Applications of Stochastic and Queueing Models to Operational Decision Making

by

PAUL ENDERS

Submitted to the Tepper School of Business

in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

CARNEGIE MELLON UNIVERSITY

26 April 2010

Dissertation Committee:
Professor Alan Scheller-Wolf (Chair)
Professor Geert-Jan van Houtum
Professor Nicola Secomandi
Professor Soo-Haeng Cho

Abstract

An operations manager makes operational decisions in the face of a, by definition, uncertain future. In this thesis we develop tools that can improve the quality of operational decision making by modeling the stochastic environment and analyzing the trade-offs that the operations manager faces within this environment. We examine three specific settings:

The question of how to best leverage technology is fundamental to almost any industry. Using real data from EQT Corp. (an integrated natural resources company operating natural gas wells throughout the Appalachian basin) we analyze the interaction between the real options to scale different technologies and the real option to scale the extraction rate. We find that the values of these options are highly interdependent and their optimal use is rather complex. We bring to light data-driven managerial principles guiding the use of these options and provide a very effective heuristic control policy.

Prioritizing demand streams is common in inventory management. In many settings (e.g. a central warehouse), some demands can be backordered while others are lost when not immediately satisfied. A critical level (CL) policy reserves some inventory for future high-priority demand by backordering current, lower-priority, demands. We develop an efficient algorithm to find the optimal CL policy in this setting, and compare the performance to the globally optimal policy. We find that although the CL policy performs (slightly) worse, it is almost insensitive to variations in the lead time distribution.

Emergency Department (ED) demand for care is by its very nature hard to predict accurately. As ED capacity is regularly outstripped by demand, EDs attempt to decrease the inflow of patients during such periods of “crowding.” We use real data to model the Pittsburgh (PA) Emergency Medial Services (EMS) system and evaluate the impact of several coordination mechanisms between ambulances and/or hospitals on the timeliness of care and total hospital revenues. We find that coordination mechanisms in which hospitals share certain indicators with EMS crews can significantly outperform the coordination mechanisms currently used in practice in term of quality of care, without being detrimental to hospital revenues.

to Silvia,

my parents Co & Trees,

and my brother Remco

Acknowledgements

During my research endeavors I have been fortunate to work with many talented and inspiring individuals. For almost 10 years now I have had many interesting and engaging conversations with Geert-Jan van Houtum. His enthusiasm for research was what first got me interested in pursuing a Ph.D. and he has been supportive throughout. Geert-Jan was also responsible for getting me in contact with Alan Scheller-Wolf back in June 2004. At that point I could not have foreseen how these two people would form my career and academic development. Alan was always there for me throughout my time at Tepper. Not only was he always interested in research problems that we were attacking, he was also genuinely interested in my personal development and well-being. I would like to thank Alan & Geert-Jan for all the time they invested to help me develop into the person I am now.

This thesis also greatly benefitted from discussions with many faculty members at Carnegie Mellon University. First and foremost I would like to thank Nicola Secomandi and Soo-Haeng Cho for serving on my committee and making many valuable suggestions. Ivo Adan has been an inspiring co-author on one of my chapters. Furthermore I would like to thank the CMU faculty for creating an inspirational and challenging environment. Specifically I would like to thank Egon Balas, Bahar Biller, Mor Harchol-Balter, Robert Hampshire, Sunder Kekre, Javier Pena, François Margot, Fallaw Sowell, Willem-Jan van Hoeve, and Sridhar Tayur.

My time at Tepper has greatly benefited from the support of the administrative staff. Above all there is Lawrence Rapp. Lawrence has made life at Tepper flow smoothly and was always there for me with a smile and a joke to crack. His dedication to the Ph.D. students is remarkable and greatly appreciated. Every time I was involved in teaching undergraduate or MBA students I greatly benefited from the help of Rosanne Christy and Jess Schaeffer. Rosanne and Jess have helped me navigate the trenches of teaching and administrative procedures. I also want to thank Wendy Hermann for her trust in my teaching abilities and everybody in the MBA student services office for helping me deal with the administrative burden that comes with teaching. Although my first encounter with the Tepper Career Opportunities Center has only been within the last year, I do want to thank them for their help in my job hunt, I owe specific thanks to Leslie Kromer.

For each of the chapters in this thesis I have benefited from inspiring conversations with practitioners. At EQT Corp. I want to thank Stuart Olson, Fil Sciullo, and Tim Dugan for their helpful suggestions on the second chapter of this thesis. Chapter 3 was motivated by a problem faced by ASML and special thanks go out to Harold Bol, Harrie de Haas, and Eric Messelaar. The key research questions for the fourth chapter were suggested to us by Shane Henderson at Cornell University. I would like to thank Shane for suggesting such an interesting research topic. This chapter could not have come this far without the guidance into the EMS world provided by Daniel Patterson, Frank Guyette, and Dave Hostler.

Special thanks goes out to my friends that have supported me throughout the years. This includes those that have known and supported me for almost 15 years: Marieke, Paul, Sander, Marijn, Richard, Judith, and all those others in Oosterhout and surroundings. Throughout college I have made friends who were at first surprised about my decision to pursue a Ph.D. but have supported me throughout: Bob, Francine, Jasper, Jérôme, Koen, Mark, Niels, Onno, Robbert, and many others. In Pittsburgh my friendships have developed from the day of our arrival on August 3rd 2005 when Masha picked us up from the airport. Masha, without your friendship and support I could have never made it through the last 5 years. You were there for me when I was stuck and knew how to motivate me even in the middle of the night during the last days of frantic thesis writing. Special thanks also needs to go out to our "lunch group": Masha, Sam, Ben, John, Judy, Federico, Peter, and Zed. Whether we were eating inside or out, our lunches always provided a welcome distraction from cubicle thoughts! Finally I would like to thank all of you that tolerated my ranting, hung out with me at conferences, crawled Pittsburgh bars, shivered at football games, trained for my first (and last?) marathon, you know who you are[1]!

I am thankful to my parents and my brother as they have always been there for me. Co & Trees, you taught me how to live my life and make the important decisions. Your guiding principles have helped me get to this point and will sure help me in the years to come. Even though the last years of us living far away were emotionally hard on you, your support never wavered, thanks!.

Finally I would like to thank my wife Silvia. Although somewhat skeptical at first about moving to the USA to pursue our Ph.D.s she jumped into the adventure wholeheartedly. Without your support, motivation, drive, and love, I would have not even made it through the first months at Tepper, let alone the last. You were always there for me, and I promise I will always be there for you.

[1] If not (in random order): Melanie, Loes, Nagesh, Liese-Lotte, Leo, Kate, WJ , Alina, Marieke, David, Vince, Ilse, Jesse, Yvonne, Ben, Vish, Rich, Marloes, Ans, Eefje, Marchien, Rob, Tava, Laurie, Joanna, Nathalie, Hans, Jason, Jesper, Joris, Kate, Aafke, Lauren, Sasha, Allison, Ingeborg, Joost, Adam, Jorn, Grace, Jasper, Amanda, Fatos, Shruti, Chris, Ryan, Margaret, Marianne, Babette, Jeannine, Boris , Holly, Varun, Lizet, JZ, Bram, Donna, Diana, Judy, Henry, Andrew, Jenn, Ronald, Chris, John, Jess, Carolyn, Danya, Candace, Brandon, Atul, Erinn, Erik, Erkut, Ilya, Dr. C., Irene, Ingrid, Lisa, Wemke, Jeanne, Billie, . . .

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Operational decisions are made on a day-to-day basis (Simchi-Levi et al. 2008) and include a wide range of choices made in the businessLength of stay data (days) by Acuity world. These decisions typically involve expectations about the future; hence they involve uncertainty about this future. Stochastic models are used to inform a present-time decision maker about how the uncertain future may affect the outcome of his decision. In this thesis we explore stochastic models, including queueing models, and analyze how these models can be used to increase the quality of decisions made. Specifically, we consider decisions regarding maintenance of assets, providing service to customers, or where to optimally route ambulance patients. This thesis analyzes the different trade-offs seen by an operations manager and finds optimal strategies for the corresponding problems.

Throughout this thesis we will analyze real-world problems which all require the operations manager to deal with an uncertain future. In Chapter 2, we analyze the optimal technology deployment decision for a natural gas property. In this setting the uncertainty we consider is the volatility of the natural gas price, which is complicated by the seasonality of the cost structure. In Chapter 3, we consider the problem of an inventory manager who wants to differentiate between classes of customers. We analyze an inventory management policy that helps the inventory manager minimize his cost in the face of uncertainty in demand. In Chapter 4, we consider the problem of the coordination of Emergency Medical Services (EMS). Not only is the demand for EMS support random, one also deals with random treatment times once a patient has been transported to a hospital. In Chapter 5, we conclude. In the remainder of this introduction we introduce the subsequent chapters in more detail and illustrate what binds them together.

Interaction between Scaling Options in Natural Gas Production

This first study is the outcome of a research project studying questions of technology utilization and production management with managers at EQT Corp., an integrated natural resources company that operates natural gas wells throughout the Appalachian basin. One of the primary operational problems encountered when drilling for natural gas is that liquids gather in the well bore. Once too much liquid has accumulated, the well stops producing until the liquid is removed, which can be done in a variety of ways. We use a real options framework to model two different types of technologies that can be used, either separately or together, to *(i)* monitor a well, in order to react to gathering liquids quickly, and *(ii)* deal with the accumulated liquids.

In addition, the operations manager has the option to scale the natural gas extraction rate, by pausing production. This could be advantageous because natural gas is a commodity sold into the volatile spot market. In the face of this volatility the operations manager needs to form expectations about the future, in order to make decisions about technology deployment and extraction scaling in the present. We develop a model that provides data-driven managerial principles regarding the trade-offs that need to be made by the operations manager when making technology choices.

We use a stochastic dynamic program to bring these principles to light. Since the optimal decision structure is rather complex, we also study a simplification of the optimal deployment and extraction policy and show that this simplified policy can still obtain almost 100% of the optimal value. Finally, we determine when the technology and extraction scaling options that we analyze are complements or substitutes. The principles that we develop can not only help an operations manager who is drilling for natural gas, but also can be helpful to managers of other natural resource production processes, e.g. the extraction of oil and mining.

Inventory Rationing for a System with Heterogeneous Customer Classes

In many business settings a product is sold to different classes of customers, who may have different values to the business. In this setting one can imagine that customers might receive different levels of service, commensurate with their respective values. We consider a case in which the demand that the operations manager has to satisfy originates from two classes of customers, that are primarily distinguished by their willingness to wait for the product. In such a setting the operations manager has to deal with several sources of uncertainty: Not only does he face uncertainty regarding the delivery lead times of his own outstanding orders, he also faces uncertain customer demands.

Uncertainty in delivery lead times is typically dealt with using inventory. We propose that the operations manager imposes a specific type of inventory allocation policy, a *critical level policy*, to deal with the uncertainty in customer demands. A critical level policy reserves inventory for potential future demands from more important customers by backordering the less important demand as soon as inventory is at or below a certain critical level.

We model this problem using a continuous review base stock replenishment policy in which demand and lead times are stochastic. Modeling the system as a Markov chain, we develop an exact and efficient procedure to determine the performance of any given parameter pair for the base stock and critical level policies. We then leverage this procedure to find the parameters of the optimal policy. Next we compare the performance of our critical level policy with two more naïve allocation policies as well as the globally optimal policy. From this comparison we learn that the critical level policy performs rather well. Furthermore, we find that although the globally optimal policy is rather sensitive to the degree of replenishment lead time variability, the critical level policy appears to be fairly robust. Hence the critical level policy appears to be an effective tool for an operations manager dealing with these sources of uncertainty while wanting to differentiate between different classes of customers.

Ambulance Traffic Coordination

Demand for emergency medical services is almost by definition uncertain: As a patient you typically do not see an emergency coming. However, you still rely on the health care system to effectively and efficiently take care of you once you need their services. We consider the emergency services consisting of both the ambulance service that may provide transportation to the hospital, as well as the emergency care provided in the Emergency Department (ED) at the hospital. The ED typically is part of a larger hospital setting and relies on the Inpatient Department (ID) of the hospital to provide care to patients once they have received their initial treatment at the ED. As patients may also visit the ID for elective procedures, the emergency services is one part of a complex system. In addition, supply and demand of emergency services are often unbalanced, the operations manager must manage the supply of emergency services such that effective and efficient care can be provided.

A common way trying to match demand with supply within emergency service systems is by redirecting incoming ambulance traffic away from a hospital that is already short in resources. This is typically done using a *diversion policy* which requests that ambulances take their patients elsewhere. We develop a simulation model to analyze this and various other mechanisms that the operations manager could use to disseminate information about the speed with which care can be provided to incoming patients. The mechanisms we evaluate range from the case in which no information is provided by the hospital and EMS crews update the information about the status of a hospital on their own, to the case in which hospitals communicate detailed indicators, capturing the current patient load, as well as metrics that aim to take future developments into account. Our simulation model is able to capture many real-world complexities. We calibrate our model to the situation that exists in Pittsburgh (PA): Here 7 EDs are located within a small geographical area.

Due to this proximity, there is the potential for routing ambulance traffic to the hospital that is best able to provide care to a patient. We show that several of the commonly used policies to avoid or reduce crowding in EDs perform reasonably well. However, in reality these policies cannot be implemented in isolation as EMS crews typically use their own experience to take patients to a hospital where they believe expedient care can be provided. We illustrate that the use of outdated information (or experience) by EMS crews can significantly decrease the quality of care. We introduce two coordination mechanisms that significantly outperform the other mechanisms by spreading demand for emergency care more evenly, without decreasing hospital revenues.

In these three chapters we deal with situations in which an operations manager has to cope with uncertainty while having to provide an adequate level of service. Whether the operations manager is operating gas wells, managing inventory, or coordinating ambulance traffic, taking this uncertainty into account is crucial to him making effective decisions.

Chapter 2

Interaction between Scaling Options in Natural Gas Production[1]

2.1. Introduction

One of the fundamental questions in almost any industry is how to best leverage the use of technology; companies are typically faced with an assortment of technology options and must decide how to dynamically manage them to maximize the value of their assets. This is particularly true in the extraction industry: For example production managers at an oil or natural gas production company are constantly challenged with evaluating technology adoption choices to deal with liquid load-up, a phenomenon that retards extraction from a well.

Two types of technologies that help deal with liquid load-up in natural gas wells are *(i)* flow *enhancement technology* (ET) that increases the instantaneous production level as well as the total extractable reserves, and *(ii) communication technology* (CT) that increases

[1] This chapter has been published as Enders et al. (2010) .

operational efficiency, boosting the instantaneous production level, but also increasing the rate at which a reservoir is depleted. ET examples include *continuous flow*, whereby one simply uses the gas flow to bring liquids to the surface, and *soap and blow*, which comprises periodically manually dropping a soapstick into the well to transform the liquids into foam, which is more easily carried to the surface by the gas flow. CT examples include *manually* collecting production (flow rate) data and monitoring a well's status (operating or not) on a periodic basis by a human operator, and *automated continuous* data collection and monitoring of the well through electronic data transmission.

These examples illustrate that ET and CT can be deployed at different "levels" that can be modified over time. These modifications may be contingent on the natural gas flow rate and the prevailing natural gas price, which fluctuates over time, as well as the technology adjustment costs, which may be seasonal. Thus, one can interpret the choices of ET and CT levels as real options to scale the production level (Trigeorgis 1999, p. 2). A third real option that is available to natural gas production managers, and more generally commodity production managers, is the option to scale the extraction rate by *pausing* production. This option consists of temporarily ceasing production from a well, often in anticipation of an increase in price.

While versions of these three real options have been considered in *separate* streams of literature, the study of their interaction has received very little attention. We explore the nature of this interaction in cooperation with production managers at EQT Corp., an integrated energy company with emphasis on natural gas supply, transmission, and distribution in the Appalachian area (`www.eqt.com`). To do so we develop and analyze a stochastic dynamic programming model, with uncertain natural gas price evolution modeled as in Jaillet et al. (2004). Our work significantly extends the existing models of well behavior (e.g., Tarek 2006) to capture how modifying technology levels affects the flow rate evolution. Our work also stands out for its unique use of a novel data set that includes over 20 years of production data from a natural gas well that is part of the Eastern Kentucky production portfolio of EQT Corp. Our model integrates this production data with financial data describing the well's cost structure and natural gas price data available from the New York Mercantile Exchange (NYMEX), in order to establish the marginal and joint values of the three scaling options. Our model also includes the option to abandon a well, but we do not focus on its valuation, as our main interest is on how these three scaling options interact with each other.

We find that all options, especially the ET scaling option, have substantial value by themselves. That the application of technology to natural gas extraction could add significant value is to be expected. However, our finding that pausing could have significant value, under certain conditions, was surprising to the EQT Corp. managers who participated in this study. More broadly, pausing is often ignored in practice (Slade 2001, Dugan 2006, Sciullo 2006, Al-Harthy 2007), because the conventional wisdom among managers is often to maximize the production rate, under the premise that the value of any gas not produced during a given time period is effectively lost due to financial discounting.

When used together, the ET and CT scaling options behave as complements, in the sense that each option is worth more when used in combination with the other option than when used alone. This is also true of the combination of CT scaling and pausing options. In contrast, the ET scaling and pausing options behave as substitutes. Moreover, the ET scaling option and the portfolio that includes the CT scaling and pausing options are complements, and so are the CT scaling option and the portfolio that includes the ET scaling and pausing options; this means that each technology scaling option is worth more when used together with the remaining technology and extraction scaling options rather than alone. In contrast, the pausing option and the portfolio that includes the ET and CT scaling options are substitutes. Finally, the option value of the portfolio that includes all the technology and extraction scaling options exceeds the sum of the individual values of these options. The insights behind these results are not immediate; we provide detailed supportive explanations.

We also find that stochastic variability in the natural gas price increases the value of all the scaling options, but in unique ways. The pausing option derives almost its entire value from this stochastic variability, which is surprising given the marked seasonality (deterministic variability) in the natural gas forward curve. In contrast, stochastic variability only slightly increases the values of the technology scaling options, which seems expected.

Finally, we show that the optimal scaling option deployment policy is rather complex. We thus seek a simpler yet effective approach to managing the portfolio of scaling options. We find that immediately setting the two technology scaling options at their highest levels and afterwards managing the pausing option in conjunction with the abandonment option yields essentially the same value as the complex optimal policy. This approach is useful because it drastically simplifies the deployment of the portfolio of scaling options with almost no decrease in value.

Our results contribute to the practice-based literature in operations management by providing data-driven managerial *principles* (Fisher 2007, p. 374) pertaining to the value and management of technology and extraction scaling options in natural gas production. These principles are specifically significant to natural gas production managers and, potentially, to managers of other natural resource production processes, such as the extraction of oil and mining, two activities that encompass similar volatile environments and production processes.

We proceed as follows. We review the relevant literature in Section 2.2, introduce our model in Section 2.3, and discuss the data used in our study in Section 2.4. We present our results on the valuation and deployment of the technology and extraction scaling options in Section 2.5 and Section 2.6. We conclude in Section 2.7 where we also discuss limitations of our work and potential additional research avenues.

2.2. Literature Review

Early on, Grayson (1960) recognized the need for a quantitative approach to support drilling decisions by oil and gas operators, to step away from rules of thumb and pure expert judgment that were commonly used in practice. Durrer and Slater (1977) provide a thorough review of operations research methods used in petroleum and natural gas production up to the late 1970's. Most of the more recent literature on the optimal operation and valuation of commodity projects, including oil and gas ventures, uses the real option approach to value and optimize managerial flexibility (Dixit and Pindyck 1994, Smith and McCardle 1999, Trigeorgis 1999).

Brennan and Schwartz (1985) consider the decision to open, close or abandon a mine using replicating portfolios. Olson and Stensland (1988) analyze the optimal shutdown decision when the extraction cost is fixed, as in Clarke and Reed (1990), who consider oil well valuation and abandonment decisions using stopping rules. Smith and McCardle (1998) use the integrated option pricing and decision analytic approach of Smith and Nau (1995) to value oil properties. They consider the decision to open, close or abandon a property, and provide thresholds for each decision in terms of production rates and prices. Lund (2000) analyzes the value of flexibility in offshore petroleum projects. Kamrad and Ernst (2001) study the optimal production policy with yield and price uncertainty for a mining concern when the resource to be exploited is nonhomogeneous. Cortazar et al. (2001) evaluate the exploration, development, and extraction phases of a copper mine with investment timing flexibility, and open, close or abandonment decisions during the extraction phase. Lumley and Zervos (2001) consider the effect of switching costs in pausing and restarting a mining project. Similar to most of these papers, we study the option to pause production, i.e., scale the extraction rate. Different from these papers, we also investigate the interaction between this real option and the real options to scale the ET and CT levels in natural gas production.

ET and CT are discussed in detail in the natural gas and petroleum engineering literature, for example, by Coleman et al. (1991a – 1991c), Clegg et al. (1993), and Coşkuner and Strocen (2003). However, these authors do not quantify the economic value of optimally managing these technologies as real options, as we do in this chapter. Moreover, we extend the commonly used models of natural gas production (e.g., Tarek 2006) to incorporate the effects of scaling the ET and CT levels on the flow rate evolution, consistent with the empirical observations available in the petroleum engineering literature (Coşkuner and Strocen 2003).

The CT scaling option that we investigate is akin to the optimization of preventive maintenance decisions studied by Kamrad and Lele (1998) in a manufacturing environment with random yield and output prices. Unlike their model, in our model the scaling of CT is reactive and impacts the evolution of the extraction rate, not only the instantaneous production rate. Moreover, different from Kamrad and Lele (1998), we investigate the interaction of the CT scaling option with the ET scaling and production pausing options.

Our work is also related to the paper by Moel and Tufano (2002) who study opening and closing decisions of gold mines' managers as a function of industry and firm level characteristics. But their focus is descriptive, and their aim is to empirically test whether gold mines' managers opening and closing decisions are consistent with the basic predictions laid out in Brennan and Schwartz (1985). Our focus is prescriptive, and our aim is to bring to light model-based and data-driven principles that could be used by managers to inform their decisions on how to manage technology and extraction scaling options in the production of natural gas and other commodities. Moreover, Moel and Tufano (2002) do not include technology scaling decisions in their analysis. This is an important distinction as these scaling options significantly increase the value of the property and managing them optimally is not straightforward.

2.3. Model

In this section we present our stochastic dynamic programming model that optimizes technology and extraction scaling decisions during a finite time horizon. Let T denote the end of this time horizon, which is assumed to be divided into a set of equal length time periods (see Figure 2.1). We normalize this length to one. Even though we assume that the well's flow rate and the natural gas price evolve in continuous time, in our stochastic dynamic program decisions are made only at times $0, 1, \ldots, \mathsf{T}-1$, the start of time periods $1, \ldots, \mathsf{T}$. (That is, the stages of this model correspond to the times $0, 1, \ldots, \mathsf{T}$.) The state of our stochastic dynamic program at time t is the vector (p, q, x, y), which includes the natural gas spot price, $p \in \mathbb{R}_+$; the well's flow rate, $q \in \mathbb{R}_+$; and the ET and CT levels, x and y. We assume that there are $X+1$ and $Y+1$ possible ET and CT levels, which belong to sets $\mathcal{X} := \{0, \ldots, X\}$ and $\mathcal{Y} := \{0, \ldots, Y\}$, respectively. Thus, the state space in any stage is the set $\mathcal{S} := \{(p, q, x, y) : p \in \mathbb{R}_+, q \in \mathbb{R}_+, x \in \mathcal{X}, y \in \mathcal{Y}\}$.

At time T, if still operational, the well is abandoned (which could be required by contractual agreements). At each time $t < \mathsf{T}$, the production manager decides how to operate the well in time period $t+1$. Specifically, this manager decides whether to

1. *Produce* using the currently deployed technology.
2. *Abandon* the well and forego the opportunity of ever producing from it again. (When abandoning, all equipment at the well is removed.)
3. *Pause* production and leave all the equipment as is, to potentially resume production or abandon the well at a future time.
4. *Invest* or *divest* in ET and/or CT, and either produce or pause during time period $t+1$.

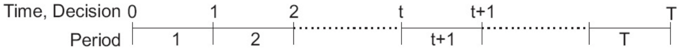


Figure 2.1: Timeline

We denote by $a := (a_1, a_2, a_3)$ the action vector in a given state and stage. This vector specifies the new ET level, a_1, the new CT level, a_2, and whether to produce, $a_3 = 1$, pause, $a_3 = 0$, or abandon, $a_3 = -1$. We denote the action set at time $t < \mathsf{T}$ by $\mathcal{A}(t) := \{(a_1, a_2, a_3) : a_1 \in \mathcal{X}, a_2 \in \mathcal{Y}, a_3 \in \{-1, 0, 1\}\}$. Decisions on the technology levels made at time $t < \mathsf{T}$ are immediately implemented, so the system in period $t + 1$ operates at technology levels a_1 and a_2. This is reasonable because of the limited magnitude of the changes that need to be made to the well.

The cash flows in stage t are determined by the gas price, p_t, and the flow rate, q_t, at time t, taking into account the instantaneous implementation of the decision a at time t and any cost associated with this decision. Since many gas (and oil) wells are located in remote areas, weather conditions can significantly impact their operational and investment costs. Therefore, the relevant costs in our model are time dependent. We consider:

- *Preventive maintenance cost* c_t^M that is associated with basic seasonal maintenance jobs. This cost is paid in every stage up to abandonment.

- *Operating cost* $c_t^O(x, y)$ that depends on the currently employed technology levels x and y. When a well is paused or abandoned, no operating cost in incurred.

- *Investment cost* $c_t^I(x, y, a_1, a_2)$, to change the technology levels from (x, y) to (a_1, a_2). We assume the investment cost is separable in technology type, i.e., $c_t^I(x, y, a_1, a_2) = c_t^{IE}(x, a_1) + c_t^{IC}(y, a_2)$, where $c_t^{IE}(x, a_1)$ and $c_t^{IC}(y, a_2)$ are the costs of changing the ET level and CT level, respectively. The investment cost is positive when the technology level is increased and *may* be negative if the technology level is decreased, due to the salvage value of the equipment.

- *Abandonment cost* c_t^A that may be positive or negative depending on whether the well is sold or needs to be dismantled at no salvage value. When a well is abandoned, the installed ET and CT need to be dismantled at costs $c_t^{IE}(x, 0)$ and $c_t^{IC}(y, 0)$, respectively.

The stochastic component of our model is the price process. We assume that the natural gas spot price evolves according to an exogenously specified stochastic process that is not affected by the flow rate evolution and the manager's operational decisions. This is consistent with other models used in the literature (e.g., Smith and McCardle 1998). In our numerical analysis in Sections 2.5 and 2.6 we use the seasonal mean reverting price process of Jaillet et al. (2004). Schwartz (1997) and Seppi (2003) provide excellent reviews of the typical reduced-form models of the evolution of commodity prices that are employed in the real option literature.

We model the well's flow rate as a deterministic process. Although other authors, such as Smith and McCardle (1998), have used stochastic models of flow rate evolution, our choice is motivated by modeling simplicity, i.e., it allows us to capture the effect of technology scaling on the flow rate evolution in a parsimonious fashion, as discussed below. In addition the deterministic model corresponds well with models used in practice, as discussed below. We leave the extension to a stochastic case of our model to future research, as pointed out in Section 2.7.

The flow rate at time $t+1$, q_{t+1}, with $t < \mathsf{T}$, depends on the time t flow rate, q_t, and the action vector, $\mathcal{A}(t)$. If the well is abandoned then q_{t+1} is set equal to 0. Otherwise, if ET and CT are unmodified, then q_{t+1} is equal to q_t if the well is paused, and changes to $q_t \exp[-\mu(x,y)]$ if the well is producing, where $\mu(x,y) > 0$ is the technology dependent decline rate of the well's flow rate. When the ET and CT technology levels are unchanged, this model of exponential flow rate decline is commonly used in the petroleum engineering literature (see, e.g., Tarek 2006); in particular, the positive decline rate is consistent with the behavior of mature wells, as we model here. If the well is not abandoned, modifying either one of the current ET or CT levels affects both q_t and $\mu(x,y)$. We capture these effects by embedding petroleum engineering models of natural gas well liquid load-up behavior (Clegg et al. 1993 and Coleman et al. 1991a-1991d) into the exponential flow rate decline model.

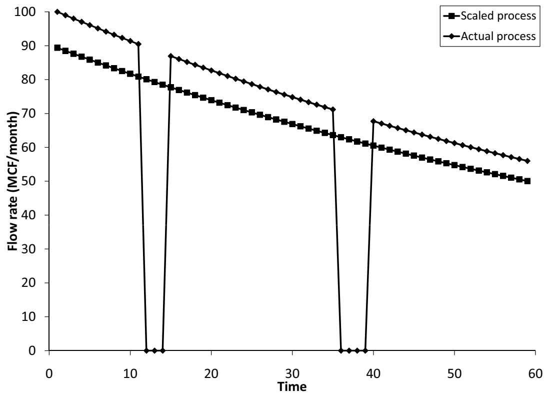


Figure 2.2: Sketch of the actual and scaled flow rate processes.

We introduce the notion of nominal flow rate, denoted by $\bar{q}_t$. This is the flow rate at an idealized CT level that would ensure a production yield equal to 100%. For ET level x, the nominal flow rate evolves according to the exponential model with decline rate $\bar{\mu}(x)$ (we discuss how we model this rate below). We relate the flow rate q_t to the nominal flow rate $\bar{q}_t$ by using the production yield $\rho(y) \in (0, 1)$, which depends on the employed CT level y. Specifically, if the time t nominal flow rate is $\bar{q}_t$ and the ET and CT levels are x and y, then the flow rate and its decline rate satisfy the identities $q_t \equiv \rho(y)\bar{q}_t$ and $\mu(x, y) \equiv \rho(y)\bar{\mu}(x)$. This means that our representation of the flow rate is a smoothed version of an underlying flow rate model with load-up events (see Figure 2.2).

We model the production yield $\rho(\cdot)$ as follows. CT affects how quickly one can respond to a situation in which a well is loaded up with liquid. Wells that are operational are assumed to load-up according to a geometrically distributed time to failure with mean $1/\lambda$, where $\lambda > 0$. Loaded-up wells are assumed to return to operational status through corrective maintenance, i.e., recover, with a geometrically distributed time with mean $1/\eta(y)$, where $\eta(y) > 0$ depends on the employed CT level y. We define the production yield at CT level y as the fraction of time that a well is operational: $\rho(y) := 1/[1 + \lambda/\eta(y)]$. We assume that this quantity increases in the CT level y by assuming that $\eta(y)$ does so, which is natural, i.e., the mean recovery time decreases in the CT level y.

We now discuss the effect of changing the ET and CT levels from (x, y) to (a_1, a_2) at time t, when the flow rate is q_t. Using the identities that relate the flow rate to the nominal flow rate and their decline rates, the CT change from level y to level a_2 has two effects: *(i)* it modifies the flow rate from q_t to $\rho(a_2)q_t/\rho(y)$, as $q_t/\rho(y)$ is the nominal flow rate corresponding to q_t; *(ii)* it modifies the decline rate from $\mu(x, y) \equiv \rho(y)\bar{\mu}(x)$ to $\rho(a_2)\bar{\mu}(x)$. Notice that if $a_2 > y$, then this CT change increases the rate of decline of the flow rate; this implies that the new flow rate also increases.

The ET change from level x to level a_1 affects both the flow rate and its decline rate. Notice that if $a_1 > x$, then this change instantaneously raises the flow rate and reduces its decline rate; this is consistent with the empirical observations made by Coşkuner and Strocen (2003), which are reflected in our data set, as discussed in Section 2.4. We use the given function $\delta(x, a_1)$, with $\delta(x, a_1) > 1$ for $x < a_1$, to model the effect on the flow rate q_t of changing the ET level from x to a_1 as follows: q_t immediately changes to $\delta(x, a_1)q_t$. To model the effect on the decline rate of changing the ET level from x to a_1, we first must model the decline rate of the nominal flow rate $\bar{\mu}(0)$. We use the function $\gamma(x)$ for this purpose: We assume that $\gamma(0) = 1$ and that $\gamma(\cdot)$ is an increasing function. We then define the decline rate of the nominal flow rate at ET level x, relative to its value at level 0, as $\bar{\mu}(x) := \gamma(x)\bar{\mu}(0)$. Changing the ET level from x to a_1 changes the decline rate of the nominal flow rate from $\gamma(x)\bar{\mu}(0)$ to $\gamma(a_1)\bar{\mu}(0)$. Thus, this change modifies the decline rate of the flow rate from $\rho(y)\gamma(x)\bar{\mu}(0)$ to $\rho(y)\gamma(a_1)\bar{\mu}(0)$.

We aggregate the effects of changing technologies through the flow rate transition function $f_t(q_t, x, y, a)$; this function yields the flow rate at time $t + 1$: $q_{t+1} = f_t(q_t, x, y, a)$. We define this function as follows:

$$f_t(q_t, x, y, a) := \begin{cases} 0 & \text{if } a_3 = -1, \\ \delta(x, a_1)\frac{\rho(a_2)}{\rho(y)}q_t & \text{if } a_3 = 0, \\ \delta(x, a_1)\frac{\rho(a_2)}{\rho(y)}\exp[-\rho(a_2)\gamma(a_1)\bar{\mu}(0)]q_t & \text{if } a_3 = 1. \end{cases} \tag{2.1}$$

If the well is abandoned the flow rate drops to 0; otherwise, the flow rate immediately changes to $\delta(x, a_1)\rho(a_2)q_t/\rho(y)$ and stays at this level if the well is paused, and declines exponentially from this level with decline rate $\rho(a_2)\gamma(a_1)\bar{\mu}(0)$ if the well produces during the next period.

