-squared automatic interaction detection (CHAID) implemented in AnswerTree, the M5'algorithm implemented in the Waikato Environment for Knowledge Analysis (WEKA), and a neural network also implemented in WEKA [22].

The class average estimator takes the mean of all previous instances of an operation class as the estimate for the duration of future operations of the class. Essentially, the class average estimates each patient's LoS with the average for his group. Patients in multiple classes were included when determining the mean of all applicable classes and the expected values. For testing purposes, we assigned the longest length of stay applicable. This is similar to how operation duration is currently estimated. The hospital database keeps track of the previous patient's LoS and uses the average of the last five patients of a type, excluding the longest and the shortest. One problem with this approach is the large number of multiple-class patients.

We used least-squares linear regression on the entire data set without any grouping.

The variables included the operation classes, age, race, and sex. We also used least median regression. Unlike least squares regression, least median regression minimizes the median of the error instead of the root mean squared error.

In addition to the standard regression approaches, we used three data mining techniques. CHAID is a method for clustering large data sets into groups. CHAID creates groups that have little variation with in the group relative to the variation between the groups. M5' is a model tree algorithm. This algorithm generates a series of rules that divides the data set into groups and creates a linear model for each of the groups. A neural network model with back propagation was tested in WEKA. To reduce the error due to over-training, the learning rate decayed. This parameter was found to significantly reduce error.

5.3.3 Testing Results

In order to provide a fair comparison, a random sample of about 34% of the cases was removed from the total data set to serve as a test set. The remaining 66% of the data set was used as the training set. The models are generated on the training set and then the error is calculated on the test set. Ten different training/testing sets were used and the results were averaged over these different runs.

When comparing methods, it is important to analyze both the error and the number of rules generated. Clearly, a model with a large error will not be useful in making predictions. The problems that arise from using many groupings may be less obvious. A large number of rules may indicate over-training. Over-training is a problem because it can lead to estimates that are too specific to the data set used for the model training. Essentially, this means that the estimates fit the training set well but not other test sets. Thus, one can be deceived about the true accuracy of a method if the training set is used for testing. By using a separate set for training and testing, over-training should be detected and avoided. By comparing the six methods on a test set, over-training is accounted for in the error. The number of rules that we generated also can be considered a measure of the ease of use for the estimates. A large number of rules and models will, regardless of accuracy, create an unwieldy system that cannot be implemented easily in a hospital setting.

We looked at two error metrics: mean absolute error (MAE) and root mean squared error (RMSE). MAE is less sensitive to outliers than RMSE and a comparison of these two metrics can provide an indication of the impact outliers are having on the estimation accuracy. The formula for MAE is (5.1) and (5.2) is for RMSE. In these equations, n is the number of observations, x_i is the actual value, and $\hat{x}_i$ is the estimated value.

$$MAE = \sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{n} \tag{5.1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \hat{x}_i)^2}{n}} \tag{5.2}$$

Table 5.2 provides a detailed comparison of the methods. Table 5.2 reports the error for each method on the unmodified data set (total), the data set with patients more than three standard deviations from the mean removed (truncated), and on the lognormal transformed data set (transformed).

	Neural Networks	Linear Regression	Median Regression	Group Mean	Model Tree	CHAID
Total MAE	4.31	4.17	3.87	4.26	4.42	4.60
Total RMSE	9.11	8.75	10.33	9.03	9.24	9.67
Truncated MAE	4.20	4.31	3.69	4.24	4.31	4.34
Truncated RMSE	6.11	6.21	7.43	6.17	6.12	6.43
Transformed MAE	.570	.571	.595	.583	.584	.592
Transformed RMSE	.756	.751	.802	.766	.770	.769
Parameters	406	25	25	21	10	5

Table 5.2: The prediction results for total length of stay

5.3.4 Discussion

As mentioned earlier, there are two factors that need to be considered when comparing prediction methods: accuracy and complexity. A model that lacks sufficient accuracy is worthless and a model that is too complex to use is equally devoid of value.

For each approach, the mean absolute error (MAE) is roughly four days on both the training and testing sets. On the full set, the root mean squared error (RMSE) is between roughly eight and ten for each method. The RMSE is significantly more sensitive to outliers than the MAE. Because the RMSE is more than two times the MAE, it is clear that outliers are a driver of the error. The truncated set performed much better in terms of the RMSE, but the MAE error did not significantly improve. The error on the transformed data set can be thought of as percent error (i.e., .57 indicates 57% error). It is clear from Table 5.2 that the transformation did not significantly improve the estimates.

Unfortunately, these results are not sufficiently accurate to predict capacity. If the error is 4 days, when the average length of stay is 10 days, there is a significant problem with accuracy. None of the methods performed substantially better than the others. Because we did not find a sufficiently accurate estimate of patient length of stay, an alternate approach is needed to predict discharge volume.

5.4 Discharge Volume Prediction

Because the individual LoS predictions were not sufficiently accurate to provide a good estimate of the discharge volume, we decided to use a probabilistic approach based on survival analysis. We first want to know the probability that a patient will be discharged from the unit a few days from today given that patient's current length of stay. To do this, we used the CHAID clusters and then made a posterior distribution for each group. Then we need to aggregate these probabilities using a binomial-like distribution to estimate the total number of discharges a few days from today. In this section, we will describe the approach and test it both using simulation and in the hospital setting.

5.4.1 CHAID Groupings

We used the clusters produced by the CHAID approach to classify each patient into a group. The CHAID groupings produced a set of groups that can be easily interpreted. The first group consisted of patients that had no surgery. These patients generally were those with ventricular assistance devices (VADs) previously installed. The VAD patients visit the ICU for their periodic check ups. The second group contained patients with fairly minor surgeries. The third group consisted of bypasses of one, two, or three arteries. The

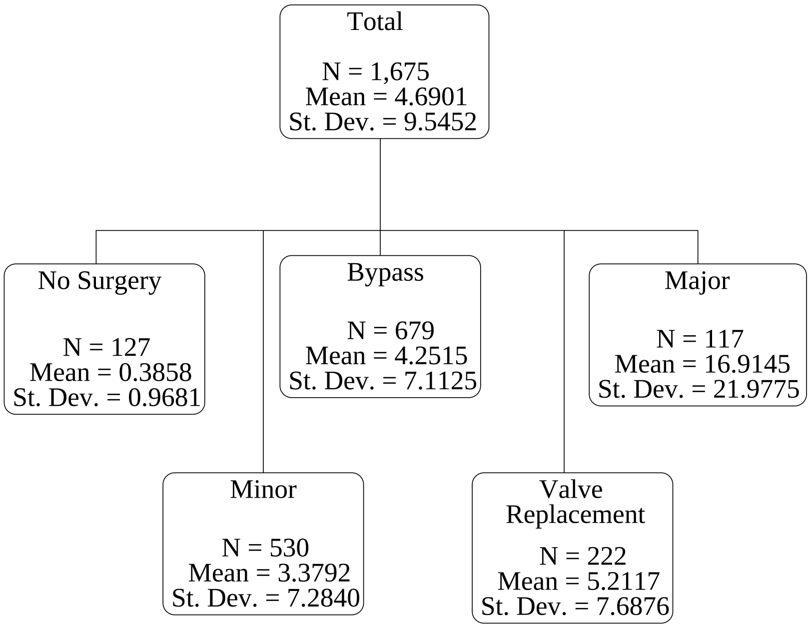

Figure 5.2: The tree formed by the CHAID clusters.

	Sample Size	Mean	Standard Deviation
No Surgery	127	0.3858	0.9681
Minor	530	3.3792	7.2840
Bypass	679	4.2515	7.1125
Valve Replacement	222	5.2117	7.6876
Major	117	16.9145	21.9775
Total	1675	4.6901	9.5452

Table 5.3: Statistics about each of the groups.

forth group was predominately valve replacement types (mitral, aortic, etc.), but also more serious bypasses (four or more arteries). We labeled the final group Major. This group included operations like VADs, heart transplants, and lung transplants. When a patient had operations in multiple groups, the patient was placed into the most serious applicable group. Statistics for the groupings can be found in Table 5.3. Figure 5.2 displays the groupings in a tree format.

5.4.2 Posterior Distribution

No parametric distribution provided a good fit for any of the length of stay groups. The tail is too large for either a lognormal or an exponential distribution. In order to construct a posterior distribution, we used Kaplan-Meier estimators as seen in equation (5.3). In this equation, $p_{g,j,k}$ is the probability that a group g patient has been discharged from the current unit after j days, given that he has already spent k days in the hospital. $n_{g,i}$ is the number of group g patients discharged with a length of stay of exactly i days in the primary data set.

$$p_{g,j,k} = 1 - \frac{\sum_{i=j+k}^{\infty} n_{g,i}}{\sum_{i=k}^{\infty} n_{g,i}} \text{ such that } j \geq 0. \tag{5.3}$$

5.4.3 Aggregation

Now that we have the posterior distributions, we need a way to put them together to estimate unit census. Let X_j be the number of patients discharged from the unit j days after today and x_i be the event that patient i is discharged by j days from today. If there are m patients in the hospital on day k,

$$X_i = \sum_{i=1}^{m} x_{i,j} \text{ such that}$$

$$x_{i,j} = \begin{cases} 1 & \text{with probability } p_{g_i,j,k_i}, \\ 0 & \text{otherwise.} \end{cases}$$

where g_i is the group of patient i and k_i is patient i's current length of stay.

We assume that $Cov(x_{i,j}, x_{l,j}) = 0 \ \forall i \neq l$ and j. This condition means that whether or not patient i is in the hospital j days from today is not contingent upon patient l's discharge status. The expected number of discharges between today and j days from today, $E(X_j)$ is

$$\begin{aligned} E(X_j) &= E\Big(\sum_{i=1}^{m} x_{i,j}\Big) \\ &= \sum_{i=1}^{m} E(x_{i,j}) \\ &= \sum_{i=1}^{m} p_{g_i,j,k_i}. \end{aligned}$$

The variance of X_j, $Var(X_j)$, is

$$\begin{aligned} Var(X_j) &= Var\Big(\sum_{i=1}^{m} x_{i,j}\Big) \\ &= \sum_{i=1}^{m} Var(x_{i,j}) + 2\sum_{i=1}^{m-1}\sum_{l=i+1}^{m} Cov(x_{i,j}, x_{l,j}) \\ &= \sum_{i=1}^{m} p_{g_i,j,k_i}(1 - p_{g_i,j,k_i}) + 0 \\ &= \sum_{i=1}^{m} p_{g_i,j,k_i}(1 - p_{g_i,j,k_i}). \end{aligned}$$

5.4.4 Testing

We tested the approach in two ways: first using a simulation and then in the real-time hospital setting. The simulation was constructed in Python. We used three error metrics to analyze the prediction approach: Bias (equation (5.4)), MAE (equation (5.1)), and RMSE (equation (5.2)). In equation (5.4), $Bias(j)$ is the bias of estimates made j days prior to the observation, $X_{v,j}$ is the actual census of day v after j days and $\hat{X}_{v,j}$ is the estimate for this observation. n is the number of days the observed. A large negative value for $Bias(j)$ would imply the estimate was consistently larger than the actual. A large positive value implies underestimation of the observation. If $Bias(j)$ is small, it would indicate very little systematic bias in the discharge estimates.

$$Bias(j) = \sum_{v=1}^{n} \frac{X_{v,j} - \hat{X}_{v,j}}{n} \tag{5.4}$$

We also include the percent of predictions that were overestimates, accurate, or underestimates. An estimate is considered to be accurate if $|X_j - \hat{X}_j| < .5$.

Simulation Tests

To test the prediction approach, we simulated the CSICU census using the primary data set. We ran 10,000 runs with 12 patients per run. Each run was generated independently. To simulate a patient, we randomly sampled patients from the data set. Each patient had an ICU LoS and group. We assigned each patient a current LoS between $[0, ICULoS)$. X_j was taken to be the number of patients remaining in the unit j days after the estimate was

j	1	2	3
Bias	.0205	.0050	-.0042
MAE	1.141	1.265	1.265
RMSE	1.427	1.598	1.578
Underestimates	38.9%	38.1%	37.4%
Accurate	34.3%	37.5%	38.2%
Overestimated	26.8%	24.4%	24.4%

Table 5.4: Statistics about the simulation experiment.

made calculated using the posterior LoS distribution. Appendix C explains why the ICU LoS sampled for each patient cannot be used to calculate an unbiased discharge count. The estimate, $\hat{X}_j$, was calculated using $\hat{X}_j = E(X_j)$. The error was taken to be the difference between the estimate and the actual.

The results for the simulation comparison can be found in Table 5.4. The one day results are promising. There is no strong bias and the error is fairly small. The multiday error is sufficiently small that the prediction approach could be used for planning purposes. The bias is sufficiently small that it is most likely a result of rounding error. In general, there appeared to be more overestimates of the discharge. This skew is likely a result of the census distribution's skewness. Essentially, the ICU is more likely to be full than empty, because of the length of stay distribution, thus the census distribution will be skewed.

Live Tests

In the summer of 2007, we applied the prediction approach to estimate the CSICU discharge volume a few days in advance. We used the operation postings (the hospital's internal list of operations) to classify patients and the internal bed tracking to determine

	1 Day	2 Days	3 Days
Absolute Error	1.15	1.80	2.50
Bias	-0.17	-0.49	-0.87
% Overestimated	17%	20%	0%
% Accurate	33%	30%	25%
% Underestimates	50%	50%	75%

Table 5.5: Statistics about each of the groups.

the available census. We tracked the error (as measured by the difference between the prediction and the actual) for 20 days. The results can be found in Table 5.5.

The results indicate a larger bias and higher error than in the simulation experiments. The evidence of a systematic underestimation of the number of discharges is particularly problematic.

5.4.5 Discussion

There are a few explanations for the difference between the simulation performance and the actual performance.

First, the observations in the hospital tests are not independent because many of the same patients are present in each of the observations. This is problematic because the error for the observations will be autocorrelated. This makes the error estimates unreliable.

Second, the error appeared to be larger on days with more scheduled cases. Closer inspection revealed that more patients were discharged than expected when the volume of scheduled cases was high. This implies that the posterior distribution is dependent upon the scheduled volume. This also indicates that our assumption that $Cov(x_i(t), x_j(t)) = 0 \ \forall i, j, t$

was flawed. Upon reflection, this should not be too surprising because the surgeons are pressured to discharge more patients when there are more surgical cases scheduled.

5.5 Conclusions and Further Work

Our results highlight the difficulty with forecasting discharges and patient length of stay. While our approach was not sufficiently accurate to be used for its original purpose of predicting discharge volume a few days in advance, it does indicate some important considerations for future work. The simulation results indicate that the predictions can be accurate enough if the posterior distribution are a good fit. A close inspection of the error found that the volume of discharges was higher on days with a high scheduled caseload. Further work is needed to determine if this is statistically significant. Additional work is also needed to find how this effect alters the posterior distribution. Predicting discharge volume is very important for managing patient flow and this work provides an outline that can be duplicated in other hospitals.

Chapter 6

The Impact of Scheduling on Post-Operative Length of Stay

We investigate the relationship between surgeon incentives and discharge practices. More specifically, we look for evidence that patients are being discharged sooner because of capacity constraints. There have been several studies on the impact of physician incentives. In contrast to previous studies, we will focus on how the assignment of operating room time impacts patient length of stay and patient discharges. In addition to impacting patient care, these effects can reduce hospital efficiency.

6.1 Introduction

In a health care system that uses a fee for service model, surgical services tend to be a primary driver of hospital revenues and profits. Profits from surgical services are used to cross-subsidize less profitable, but still vital aspects of hospital operations. Surgeons also derive a large portion of their personal income from the surgeries that they perform and they make more money by doing more operations. Therefore, surgeons and hospitals have

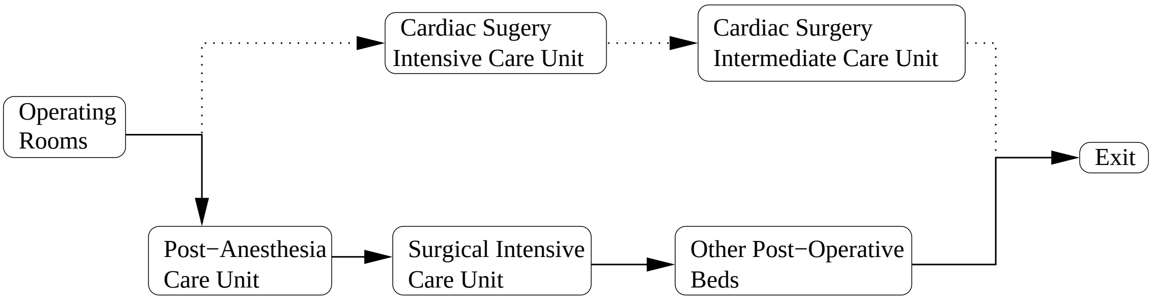

Figure 6.1: The post-operative path taken by patients. The dotted line is the path taken by cardiac surgery patients and the bold path is taken by most other patients from other service lines (e.g. general surgery or orthopedics).

an incentive to ensure that surgeries can be performed.

While an operating room is the most obvious resource necessary to perform surgery, downstream bed capacity is also required for the patient to recover. Figure 6.1 shows two post-operative paths: the bold line for most types of surgery and the dotted line for cardiac surgery patients. Immediately following surgery, most patients move to the post-anesthesia care unit (PACU) where they spend an hour or two recovering from the anesthesia. Usually, these patients require a day or two in the surgical intensive care unit (SICU). Once they have sufficiently recovered, the patients move to a step down unit for a few days before they are discharged. Cardiac surgery requires specialized resources beyond those in the SICU and so there is a dedicated cardiac surgery intensive care unit (CSICU) and a cardiac surgery intermediate care unit (CSIMC), the step down unit of cardiac surgery. Cardiac surgery patients move directly to the CSICU following surgery and then step down to the CSIMC before discharge.

A segment of operating room time (frequently referred to as a “block”) is a guarantee that a specific operating room will be available to the service line on a specific day. Each surgical service line (neurosurgery, orthopedic surgery, etc.) is allotted a certain quantity of blocks throughout the week. The surgeons within the service line then schedule their individual cases into the block time designated for the service line. Each service line's schedule of cases is developed independently of other service lines and generally without explicit regard to the number of patients in the hospital (at least in some part because these cases are scheduled in advance and the available capacity is not known at that time). The term "block schedule" refers to the schedule of all the block time for all of the surgical service lines. The block schedule is developed by a committee that considers a variety of physical constraints and surgeon preferences, but, in general, they do not take into account system efficiency. Thus, neither the block schedule nor the individual operating room schedule accounts for systemwide efficiency or capacity issues.

If the hospital does not have sufficient downstream bed capacity on the day of surgery, surgical cases are either canceled or delayed. Either of these is an expensive option and creates problems for both the staff and patients. Therefore, it is in each surgeon's interest to ensure that there is capacity available on days when the surgeon is scheduled to do cases. Previous data analysis found evidence that there are more discharges on days when the surgeons have blocks [29]. Because the surgeons make the operating schedule independent of the future state of the system, it seems reasonable that one mechanism by which surgeons could ensure bed availability would be to adjust in the length of stay for their patients. For example, if the hospital's post-operative beds are full, a surgeon might discharge patients earlier than under normal conditions to make room for his cases on that day. This could

Day of Week	SGL Blocks	SCS Blocks
Sunday	0.0	0.0
Monday	1.0	2.5
Tuesday	2.0	3.0
Wednesday	1.6	2.5
Thursday	2.0	2.0
Friday	2.9	2.0
Saturday	0.0	0.0

Table 6.1: The current number of blocks given to SGL and SCS for each day of the week.

lead to patients returning to the ICU a few days later in their or an increased rate of readmissions to the hospital. It also could lead to an increase in the stress and workload of clinical staff in the downstream beds because they would be caring for higher acuity patients. To determine if this is occurring, we looked at historical discharge data to see if the day of discharge was related to length of stay.

The hospital analyzed in this paper is a large academic tertiary-care medical center. This hospital uses a block schedule negotiated every three months with more than 15 blocks assigned to the various service lines most days. The schedules of block time for the General Surgery (SGL) and Cardiac Surgery (SCS) groups can be found in Table 6.1. It is important to note that the number of blocks is not uniform across the week. The fractions of a block indicate fractions of a day that the operating room is set aside for the service line. For the time period reflected in the data, the block schedule did not change for general and cardiac surgeries.

6.2 Background

Understanding the variation in length of stay is of great interest to hospital administrators and health policy makers. The literature has looked into understanding this problem and we will discuss some of the papers here. We will also relate them to the problem of detecting and explaining day of week variations in length of stay and discharge volume.

Both Singer et al. [23] and Strauss et al [24] discuss the rationing of intensive care beds to maintain the flow of post-operative patients in the hospital. Both of these papers investigate how physicians respond to a decrease in bed availability by reducing length of stay and increasing discharges. McManus et al. [25] found that the primary driver of the decreases in bed availability was scheduled admissions, not emergency arrivals. We extend these works by taking a rigorous look at the impact of the surgical block schedule (therefore the variability resulting from scheduled admissions) on patient length of stay.

There appears to be a consensus in the literature that financial incentives impact medical care. Mitchell [26] found that physician ownership of hospitals impacted utilization and practice patterns. Levin and Rao [27] found that financial incentives resulting from physician ownership of imaging equipment led to self-referrals and an overutilization of imaging. Both of these works focus on how direct financial incentives impact physician behavior. Like all economic actors, physicians respond to financial incentives. We will focus on more indirect incentives that arise from constrained hospital resources. If there is insufficient post-operative capacity available, surgeons generally cannot perform surgeries because there is nowhere for the patients to recover. Because surgeons are generally paid per surgery, there is a financial (and medical) incentive to ensure that post-operative capacity is available on days when they have surgical block time.

Length of stay prediction is an important aspect of hospital management because an accurate estimate is needed to determine resource requirements such as nurse staffing. Zernikow et al. [20] used a neural network to predict the length of stay of preterm neonates (babies born prematurely). They found that data from the first day of life could be used to predict length of stay in the neonatal intensive care unit. While their model had a high correlation to the actual length of stay, other error metrics were relatively large. Mobley et al. [21] also used a neural network to predict length of stay for post-coronary care patients. Their model also had a high correlation, but other error metrics were very large relative to the actual length of stay. Neural networks were used to predict psychiatric length of stay by Davis et al. [28]. Their neural network model was more accurate than initial estimates of patient length of stay made by the treatment team. None of these papers considered the role of physician incentives in determining patient length of stay. Because the physician is ultimately the one who is responsible for making the discharge decision, these works omit a key component in determining the length of stay (obviously, the health state of the patient is a primary determinant). Price et al. [29] used data mining and survival analysis to predict CSICU capacity a few days in advance. They found evidence that the day of week impacted discharge policy. In this chapter, we will look for impacts of the day of week on discharge practices and explain some possible causes related to physician incentives.

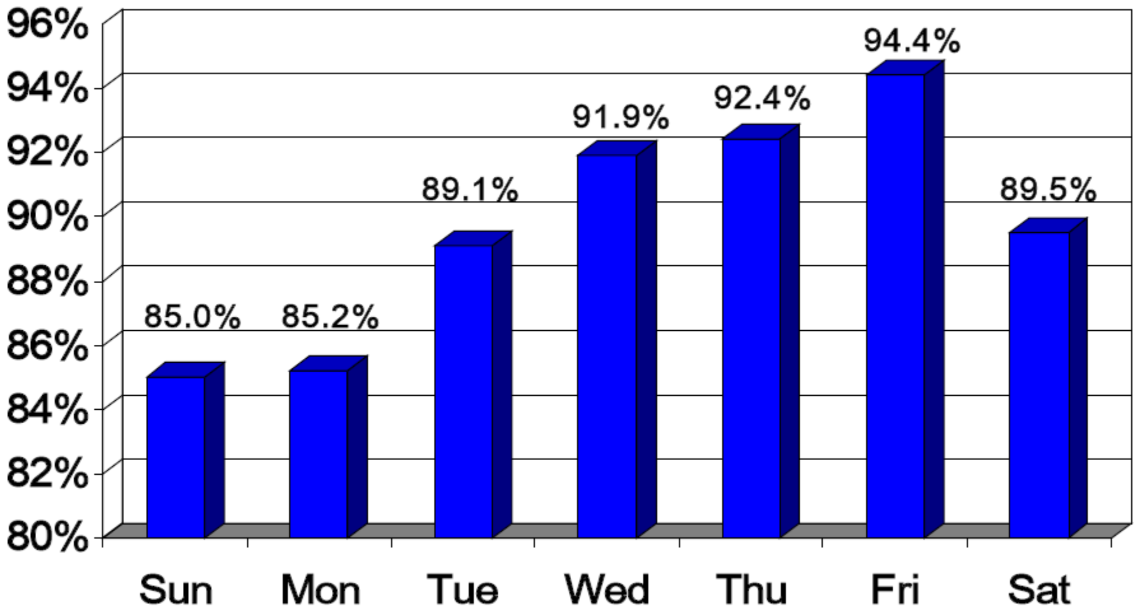

Figure 6.2: The hospital's average utilization of post-operative beds.

6.3 Data and Methodology

In this study, we used data from the financial department of a large academic tertiary-care medical center. The data set used in this work consisted of roughly 51 weeks of surgical discharge data during the 2007 fiscal year (July 1st 2006 to June 30th 2007). Figure 6.2 is a histogram of the average utilization of the post operative bed utilization. We see in this histogram that the utilization increases later in the week. As discussed in the previous section, available capacity is an important consideration in determining length of stay. Unfortunately, time series data on utilization was not available. The variables included in the analysis are listed below along with a brief description. Table 6.2 provides a summary of the average values of each of the variables for both SGL and SCS.

	SGS	SCS
Race	0.73	0.67
Sex	0.40	0.64
Level	0.89	2.11
Age	52.54	59.58
LoS	6.20	11.05
Ln(LoS)	1.29	2.04
Arrivals	1.65	1.81
Discharges	1.67	1.82

Table 6.2: Patient characteristic statistics. These represent the mean values for each variable in the data set for the respective service line.

1. *Race* is taken to be 0 or 1. 1 represents white, while 0 represents other. The primary other is black and there are not enough Hispanic, Asian, or others to warrant further divisions or indicator variables.

2. *Sex* is taken to be 0 or 1. 1 represents male.

3. *Level* is the APR-DRG Severity of Illness Index and ranged from 0-3. 0 is for minor, 1 for moderate, 2 for major, and 3 for extreme.

4. *Age* is a non-negative integer representing the patent's age in years rounded down (so 59 and 11 months is taken to be 59).

5. *LoS* is a positive real number representing the length of stay of the patient.

6. *Ln(LoS)* is a natural logarithm transformation of LoS and is a real number. We used this transformation because the length of stay had a long tail and the logarithm transformation produced a distribution that was closer to normal.

7. *Arrivals* is a non-negative integer representing the number of patients arriving on a given day. This is the number of patients that were admitted following surgery, not the total number of surgical cases. Not all patients are admitted after surgery because some are transferred to another hospital or discharged immediately because the surgery was relatively minor.

8. *Discharges* is a non-negative integer representing the number of patients discharged on a given day.

9. *Blocks* is the number of operating rooms assigned to a service line on a given day (the block schedules for SCS and SGL can be found in Table 6.1).

In order to provide an in depth look at the system, our analysis will focus separately on General Surgery and Cardiac Surgery. First, we used Analysis of Variance (ANOVA) to see if there were differences in discharge volume, length of stay, and patient acuity level between the days of the week. Next, we looked at the correlations between the variables to get an indication of the possible relationships between them. Finally, we did a series of regressions on Ln(LoS) and discharges to determine the magnitude and significance of the relationships. Unless otherwise stated, the results are to a 95% confidence. The regressions cannot prove causality but they can provide evidence of the impact of incentives. The results and interpretation for General Surgery are provided first, followed by Cardiac Surgery.

6.4 General Surgery

General surgery includes a variety of surgical operations such as laparoscopy, hernia repair, and gall bladder removal. The general surgeons at this hospital emphasizes minimally invasive procedures. SGL patients generally use the Surgical Intensive Care Unit (SICU) for some portion of their post-operative length of stay. Most other surgical service lines (e.g., orthopedics or surgical oncology) also use the SICU. Therefore, the available capacity is not entirely a function of SGL volume and discharge policies. In some sense, SGL must compete with the other service lines for beds and to be assured of having available beds when occupancy is high, the general surgeons must discharge their patients.

Figure 6.3 has some basic statistics about the general surgery service line. There appear to be differences in discharge volume, Ln(LoS), and patient acuity level among the days of the week. This will be tested using ANOVA. It also appears that some of the variables could be related. We will investigate those relationships using an analysis of the correlations and regression.

6.4.1 Differences Throughout the Week

We will use ANOVA to determine if there are statistically significant differences between characteristics of patients discharged on different days of the week. In this section, there are three sets of ANOVA results. The first is the ANOVA for the volume of discharges by the day of week. The second is for average length of stay and the third is for average

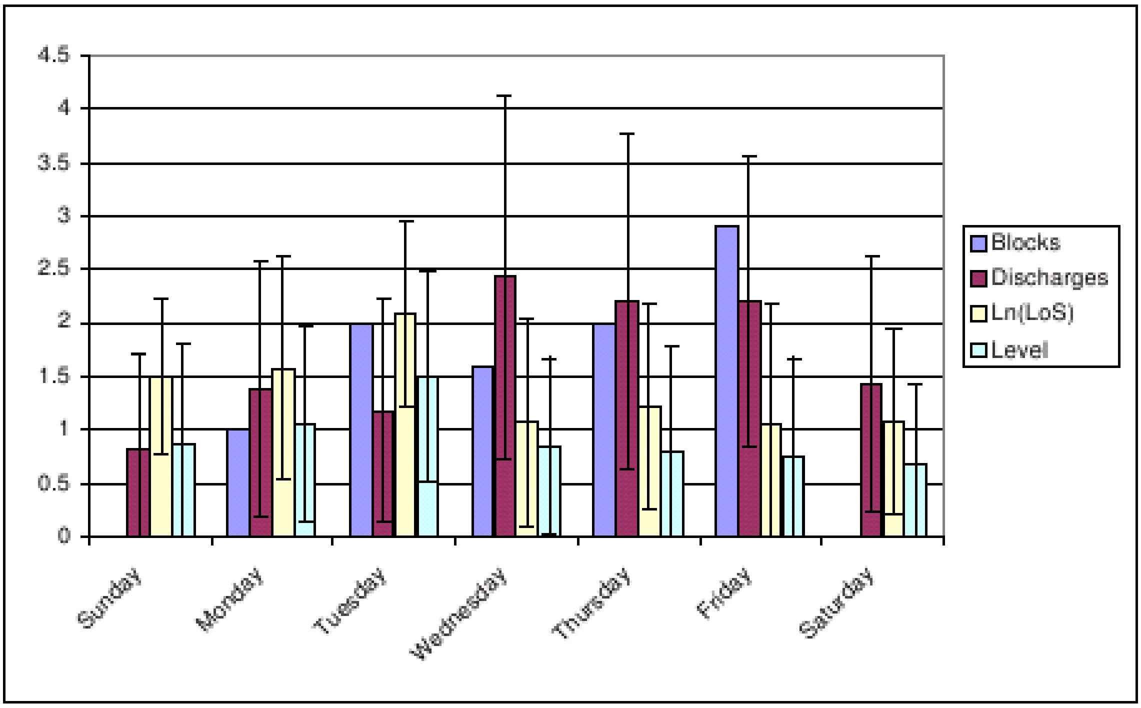

Figure 6.3: Basic statistics about SGL by day of week. The bars mark one standard deviation from the mean. The number of blocks ranges from one to three during the week and there are no blocks on weekends. During the week the average number of discharges ranges from 1 to 2.5 and is substantially lower on weekends. The length of stay is quite variable across the week as is the level of acuity.

patient acuity or level. In each case, the ANOVA rejects the null hypothesis that the groups have identical means. This means there are statistically significant differences between the days of the week.