We now formulate our stochastic dynamic program. We assume that there exists a natural gas futures market, e.g., NYMEX, that is arbitrage free and complete. Thus, we employ a risk-neutral valuation approach (Luenberger 1998, Smith 2005). We use the one period risk free discount factor α. We let $\tilde{p}_t$ be the random variable denoting the natural gas price at time t. We denote by $1\{\cdot\}$ the indicator function that is 1 if its argument is true and 0 otherwise. We let $r_t(p, q, x, y, a)$ denote the following immediate payoff function associated with action $a \in \mathcal{A}(t)$ in state $(p, q, x, y) \in \mathcal{S}$ at time $t = 0, \ldots, \mathsf{T} - 1$:

$$r_t(p,q,x,y,a) := \begin{cases} -c_t^I(x,y,a_1,a_2) - c_t^A & \text{if } a_3 = -1, \\ -c_t^M - c_t^I(x,y,a_1,a_2) & \text{if } a_3 = 0, \\ -c_t^M - c_t^I(x,y,a_1,a_2) - c_t^O(a_1,a_2) + p\delta(x,a_1)\dfrac{\rho(a_2)}{\rho(y)}q & \text{if } a_3 = 1. \end{cases}$$

The optimal value function of our model, V_t, satisfies the following Bellman equations:

$$\begin{aligned} V_t(p,q,x,y) &= \max_{a \in \mathcal{A}(t)} v_t(p,q,x,y,a),\ \forall t = 0,1,\ldots,\mathsf{T}-1,\ (p,q,x,y) \in \mathcal{S} \\ v_t(p,q,x,y,a) &:= r_t(p,q,x,y,a) + 1\{a_3 \neq -1\}\alpha\mathbb{E}[V_{t+1}(\tilde{p}_{t+1}, f_t(q,x,y,a), a_1, a_2)|p_t] \\ V_{\mathsf{T}}(p,q,x,y) &:= -c_{\mathsf{T}}^I(x,y,0,0) - c_{\mathsf{T}}^A,\ \forall (p,q,x,y) \in \mathcal{S}. \end{aligned}$$

2.4. Data

In this section we discuss how we, in cooperation with production managers at EQT Corp., assessed the costs of operating our subject natural gas well located in Eastern Kentucky owned and operated by EQT Corp., the costs of investing and divesting in technology, and the well performance parameters. We focus our attention on this particular well as it

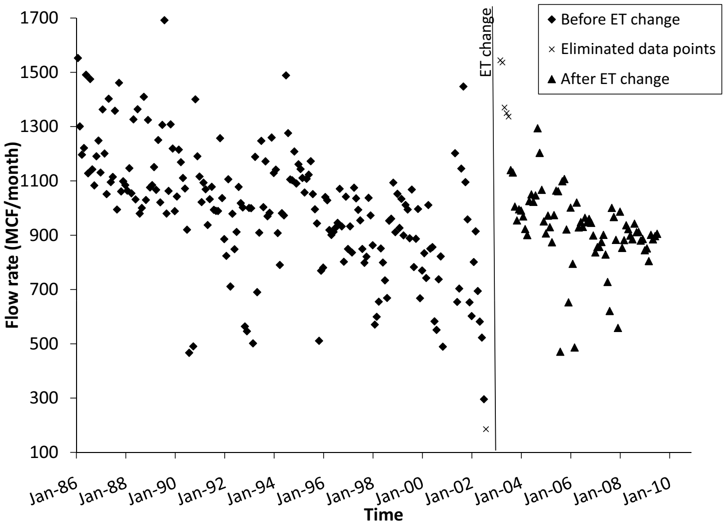


Figure 2.3: Historical (monthly) production data for the well studied in this chapter.

provides insights into the effect of technology change. We also discuss the estimation of the parameters of the natural gas price process.

Our well was drilled in the late 1950's. Although the flow rate of wells may show non-exponential decline behavior early in a well's life cycle (Garb and Larson 1989), we can safely assume our well has entered its phase of exponential flow rate decline. Our well was operated as a continuous flow well until 2003; in January 2003 it was equipped with soap and blow technology. (These types of ET technologies are described in Section 2.1.) We let 0 and 1 denote the continuous-flow and soap-and-blow ET levels, respectively. Data on the well's operational status has always been obtained periodically; we let this data collection approach be CT level 0. The possibility of employing an automated and continuous data collection approach, which we indicate by CT level 1, was being investigated at EQT Corp. at the time of this study.

Figure 2.3 shows monthly production data from the well from 1986 to 2009 (MCF = one thousand cubic feet). After the technology change in January 2003, the flow rate of our well increased and its decline rate decreased, which is consistent with our flow rate modeling approach discussed in Section 2.3. We used this data to estimate the parameters of our well flow rate evolution model.

After consultation with production managers at EQT Corp. we eliminated production rates of zero because they presumably originated from a failure of the measurement device. We also excluded "very low" production rates, two data points before the technology change, because they were deemed part of irregular well behavior caused by external factors. In addition, the well became unoperational for the five months before the change in technology. This led to a build-up of pressure and thus likely caused elevated production rates immediately after the well was brought back online. Thus, we eliminated the first five exceptionally high production rates after production resumed.

We estimated the parameters of the process describing the flow rate evolution using linear regression on the natural logarithms of the flow rate data. The estimate for $\mu(0,0)$ is 8.32% per year, and that of $\mu(1,0)$ is 3.19% per year; hence $\gamma(1) = 0.38$. Thus upgrading ET from continuous flow to soap and blow leads to significant operational improvement. We averaged the twelve retained data points immediately before and after the ET change in January 2003 to estimate the flow rate before and after this change. These values are 27.8 MCF/Day and 33.36 MCF/Day, which provides an estimate of 1.20 for $\delta(0,1)$. The estimate is consistent with the observations of Coşkuner and Strocen (2003), on which our model of flow rate behavior is based.

The production data from our example well did not have high enough granularity to estimate the yield function $\rho(\cdot)$, so we used up to 50 years of production data for a set of 135 of EQT Corp.'s soap and blow wells to estimate the yield. We identified periods during which these wells were loaded up and estimated $\rho(0)$ as the average fraction of time that the wells were producing, i.e., we directly estimated the quantity $1/[1 + \lambda/\eta(0)]$. This leads to an estimate for $\rho(0)$ of 90%. As no data was available on the effect of continuous monitoring, we assumed $\rho(1)$ to be equal to 99%, as this CT level would enable a more timely resolution of liquid load-up situations.

Table 2.1: Cost parameters for summer (winter).

		CT (monitoring frequency)	
		Periodic (0)	Continuous (1)
ET	Continuous flow (0)	100 (105)	110 (115.5)
	Soap and blow (1)	250 (262.5)	275 (288.8)

(a) Operating cost ($/month).

	ET	CT
Investment (0 → 1)	1,000 (1,050)	8,000 (8,400)
Divestment (1 → 0)	300 (315)	-2,667 (-2,540)

(b) Technology investment and divestment costs ($).

The relevant costs were estimated using various sources within EQT Corp. For confidentiality, these costs do not specifically represent those of our well, but are representative for this type of well. We distinguish between summer (April through October) and winter (November through March). Costs during the summer are as follows (winter costs are 5% higher): The maintenance cost is $321 per month. The abandonment cost is $36,250. The cost of operating the well is $100 per month with continuous flow ET and $250 per month with soap and blow ET, both at the lowest CT level, i.e., periodic monitoring. We estimated a 10% higher operating cost with continuous monitoring. The investment cost to move from continuous flow to soap and blow is $1,000. The divestment cost to move from soap and blow to continuous flow is $300. The CT that enables continuous monitoring costs $8,000 to install and has a salvage value of $2,677. Table 2.1 summarizes these costs (the winter cost are in parentheses).

We use NYMEX natural gas futures prices and prices of options on natural gas futures traded on 5/29/2009 to estimate the parameters of the price model of Jaillet et al. (2004). Figure 2.4 displays the NYMEX natural gas forward curve on this date, along with the fit

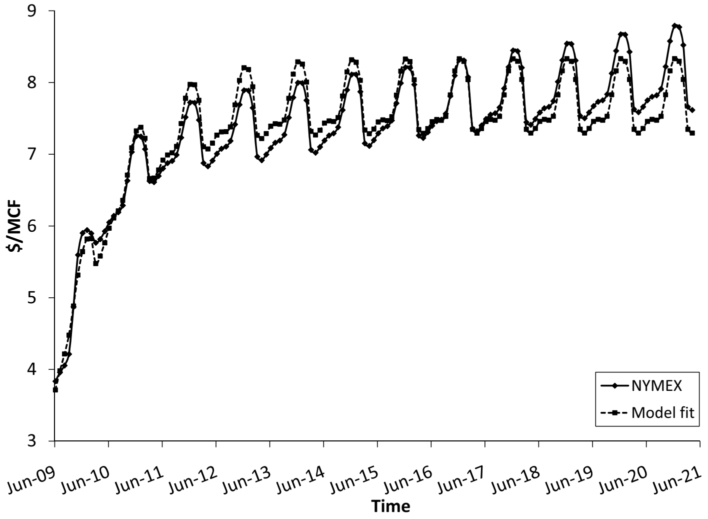


Figure 2.4: NYMEX natural gas forward curve on 5/29/2009 and model fit.

of our model. This curve exhibits significant seasonality. The model of Jaillet et al. (2004) features twelve deterministic monthly seasonality factors and a deseasonalized spot price, whose natural logarithm evolves as a single factor mean reverting process. The parameters of this model are the current level of this factor and its volatility, speed of mean reversion, and long term level; this last parameter is however not needed for valuation purposes (Jaillet et al. 2004). We used the method described by Lai et al. (2009) to estimate these parameters. This yielded an estimate of the natural gas spot price on 5/29/2009 equal to \$3.38 (the actual spot price was \$3.92), estimates of the volatility and speed of mean reversion parameters equal to 67% and 1.05, and estimated seasonality factors (displayed in Table 2.2) ranging between 0.946 (May) and 1.081 (January).

Table 2.2: Seasonality factors of the natural gas price.

Month	Jan	Feb	Mar	Apr	May	Jun
Seasonality factor	1.081	1.076	1.043	0.953	0.946	0.955
Month	Jul	Aug	Sep	Oct	Nov	Dec
Seasonality factor	0.967	0.971	0.970	0.977	1.016	1.059

2.5. Valuation Results

In this section we discuss the scaling option valuation results obtained by applying our stochastic dynamic program to the data described in Section 2.4 (we do not discuss the value of the abandoning option as our focus is on the technology and extraction scaling options).

We consider a 100-year horizon, with monthly stages. We use an annualized risk free discount rate of 5%. We implement a discrete time and space version of the price evolution model using a trinomial tree constructed by applying the method described by Jaillet et al. (2004), using monthly time intervals and the estimated natural gas spot price on 5/29/2009 as the initial price level. As the natural gas forward curve on 5/29/2009 only includes 144 monthly maturities, we reuse the last 12 prices in this forward curve to generate a 100 year forward curve that we use to calibrate this trinomial tree.

We also implement a discrete time and space version of the flow rate evolution, because modifications to the flow rate due to pausing and technology changes preclude the use of a single decline curve. We set the starting flow rate level equal to 27.8 MCF/Day; this is the average flow rate level of the 12 retained data points immediately preceding the ET level change discussed in Section 2.4. We use a grid in which the natural logarithms of the flow rate at the start of each month are equally spaced. The lowest flow rate considered is 0.001 MCF/Day, and each next value on the grid is a 1% increase on the previous. The largest value on the grid for the ET level 0 is 30.59 MCF/Day, which is the smallest grid value that exceeds 27.8 MCF/Day multiplied by 0.99/0.90 (as the time 0 CT level is 0). For the ET level 1 the largest value on the grid is 36.96 MCF/Day, which is the smallest entry on the grid that exceeds 30.59 MCF/Day multiplied by 1.20. If transitioning from one stage to the next according to (2.1) leads to a flow rate that is not on the grid, we linearly interpolate between the optimal value functions corresponding to the adjacent grid values.

The value of an option is the difference between the optimal value functions of our model in the initial stage, given the initial flow rate and price, with and without this option as part of the available decisions. The option value of a portfolio of options is calculated in a similar manner. We consider both the intrinsic and total (intrinsic plus extrinsic) values of an option. The former is the option value when the natural gas spot price evolves deterministically (approximated by setting the volatility parameter equal to 0.00001). The latter includes the effect of positive price volatility, i.e., set to its estimated value discussed in Section 2.4.

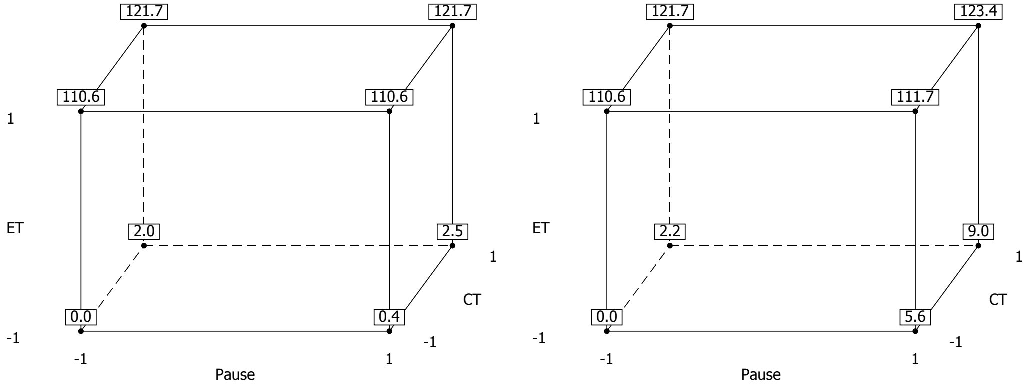
We find it useful to display our results using the cubes in Figure 2.5. The three axes of each cube correspond to the availability of an option (at -1 an option is not available, at $+1$ it is). Panels (a) and (b) of Figure 2.5 show the intrinsic and total option values, respectively, measured as percent increases from the base value of \$445,478. This is the value of the well when *(i)* it cannot be paused, *(ii)* no technology is available, and *(iii)* there is no price volatility. We select this as the base case because it allows us to distinguish between extrinsic and intrinsic values, and it yields the lowest value function for our model, thus making all the considered option values positive.

We discuss the option valuation results in detail in Sections 2.5.1-2.5.3, and provide a summary of our valuation results in Section 2.5.4. We point out that the real option valuations and managerial insights that we discuss are specific to the data that we use.

2.5.1 Technology Scaling Options

The left vertices of the cubes in Figure 2.5 display the valuation results of the technology scaling options, in isolation from the extraction scaling option; i.e., pausing is not available.

The ET scaling option has the largest total option value, increasing the well's total value by 110.6%, while the CT scaling option yields a 2.2% increase in total value (see Figure 2.5(b)). The value of the ET scaling option is entirely intrinsic. This value is obtained through access to larger reserves – recall that the flow rate level increases and its decline rate decreases when the ET level increases. The value of the CT scaling option



(a) Percent intrinsic value increase

(b) Percent total value increase

Figure 2.5: Valuation results; the displayed figures are percent increases over the well's intrinsic value without scaling options ($445,478)

exhibits very little sensitivity to price volatility: its intrinsic value is 2.0% in Figure 2.5(a), and its total value is 2.2% in Figure 2.5(b). As CT affects the speed at which one extracts natural gas, the high CT level allows one to produce at a higher rate when prices are high; conversely the low CT level allows one to reduce production, i.e., save natural gas, when prices are low. However, the additional value brought about by this flexibility in the presence of price volatility is minimal.

The total value of the ET scaling option in the presence of the CT scaling option increases from 110.6% to 121.7% − 2.2% = 119.5%. The total value of the CT scaling option in the presence of the ET scaling option increases from 2.2% to 11.1%. Thus, the two technology scaling options can be seen as complements: Considering both options simultaneously adds 121.7% to the well's total value, 8.9% more than the sum of the two options' individual values. As both technology options yield a percent increase in the immediate flow rate, their complementarity stems from the benefit of adding a percent increase to a higher flow rate, i.e., the flow rate obtained after increasing the level of one of the two technologies.

2.5.2 Extraction Scaling Option

The front, bottom vertices of the cubes in panels (a) and (b) of Figure 2.5 display the values of the extraction scaling option; i.e., pausing without any technology option. Pausing has little value without price volatility: The total value of this option is 5.6%, compared with an intrinsic value of 0.4%. Pausing can be used to delay production from a well during periods with low prices (it has other uses as well; see Example 3 in Section 2.6.1); i.e., to "save up gas" by postponing production until the price rises. Thus, pausing can be interpreted as a storage option, used when prices are low, which also saves operating costs. However, deterministic variations in prices, as expressed by the marked seasonality of the natural gas forward curve (see Figure 2.4 in Section 2.4), are not sufficient for this storage option to have significant value, which instead requires stochastic price variability.

2.5.3 Portfolio of Technology and Extraction Scaling Options

The right, top, and rear vertices of the cubes in Figure 2.5 display the option value of a portfolio of technology and extraction scaling options. Lowering the CT level and pausing have a similar function, i.e., they help "save up" natural gas, but pausing also saves the operating cost and does not require capital outlays. Given their similarity one would expect these options to behave as substitutes. However this is not the case here. Combining pausing with the CT scaling option increases the total value of the pausing option from 5.6% to 6.8% and the total value of the CT scaling option from 2.2% to 3.4%. The CT scaling and pausing options can hence be seen as complements, as their combined value of 9.0% exceeds the sum of their individual values, 7.8%.

This is consistent with the following observations. When the CT scaling option is available, operating CT at its high level increases the decline rate of the well's flow rate. This exhausts the well sooner, which extends the amount of time during which pausing can be used without the wasting of production (the interval between the time the well is exhausted and the end of the horizon or the time when financial discounting renders production virtually worthless). Conversely, pausing can be used to limit the costly reduction of the CT level, i.e., from its high to its low level, whenever the well is operated at the high CT level and this reduction would be otherwise beneficial, e.g., when prices are low (recall that the salvage value of the continuous monitoring technology is significantly lower than its associated initial capital outlay).

Using the pausing option in conjunction with the ET scaling option has an opposite effect. The total values of these options when used in isolation are 5.6% and 110.6%, respectively. However, once these options are jointly used their total values decrease to 1.1% and 106.1%, respectively. The pausing and ET scaling options can hence be seen as substitutes. This substitution effect can be explained as follows. Deploying ET at its high level increases the immediate flow rate and decreases its rate of decline. When the flow rate is sufficiently high, as it appears to be in our computational experiments, using the pausing option in conjunction with the high ET level would indeed amount to wasting production, by postponing it until much later in the horizon.

When we consider pausing in conjunction with the two technology scaling options, the substitution effect between pausing and the ET scaling option dominates the complementarity effect between pausing and the CT scaling option, as the total value of pausing in the absence of ET and CT, 5.6%, is higher than its total value in the presence of these two options, 1.6%. Hence, the pausing option and the portfolio that includes the two technology scaling option are substitutes.

Comparing the relevant vertices of the cubes in Figure 2.5 indicates that there is a complementarity relationship between each technology scaling option and each pair of the other two options. This reinforces our previous finding that the technology scaling options are complements. It also suggests that the complementarity between these options is stronger than the substitution relationships previously described.

Finally, we observe that the total option value of the portfolio that includes all the three scaling options, 123.4%, whose intrinsic value is 121.7%, is higher than the 110.6% +

$2.2\% + 5.6\% = 118.4\%$ increase brought about by employing the three options individually. Although this result is implied by the complementarity between the CT scaling and pausing options and the complementarity between the ET scaling option and the portfolio that includes the two other scaling options, its magnitude is notable.

2.5.4 Summary

We summarize our valuation results as follows: When used in isolation, the ET, CT, and extraction scaling options increase the value of the well by 110.6%, 2.2%, and 5.6%, respectively. The relative values of these options reflect their usages: Exercising the ET scaling option increases the size of the extractable reserve, a first order effect; exercising the CT scaling or the pausing options improves operational efficiency and effectiveness, respectively, second order effects. Our finding on the value of pausing contrasts with practice, where this option is typically equated to wasting production, and thus avoided if possible. The values of the technology scaling options are almost entirely intrinsic, while the value of pausing is almost fully extrinsic. Individual options and the portfolio that includes the two other relevant options are complements, excluding the ET scaling option and portfolio that consists of the CT scaling and pausing options. Likewise, the option value of the portfolio that includes all the three scaling options exceeds the sum of the values of the individual options.

2.6. Deployment Policies

In this section we study how to effectively deploy the technology and extraction scaling options together with the abandonment option; i.e., we analyze the optimal exercise of these options. We begin in Section 2.6.1 by studying the optimal deployment results of our model numerically, finding that the optimal deployment of the three scaling options is complicated. This prompts us to explore a simpler – more practical – approach to their deployment in Section 2.6.2, which captures most of their combined value.

2.6.1 Optimal Deployment Policy

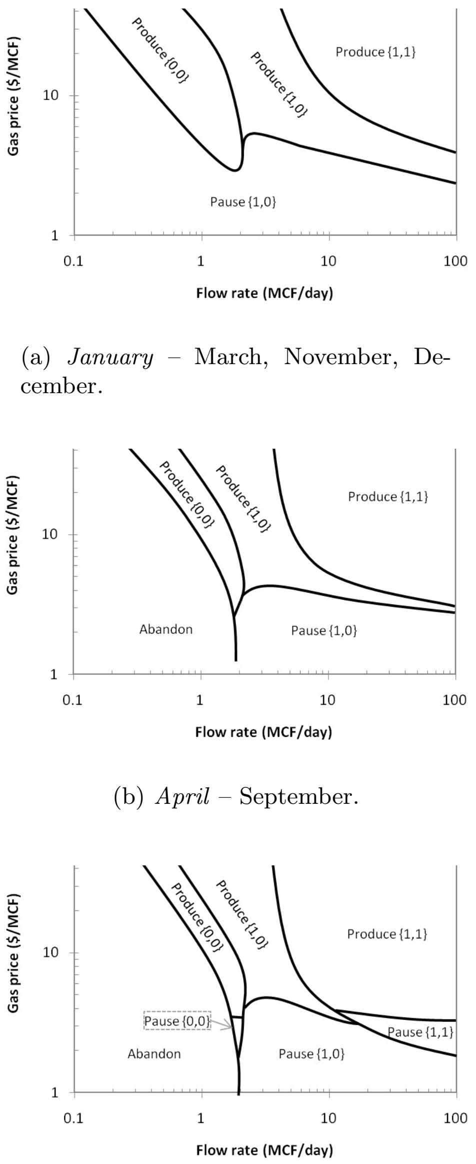
In this subsection we illustrate numerically that optimally deploying the three scaling options is rather complicated. We consider the same setting as in Section 2.5. Regarding the optimal deployment of the three scaling options, one might expect:

1. The optimally employed ET and CT levels to increase in the gas price and the flow rate; and
2. Pausing to be used primarily when prices are low (Smith and McCardle 1998, p. 211, and our discussion in Section 2.5.2).

However, this is not true in general, as shown in Examples 1-4 below. These examples are based on Figure 2.6, which illustrates the optimal deployment of the technology and extraction scaling options, as well as the abandonment option, in different time periods as a function of the flow rate and the natural gas price. The labels associated with each region of this figure indicate the type of optimal action in each such region, and also indicate the optimal deployed levels of ET and CT as the pair $\{x, y\}$. Figures 2.6(a) and 2.6(b) pertain to January and April, respectively, but they are also representative of the other winter and summer months indicated; Figure 2.6(c) is October. For ease of interpretation, Figure 2.6 uses log scales, but the labels and the tickmarks are in the original units to facilitate intuition about the magnitudes displayed. In these examples the initial technology level pair is $\{1, 0\}$. Similar results are obtained with different initial technology levels.

Example 1 *Consider any flow rate below 2 MCF/Day in Figure 2.6(a). At this flow rate, the optimally deployed ET level is nonmonotonic in the natural gas price (a similar behavior occurs for the CT level when the starting technology level pair is* $\{0, 1\}$*).*

Example 2 *Consider any price above 3 \$/MCF in Figure 2.6(a). At this natural gas price, the optimally deployed ET level is nonmonotonic in the flow rate (we observe a similar behavior for the CT level when the starting technology level pair is* $\{0, 1\}$*).*



(a) *January* – March, November, December.

(b) *April* – September.

(c) *October*.

Figure 2.6: Optimal technology deployment and operating decisions in different stages as functions of the flow rate and the natural gas price with the ET level equal to 1 and the CT level equal to 0. The captions of the panels indicate the period(s) for which the panel is representative: the month pertaining to the displayed panel is italicized.

Example 3 *Consider a price roughly equal to 3 \$/MCF in Figure 2.6(a). At this natural gas price, the optimal pausing decision is nonmonotonic in the flow rate. For gas prices near this level it is optimal to pause at low flow rates, produce at flow rates between 1 and 2 MCF/Day and at very high flow rates, while it is optimal to pause for flow rates between 2 and 30 MCF/Day. At low flow rates, and even for high natural gas prices, pausing is used to delay abandoning until summer, as shown in Figure 2.6(b), when abandoning is cheaper. (At high flow rates pausing is used to save up gas, as discussed in Section 2.5.2.) This exemplifies how the seasonality of the cost structure affects the optimal policy. This is further illustrated in Example 4.*

Example 4 *Figure 2.6(c) shows that in the "Pause $\{0, 0\}$" and "Pause $\{1, 1\}$" regions it is optimal to change one of the technology levels, which are initially at the $\{1, 0\}$ levels, and delay production at the new technology level. This occurs because in winter, which starts in November, the cost structure is less favorable than in summer.*

2.6.2 Heuristic Deployment Policy

Given the complex structure of optimally deploying the three scaling options, we seek a simpler, but still effective approach to their deployment. We start by studying the benefit of actively managing the deployment of the technology scaling options over time, comparing our valuation results discussed in Section 2.5 with those obtained by running our model by fixing ET and CT at given levels throughout the time horizon. Figure 2.7 displays the relevant total value increases relative to the base case. Note that we still actively manage the pausing and abandonment options (the former option when it is available).

Comparing the cubes in Figures 2.5(b) and 2.7 reveals that actively managing the technology scaling options yields only a very small increase in the value of the well. Specifically, only in the absence of the ET scaling and the pausing options is there a positive difference (for the CT scaling option) in valuations, and even in this case the difference is very small. Hence, essentially *all* of the value increase brought about by using an optimal deployment

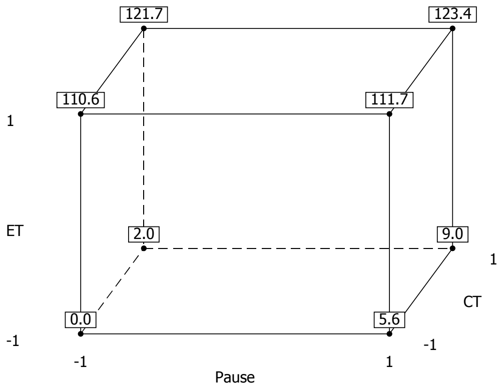


Figure 2.7: Valuation results with fixed technology scaling options; the figures displayed are the percent total value increases relative to the base case ($445,478).

policy can be attained by the following simpler policy: *(i)* exercise the two technology options immediately by switching to their high levels, and *(ii)* optimally manage the pausing and abandonment options thereafter. Though not explicitly illustrated here, we also expect this policy simplification to be valuable when managing a portfolio in which a larger set of technology levels is considered.

2.7. Conclusions

In this chapter we study the interaction between technology and extraction scaling options in natural gas production by developing and applying a stochastic dynamic program that uses empirical production and financial data. Our research is grounded in practice; it was developed in conjunction with production managers at EQT Corp. We bring to light basic managerial principles related to the drivers of these options' values and their deployment. These principles are likely to have managerial relevance in other commodity production settings, such as oil extraction and mining.

Our work has the potential to be extended in several dimensions. We employ a single-factor mean reverting model to describe the evolution of the price of natural gas. Although mean reversion in commodity prices is well documented in the literature (Clewlow and

Strickland 2000, Chapter 2, Smith and McCardle 1999), more elaborate, multifactor models of commodity price evolutions are also available, such as the two-factor reduced-form and equilibrium models of Schwartz and Smith (2000) and Routledge et al. (2000), respectively. Our model uses a deterministic representation of the evolution of the flow rate of a natural gas well. It would be interesting to extend our work by using a stochastic model of production, as is often done in the real option literature (see, e.g., Smith and McCardle 1998). This extension might benefit from the application of the method described by Hahn and Dyer (2008) to model the discrete time and space evolution of correlated mean reverting processes. In addition, our analysis is based on data pertaining to a specific natural gas well; it would be of interest to examine the dependence of our conclusions on this data by replicating our analysis based on data from other wells.

More broadly, one could extend our work to consider the strategic management of the operations of a natural gas field, which includes multiple wells. This presents additional research opportunities related to modeling the transfer of technology across wells, the dependence among the flow rates of different wells, and the staffing and routing of maintenance operators. Moreover, when a production firm has a pre-committed production level and operates a gas field whose wells' flow rates are subject to natural decline, there may be multiple options to maintain production at the pre-committed level, including drilling more wells. Our research suggests technology deployment as an alternative approach to meet the pre-committed production level. It would be interesting to evaluate the financial effect of this flexibility. This would require an even more complete well and operations model to asses the trade-offs between investing in current wells or drilling new ones. Such a strategic model could start at the pre-drilling phase, in which case the initial investment costs may depend on well characteristics that may only be discovered after drilling occurs.

Chapter 3

Inventory Rationing for a System with Heterogeneous Customer Classes[1]

3.1. Introduction

Managers often face demand, and thus differentiate between, customers that expect different standards of service. One class of industries that recognizes and implements customer differentiation along these lines are those that deliver and maintain expensive capital goods requiring high up-times; examples include defense systems (e.g. Deshpande et al. 2003a and 2003b), semiconductor manufacturing equipment (e.g. Kranenburg and Van Houtum 2008), and mobile phone operating systems (e.g. Möllering and Thonemann 2008). In all these cases customers are assigned a priority level based on equipment criticality or demand type, for example demand from a machine that is down may have higher priority than a replenishment demand from a stockpoint in the network.

Various tools have been developed to differentiate between such customers: a common

[1] This chapter is joint work with Ivo Adan, Geert-Jan van Houtum, and Alan Scheller-Wolf.

approach is the use of a *critical level* policy that reserves some inventory for the more important customer class. Specifically, current state information (e.g. amount of inventory in hand) is used to deny some customers access to inventory, in order to reserve this stock to serve more important demands that have yet to arrive. This type of policy yields considerable benefits when compared to cases in which all customers receive the same level of service, or when separate inventories are kept for each customer class. These types of problems have been studied under varying assumptions: see e.g. Veinott (1965), Topkis (1968), Ha (1997a), Cattani and Souza (2002), Dekker et al. (2002), De Véricourt et al. (2002), Deshpande et al. (2003b), Möllering and Thonemann (2008), and Kranenburg and Van Houtum (2007, 2008). An assumption common to this literature is that all customer classes behave similarly, i.e. either all classes leave and the sale is lost when not immediately satisfied, or all are willing to wait and are backordered.

We study a mixed problem in which one customer class leaves when demand is not satisfied immediately, while the other customer class is willing to wait while demand is backordered. This characteristic, in fact, may be the basis of customer differentiation. We see, at least, three application areas for our model:

1. Consider a retailer that faces demand from both loyal (demanding), long term customers with high service level requirements and occasional walk-in customers. One can imagine the retailer holding back some inventory to serve anticipated demands from loyal customers while turning down walk-in customers. A situation like this is described by Gans and Savin (2007) for rental cars.

2. When operating a physical store in combination with an online shop the customers issuing their demand in the store observe actual inventory and may leave unsatisfied if the desired item is not available. The online customers can be backlogged when inventory is low and still be considered satisfied, since they anticipated some lead time anyways. Cattani and Souza (2002), Swaminathan and Tayur (2003), and more recently Duran et al. (2008), identify the opportunity to differentiate between types.

3. An OEM may operate a central warehouse as well as a network of local warehouses from which it serves its customers. The central warehouse must satisfy replenishment demands from the local warehouses as well as emergency demands directly from customers; these latter demands occur when a customer's machine is down and the nearest local warehouse does not have the desired item. In this case the emergency demand has priority over replenishment demands, which can be delayed. This situation was recognized by Alfredsson and Verrijdt (1999), Deshpande et al. (2003a, 2003b), and Möllering and Thonemann (2008), among others.

As mentioned above, nearly all previous papers consider homogeneous customer behavior – either all customers are willing, or all customers are unwilling to wait. Because of this, the methods and models used by other papers cannot be readily extended to the heterogeneous customers that we consider. We further discuss the related literature in Section 3.2.2.

In this chapter we make several contributions.

1. We are the first to thoroughly analyze using a critical level in response to customer classes reacting differently to being denied an item. This important characteristic in practice has only been modeled in a limited fashion before.

2. We develop an exact evaluation procedure for a given CL policy using matrix analytic methods, and prove monotonicity properties of the main performance measures via sample path analysis.

3. Using these monotonicity properties we develop an efficient optimization procedure, which avoids enumerating over large numbers of potentially optimal policies.

4. We demonstrate the near-insensitivity of the performance of the optimal CL policy to lead time distribution variability. This near-insensitivity implies that the assumption of exponential lead times that is needed in the analysis has little effect on the solution.

5. Finally, we benchmark the performance of the optimal CL policy against the globally optimal (state dependent) policy and two alternative, more naïve, policies. This

provides insights into when it makes sense to use a critical level policy. The comparison of CL policies to the globally optimal is surprisingly absent in the literature: Kaplan (1969), Dekker et al. (2002), Ha (1997a), and Möllering and Thonemann (2008) all compare to more naïve policies only. And, when comparisons to the globally optimal policy are made, this is often only done in oversimplified systems (see e.g. Benjaafar et al. 2006, who consider only 1 customer class). We show that our CL policy improves significantly upon the more naïve policies *and* performs near optimally.

The remainder of the chapter is structured as follows. First we will introduce our model and review the related literature in Section 3.2. Section 3.3 develops an evaluation procedure to compute the performance of a CL policy at any desired level of exactness. An efficient optimization algorithm that bounds the enumeration space using monotonicity results is presented in Section 3.4. Section 3.5 details our numerical experiment comparing the performance of the optimal, CL and more naïve, policies, provides insight into sensitivity with respect to lead time variability, and studies the efficiency of our bounds. Furthermore, some insight into the structure of the globally optimal policy is provided. Section 3.6 introduces several extensions and outlines how some of these can be incorporated in our model with relative ease. Section 3.7 presents our conclusions.

3.2. Model and related literature

In this section we first describe of our model as well as our main assumptions. Then, having detailed our model, we briefly review the related literature and how it compares with our model.

3.2.1 Model description

We consider a single stockpoint where a single product is kept on stock. Customer classes are denoted by $j = 1, 2$; class 1 has the highest priority and its demand is lost if not

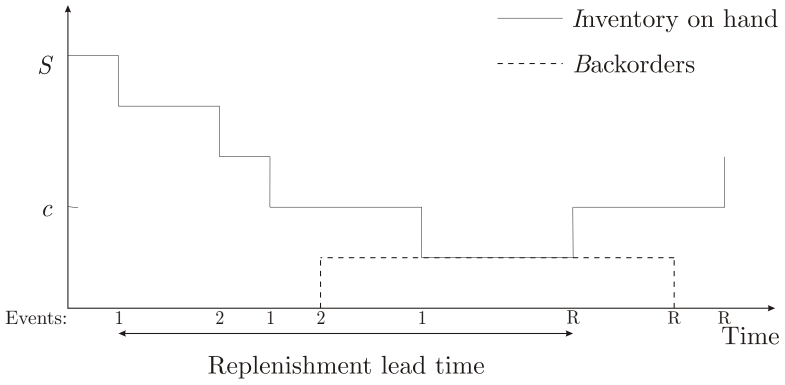


Figure 3.1: An illustration of the critical level policy

immediately satisfied from stock. Class 2 has lower priority and its demand is backordered if not immediately satisfied. Demands of class j arrive according to a Poisson process with rate λ_j, and the total demand rate is denoted by $\lambda = \lambda_1 + \lambda_2$. Inventory is controlled using a continuous review *critical level* (CL) policy, which reserves inventory for the most important customer class by backordering class 2 as soon as inventory drops below a certain *critical level*. Backorders are delivered as soon as inventory on hand increases *above* the critical level.