Table 6.3 contains statistics about the number of discharges by day of week. The average number of discharges is much higher for Wednesday, Thursday, and Friday. Additionally, the variance in the number of discharges is much higher for Wednesday and Thursday than other days of the week. Table 6.4 contains statistics on the length of stay and the acuity level of patients separated by the day of discharge. Patients discharged on Tuesdays have a

	Sample Size	Average	Variance
Sunday	51	0.82	0.79
Monday	51	1.39	1.44
Tuesday	51	1.18	1.07
Wednesday	51	2.43	2.89
Thursday	51	2.20	2.44
Friday	51	2.20	1.84
Saturday	51	1.43	1.45

Table 6.3: Discharge volume statistics. There is a clear increase in the number discharges later in the week.

	Sample Size	Ln(LoS) Average	Ln(LoS) Variance	Level Average	Level Variance
Sunday	42	1.50	0.54	0.86	0.91
Monday	71	1.58	1.08	1.06	0.85
Tuesday	60	2.08	0.76	1.50	0.97
Wednesday	124	1.07	0.94	0.85	0.68
Thursday	113	1.22	0.94	0.80	0.98
Friday	117	1.05	1.29	0.76	0.84
Saturday	75	1.08	0.77	0.67	0.58

Table 6.4: Ln(LoS) and Level Statistics. Patients discharged on Wednesdays, Fridays, and Saturdays have significantly lower lengths of stay than other days. Patients discharged on Tuesdays have a significantly longer length of stay and higher acuity on average than patients discharged on other days of the week.

longer length of stay and higher level of acuity than patients discharged on other days. The individual ANOVA results will be discussed below and ANOVA statistics can be found in Appendix D.

Discharge Volume

Table 6.3 has the discharge statistics by day of week. An ANOVA rejects the notion that the discharge volume is identical throughout the week. Later in the week, the volume of discharges picks up. At first glance, that would seem to be natural because many of the patients arriving earlier in the week would be ready to leave by the end of the week. However, this ignores the facts that most of the blocks are later in the week, the peak in average length of stay is three days long (not five days), and relatively few SGL patients stay in the hospital for one week.

Ln(LoS)

The pattern for Ln(LoS) in the Table 6.4 data is similar to that of the number of discharges. An ANOVA rejects the hypothesis that the length of stay of patients being discharged is constant throughout the week. The average length of stay drops off later in the week when the average number of discharges picks up. Therefore, there are more patients with shorter lengths of stay being discharged later in the week when the number of blocks for SGL reaches a peak. This dip in length of stay also corresponds to the higher level of utilization at the end of the week seen in Figure 6.2.

Level

The ANOVA determined the acuity level is not constant throughout the week. The level appears to peak on Tuesday and then decrease later in the week. This is not very surprising since length of stay peaked on Tuesday and patients with higher levels of acuity would be expected to stay longer in the hospital. These results imply that patients with the most severe conditions are discharged more frequently on Mondays and Tuesdays after having spent the weekend in the hospital.

6.4.2 Relationships Between Variables

This section contains a correlation analysis of the variables. Table 6.5 has the correlations for individual patent's statistics (e.g., *LoS* is the patients' length of stay). The key finding in Table 6.5 is the strong correlation between *Level* and *LoS*. This implies that medical necessity is the primary determinant of length of stay.

Table 6.6 shows the correlations for the average values of patients discharged on a given date. For example, *aLoS* in Table 6.6 is the average length of stay of patients discharged on a given date. So if one patient is discharged today with a length of stay of one day and another patient is discharged with a length of stay of three days, the average length of stay of patients discharged today will be two days. In this table, *Arrivals* is correlated with *Blocks* and *Discharges* a day or two in advance. *Blocks* is also correlated with *Discharges.*

6.4.3 Models on Variables

The correlations indicated some possible relationships that require further investigations. This section contains the statistics from regressions on the patient characteristics and

	Race	*Level*	*Sex*	*Age*
LoS	-.0582	.5180	.0480	.1720
Ln(*LoS*)	-.0960	.6102	.1146	.2187
Age	.1398	.1932	.1532	-
Sex	-.0022	.1763	-	-
Level	-.0881	-	-	-

Table 6.5: This is the correlation table for individual patient statistics. The strongest correlations by far are with the length of stay measures and *Level.* This indicates that the level of acuity is the primary driver for patients' length of stay.

	Arrivals	*Blocks*	*Discharges*
aLevel	.1386	.0748	.0119
$aLoS_{-1}$	-.0313	.0485	-.0794
aLoS	.0476	-.0059	-.0460
$aLoS_{+1}$	.0182	-.1343	-.0251
$aLnLoS_{-1}$	.0165	.1330	.0303
$aLn(Los)$	.1112	-.0291	-.0799
$aLnLoS_{+1}$	-.0385	-.2698	-.0896
$Discharges_{-2}$	-.2256	-.1085	.0502
$Discharges_{-1}$	-.0064	.0801	.0523
Discharges	.1414	.2655	-
$Discharges_{+1}$	.3060	.2326	-
$Discharges_{+2}$	.2447	-.0234	-
$Blocks_{-1}$	-.1137	-	-
Blocks	.4456	-	-
$Blocks_{+1}$	.4528	-	-

Table 6.6: Correlation table for average daily statistics. For this table, *aLevel* indicates the average level of patients discharged and x_{-1} indicates the value of x the day before. The number of blocks has a negative correlation with the length of stay tomorrow. The number of blocks correlates positively with the number of discharges today and discharges tomorrow. Taken as a whole, these could imply that on days after SGL has blocks the hospital must discharge some of these patients sooner than on other days to make room for other incoming patients. This could in part be due to the correlation with *Arrivals. Arrivals* is strongly correlated with *Blocks* and *Discharges.*

their length of stay. We include three models: two on length of stay and one on the number of discharges.

SGL Model 1: Ln(LoS)

To determine if the number of blocks impacts the length of stay of patients, it is necessary to control for patents' level of acuity and other characteristics. The first model that we will use, SGL model 1, accounts for patient characteristics and looks for a relationship between blocks and length of stay.

$$Ln(LoS) = a_0 + a_1 * Age + a_2 * Sex + a_3 * Level + a_4 * Race + a_5 * Blocks.$$

Appendix D has goodness of fit statistics and statistics on the coefficients. The coefficient for both *Level* and *Age* are statistically significant and positive. This implies that more severe patients have longer lengths of stay and older patients stay longer. *Blocks* is not statistically significant.

SGL Model 2: Ln(LoS) with Day of Week Variables

Because the coefficient of *Blocks* was not statistically different from zero, we need to look into the day of week the patients were discharged to determine if the day of week impacts length of stay. For this set of regressions, the day of discharge was included in the analysis as an indicator variable. This added six variables to the analysis: the days of the week, *Monday* through *Saturday. Monday* is one if the patient was discharged on a Monday and zero otherwise. If the patient was discharged on a Sunday, all the day of week indicator variables are zero. Appendix D has goodness of fit and coefficient statistics for this model.

$$Ln(LoS) = a_0 + a_1 * Age + a_2 * Sex + a_3 * Race + a_4 * Level + a_5 * Monday + ...$$

Again, *Level* and *Age* are statistically significant and positive. For this model, the coefficients for *Wednesday, Thursday, Friday,* and *Saturday* were found to be statistically significant and negative. A negative coefficient indicates the length of stay is shorter on these days. In particular, the coefficients on *Wednesday* and *Friday* are the highest in absolute magnitude. *Wednesday*'s coefficient could be a response to the increase in utilization resulting from the surgeries earlier in the week (as seen in Figure 6.2). *Friday*'s coefficient could imply patients are being discharged sooner than other days to get them out of the hospital before the weekend.

SGL Model 3: Discharges

While SGL Model 1 suggests that block time does not significantly impact length of stay, SGL Model 2 suggests that length of stay is shortened later in the week when utilization was highest (as seen in Figure 6.2). Neither of these explain how the block schedule affects the discharge patterns. To investigate this relationship, we developed SGL Model 3. *aLevel* is the average acuity level of patients being discharged on a given day of week. Appendix D contains the regression statistics and the coefficients for SGL Model 3. For

SGL Model 3, we used:

$$Discharges = a_0 + a_1 * aLevel + a_2 * Arrivals + a_3 * Blocks.$$

Both the coefficients for *aLevel* and *Blocks* are statistically significant. This implies that there are more discharges on days with more blocks even when accounting for differences in *aLevel*.

6.4.4 Implications

The ANOVA results indicate that the volume of discharges, the average length of stay, and the average acuity level vary significantly by day of week. These differences warrant some explanation. The correlations indicate that acuity level is strongly correlated with patient length of stay. Additionally, the number of blocks is correlated with the number of discharges. While we cannot prove causality without a controlled experiment, a causal relationship seems feasible. Because the number of blocks is determined in advance and determined independent of volume considerations, this correlation implies that the number of blocks drives the number of discharges. However, because the average acuity level of acuity of patients discharged varied by day of the week, regression analysis was needed to determine if the block/discharge correlation was because of the acuity level varying or another process.

In each of the regressions, *Level* was included as a variable so that patient acuity would be taken into account. The regression in SGL Model 1 finds that the number of blocks does not significantly impact length of stay when the acuity level is included. However, in SGL Model 2 the coefficients for *Wednesday* through *Saturday* are significant and negative. In particular, the coefficients for *Wednesday* and *Friday* are large in absolute magnitude. This indicates that the LoS is shorter later in the week. SGL Model 3 indicates that the number of discharges increases significantly on days with more blocks. So the number of discharges increases on days with more block time and length of stay decreases on days later in the week.

One explanation for the regression coefficients discussed above involves surgeon incentives. As the week progresses, the number of patients in post-operative beds increases (SGL shares beds with other services). In order to perform additional cases later in the week, surgeons increase the number of patients discharged by reducing patients' lengths of stay on these days. Thus, hospital occupancy has a strong impact on the discharge practices. This explanation accounts for all of the effects in the data discussed above.

6.5 Cardiac Surgery

Most of the cardiac surgeries performed at the hospital are either valve replacements or bypasses. The hospital also has heart and lung transplants. An increasing number of these surgeries are done with robot assistance. The robotic surgeries are thought to have a shorter length of stay, but the surgical method (robotic versus traditional) was not included in the data set. As seen in Figure 6.1, cardiac surgery does not use the SICU, unlike SGL. In-

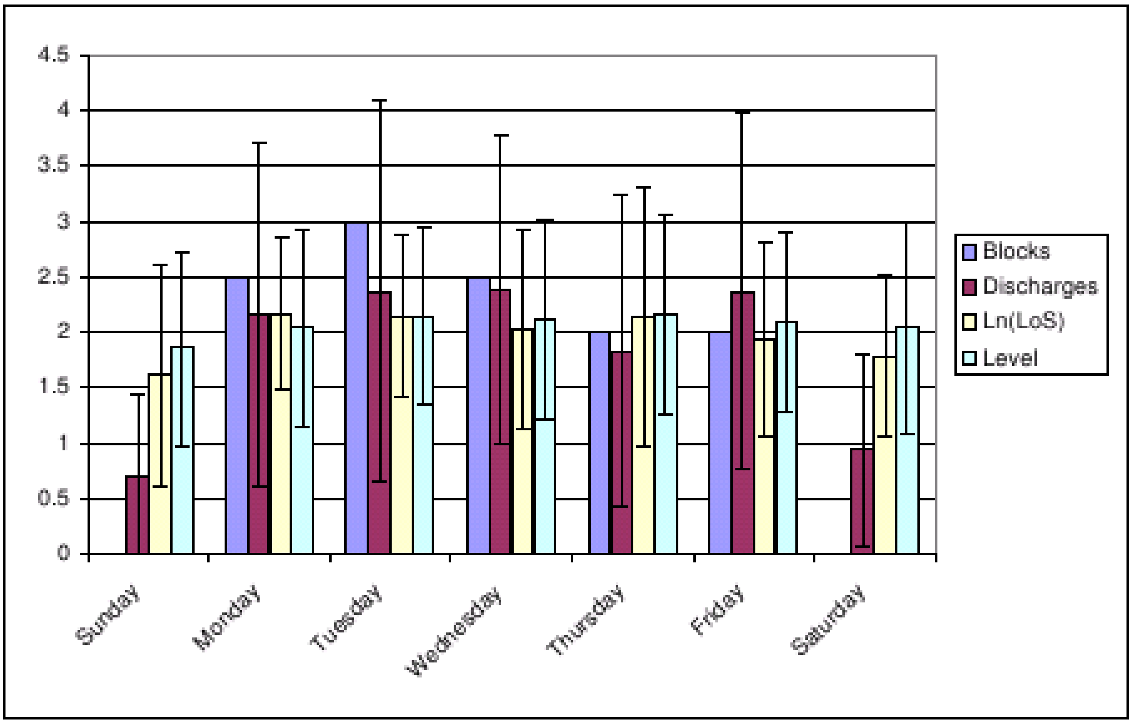

Figure 6.4: Basic statistics about SCS by day of week. The bars mark one standard deviation from the mean. The number of blocks ranges from two to three during the week and there are no blocks on weekends. During the week the average number of discharges ranges from 1.5 to 2.5 and is substantially lower on weekends. The average level and the average of the natural log of length of stay appear to move together, but there also appears to be much more variability in the length of stay.

stead, SCS uses the Cardiac Surgery Intensive Care Unit (CSICU) and the Cardiac Surgery Intermediate Care Unit (CSIMC) for post-operative patients. Essentially, this decouples the cardiac surgery volume from the other surgical service lines and the utilization impacts seen in Figure 6.2. General statistics by day of week can be seen in Figure 6.4. The utilization statistics for cardiac surgery post-operative units were not available and so a figure comparable to Figure 6.2 could not be made.

	Sample Size	Sum	Average	Variance
Sunday	51	35	0.69	0.58
Monday	51	110	2.16	2.41
Tuesday	51	121	2.37	2.96
Wednesday	51	122	2.39	1.96
Thursday	52	95	1.83	1.99
Friday	52	123	2.37	2.59
Saturday	52	49	0.94	0.76

Table 6.7: Volume Statistics. There is a clear increase in the number discharges later in the week.

6.5.1 Differences Throughout the Week

In this section, there are three sets of ANOVA results used to determine if there are differences in the characteristics of SCS patients by day of discharge. The first is the ANOVA for the volume of discharges by the day of week. Statistics about the discharges by day of week can be found in Table 6.7. The ANOVA finds statistically significant differences and it is clear that the weekends have substantially fewer discharges than weekdays. The second ANOVA is for average length of stay and the third is for average patient acuity level. Table 6.8 contains statistics on the average length of stay and acuity level of patients by day of discharge. The acuity level appears to be constant throughout the week, but the length of stay looks to be lower for patients discharged on weekends.