We impose a static base stock level denoted by S, and let c denote the critical level, with $S, c \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Replenishment orders are assumed to have a exponential lead times[2] with mean μ^{-1}. Orders need not arrive in the order in which they are placed. An illustration of the behavior of inventory and backorder levels under this policy can be found in Figure 3.1. Events $j = 1, 2$ denote demands from customer class j, and R denotes the arrival of a replenishment order. In Section 3.5.2 we generalize our analytical results to lead times that have higher variability, distributed as degenerate hyperexponential random variables. Furthermore, using simulation we investigate the effect of lead times with higher

[2] The assumption of exponential lead times considerably simplifies the analysis. Furthermore, we expect that the performance of the optimal CL policy will be fairly insensitive to the distribution of the lead time, because our model is a combination of the $M|G|\infty$ and $M|G|C|C$ queueing models, both of which have steady state queue length distributions known to be insensitive to the distribution of the service time, see e.g. Cohen (1976).

and lower variability in Section 3.5.2.

We seek to minimize the infinite horizon expected cost of a policy, $C(S,c)$, which can be separated into three different types of cost. First, a one-time penalty cost $p_j \geq 0$ is incurred whenever a demand of class j is not immediately satisfied from stock. Second, a backorder cost $b \geq 0$ is incurred per unit per unit time a (class 2) backorder exists. Third, an inventory holding cost $h \geq 0$ is charged per unit per unit time an item is on hand. We denote the fraction of demand from class j that is immediately satisfied from stock (the fill rate) by $\beta_j(S,c)$, the average number of backorders by $B(S,c)$, and the average inventory by $I(S,c)$. This leads to the following optimization problem:

$$\begin{aligned}\min_{S,c} C(S,c) &= \min_{S,c} \{p_1\lambda_1(1-\beta_1(S,c)) + p_2\lambda_2(1-\beta_2(S,c)) + bB(S,c) + hI(S,c)\} \quad (3.1)\\ s.t. \quad & c \leq S,\\ & S,c \in \mathbb{N}_0.\end{aligned}$$

To solve (3.1) we first develop an efficient, exact procedure to determine the cost of a given CL policy, $C(S,c)$, in Section 3.3. Then we develop an efficient optimization procedure that eliminates large sets of potentially optimal values for S and c, bounding our enumeration space, in Section 3.4. We find the optimal S and c by enumerating over the reduced space.

3.2.2 Related literature

The policy we describe in Section 3.2.1 belongs to the class of rationing or critical level policies. Veinott (1965) introduced the CL policy, and since then the performance of such policies has been extensively studied. We focus on the case with a single static critical level. This in contrast to e.g. Evans (1968), Topkis (1968), Kaplan (1969), Melchiors (2003), and Teunter and Klein Haneveld (2008); in their papers the critical level depends on the time remaining until the next replenishment arrives. This also is different from policies in which the critical level is state dependent, e.g. Benjaafar et al. (2006). A single static critical level is easy to explain to practitioners and to implement, as it does not depend on the progress of items beyond your control, i.e. in the replenishment pipeline. In Section 3.5.3 we will compare the optimal state dependent policy with our policy.

Within the class of papers having a single, static, critical level, we distinguish problems by the way customers react to unsatisfied demand. Studies in which demand from each class is lost when not immediately satisfied have been performed by Ha (1997a), Ha (2000), Melchiors et al. (2000), Dekker et al. (2002), Frank et al. (2003), and Kranenburg and Van Houtum (2007). Ha (1997a) studies a continuous review model with a Poisson demand processes, and a single exponential replenishment server. He proves the optimality of CL policies and shows that both the base stock level and the critical level are time-independent. In Ha (2000), Ha (1997a) is extended to include Erlang distributed lead times. Dekker et al. (2002) consider a model similar to the one studied by Ha (1997a) but assume an ample exponential replenishment server; they derive exact procedures for determining the optimal CL policy. Melchiors et al. (2000) generalizes Dekker et al. (2002) by including a fixed order quantity. They optimize the order quantity, base stock level, and the critical level. Frank et al. (2003) consider periodic review models with fixed lead times (i.e. ample replenishment servers) for which they find the optimal policy parameters. Many of the solution approaches described above are computationally expensive for more than two demand classes. Kranenburg and Van Houtum (2007) divide larger problems into subproblems, and develop efficient heuristic algorithms for these subproblems (one for each customer class). These heuristics are tested on a large testbed and shown to perform well. This increase in speed allows for application in a multi-item setting as demonstrated in Kranenburg and Van Houtum (2008).

The other primary subclass is that in which demands from both classes are backordered when they cannot be met from stock. This is studied by Nahmias and Demmy (1981), Ha (1997b), Dekker et al. (1998), De Véricourt et al. (2002), Deshpande et al. (2003b), Duran et al. (2008) and Möllering and Thonemann (2008). Nahmias and Demmy (1981) are the first to evaluate the performance of a system with two classes that are backordered when not immediately satisfied. They assume that there is at most one outstanding replenishment order to facilitate their analysis; this assumption remains common to date in this stream of literature. Ha (1997b) and De Véricourt et al. (2002) derive the optimal allocation policy in a make-to-stock capacitated assembly system in which demands from all classes (two classes in Ha 1997b, n classes in De Véricourt et al. 2002) are backordered if not immediately satisfied. Other than customer behavior when a demand is not immediately satisfied De Véricourt et al. (2002) use the same assumptions as Ha (1997a). Dekker et al. (1998) derive an approximation to the performance of a given policy under Poisson demands and deterministic lead times under a lot-for-lot inventory management policy. Deshpande et al. (2003b) study a problem with two customer classes and rationing under a (Q, r) policy, also clearly outlining what complicates the problem when demands are backordered: (i) one has to determine the order in which backorders are optimally cleared, and (ii) if the optimal clearing mechanism is used extensive state information is needed. Möllering and Thonemann (2008) study a periodic review model with arbitrary, discrete, demand distributions and a lead time that is an integer multiple of the review period. Duran et al. (2008) consider the finite horizon problem for which they find the optimal policy in terms how much inventory to reserve, how many demands to backorder (the alternative is to reject them) and what level to order-up-to.

We study the combination of these two subclasses of policies; demand from one class is lost and the other is backordered. So far, this policy has received little attention in the literature. It is one of several policies compared by Cattani and Souza (2002), who assume Poisson demand, and a single, exponential, replenishment server. They determine the parameters of the optimal policy through exhaustive search over a suitably large state space. Compared to Cattani and Souza (2002), our replenishment system can operate either a single, several parallel, or an ample number of replenishment servers. We will focus on the ample server case as this captures practical settings we wish to model (see below), and as the other cases are special (and easier) cases. Furthermore we avoid enumeration over a suitably large state space by the development of bounds on the cost of a policy. Hence, Cattani and Souza (2002) can be thought of as containing a special case of our problem.

In a practical setting, the ample server assumption is motivated, for example, by the problems studied by Gans and Savin (2007) or Kranenburg and Van Houtum (2008). Gans and Savin (2007) studies a car-rental problem in which every car that has been rented has an exponentially distributed rental period. In Kranenburg and Van Houtum (2008), like many other papers in the spare parts literature, lead times are negotiated with suppliers such that the supplier is required to deliver within a specified window, no matter how many orders are issued. Suppliers are able to meet these requirements as they generally supply a variety of items to different customers and hence have ample capacity when observed from the point of view of a single item.

Throughout the literature several assumptions on lead times have been made; we assume exponential lead times initially, and then generalize to degenerate hyperexponential. In addition, we determine the cost of the globally optimal policy, without assuming static critical levels, using dynamic programming, and compare the performance of our policy to the globally optimal policy and two alternative, more naïve, policies. We also compare the robustness of both the CL and the globally optimal policy and establish for the first time that the CL policy is typically more robust to changes in lead time variability than the globally optimal policy. In fact, the optimal CL policy determined under exponential lead times may even outperform the globally optimal policy determined under exponential lead times when they are utilized in situations with non-exponential lead times.

There are some other related fields in the literature that deserve mentioning. In the revenue management literature policies similar to the critical level policy are commonplace. Most closely related are booking limits, these limit access to parts of the inventory to specific demand classes. For a review of the literature we refer the reader to Talluri and Van Ryzin

(2004). In general the revenue management literature deals with a perishable item (like a hotel room or an airline seat) that can only be sold once, while we deal with inventory that can be utilized at any point in time. Wang (2008) study a setting in which customers are willing to accept flexible delivery, up to a certain deadline. A key differentiator of our work is that we partition the customers in those that are and are not willing to wait while their customers are always willing to wait and differ in how long they want to wait. In addition we focus on different penalty cost to differentiate between customers of specific types. In spare-parts management the concept of lateral transshipments (see e.g. Paterson et al. 2009, Wong et al. 2006) relates to our work. Specifically the allocation of inventory to "own" demand vs demand from another location. Van Wijk et al. (2009) derive the optimal policy for lateral transshipments between warehouses. Our model resembles theirs, except for a key assumption, which we will highlight when discussing the optimal policy structure in Section 3.5.3.

3.3. Evaluation

Our model, under CL policy, (S, c), can be described by a Markov process with states (m, n), where $m \in \mathbb{N}_0$ represents the number of items on hand, $n \in \mathbb{N}_0$ the number of items backordered. The state space and transition scheme of this policy is depicted in Figure 3.2.

In Figure 3.2 two categories of transitions can be recognized. First, demand-related transitions that decrease the amount of stock or increase the number of backorders: Transitions from $(m, 0)$ to $(m-1, 0)$ occur at rate λ as long as $m > c$ (both classes are served). If $0 < m \leq c$ transitions from (m, n) to $(m-1, n)$ occur at rate λ_1 and transitions to $(m, n+1)$ occur at rate λ_2 (class 1 is served, class 2 is backordered). If $m = 0$ the only demand related transition is from $(0, n)$ to $(0, n+1)$ which occurs at rate λ_2, since class 1 demand is lost. Second, we have supply related transitions that decrease the number of backorders or increase the amount of inventory: All supply related transition occur at rate

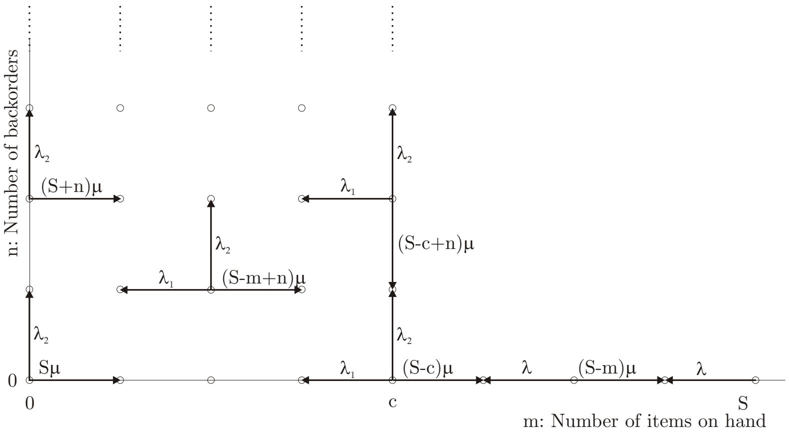


Figure 3.2: Transition scheme of our critical level policy

$(S-m+n)\mu$ since there are $S-m+n$ outstanding orders. If $m=c$ and $n>0$ these transitions go from (m,n) to $(m,n-1)$ (a backorder is cleared); all other supply related transitions result in a transition from (m,n) to $(m+1,n)$.[3] Note that states with both $m>c$ and $n>0$ are transient.

Let $\pi_{m,n}$ denote the steady state probabilities of our Markov chain. Since both customer classes arrive according to a Poisson process we can use PASTA (Wolff 1982) to evaluate the cost as defined in (3.1). To do so we need four performance measures, expressed in terms of $\pi_{m,n}$ as follows:

$$\beta_1(S,c) = 1-\sum_{n=0}^{\infty}\pi_{0,n}, \tag{3.2}$$

$$\beta_2(S,c) = \sum_{m=c+1}^{S}\pi_{m,0}, \tag{3.3}$$

$$I(S,c) = \sum_{m=1}^{S} m\pi_{m,0}+\sum_{m=1}^{c}\sum_{n=1}^{\infty} m\pi_{m,n}, \tag{3.4}$$

$$B(S,c) = \sum_{m=0}^{c}\sum_{n=1}^{\infty} n\pi_{m,n}. \tag{3.5}$$

[3] In case of $N \in \{0,1,\ldots,\}$ parallel replenishment servers the replenishment rates are $N\mu$ at maximum.

In the next two subsections we develop an efficient procedure to evaluate the above performance measures.

3.3.1 Structure of the Markov process

Our solution procedure exploits the structure of our Markov process. We partition the set of all states into *levels* according to the number of backorders n. Level n consists of the following states:

$$\begin{array}{ll} \{(0,0),(1,0),\ldots,(c,0),\ldots,(S,0)\} & \text{for level } n=0, \\ \{(0,n),(1,n),\ldots,(c,n)\} & \text{for level } n>0. \end{array}$$

According to this partitioning, the generator Q of the Markov process is given by:

$$Q = \left(\begin{array}{cccccc} B_0 & B_1 & 0 & 0 & 0 & \cdots \\ B_{-1} & A_0(1) & A_1 & 0 & 0 & \cdots \\ 0 & A_{-1}(2) & A_0(2) & A_1 & 0 & \cdots \\ 0 & 0 & A_{-1}(3) & A_0(3) & A_1 & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{array} \right), \tag{3.6}$$

where B_0, B_{-1} and B_1 are matrices of size $(S+1)$x$(S+1)$, $(c+1)$x$(S+1)$ and $(S+1)$x$(c+1)$ respectively; and $A_0(n)$, $A_{-1}(n)$ and A_1 are matrices of size $(c+1)$x$(c+1)$. A more detailed description of these matrices is given in Appendix A.1.

Note that Q is a Quasi-Birth-Death process. In case of level *independent* matrices, i.e., $A_{-1}(n) \equiv A_{-1}$ and $A_0(n) \equiv A_0$, standard Matrix Analytic Methods (MAM) can be applied to compute the steady state distribution (see e.g. Neuts 1981, Lautouche and Ramaswami 1987). In our case, the matrices $A_{-1}(n)$ and $A_0(n)$ *do* depend on level n, which complicates the computation of the steady state distribution (see e.g. Bright and Taylor 1995). However, the process' characteristic that there is only one transition from level n to $n-1$, from (c,n) to $(c,n-1)$ considerably simplifies our analysis. This enables us to determine the $\pi_{m,n}$

exactly, via recursion, as we demonstrate below.

Let $\boldsymbol{\pi}_n$be the vector of steady state probabilities at level n:

$$\begin{aligned} \boldsymbol{\pi}_0 &= (\pi_{0,0}, \pi_{1,0}, \ldots, \pi_{c,0}, \ldots, \pi_{S,0}) \\ \boldsymbol{\pi}_n &= (\pi_{0,n}, \pi_{1,n}, \ldots, \pi_{c,n}) \qquad\qquad , n \in \mathbb{N} \end{aligned}$$

Let $\tilde{\boldsymbol{\pi}}_n$be the solution to:

$$\tilde{\boldsymbol{\pi}}_0 B_0 + \tilde{\boldsymbol{\pi}}_1 B_{-1} = 0 \qquad \text{for } n = 0 \tag{3.7}$$

$$\tilde{\boldsymbol{\pi}}_0 B_1 + \tilde{\boldsymbol{\pi}}_1 A_0(1) + \tilde{\boldsymbol{\pi}}_2 A_{-1}(2) = 0 \qquad \text{for } n = 1 \tag{3.8}$$

$$\tilde{\boldsymbol{\pi}}_{n-1} A_1 + \tilde{\boldsymbol{\pi}}_n A_0(n) + \tilde{\boldsymbol{\pi}}_{n+1} A_{-1}(n+1) = 0 \qquad \text{for } n \geq 2 \tag{3.9}$$

$$\tilde{\pi}_{S,0} = 1, \tag{3.10}$$

where $\tilde{\boldsymbol{\pi}}_n$ is defined similar to $\boldsymbol{\pi}_n$, but now in terms of $\tilde{\pi}_{m,n}$ instead of $\pi_{m,n}$. Note that in (3.10) $\tilde{\boldsymbol{\pi}}_n$ is normalized by setting $\tilde{\pi}_{S,0} = 1$ instead of using $\sum_{n=0}^{\infty} \boldsymbol{\pi}_n \boldsymbol{e}$=1, which cannot be determined *yet*.

Note that equations (3.7)-(3.9) relate the steady state probabilities of level n to those of levels $n-1$ and $n+1$ as is standard in applying MAM, however, as $A_0(n)$ and $A_{-1}(n+1)$ still depend on n, we cannot readily apply MAM. But the following lemma offers a solution methodology. Define the $(c+1) \times (c+1)$ matrix A as:

$$A = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{3.11}$$

Lemma 1 *The $\tilde{\boldsymbol{\pi}}_0$, and $\tilde{\boldsymbol{\pi}}_1$ can be determined by solving:*

$$\begin{aligned} \tilde{\boldsymbol{\pi}}_0 B_0 + \tilde{\boldsymbol{\pi}}_1 B_{-1} &= 0 \qquad\qquad \textit{for } n = 0 \\ \tilde{\boldsymbol{\pi}}_0 B_1 + \tilde{\boldsymbol{\pi}}_1 A_0(1) + \lambda_2 \tilde{\boldsymbol{\pi}}_1 A &= 0 \qquad\qquad \textit{for } n = 1 \\ \tilde{\pi}_{S,0} &= 1, \end{aligned}$$

and for $n > 1$ the $\tilde{\boldsymbol{\pi}}_n$ follow from:

$$\tilde{\boldsymbol{\pi}}_n = -\tilde{\boldsymbol{\pi}}_{n-1} A_1 (A_0(n) + A\lambda_2)^{-1} \qquad \textit{for } n \geq 2. \tag{3.12}$$

The proof of Lemma 1, along with all other proofs, can be found in Appendix A.2.

In Lemma 1 the steady state probabilities of levels 0 and 1 can be solved explicitly, and the steady state probabilities of level n are expressed in terms of level $n-1$ only, leveraging the special structure in our Markov process. The original $\pi_{m,n}$ can then be calculated as follows:

$$\boldsymbol{\pi}_n = \frac{\tilde{\boldsymbol{\pi}}_n}{\sum_{n=0}^{\infty} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}.$$

As is standard in MAM, the infinite sums in the performance measures introduced in (3.2) through (3.5) need to be truncated. Next, we develop (tight) bounds for this truncation error, along with our general solution procedure.

3.3.2 Solution procedure

The performance measures can be written in terms of $\tilde{\boldsymbol{\pi}}_n$, again using PASTA:

$$\begin{aligned} \beta_1(S,c) &= \frac{\sum_{m=1}^{S} \tilde{\pi}_{m,0} + \sum_{n=1}^{\infty} \sum_{m=1}^{c} \tilde{\pi}_{m,n}}{\sum_{n=0}^{\infty} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \\ \beta_2(S,c) &= \frac{\sum_{m=c+1}^{S} \tilde{\pi}_{m,0}}{\sum_{n=0}^{\infty} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \end{aligned}$$

$$\begin{aligned} I(S,c) &= \frac{\sum_{m=1}^{S} m\tilde{\pi}_{m,0} + \sum_{n=1}^{\infty}\sum_{m=1}^{c} m\tilde{\pi}_{m,n}}{\sum_{n=0}^{\infty} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \\ B(S,c) &= \frac{\sum_{n=1}^{\infty} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}{\sum_{n=0}^{\infty} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}. \end{aligned}$$

Lower bounds for the infinite sums appearing in these expressions are easily found by truncation. Upper bounds follow from the next lemma.

Lemma 2 *For all $\ell \geq 1$,*

$$0 \leq \sum_{n=c+\ell}^{\infty} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e} \leq U(\ell)$$

where

$$U(\ell) = (\tilde{\pi}_{0,\ell}, \tilde{\pi}_{1,\ell+1}, \ldots, \tilde{\pi}_{c,\ell+c})\, \boldsymbol{e}\ (S+\ell)! \left(\frac{\mu}{\lambda}\right)^{S+\ell} \left[\frac{\lambda}{\mu}\phi(S+\ell-1) - (S-c)\phi(S+\ell)\right],$$

and

$$\phi(\ell) = \sum_{k=\ell}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} = e^{\frac{\lambda}{\mu}} - \sum_{k=0}^{\ell-1} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}. \tag{3.13}$$

The intuition behind Lemma 2 is that the definition of *diagonal layers*: $\{(0,n), (1,n+1),\ldots,(c,n+c)\}$, for $n \geq 0$ highlights a structural property of the Markov process. The transition rate from each of the states on a diagonal layer to the right (i.e. down to diagonal layer $n-1$) is $(S+n)\mu$ and the flow to the left (i.e. up to the diagonal layer $n+1$) is upper bounded by λ. This structure is exploited in upper bounding the probability mass above the truncation level. The proof of Lemma 2 is given in Appendix A.2.2.

Now we can bound our performance measures from above and below by either ignoring the mass above the truncation level or using the bound from Lemma 2:

$$\begin{aligned} \frac{\sum_{m=1}^{S} \tilde{\pi}_{m,0} + \sum_{n=1}^{c+\ell}\sum_{m=1}^{c} \tilde{\pi}_{m,n}}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e} + U(\ell+1)} &\leq \beta_1(S,c) \leq \frac{\sum_{m=1}^{S} \tilde{\pi}_{m,0} + \sum_{n=1}^{c+\ell}\sum_{m=1}^{c} \tilde{\pi}_{m,n} + U(\ell+1)}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \\ \frac{\sum_{m=c+1}^{S} \tilde{\pi}_{m,0}}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e} + U(\ell+1)} &\leq \beta_2(S,c) \leq \frac{\sum_{m=c+1}^{S} \tilde{\pi}_{m,0}}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \\ \frac{\sum_{m=1}^{S} m\tilde{\pi}_{m,0} + \sum_{n=1}^{c+\ell}\sum_{m=1}^{c} m\tilde{\pi}_{m,n}}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e} + U(\ell+1)} &\leq I(S,c) \leq \frac{\sum_{m=1}^{S} m\tilde{\pi}_{m,0} + \sum_{n=1}^{c+\ell}\sum_{m=1}^{c} m\tilde{\pi}_{m,n} + cU(\ell+1)}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \\ \frac{\sum_{n=1}^{c+\ell} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e} + U(\ell+1)} &\leq B(S,c) \leq \frac{\sum_{n=1}^{c+\ell} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e} + U(\ell+1)}{\sum_{n=0}^{c+\ell} \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}}, \end{aligned}$$

where $\ell \geq 1$. To compute the performance measures at the desired level of accuracy we start with $\ell = 1$ and increase ℓ, one unit at a time, until the upper and lower bound for each of the performance measures are sufficiently close. As $U(\ell+1)$ is expected to decrease very rapidly as $\ell \to \infty$, the bounds may already become tight for moderate values of the truncation level $c+\ell$. In the numerical experiment to be introduced later we observe that, for a maximum distance between the upper and lower bounds of 10^{-6}, ℓ varies between 6 and 29 with a mean of 11.9. Also, ℓ seems to increase in c as, with a higher critical level, there is more mass below c, which leads to the need of evaluating more levels for accuracy. Hence we conclude that, in practice, the bounds become tight very rapidly.

These truncation error bounds, and their quality, is important as exact recursive calculation of the steady state probabilities involves matrix inverses (3.12), which become more costly for larger values of c. Using these truncation error bounds we limit the number of matrix inversions.

3.4. Optimization

Recall that our goal is to find the parameters of the optimal CL policy, i.e. the optimal values of S and c. So far we can determine the performance of a given CL policy, i.e. for given S and c we can determine the cost (as defined in (3.1)). In this section we build this evaluation technique into a procedure for finding the optimal S and c parameters.

Even though our evaluation procedure avoids many matrix inversions, some are unavoidable to precisely evaluate the performance of each S and c. Here, we develop two sets of lower bounds on the optimal cost in terms of S and c respectively. These bounds allow us to eliminate candidate solutions and hence bound the enumeration space. We target to eliminate candidate solutions with large S and c as the size of the matrices that need to be inverted grows in S and c.

For the development of these bounds we first need monotonicity results for several performance measures with respect to the critical level, c. We obtain these monotonicity results using sample path arguments. Define $X(S,c)$ as the average pipeline stock, i.e. the average number of as yet undelivered items that have been ordered from a supplier. We prove the following monotonicity results:

Theorem 1 *The performance measures depend on c in the following manner:*

$$\begin{aligned} B(S,c) &\le B(S,c+1) && \textit{i.e. } B(S,c) \textit{ is monotonically increasing in } c, \\ X(S,c) &\le X(S,c+1) && \textit{i.e. } X(S,c) \textit{ is monotonically increasing in } c, \\ \beta_2(S,c) &\ge \beta_2(S,c+1) && \textit{i.e. } \beta_2(S,c) \textit{ is monotonically decreasing in } c. \end{aligned} \tag{3.14}$$

The results in Theorem 1 are in line with the literature for homogeneous customer classes, e.g. Ha (1997a), Dekker et al. (2002), Deshpande et al. (2003b), Kranenburg and Van Houtum (2007), and Möllering and Thonemann (2008), but our proof is more involved due to customer heterogeneity. Using the results from Theorem 1 we develop an efficient nested procedure to solve the optimization problem from Section 3.2. First, we need a lower bound for the costs:

Lemma 3 *A lower bound for $C(S,c)$ is given by:*

$$C(S,c) \ge C_{LB}(S,c) := p_2\lambda_2(1-\beta_2(S,c)) + (b+h)B(S,c) + h\left(S - \frac{\lambda}{\mu}\right)$$

The proof of Lemma 3 can be found in Appendix A.2.4. Next, consider the minimum cost for a fixed value of S, $\hat{C}(S)$, and notice that this can be bounded as follows:

Corollary 1 *A lower bound function for $\hat{C}(S)$ is given by*[4]*:*

$$\hat{C}(S) \ge \hat{C}_{LB}(S) := h\left(S - \frac{\lambda}{\mu}\right)$$

[4] A stronger lower bound is given by $p_2\lambda_2(1-\beta_2(S,0))+(b+h)B(S,0)+h\left(S-\frac{\lambda}{\mu}\right)$ but our computational experience is that the computational gain from this bound does not outweigh the additional computational cost of computing the bound value.

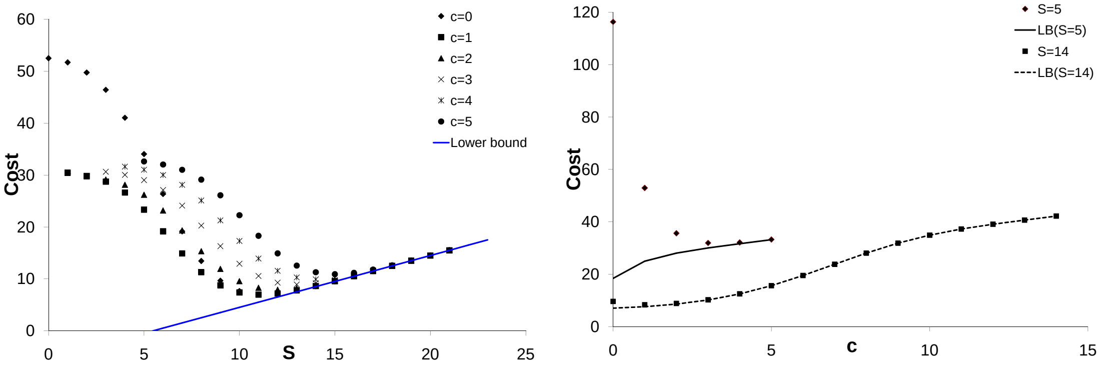
To find the optimal S, we increase S, one unit at a time, starting from $S = 0$. Let the optimal cost up to a certain value for S be denoted by $\hat{C}^*(S)$. We keep increasing S until $\hat{C}^*(S) \leq \hat{C}_{LB}(S+1)$ (note that $\hat{C}_{LB}(S)$ increases in S).

We now have a way of bounding S, but for a fixed value of S we also want to limit the number of values for c that we need to evaluate. This can also be done by using Lemma 3. Using the monotonicity properties from Theorem 1 we know that $C_{LB}(S,c)$ is increasing in c for given S. Thus, for given S, we increase c, one unit at a time, starting from $c = 0$. Let $\widetilde{C}^*(c)$ be the optimal cost for a given S up to a certain value for c. We stop increasing c as soon as $\widetilde{C}^*(c) \leq C_{LB}(S, c+1)$. A summary of our procedure is given in Algorithm 1.

Figures 3.3(a) and 3.3(b) provide illustrative numerical examples of both of our lower bound functions. Figure 3.3(a) shows the cost, $C(S,c)$, for $S \in \{0,\dots,21\}$ and $c \in \{0,\dots,\min(S,5)\}$; higher values for c are not displayed for clarity. In this figure, one can observe two things: (i) the optimal cost function, $\hat{C}^*(S)$, is not convex in S which makes optimization difficult, and (ii) the lower bound is rather tight after the minimum has been reached. For another instance, Figure 3.3(b) shows the cost, $C(S,c)$, for $S = 5$ and $S = 14$ with $c \in \{0,\dots,S\}$. Here again one can see that the bound is tight, especially when needed, i.e. for c large. Using our lower bounds we are able to eliminate large parts of our enumeration space. The general performance of these bounds is analyzed in more detail in Section 3.5.4.

Algorithm 1: Find an optimal CL policy(S, c)

```
$S \leftarrow 0$
$\hat{C}^*(S) \leftarrow \infty$
while $\hat{C}^*(S) > \hat{C}_{LB}(S+1)$
do
    $c \leftarrow 0$
    $\widetilde{C}^*(c) \leftarrow \infty$
    while $\widetilde{C}^*(c) > C_{LB}(S, c+1)$
    do
        $C(S,c) \leftarrow EvaluatePolicy(S,c)$
        if $C(S,c) < \widetilde{C}^*(c)$
          then $\widetilde{C}^*(c) \leftarrow C(S,c)$
        $c \leftarrow c+1$
    $S \leftarrow S+1$
```



(a) $\hat{C}_{LB}(S)$ illustration, instance 985

(b) $\widetilde{C}^S_{LB}(c)$ illustration for $S \in \{5, 14\}$, instance 1485

Figure 3.3: Example of both lower bound functions.

3.5. Numerical Experiments

In this section we conduct numerical experiments to gain insight into the performance of our CL policy. Specifically, we seek to answer questions regarding: the performance of the optimal CL policy as compared to the globally optimal policy and more naïve policies, the sensitivity of the optimal CL policy to the assumed lead time distribution, the structure of the globally optimal policy, and the quality of the bounds developed in the previous section. These questions will be answered in Sections 3.5.1 through 3.5.4, respectively. Although numerical results have been obtained by many papers in this line of research, to our knowledge none compares the globally optimal policy, an advanced heuristic (the CL policy) and naïve policies.

For all of the numerical experiments we create a set of 1500 instances, shown in Table 3.1. All instances share some common settings, i.e. $\mu = 1$, $h = 1$ and the level of accuracy in the evaluation of a given policy, i.e. the distance between the upper and the lower bound of the performance measures, $\leq 10^{-6}$. We vary both the magnitude of demand as well as the relative share of each customer class, to provide insight into how the customer base affects the performance of different policies. Furthermore, the cost parameters are changed, both in magnitude and in relation to each other, as this gives insight into the effect of disparate customer valuations.

3.5.1 Comparison of CL to globally optimal and more naïve policies

Our aim in this section is to analyze whether a static CL policy is an effective way to differentiate between customers. To do so, we compare the CL policy to the globally optimal policy (OPT). To find the OPT policy, we formulate a Markov Decision Process (MDP) and solve it using linear programming, as outlined by Puterman (1994). The OPT policy does not assume a single, static, critical level but is allowed to make state-dependent decisions with respect to class 2 demand and backorders. Details on the MDP formulation and

		λ_1	λ_2
low	$\lambda_1 >> \lambda_2$	0.1	0.01
	$\lambda_1 > \lambda_2$	0.1	0.05
	$\lambda_1 = \lambda_2$	0.1	0.1
	$\lambda_1 < \lambda_2$	0.05	0.1
	$\lambda_1 << \lambda_2$	0.01	0.1
middle	$\lambda_1 >> \lambda_2$	1	0.1
	$\lambda_1 > \lambda_2$	1	0.5
	$\lambda_1 = \lambda_2$	1	1
	$\lambda_1 < \lambda_2$	0.5	1
	$\lambda_1 << \lambda_2$	0.1	1
high	$\lambda_1 >> \lambda_2$	5	0.5
	$\lambda_1 > \lambda_2$	5	2.5
	$\lambda_1 = \lambda_2$	5	5
	$\lambda_1 < \lambda_2$	2.5	5
	$\lambda_1 << \lambda_2$	0.5	5

(a) Demand parameters.

p_1	
1	$p_2 \in \{0.01, 0.05, 0.1, 0.5\}$
	$b \in \{0.01, 0.05, 0.1, 0.2, 1\}$
5	$p_2 \in \{0.05, 0.25, 0.5, 2.5\}$
	$b \in \{0.05, 0.25, 0.5, 1, 5\}$
10	$p_2 \in \{0.1, 0.5, 1, 5\}$
	$b \in \{0.1, 0.5, 1, 2, 10\}$
20	$p_2 \in \{0.2, 1, 2, 10\}$
	$b \in \{0.2, 1, 2, 4, 20\}$
50	$p_2 \in \{0.5, 2.5, 5, 25\}$
	$b \in \{0.5, 2.5, 5, 10, 50\}$

(b) Cost parameters.

Table 3.1: Demand and cost parameters for the numerical experiments.

solution can be found in Appendix A.3.

To broaden our comparison, we also consider two, somewhat naïve policies, both of which: *(i)* have been used for comparison against CL-type policies (see, e.g. Deshpande et al. 2003b and Möllering and Thonemann 2008); and *(ii)* are commonly used in practice (see e.g. Dekker et al. 2002, Deshpande et al. 2003b, and Möllering and Thonemann 2008). These policies are:

- First Come First Served (FCFS): All demands are served as long as there is inventory, when inventory equals zero class 1 demand is lost and class 2 demand is backordered. In effect this is a critical level policy with $c = 0$.
- Separate Inventories (SI): Each customer class is served from its own "reserved" inventory.

The CL policy can be seen as a combination of these alternatives, it utilizes inventory pooling while also "reserving" inventory.