Discharge Volume

Basic statistics about the volume of discharges can be found in Table 6.7. The ANOVA rejects the notion that the volume is identical throughout the week. The statistics for the

	Sample Size	Ln(LoS) Average	Ln(LoS) Variance	Level Average	Level Variance
Sunday	35	1.62	1.01	1.86	0.77
Monday	110	2.17	0.46	2.04	0.77
Tuesday	123	2.15	0.54	2.15	0.64
Wednesday	122	2.03	0.80	2.12	0.80
Thursday	95	2.14	1.39	2.17	0.80
Friday	124	1.94	0.76	2.09	0.65
Saturday	49	1.79	0.53	2.04	0.91

Table 6.8: Level Statistics.

ANOVA can be found in Appendix E. The number of discharges is lower on Thursday than on other weekdays and much lower on weekends than weekdays.

Ln(LoS)

Table 6.8 contains data on the average length of stay for patients discharged on each day of week. Similar to the pattern in the number of discharges, the average length of stay drops off sharply on the weekends. An ANOVA rejects the notion the length of stay is constant across the week. The ANOVA statistics are in Appendix E.

Level

Here we are looking at the acuity of patients by day of discharge. In Table 6.8, the level appears to be relatively constant throughout the week and the ANOVA fails to reject a constant acuity level. This is somewhat surprising because the length of stay varies across the week and presumably the level is a major factor in determining the length of stay. The statistics for the ANOVA can be found in Appendix E.

	Race	*Level*	*Sex*	*Age*
LoS	-.1503	.3722	-.0252	-.1454
Ln(LoS)	-.1570	.4453	-.0165	-.0190
Age	.2680	.0450	.0252	-
Sex	.1425	-.0952	-	-
Level	-.1217	-	-	-

Table 6.9: Correlation table for individual patient statistics. The strongest correlations by far are with the length of stay measures and the level. This indicates that the level is the primary driver for a patient's length of stay.

6.5.2 Relationships Between Variables

This section contains a correlation analysis of the variables. Table 6.9 has the correlations for individual patient's statistics (i.e., *LoS* is the patients length of stay). As with SGL, *Level* is correlated with *LoS*.

Table 6.10 contains the correlations for the patients discharged on a given day. The correlation between *Discharges* and *aLn(LoS)* is particularly high. This implies that the average length of stay is higher on days with more discharges, which is the opposite of SGL. The correlation between *Blocks* and *Discharges* is also strong and positive. Thus, there are generally more discharges on days with more blocks.

6.5.3 Models on Variables

The correlations indicate several interesting relationships that require further investigation to explain more fully. This section contains regression analysis on SCS patients. We include three models: two on the length of stay and one on the number of discharges. The statistics for each of the regressions can be found in Appendix E.

	Blocks	*Discharges*	*Arrivals*
$aLevel$	.2416	.5263	.1964
$aLoS_{-1}$	.0776	.0992	.0512
$aLoS$	.2639	.4175	.1547
$aLoS_{+1}$	.1584	.0150	.0494
$aLn(LoS_{-1})$	.1585	.0002	-.0018
$aLn(Los)$	.3109	.5752	.1892
$aLn(LoS_{+1})$	.0635	.0472	.0202
$Discharges_{-2}$	-.2063	-.0525	-.1364
$Discharges_{-1}$	.0576	.0285	-.0055
$Discharges$	.4193	-	.3856
$Discharges_{+1}$	.2042	-	.1288
$Discharges_{+2}$	-.0966	-	-.0473
$Blocks_{-1}$	-	-	.2258
$Blocks$	-	-	.5917
$Blocks_{+1}$	-	-	.1647

Table 6.10: Correlation table for average daily statistics. For this table, level indicates the average level of patients discharged and x_{-1} indicates the value of x the day before. *Discharges* is strongly correlated with *Level* and *aLn(LoS).* This is very different from SGL where *Discharges* is essentially independent of *aLevel* and *aLn(LoS). Blocks* is positively correlated with *aLevel, Ln(LoS),* and *Discharges.* This is also different from SGL.

SCS Model 1: Ln(LoS)

This model is essentially the same as SGL model 1. To determine how the number of blocks impacts the length of stay, we control for several patient characteristics in the model.

$$Ln(LoS) = a_0 + a_1 * Age + a_2 * Sex + a_3 * Level + a_4 * Race + a_5 * Blocks.$$

The coefficients for *Level, Race,* and *Blocks* are statistically significant and positive. In this case, a positive coefficient indicates $Ln(LoS)$ increases with the variable. A positive coefficient for *Blocks* is somewhat counterintuitive because it indicates that patients are not being discharged sooner than on days with more cardiac surgery block time to make room for new cases, as appears to be the case with SGL.

SCS Model 2: Ln(LoS) with day of week variables.

For this set of regressions, the day of discharge was included in the analysis. This added six variables to the analysis: the days of the week Monday through Saturday. *Monday* is one if the patient was discharged on a Monday and zero otherwise. If the patient was discharged on a Sunday all the day of week indicator variables are zero.

$$Ln(LoS) = a_0 + a_1 * Age + a_2 * Sex + a_3 * Race + a_4 * Level + a_5 * Monday + ...$$

Again, *Race,* and *Level* are statistically significant. In this case, the coefficients for *Monday* through *Thursday* are statistically significant and positive. The coefficient for *Monday* is particularly large. This could imply that some patients are spending additional time in the hospital over the weekends instead of being discharged.

SCS Model 3: Discharges

Like in SGL Model 3, we would like to test how the number of discharges is related to the number of blocks and arrivals, while controlling for the level of patients. For this model, *aLevel* is the average acuity level of patients being discharged on a given day of the week.

$$Discharges = a_0 + a_1 * aLevel + a_2 * Arrivals + a_3 * Blocks$$

The coefficients for *aLevel, Arrivals* and *Blocks* are statistically significant and positive. This implies that there are more discharges on days with more blocks or more arrivals.

6.5.4 Implications

The ANOVA results indicate that discharge volume and patients' Ln(LoS) vary across the week but the acuity level of patients does not vary significantly between days. Like with SGL, *Level* and *Ln(LoS)* are strongly correlated. The correlation between *Blocks* and *Discharges* is positive, as expected, but the correlation between *Ln(LoS)* and *Blocks* is also positive, unlike SGL. Because *Level* is strongly correlated with *LoS* and *Level* does not vary significantly by day of week, additional explanation is required to determine why the length of stay and discharge volume vary by day of week. Because *Level* is essentially constant throughout the week, it might be expected that *LoS* would also be constant. This is not the case and requires some explanation.

Looking at the coefficients of *Blocks* in the SCS Model 1 regression, we see that as the number of blocks increases the expected length of stay also increases. Because the number of blocks varies across the week, this could explain why length of stay varies by day of week, at least in part. Essentially, this means that the number of blocks drives the length of stay variation across the week.

The coefficients in the SCS Model 2 regression fit with the earlier correlation and SCS Model 1 findings. The days Monday through Thursday have the most blocks and have the largest coefficients. The fact that the coefficient of *Monday* is greater than for the other days suggests a “weekend effect,” where patients, instead of being discharged on weekends, spend an extra day or two in the hospital and are discharged on Monday.

The results from SCS Model 3 indicate that the number of discharges tends to increase with increases in the number of blocks or the number of arrivals (new surgeries). This could imply that surgeons are discharging more patients on busier days to make room for the incoming patients.

Both the volume of discharges and patients’ average length of stay increase on days when there are more blocks. This could be explained if surgeons primarily discharge patients when they are pressured to do so by incoming cases from the blocks of operating room time. Thus, patients’ lengths of stay might be artificially increased.

One explanation for these results could involve the surgeons’ schedules. Cardiac surgeons are quite busy with cases, clinic time, research, and teaching. If surgeons spend more time checking in on their patients on days when they are in the operating suite (the OR suite is only a few floors away from the CSICU, unlike the clinic which is held in a different building), there would not be as many discharges on days with less block time (or no block time like weekends). Unlike with SGL, SCS does not share its post-operative units. Therefore, there is no external pressure from other service lines to discharge patients sooner. This explanation would account for the effects seen in the data.

6.6 Comparisons

Surgeons factor in several considerations when making discharge decisions. As discussed earlier, the primary factor for this decision is the patient's medical state, but other factors like block time, day of week, etc. also influence discharge practices. A key difference between the two service lines we investigated is the management of the post-operative bed units. General surgery shares beds with other surgical services, while cardiac surgery has dedicated post-operative units.

Let us assume patients can be discharged on the optimal day, a day early, or a day late without serious implications. Furthermore, let us assume the average length of stay is optimal.

For the most part, general surgeons do not have the option of holding patients longer than their optimal length of stay because of the external pressures from other service lines. Thus, when the utilization is high, the surgeons' options are to discharge the patient on the optimal day or a day early. When the surgeons want to improve the probability that they will be able to create beds for their incoming patients, they will discharge patients a day earlier than usual, if possible (they are not guaranteed to get a bed, but it does improve the likelihood that a bed will be available). Thus, the average length of stay is shorter on days with higher utilization (e.g., days later in the week) as seen in our analysis. Cardiac surgeons, on the other hand, have less external pressure and so they have the option of discharging patients a day later than optimal. Thus, cardiac surgeons can delay discharges until they need the bed for one of their incoming surgical patients. Discharging a cardiac surgery patient today increases the probability that a cardiac surgery post-operative bed will be available for a new surgery. Because the throughput is higher for general surgery than cardiac surgery, the likelihood that a specific surgeon's discharge will lead to an open bed for that surgeon is higher for cardiac surgery. Additionally, because there is a slight risk to discharging patients before the optimal day, the surgeons have a bias toward discharging patients later. This creates a positive relationship between the number of blocks and the length of stay as seen in the analysis.

6.7 Conclusion

Our results indicate Cardiac Surgery and General Surgery had two very different approaches to managing the volume in their service line. SGL patients tended to have shorter lengths of stay on days when the utilization rate was highest (which corresponded to the days with the most SGL blocks). This fact, coupled with the increase in discharge volume on days with more block time, implies that surgeons discharge patients sooner to make room for block time. SCS discharge volume also increased on days with more block time, but unlike with SGL the length of stay was longer on days with more block time and arrivals. This implied that cardiac surgeons did not discharge patients quickly unless post-operative space was needed. This would mean that cardiac surgeons generally discharge patients when they need the room for more patients. Essentially both tend to discharge more patients on days when they have more block time, but the length of stay results imply

SGL surgeons increase discharges by reducing length of stay. On the other hand, it appears that SCS surgeons increase discharges when they have block time by keeping patients until they need the room.

This work implies that surgeons' incentives can impact the discharge policy. Additionally, physician incentives should be taken into account when predicting length of stay and capacity. If post-operative beds are shared, there is an incentive to discharge patients when the service line has block time and the post-operative bed utilization is high. If there is a dedicated post-operative unit, there is no incentive to discharge patients unless the space is required for incoming patients, potentially increasing patient length of stay. For the same reasons, these incentives should be included when generating the block schedule. This work can also be used to develop performance expectations related to discharge practices. Further work should be done to determine the impact that these processes have on patient care and how the surgeons' incentives can be aligned with the hospitals' to improve the flow of patients through the post-operative units.

Chapter 7

Reducing Boarding in a Post-Anesthesia Care Unit

When operating room schedules in hospitals are produced, the constraints and preferences of surgeons and hospital workers are a primary consideration. The downstream impact on post-operative bed availability is often ignored. This can lead to the boarding of patients overnight in the post-anesthesia care unit (PACU) because intensive care unit (ICU) beds are unavailable.

In this paper, we apply integer programming and simulation to develop improved surgical scheduling assignments. The goal is to do a better job of balancing new surgeries with hospital discharges in order to reduce the variability of occupied beds from one day to the next and, as a result, to reduce boarding in the PACU.

7.1 Introduction

Operating room (OR) time is a very valuable and scarce resource in hospitals. When an OR is scheduled, a surgical team of surgeons, anesthetists, nurses, and skilled hospital workers is assigned to the room.

Surgical services are a key generator of revenue. They are frequently expected to subsidize vital service lines (i.e., other hospital departments) that are less profitable. If hospitals can learn to utilize their OR time more effectively, they can serve more patients and serve them better. In this paper, we seek to make progress in this direction by reducing the impact of the bottleneck caused by limited post-operative bed availability.

First, we review the paths taken from OR to recovery by surgical patients. Immediately following surgery, patients recover in the post-anesthesia care unit until the effects of anesthesia wear off. After an hour or two of recovery in the PACU, patients typically move to a downstream bed, as seen in Figure 7.1. In particular, most patients move to an ICU bed for a few days and then to a NonICU bed. Some patients do not require an ICU bed, so they move directly to a NonICU bed. The NonICU beds can be intermediate care or floor beds. When patients are discharged from the NonICU beds, they go home or to another facility. The exact path and length of stay (LoS) varies greatly between service lines and to a lesser extent within a service line.

If some patients cannot leave the PACU by the end of the day because downstream beds are unavailable, they must spend the night in the PACU. This is referred to as PACU boarding. Therefore, the PACU functions as a buffer that limits the impact of post-operative bed

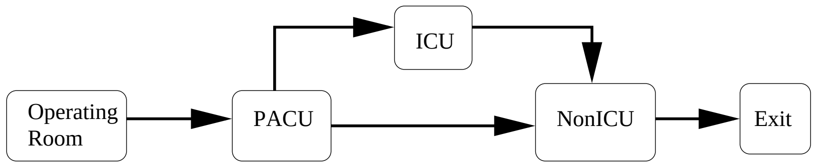

Figure 7.1: The possible paths for post-operative surgical patients. Most patients go to the intensive care unit after they recover in the PACU, but some go straight to an intermediate care bed.

constraints on the OR schedule. PACU boarding is undesirable for several reasons. First, it increases stress on the staff because it is unplanned and an overnight shift for the PACU must be arranged, usually at the last minute. Second, there is evidence that PACU boarding extends a patient's length of stay and increases hospital costs (see Zollinger et al. [30]). The total length of stay is increased because the PACU cannot provide the specialized care delivered in an ICU. Some treatments required by a recovering surgical patient cannot begin until the patient has entered the ICU. So, the time spent in the PACU does not replace time in the ICU. Thus, the time spent boarded in the PACU adds to the total length of stay. The extra time in the hospital increases the total cost of care (extra resources are devoted to the patient) and this cost is transferred to the patient. Third, PACU boarding can have a negative impact on new surgical cases.

The negative impact emerges in the following way. Sometimes, post-operative beds do not become available quickly enough. The PACU becomes full and cannot accept additional patients from the OR. Therefore, patients must recover in their ORs and all ORs go on hold. This means that new surgeries are not allowed to begin in any OR.

The limited capacity of the PACU becomes a major bottleneck as it impacts the utilization of the OR. In order to address the boarding problem, an accurate model of patient flows is required. This model can be used to test proposed solutions. We were asked to develop this model by hospital administrators at the University of Maryland Medical Center (UMMC).

A key factor related to boarding is the variability in demand for beds from one day to the next. The variability in patient flow can be divided into two types: natural and artificial. Emergency surgeries create a natural variability in the number of occupied beds (the hospital census) that is beyond the control of the surgeons and hospital administrators. In contrast, imbalances in the scheduled caseload across the days of the week generate an artificial variability which can lead to boarding; this can potentially be controlled. The overall system may be operating below capacity. However, if, on certain days of the week (due to scheduling), there are significantly more surgical patients arriving than the average number, then boarding is more likely to occur. In Figure 7.2, we show how boarding varied in 2007 and 2008 at UMMC.