We compare policies $i \in \{CL, FCFS, SI\}$ to the OPT policy by comparing their cost. Let C^{*i} be the cost of the optimal policy in class i; we then compare:

$$\frac{C^{*i} - C^{*OPT}}{C^{*OPT}} \cdot 100, \tag{3.15}$$

for all our instances. For the analysis in this *sub*section we drop 268 instances where no demand is satisfied from stock in the CL nor the OPT policy (optimal CL policy has $S = c = 0$ *and* the cost are equal to the cost of the OPT solution), thus worsening the relative performance of our CL policy compared to the OPT policy.

	CL	FCFS	SI
Average % difference from optimal	2.09	7.17	27.21
Std. Dev.	4.04	11.22	15.89
# instances different	681	829	1232
Average % difference from optimal	3.82	10.73	27.21
Std. Dev.	4.82	15.89	12.26

Table 3.2: Performance of the CL, FCFS, SI policies versus the OPT policy.

First, we compare the average performance over the remaining 1232 instances. Table 3.2 lists the average % difference as defined in (3.15). We see that the CL policy is on average only 2.09% from the OPT policy, and achieves the OPT policy cost in 45% of the instances. Compared to OPT, FCFS and SI perform 7.17% and 27.21% worse, respectively. Focussing only on those instances in which difference is nonzero (681 instances for CL), we see that the optimality gap does not increase dramatically. Not only are FCFS and SI significantly outperformed by CL, but they are more unpredictable as well (larger Std. Dev.).

Next we investigate those instances in which the OPT policy outperforms the CL policy to determine why the CL policy falls short. To do so we have found it useful to look at two metrics for each instance: The ratio between class 1 and class 2 demand, $\frac{\lambda_1}{\lambda_2}$, and the cost ratio between old and new backorders, $\frac{\lambda_2 p_2}{b}$. The latter ratio is an indicator of the rate at which cost are incurred when denying a class 2 demand an item versus not clearing a current backorder. Table 3.3 shows that the largest differences between the CL and the OPT policies occur when class 2 makes up a large fraction of demand, i.e. $\frac{\lambda_1}{\lambda_2}$ is small[5],

[5] There are no entries having both $\frac{\lambda_1}{\lambda_2} = 10$ and $100 < \frac{\lambda_2 p_2}{b} \leq 1000$ in our test bed.

highlighting the importance of the size of class 2. The second index is more subtle.

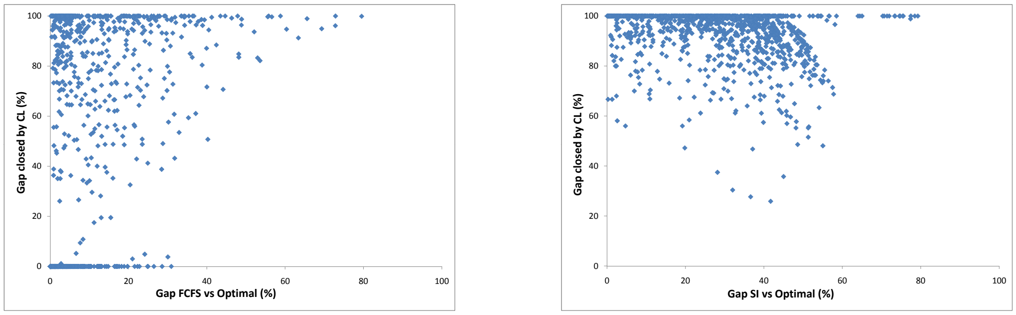
When the rate at which cost for a new backorder is accrued is much higher than the rate at which existing backorders accumulate cost, i.e. $\frac{\lambda_2 p_2}{b} \gg 1$, big differences between the CL and OPT policies arise. This occurs because the OPT policy has the opportunity to differentiate it's decision regarding clearing backorders state-by-state. When $100 < \frac{\lambda_2 p_2}{b} \leq 1000$, deciding to backorder any new incoming class 2 demand would lead to accruing cost at a much higher rate than the current backorder cost rate. Therefore, there are states with positive backorders in which the OPT policy serves new demand, but if a replenishment order comes in the item is added to inventory, to protect against future backorders. The CL policy does not allow for this flexibility as there is only a single critical level.

	$\frac{\lambda_1}{\lambda_2} = 0.1$	$\frac{\lambda_1}{\lambda_2} = 0.5$	$\frac{\lambda_1}{\lambda_2} = 1$	$\frac{\lambda_1}{\lambda_2} = 2$	$\frac{\lambda_1}{\lambda_2} = 10$
$\frac{\lambda_2 p_2}{b} \leq 1$	0.14	0.39	0.25	0.19	0.14
$1 < \frac{\lambda_2 p_2}{b} \leq 10$	5.55	3.89	3.72	2.18	1.09
$10 < \frac{\lambda_2 p_2}{b} \leq 100$	12.79	10.03	8.80	4.77	1.72
$100 < \frac{\lambda_2 p_2}{b} \leq 1000$	18.27	14.45	11.42	6.71	

Table 3.3: Mean % difference. Under what parameter settings does the OPT policy outperform the CL policy?

We performed the same comparison between CL, FCFS and SI. From this we see that CL improves most upon FCFS and SI when demand is balanced between class 1 and class 2, but class 2 is still sizeable (table omitted for brevity). In this case the classes are competing for the same inventory and CL can thus actually make a difference.

Finally, it is interesting to see how much CL improves on FCFS and SI for each problem instance: How much of the distance from the optimal cost is closed by using a CL policy instead of FCFS or SI is shown in Figures 3.4(a) and 3.4(b). We see that a CL policy is *always* able to close some of the gap of SI (due to pooling) and is able to close more of the gap whenever the gap is large. Thus introducing something as simple as a single static critical level into an inventory management system may yield large benefits.



(a) Optimality gap closed by CL from FCFS

(b) Optimality gap closed by CL from SI

Figure 3.4: Comparison of CL, FCFS, SI.

3.5.2 Sensitivity to lead time variability

For analytical tractability we assumed an exponential lead time throughout this chapter, which is in line with many of the other papers in this stream of literature (see Section 3.2). However, it might be the case that the true lead time is more or less variable than exponential. In this subsection we will analyze the robustness of both the CL and the OPT policy to changes in variability in the lead time distribution. Specifically, we will examine how a policy that is found under one lead time assumption (e.g. exponential), performs if the lead times follow a different distribution (e.g. degenerate hyperexponential, H^*). As a measure of variation we use the squared coefficient of variation $\left(C^2 = \frac{var}{mean^2}\right)$. Here we return to analyzing the full set of 1500 instances.

Thus, throughout this section we will at times use different assumptions to find the parameters for a policy than for actually evaluating it (e.g. we will find a policy assuming exponential lead times but evaluate its performance assuming actual lead times are H^*). As a general rule we will use superscripts to indicate the policy and under which lead time assumptions a policy was identified, and parentheses to indicate under which conditions it was evaluated. For example

$$C^{*CL,EXP}(H^*) \tag{3.16}$$

would represent the cost of the optimal CL policy identified under the exponential lead time assumption and evaluated under H^* distributed lead times. When both sets of assumptions are the same, the $(\cdot)$ term is omitted, as is the policy index when this can be done without confusion.

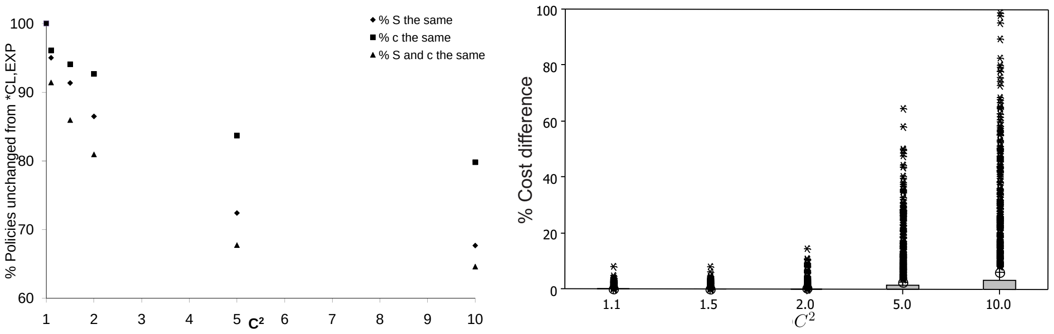
In the subsections to come we will make the following comparisons. In Section 3.5.2 we analytically explore the sensitivity of the CL policy to higher variability lead times, using the H^* distribution. In Sections 3.5.2-3.5.2, we use simulation to extend our comparison to more general variabilities, using a Weibull distribution. In Section 3.5.2 we look at the sensitivity of the optimal CL policy; in Section 3.5.2 we compare the sensitivity of both the CL and the OPT policy.

Sensitivity of CL policy to higher variability lead times

The procedures introduced in Sections 3.3 and 3.4 only require slight modifications to accommodate a degenerate hyperexponential distribution (H^* distribution), which allows for *larger*, or equal, variability than exponential[6]. (The modifications are outlined in Appendix A.4.) We first compare the *parameters* of the optimal CL policy for $C^2 \in \{1.1, 1.5, 2, 5, 10\}$ with those of the optimal CL policy under an exponential lead time ($C^2 = 1$). To do so, let S^{*,H^*} and c^{*,H^*} be the optimal CL parameters found assuming H^* lead times with $C^2 \in \{1.1, 1.5, 2, 5, 10\}$.

In Figure 3.5(a) we plot the % of instances for which $S^{*,EXP} = S^{*,H^2}$, $c^{*,EXP} = c^{*,H^2}$, or both, for varying values of C^2. With only a slight increase in variability ($C^2 = 1.1$) we see that in almost 4% of the instances the optimal c changes and in 5% the optimal S value changes. However, as C^2 increases from 5 to 10, the effect is much less pronounced. Numerically, we find that, the optimal values of S and c may increase or decrease, but

[6] Note that the H^* distribution with $C^2 = 1$ is the exponential distribution.



(a) Sensitivity of S^{*,H^*}, c^{*,H^*} to C^2.

(b) % Cost difference from implementing the optimal CL policy found using exponential lead times, compared to the optimal CL policy at higher C^2.

Figure 3.5: Sensitivity to lead time variability, all instances.

decreases become more numerous as variability grows. From Figure 3.5(a) it is clear that over the range of C^2 values tested the optimal S value is more sensitive to an increase in variability than the optimal c value.

Even though the optimal S and c change, we expect that the cost function is flat around the optimal. To explore this, we analyze how much the cost increases when implementing the optimal CL solution assuming exponential lead times in a more variable, H^*, environment (in which the solution may not remain optimal). Specifically (using notation along the lines of (3.16)), for the CL policy, we calculate

$$\frac{C^{*,EXP}(H^*) - C^{*,H^*}}{C^{*,H^*}} \cdot 100, \tag{3.17}$$

for varying C^2, i.e. $C^2 \in \{1.1, 1.5, 2, 5, 10\}$. Figure 3.5(b) displays a boxplot[7] of (3.17) across our 1500 instances. Clearly, on average, the effect of variability is small. Even when the lead time distribution is 10 times as variable, the mean cost increase is only 5.97%; if 5 times as variable it is 2.69%. For $C^2 \leq 2$ the maximum difference is 14.30% and the mean difference is less than 0.37%. This indicates that, even when an exponential distribution is

[7] See e.g. Montgomery and Runger (1999). The mean is indicated by $\bigoplus$.

wrongly assumed for the lead time, and the true lead time is more variable, the expected increase in cost for a CL policy is small, even up to relatively high variabilities. This suggests that the performance of the CL policy is robust with respect to lead time variability.

Sensitivity of CL policy to higher and lower variability lead times, using simulation

To gain insight into the sensitivity of the CL policy with respect to a wider range of lead time distributions, simulation is used in this subsection. Our procedure is as follows. Let $\mathcal{P}$ be the set of CL policies with cost within 10% of the optimal CL cost under exponential lead times:

$$(S,c) \in \mathcal{P} \Leftrightarrow C(S,c)^{EXP} \leq 1.1C^{*,EXP}.$$

Using simulation, we evaluate the performance of all $(S,c) \in \mathcal{P}$ under lead times that follow a Weibull distribution, i.e. for all policies in $\mathcal{P}$ we calculate $C^{EXP}(Weibull)$ with $C^2 \in \{\frac{1}{32}, \frac{1}{8}, \frac{1}{2}, 1, 2, 5, 10\}$[8]. We use 30 simulation runs with 20,000 customer arrivals each. On average, $|\mathcal{P}| = 4.59$, with at least 1, the optimal, and at most 27 policies, across our 1500 instances.

At this point we need slightly more specific notation, again defined as outlined in (3.16). Let $C^{(S,c)\in\mathcal{P},EXP}(Weibull)$, and $\sigma^{(S,c)\in\mathcal{P},EXP}(Weibull)$ be the mean and standard deviation of cost for the 30 simulation runs of a specific instance. Across all 1500 instances and seven C^2 values, the mean coefficient of variation of the resulting cost:

$$\frac{1}{|(S,c) \in \mathcal{P}|} \sum_{\forall (S,c)\in\mathcal{P}} \frac{\sigma^{(S,c)\in\mathcal{P},EXP}(Weibull)}{C^{(S,c)\in\mathcal{P},EXP}(Weibull)}$$

equals 0.00942 (maximum observation 0.07387, minimum observation: $4.29 \cdot 10^{-6}$). We use

[8] Comparisons have also been made with Weibull ($C^2 \in \{\frac{1}{16}, \frac{1}{4}, 1.1, 1.5\}$), deterministic ($C^2 = 0$), Erlang-k ($C^2 \in \{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$) degenerate hyperexponential ($C^2 \in \{1, 1.1, 1.5, 2, 5, 10\}$), hyperexponential distribution ($H^2, C^2 \in \{1, 1.1, 1.5, 2, 5, 10\}$), and lognormal distribution ($C^2 \in \{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 1.1, 1.5, 2, 5, 10\}$) and results were very similar. For brevity we choose not to include these in this chapter.

these values to construct ±3σ confidence intervals for the cost of all policies. Of particular importance are the upper and lower bound for the confidence intervals for each C^2: $UB(S,c) = C^{(S,c),EXP}(Weibull) + 3\sigma^{(S,c),EXP}(Weibull)$, $LB(S,c) = C^{(S,c),EXP}(Weibull) - 3\sigma^{(S,c),EXP}(Weibull)$, which lead to the following classifications for policies (S,c):

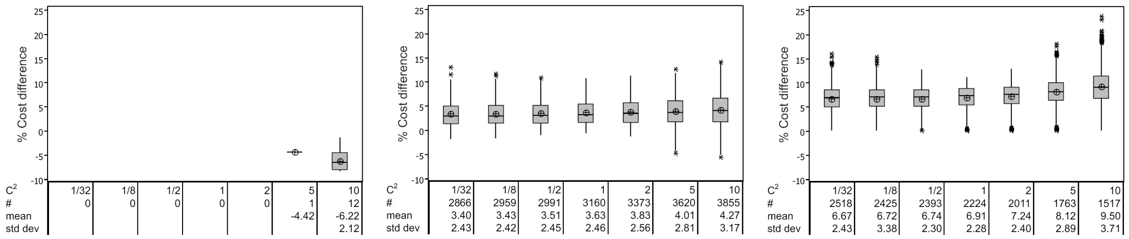
- *Alternative lower cost:* If $UB((S,c) \in \mathcal{P}) < LB((S^*,c^*)$,
- *Alternative higher cost:* If $UB((S^*,c^*)) < LB((S,c) \in \mathcal{P})$,
- *Insignificantly different:* All other cases, i.e. the confidence intervals overlap.

Using these classifications, Figure 3.6 summarizes the % difference between the mean costs:

$$\frac{C^{(S,c)\in\mathcal{P},EXP}(Weibull) - C^{*,EXP}(Weibull)}{C^{*,EXP}(Weibull)} \cdot 100,$$

for our range of C^2 values. This figure reports on *all* policies in $\mathcal{P}$ that are different from the optimal CL policy as determined under exponential lead times. At least 28% of the alternative policies are significantly more expensive (Figure 3.6(c), $C^2 = 10$), whereas at most 0.22% leads to a decrease in cost (Figure 3.6(a), $C^2 = 10$). An immediate observation is that we find *no* significantly better policies when the lead time distribution is less variable than exponential, or slightly more variable (i.e. $C^2 \leq 2$), and even when $C^2 > 2$ very few are significantly better (a maximum of 12 policies when $C^2 = 10$, in Figure 3.6(a)). Overall, as C^2 increases the alternative policies get closer to the cost of the policy that was optimal under the assumption of exponential lead times. For example, for $C^2 = 2$ in 2011 of the 5384 policies (37.35%) the alternative policies have significantly higher cost, compared to only 1517 (28.24%) when $C^2 = 10$.

In summary, we see that the optimal CL policy obtained under the assumption of exponential lead times performs very well under a wide range of lead time distributions (recall, the results for distribution families other than Weibull are very similar). Only when the variability is very high we do find other policies that lead to lower cost (in this case we could potentially solve the model with H^* distribution for lead times). Furthermore,



(a) Alternative lower cost (b) Insignificantly different (c) Alternative higher cost

Figure 3.6: % Cost difference for policies $(S, c) \in \mathcal{P}$ compared with the cost of the optimal policy determined using the assumption of exponential lead times, both evaluated under Weibull lead times.

given that policy evaluations take minutes using simulation versus a split second using our analytical approach, we find no practical reason not to use the exact solution found under exponential lead times.

Sensitivity of the optimal costs

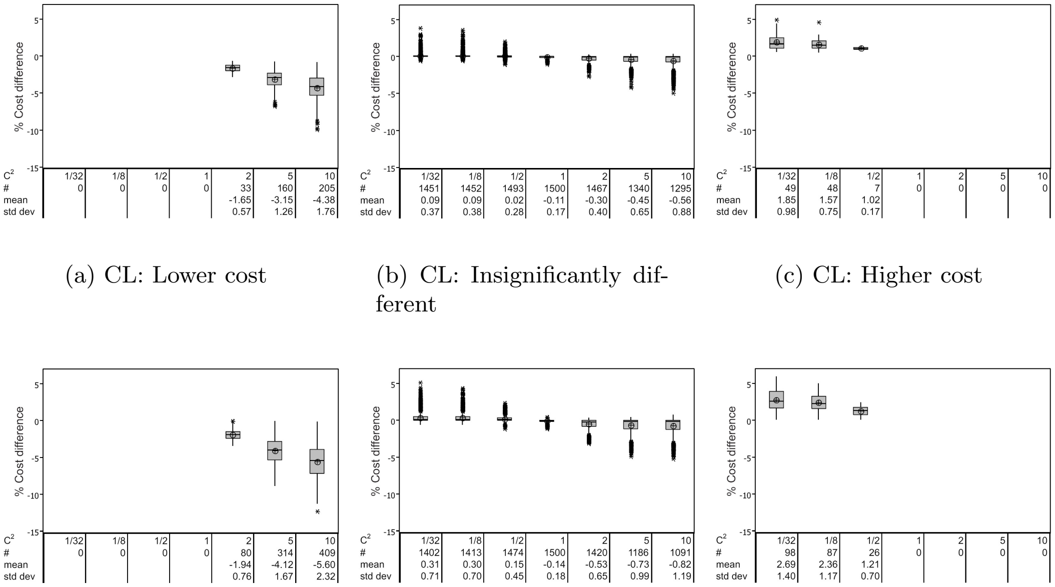
Having established that the optimal CL policy *parameters* are largely insensitive to changes in lead time variability, we now compare the robustness with respect to *costs* of the optimal CL and OPT policies, using simulation. We compare the cost of *only* the optimal policies. Specifically, using the notation as defined in (3.16), for both the CL and the OPT policies we calculate

$$\frac{C^{*,EXP}(Weibull) - C^{*,EXP}}{C^{*,EXP}} \cdot 100$$

over 1500 instances under 7 different variabilities ($C^2 \in \{\frac{1}{32}, \frac{1}{8}, \frac{1}{2}, 1, 2, 5, 10\}$)[9], using 30 simulation runs of 20,000 customer arrivals each. Once again, results for other distributions (i.e. deterministic, Erlang-k, H^*, H^2, and Lognormal) are similar over the same range of C^2 values and are omitted.

In a similar fashion as in Section 3.5.2 we construct $\pm 3\sigma$ confidence intervals to determine whether the % difference of costs is significantly different from 0. The results are summarized

[9] $C^2 \in \{\frac{1}{16}, \frac{1}{4}, 1.1, 1.5\}$ were also evaluated but omitted for brevity.



C^2	1/32	1/8	1/2	1	2	5	10
#	0	0	0	0	33	160	205
mean					-1.65	-3.15	-4.38
std dev					0.57	1.26	1.76

(a) CL: Lower cost

C^2	1/32	1/8	1/2	1	2	5	10
#	1451	1452	1493	1500	1467	1340	1295
mean	0.09	0.09	0.02	-0.11	-0.30	-0.45	-0.56
std dev	0.37	0.38	0.28	0.17	0.40	0.65	0.88

(b) CL: Insignificantly different

C^2	1/32	1/8	1/2	1	2	5	10
#	49	48	7	0	0	0	0
mean	1.85	1.57	1.02				
std dev	0.98	0.75	0.17				

(c) CL: Higher cost

C^2	1/32	1/8	1/2	1	2	5	10
#	0	0	0	0	80	314	409
mean					-1.94	-4.12	-5.60
std dev					0.76	1.67	2.32

(d) Opt: Lower cost

C^2	1/32	1/8	1/2	1	2	5	10
#	1402	1413	1474	1500	1420	1186	1091
mean	0.31	0.30	0.15	-0.14	-0.53	-0.73	-0.82
std dev	0.71	0.70	0.45	0.18	0.65	0.99	1.19

(e) Opt: Insignificantly different

C^2	1/32	1/8	1/2	1	2	5	10
#	98	87	26	0	0	0	0
mean	2.69	2.36	1.21				
std dev	1.40	1.17	0.70				

(f) Opt: Higher cost

Figure 3.7: % Cost difference between the cost of the optimal CL and OPT policies, as determined using the assumption of exponential lead times, evaluated under Weibull lead times.

in Figure 3.7. There are several observations we can make.

First, we observe that the optimal cost of both the CL ad OPT policies is rather insensitive to the variability of the lead time distribution. Even for $C^2 = 10$, more than 72% (CL) and 86% (OPT) of the instances have cost that are insignificantly different. Second, we see that the cost of the OPT policy is *less* robust than the cost of the CL policy in the face of changing lead time distributions. Whereas, for example, under the CL policy, 104 instances have a significantly higher cost when lead times are Weibull distributed (Figure 3.7(c)), for the OPT policy there are 211 instances (Figure 3.7(f)). Note that all of these occur under *lower* variability lead times. Third, the spread in differences is also larger for the globally optimal policy; compare the interquartile ranges and standard deviations in the two rows of Figure 3.7. Finally, when lead time variability increases, the globally optimal policy is again more sensitive, but now to its benefit (Figures 3.7(a) and 3.7(d)): costs may be significantly lower under $C^2 > 1$ as compared to the exponential baseline.

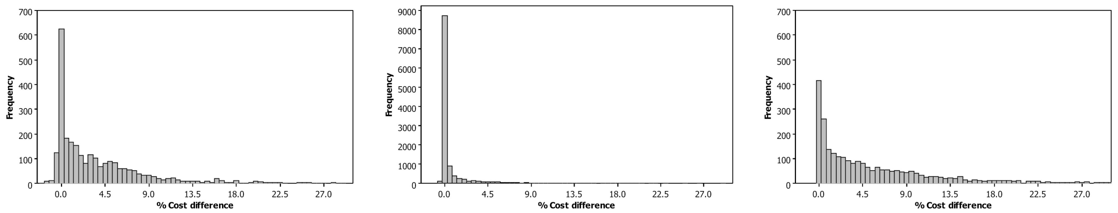
In fact, for both the OPT and CL policies, the effect of variability is such that if the variability increases, the cost tend to *decrease*. The intuition behind this is that for constant mean, as variability increases, the median lead time *decreases*. Occasional long lead times will decrease inventory and/or increase the number of backorders temporarily, but as long as there are items on hand this will not affect the number of demands rejected. In contrast, the increased fraction of shorter lead times increases availability and hence service. This has also been observed in the exact results for the H^* distribution. Note that having ample, or at least multiple, servers is crucial for the existence of this effect.

Finally, we analyze how the cost of the OPT policy and the cost of the optimal CL policy determined under the assumption of exponential lead times change *relative* to each other as variability changes. We simulate 30 runs of $20,000$ arrivals each, with Weibull lead times under varying C^2 ($C^2 \in \{\frac{1}{32}, \frac{1}{8}, \frac{1}{2}, 1, 2, 5, 10\}$). We calculate:

$$\frac{C^{*CL,EXP}(Weibull) - C^{*OPT,EXP}(Weibull)}{C^{*OPT,EXP}(Weibull)}, \tag{3.18}$$

i.e. the % difference between the optimal policies in each class once subjected to other lead time distributions. Recall from Section 3.5.1 that for $C^2 = 1$ the average % difference is 2.09% (for a subset of 1232 instances where the policies are nontrivially different, for all 1500 this is 2.04%).

A paired t-test (with a significance level of 95%) is used to determine how the difference between the CL and OPT policies changed, relative to their difference under exponential lead times. Figure 3.8 displays histograms for (3.18) across our 1500 instances, and seven C^2 values. The three panels distinguish whether the difference decreased (3.8(a)), increased (3.8(c)), or did not significantly change (3.8(b)). The most interesting observation can be made in Figure 3.8(a). This subfigure contains those instances in which the difference



(a) Difference decreased
2572 instances

(b) Insignificantly different
11,450 instances

(c) Difference increased
2478 instances

Figure 3.8: Histogram of % cost difference between the cost of the optimal CL policy as determined under exponential lead times compared to the globally optimal policy as determined under exponential lead times, both evaluated under Weibull lead times.

decreased significantly; there are 207 instances in which the difference is negative. These are cases in which, under Weibull lead times, implementing the optimal CL policy leads to *lower* cost than implementing the OPT policy. Thus, not only is the CL policy more robust to changes in the lead time variability, it may also outperform the globally optimal policy when lead time variability is mis-specified.

Conclusions with respect to sensitivity

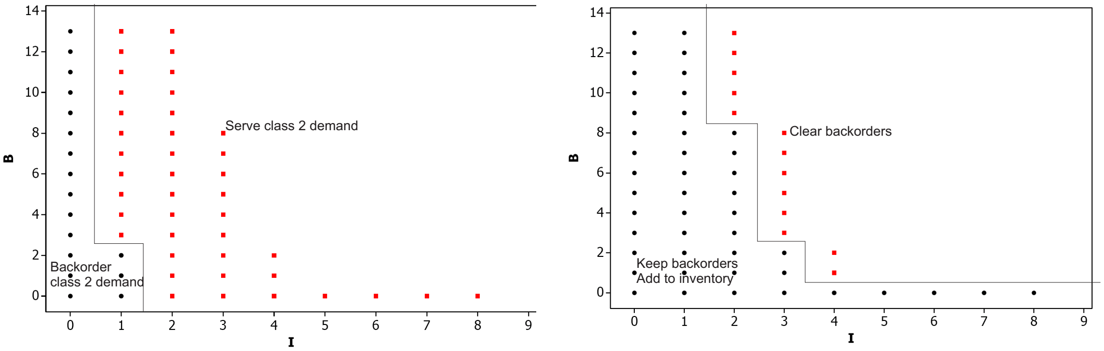
From the above discussion we conclude that the optimal CL policy found under the assumption of exponential lead times is rather robust to changes in the variability of the lead times: When the distribution changes to H^* more than 70% of the optimal CL policies remain unchanged, even when C^2 increases to 10. In addition, for more general lead times, when the lead time distribution has $C^2 \leq 2$ we find, via simulation, *no* CL alternative policies with significantly lower cost in the neighborhood of the optimal CL policy found under exponential lead times. But, when the lead time variability increases, i.e. for $C^2 > 2$, there are a few instances where the CL policy that is optimal under exponential lead times is outperformed by an alternative CL policy.

Even though the cost of the OPT as well as the CL policy appear to be largely insensitive to lead time variability, we find that the OPT policy (found assuming exponential lead times) is more insensitive to changes in the lead time distribution. For example, for insignificantly different policies, when $C^2 = \frac{1}{32}$ the standard deviation of the cost of the OPT policy is 0.71 versus 0.37 for the CL policy. For $C^2 = 10$ these values are 1.19 for OPT versus 0.88 for CL. In fact, when lead times are *not* exponential, the OPT policy is even sometimes *outperformed* by the optimal CL policy found assuming exponential lead times. In the next subsection we explore the structure of the OPT policy to determine why it tends to be less robust than the optimal CL policy.

3.5.3 Structure of the optimal policy

In this section we provide some insight into the structure of the OPT policy by examining a representative instance in detail. Our model closely relates to the model of Van Wijk et al. (2009) in which the optimal lateral transshipment policy is derived. However, key to their analysis is the concept of "proportional allocation" of incoming replenishment orders. In our setting this would require incoming replenishment orders to be allocated to inventory or clearing backorders according to probabilities proportional to the number of backorders and inversely proportional to the inventory level. Under this proportional allocation the structure of the optimal policy can be proven. Note that the decision how to allocate incoming replenishments cannot be optimized in this setting, as that would violate the structural properties of the value function. Hence we resort to numerical analysis of the optimal policy as that still brings to light how improvements over the CL policy can be obtained. The instance we discuss here has the following parameters, $\lambda_1 = 5$, $\lambda_2 = 5$, $p_1 = 1$, $p_2 = 0.5$, and $b = 0.01$. The parameters of the optimal CL policy are $S^* = 11$, and $c^* = 1$. Figure 3.9 displays the OPT policy for this instance.

First, we observe that the maximum level of inventory in the OPT policy is 8, as compared to an order up to level of 11 in the CL policy. Second, the OPT structure does not show a single critical level, it varies depending on the pipeline stock. If we fix an inventory level, say $I = 1$ (on the horizontal axes), we see that for low levels of backorders



(a) Backorder or serve an incoming class 2 demand?

(b) Clear a backorder or add to inventory?

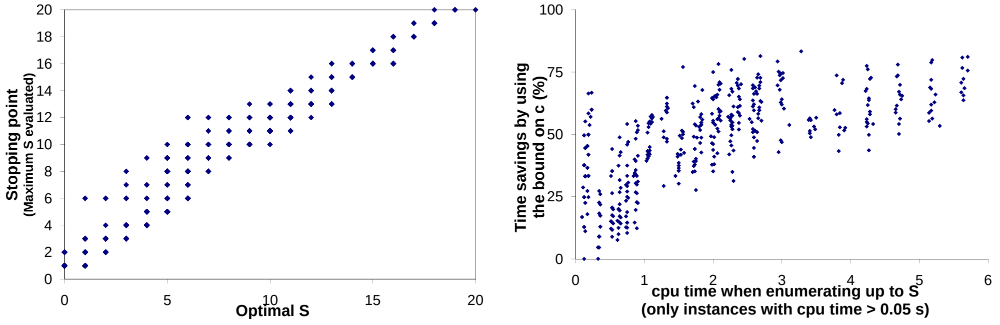
Figure 3.9: Example of globally optimal actions.

($B \leq 2$) arriving class 2 demand gets backordered, even though there is stock on hand (Figure 3.9(a)). Furthermore, if a replenishment arrives, no backorders would get cleared ($B \leq 14$, Figure 3.9(b)). As the number of backorders increases, for fixed I (equivalently as the pipeline stock increases), we observe a threshold above which a class 2 demand would get served ($B \geq 3$, Figure 3.9(a)) or a backorder would get cleared (for some $B > 14$, not displayed for clarity, Figure 3.9(b)). This is because for these high levels of pipeline stock, the policy "expects" another replenishment soon. Also, for certain states, for example at $I = 2$ and $B = 2$, we see that optimally a new class 2 demand is served while a backorder would *not* be cleared. This recalls the observation made in Section 3.5.1: The rate at which cost for new backorders accrues is $p_2\lambda_2 = 2.5$ while a current backorder (if not cleared) accrues costs at rate 0.01. Thus, at low inventory levels, the OPT policy will serve new demands while also stockpiling inventory, leaving extant backorders unsatisfied. This stockpiling offers protection against losing a class 1 demand or needing to backorder a future class 2 demand. The CL policy does not have this flexibility to distinguish between these two actions. Note that one undesirable feature of the optimal policy is that, through this flexibility, it may happen that class 2 customers may end up being served out of order.

Thus, the OPT policy saves on inventory by changing the optimal decisions once inventory becomes low while at the same time keeping the service levels high. Recall however that the OPT policy is less robust with respect to lead time variability. This is because its power relies on conditioning on expected incoming replenishments, which change as C^2 changes.

3.5.4 Computational efficiency

The bounds developed in Section 3.4 allow us to limit our enumeration space. To evaluate the quality of the two bounds we use different measures, as there does not exist a maximum enumeration value for S, as there exists for c (given S).



(a) Effect of bounding S.

(b) Effect of bounding c.

Figure 3.10: Effect of bounds on enumeration.

First the performance of the bound from Corollary 1 is evaluated in Figure 3.10(a). This figure plots the largest evaluated S value versus the optimal S value over all 1500 instances. On average, only 0.933 additional S values were evaluated (standard deviation: 0.926). Thus, the bound on the enumeration over S values is rather effective. The performance of the bound on enumeration over c from Lemma 3 is evaluated in Figure 3.10(b). Here we plot the % of cpu time saved against the cpu time needed when enumerating up to S to find the optimal c. Average savings are 44.8% (standard deviation: 21.1%), which is also a significant gain. From these results we conclude that our bounding procedures are very effective in truncating the enumeration space.

3.6. Extensions

There are several opportunities for extensions of this work. In this section we will present some, and where possible outline how these fit within our model.

Throughout this chapter we assumed an infinite number of replenishment servers and mentioned early on that the case with a single or several parallel servers is a special case. This special case is easier to solve: The generator of the Makov process for finitely many servers has repeating submatrices and could be evaluated either by our procedure, or by straightforward application of Matrix Analytic Methods. Since our monotonicity results require uniformization, these would become easier if there is a maximum replenishment speed. The bounds developed to avoid complete enumeration should be slightly modified but would remain conceptually similar.

A second extension is the incorporation of service level constraints, e.g. a minimum fraction of demand that should be satisfied immediately from stock. Although our evaluation procedure could still be used, the optimization procedure would have to be modified, as we can no longer use the bounds on cost to truncate our search space but would need to truncate based on service levels. This should be possible using our monotonicity results.