We can also focus on the utilization of hospital beds over the days of the week. In Figure 7.3, the pattern of rising utilization throughout the week followed by a drop over the weekend is largely a product of artificial variability (artificial because it is a result of human decision-making as opposed to natural processes). The spike in utilization late in the week indicates that the arrival of patients is not matched with patient discharges. If the flow of patients into the post-operative units was matched with the flow out, the utilization level would not have such a spike. Figures 7.2 and 7.3 both indicate that there are problems with the flow of patients through the hospital system.

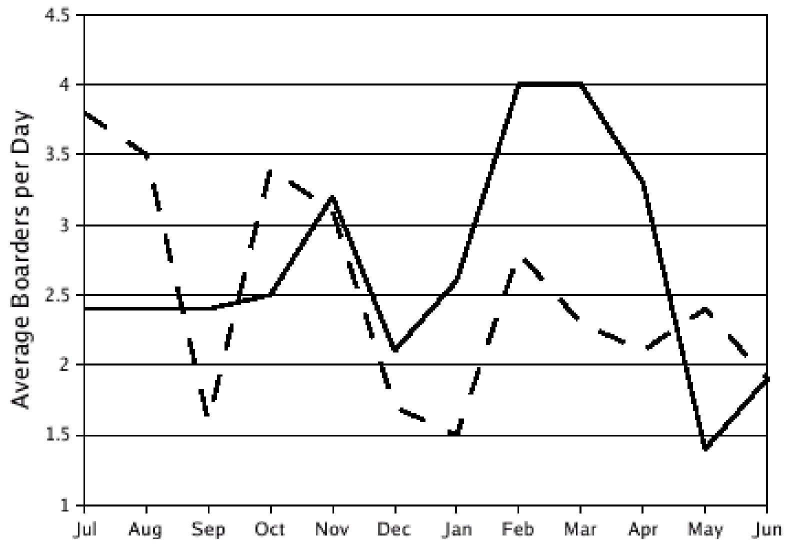

Figure 7.2: The average number of boarders in the PACU per day for each month in the 2007 (dashed line) and 2008 (solid line) fiscal years.

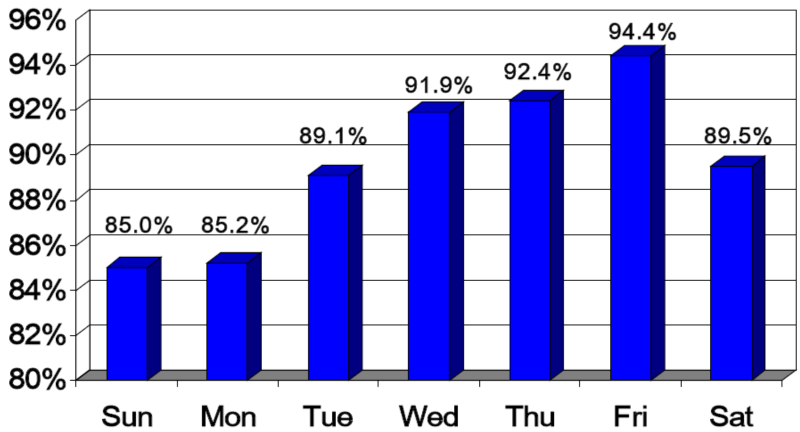

Figure 7.3: The average percent utilization of post-operative beds by day of week.

One source of artificial variability, that is a potential roadblock to the effective scheduling of surgeries, is the use of block schedules. Blocks of OR time are assigned to various surgical services. Each block gives a particular service line (e.g., orthopedics) a specific operating room on a specific day. Individual surgeons are then assigned operating time inside the relevant block for their cases.

In general, surgeons prefer a block schedule because it provides them with a predictable work week (they know exactly which days they will be in the OR). Unfortunately, block schedules can produce problems. If several high-volume service lines are given blocks on the same day, the overall volume (number of surgeries) on these days is likely to be above average. Without sufficient hospital discharges to accommodate the volume of arriving patients, boarding and OR holds are more likely. In order to address this problem, we have developed a new approach to surgical scheduling.

7.2 Literature Review

Several authors have addressed the issues of scheduling block time and the variability of hospital surgical volume. Some have used mathematical programming techniques to improve scheduling.

Blake and Carter [31] use a goal programming model to assign resources to various service lines in a large Canadian hospital with the intent of reducing costs. Contrary to the conventional wisdom in the hospital, the results of their analysis indicated that non-vital departments such as ophthalmology should not be cut back because they bring in sufficient revenue without using too many resources. They also determined that, the hospital's thoracic surgery department, originally thought to be vital, should be cut back because it was too resource intensive. In related work, Blake and Donald [32] use integer programming to assign surgical resources with the purpose of maximizing revenue. Sier et al. [33] use simulated annealing to assign surgical resources (e.g., operating room time) in order to maximize profit.

Jebali et al. [34] assign OR time to specific surgeons based on their availability and other requirements to minimize the cost of running the hospital. First, surgeons are given hours in a specific OR for a given day. Second, the assigned operations are sequenced to minimize overtime. The authors use mixed integer programming and find this approach to be very effective in minimizing costs related to the operating room.

Dexter et al. [35] use an on-line bin-packing approach to assign surgical cases with the goal of maximizing the use of OR capacity. In simulated tests, their approach results in a higher utilization and throughput. However, the simulation did not take into account the downstream bottlenecks in the system.

McManus et al. [25] find large variability in the daily surgical caseloads and relate this variability to problems in the downstream ICU. The most interesting result of their work involves the nature of the variability. The variability of the patient flow is more highly correlated with the scheduled caseload than the unscheduled (emergency) volume. This means that most of the variability comes from the schedule developed within the hospital. Thus, their statistical analysis indicates that improving the surgical schedule is needed to reduce the variability of the patient flow and related problems.

Belien and Demeulemeester [36] outline several approaches to generating surgical schedules that minimize the maximal bed occupancy. This problem is analogous to minimizing the maximum expected number of boarders in the PACU. They develop several integer programming models to minimize the maximum occupancy using hospital data. Their work uses hospital data, but they do not test their models using simulation.

Santibanez et al. [37] use a mixed integer programming model to explore trade-offs in efficiency when assigning block time. They seek to develop schedules that minimize the maximum occupancy. In practice, it is difficult to generate an optimal solution. Most hospitals do not have access to the software or expertise to use a mathematical programming approach.

In the remainder of this chapter, we develop a flexible approach that generates a surgical schedule for UMMC. In Section 7.3, we describe the data set that we use in our analysis. In Section 7.4, we argue that the grouping of surgical service lines provides a desirable degree of flexibility. In Section 7.5, we formulate an integer program (IP) to develop a surgical block schedule. We use the optimal solution from this IP as a starting point to develop rules of thumb for constructing a more flexible block schedule in Section 7.6. In Section 7.7, we perturb the historical block schedule by swapping blocks using our rules of thumb as a guideline. In Section 7.8, we compare a number of different schedules using simulated data.

7.3 Data Set

Our data set contained detailed information on every surgical patient from January 2007 to May 2007 at UMMC including the length of time the operating room was occupied for

Service Line	Surgery Time (Minutes)	Mean LoS (Days)	Mean Patients per Week	Standard Deviation of Patients per Week
Gynecology	108.36	1.18	11	3.42
Ophthalmology	88.90	0.33	10	4.46
Urology	160.71	1.85	14	3.23
General	179.37	3.08	22	5.71
Oral	191.56	3.60	19	3.68
Otolaryngology	138.09	2.49	29	6.06
Plastic	194.08	3.22	17	4.99
Vascular	182.09	4.31	9	2.48
Neurosurgery	229.66	5.35	27	3.29
Organ Transplant	216.74	6.91	13	4.61
Orthopedics	207.71	4.88	54	7.69
Surgical Oncology	225.82	4.74	10	3.74
Thoracic	202.38	5.15	7	1.73

Table 7.1: Summary statistics for 13 service lines from January 2007 to May 2007.

each surgery (measured as the time the patient enters the OR until the patient leaves the OR) and the time the patient spent in a post-operative bed (measured as the time from PACU discharge to discharge from the hospital). Summary statistics for 13 service lines are given in Table 7.1.

It is important to note the variability in the number of patients per week within each service line. The last two columns in Table 7.1 contain the average number of patients per week and the standard deviation of the number of patients per week for each service line. There is clearly a difference in both the average flow of patients and the variability of the flow across the service lines. To better illustrate this variability, in Figure 7.4, we show the number of patients per week for the orthopedics, general, and ophthalmology service lines. As discussed in the work by McManus et al. [25], high variability can cause problems with scheduling because the exact amount of OR time required is dependent on the number and

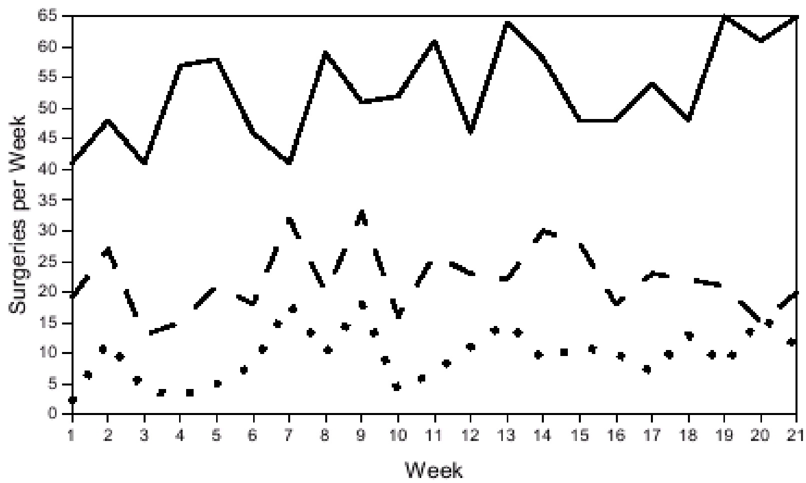

Figure 7.4: The number of surgeries per week performed by the orthopedics (solid line), general (dashed line), and ophthalmology (dotted line) service lines between January 2007 and May 2007.

the specific nature of the cases. The greater variability in each service line's case volume increases the difficulty in determining the post-operative resource requirements (ICU beds, staffing levels, etc.). This variability can also exacerbate the problem of bed capacity. For example, if multiple service lines have high volume weeks, it might not be possible to accommodate all patients without boarding.

7.4 Grouping the Service Lines

UMMC administrators preferred to block schedule groups of service lines instead of individual service lines. This provided them with flexibility when developing a schedule. With this preference in mind, we clustered the patients into three groups with similar volume and length-of-stay characteristics. The first group has gynecology, ophthalmology, and urology. Patients in this group have relatively short lengths of stay with relatively high volumes. The second group contains general, oral, otolaryngology, plastic, and vascular surgeries. The third group has neurosurgery, oncology, organ transplant, orthopedics, and thoracic surgeries. This group has patients with lower volumes and longer lengths of stay. We did not include cardiac surgery and pediatric surgery because these service lines have their own ICU and NonICU beds and do not use the PACU.

The block schedule from January 2007 to August 2007 is displayed in Table 7.2. We refer to this as the historical schedule. The base unit of the schedule is one operating room for an entire day every week. Some service lines receive an operating room for only the morning or the afternoon, which is denoted by a half-block. The schedule is made for a five-week period and some groups are not given the operating room every week. Because one half-day per five-week period is the smallest unit allowed, we will work with a tenth of a block.

Table 7.3 has data about each of the groups collected from the primary data set and the block schedule. The minimum, maximum, and required OR times were determined from the historical block schedule. A minimum number of blocks must be assigned to certain service lines each day because they need available operating room time every day to meet demand. For example, neurosurgery has two blocks every day of the week. Each service line has a maximum number of blocks each day. In order to meet demand, the schedule is required to assign a certain number of blocks to each service line. In Figure 5, we show a graphical representation of the historical schedule in terms of each group. The size of

Service Line	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Gynecology	1.0	1.0	0.6	1.0	0.0	3.6
Ophthalmology	0.3	1.0	0.0	0.6	0.2	2.1
Urology	1.0	1.0	1.0	1.0	1.4	5.4
Group 1 Totals	2.3	3.0	1.6	2.6	1.6	11.1
General	1.0	2.0	1.6	2.0	2.9	9.5
Oral	2.0	1.0	0.0	1.0	1.6	5.6
Otolaryngology	2.0	2.0	2.0	1.0	1.0	8.0
Plastic	1.4	0.5	0.6	0.0	0.5	3.0
Vascular	0.5	1.0	1.0	1.0	1.0	4.5
Group 2 Totals	6.9	6.5	5.2	5.0	7.0	30.6
Neurosurgery	2.0	2.0	2.0	2.0	2.0	10.0
Organ Transplant	0.4	0.6	1.0	1.0	1.2	4.2
Orthopedics	2.0	2.0	1.4	2.0	1.4	8.8
Surgical Oncology	0.5	0.0	0.5	2.0	1.6	4.6
Thoracic	0.6	1.0	1.0	1.0	0.0	3.6
Group 3 Totals	5.5	5.6	5.9	8.0	6.2	31.2
Total	14.7	15.1	12.7	15.6	14.8	72.9

Table 7.2: The block schedule for January 2007 to August 2007. Each entry is the number of operating rooms blocked for a service line on each day of the week over five weeks.

	Minimum Blocks	Maximum Blocks	Total Blocks	ICU LoS	NonICU LoS	Total LoS	Patients per Block
Group 1	1.0	5.0	11.1	0	1	1	2.5
Group 2	2.0	13.0	30.6	1	2	3	1.5
Group 3	3.0	15.0	33.2	2	3	5	1.25

Table 7.3: Basic statistics about each group. Minimum, maximum, and total blocks were determined from the historical block schedule. The values for ICU LoS, NonICU LoS, and Total LoS are the median values and were rounded to the nearest whole day. The number of patients per block was estimated by dividing the length of surgical day by the mean set up time plus the mean case time plus the mean clean up time.

each group is proportional to the number of blocks assigned to that group on that day. The number of blocks assigned to each group is also displayed in Figure 7.5.

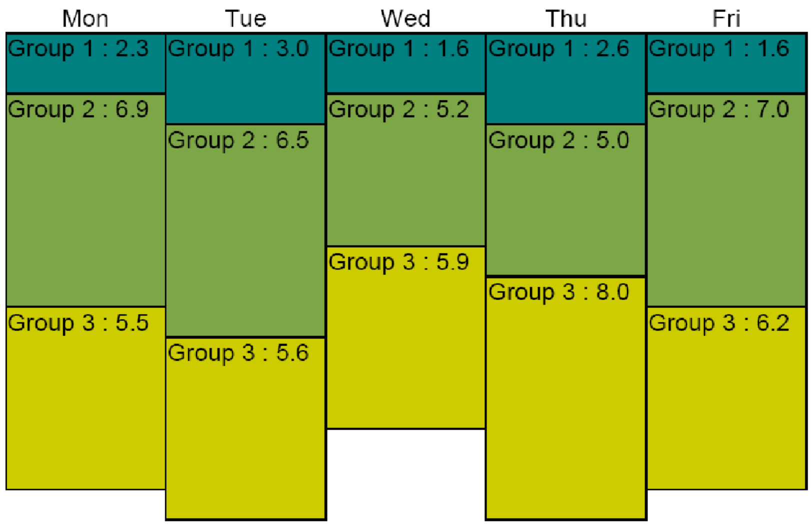

Figure 7.5: A graphical representation of the historical block schedule.

7.5 Developing Block Schedules

The problem of PACU borders arises when more patients arrive into the PACU than leave in one day. If the arrivals into the PACU were exactly matched with hospital departures, PACU boarding could be avoided and the utilization of beds could remain high because each discharge would make room for a new surgical case. In practice, there are two flows of patients: the ICU flow and the NonICU flow. Boarding is due to insufficient ICU beds or insufficient NonICU beds. Furthermore, if the NonICU beds are full, then the ICU patients cannot be transferred to a NonICU bed. This leaves a patient in an ICU bed when a NonICU bed should be used, because there are no available NonICU beds. Keeping patients in ICU beds when they do not need that level of care is wasteful because the ICU beds are an extremely scarce resource and this exacerbates the problem with the ICU flow. When ICU patients can't move from ICU beds to NonICU beds, there is a higher probability there will be PACU boarders because the ICU capacity has effectively been reduced.

We need to determine how to balance the flow of arrivals into the units with the flow of departures out of the units. One way to achieve this balance is to improve the block schedule. We now develop an integer program for the block assignment problem. The blocks are assigned to match the flow into the ICU with the flow out of the ICU. Our initial experimentation indicated that including the flow of NonICU beds in the IP did not improve the results. Essentially, if the patient flow into and out of ICU beds is well-balanced, then the flow into and out of the NonICU beds also tends to be well-balanced. Furthermore, the PACU boarding of patients who belong in ICU beds is more detrimental to patient care. Therefore, we focus on the flow of ICU patients in our IP model. An additional benefit is that the size of the IP is reduced.

Let b_{ij} be the number of blocks assigned to group i on day j and n_i be the total number of blocks group i requires. We define n_{OR} to be the number of ORs available. Based on the UMMC data, there were 16 ORs available daily. The minimum and maximum number of blocks that can be assigned to group i are Min_i and Max_i, respectively. The average ICU length of stay for patients in group i is given by μ_i. The expected value for the number of patients in one block of group i is represented by λ_i. The expected number of arrivals on day j is the sum over all the groups of the number of blocks for a group multiplied by the expected number of patients per block ($\sum_i \lambda_i b_{ij}$). If patients of group i are expected to stay μ_i days in the ICU, the patients discharged from the ICU on day j are expected to have arrived at the hospital on day $j - \mu_i$. Thus, the expected number of discharges on day j will be $\sum_i \lambda_i b_{i(j-\mu_i)}$. ICU_j^- (ICU_j^+) is the negative (positive) change in the ICU census (the number of occupied ICU beds) on day j. If the ICU was full on day $j-1$, ICU_j^+ would be the number of patients boarded in the PACU on day j because ICU_j^+ is the number of patients arriving in the ICU minus the number discharged. The values of $Min_i, Max_i, n_i, \lambda_i,$ and μ_i are given in Table 3. ICU_j^-, and ICU_j^+ are nonnegative integer variables. On each weekday j, b_{ij} can take nonnegative integer values between Min_i and Max_i for each service line i.

Our IP is given by the following objective function and constraints.

$$\text{Min} \sum_j ICU_j^+ \tag{7.1}$$

s.t.

$$\sum_j b_{ij} = n_i \ \forall i \tag{7.2}$$

$$\sum_i b_{ij} \leq n_{OR} \ \forall j \tag{7.3}$$

$$\sum_i \lambda_i b_{ij} - \sum_i \lambda_i b_{i(j-\mu_i)} + ICU_j^- - ICU_j^+ = 0 \ \forall j \tag{7.4}$$

$$Min_i \leq b_{ij} \leq Max_i \ \forall i, j \tag{7.5}$$

$$ICU_j^-, ICU_j^+, b_{ij} \geq 0 \text{ and integer } \forall i, j. \tag{7.6}$$

The objective (7.1) minimizes the number of expected arrivals (new surgical cases, $\sum_i \lambda_i b_{ij}$) beyond the expected number of discharges (patients that have spent enough time in the

ICU to have recovered, $\sum_i \lambda_i b_{i(j-\mu_i)}$). The objective function does not explicitly minimize the number of boarders; it matches the flow of patients into the ICU with the flow out of the ICU. Because of the high utilization rate of the post-operative beds, this objective is roughly equivalent to minimizing the number of boarders, but the formulation is simpler. Note that if we could could push the objective function value to near-zero, PACU boarding would essentially vanish. That is, by reducing the spikes in the ICU census, we reduce PACU boarding. Constraint (7.2) ensures that each group's demand for blocks is met. Constraint (7.3) ensures that there can be no more blocks assigned in one day than there are available operating rooms. Constraint (7.4) matches the expected number of ICU patients arriving ($\sum_i \lambda_i b_{ij}$) with the expected number of ICU patients being discharged from the ICU ($\sum_i \lambda_i b_{i(j-\mu_i)}$). The $ICU_j^- - ICU_j^+$ portion of equation (7.4) determines the imbalance between arrivals and departures on day j. Certain service lines require a minimum number of blocks in a day. For staffing reasons, each group has a maximum number of blocks that can be used in one day. Constraint (7.5) restricts the number of blocks assigned to a group to be within the maximum and minimum allowable values. Constraint (7.6) ensures that ICU_j^-, ICU_j^+, and b_{ij} are nonnegative and integer.

The IP for the UMMC problem has 15 variables (5 weekdays $\times$ 3 groups) and 28 constraints (three from (7.2), five from (7.3), five from (7.4), and 15 from (7.5)). Our IP is simpler than the formulations in [36, 37]. The exact formulation used for UMMC can be found in Appendix **??**.

It is important to clarify how we adjust the data and model to exclude consideration of

Service Line	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	3.1	5.0	1.0	11.1
Group 2	5.8	12.0	7.1	3.7	2.0	30.6
Group 3	9.2	3.0	3.0	3.0	13.0	31.2
Total	16.0	16.0	13.2	11.7	16.0	72.9

Table 7.4: The number of blocks assigned to each group on Monday to Friday. Entries are based on the optimal solution to the integer program.

NonICU beds. Suppose Group 1 patients typically move directly from the PACU to NonICU beds. In our IP model, they exit the hospital system from the PACU. Suppose other patients typically spend three days in ICU beds and then two days in NonICU beds. In our model, after three days in ICU beds, these patients would exit the hospital system.

Using the data from January 2007 to May 2007 to determine the required number of blocks, the integer program was solved using CPLEX [38] and SCIP [39], a non-commercial IP solver (see http://scip.zib.de/ for details). We chose a freely available, non-commercial solver (as opposed to a commercial solver) to show administrators that sophisticated, no-cost software was readily available to solve their problem.

In Table 7.4, we show the number of blocks assigned to each service line on Monday to Friday based on the optimal solution to our IP. A graphical representation of the block schedule is shown in Figure 7.6. We point out that we scaled the b_{ij} variables when we solved our IP in order to maintain integrality. For example, the optimal value for Group 2 on Monday in the IP solution is 58 which is 5.8 blocks in Table 7.4.

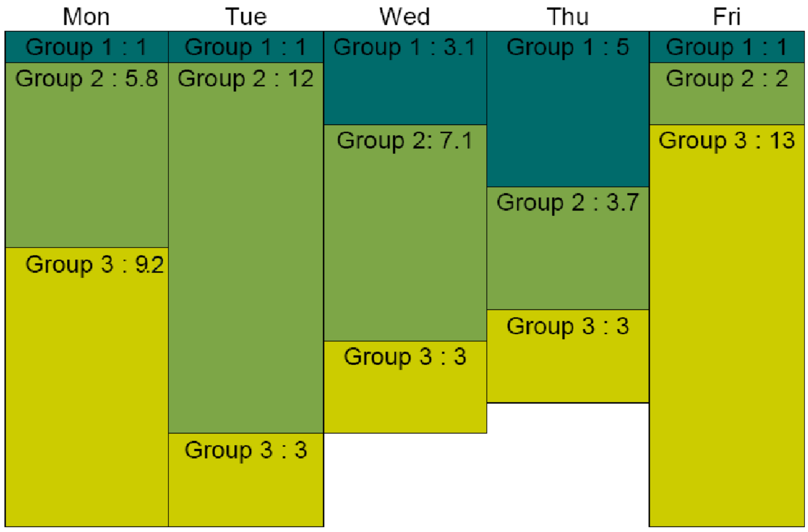

Figure 7.6: A graphical representation of the number of blocks assigned to each group as determined by the optimal solution to the IP.

7.6 Rules of Thumb

The hospital administrators liked the IP model, but wanted more flexibility in setting block schedules. The IP cannot take all constraints and preferences into account because not all are known until the schedule is negotiated with the heads of each surgical service line. By using groups of service lines instead of individual service lines, flexibility was increased, but that still might not be good enough. If a surgical service line cannot accept the IP schedule for any reason, the schedule is infeasible. For instance, if the clinic schedule prevents the Group 1 service lines from assigning their designated blocks on Thursdays, the IP model would be infeasible. While it is unrealistic to expect a series of IPs to be formulated and solved over the course of the block schedule meeting (which lasts several hours), the optimal solution to the IP can be used to provide managerial insight and serve as a starting point for an implementable schedule. In other words, the patterns emerging from the IP solution (e.g., see Figure 6) were useful and revealing, but typically not immediately implementable. If a flexible set of rules of thumb could be generated, these rules could be applied during the meeting to construct a final schedule.

First, we made some basic observations about the IP solution. Most of Group 1's blocks should be on Wednesday and Thursday. For Group 2, Tuesday is a high volume day and Monday, Wednesday, and Thursday are moderate volume days. Monday and Friday are high volume days for the Group 3 service lines. Given these observations, we created general guidelines or rules of thumb and demonstrated their use to administrators. The rules and incremental results are presented next.

1. Meet the minimum daily demand requirements for each service line using the minimum block numbers from Table 7.3. In this case, we assigned more blocks to Group 3 because a primary complaint with the IP schedule involved the lack of Group 3 mid-week block time.

Service Lines	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	1.0	1.0	1.0	5.0
Group 2	2.0	2.0	2.0	2.0	2.0	10.0
Group 3	4.0	4.0	4.0	4.0	4.0	20.0
Total	7.0	7.0	7.0	7.0	7.0	35.0

2. Split the block time for Group 1 between Wednesday and Thursday.

3. Schedule as many Group 2 blocks as possible on Tuesday. Hospital administrators thought that there were too many blocks assigned to the Group 2 service lines on

Service Lines	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	4.1	4.0	1.0	11.1
Group 2	2.0	2.0	2.0	2.0	2.0	10.0
Group 3	4.0	4.0	4.0	4.0	4.0	20.0
Total	7.0	7.0	10.1	10.0	7.0	41.1

Tuesday (12 blocks in the IP solution). Therefore, we reduced the maximum capacity from 12 blocks to 9 blocks.

Service Lines	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	4.1	4.0	1.0	11.1
Group 2	2.0	9.0	2.0	2.0	2.0	17.0
Group 3	4.0	4.0	4.0	4.0	4.0	20.0
Total	7.0	14.0	10.1	10.0	7.0	48.1

4. Spread the remaining Group 2 block time across Monday, Wednesday, and Thursday.

Service Lines	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	4.1	4.0	1.0	11.1
Group 2	6.0	9.0	6.0	6.0	3.6	30.6
Group 3	4.0	4.0	4.0	4.0	4.0	20.0
Total	11.0	14.0	14.1	14.0	8.6	61.7

5. Schedule as many of the Group 3 blocks as possible on Monday and Friday. Because of the limited OR capacity, the maximum number of blocks that can be scheduled on one day is 16. Thus, no more than 9 blocks may be assigned to Group 3 on Monday.

A graphical representation of the block schedule after applying the five rules of thumb is shown in Figure 7.7. This figure looks quite similar to the schedule in Figure 7.6. This

Service Lines	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	4.1	4.0	1.0	11.1
Group 2	6.0	9.0	6.0	6.0	3.6	30.6
Group 3	9.0	4.0	4.0	4.0	10.2	31.2
Total	16.0	14.0	14.1	14.0	14.8	72.9

similarity indicates that the rules gave additional flexibility to the administrators, but resulted in a schedule that has a similar structure to the optimal schedule developed by the IP.

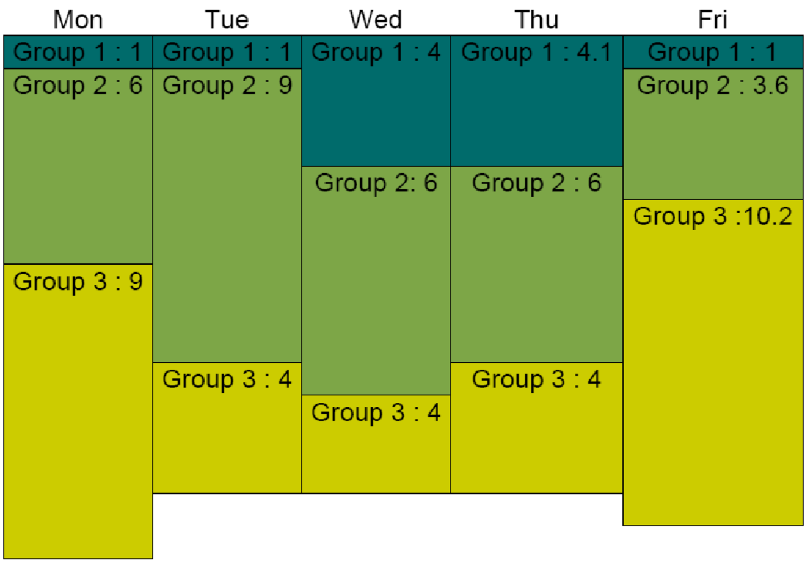

Figure 7.7: A graphical representation of a block schedule constructed using the rules of thumb approach.

7.7 Revising the Historical Schedule

Because the historical schedule is very different from the IP schedule, the hospital administrators were concerned that the heads of the surgical service lines would reject the schedules developed by the IP model or the rules-of-thumb approach. The total absolute

Service Lines	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.3	3.0	2.6	2.6	1.6	11.1
Group 2	5.9	7.5	5.2	7.0	5.0	30.6
Group 3	7.5	4.6	4.9	6.0	8.2	31.2
Total	14.7	15.1	12.7	15.6	14.8	72.9

Table 7.5: Entries are the number of blocks assigned to each group on Monday to Friday based on the swaps using the schedule from the rules-of-thumb approach.

deviation of the historical schedule from the IP schedule ($\sum_{i,j} |IP\ b_{ij} - historical\ b_{ij}|$) is 43.6 blocks (there are a total of 72.6 blocks). The total absolute deviation for the rules of thumb approach from the IP schedule is 16.2 blocks (33.2 blocks from the historical schedule). Because the various stakeholders (e.g., administrators and surgeons) have different objectives, a complete revision of the block schedule, even with the added flexibility of the rules-of-thumb approach, might be difficult to implement. Therefore, we began to look for small revisions to the historical schedule that could result in significant improvements to its effectiveness. Using the rules-of-thumb schedule as a guide, we looked for swaps (exchanges) that could reduce the number of boarders. Based on the suggestions of the hospital administrators, we considered three types of swaps.

1. One Group 1 block on Monday for one Group 3 block on Wednesday.

2. One Group 2 block on Monday for one Group 3 block on Tuesday.

3. Two Group 3 blocks on Thursday for two Group 2 blocks on Friday.

The results from revising the historical schedule are given in Table 7.5.

7.8 Comparisons on Simulated Data

There are two concerns with the IP model: it uses deterministic values for length of stay and it omits the NonICU flow. We constructed a simulation model to test the IP schedule against other scheduling approaches. Discrete event simulation allows us to use the data set covering five months of surgical cases for additional experiments. Without simulation we could not compare schedules because the historical data (based on expected values) does not account for the stochastic nature of a patient's length of stay (recall that in Sections 7.5, 7.6, and 7.7, we assume that the length of stay for each group of patients is deterministic.)

As discussed in Section 5, the IP focuses on the flow of patients into and out of the ICU. While this is the primary source of PACU boarding, the flow of patients into and out of the NonICU beds must also be taken into account. A simulation allows us to include this flow and conduct an analysis using thousands of months of hypothetical data.