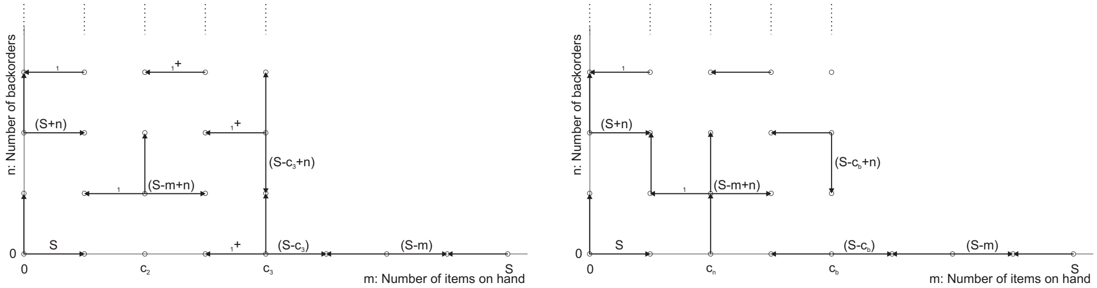
3.6.1 Multiple Customer Classes

Increasing the number of customer classes would also be an interesting extension. However, when the number of classes that gets backordered when not immediately satisfied increases the state space increases geometrically (see e.g. Deshpande et al. 2003b).

If instead the number of classes is increased and only the lowest priority class gets backordered, the model can be analyzed after some modifications. Let customer classes be denoted by $j = 1, \ldots, J$; the priority of the classes decreases in j. Class Js demand is backordered if not immediately satisfied (i.e. it has the lowest priority and the inventory level is at or below c_J). Demand for all classes $j < J$ is lost as soon as inventory is at or below c_j, where $c_1 \triangleq 0$. Demands for each class arrive according to a Poisson process with rate $\lambda_j$and the total demand rate is denoted by $\lambda = \sum_{j=1}^{j=J} \lambda_j$. The transition scheme for this policy is displayed in Figure 3.11(a).

For the evaluation of our new Markov process we can use the procedure as outlined in Section 3.3. Lemma 1 straightforwardly holds, and in the proof of Lemma 2 several negative terms are added to the left hand side of (A.8), but these can then also be omitted when moving to an inequality in (A.9).

When searching for the optimal policy $(S, c_1, \ldots, c_J)$, we can still use the lower bounds from Lemma 3 and Corollary 1 to bound the search space for S and c_J. However, for the critical levels for the lost sales classes other bounds would need to be developed or exhaustive search can be used.



(a) Multiple lost sales classes.

(b) Dual critical levels.

Figure 3.11: Transition diagram for extensions to our policy.

3.6.2 Differentiation between New Demands and Replenishments

Guided by the structure of the optimal policy as displayed in Figure 3.9 we explore a related extension. A key observation is that incoming class 2 demands are served at inventory levels at which backordered class 2 demands are not cleared. To mimic this feature we explore a policy with dual critical levels: If the inventory level is at or below c_n, incoming class 2 demands are backordered; Incoming replenishment orders are used to clear backorders only if the inventory level equals c_b (if inventory is below c_b priority is given to inventory replenishment, and if inventory is above c_b there are no backorders). Note that, given our cost structure, it would never be beneficial to have $c_n > c_b$ as this would "replace" a current backorder with a new one, while incurring the one time penalty cost p_2. The transition diagram is illustrated in Figure 3.11(b).

This modification "prioritizes" new demand over old ones and its transition scheme is displayed in Figure 3.11(b). The structure needed in Lemma 1 is preserved, with slight modifications to the transition matrices. Also, the concept of diagonal layers as used in the proof of Lemma 2 can be applied in this case, allowing the calculation of the performance measures with arbitrary precision.

For the optimization routine to work Theorem 1 should be specialized for the different critical levels. The bounds to limit the search for the optimal base stock level will hold. For each base stock level one would have to search over a space of size $(S+2)(S+1)/2$. Although we expect that bounds on this search space can be developed, this is beyond the scope of this chapter.

Note that, under this modification, class 2 demand may be served out of order. Although this would be feasible in an online setting, in many other business settings this would lead to significant customer dissatisfaction.

3.7. Conclusions

In this chapter we considered a single location, single item, continuous review inventory model with two types of customers. A critical level (CL) was used to differentiate among these customers. Modeling this problem as a Markov process allowed us to exploit special structure to develop an exact and efficient procedure for the evaluation of the performance of a given policy. Leveraging our evaluation procedure, an efficient optimization procedure was developed. This procedure exploits the property - which we proved - that some key performance measures are monotone in the critical level. As far as we know this is the first study to consider a system in which different customer classes behave differently when not immediately served, save for one paper that considers the case with a single replenishment server. Such heterogeneous customer behavior is important; it can be observed in practice.

We compared the performance of the CL policy with the globally optimal (state dependent) policy and two, more naïve, alternatives. On average the CL policy performs near optimal (2.09%) and is able to close a large part of the optimality gap from the more naïve policies, while still being easy to implement. When class 2 (the lower priority, backordered class) is relatively large, and/or has a high penalty cost rate as compared to the cost of holding extant backorders, the CL policy is further away from the optimal policy. However, large benefits over the alternatives are still achieved. Furthermore, we show that our CL policy is largely insensitive to the amount of variability in the lead times: (*i*) The optimal CL policy found assuming exponential lead times remains optimal in many cases as C^2 of the lead time varies between $\frac{1}{32}$ and 10. (*ii*) As the lead time variability changes, the optimal CL policy found under exponential lead times may even outperform the globally optimal policy found under exponential lead times, and in general is more robust to changes in lead time variability than the globally optimal policy.

In summary, using a critical level policy, though sub-optimal, achieves good performance when compared to alternative policies and is furthermore largely insensitive to changes in lead time variability. Our evaluation and optimization procedures allow for efficiently finding the optimal parameters of a CL policy (much faster than solving an MDP), thus claiming much of the improvement of the more complex optimal state dependent policy at a fraction of the operational cost. This should prove crucial as the problem size grows, considering more customer classes, more items, or more locations.

Chapter 4

Ambulance Traffic Coordination[1]

4.1. Introduction

Over the last decade the number of Emergency Departments (EDs) in the US has decreased while the number of patients seeking care at EDs has increased: Pitts et al. (2008) report that, in the United States, the annual number of ED visits increased by 32% to 119.2 million while the number of EDs has decreased by 4.63% to 3,833 between 1996 and 2006. Another troubling ED trend is that at the same time the number of visits is increasing and resources are decreasing, there has been an increase in acuity of ED patients, (i.e. the urgency to be seen by a physician, which ranges from immediate –within 15 minutes– to nonurgent –up to 24 hours), which affects the distribution of patient arrivals to the ED. Lambe et al. (2002) report an increase of 59% in the number of critical patients and an increase of 36% in the number of urgent patients in California between 1990 and 1999. Patients of higher acuity have longer care time (Pitts et al. 2008), laying claim on the scarce resources for longer amounts of time. Both these trends contribute to *crowding* at the EDs.

The terms crowding and overcrowding are generally used interchangeably. We use the term crowding as defined by the American College of Emergency Physicians (2006): "Crowd-

[1] This chapter is joint work with Masha Shunko, Soo-Haeng Cho, and Alan Scheller-Wolf.

ing occurs when the identified need for emergency services exceeds available resources for patient care in the emergency department, hospital, or both." The fact that decreasing supply and increased demand as described earlier leads to actual crowded situations is documented throughout the literature; for example Andrulis et al. (1991) and Richards et al. (2000) report that 10 to 30 percent of US hospitals report daily crowding. Schneider et al. (2003) conducted a survey to measure the extent of crowding and found that the number of ED patients per treatment space in the U.S. on Monday, March 12, 2001 at 7PM was 1.1 on average and 52% of EDs reported having more than one patient per treatment space. Crowding can have dramatic consequences: Crowded EDs lead to increased patient mortality (Cameron 2006) and a decreased level of care (Pines and Hollander 2008). In addition, Blomkalns and Gibler (2004) report that emergency physicians and nurses often work in an environment in which patients are on stretchers in hallways, decreasing satisfaction for both health care providers and patients.

In an attempt to combat crowding, incoming ambulances are often redirected to nearby EDs, a practice known as *ambulance diversion*. Approximately 500,000 ambulances are diverted annually in the United States (Burt et al. 2006). Although ambulance diversion can reduce waiting room crowding, it can also contribute to prehospital deaths, and may also increase the total out-of-service time (Carter and Grierson 2007), i.e. the total time it takes an ambulance from the alarm (e.g. 911 call) to the moment it is back in service, ready to respond to another call. However, the literature suggests that the significance of this effect depends on local conditions: Neely et al. (1994) found that for diverted patients, transportation times were 5.0 to 11.6 minutes longer and transportation distances were 1.3 to 4.6 miles further than for nondiverted patients in an urban area with a population of 600,000. In Toronto, increased transportation times were observed for patients with chest pain in crowded situations (Schull et al. 2002a). However, Carter and Grierson (2007) found no significant effect on any of the factors constituting total out-of-service times in Winnipeg (MB, Canada).

In addition to potentially increasing transportation times due to diversion, crowded EDs may reduce the availability of Emergency Medical Services (EMS) in a community in other ways, as EMS crews may find themselves delayed in EDs for extended periods of time, unable to transfer their patient from the ambulance to the care of the ED (Eckstein et al. 2005). In Los Angeles (CA) EMS crews find themselves waiting for an ED stretcher in 13% of their transports to EDs (10% of these situations exceed 1 hour) as reported by Eckstein and Chan (2004). Thus, a diverted ambulance might be subject to a longer transportation time but a shorter turnaround time at the hospital. This illustrates that diversion may be particularly beneficial in areas where the negative effect of diversion (longer travel time) is minimized. This could be either due to several EDs being located in a small geographic area, or there being abundant ambulance capacity. Some hospitals have a policy to never divert (e.g. UPMC in Pittsburgh, Guyette 2009). Although this can be perceived as a service to the community, there may be cases in which diversion is in the best interest of patients (Williams 2006, Cheung et al. 2006).

In the US EDs do not just care for critically ill patients but also function as a safety net of the health care system (Hoot and Aronsky 2008): By the Emergency Medical Treatment and Active Labor Act (EMTALA), an unfunded congressional mandate, EDs are required to care for patients with no other source of primary care (Moskop et al. 2009). This can also contribute to crowding. However, problems with crowded EDs have also been reported in Spain, Australia, Canada, and the United Kingdom (Bradley 2005), countries that have some form of public/universal health coverage and hence that would not expect to see patients at their EDs for low acuity medical treatment. Hence, crowding and diversion can be seen as problems occurring internationally, and not solely due to idiosyncrasies of the US healthcare system.

Thus, it is not surprising that researchers have considered this problem. At a high level, Asplin et al. (2003) develop a conceptual model that identifies three areas from which the problem of crowded EDs can be approached: input, i.e. managing the demand for emergency services; throughput, i.e. measures within the ED to provide expedient service; and output, i.e. ensuring that patients leave the ED as soon as possible. Shah et al. (2006) identify three solution themes: increased resources, demand management, and operations research.

In this chapter we focus on potential solutions on the input side of the problem from a demand management perspective. Specifically, we seek to inform decisions regarding ambulance destination control. We incorporate specific characteristics of the situation in Pittsburgh (PA) and focus specifically on a 2-mile circle in the center of Pittsburgh that hosts 7 EDs operated by two hospital systems (see Figure 4.1) that service patients from all over Allegheny county (and beyond). Of these 7 EDs, 2 are operated by the West Penn Allegheny Health System (WPAHS), and 5 are operated by the University of Pittsburgh Medical Center (UPMC). This specific setting significantly reduces one of the negative effects of diversion: Diversion does not require significantly longer distances as all EDs are located in a small geographic area. In addition, the City of Pittsburgh EMS is responsible for prehospital emergency care throughout the City of Pittsburgh. Having a centrally coordinated EMS service could allow for coordination of ambulance traffic, similar to the instances described by Barthell et al. (2003) in Milwaukee (WI) and Lagoe et al. (2003) in Syracuse (NY). But our setting does feature other complicating factors: Differences in ED size, hospital capacity and the fact that the hospital systems compete for market share.

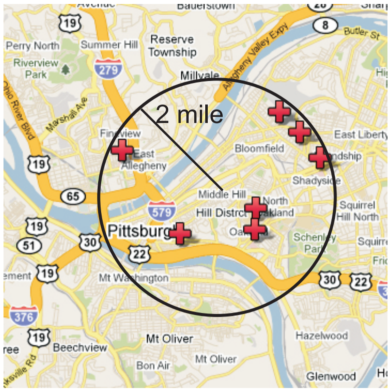


Figure 4.1: Concentration of Pittsburgh hospitals

In this chapter we explore two methodologies to evaluate methods of dealing with crowded EDs. We first develop a queueing model to study decision-making strategies in a setting with two hospitals that use ambulance diversion as a means to avoid or reduce crowding at their EDs. This policy is similar to the critical level model introduced in Chapter 3: Using a diversion level, a hospital attempts to redirect incoming ambulance traffic to another hospital when a certain criterion is met (i.e. there are too many patients in the ED waiting to be transferred to the Inpatient Department, ID). Unfortunately we run into similar problems as outlined in Section 3.6.1, this Markov chain is multi-dimensional and lacks sufficient structure to allow for obtaining a solution without truncation. However, we are still able to illustrate some trade-offs numerically. In this model we maximize the profit for individual hospitals using hypothetical parameter values.

To better capture many of the complexities that are present in the real world we then develop a simulation model. In this simulation model we are able to incorporate some features specific to the situation in Pittsburgh (PA). Specifically, seven EDs are located near the center of the city, within a circle with a 2-mile radius, which allows for diversion or coordination of ambulance traffic with minimal increases in transportation times. In the simulation model we use real data from various sources, enabling us to explore different coordination mechanisms that can be used to disseminate information about hospital (ED and ID) status to EMS crews. This information can then be used to help decide to which hospital to take a patient. In the simulation model we maximize either the quality of care delivered or the revenues generated by hospitals.

We find that the commonly used coordination mechanism (ambulance diversion) performs reasonably well. However, we suspect, and have supporting anecdotal evidence (Hostler 2010), that when hospitals set diversion levels too high (from the perspective of EMS crews) the EMS crew might use their own experience as a guideline in determining to which hospital they should go. The mechanism that we use to capture the EMS crews' use of experience (an Exponentially Weighted Moving Average, EWMA, of observed queue lengths) can perform significantly worse, depending on how quickly the EMS crews update their beliefs. To counter this, we identify two coordination mechanisms that can significantly increase the average level of care provided without being detrimental to hospital revenues. One of these coordination mechanisms (Join the Shortest Queue, JSQ) is grounded in queueing theory, and is known to perform well in a variety of settings. The second coordination mechanism is grounded in the emergency medicine practice field and aims to predict crowding in EDs based on basic information readily available within an ED (Weiss et al. 2004). This uses measures of instantaneous utilization of the ED and ID, the number of patients in the ED requiring ventilators and measures of patient throughput.

The concept of resource pooling, which is what ambulance coordination comes down to, has been studied in different settings. In the spare-parts literature this concept is studied as lateral transshipments (e.g. Wong et al. 2006) and it also appears in the call center literature (e.g. Van Dijk and Van der Sluis 2004). A key differentiator of our paper is that resources are shared between competing entities and that the we are primarily interested in the effect it has on the level of care provided to a homogeneous customer group, i.e. patients/customers do not in principle belong to a specific hospital. In addition we study the effect of different pooling strategies which *(i)* are used in practice, *(ii)* appeal to practitioners, or *(iii)* are known to perform well from research done in related fields. We show that it is important *how* resources are shared by demonstrating that some intuitively appealing coordination mechanisms clearly perform worse than others.

The remainder of this chapter will proceed as follows. In Section 4.2 we will review the literature related to ED crowding. Section 4.3 will introduce a queueing model that can be used to optimize diversion decisions for a profit-maximizing hospital. Section 4.4 introduces a simulation model that captures significantly more real life features than the queueing model, specifically detailing the situation in Pittsburgh. Using this simulation model we analyze several coordination mechanisms for ambulance traffic. The results of these two models will be discussed in Section 4.5, after which we conclude in Section 4.6.

4.2. Literature

The issue of ED crowding and ambulance diversion first received national attention with sporadic reports of crowding in the late 1980s (Olshaker and Rathlev 2006). Over the years, a multitude of causes and potential solutions have been analyzed. Here we give a brief overview of the existing research in the field.

Early analysis by the General Accounting Office (GAO 1993) concluded that crowding was concentrated at urban safety net hospitals and that one of the main causes was ED visits for non-urgent conditions. Over the last 20 years a multitude of additional explanations for crowded EDs have been proposed (Krochmal and Riley 1994, Derlet and Richards 2000, Schull et al. 2002b, Schafermeyer and Asplin 2003, Schull et al. 2003, Reid 2005, Olshaker and Rathlev 2006, Patel et al. 2006, Allon et al. 2009, Moskop et al. 2009, Olshaker 2009). The main hypotheses are: *(i)* Caring for admitted patients in the ED (boarding); *(ii)* Increasing acuity; *(iii)* Increasing demands because of the uninsured; *(iv)* Inadequate space in EDs (downsizing of hospital capacity); *(v)* Nursing shortages; *(vi)* Delays in arrival of specialists; *(vii)* Aging population; *(viii)* Limited availability of off-hour primary care physicians; *(ix)* And more technology for which one gets referred to the ED. However, more recently the inability to transfer ED patients to inpatient beds (the first of the themes listed above) has been identified as the most likely root cause of ED crowding (Schull et al. 2003, Olshaker and Rathlev 2006, Moskop et al. 2009). If ED patients cannot be transferred to inpatient beds, they stay in the ED, and are classified as *boarding patients.* According to Schafermeyer and Asplin (2003) two themes have emerged:

1. “Emergency department crowding is not a problem that can be fixed in the ED itself; rather, it is the result of a complex series of policy and market forces that have created a mismatch between the supply of and demand for emergency services”

2. “Despite the complexity of the problem, it is too serious to ignore and unlikely to go away without thoughtful interventions to alleviate the problem.”

Several solution approaches have been proposed in the literature. Asplin et al. (2003) and Shah et al. (2006) introduce complementary frameworks to classify solutions. Asplin et al. (2003) introduce a conceptual model that distinguishes between the focus on input, throughput, or output of the ED system. Conversely, Shah et al. (2006) classify solution themes by increased resources, demand management, or operations research. We focus our literature review on those papers that, like us, focus on the intersection of input and demand management.

Barthell et al. (2003) and Lagoe et al. (2003) report on the implementation of an Internet tool in Milwaukee (WI) and Syracuse (NY) respectively, that disseminates information about hospital and EMS statuses to all entities in the local health care system on a daily basis. In case there are significant changes, status updates can be provided more regularly. Both implementations report significant decreases in the number of diversion hours. Vilke et al. (2004a) discuss an experiment in which, in a two hospital system, one hospital is allocated additional resources to avoid diversion completely for a week. They find that the reduction of diversion hours at one hospital reduces the number of required diversion hours at the other hospital to zero as well. This indicates that there is a strong interdependence effect. Vilke et al. (2004b) and Patel et al. (2006) describe formal coordination mechanisms and diversion guidelines that have helped to significantly reduce ambulance diversion in San Diego (CA) and Sacramento (CA). Shah et al. (2006) describe a trial, in Rochester (NY), in which EMS crews would contact an EMS destination-control physician for patients requesting transport to either of the two hospitals in the study. The physician would then direct the ambulance based on patient and system characteristics. Reductions of diversion hours at the two hospitals, 41% at a university hospital and 61% at a community hospital, were reported.

It appears that the most tightly integrated and coordinated mechanisms for patient routing have been developed outside the US. Sprivulis and Gerrard (2005) and Larson (2008) describe coordination efforts in Western Australia and Edmonton (AB, Canada) respectively. Both studies report on the use of a real-time Internet-accessible information system that provides information on ED status (occupied spaces, emergency inpatients, waiting room patients, etc.) that helps EMS crews make more informed decisions. Both these studies report significant decreases in diversion hours and a more balanced workload between different hospitals. The fact that these studies have been carried out outside the US could potentially be explained by the structure of the different countries' respective health care systems. Both Australia and Canada have systems with universal health care coverage, which ensures that the patients can be treated everywhere at the same cost to the patient. This also ensures that EDs will see fewer nonurgent visits; moreover the universal coverage also removes some of the tension that exists between privately-operated hospitals, which complicates coordination in the US system. In this chapter we generate insights on the potential benefits of coordination between privately operated US hospitals.

For a very extensive review of crowding in EDs and the surrounding policy domain we refer the reader to an excellent review by Bradley (2005).

4.3. Queueing Model

Ambulance diversion is a common approach to reducing crowding in EDs; in particular, hospitals start diverting ambulances when the number of boarding patients exceeds a prespecified threshold (Allon et al. 2009), we call this threshold a *diversion level.* In this section we introduce a queueing model that allows us to generate insights into how diversion levels should be set. We make several simplifying assumptions to keep our model tractable. These assumptions will be relaxed in Section 4.4, where we more realistically model the situation as it exists in Pittsburgh. The results obtained from both models will be presented in Section 4.5. A schematic overview of the patients and how they flow through a hospital can be found in Figure 4.2.

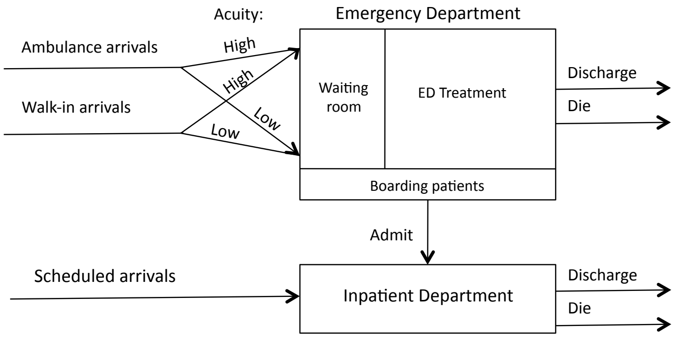


Figure 4.2: Schematic flow of patients through a hospital (ED & ID).

In this section we consider 2 hospitals, indexed by j (a table of notation is provided in Appendix B.1). At each hospital patients arrive to the ED through 2 channels: *(i)* Ambulance arrivals, and *(ii)* Walk-in arrivals. Upon arrival, each patient is triaged and classified as "high" acuity or "low" acuity. Once triaged, a patient's progress through the ED becomes independent of his mode of arrival and is governed by his acuity. In addition, we consider scheduled arrivals into the Inpatient Department (ID) of the hospital.

Each hospital has a limited number of beds, which we consider as the primary capacity constraint. We consider three types of beds: *(i)* Regular ED beds, B_{jED}, *(ii)* Surge ED bed capacity, B_{jSED}, which is additional, and more expensive, capacity that can be used during periods of peak demand, and *(iii)* Beds in the ID, B_{jI}, to be used for patients that have transferred from the ED to the ID as well as for scheduled arrivals to the ID.

The state space of hospital j is described by the number of high acuity patients at the ED (n_{jH}), the number of low acuity patients at the ED (n_{jL}), and the number of inpatients (n_{jI}), either in the ID or boarding in the ED. We seek to determine the "best" diversion level for each of the hospitals, and analyze how these policies are interdependent between the two hospitals. We now describe the patient arrival process, the treatment/departure process and the cost components that we consider in our model, before constructing and solving the Markov chain.

4.3.1 Arrivals

Hospital j observes ambulance arrivals at rate λ_{jA} and walk-in arrivals at rate λ_{jW}. Upon arrival these patients are triaged, leading to effective arrival rates of $\lambda_{jAH} := \lambda_{jA}P_{jAH}$, and $\lambda_{jAL} := \lambda_{jW}P_{jWH}$, for high and low acuity ambulance patients at hospital j; and λ_{jWH}, λ_{jWL} for high and low acuity walk-in patients at hospital j. Here P_{jAH} (P_{jWH}) are the probabilities that an ambulance (walk-in) patient is triaged as high acuity at hospital j. Scheduled patients arrive with rate λ_{jI}. All arrivals are assumed to have exponentially distributed interarrival times and are independent of other arrivals. After probabilistic splitting (at the point of assigning an acuity) the interarrival times are still exponentially distributed. We then merge the arrival streams by acuity to obtain two arrival streams for high and low acuity patients with rates $\lambda_{jH,}$ and $\lambda_{jL.}$ When a hospital is on diversion, it attempts to redirect ambulance arrivals to the other hospital; walk-in arrivals are unaffected. In the situation that we consider here there are three options for the system: *(i)* Neither of the hospitals is on diversion: Each sees their own arrivals; *(ii)* One hospital is on diversion: The hospital on diversion observes a smaller arrival stream as ambulances get rerouted to the other hospital (specifically, a fraction p_{jdiv} of ambulance traffic that would prefer to go to the hospital on diversion will go to the other hospital); *(iii)* Both hospitals are on diversion: Each sees its own arrivals. Based on the diversion status the actual arrival rate observed by the hospital differs from $\lambda_{jH,}$ and $\lambda_{jL,}$ and will be defined as we set up the Markov chain in Section 4.3.4.

4.3.2 Departures

Patients that are in the hospital depart after receiving a certain (random) amount of treatment. The treatment rates for high acuity, low acuity, and scheduled arrivals are exponentially distributed and respectively have rates: μ_{jH}, μ_{jL}, μ_{jI}. Once treatment in the ED is complete, a high acuity patient is released/discharged from the hospital with probability p_{jHR} (low acuity patient: p_{jLR}). A fraction p_{jHA} of high acuity patients are admitted to the ID (p_{jLA} for low acuity). Finally, there is a risk that a patient dies while undergoing treatment, this happens with probability p_{jHD} or p_{jLD} for high and low acuity patients, respectively. In addition, due to the severity of their injury, high acuity patients that are waiting (and are not treated at that time) die with rate μ_{jHD}. All treatment rates are independent of each other.

During treatment we assume that the following service policy applies. Patients that originally arrived to the ID (scheduled arrivals) are treated at the Inpatient Department (i.e. only occupy beds in the ID). Any patient that should be transferred from the ED to the ID is transferred whenever this is possible. If the ID is full, these patients stay in the ED (and occupy a bed there, these are boarding patients). As soon as a bed becomes available in the ID, a boarding patient is transferred. As interarrival and treatment times are assumed to be exponentially distributed, it is immaterial which patient is transferred, but in reality it would probably be the patient that has been boarding the longest. This is also reflected in the calculation of the National Emergency Department Overcrowding Score (NEDOCS, `http://www.nedocs.org`). Beds in the ED that are not occupied by boarding patients are used to treat ED patients. High acuity patients get *pre-emptive resume* priority over low acuity patients: We assume that the treatment of a low acuity patient that was interrupted resumes treatment from where it was left off.

4.3.3 Cost Components

In this section we consider hospitals that are primarily self-interested; we therefore evaluate the performance of the diversion policy from a profit perspective. In order to do this, we assign cost and revenue rates to each state of our system. We are aware that some of these factors cannot easily be expressed in monetary units; therefore we evaluate a range of parameter values.

The first category of cost factors are those directly related to the use of our main unit of capacity: Beds. Cost is accrued at a rate of c_{jnE} when a normal ED bed is occupied. Using surge capacity in the ED is more expensive and accrues cost at a rate of c_{jnS}. Using a bed in the ID costs c_{jnB} per period. Furthermore, we consider three other categories of cost. We assign a cost of c_{jBB} for each scheduled patient at the ID that needs to be rescheduled when the ID is full (the rescheduling is considered out of the scope of our research, hence we assume that from our perspective this patient is "lost"). We incur a cost of c_{jDW} for patients that pass away while waiting and c_{jDT} for patients that pass away while being treated. The waiting cost is c_{jW} per period.

Revenues are generated dependent on the status of the patient: A high acuity patient in the ED generates a revenue of r_{jnH} per period and a low acuity patient generates r_{jnL} per period. Patients in the ID generate revenue at a rate of r_{jnB}. Here we need to make an important distinction: Patients that have been officially discharged from the ED for admittance to the ID but have not yet been transferred to the ID are boarding in the ED. These boarding patients generate revenue at the rate r_{jnB} per period as the services that they are provided qualify as being provided in an inpatient bed, even though a (more expensive) ED bed (or ED surge bed) is being occupied.

4.3.4 Setup of the Markov chain

As we assume that interarrival times and treatment times are exponentially distributed, we can now formulate our problem as a Markov chain. For tractability we limit ourselves to the case of two hospitals, i.e. $j \in \{1, 2\}$. This allows us to calculate the steady state probability distribution, which in turn allows us to calculate the cost of a given policy. We first describe the specific state-dependent arrival and departure processes before describing how we solve the Markov chain.

Arrivals

We assume hospitals use a diversion level policy (Allon et al. 2009). Under diversion, a patient that originates as an arrival to one hospital may end up at another hospital. We assume that only ambulance patients can be diverted. Whether an ambulance patient is diverted depends on the state of the system. In our queueing model, hospitals use use a binary diversion indicator, $D_j \in \{0,1\}$. Let the number of boarding patients in the ED of hospital j be given by $n_{jB} := \max(0, n_{jI} - B_{jI})$. We use $\hat{n}_j$ to represent the threshold on the number of boarding patients in the ED of hospital j, such that above this threshold, the hospital goes on diversion. We let p_{ij} be the fraction of patients originally going to hospital i that go to hospital j instead. Using diversion level $\hat{n}_j$ has the following effect on the system:

Table 4.1: Effect of diversion.

	$n_{1B} < \hat{n}_1$	$n_{1B} \geq \hat{n}_1$
$n_{2B} < \hat{n}_2$	$D_1 = D_2 = 0$ $p_{12} = p_{21} = 0$	$D_1 = 1,\ D_2 = 0$ $p_{12} = p_{jdiv},\ p_{21} = 0$
$n_{2B} \geq \hat{n}_2$	$D_1 = 0,\ D_2 = 1$ $p_{12} = 0,\ p_{21} = p_{jdiv}$	$D_1 = D_2 = 0$ $p_{12} = p_{21} = 0$

This diversion model can be seen as a variant of the Join the Shortest Queue (JSQ) queueing policy. In JSQ, jobs (patients) will choose to go to the server (hospital) with the shortest queue in an attempt to minimize their waiting time. The variant of JSQ that most resembles our setting would be Threshold Jockeying (TJ), in which case jobs (patients) will only transfer to the shorter queue if it is a least a certain amount (threshold) shorter. For an overview of several of the variants of JSQ policies, see Van Houtum et al. (1998). In our setting, the decision is not based on queue length but on the number of boarding patients. However, as boarding patients are generally seen to be the main cause of ED crowding (Schull et al. 2003, Olshaker and Rathlev 2006, Moskop et al. 2009) they can be seen as a proxy for queue length.

The effect of diversion is described in Table 4.1. This directly affects the arrival rates in our model. Hence, we need a translation from base arrival rates (original demands for care) to actual (or net) arrival rates. Let Λ_j^H, Λ_j^L be the net arrival rates of high and low acuity patients into hospital j. This is a function of the congestion level of hospital j as follows:

$$\Lambda_j^H = \lambda_{jA}(1-p_{ji})P_{jAH} + \lambda_{iA}p_{ij}P_{iAH} + \lambda_{jW}P_{jWH}$$
$$\Lambda_j^L = \lambda_{jA}(1-p_{ji})P_{jAL} + \lambda_{iA}p_{ij}P_{iAL} + \lambda_{jW}P_{jWL}$$

Departures

The other state transition is that of departing patients. To determine the effective departure rates we need to know the number of patients in the ED that are receiving emergency treatment, which depends on the effective capacity of the ED, which may be reduced by the number of boarding patients. We use the number of ED beds as our main unit of capacity. This is in line with the literature (Lambe et al. 2002, Schneider et al. 2003, e.g.) and our interviews with EMS professionals. Let $\overrightarrow{n}_{jI}$, $\overrightarrow{n}_{jH}$, and $\overrightarrow{n}_{jL}$ be the number of ID, high acuity ED, and low acuity ED patients that are currently receiving treatment at hospital j:

$$\overrightarrow{n}_{jI} = \min(n_{jI}, B_{jID}),$$
$$\overrightarrow{n}_{jH} = \min(n_{jH}, B_{jED} + B_{jSED} - n_{jB}),$$
$$\overrightarrow{n}_{jL} = \min(n_{jL}, B_{jED} + B_{jSED} - n_{jB} - \overrightarrow{n}_{jH})$$

i.e. the number of high acuity patients that can receive treatment is constrained by the capacity of the ED and by the number of boarding patients that arrived to the ED earlier but have not transferred to the ID yet, though they have completed ED treatment. The number of low acuity patients receiving treatment at any point in time is constrained not just by the ED capacity and boarding patients, but also by the number of high acuity patients receiving treatment. Now the actual treatment rates are given by $\overrightarrow{n}_{jH}\mu_{jH}$, $\overrightarrow{n}_{jL}\mu_{jL}$, and $\overrightarrow{n}_{jI}\mu_{jI}$. Furthermore, $n_{jH} - \overrightarrow{n}_{jH}$ high acuity patients are waiting for treatment and die at rate $(n_{jH} - \overrightarrow{n}_{jH})\mu_{jHD}$.

Solving the Markov chain

In order to determine the steady state performance (in terms of costs and revenues) of our system we now need to determine its steady state distribution. Let π_S be the steady state probability of being in state $S := (n_{1H}, n_{1L}, n_{1I}, n_{2H}, n_{2L}, n_{2I})$. Note that $0 \leq n_{jH}$, $0 \leq n_{jL}$, and $0 \leq n_{jI} \leq B_{jI} + B_{jED} + B_{jSED}$. This leaves us with 4 queues of potentially infinite length. A common approach, introduced in Chapter 2, to solve a Markov chain that is infinite in one dimension is Matrix Analytic Methods (MAM, see e.g. Neuts 1981). However, MAM becomes very difficult to apply when there is more than a single infinite dimension. In a typical JSQ setting one can reduce the number of infinite length queues by looking at the difference between queue lengths. However, in our setting, with 4 infinite length queues and priorities between high and low acuity patients in each hospital, there is no apparent way in which we can reduce the Markov chain to being infinite in fewer dimensions.

Therefore, we truncate the state space by imposing upper bounds on the sum of n_{jH} and n_{jL}, reducing their domains such that $0 \leq n_{jH} + n_{jL} \leq \bar{n}_j$. Transitions beyond these upper bounds are cut-off, which is reasonable if $\bar{n}_j$ is sufficiently large. In selecting parameter values we ensure that the probability of being in boundary states is small, which can be done in a similar fashion as described in Appendix A.3. Having obtained a finite dimensional Markov chain we can write out the balance equations for each state and solve the resulting system of approximately $(\bar{n}_1(B_{1I} + B_{1ED} + B_{1SED})\bar{n}_2(B_{2I} + B_{2ED} + B_{2SED}))$ equations. The solution is given in steady state probabilities; the total profit M_j for hospital j can now be calculated by adding up the long run average profits of all states (steady state probability of a state multiplied with the per period cost of the state).

In Section 4.5 we use this Markov chain approach to obtain insights on how diversion levels should be set in a two hospital setting. Realizing that this Markov chain is not able to capture many of the intricate details of real ED operations we now introduce a simulation model that allows us to analyze a more realistic setting.

4.4. Simulation Model

In this section we detail a simulation model with the goal of analyzing a realistic setting, the results of which are discussed in Section 4.5.2. For reasons described at the end of Section 4.3 we focus on maximizing the quality of care (as will be operationalized on Page 104) or the hospital revenue in our simulation model.

In the simulation model we approximate the situation as it exists in Pittsburgh: We model the seven hospitals within city limits that operate EDs, assuming that all patients that arrive by ambulance are transported from a location within Allegheny county. Pittsburgh is located approximately within the center of the county, as can be seen in Figure 4.3.

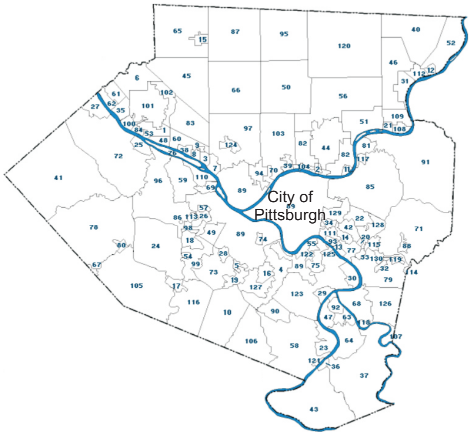


Figure 4.3: Allegheny County (2010)

We use the Arena software package to build and run our simulation. Figure 4.4 displays a high-level process flow chart of our simulation model; this is a simplified representation of the actual Arena simulation model. The simulation model consists of five main modules: Demand generation, Transportation, Coordination, ED care, and ID care. Throughout this section we discuss details of each of the modules, as the patient flows through the process. After discussing the patient flow we discuss the hospital revenues generated throughout the process in Section 4.4.6.

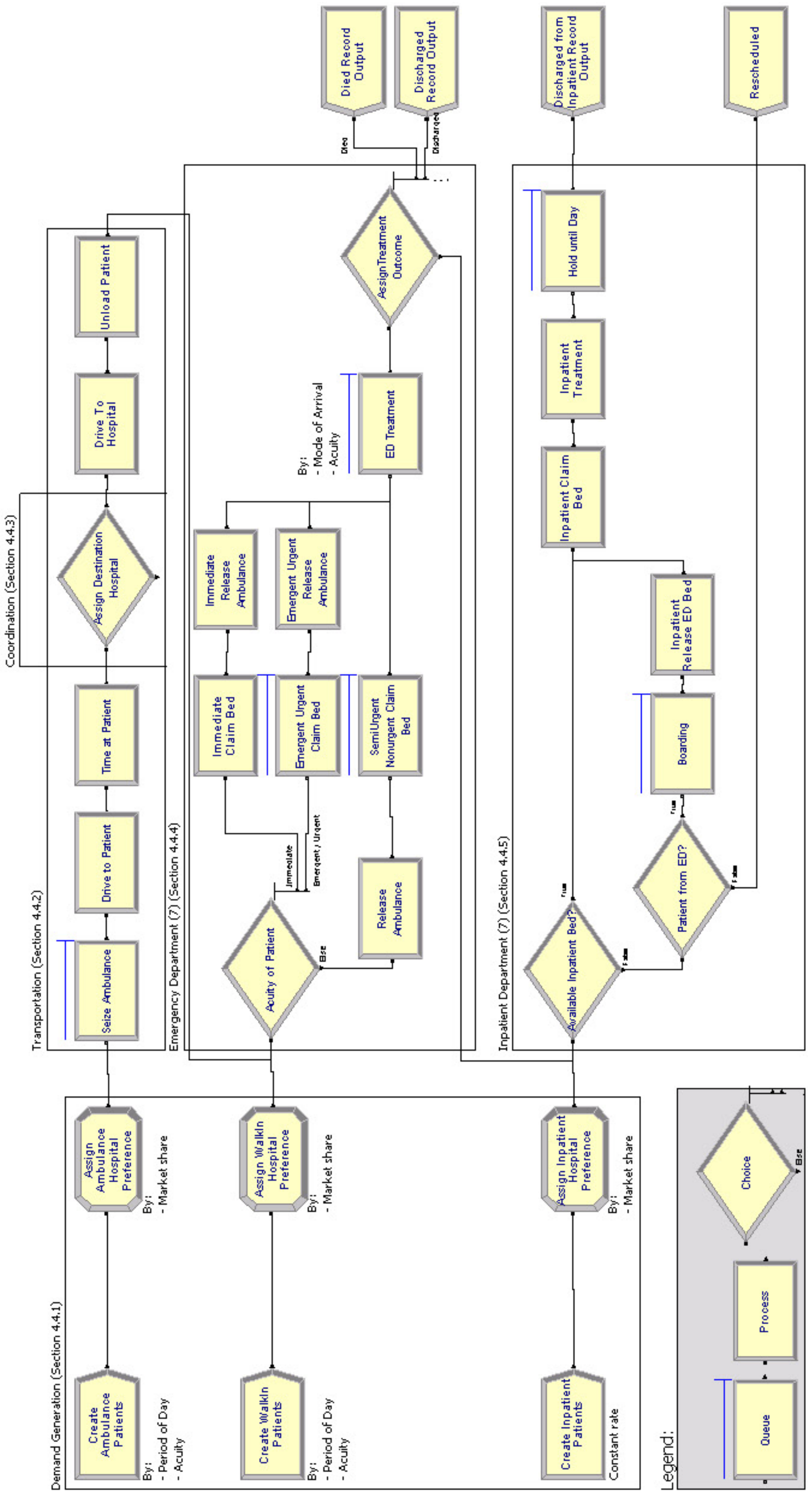


Figure 4.4: Process flow chart of our simulation model.

4.4.1 Demand generation

In our model we consider two types of patients, those that use services at the ED and those that arrive at the ID. In this section we outline the underlying model of both types of patient arrivals.

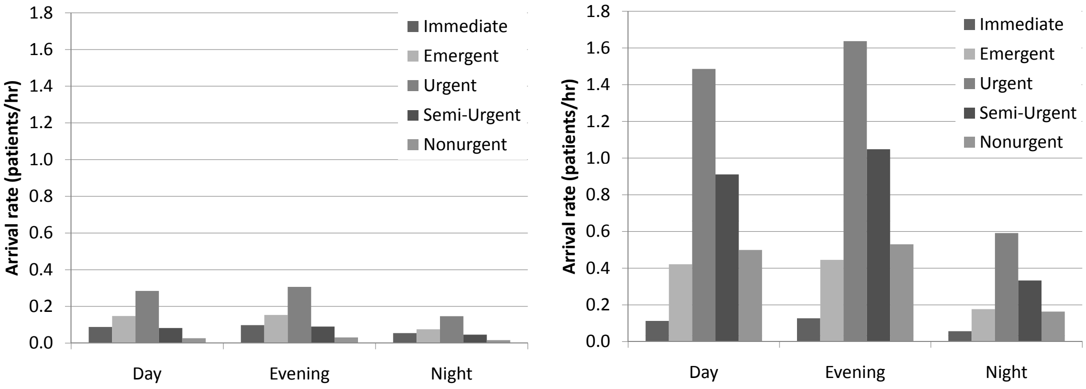
ED arrivals

In our model we will distinguish ED patients by their acuity, i.e. how urgently they need care. In practice, ED arrivals are non-stationary in several ways. Not only may the arrival rate vary over time, but the distribution of acuity or mode of arrival may also shift. In this section we first determine which factors significantly influence the arrival patterns observed by an ED and then determine the arrival rates, split by the significant factors. We capitalize names of variables when referring to parameters in our estimation models.

To determine which factors have a significant influence on the arrival pattern we use data from the National Health Ambulatory Medical Care Survey (NHAMCS) from 2006 and 2007. From the literature (e.g. Nawar et al. 2007, Pitts et al. 2008) and conversations with EMS professionals (Patterson 2010) we identified the following potentially significant factors: Month (January, February, . . .), Day of week (Monday, Tuesday, . . .), Period of day (Day: 7-15hrs, Evening: 15-23hrs, Night: 23-7hrs), Acuity (Immediate, Emergent, Urgent, Semi-Urgent, Nonurgent) and Mode of arrival (Ambulance, Walk-in). Time of day was also identified as a potentially influential factor but eliminated as the data was not granular enough to run an estimation model with this number of factors and Time of day included; therefore we use Period of day as a proxy. The NHAMCS data contained two additional Modes of arrival: Unknown and Public service (non-ambulance). We consider ED arrivals from which the mode of arrival was Unknown or that arrived by a public service (non-ambulance, e.g. police) to be walk-in arrivals. We believe that this simplification will not impact our insights regarding ambulance coordination (this is in line with e.g. Falvo et al. 2007).

The NHAMCS data do not contain the date of ED visit, only the day of the week and the month of the visit are recorded. Therefore, we take the total number of arrivals on a particular day of the week during each month and divide this by the number of occurrences of that day in that month. For example, in the NHAMCS data for January of 2006 we find that there were 587 arrivals on Mondays, and there were 5 Mondays in that month. Hence, on a typical Monday in January we expect 587/5 patient arrivals. These 117.4 arrivals on a typical Monday in January 2006 arrived during different Periods of the day, had different Acuity levels, and Modes of arrival. We now use an unbalanced ANOVA to test which factors have a statistically significant impact on the mean number of arrivals. To ensure that the error terms are normally distributed, we apply a natural log transformation (pp. 90-91 Tabachnick and Fidell 2001) to the mean number of arrivals. For model fit details we refer the reader to Appendix B.4.1. With all five factors we obtain an R^2 of 72.18% and conclude that the Day of week is not significant ($p = 0.370$). We also estimate a simplified model, eliminating the Month factor. For this model we obtain an R^2 of 71.94%, where the Day of week is again not significant ($p = 0.615$). Hence we assume the mean number of arrivals to be affected by 3 factors (Period of day, Acuity, Mode of arrival) for our simulation model. The mean number of arrivals as a function of these three factors is displayed in Figure 4.5, for the total of 59,504 retained arrivals (arrivals of Unknown acuity or where the Arrival time had not been recorded were discarded).

As the NHAMCS data do not identify subsequent arrivals at a hospital (for privacy reasons), we are unable to estimate a distribution for the interarrival times. Therefore we assume exponentially distributed interarrival times in line with several papers in the literature (e.g. Green and Nguyen 2001, Allon et al. 2009) (This would be easy to generalize). As the NHAMCS is a national data set, we scale the arrival pattern obtained from the data set to match the mean patient volumes for each of the Pittsburgh EDs (given in Table 4.2).



(a) Ambulance arrivals.

(b) Walk-in arrivals.

Figure 4.5: Mean number of ED arrivals per hour, by Acuity and Period of day, in the NHAMCS data of 59,504 arrivals in 2006-2007.

ID arrivals

Arrivals to the IDs cannot be obtained from the NHAMCS database. We obtained the number of discharges for our set of hospitals from "Profiles of U.S. Hospitals," a public data source published by Thomson Reuters (2009). This number of discharges is the number of patients discharged from a hospital during a year; we take this to be the number of patients treated, see Table 4.2. We assume that these patients arrive homogeneously between 7 a.m. and 3 p.m. on every weekday (260 weekdays per year). No detailed data on the Shadyside hospital was available. However, the characteristics of the patients at Shadyside most resemble those at Presbyterian (Ruffner 2010), hence we scale the number of discharges by the number of beds and obtain 24,045 discharges at Shadyside per year.

For those ED patients that require ambulance transportation we model the EMS process in the next section.

4.4.2 Transportation

In this section we describe the process that occurs between the moment that a 911 operator decides to dispatch an ambulance – the alarm – and the moment that an ambulance returns

Table 4.2: Pittsburgh hospitals.

Name	System	Beds	Patients/yr
Children's Hospital of Pittsburgh	UPMC	41(1)	65,000(1)
Magee-Women's Hospital	UPMC	14(1)	18,000(1)
Mercy	UPMC	28(2)	57,000(2)
Presbyterian	UPMC	34(2)	50,000(2)
Shadyside	UPMC	31(2)	45,000(2)
Allegheny General Hospital	WPAHS	30(2)	45,000(2)
The Western Pennsylvania Hospital	WPAHS	25(2)	35,000(2)

Sources:
(1): Personal inquiry at ED
(2): Guyette (2009)

(a) Emergency Departments.

Name	System	Beds	Patients/yr	Utilization (%)	LOS (days)
Children's Hospital of Pittsburgh	UPMC	260	13,567	77.7	5.5
Magee-Women's Hospital	UPMC	278	27,488	71.7	2.7
Mercy	UPMC	364	17,725	65	4.9
Presbyterian	UPMC	1263	58,740	76	6.0
Shadyside	UPMC	517(4)			
Allegheny General Hospital	WPAHS	583	29,132	74.4	5.5
The Western Pennsylvania Hospital	WPAHS	476	20,316	50.6	4.4

Sources:
(3): Profiles of U.S. Hospitals (Thomson Reuters 2009)
(4): `http://www.upmc.com/HospitalsFacilities/Hospitals/Shadyside/Pages/default.aspx`

(b) Inpatient Departments(3).

to service and is ready to respond to another alarm. During this complete period an ambulance is considered to be out-of-service. In line with Spaite et al. (1993) we recognize 4 intervals within this total out-of-service period: *(i)* Response time, from the alarm until arrival at the scene of the patient; *(ii)* Scene time, from arrival at the scene until departure to a hospital; *(iii)* Transport time, from leaving the scene to arrival at the hospital; *(iv)* Turnaround time, from arrival at the hospital until the ambulance is back in service. In this subsection we outline the data and our modeling approach for these 4 components of the out-of-service period.

We assume that walk-in patients transport themselves to a hospital of their preference, without considering the status of the hospital. To determine the preference of walk-in and transported patients for a specific hospital, we use the market share of an ED in Table 4.2.

Response time

Although almost each municipality within Allegheny county has their own EMS agency, the calls are dispatched through a centralized 911 facility. As modeling individual EMS agencies is beyond the scope of our research, we assume that ambulances are spread uniformly by area across the county, hence minimizing the maximum distance to any call (minimizing the maximum distance is a common metric in the Operations Research literature regarding the dispatch of emergency services, see e.g. Toregas et al. 1971). Furthermore, we assume the location of ambulances are continually re-evaluated to minimize this maximum distance. As ambulances become out-of-service, the remaining, in-service, ambulances each need to cover a larger area, resulting in an increased maximum response time. We approximate the response time by using a square grid representing the coverage area. An ambulance located in the center of a square area of w_a^2 miles, has an average a response distance, d_r, within its coverage area, of:

$$d_r = \int_{x=0}^{w_a} \int_{y=0}^{w_a} \left(\frac{1}{w_a}\right)^2 \sqrt{\left(x - \frac{w_a}{2}\right)^2 + \left(y - \frac{w_a}{2}\right)^2} dxdy,$$

which simplifies to:

$$\begin{aligned} d_r &= 1/6w_a\left(\sqrt{2} + \sinh^{-1}(1)\right) \\ &= 0.38w_a \end{aligned}$$

Now let w_{All} be the width of Allegheny county, estimated to be 27 miles (assuming the county, with an area of 730 square miles U.S. Census 2000b is square). The City of Pittsburgh has 13 ambulances serving 312,819 residents (U.S. Census 2000b). We approximate the number of ambulances in the county by scaling the number of Pittsburgh ambulances by the number of residents of the county (1,215,103, U.S. Census 2000b), leading to an estimate of 50 ambulances. Let N_A be the number of ambulances that are in-service. If the

N_A, in-service, ambulances each cover a square area, the area covered by an ambulance is w_{all}^2/N_A, which has a width of $w_{All}/\sqrt{N_A}$.

We assume that an ambulance with signals (light and sirens) drives at $v_A = 35$ MPH[2]. Now the expected response time to any patient, t_r, equals:

$$\begin{aligned} t_r &= 0.38\tfrac{w_{All}}{\sqrt{N_A}}/v_A \\ &= \tfrac{0.30}{\sqrt{N_A}} \end{aligned} \tag{4.1}$$

As unforseen factors influence this we assume this time to be exponentially distributed with mean t_r. The exponential distribution captures the fact that the variance increases as the mean increases, i.e. the further an ambulance has to drive, the more unforseen events that may delay it. Again, this is easy to generalize.

Scene time

Upon arrival at the scene, the EMS crew triages and potentially treats the patient. We use data from Carter and Grierson (2007) as an empirical distribution of scene time. These data are displayed in Figure 4.6.

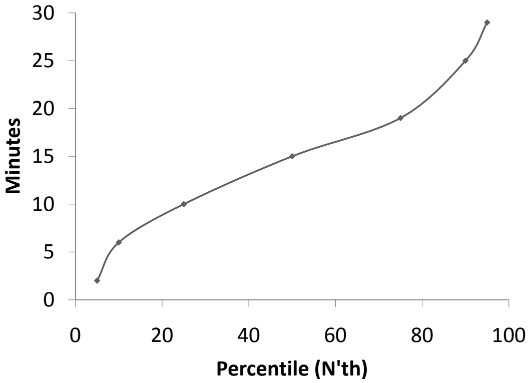


Figure 4.6: EMS Scene time distribution from Carter and Grierson (2007)

[2] In the center of the range suggested by Inoue et al. (2006), and also in the center of the 30-40 MPH range reported to the Portsmouth (UK) council meeting (Portsmouth City Council 2006)

Transport time

Once all activities at the scene have been completed, the ambulance leaves for a hospital. At this point the EMS crew decide on their destination hospital based on several factors. Two major factors are the patient's medical condition and patient preference (Hostler 2009). A patient with a specific medical condition might need to go to a specific ED (e.g. level *I* trauma center). Patient preferences may be guided by past experience, rankings (e.g. by the Joint Commission on Accreditation of Healthcare Organizations, JCAHO), insurance status, etc. In addition, the EMS crew may have preferences based on past experience or knowledge of ED status, but an EMS crew cannot transport a patient to an alternative hospital without consent of the patient. In Section 4.4.3 we describe different coordination mechanisms that may inform the EMS crew of current ED status.

As all hospitals that we consider are located in a limited geographical area (see Figure 4.1) we assume they are co-located at the center of our 2-mile circle. These hospitals see ED patients from within the city but also the surrounding county. In our model we include the 7 co-located hospitals and assume that all ED patients they see come from locations in Allegheny county[3]. To sample realistic driving distances we obtained the population and location of each of the 91 zipcodes in Allegheny county (U.S. Census 2000a). Using the coordinates of the center of each zipcode we calculated the distances to the center of our 2-mile circle. The probability of a distance (i.e. zipcode) being sampled is given by the fraction of the county's population that lives in the corresponding zipcode.

Turnaround time

An ambulance that has transported a patient to a hospital is responsible for the patient until care is transferred to ED personnel. The turnaround time is the time from arrival at the hospital, which includes the time for transferring care of the patient to ED personnel,

[3] Some patients may be coming from outside the county, but in these cases other hospitals are more likely to be nearby and hence preferred. As these other hospitals are at a significant distance from the seven that we consider we assume that they would not be part of potential mechanism to coordinate traffic.

potentially cleaning and restocking the ambulance, and returning to their station.

According to the EMTALA a hospital should accept responsibility for any patient presented to its doorstep. In reality, however, EMS crews may find themselves waiting at the ED for protracted periods because beds, stretcher spaces, or personnel are not always immediately available (Eckstein et al. 2005). To mimic the transfer of a patient from EMS crews to ED personnel we recognize the following three priority levels, based on interviews with EMS practitioners:

- Immediate: These patients will always be admitted to the ED immediately.
- Emergent & Urgent: These patients should be seen within 60 minutes (15 for Emergent and 60 for Urgent). The ambulance will wait until care of these patients has been transferred to ED personnel. These patients get priority over the next category.
- Semi-Urgent & Nonurgent: These patients need to be seen within 24 hours (2 hours for Semi-Urgent, 24 hours for Nonurgent). As long as there are any Emergent or Urgent patients waiting these patients will not be treated. We assume that the ambulance will leave without waiting until care of semi-urgent and nonurgent patients has formally been transferred.

Once an ambulance is no longer responsible for the patient, is cleaned, and restocked, it can return to service (in the simulation model we only model the transfer of care, not the time for cleaning and restocking). At this point the ambulance drives to a location that minimizes the maximum distance as discussed before. The ambulance locations are uniformly spaced across the county and hence the driving time can be determined similarly to the driving time to a patient (see equation (4.1)). If we assume an average speed of 25 MPH on this return trip, not using signals, we get an average return driving time of $0.38 w_{All}/25 = 0.38$ hours, which we again assume is exponentially distributed[4].

[4] The U.S. Environmental Protection Agency tests the fuel economy of cars at 21.2 MPH and 48.3 MPH, representing city and highway driving respectively. As it cannot be determined what fraction of an ambulances distance is either city or highway, we feel that a speed of 25 MPH is reasonable since Allegheny county is mostly considered an urban area (`http://www.fueleconomy.gov/FEG/fe_test_schedules.shtml`)

4.4.3 Coordination

Several factors influence the decision of to which hospital to transport a patient. Information about hospital status can be important for EMS crews as taking a patient to a crowded ED, or a hospital with an overflowing ID, decreases the quality of care received by the patient (as waiting for care is detrimental) and may also increase the turnaround time for the ambulance. Not all patients can be convinced to be transported to an ED that is their first preference though. In Pittsburgh around 90% of patients have an ED preference based on their insurance or home location, and up to 50% of those with a preference may be convinced to go elsewhere (Hostler 2010). Hence, about 55% of patients can be convinced to be taken to an ED based on the judgement of the EMS crew[5]. We use p_{div} to denote the fraction of patients that can be convinced to be transported to an ED that was not their initial preference[6] and let p_{div} vary between 25% and 75%. The EMS crew base their advice on the information that is available to them at the moment that they leave the scene. We recognize the following coordination mechanisms that each reveal different information to the EMS crew (these scenarios are summarized in Table 4.3):

Myopic hospitals: In this scenario no information about hospital status (ED nor ID) is available to EMS crews or patients. We randomly assign a patient to the ED with a probability proportional to the market share of ED patients as displayed in Table 4.2.

EMS coordination: As EMS crews visit EDs they observe the number of patients waiting at the ED. As observations by one EMS crew are not immediately relayed to all other crews we model the estimated queue length as an Exponentially Weighted Moving Average (EWMA) of the observed queue lengths. The EWMA uses a parameter α that discounts past observations[7]. In our setting a higher α means that information about queueing is relayed faster. We consider both regularly updated in-

[5] As UPMC Health Plan has about 1.4 million members (UPMC website 2010), this is a large player in the Pittsburgh area. Hence insurance preference might not strictly be tied to a specific hospital but rather a hospital system. This is, however, not generally observed by practitioners (Hostler 2010)

[6] Similar to p_{jdiv}, but for the case where the fraction is independent of the initial hospital preference

[7] As α increases more weight is given to the most recent observation.

formation ($\alpha \in \{0.4, 0.6\}$) and long run average information ($\alpha = \{0.01, 0.05)$). The average age, a measure of how "old" the estimate is, for an EWMA is $\frac{1-\alpha}{\alpha}$. Hence, for our range of α values we use observations that are on average between 0.67 and 99 ambulance arrivals old, where the latter reflects a long run average.

Hospital coordination: Hospitals that want to avoid crowding may benefit from coordination of ambulance traffic (fewer patients, less congestion, better patient care, etc.). We consider two coordination mechanisms through which hospitals provide information about their status:

1. Join the Shortest Queue (JSQ): In this coordination mechanism hospitals communicate their actual queue length; whenever there are no patients waiting the number of available ED beds is communicated. EMS crews advise patients to go to the hospital that has the shortest queue, or the most empty beds. This coordination mechanisms is similar to an EWMA estimate with $\alpha = 1.0$ with additional information if there are empty beds. Note however that an EWMA estimate with $\alpha = 1.0$ reflects the situation when an ambulance last visited an ED, while JSQ operates on current information.
2. NEDOCS: This coordination mechanism summarizes information about the instantaneous utilization[8] of the the ED, the instantaneous utilization of the ID, the time that the longest boarding patient has been boarding and how long the last patient that was admitted to the ED had waited. A lower NEDOCS score indicates a lower level of crowding. As the NEDOCS score also incorporates information from the ID it is "forward looking" in the sense that a full ID may soon lead to boarding in the ED.

Diversion signalling: A common policy in practice is for hospitals to go "on diversion," i.e. they request ambulances not to bring any more patients to their ED. In our model

[8] Note that the use of the word utilization here means the fraction of beds that are occupied, rather than the fraction of the population that uses a specific service, as is common in the emergency medicine literature.

hospitals set their diversion status based on the number of boarding patients (as in our queueing model, Section 4.3, and is common in practice, see e.g. Medical Advisory Committee, Pennsylvania 2004).

We recognize that hospitals and patients can have different perspectives on setting diversion status:

1. From the short run patient's perspective diversion levels should be set so as to maximize their quality of care. We consider two cases that reflect the Pittsburgh situation: a scenario in which only WPAHS sets their diversion level to maximize patient health outcomes (and UPMC never diverts), and a scenario in which both WPAHS and UPMC set their diversion levels to maximize the quality of care.
2. Diversion levels can also be used to maximize hospital revenues (or profit). Although Williams (2006) suggests that scheduled patients provide a certain revenue compared to the uncertainty inherent in ED arrivals, Guyette (2009), and Patterson (2010) argue that ED patients bring large revenues to both the ED and the ID. A diversion level can thus be used to strike a balance between these two revenue streams. For this objective we again consider two cases in which either just WPAHS, or both WPAHS and WPAHS use diversion levels. A hospital is assumed to set its diversion level so as to maximize its revenue. (Recall that a full ID may cause scheduled arrivals to leave the system)

Table 4.3: Summary of coordination mechanisms.

p_{div}	**Scenario**	**Usage**
{25%, 50%, 75%}	Myopic hospitals	UPMC & WPAHS
{25%, 50%, 75%}	EMS coordination, recent information EWMA: $\alpha \in \{0.4,\ 0.6\}$	UPMC & WPAHS
{25%, 50%, 75%}	EMS coordination, long run average EWMA: $\alpha \in \{0.01,\ 0.05\}$	UPMC & WPAHS
{25%, 50%, 75%}	Join the Shortest Queue (JSQ)	UPMC & WPAHS
{25%, 50%, 75%}	NEDOCS	UPMC & WPAHS
{25%, 50%, 75%}	Diversion Signalling, maximize quality of care	WPAHS only
		UPMC & WPAHS
{25%, 50%, 75%}	Diversion signalling, maximize hospital revenue	WPAHS only
		UPMC & WPAHS

4.4.4 Emergency Department care

We approximate the capacity of an ED by the number of beds in the ED. Not only is this common in the literature (e.g. Lambe et al. 2002, Schneider et al. 2003, Eckstein et al. 2005), it is also one of the most “rigid” resources required in the treatment of ED patients, staffing levels for example can be reviewed and adjust hourly (Patterson 2010). In Section 4.4.2 we have described how ambulance patients arrive at the ED, and walk-in patients transport themselves. In this section we analyze the duration of time a patient occupies a bed, which we refer to as the treatment time, as well as the outcome of the ED visit.

ED treatment time

Patients undergo a variety of treatments while in the ED. We use the NHAMCS data from 2006 and 2007 to analyze which factors have a significant impact on the treatment time of a patient and then estimate the parameters of a probability distribution for treatment times.

The NHAMCS data include: *(i)* Arrival time, *(ii)* Wait time until seen by a physician, and *(iii)* Length of visit (time between arrival and the ED discharge).

We use the difference between the Length of visit and Wait time as an estimate for the treatment time, but the actual time during which a bed is out-of-service may be significantly longer (as we will address in Section 4.5.2), as a patient may have been assigned a bed long before he is first seen by a doctor. However, we were unable to obtain more precise data, and believe that the treatment times obtained in this way provide reasonable base estimates for the treatment time distribution.

We now use an unbalanced ANOVA to test which factors have a statistically significant impact on the treatment time. To ensure that the error terms are normally distributed, we apply a natural log transformation (pp. 90-91 Tabachnick and Fidell 2001) to the treatment time. Similar to in Section 4.3.1 we suspect that the factors Acuity, Mode of arrival, Period of day, Month, and Day of week may drive treatment times. The details of the estimation can be found in Appendix B.4.2. The model with 5 factors model obtains an R^2 of 10.17% and all factors are significant. However, a simplified model that includes only Acuity and Mode of arrival still obtains an R^2 of 9.46%, hence we proceed with a model that only includes these two factors[9].

As we have determined the driving forces behind the length of the treatment we use Arena Input Analyzer (AIA), a tool within the Arena simulation package (Kelton et al. 1998), to estimate the parameters of various distributions, and select the most appropriate distribution. The data was well represented by a Gamma distribution, hence in our simulation model the treatment times are modeled as Gamma random variables with parameters as displayed in Table 4.4[10].

[9] Using a model that only includes Acuity we obtain an R^2 of 5.99%.

[10] We considered the Gamma, Weibull, Erlang, Exponential, Lognormal, Beta, Normal, Triangular, and Uniform distributions. The Gamma distribution is top ranked, by Squared Error, in all but three cases. For two of these three cases (Semi-Urgent and Nonurgent walk-in patients) the Gamma distribution still had a p-value below 0.005. For the third case (Nonurgent ambulance patients) the Squared Error for the top ranked (Erlang) distribution was 0.00106 (p-value of 0.0989) and for the Gamma the Squared Error was 0.00108 (p-value of 0.0486).

Table 4.4: Parameters for ED treatment times (minutes) by Acuity (NHAMCS).

	Ambulance				Walk-In			
	Mean	Std Dev	Scale	Shape	Mean	Std Dev	Scale	Shape
Immediate	259	330	230	1.13	192	270	177	1.08
Emergent	291	346	242	1.2	188	243	175	1.08
Urgent	252	297	197	1.28	159	212	144	1.11
Semi-urgent	207	219	172	1.2	111	156	109	1.03
Nonurgent	202	239	203	0.995	94.4	138	99.3	0.951

ED treatment outcome

Patients who leave the ED may do so in different ways: *(i)* A fraction of the patients are discharged from the ED and leave the hospital altogether, *(ii)* a fraction are admitted to the ID, and *(iii)* a remainder pass away as a result of the condition for which they sought treatment. In the literature the ID admission rate is often averaged across all patient characteristics (Falvo et al. 2007), and the number reported to us in interviews was too (Guyette 2009). However, the admission rates in the NHAMCS data for 2006 and 2007 are broken down by acuity and we observed a high variability by acuity. We thus estimate the fractions of patients that leave the ED in each way from the NHAMCS data from 2006 and 2007. Figure 4.7 displays our estimates for the different fractions for ways in which treatment in the ends.

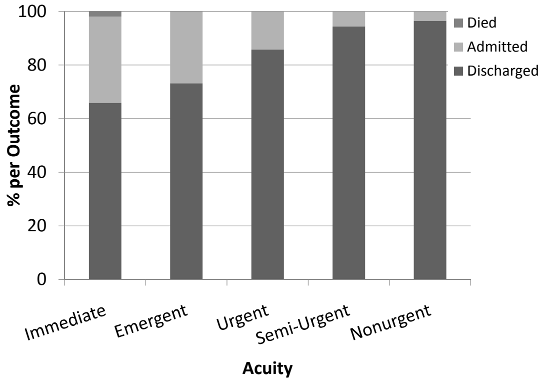


Figure 4.7: ED treatment outcome for patients of different Acuity

In order to compare the quality of care we use a "discounted quality" metric. Although the reputational, or perceived quality level may differ by hospital, all hospitals in the Pittsburgh area are fully certified and have various JCAHO awards. Therefore we set a uniform basic quality level which is then discounted by the waiting time: $100exp(-\text{Waiting time})$. In Section 4.5.2 we use this metric to evaluate the quality of care delivered, and use the average discounted quality of care to capture the level of care provided by all hospitals together, on average.

4.4.5 Inpatient care

The ID of a hospital sees two different types of patients. One type are patients that have been admitted from the ED: A bed in the ID is requested for an ED patient once his/her treatment in the ED is over. If an ED patient requests a bed in the ID and none are available, the patient remains in the ED (i.e. the patient is boarded in the ED). The second type are scheduled patients, their arrival process has been described in Section 4.4.1. In this section we will analyze the Length Of Stay (LOS) distribution for both types of patients.

Length of stay for admitted ED patients

The NHAMCS data set records the LOS for patients that have been admitted from the ED. Unfortunately we were not able to fit a model that has normally distributed error terms and explains a reasonably large percentage of variability (see Appendix B.4.3 for details). Following the logic for treatment outcomes (Section 4.4.4) we assume that Acuity is the main factor determining the length of ID stay. We now use AIA to estimate the distribution of LOS for each of the Acuity categories. Using Squared errors AIA ranked the Lognormal (3 Acuity levels), Exponential, and Gamma distributions as best fitting distributions. However, we note that the exponential distribution has a p-value below 0.005 for the estimate of the distribution of LOS for all Acuities[11]. We thus assume that the LOS

[11] As Thomson Reuters (2009) only reports mean LOS we prefer a distribution with only a single parameter, the exponential distribution has only one parameter and shows a good fit.

for an admitted ED patient is exponentially distributed with the parameters as displayed in Table 4.5.

Table 4.5: Length of stay data (days) by Acuity (NHAMCS).

	Mean	Std Dev	# Obs	Shift	Mean
Immediate	6.19	6.68	844	0.5	5.69
Emergent	5.69	7.07	1956	0.999	4.7
Urgent	5.55	5.72	3422	0.999	4.55
Semi-urgent	5.24	4.81	760	0.5	4.74
Nonurgent	5.57	5.59	240	0.5	5.07
Inpatient	5.63	6.15	7222	0.999	4.63

Length of stay for scheduled patients

LOS for scheduled patients could not be obtained from NHAMCS. However, Thomson Reuters (2009) provides the average LOS for all but one of the Pittsburgh hospitals (see Table 4.2). We observe that the LOS varies from 2.7 days (Magee-Women's hospital) to 6.0 days (Presbyterian). We incorporate these differences into our model of patient characteristics seen by the hospitals, these are likely also reflected in the capacities of their IDs (for which we do have data). For Shadyside hospital only the number of beds is available[12]. However, Ruffner (2010) estimates that Shadyside is comparable to Presbyterian in the sense that the LOS is similar and the number of patients per year is proportional to the number of beds.

For the distribution of the LOS for scheduled patients we reverse the argument made by Falvo et al. (2007): Falvo et al. (2007) estimates that the amount of revenues of an admitted ED patient is similar to the revenues generated by a scheduled arrival. Reversing their argument we assume that the LOS distribution is similar and hence assume that the LOS for a scheduled admit is exponentially distributed with a mean equal to the average LOS as reported by Thomson Reuters (2009)[13].