The simulation was designed in MATLAB [40]. In each block of OR time, a random number of cases were performed based on the historical volume distribution for the group assigned to the block. These volume distributions were created using the number of cases per block for each group from the primary data set (January 2007 to May 2007). Each case represents a single patient with a total LoS based upon the empirical LoS distribution for the group. Consistent with administrators' impressions of length-of-stay patterns and the available data, the ICU LoS was randomly distributed between one-sixth and one-half of the total LoS (we assumed a uniform distribution). The remaining LoS was used for the NonICU LoS. Each patient's LoS values were rounded to the nearest whole day.

We are interested in the steady-state behavior of the post-operative patient flow. In order to eliminate transient effects, the simulation was warmed up for five weeks (35 days). Data was then collected for the next 10 weeks (70 days). The simulation was run 10,000 times for each scheduling approach and the results were averaged.

For our analysis, we focused on the number of boarders and the variability in the census. The census is the number of patients in a unit and the capacity is the number of patients that a unit can hold. We calculated the number of boarders on a given day i using equation (7.7):

$$Boarders_i = \text{Max}(Census_{ICU} - Capacity_{ICU}, 0). \tag{7.7}$$

Patients are boarded when the census ($Census_{ICU}$) is greater than the hospital's capacity ($Capacity_{ICU}$). In addition to boarders, we examined the standard deviation of the ICU census to determine if the variability had been reduced.

The simulation results are given in Table 7.6. The hospital had 31 ICU beds for the 13 service lines. We report the mean, 5^{th} percentile, and 95^{th} percentile of the number of boarders per day and the mean and standard deviation of the ICU census. In Table 7.6, we show results from the historical schedule, the schedule constructed from the IP solution, the schedule developed using the rules-of-thumb approach, and the revision of the historical schedule. We also give the results of an even schedule that distributes each group's blocks equally across the week. The schedules produced by Even, IP, Thumb, and Revised do a better job of reducing the number of boarders and reducing the standard deviation in the census than the historical schedule. The fact that there was very little difference be-

Schedule	Bottom 5% Boarders	Mean Boarders	Top 95% Boarders	Mean Census	Census Standard Deviation
Historical	3.357	4.670	6.057	30.933	11.702
Even	3.200	4.501	5.886	30.939	11.365
IP	2.700	4.003	5.500	30.943	10.758
Thumb	2.729	4.024	5.457	30.939	10.649
Revised	3.071	4.316	5.686	30.929	11.107

Table 7.6: These are the averages for 10,000 runs of the simulation representing a 10-week period with a capacity of 31 beds. Historical refers to the block schedule currently in use. Even distributes the blocks evenly across the week. IP is the block schedule generated from the IP model. Thumb is the schedule based on the rules of thumb derived from the IP model solution. Revised is the revision of the historical schedule generated by applying swaps. Boarders were calculated using equation (7.7). Bottom 5% Boarders (Top 95% Boarders) is the 5^{th} (95^{th}) percentile of boarders from the 10,000 runs. Mean Boarders is the mean number of boarders per day. Mean Census is the average daily ICU census and Census Standard Deviation is the standard deviation in the daily census.

tween the IP schedule and the rules-of-thumb schedule indicates that the IP solution is not very sensitive to minor changes as long as the general guidelines from the rules-of-thumb approach are followed. The lack of sensitivity is important because the schedule developed from the IP model might not be implementable. By making simple swaps to the historical schedule, the revised historical schedule achieves significant reductions in boarders over both the historical and the even schedules. Furthermore, the revised schedule has a much better chance of acceptance by surgeons, hospital staff, and administrators because it only requires a few swaps of blocks between groups.

The efficiency of a change in the schedule can be thought of as the average reduction in boarders per week per swap from the historical schedule. That is, efficiency = $\frac{\text{Weekly Reduction}}{\text{Swaps}}$. If the hospital decided to maintain the current level of boarding as in the historical schedule by admitting additional surgical patients, there would be an increase in the

Schedule	Swaps	Weekly Reduction	Standard Deviation	Efficiency	Potential Increase in Net Revenue
IP	22	4.69	10.76	0.21	\$3.5 million
Thumb	16	4.55	10.65	0.28	\$3.4 million
Revised	4	2.45	11.11	0.61	\$1.8 million

Table 7.7: Comparing schedules in five ways. Swaps is the number of block swaps required to transform the schedule into the historical schedule. Weekly reduction is the average number of fewer boarders per week as determined by the simulation. Efficiency is the weekly reduction divided by swaps. Potential increase in net revenue was determined by multiplying the weekly reduction in borders by the number of weeks in a year (50) and an estimate of the hospital's average net revenue per surgical case ($\$15,000$). If the hospital maintained the same level of boarding as in historical system by increasing the patients in the system, the potential increase in net revenue would be the resulting increase in the hospital's profit.

hospital's surgical volume resulting in an increase in the net revenue from the additional surgeries. An estimate of this value can be calculated, by multiplying the estimated additional capacity (the reduction in boarders) and the net revenue per case. Thus, the potential annual increase in net revenue would be $50\times$fewer boarders per week$\times\$15,000$ (see Table 7.7).

7.9 Conclusions and Further Work

Our work has demonstrated that a simple IP can generate a practical and implementable block schedule using rules of thumb. Our revised schedule showed that a small number of easy alterations have a large impact in reducing the number of boarders. By considering the patient flow of all service lines together, decisions can be made to reduce the number of boarders. This highlights the need for a systemwide approach to improve hospital performance.

By assigning the surgical block time to improve the flow of patients through the system, UMMC can greatly reduce the number of boarders. Our work showed that matching patient inflow and outflow can be effective in reducing boarders. In particular, even if the variability in arrivals were reduced to zero, there would still be boarders in the PACU unless the arrivals and departures were exactly matched.

Further work is needed to determine additional steps that could be taken to maintain throughput while reducing boarding. A simulation platform could be used to test the effects of adding beds, adding operating rooms, and altering the number of operating room blocks for a given service line.

UMMC has been looking into using the rules-of-thumb approach to determine how to adjust the block schedule. We have had several meetings with hospital administrators and their feedback has been very encouraging.

Chapter 8

Conclusion

Improving health care efficiency is becoming an increasingly important part of the health care policy debate in the United States. In order to expand health care access while maintaining some semblance of fiscal responsibility and the current quality, hospital systems must be more efficient. In this dissertation, we have investigated approaches to improving collection of blood donations, delivery of emergency supplies, and patient flow using techniques from mathematics and operations research.

We analyzed different optimization problems using a variety of techniques. The work on the travelling salesman problem with a center looked at the balance between different objectives and found the "best" tours for the TSPwC are highly dependent on the relative weightings. The TVP work found a similar sensitivity though the behavior was different. The hospital work also used optimization in addition to data analysis and simulation. We determined the optimal bed mix for the cardiac surgery post-operative units using a simulation model. The prediction work failed to accurately determine the future census in practical tests, but this failure highlighted some behaviors related to the discharge practices that needed some explanation. We looked into these discharge practices and found that the current scheduling approach caused problems with patient flow. In the scheduling chapter, we looked into practical approaches to improving patient flow. Through many iterations and tests, we found scheduling approaches that can practically alleviate some of the problems associated with the discharge practices that we analyzed.

Each of these pieces highlights the usefulness of applying techniques from mathematics to problems in health care.

Appendix A

Mean Return Time

To solve $\int_0^1 \sqrt{(\omega x_i + (1-\omega)x_j - x_c)^2 + (\omega y_i + (1-\omega)y_j - y_c)^2} d\omega$, we should first use some trigonometry. Looking at Figure A.1, it should be clear that the length of the dotted line is $d_n(i, j)\sec(\theta)$. Next if we integrate this over all possible θ, we have

$$\int_0^{\alpha} d_n(i, j)\sec(\theta)d\theta + \int_0^{\beta} d_n(i, j)\sec(\theta)d\theta \tag{A.1}$$

for edges where $\max(d_e(i, c), d_e(j, c)) < d_e(i, j)$. In the other case, we have

$$\int_{\beta}^{\alpha} d_n(i, j)\sec(\theta)d\theta. \tag{A.2}$$

We can calculate α and β using the law of cosines. It is also important to note that

$$\int_{\beta}^{\alpha} \sec(\theta)d\theta = ln\frac{\sec(\alpha) + \tan(\alpha)}{\sec(\beta) + \tan(\beta)}.$$

This can be used to find the mean return time for a vehicle travelling between two nodes. By averaging over each pair of nodes in a tour, we have the mean return time for a tour.

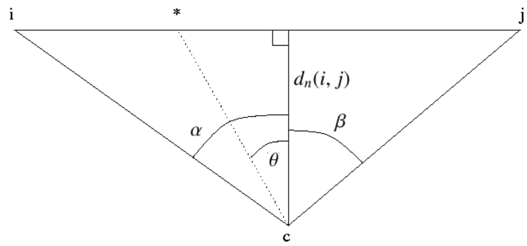

Figure A.1: Here i and j are nodes, c is the center, and $*$ is a point along the edge (i, j). $d_n(i, j)$ is the nearest point distance along this edge. α is the angle between (i, c) and $d_n(i, j)$, β is the angle between (j, c) and $d_n(i, j)$, and θ is the angle between $(*, c)$ and $d_n(i, j)$.

Appendix B

CPT Codes

This is a partial list of CPT Codes and the abbreviation used in the clustering with some explanation.

Bypasses
1C 36.11 Aortocoronary Bypass 1 Coronary Artery
2C 36.12 Aortocoronary Bypass 2 Coronary Artery
3C 36.13 Aortocoronary Bypass 3 Coronary Artery
4C 36.14 Aortocoronary Bypass 4+ Coronary Artery
1M 36.15 1 Internal Mammary-Coronary Bypass
2M 36.16 2 Internal Mammary-Coronary Bypass

Annuloplasty
AN 35.33 Annuloplasty

Heart Transplantation
Hrt 37.51 Heart Transplantation

Lung Transplantation
Lng 33.51 Unilat Lung Transplant
Lng 33.52 Bilat Lung Transplant

VAD
VAD 37.66 Insertion of Implantable Heart Assist System
VAD 86.07 Insert Vascular Assistance Device

Replacements
Rpl 35.2 Replacement of Unspecified Heart Valve
Rpl 35.21 Replacement of Aortic Valve with Tissue Graft
Rpl 35.22 Other Replacement of Aortic Valve

Rpl 35.23 Replacement of Mitral Valve with Tissue Graft
Rpl 35.24 Other Replacement of Mitral Valve
Rpl 35.25 Replacement of Pulmonary Valve with Tissue Graft

Pericardiotomy
Pc 37.12 Pericardiotomy
Pc 37 Pericardiocentesis

Valvuloplasty
Vp 35.11 Open Aortic Valvuloplasty
Vp 35.12 Open Mitral Valvuloplasty
Vp 35.13 Open Pulmonary Valvuloplasty
Vp 35.14 Open Tricus Valvuloplasty

Angioplasty
Ap 39.5 Angioplasty or Atherectomy of Other Non-Coronary Vessel(s)
Ap 0.61 Percutaneous Angioplasty or Atherectomy of Precerebral (Extracranial) Vessel(s)
Ap 0.66 Percutaneous Transluminal Coronary Angioplasty or Coronary Atherectomy

Excision
Ex 34.59 Other Pleural Excision
Ex 37.32 Heart aneurysm excision
Ex 37.33 Excision or destruction of other lesion or tissue of heart open approach
Ex 37.34 Excision or destruction of lesion or tissue of heart other approach

Repair
Rp 37.4 Heart and Pericardium repair
Rp 37.49 Other repair of heart and pericardium
Rp 35.52 Aneurysm repair NEC
Rp 35.53 Prost Repair ventricular def
Rp 35.57 Rep vess with synthetic patch
Rp 35.62 Graft repair ventric def
Rp 35.71 Atria Septa Def Rep NEC
Rp 35.72 Other and unspecified repair of ventricular septal defect
Rp 35.81 Total repair tetral fallot
Rp 35.82 Total Repair of TAPVC

Resection
Rs 38.34 Aorta Resection and Anast
Rs 38.44 Resection of abdominal aorta with replacement
Rs 38.45 Resection Thorac ves with repl
Non-implantable heart assist systems

Im 37.62 Insertion of non-implantable heart assist system
Im 37.68 Insertion of percutaneous external heart assist device

Defibrillator
D 37.94 Implant/replace cardiacdefibrillator total
D 89.49 Automatic implantable cardioverter/defibrillator(AICD) check
D 0.5 Implantation of cardiac resynchronization pacemaker without mention of defibrillator
D 37.41 Implantation of prosthetic cardiac support device around the heart

Appendix C

Simulation Sampling Bias

This appendix is included because the initial simulation approach for testing the capacity predictions resulted in a biased sampling of points. The results in this appendix highlight the importance of ensuring the appropriateness of the simulation construction for the specific problem being modeled.

If patients are sampled from a data set similar to the one used in the length of stay prediction chapter, and each patient i's current length of stay (LoS) is taken to be $c_i \sim U[1, t_i]$, where t_i is the total length of stay for patient i, then the sampling of current length of stay will result in a bias for the remaining length of stay.

Proof: Let c_k be the current LoS of a patient with total length of stay t_k. Assume m is the longest length of stay in the data set. Also, let n_i be the number of patients with length of stay i and n be the total number of patients in the data set.

Let P^* be the probability determined by simulation.

Assume$c_k \sim U[1, t_k]$. So,

$$P^*(c_k = j | t_k = i) = \begin{cases} \frac{1}{i} & \text{if } i \geq j, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Using Bayes' Theorem, we know

$$\begin{aligned} P^*(t_k = i | c_k = j) &= \frac{P^*(c_k = j | t_k = i) P(t_k = i)}{\sum_{l=1}^{m} P^*(c_k = j | t_k = l) P(t_k = l)}, \\ &= \frac{\frac{1}{i} \frac{n_i}{n}}{\sum_{l=j}^{m} \frac{1}{l} \frac{n_l}{n}}, \\ &= \frac{\frac{1}{i} n_i}{\sum_{l=j}^{m} \frac{1}{l} n_l}. \end{aligned}$$

But we know $P(t_k = i | c_k = j) = \frac{n_i}{\sum_{l>k} n_l}$.

Thus, $P^*(t_k = i | c_k = j) \neq P(t_j = i | c_k = j)$.

So, using $c_i \sim U[1, t_i]$ will not accurately simulate the remaining length of stay.

Appendix D

General Surgery

D.1 SGL ANOVA Statistics

	Degrees of Freedom	Sum of Squares	Mean Squares	F	P-value	F critical
Between Groups	6	113.59	18.93	11.12	2.4E-11	2.12
Within Groups	350	596.08	1.70			
Total	356	709.66				

Table D.1: SGL Discharge volume ANOVA rejects constant volume.

	Degrees of Freedom	Sum of Squares	Mean Squares	F	P-value	F critical
Between Groups	6	61.61	10.27	10.71	2.5E-11	2.11
Within Groups	595	570.24	0.96			
Total	601	631.85				

Table D.2: SGL Ln(LoS) ANOVA rejects constant Ln(LoS).

	Degrees of Freedom	Sum of Squares	Mean Squares	F	P-value	F critical
Between Groups	6	14.37	2.39	4.71	.0001	2.13
Within Groups	271	137.66	0.51			
Total	277	152.03				

Table D.3: SGL Level ANOVA rejects constant level.

D.2 SGL Regression Statistics

Multiple R	0.6232
R^2	0.3884
Adjusted R^2	0.3833
Standard Error	0.8052
Observations	602

Table D.4: Regression Statistics for SGL Model 1.