[12] This hospital is a single entity in the UPMC books with Presbyterian, which might explain why Thomson Reuters (2009) was not able to get specific data

[13] If we estimate the LOS for an average ED patient that is admitted (last row in Table 4.5) we get a LOS

We assume that patients are discharged from the ID every day between the hours of 7 a.m. and 3 p.m.. If a patient's stay/treatment ends between 3 p.m. and 7 a.m., the patient is kept in the ID and released the following morning[14].

4.4.6 Revenues

As the hospitals in Pittsburgh are part of two different systems (UPMC and WPAHS), coordination between the hospital systems will have an impact on their financial performance. Therefore we are interested in the revenues obtained from each patient that is seen at a hospital. Falvo et al. (2007) analyze the lost revenue for hospitals as a result of ambulance diversion or balking customers (those that leave before being seen/treated). This net revenue measure does not include additional charges by specialists in the hospital, nor does it reflect the collection rate (i.e. the fraction of billed charges that get paid).

Nevertheless, we use net revenue estimates from Falvo et al. (2007). The net revenues generated by an ED patient while in the ED depend on the Acuity level of the patient, these estimates are listed in Table 4.6. ED patients that are admitted to the ID, or scheduled arrivals to the ID, generate a net revenue of $8,551. Although Falvo et al. (2007) do not consider net revenues generated by walk-in ED patients that are admitted to the ID, we assume these net revenues are independent of the mode of arrival.

Table 4.6: Net revenue ($) per acuity category (Falvo et al. 2007).

Acuity	Net Rev. ($)
Immediate	3,209
Emergent	1,334
Urgent	963
Semi-Urgent	476
Nonurgent	317

that is within our range from Thomson Reuters (2009). If we fit the distribution the exponential distribution again has a p-value below 0.005.

[14] In runs of our simulation model with just scheduled arrivals we obtain occupancy rates which are on average 4% higher, which could be caused by patients being released in practice before their LOS ends, towards the end of the day.

4.5. Results

4.5.1 Results - Queueing model

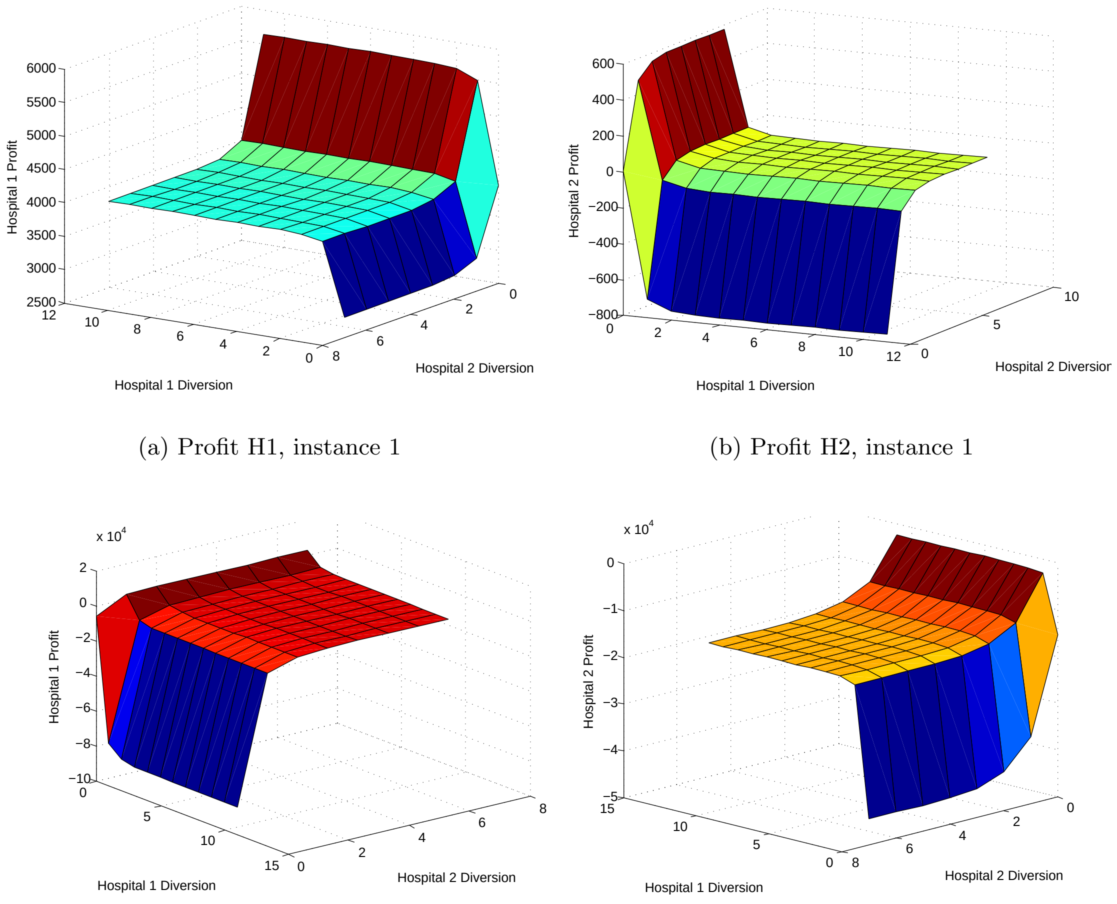
Using the analytical model as introduced in Section 4.3 we can analyze the effects of a traditional approach to minimizing ED crowding, i.e. the use of an ambulance diversion policy. By setting their diversion level, $\hat{n}_j$, hospital j requests ambulances go to hospital i instead. In this section we numerically explore some of the dependencies between optimal diversion levels and hospital characteristics that one might observe.

Setting Optimal Diversion Levels

We first explore how the diversion level impacts the optimal profit. We consider two instances (details regarding these instances can be found in Appendix B.2), which only differ in the cost assigned to patients that die in the ED and the cost of waiting. In the second instance the cost of a patient dying (waiting) is 1000 (100) times as large as in the first instance. The optimal profits (losses) are displayed in Figure 4.8. (Note: for readability the orientation of the axes are not consistent.)

What we observe in Figures 4.8(a) and 4.8(b) is that as the diversion level for one hospital increases, it increases its own profit and decreases the profit of the other hospital. This can be explained by the fact that if one requires there to be more boarding patients before redirecting ambulances, fewer ambulances get redirected. This leads to more patients being seen in your own ED, and fewer being redirected to the ED of the other hospital. If a typical patient is profitable this is desirable and hence, diversion levels will be set high. In Figures 4.8(c) and 4.8(d) we observe the opposite. In this instance a waiting or dying patient is very expensive; hence, one prefers to minimize the number of patients (to the lowest possible level, without shutting down the ED completely). In this case any patient that has to wait, however short, is unprofitable.

These functions are very steep close to a diversion level of 0 for the following reason. As our state space is 6-dimensional and infinite in four of these dimensions we need to truncate



(a) Profit H1, instance 1

(b) Profit H2, instance 1

(c) Profit H1, instance 2

(d) Profit H2, instance 2

Figure 4.8: Optimal profits as a function of diversion levels. Instance 1 has costs for waiting and mortality, instance 2 has high costs for waiting and mortality.

the state space. In order to minimize the effect of truncation we have set the parameters such that the steady state probability of states near the truncation is very low; hence the mean number of (boarding) patients is very low and a chance of diversion level from 0 to 1 has a large effect.

Effect of Capacity

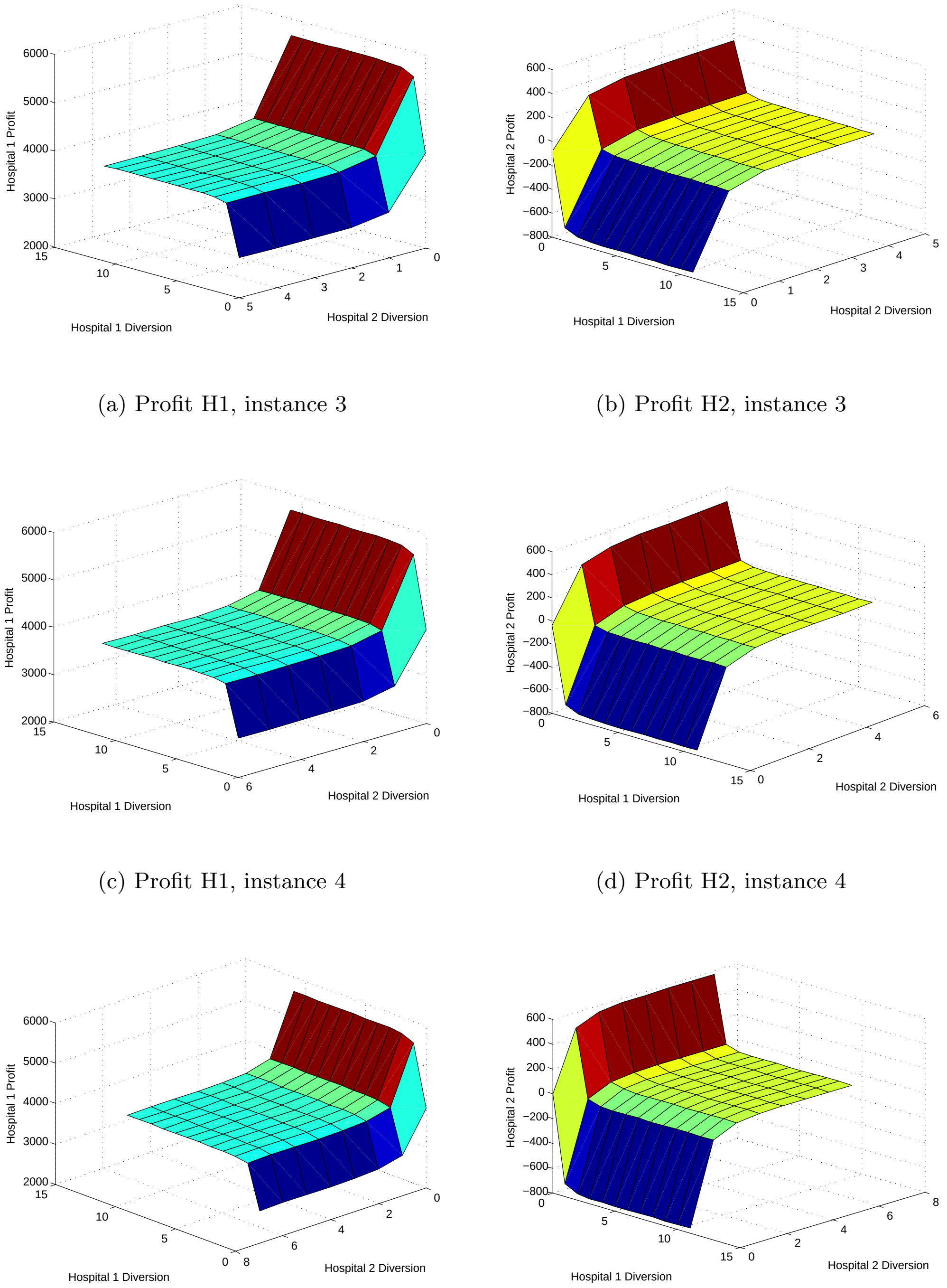
For given hospitals we now know how their diversion levels might be set. In Figure 4.9 we vary the capacity of the ED in hospital 2. Figures 4.9(a), 4.9(c), and 4.9(e) display the profits for hospital 1, and Figures 4.9(b), 4.9(d), and 4.9(f) display the profits for hospital 2 (see Appendix B.2 for details about these instances). What we see here is that the profits for hospital 1 are mostly insensitive to the capacity of hospital 2, they are, however, still sensitive to the diversion level. Recall that a higher capacity allows for a higher diversion level.

At hospital 2 we observe that the capacity directly impacts profitability, i.e. one needs a certain scale in order to avoid crowding and treatment delays, and reach a profitable level of throughput. What we observe here is in line with the literature (e.g. Allon et al. 2009), where it is found that the amount of time that a hospital spends on diversion decreases in the ED size. For hospital 2 it is optimal to set the diversion level as high as possible (i.e. very rarely divert) and the larger the ED, the less likely it is for the number of boarding patients to reach this level.

Competitive Insights

From Figures 4.8 and 4.9 we can also obtain competitive insights. When we compare Figures 4.9(e) and 4.9(f) we observe that it is optimal for both hospitals to set their diversion levels as high as possible. Now if we check what this does to the profitability of the other hospital we can make an interesting observation. Independent of where hospital 2 sets their diversion level, hospital 1 will always make a profit. However, by setting their diversion level at its highest level hospital 1 can cause hospital 2 to turn a loss. Although this situation would be optimal from a self interested point of view, it would mean that hospital 2 would go out of business at some point, which might be undesirable from the patient's perspective.

Having high diversion levels is similar to a "never divert" policy. The largest hospital system in Pittsburgh has been operating on a never divert policy "for a long time," while the smaller hospital system would divert under some circumstances (Guyette 2009). Here we see that, under certain parameter settings, this phenomenon can be reproduced in a simplified model (e.g. with the high diversion levels illustrated in Figures 4.9(e) and 4.9(f)). However, in reality there is not one, but multiple, hospitals within each hospital system.



(a) Profit H1, instance 3

(b) Profit H2, instance 3

(c) Profit H1, instance 4

(d) Profit H2, instance 4

(e) Profit H1, instance 1

(f) Profit H2, instance 1

Figure 4.9: Profits as a function of diversion levels and capacity at ED2. The capacity of the ED at hospital 2 (B_{2ED}) are 5, 6, and 7 in instances 3, 4, and 1 respectively.

Hence, if one would be able to divert to hospitals within the same system that might be beneficial. In the next section we will explore several coordination mechanisms that attempt to divert patients to a hospital where an adequate level of care can be provided, also taking into account the objectives of the individual hospitals.

4.5.2 Results - Simulation model

In this section we discuss the results from our simulation experiment. We first discuss some details about how the parameters of the simulation model were set. Then we will present the results and insights. We will evaluate the different coordination mechanisms on three metrics: average discounted quality of care, hospital revenue, and ambulance response times. Ambulance response times are a relevant measure as ambulances that are tied up at an ED are out-of-service and decrease the number of available ambulances to the public. Moreover EMS crews are important decision makers.

Simulation Parameters

As we gathered data from various sources we may have inconsistencies, which we discuss here.

When comparing the utilization of ED beds in our initial simulation runs with what is observed in reality we saw a significant difference: In our initial simulation runs we observe ED bed utilization levels in the lower end of the 50-75% range (and almost no patients waiting). Interviews with ED professionals and our personal experience however indicate that queueing is a serious issue. This difference between perceived and simulated performance measures could be explained by the imperfection of our treatment time data (as explained in Section 4.4.4). Therefore, we apply a scaling factor to the treatment time used in our simulation model to perform sensitivity analysis. We enumerate potential scaling factors until the confidence interval for the ED utilization level is centered at the desired utilization level. We will evaluate the sensitivity of our simulation outcomes for utilization levels of: 50, 75, 90, and 95%, all assuming no diversion takes place. See Figure 4.10 for the 95% confidence intervals for the ED treatment times at Allegheny General Hospital for each scaling factor (only the final scaling factors in our search are displayed). In order to obtain scaling parameters for every ED, we perform a similar analysis for each ED; for the scaling factors for other EDs and plots of confidence intervals of the corresponding ED utilizations, see Appendix B.7.

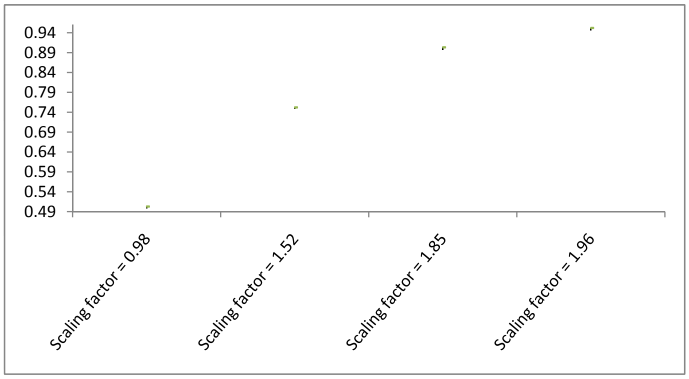


Figure 4.10: 95% Confidence intervals for the utilization at the ED at Allegheny General Hospital, for each scaling factor. Note that the confidence intervals are indeed so small that they may be hard to see.

Diversion signalling

One of the coordination mechanisms that we consider is diversion signalling. Unfortunately, no diversion level data exists for the Pittsburgh hospitals as many of them operate on a never divert policy. We know, for example, that UPMC hospitals have not used a diversion policy "for a long time" (Guyette 2009, Hostler 2009). According to Guyette (2009), Allegheny General Hospital (AGH) has used a diversion policy "up to fairly recently," but has recently abandoned this policy. Hence, "real" diversion histories cannot be obtained. Therefore, we use OptQuest (a general purpose optimizer that is able to search for the optimal parameter settings in an Arena model, Glover et al. 1999) to determine the "optimal" diversion levels. OptQuest uses a combination of scatter search (Glover 1989), tabu search (see e.g. Glover and Laguna 1997) and a neural network accelerator (see e.g. Glover et al. 1999). The exact mechanics of this optimization methodology are beyond the scope of our work, but at a high level OptQuest uses the output from the simulation model to determine in which direction it wants to direct the search for an optimal solution, and then provides new input for the simulation model (see the diagram in Figure 4.11). In other settings (e.g. Kekre et al. 2008) this approach has been shown to perform very well.

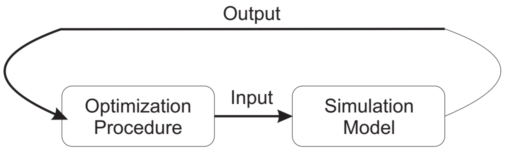


Figure 4.11: Coordination between optimization (OptQuest) and simulation (Arena)

In Section 4.4.3, when we introduced the diversion signalling coordination mechanism, we introduced two different objectives: Maximizing the quality of care and maximizing hospital revenue. For each of these objectives we consider the two cases in which either only WPAHS hospitals use diversion (as was the case in Pittsburgh until recently) or in which all hospitals use diversion signalling.

Our interviews with practitioners indicate that there is very little exchange of information between hospitals, even if the hospitals belong to the same hospital system (e.g. one UPMC ED physician could not tell us the number of beds or yearly number of patients of several other UPMC EDs); hence, we assume that all hospitals set their diversion levels in isolation (i.e. without considering the reaction of other hospitals to their change in policy). In order to analyze a system with fully coordinated diversion levels, in which all hospitals jointly optimize their diversion policies, one would need to solve a global optimization problem in 7 dimensions (one for each ED), which is outside the scope of this research and is not applicable in Pittsburgh[15]. In Appendix B.6 we list the diversion levels that were selected

[15] We did run several experiments with the goal of finding a globally optimal set of diversion levels for the case when 75% of patients could be convinced to be taken a hospital that was not their initial preference (as we expect diversion levels to have the most impact if more patients would oblige to the request by the hospital). We ran experiments for ED utilization levels of 50, 75, 90, and 95%. Although the optimization routine was able to eliminate several solutions from the set of potentially optimal solutions, still many

using optimization on a per hospital basis.

Repetitions & Run length

For every simulation experiment (each coordination mechanism is simulated with 4 utilization levels and 3 values for p_{div}) we run 20 replications of 1/2 year (183 days)[16]. Studies that perform an intervention (i.e. actually make a change to the system in a controlled setting) in how ED patients are routed use study lengths between 1 week (Vilke et al. 2004a) and 17 months (Patel et al. 2006), with several under 6 months (e.g. Sprivulis and Gerrard 2005, Larson 2008). Hence, we are confident that both from a statistical as well as a practical point of view we have picked reasonable simulation settings.

To reduce the amount of variation in our simulation outcomes we apply a variance reduction technique know as Common Random Numbers (see e.g. Law and Kelton 1999), which synchronizes the random events between the evaluation of different scenarios.

Simulation Results

In this section we analyze the effect of the coordination mechanisms outlined in Table 4.3 in Section 4.4.3. We evaluate their effect with regard to several metrics: Average discounted quality of care for all patient in the simulation (i.e. across the county), Hospital revenues, and Ambulance response times. However, first we discuss how the optimal diversion levels are set.

When we compare how the different objectives introduced for the diversion signalling coordination mechanism influence the optimal diversion level it stands out that there appear to be many potential solutions for which the objective values are statistically indistinguishable. In our simulation model this means that several diversion levels could be optimal.

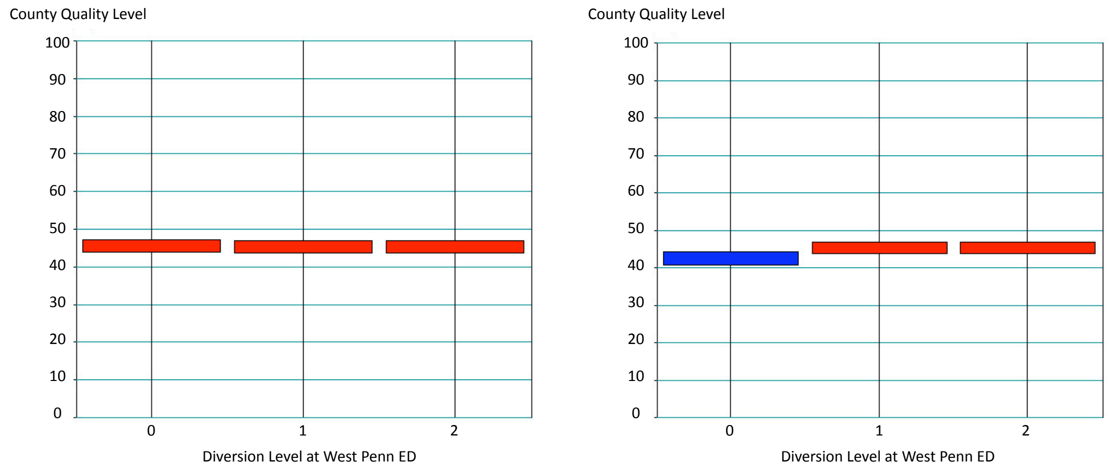
scenarios are statistically indistinguishable (even though the standard deviation of the estimate of overall quality was relatively small) and have significantly different parameter values. An example of the progress of OptQuest can be found in Appendix B.5.

[16] Based on preliminary experimentation we were confident that this would yield results that are powerful enough for our purposes. In these preliminary simulation experiments we considered the coordination mechanism in which hospitals are myopic and used ED utilization levels of 50, 75, 90, and 95%. We obtained very narrow 95% confidence intervals for the average discounted quality level (as can be seen in Figure 4.13).

OptQuest ranks the solutions by highest mean value for the objective under consideration (even though this mean may not be statistically larger than many of the other means)[17].

For an individual hospital it might appear to be optimal to divert all incoming ambulance traffic from an average discounted quality of care perspective, i.e. the diversion level is set at 0 (see Appendix B.6). However, for the instances where a diversion level of 0 provides the highest average discounted quality of care, the average discounted quality of care provided with a diversion level of 0 is statistically indistinguishable from other diversion levels. On the other hand, when the highest average discounted quality of care is achieved with a non-zero diversion level, the performance of the instance with the diversion level equal to 0 is typically statistically different (worse). Figure 4.12 shows instances in which the non-zero diversion level leads to statistically insignificant, and statistically significant better performance. We see in Figure 4.12(a) that when a small fraction of patients is willing to listen ($p_{div} = .25$), setting the diversion level at 0 (attempting to divert all ambulance traffic) has no effect, as very few patients will change their behavior. But, as p_{div} increases to 0.75 in Figure 4.12(b), the diversion level will have to be increased in order to maximize the average discounted quality of care: The hospital does not have to be as aggressive in diverting as it will have more impact as more patients oblige. Hence, setting the diversion higher balances the patients between the different EDs (setting it at 0 with many patients obliging would just overload the other EDs). For several instances, in addition to using OptQuest to find the optimal diversion level, we enumerated over all potential diversion levels and observed that the hospital revenue has a similar shape as the profit curve described in Figures 4.8 and 4.9 in Section 4.5.1, even though we dealt with a hypothetical set of parameters in that section.

[17] In reality the difference between two high diversion level may not be very large, as there may be practically no difference between never diverting or only diverting when more than 50% of the ED beds are occupied by boarding patients.



(a) $p_{div} = 0.25$, not statistically different.

(b) $p_{div} = 0.75$, diversion level of 0 performs statistically worse.

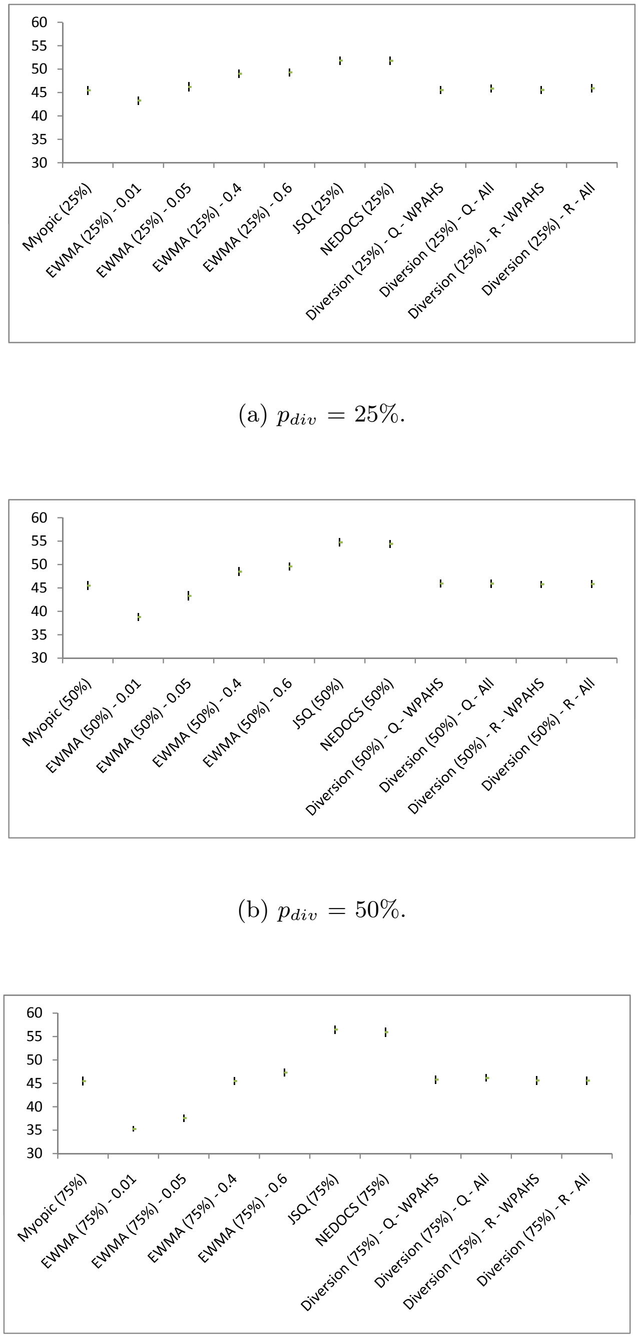
Figure 4.12: Confidence intervals for the average discounted quality as a function of the diversion level at The Western Pennsylvania Hospital.

Average discounted quality of care

In Section 4.4.4 we introduced a discounted quality metric: $100exp(-\text{Waiting time})$. We now explore which coordination mechanism is best able to deliver a good average discounted quality of care over the entire county, i.e. all patients. In Figure 4.13 we plot confidence intervals for the average discounted quality for each of the coordination mechanisms for varying levels of p_{div}, i.e. the likelihood that a patient can be persuaded to be taken to an ED that was not his or her initial preference. In Figure 4.13 we limit ourselves to the case of a utilization of 95%. Qualitatively similar effects are observed for other utilization levels, these graphs are displayed in Appendix B.8.

In Figure 4.13 the JSQ and NEDOCS coordination mechanisms are identified to perform best. Let's first focus on Figure 4.13(c). If we compare the different versions of the EWMA[18] coordination mechanisms we see that the performance deteriorates as the α pa-

[18] Recall that in this mechanism no information is disseminated by the hospitals and EMS crews talk among each other to update the estimate of the queue length at a hospital, then the hospital with the lowest queue length estimate is selected.



(a) $p_{div} = 25\%$.

(b) $p_{div} = 50\%$.

(c) $p_{div} = 75\%$.

Figure 4.13: Average discounted quality for a 95% utilization level

rameter decreases. This can be explained as follows: As α increases the average age of the observed queue lengths in the estimate increases. For $\alpha = 0.01$ the average age of the observations in the EWMA estimate is almost 100 arrivals old. An average ED in our set sees 7,280 ambulance arrivals per year ($45,000 * 0.16$, where 0.16 is the fraction of patients that arrives by ambulance), which translates to roughly 20 patients per day. Hence, the average observation used by EMS crews to determine their best choice of ED is about 5 days old, while the situation in that ED may have changed significantly. Another potential drawback could be that the EWMA queue estimate causes the arrival stream into an ED to be rather bursty. That is, if an ED currently has the highest estimated queue length it will attract little ambulance traffic in the near future (as all estimates only slowly get updated as α is low). Almost “automatically” the other queues will lengthen and this queue will become the shortest, by the EWMA estimate (even though walk-in arrivals may have been coming in throughout the period that ambulance traffic was being diverted), and attract a significant amount of ambulance traffic. As the decision of which ED to go to is made before the patient is transported to the ED several ambulances may have based their decision on the same, old, information. Thus directing traffic based on old information can lead to unnecessary diversion, bursty arrivals, and lower quality. As α approaches 1 we get close to approximating the JSQ coordination mechanism (note that our version of JSQ is not a “true” JSQ policy as we also take into account the number of empty beds, if any, and that JSQ also differs from EWMA in that JSQ uses real time information rather than observations when an ambulance last visited a hospital).

The coordination mechanisms that use a diversion level (either maximizing average discounted quality or revenues) show a performance very similar to the performance of the myopic policy which never diverts. This was to be expected given that many solutions obtained by OptQuest often proved to be statistically very similar to one another implying that all share very little diversion. The fact that diversion levels have very little effect on the average discounted quality could provide comfort to the managers at various Pittsburgh hospitals that decided to use a "never divert" policy (Guyette 2009): We show they are detrimental to the quality of care, and maximize revenues from ED operations.

The very strong performance of both JSQ and NEDOCS can be explained as follows. These metrics have a higher information content than the diversion level policies in the sense that they do not just communicate whether an ED is crowded but also provide a measure to what degree an ED is crowded, i.e. even when the ED is very empty or very full one is able to rank EDs by these metrics while a diversion signal would either be "on" or "off." This higher information content allows these measures to be informative also when none or all EDs are on diversion. Furthermore, the NEDOCS mechanism also provides insight into future developments as it also captures the state of the ID, providing information on how likely the flow through the ED is going to keep moving. Another relative strength of JSQ and NEDOCS as compared to the EWMA measures is that they have current information, rather than observations from past visits by ambulances.

We illustrated that the policy most commonly used to mitigate crowded EDs (diversion signalling) has limited effect on the average health outcomes as defined by our average discounted quality measure, i.e. it often performs very similar to the policy in which all patients go to their "preferred" hospital as determined by ED market shares. However, we show that some others forms of coordination (JSQ, and NEDOCS) can significantly increase the level of patient care by decreasing the waiting times[19]. In the specific setting in Pittsburgh it might not be feasible to use the JSQ or NEDOCS measure on a city wide scale. However, using them to coordinate ambulance arrivals between hospitals within the same system may already deliver significant benefits.

We see the same qualitative effect in all scenarios. Comparing Figures 4.13(a) and 4.13(b) with Figure 4.13(c), we see that the aforementioned effects are less pronounced as p_{div} is lower. This is not unexpected, as fewer patients are willing to adjust their hospital preference if so advised by an EMS crew, the lesser effect these coordination mechanisms can

[19] As we took the base quality levels to be equal for all EDs, any improvement in average discounted quality is due to decreased waiting times.

have (although even for $p_{div} = 25\%$ the performance increase of JSQ and NEDOCS over the other policies is statistically significant).

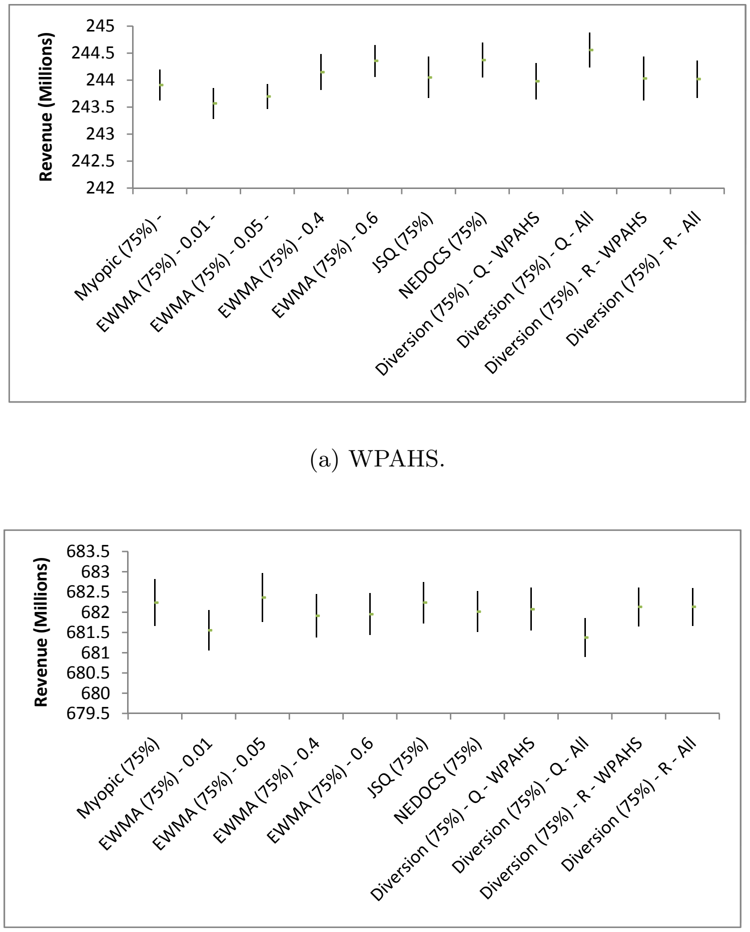
Hospital revenues

When we defined coordination mechanisms in Section 4.4, we recognized that hospitals may be maximizing their revenue (or profit) rather than the average discounted quality of care. So far we have evaluated the effect of coordination mechanisms on the average discounted quality of care. We now discuss how hospital revenues are impacted. We limit ourselves to discussing the cases with a utilization of 95% and $p_{div} = 0.75$ as hospital systems would be impacted most if more patients are diverted.

The 7 hospitals that we used in our study are part of two hospital systems, WPAHS and UPMC. Figure 4.14 displays the revenues for each hospital system. For WPAHS only the myopic and the EWMA scenarios with low α are positively identified as having statistically lower revenues than the other scenarios. All other coordination mechanisms have revenues that are not statistically different from one another.