	Coefficients	Standard Error	t stat	P-value
Intercept	0.4818	0.1378	3.50	.0001
Age	0.0078	0.0023	3.40	.0001
Sex	-0.0082	0.0688	-0.12	.9050
Level	0.6455	0.0367	17.58	5E-56
Blocks	-0.0476	0.0344	-1.38	0.167

Table D.5: Coefficient table for SGL Model 1.

Multiple R	0.6471
R^2	0.4187
Adjusted R^2	0.4089
Standard Error	0.7883
Observations	602

Table D.6: Regression Statistics for SGL Model 2.

	Coefficients	Standard Error	t stat	P-value
Intercept	0.7365	0.1673	4.40	.0001
Age	0.0074	0.0023	3.27	0.001
Sex	-0.0048	0.0674	-0.07	0.943
Level	0.6062	0.0369	16.41	.0001
Race	-0.1356	0.0742	-1.83	0.068
Monday	-0.1112	0.1547	-0.72	0.473
Tuesday	0.1156	0.1611	0.72	0.474
Wednesday	-0.4615	0.1411	-3.27	0.001
Thursday	-0.2887	0.1434	-2.01	0.044
Friday	-0.4045	0.1421	-2.85	0.005
Saturday	-0.3634	0.1531	-2.37	0.018

Table D.7: Coefficient table for SGL Model 2. Wednesday, Thursday, Friday, and Saturday have significant coefficients. Each is negative, suggesting shortening the length of stay is the mechanism by which the number of discharges is increased on the "stressed" days later in the week.

Multiple R	0.3925
R^2	0.1541
Adjusted R^2	0.1470
Standard Error	1.3092
Observations	360

Table D.8: Regression Statistics for SGL Model 3.

	Coefficients	Standard Error	t stat	P-value
Intercept	0.8685	0.1299	6.69	.0001
aLevel	0.3058	0.0769	3.97	.0001
Arrivals	0.0035	0.0493	0.07	.9433
Blocks	0.5513	0.0933	5.91	.0001

Table D.9: Coefficient table for SGL Model 3.

Appendix E

Cardiac Surgery

E.1 SCS ANOVA Statistics

	Degrees of Freedom	Sum of Squares	Mean Squares	F	P-value	F critical
Between Groups	6	159.13	26.52	14.01	2.6E-14	2.12
Within Groups	353	668.13	1.89			
Total	359	827.26				

Table E.1: SCS Volume ANOVA rejects equal volume.

	Degrees of Freedom	Sum of Squares	Mean Squares	F	P-value	F critical
Between Groups	6	14.71	2.45	3.21	0.004	2.11
Within Groups	651	496.48	0.76			
Total	657	511.19				

Table E.2: SCS Ln(LoS) ANOVA rejects equal Ln(LoS).

	Degrees of Freedom	Sum of Squares	Mean Squares	F	P-value	F critical
Between Groups	6	3.55	0.59	.79	.5742	2.11
Within Groups	271	484.61	0.74			
Total	277	488.16				

Table E.3: SCS Level ANOVA does not reject constant level.

E.2 SCS Regression Statistics

Multiple R	0.4745
R^2	0.2251
Adjusted R^2	0.2192
Standard Error	0.7794
Observations	658

Table E.4: Regression Statistics for SCS Model 1.

	Coefficients	Standard Error	t stat	P-value
Intercept	0.9938	0.1502	6.62	8E-11
Age	-0.0009	0.0177	-0.53	0.60
Sex	0.0754	0.0643	1.17	0.24
Level	0.4409	0.0358	12.31	2E-31
Race	-0.1918	0.0683	-2.81	0.005
Blocks	0.1207	0.0347	3.48	0.0005

Table E.5: Coefficient table for SCS Model 1. It appears that level, race, and the number of blocks are all significant.

Multiple R	0.4817
R^2	0.2321
Adjusted R^2	0.2202
Standard Error	0.7789
Observations	658

Table E.6: Regression Statistics for SCS Model 2.

	Coefficients	Standard Error	t stat	P-value
Intercept	0.9463	0.1815	5.21	3E-7
Age	-0.0009	0.0018	-0.49	0.621
Sex	0.0771	0.0647	1.19	0.233
Level	0.4413	0.0358	12.31	2E-31
Race	-0.1922	0.0689	-2.79	0.005
Monday	0.4578	0.1517	3.02	0.003
Tuesday	0.3798	0.1504	2.53	0.012
Wednesday	0.2952	0.1500	1.97	0.050
Thursday	0.3638	0.1551	2.35	0.019
Friday	0.1932	0.1502	1.29	0.199
Saturday	0.0755	0.1729	0.44	0.663

Table E.7: Coefficient table for SCS Model 2. Monday, Tuesday, Wednesday and Thursday each have statistically significant coefficients.

Multiple R	0.6340
R^2	0.4020
Adjusted R^2	0.3969
Standard Error	1.1806
Observations	356

Table E.8: Regression Statistics for SCS Model 3.

	Coefficients	Standard Error	t stat	P-value
Intercept	-0.0578	0.1381	-0.42	.6760
aLevel	0.6574	0.0610	10.78	1E-23
Arrivals	0.1790	0.0523	3.43	.0007
Blocks	0.2720	0.0696	3.91	.0001

Table E.9: Coefficient table for SCS Model 3.

Appendix F

IP Formulation

This appendix contains the integer program based on the data provided by UMMC. Because some of the data was fractional we needed to scale the parameters in the IP formulation. Each n_i in equations (F.2a-c) was multiplied by 10 to ensure integrality of the b_{ij}. In equation (F.3), n_{OR} was multiplied by 10 to scale with the n_i. The λ_is were multiplied by 100 in equations (F.4a-e) to ensure the integrality of the ICU_j^+ and ICU_j^-. There are an average of 1.25 ICU admissions per block of Group 3 OR time, so the scaling results in a λ_3 of 125. Likewise, λ_2 is 150 because of the scaling. In equations (F.5a-c) the Min_i and Max_i are multiplied by 10 to scale with the n_i. The other consideration is the effect of the weekend on the patient flow.

Our IP model takes the average ICU length of stay for each group to be the ICU length of stay for each patient in that group. Therefore, in our model, the effects of the weekend are not felt after Tuesday (the ICU length of stay for Group two was one day and for Group 3 was two days). Because the hospital has very few surgeries on the weekend, not many of the patients discharged from the ICU on Saturday or Sunday are replaced with new patients. Equations (F.4a) and (F.4b) represent Monday and Tuesday, respectively, and take into account these empty beds. Most of the Group 2 patients from Friday and the Group 3 patients from Thursday and Friday have been or are ready to be discharged by Monday. So, these beds have become available for the Monday block schedule. Though relatively small in number, some surgical patients arrive over the weekend. To account for the small number of weekend arrivals, we add 375 (this is from the minimum demand for group three block time, three Group 3 blocks are required each day with 1.25 patients per block on average scaled by 100 to match the λ_is) to the other arrivals and discharges in equation (F.4a). The 3.75 patients that arrived on Sunday will not be discharged until Tuesday. The -375 term in equation (F.4b) represents the patients from Sunday being discharged on Tuesday.

$$\text{Min} \sum_{j=1}^{5} ICU_j^+ \tag{F.1}$$

$$\text{s.t.} \sum_{j=1}^{5} b_{1j} = 111 \tag{F.2a}$$

$$\sum_{j=1}^{5} b_{2j} = 306 \tag{F.2b}$$

$$\sum_{j=1}^{5} b_{3j} = 332 \tag{F.2c}$$

$$b_{1j} + b_{2j} + b_{3j} \leq 160 \;\; j = 1, ..., 5 \tag{F.3}$$

$$150b_{21} + 125b_{31} + 375 - 150b_{25} - 125b_{34} - 125b_{35} + ICU_1^- - ICU_1^+ = 0 \tag{F.4a}$$

$$150b_{22} + 125b_{32} - 150b_{21} - 375 + ICU_2^- - ICU_2^+ = 0 \tag{F.4b}$$

$$150b_{23} + 125b_{33} - 150b_{22} - 125b_{31} + ICU_3^- - ICU_3^+ = 0 \tag{F.4c}$$

$$150b_{24} + 125b_{34} - 150b_{23} - 125b_{32} + ICU_4^- - ICU_4^+ = 0 \tag{F.4d}$$

$$150b_{25} + 125b_{35} - 150b_{24} - 125b_{33} + ICU_5^- - ICU_5^+ = 0 \tag{F.4e}$$

$$10 \leq b_{1j} \leq 50 \ \ j = 1, ..., 5 \tag{F.5a}$$

$$20 \leq b_{2j} \leq 130 \ \ j = 1, ..., 5 \tag{F.5b}$$

$$30 \leq b_{3j} \leq 150 \ \ j = 1, ..., 5 \tag{F.5c}$$

$$ICU_j^-, ICU_j^+, b_{ij} \geq 0 \text{ and integer } i = 1, 2, 3; \ \ j = 1, ..., 5. \tag{F.6}$$

Service Line	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Group 1	1.0	1.0	3.1	5.0	1.0	11.1
Group 2	5.8	12.0	7.1	3.7	2.0	30.6
Group 3	9.2	3.0	3.0	3.0	13.0	31.2
Total	16.0	16.0	13.2	11.7	16.0	72.9

Table F.1: The optimal solution to the integer program.

Bibliography

[1] Andersen GF, Hussey PS, Frogner BK, and Waters HR. Health spending in the United States and the rest of the industrialized world. *Health Affairs* 24(4) 2005:903-914.

[2] Schoen C, Osborn R, Huynh PT, Doty M, Zapert K, Peugh J, and Davis K. Taking the pulse of health care systems: Experiences of patients with health problems in six contries. *Health Affairs* W5 2005:509-525.

[3] Chernew ME, Hirth RA, and Cutler DM. Increased spending on health care: How much can the United States afford. *Health Affairs* 22(4) 2003: 15-25.

[4] Federal Trade Commission and the Antitrust Division of the Department of Justice. 2003.

`http://www.usdoj.gov/atr/public/health_care/204694/exec_sum.htm`

[5] Lipowski, A. and Lipowska, D. Traveling salesman problem with a center. *Physical Review E.* **71** (2005) 067701.

[6] Ahmed, E. and Elettreby, M.F. On Combinatorial Optimization Motivated by Biology. *Applied Math and Computation.* **172** (2006) 40-48.

[7] Gutin, G. Punnen, A.P. *The Traveling Salesman Problem and Its Variations.* Springer, 2002.

[8] Helsgaun, K. An Effective Implementation of the Lin-Kernighan Traveling Salesman Heuristic. *European Journal of Operational Research* **126**(1), 106-130 (2000).

[9] Grundel D.A. and Jeffcoat D.E. Formulation and solution of the target visitation problem. *Proceedings of the AIAA 1st Intelligent Systems Technical Conference,* 2004.

[10] Arulselvan A., Commander C.W., and Pardalos P.M. A Random Keys Based Genetic Algorithm for the Target Visitation Problem. *Advances in Cooperative Control & Optimization.* LNCIS 369, 389-397 (2007).

[11] Saaty T.L. Risk-Its Priority and Probability: The Analytic Hierarchy Process. *Risk Analysis* **7**(2), 159-172 (1987).

[12] Bartlett D. L., Steele J.B., *Critical Conditions*, New York: Doubleday/Random House (2004).

[13] Division of Advocacy and Health Policy. “A growing crisis in patient access to emergency surgical care.” Bulletin of the American College of Surgeons, 91, 9-18 (2006).

[14] Jacobson S., Hall S., and Swisher J. “Discrete-event simulation of health care system.” in *Patient Flow: Reducing Delay in Healthcare Delivery* (R. Hall, ed.), 211-252, New York: Springer (2006).

[15] MedModel. ProModel Corporation, 556 East Technology Avenue, Orem, UT 84097 (www.promodel.com).

[16] Schmitz H. and Kwak N. “Monte Carlo simulation of operating-room and recovery-room usage.” Operations Research, 20(6), 1171-1180 (1972).

[17] Vasilakis C. and El-Darzi E. “A simulation study of the winter bed crisis.” Health Care Management Science, 4(1), 31-36 (2001).

[18] Farmer RDT and Emami J. Models for forecasting hospital bed requirements in the acute sector. *Journal of Epidemiology & Community Health.* 44: 307-312. 1990.

[19] Postma MJ, Ruwaard D, Jager HJC, and Dekkers ALM. Projecting utilization of hospital in-patient days in The Netherlands: A time-series analysis. *IMA Journal of Mathematics Applied in Medicine & Biology.* 12:185-202. 1995.

[20] Zernikow B, Holtmannspotter K, Michel E, Hornschuh, Groote K, and Hennecke KH. Predicting length-of-stay in preterm neonates. *European Journal Pediatrics.* 158:59-62. 1999.

[21] Mobley BA, Leasure R, and Davidson L. Artificial neural network predictions of lengths of stay on a post-coronary care unit. *Heart and Lung.* 24(3):251-256. 1995.

[22] Witten IH and Frank E. *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, San Francisco, 2005.

[23] Singer DE, Carr PL, Mulley AG, and Thibault GE. Rationing intensive care – physician responses to a resource shortage. *New England Journal of Medicine.* 309(19): 1155-1160. November 1983.

[24] Strauss MJ, LoGerfo JP, Yeltatzie JA, Temkin N, and Hudson LD. Rationing of intensive care unit services: an everyday occurrence. *Journal of the American Medical Association.* 255(9): 1143-6. March 1986.

[25] McManus M, Long M, Cooper A, Mandell J, Berwick D, Pagano M, and Litvak E. Variability in Surgical Caseload and Access to Intensive care services. *Anesthesiology.* 98(6): 1491-6. 2003.

[26] Mitchell, Jean. Do financial incentives linked to ownership of specialty hospitals affect physicians' practice patterns? *Medical Care.* 46(7):732-737. July 2008.

[27] Levin DC and Rao VM. Turf wars in radiology: the overutilization of imaging resulting from self-referral. *Journal of the American College of Radiology.* 1(3): 169-172. 2004.

[28] Davis GE, Lowell WE, and Davis GL. A neural network that predicts psychiatric length of stay. *Clinical Computing.* 10:87-92. 1993.

[29] Price C, Babineau T, Golden B, Harrington M, and Wasil E. Capacity management in a cardiac surgery line, presented at INFORMS Annual Meeting. November 2007.

[30] Zollinger TW, Saywell RM, Smith CP, Highland D, Pfeiffer D, and Kelton GM. 1999. Delays in Patient Transfer: Postanesthesia Care Nursing. *Nursing Economics,* **17(5)**, 283-90.

[31] Blake J, Carter M. 2002. A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research,* **140**, 541-561.

[32] Blake J, Donald J. 2002. Mount Sinai hospital uses integer programming to allocate operating room time. *Interfaces,* **32(2)**, 63-73.

[33] Sier D, Tobin P, McGurk C. 1997 Scheduling surgical procedures. *Journal of the Operational Research Society,* **48(9)**, 884-891.

[34] Jebali A, Alouane A, Ladet P. 2006. Operating room scheduling. *International Journal of Production Economics.* **99**, 55-62.

[35] Dexter F, Macario A, Traub R, Hopwood M, Lubarsky D. 1999. An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesthesia and Analgesia.* **89(1)**, 7-20.

[36] Belien J, Demeulemeester E. 2007. Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research.* **176**, 1185-1204.

[37] Santibanez P, Begen M, and Atkins D. 2007. Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority. *Health Care Management Science.* **10**, 269-282.

[38] ILOG, Inc. 1195 West Fremont Ave, Sunnyvale, CA 94087-3832 (www.ilog.com).

[39] Achterberg, Tobias. 2007. *Constraint Integer Programming*, Ph.D. thesis. Technische Universitat Berlin, Berlin.

[40] The MathWorks, Inc. 3 Apple Hill Drive, Natick, MA 01760-2098 (www.mathworks.com).