From the results in Figure 4.14 we can conclude that those coordination mechanisms that improve the quality of care and response times do not lead to a statistically significantly lower revenue, at a hospital system level.

However, as mentioned earlier, the hospitals in Pittsburgh are not very tightly integrated, even within a hospital system. Given this low level of integration we also evaluate individual hospital revenues as well as individual ED revenues. Figure 4.15 displays the revenues across different coordination mechanisms for Allegheny General Hospital (AGH); for results on the other hospitals in our study see Appendix B.9. At the ED level there are several coordination mechanisms that have revenues that are statistically lower than the maximum possible level. These coordination mechanisms are: Myopic, EWMA(0.01), and Diversion signalling (all except the case where WPAHS implements diversion levels that maximize the optimal discounted quality level). Note however that, although these differences are statistically significant, the maximum percentage difference between the center of



(a) WPAHS.

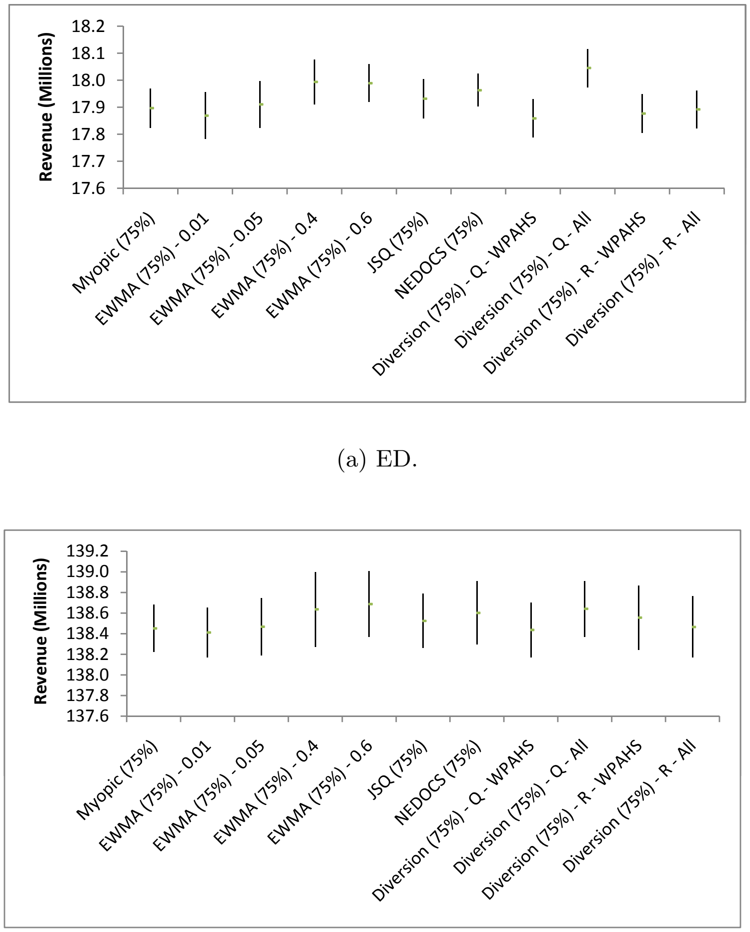
(b) UPMC.

Figure 4.14: Hospital system revenues for a utilization of 95% and $p_{div} = 75\%$.

the confidence intervals is 1.0%. Although a change in policy could thus have an effect on ED revenues, this effect is small. At a hospital level, taking into account revenues from ED and scheduled patients, the statistical significance disappears[20]. The fact that we observe only a limited effect on hospital revenues can be explained from a redistribution perspective. That is, using any of the mechanisms we analyze, traffic gets diverted away from a congested ED. However, these patients are still treated somewhere, thus diverted traffic increases the patient volume at another ED. In reverse, the ED that was once congested will

[20] The maximum percentage difference for AGH shrinks to 0.2%

attract additional traffic at times that other EDs are more congested. Hence, the amount of patient treatment that needs to be performed merely gets distributed more equally over time, an increased number of patients arrive at an ED when it is not congested versus a decreased arrival rate when an ED is congested. This roughly balances out and keep the overall revenues practically stable.



(a) ED.

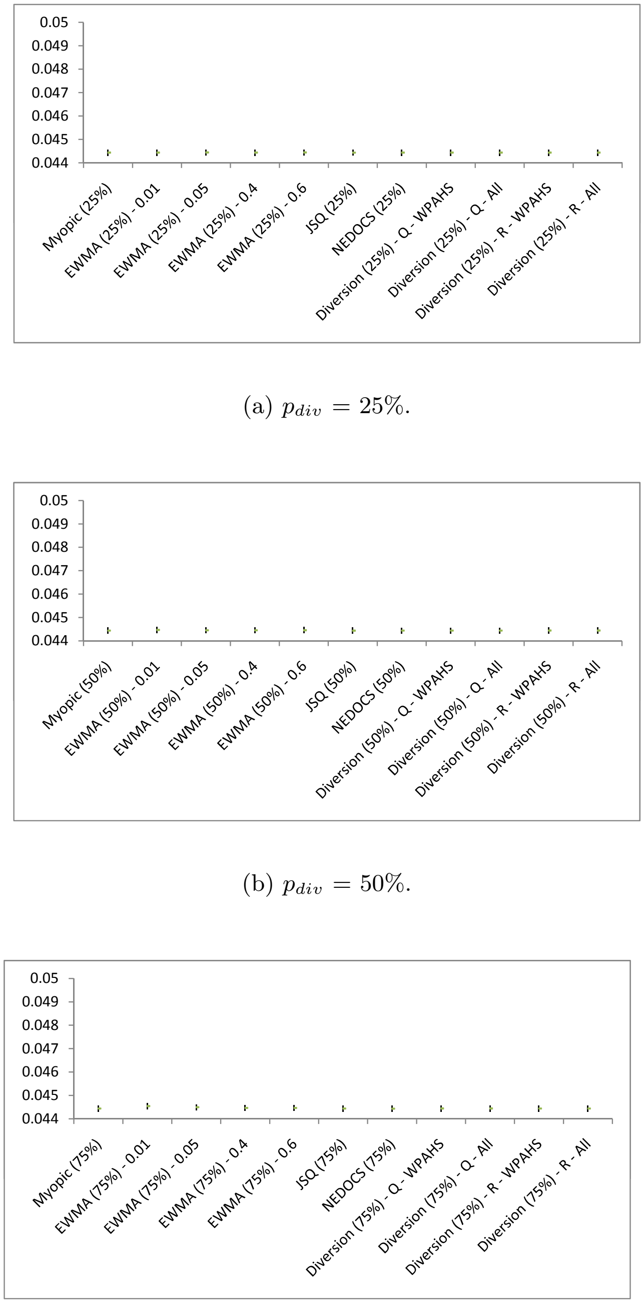
(b) ED & ID.

Figure 4.15: AGH revenues for a utilization of 95% and $p_{div} = 75\%$.

Ambulance response times

Eckstein et al. (2005) report that ambulance crews may be stuck in an ED for prolonged periods of time. We now investigate the effect of our coordination mechanisms on the response time of ambulances. Recall that ambulances that are stuck in an ED reduce the number of available ambulances in the field, hence increasing the driving distance and time to patients.

Figures 4.16 and 4.17 plot the response times for ED utilization levels of 50% and 95%. In Figures 4.16(a) through 4.16(c) we see that only if a large fraction of patients are willing to reconsider their hospital preference do we observe some effect on ambulance response times. Comparing the results for a utilization of 50% and those for 95% one should note the ratio between the shift of the mean (large for the 95% utilization case) as a function of the information sharing mechanism and the inherent variability in the output. In case the EDs are heavily utilized (as anecdotal evidence suggests they are, Guyette 2009) the effectiveness of EMS crews can be greatly enhanced by providing them with good crowding indicators (e.g. JSQ or NEDOCS) rather than having EMS crews rely on their own past experience (i.e. EWMA with $\alpha = 0.01$ or 0.05).

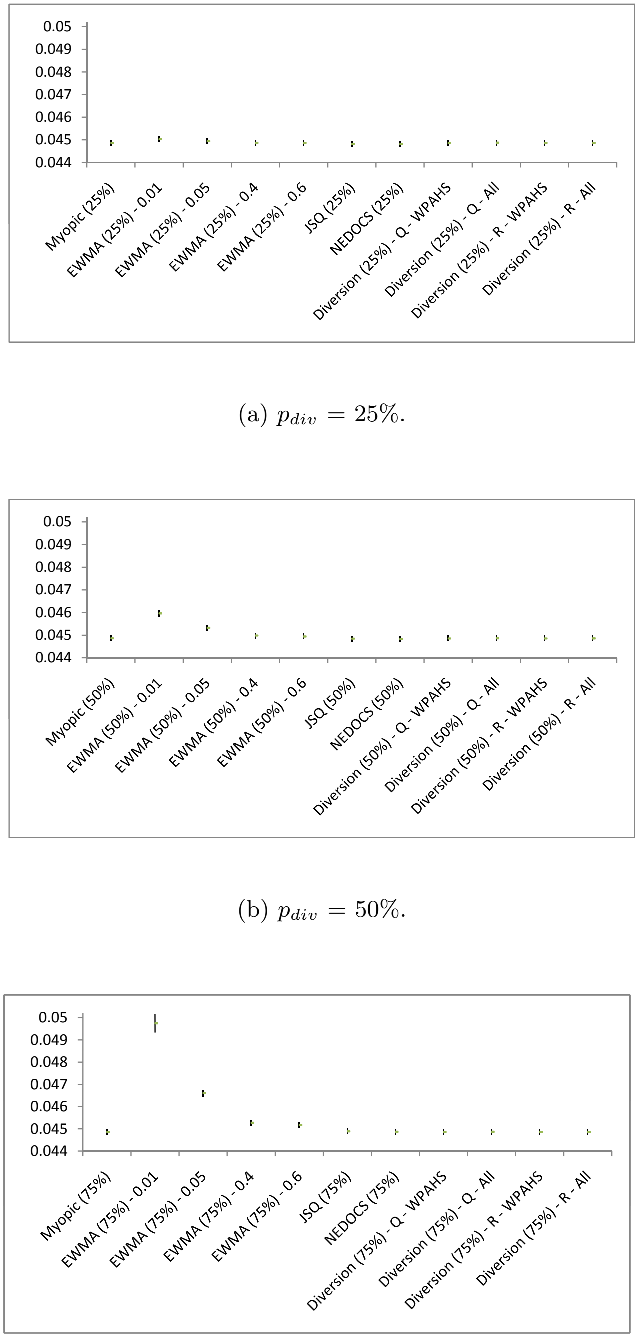


(a) $p_{div} = 25\%$.

(b) $p_{div} = 50\%$.

(c) $p_{div} = 75\%$.

Figure 4.16: Ambulance response times for an ED utilization level of 50%.



(a) $p_{div} = 25\%$.

(b) $p_{div} = 50\%$.

(c) $p_{div} = 75\%$.

Figure 4.17: Ambulance response times for an ED utilization level of 95%.

4.6. Conclusions

In this chapter we studied the effectiveness of different mechanisms to provide EMS crews and ambulance patients with information about the status of hospitals. This information is used to decide which hospital will be best able to provide care to the patient with minimal waiting times, hence improving the quality of care delivered. We investigated a commonly used coordination mechanism, diversion levels, using a simplified queueing model and then studied a wider variety of coordination mechanisms, including the diversion level mechanism, using a simulation model applied to the situation as it exists in Pittsburgh (PA).

We were able to capture many real-world complexities in our simulation model that could not be captured in the queueing model, however, there are still certain features that could not be captured. Our model could be extended to capture some of these complexities:

In our models we considered a generic hospital. Although we incorporated data on size (# beds), patients volumes, and length of stay from specific hospitals they may differ on more dimensions. Many hospitals provide specific types of care (e.g. Presbyterian is strong in transplant surgery and Shadyside has a larger cancer center) or provide care to specific patients populations (e.g. Magee-Women's Hospital and Children's Hospital of Pittsburgh). These differences in patient population may have an effect that is not captured in our data or model at this point. With insight into patient characteristics we could also obtain better estimates for ED treatment times, helping to better connect data from different sources.

Patient transportation times have been modeled based on how the population is distributed across the county. However, the population may shift significantly across the county as people go to work. Not only may the location of people change, also driving speeds of ambulances may be affected by such things as rush hour. Although we are confident that we managed to capture the first order effects on the level of patient care, more detailed data might help quantify the effect of coordination mechanisms on response times.

In the specific situation in Pittsburgh patients appear to have a rather strong hospital preference. Anecdotal evidence (Hostler 2009) suggests that this might be different in other regions, but also that the loyalty might be to a specific hospital system (e.g. based on insurance coverage) rather than by specific hospitals. If people actually have a hospital system preference this could motivate hospitals to divert ambulances within their own hospital system, rather than diverting them at a county level. This would keep the revenues within the same system while improving the quality of care provided.

Finally, in our simulation study we modeled a system of 7 EDs. Although EDs in Pittsburgh have very little communication, it could be interesting to see what the effect is of a diversion level policy in which the diversion levels are set taking into account the diversion levels at all other EDs. Although this could lead to interesting insights, the diversion level policy would still suffer from the limited amount of information contained in a single diversion signal, as opposed to the information contained in the JSQ or NEDOCS coordination mechanisms.

The two approaches developed to analyze policies to direct ambulance traffic can be used to inform policy and decision makers in the field. In a stylized setting we showed the dependencies of the parameter settings in a commonly used policy to avoid congested EDs: ambulance diversion. We showed how the optimal diversion levels depend on the estimates for various cost parameters included in our model, and that competing hospitals could set their diversion levels such that the revenues, and quality of care at the other hospital are decreased. In this sense diversion levels can be used as a competitive mechanism.

In our simulation study we capture many of the features that are present in a real-world setting. Using the situation as it exists in Pittsburgh, we estimated realistic models for patient arrivals and the distribution across patient types using data from the National Hospital Ambulatory Medical Care Survey, Pittsburgh hospitals, and EMS professionals. We used these models to simulate the performance of a set of 7 hospitals in the Pittsburgh area. In our hospital models we incorporated details about the operations within the ED and the ID, as the interactions between ED and ID have a significant effect on ED operations.

We introduced and analyzed several coordination mechanisms. These coordination mechanisms vary significantly in terms of how much information hospitals share. We show that several of the mechanisms that are commonly used in practice (never divert, or use of diversion levels) perform reasonably well. However, if a hospital never diverts ambulance traffic EMS crews may start to rely on their own, and other crews' experience. The mechanism that we used to capture the reliance on past experience led to a significant decrease in the average quality of care.

As hospitals are not able to restrict the use of EMS crew's experience, their "never divert" policies will be overtaken by mechanisms in which EMS crews make their own inference about the status of a hospital (this may even be the case if diversion levels are set too high, from the perspective of EMS crews). Hence, policies that provide too little information to EMS crews are likely to be replaced in the field by a policy that may significantly decrease the average discounted quality of care provided.

Two of our coordination mechanisms (JSQ and NEDOCS) disseminate more detailed information about the status of a hospital. We show that the additional richness in information significantly increases the average level of care that can be provided. These mechanisms merely redistribute the arrivals of patients over time. Hence, the quality of care provided can be increased without decreasing revenues.

Chapter 5

Conclusions

In this thesis we have analyzed three situations in which an operations manager has to make decisions in the face of an uncertain future. In each of the chapters we aimed to aid decision making by providing the operations manager with a model of uncertainty. Using the models that we provide and analyze the operations manager is able to make better decisions and evaluate how these decisions could play out in the future. One unifying these is that we find that thresholds can be used effectively to help the decision maker.

In the first chapter the operations manager is tasked with operating a natural gas well. As the price of natural gas fluctuates significantly, the operations manager needs to adapt his decisions not only to the prevailing conditions, but also to expectations about the future. Using stochastic dynamic programming we analyze the optimal deployment policies for technology that can be used to enhance production and communication, and guide decisions on when gas should be produced from the well at all.

In the second chapter, customers request products at random intervals. From the perspective of the operations manager not all customers are equal, and hence he wants to protect some inventory for future requests from more important customers. We show that a critical level inventory policy is able to provide good levels of service at a reasonable cost. We compare and contrast the critical level policy with more naïve inventory policies as well as with the globally optimal policy and conclude that, while it is easy to implement, the performance of the critical level policy can be close to that of the globally optimal policy. In addition, the critical level policy is largely insensitive to the extent of variability in the replenishment lead times.

In the final chapter we consider a situation that by definition is highly uncertain: providing emergency care. Emergency Departments (EDs) often observe that demand for emergency services exceeds supply, which decreases the level of care that can be provided to patients. We compare different mechanisms that can be used to relay information on hospital status from the ED to EMS crews. EMS crews use this information to predict the level of care that their patient will receive and the crew's own delay at the ED and advise the patient in hospital selection. Our comparison highlights that *(i)* some commonly used policies result in almost indistinguishable levels of quality, *(ii)* providing no information at all to EMS crews and leaving them to base decisions on their own experience can be detrimental for the level of care, and *(iii)* providing EMS crews with certain information-rich indicators can greatly improve the quality of care, possibly without compromising hospital revenues.

The unifying theme in these three chapters is that, in the face of uncertainty, an operations manager needs to adapt a model of the uncertain future. Using a model of uncertainty the operations manager can, when provided with the proper information, greatly enhance his decision making abilities. In each of the chapters we evaluate policies/machanisms of different complexity and conclude that in many cases a simple threshold policy might yield much of the benefits of complex state-dependent policies.

The applicability of the models analyzed in this thesis is not limited to the specific settings to which we applied them. There is a great variety of operational decisions that can benefit from the principles that we have uncovered in these specific settings.

Appendix A

Inventory Rationing for a System with Heterogeneous Customer Classes

A.1. Detailed structure of the submatrices of the generator

The generator Q as defined in (3.6) consists of six submatrices each of which we describe here. First, B_0 is an $(S+1) \times (S+1)$ matrix and captures the transitions within level 0 of the Markov process. Let i and j be the matrix indices, they equate to the level of inventory + 1:

$$B_0(i,j) = \begin{cases} -(\lambda_2 + S\mu) & \text{if } i = j = 1; \\ -(\lambda + (S-i+1)\mu) & \text{if } 1 < i \leq S+1 \text{ and } j = i; \\ (S-i+1)\mu & \text{if } 1 \leq i < S+1 \text{ and } j = i+1; \\ \lambda_1 & \text{if } 1 < i \leq c+1 \text{ and } j = i-1; \\ \lambda & \text{if } c+1 < i \leq S+1 \text{ and } j = i-1; \\ 0 & \text{otherwise.} \end{cases}$$

Next, B_{-1} is an $(c+1) \times (S+1)$ matrix and describes the transition, through the arrival of a replenishment order, from level 1 to level 0:

$$B_{-1}(i,j) = \begin{cases} (S-c+1)\mu & \text{if } i = c+1 \text{ and } j = c+1; \\ 0 & \text{otherwise.} \end{cases}$$

Third, B_1 describes the transitions from level 0 to level 1 through the arrival of a class 2 demand when inventory is at or below the critical level c and is an $(S+1) \times (c+1)$ matrix:

$$B_1(i,j) = \begin{cases} \lambda_2 & \text{if } 1 \leq i \leq c+1 \text{ and } j = i; \\ 0 & \text{otherwise.} \end{cases}$$

For levels 1 and up, $A_0(n)$ describes the transitions within level n and is an $(c+1) \times (c+1)$

matrix:

$$A_0(n)(i,j) = \begin{cases} -((S+n)\mu + \lambda_2) & \text{if } i = j = 1; \\ -((S+n-i+1)\mu + \lambda) & \text{if } 1 < i \leq c+1 \text{ and } j = i; \\ (S+n-i+1)\mu & \text{if } 1 \leq i < c+1 \text{ and } j = i+1; \\ \lambda_1 & \text{if } 1 < i \leq c+1 \text{ and } j = i-1; \\ 0 & \text{otherwise.} \end{cases}$$

Fifth, the $(c+1) \times (c+1)$ matrix $A_{-1}(n)$ captures the transitions, by arriving replenishment orders, from level n to level $n-1$:

$$A_{-1}(n)(i,j) = \begin{cases} (S-c+n)\mu & \text{if } i = j = c+1; \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the $(c+1) \times (c+1)$ matrix A_1 describes the transitions from level n to level $n+1$ by the arrival of class 2 demands when inventory is at or below the critical level:

$$A_1(n)(i,j) = \begin{cases} \lambda_2 & \text{if } 1 \leq i \leq c+1 \text{ and } j = i; \\ 0 & \text{otherwise.} \end{cases}$$

A.2. Proofs of lemmas and theorems

A.2.1 Proof of Lemma 1

Consider the balance equations (3.7)–(3.10). For $n \geq 1$ the third term of the left-hand side of (3.8) and (3.9) can be written as follows:

$$\tilde{\boldsymbol{\pi}}_{n+1} A_{-1}(n+1) = (0, \ldots, 0, \tilde{\pi}_{c,n+1}(S-c+n+1)\mu) \tag{A.1}$$

As can be seen from Figure 3.2 this captures that the only flow from level $n+1$ to level n is from $(c, n+1)$ to (c, n). Now, due to global balance the following relation must hold:

$$\tilde{\pi}_{c,n+1}(S-c+n+1)\mu \quad = \quad \tilde{\boldsymbol{\pi}}_n A_1 \boldsymbol{e} \quad = \quad \lambda_2 \tilde{\boldsymbol{\pi}}_n \boldsymbol{e}, \tag{A.2}$$

where $\boldsymbol{e}$ the vector of ones of the appropriate size. Then from (A.1) and (A.2) we get:

$$\boldsymbol{\pi}_{n+1} A_{-1}(n+1) = \lambda_2 \boldsymbol{\pi}_n A, \tag{A.3}$$

where A is defined in (3.11). Substitution of (A.3) into (3.8), and (3.9) leads to:

$$\tilde{\boldsymbol{\pi}}_0 B_0 + \tilde{\boldsymbol{\pi}}_1 B_{-1} = 0 \qquad \text{for } n = 0 \tag{A.4}$$

$$\tilde{\boldsymbol{\pi}}_0 B_1 + \tilde{\boldsymbol{\pi}}_1 A_0(1) + \lambda_2 \tilde{\boldsymbol{\pi}}_1 A = 0 \qquad \text{for } n = 1 \tag{A.5}$$

$$\tilde{\boldsymbol{\pi}}_{n-1} A_1 + \tilde{\boldsymbol{\pi}}_n A_0(n) + \lambda_2 \tilde{\boldsymbol{\pi}}_n A = 0 \qquad \text{for } n \geq 2 \tag{A.6}$$

$$\tilde{\pi}_{S,0} = 1. \tag{A.7}$$

Equations (A.4) through (A.7) have a unique solution.

Now $\tilde{\boldsymbol{\pi}}_0$ and $\tilde{\boldsymbol{\pi}}_1$ can be obtained from equations (A.4) and (A.5) with $\tilde{\pi}_{S,0} = 1$, and the vectors $\tilde{\boldsymbol{\pi}}_n$ for $n \geq 2$ can be recursively calculated from (A.6) which can be rewritten as follows:

$$\boldsymbol{\pi}_n = -\boldsymbol{\pi}_{n-1} A_1 (A_0(n) + A\lambda_2)^{-1} \quad \text{for } n \geq 2,$$

where $(A_0(n) + A\lambda_2)^{-1}$ exists, as it is a transient generator. □

A.2.2 Proof of Lemma 2

For the proof of Lemma 2 we first introduce diagonal levels for $n \geq 0$ defined as the set of states: $\{(0,n), (1, n+1), \ldots, (c, n+c)\}$ and the corresponding probability vectors $\widetilde{\boldsymbol{\delta}}_n = (\tilde{\pi}_{0,n}, \tilde{\pi}_{1,n+1}, \ \ldots, \tilde{\pi}_{c,n+c})$. The balance flow between two subsequent diagonal levels can be expressed as:

$$\widetilde{\boldsymbol{\delta}}_n \boldsymbol{e} \lambda - \tilde{\pi}_{0,n} \lambda_1 = \widetilde{\boldsymbol{\delta}}_{n+1} \boldsymbol{e} (S + n + 1)\mu. \tag{A.8}$$

So, leaving out the second term on the left-hand side results in:

$$\widetilde{\boldsymbol{\delta}}_n \boldsymbol{e} \lambda \geq \widetilde{\boldsymbol{\delta}}_{n+1} \boldsymbol{e} (S + n + 1)\mu. \tag{A.9}$$

and thus the probability of being at diagonal level $n+1$ can be expressed in the probability of being at diagonal level n as follows:

$$\widetilde{\boldsymbol{\delta}}_{n+1} \boldsymbol{e} \leq \frac{\lambda}{\mu} \frac{1}{S + n + 1} \widetilde{\boldsymbol{\delta}}_n \boldsymbol{e}.$$

Now we can bound the weighted probabilities using a cut off parameter $\ell \geq 1$. For horizontal levels $n \geq c + \ell$ we know the weighted probabilities are upper bounded by the weighted probabilities of the diagonal layers. This works because the lowest diagonal layer, which gets the same weight $((c + \ell)$ as the lowest horizontal layer includes states below the lowest horizontal bounding level, and the mass decreases in the level. Furthermore the weight assigned to each state under the diagonal layer definition is at least as large as under the horizontal definition. The weighted diagonal layer can then bound the weighted horizontal levels as follows:

$$\begin{aligned} \textstyle\sum_{n=c+\ell}^{\infty} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e} &\leq \textstyle\sum_{n=\ell}^{\infty}(n+c)\widetilde{\boldsymbol{\delta}}_n \boldsymbol{e} \\ &= \textstyle\sum_{k=0}^{\infty}(k+\ell+c)\widetilde{\boldsymbol{\delta}}_{\ell+k}\boldsymbol{e}. \end{aligned}$$

Using the following result for the relation between two diagonal levels at distance k

$$\begin{aligned} \widetilde{\boldsymbol{\delta}}_{\ell+k}\boldsymbol{e} &\leq \left(\tfrac{\lambda}{\mu}\right)^{k} \tfrac{1}{(S+\ell+k)...(S+\ell+1)}\widetilde{\boldsymbol{\delta}}_{\ell}\boldsymbol{e} \\ &= \left(\tfrac{\lambda}{\mu}\right)^{k} \tfrac{(S+\ell)!}{(S+\ell+k)!}\widetilde{\boldsymbol{\delta}}_{\ell}\boldsymbol{e} \end{aligned}$$

we can bound $\sum_{n=c+\ell}^{\infty} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e}$ as follows:

$$\begin{aligned} \textstyle\sum_{n=c+\ell}^{\infty} n\tilde{\boldsymbol{\pi}}_n \boldsymbol{e} &\leq \textstyle\sum_{k=0}^{\infty} \left(\tfrac{\lambda}{\mu}\right)^{k} \tfrac{(S+\ell)!}{(S+\ell+k)!}\widetilde{\boldsymbol{\delta}}_{\ell}\boldsymbol{e}(k+\ell+c) \\ &= \widetilde{\boldsymbol{\delta}}_{\ell}\boldsymbol{e}(S+\ell)!\left(\tfrac{\mu}{\lambda}\right)^{S+\ell}\left[\textstyle\sum_{k=0}^{\infty}\left(\tfrac{\lambda}{\mu}\right)^{S+\ell+k}\tfrac{1}{(S+\ell+k)!}(k+\ell+c)\right] \\ &= \widetilde{\boldsymbol{\delta}}_{\ell}\boldsymbol{e}(S+\ell)!\left(\tfrac{\mu}{\lambda}\right)^{S+\ell}\left[\textstyle\sum_{k=0}^{\infty}\left(\tfrac{\lambda}{\mu}\right)^{S+\ell+k}\tfrac{1}{(S+\ell+k)!}(k+S+\ell)-(S-c)\textstyle\sum_{k=0}^{\infty}\left(\tfrac{\lambda}{\mu}\right)^{S+\ell+k}\tfrac{1}{(S+\ell+k)!}\right] \\ &= \widetilde{\boldsymbol{\delta}}_{\ell}\boldsymbol{e}(S+\ell)!\left(\tfrac{\mu}{\lambda}\right)^{S+\ell}\left[\tfrac{\lambda}{\mu}\phi(S+\ell-1)-(S-c)\phi(S+\ell)\right] \end{aligned}$$

where $\phi(\ell)$ is as defined in (3.13). □

A.2.3 Proof of Theorem 1

Theorem 1 has been stated formulated for Poisson arrivals and exponentially distributed lead times. We will prove Theorem 1 for general arrivals and for both exponential (§A.2.3) and degenerate hyperexponential (§A.2.3) lead times.

Exponential lead times

Both class 1 and class 2 orders arrive according to an arbitrary arrival process; let t_n denote the n-th arrival time of an order and i_n indicates whether it is an arrival of a class 1 ($i_n = 1$) or class 2 ($i_n = 2$) order. It is assumed that the sequence t_n satisfies $0 < t_1 < t_2 < \cdots$ (thus only single arrivals) and that $t_n \to \infty$ as $n \to \infty$.

The assumption that lead times are exponentially distributed allows us to sample new lead times for all items in the pipeline immediately after each arrival; let $s_{j,n}$ be the j-th lead time just after t_n, where lead times are ordered such that orders that are outstanding in both systems appear first, and those that are outstanding in only one system appear later. Further, let $m^c(t)$ denote the number of items on hand, $n^c(t)$ the number of backorders and $x^c(t)$ the number of items in the pipeline at time t in the system with critical level c; note that $m^c(t)$, $n^c(t)$ and $x^c(t)$ are step functions (with steps of size 1) and we assume these functions are right-continuous (so the number at time t is the same as the number just after time t).

Using this notation we will prove that, on the same sample path, for all $t \geq 0$, the performance measures depend on c in the following manner:

$$n^c(t) \leq n^{c+1}(t) \tag{A.10}$$

$$x^c(t) \leq x^{c+1}(t) \tag{A.11}$$

Note that this will give us that the number of backorders and the pipeline inventory are stochastically increasing in c, which is actually stronger than just the monotonic increase in the means. To prove the above relations, we fix a sample path of arrivals to both systems. Replenishment orders that are common to both systems are also coupled, i.e. we couple the resampled lead times in both systems. After every customer arrival we sample $\max(x^c(t), x^{c+1}(t))$, we then assign the first $\min(x^c(t), x^{c+1}(t))$ to both systems. The remainder is assigned *only* to the system with the highest number of outstanding orders. Hence the sequences remain coupled. The orders in the pipeline are indexed by j in the order in which they are assigned. As there may be more outstanding orders in one system than in another, these additional replenishment orders are sampled separately.

At time $t = 0$ we assume that the on-hand inventory is S, there are no backorders and the pipeline is empty, so $m^c(0) = S$ and $n^c(0) = x^c(0) = 0$.

Clearly, for all $t \geq 0$,

$$m^c(t) = S - (x^c(t) - n^c(t)), \tag{A.12}$$

and the CL policy implies that $n^c(t) = 0$ if $m^c(t) > c$. Let $m(t^-)$ denote the stock level just before t, i.e.,

$$m(t^-) = \lim_{s \uparrow t} m(s),$$

and $1_{[A]}$ the indicator which is 1 if A holds and 0 otherwise.

By induction we will prove that (A.10)-(A.11) hold for $[0, t_n)$ for all $n \geq 1$. Since $m^c(0) = m^{c+1}(0) = S$ and $n^c(0) = n^{c+1}(0) = x^c(0) = x^{c+1}(0) = 0$ and there are no events during $[0, t_1)$ (since the pipeline is empty), it follows that (A.10)-(A.11) hold for $t \in [0, t_1)$. Now assume that (A.10)-(A.11) are valid for $[0, t_n)$, so just before t_n,

$$n^c(t_n^-) \leq n^{c+1}(t_n^-), \tag{A.13}$$

$$x^c(t_n^-) \leq x^{c+1}(t_n^-). \tag{A.14}$$

Then we will show that (A.10)-(A.11) remain valid during $[t_n, t_{n+1})$. At time t_n a new demand arrives. If $i_n = 1$:

$$n^c(t_n) = n^c(t_n^-) \leq n^{c+1}(t_n^-) = n^{c+1}(t_n),$$

so (A.10) is still valid for t_n. For $x^c(t_n)$ we have

$$x^c(t_n) = x^c(t_n^-) + 1_{[m^c(t_n^-)>0]} = x^c(t_n^-) + 1_{[x^c(t_n^-)-n^c(t_n^-)<S]}.$$

If $x^c(t_n^-) < x^{c+1}(t_n^-)$, then clearly, (A.11) is valid for t_n and if $x^c(t_n^-) = x^{c+1}(t_n^-)$, then by (A.13)

$$x^c(t_n) = x^c(t_n^-) + 1_{[x^c(t_n^-)-n^c(t_n^-)<S]} \leq x^{c+1}(t_n^-) + 1_{[x^{c+1}(t_n^-)-n^{c+1}(t_n^-)<S]} = x^{c+1}(t_n). \tag{A.15}$$

If $i_n = 2$: By (A.14)

$$x^c(t_n) = x^c(t_n^-) + 1 \leq x^{c+1}(t_n^-) + 1 = x^{c+1}(t_n), \tag{A.16}$$

so (A.11) is still valid for t_n. For $n^c(t_n)$ we have

$$n^c(t_n) = n^c(t_n^-) + 1_{[m^c(t_n^-)\leq c]} = n^c(t_n^-) + 1_{[S-x^c(t_n^-)+n^c(t_n^-)\leq c]}.$$

If $n^c(t_n^-) < n^{c+1}(t_n^-)$, then clearly, (A.10) is valid for t_n and if $n^c(t_n^-) = n^{c+1}(t_n^-)$, then by (A.14)

$$n^c(t_n) = n^c(t_n^-) + 1_{[S-x^c(t_n^-)+n^c(t_n^-)\leq c]} \leq n^{c+1}(t_n^-) + 1_{[S-x^{c+1}(t_n^-)+n^{c+1}(t_n^-)\leq c+1]} = n^{c+1}(t_n).$$

Hence, (A.10)-(A.11) are valid at time t_n.

Now we will show that (A.10)-(A.11) remain valid on (t_n, t_{n+1}) for both cases ($i_n = 1$ and $i_n = 2$). Assume that $u_{j,n} := t_n + s_{j,n} < t_{n+1}$, i.e., the jth replenishment arrives before

t_{n+1}. First suppose $j \leq \min(x^c(t_n), x^{c+1}(t_n)) = x^c(t_n)$, thus we have an arrival in both systems c and $c+1$, and further assume

$$\begin{aligned} n^c(u_{j,n}^-) &\leq n^{c+1}(u_{j,n}^-), \\ x^c(u_{j,n}^-) &\leq x^{c+1}(u_{j,n}^-). \end{aligned} \tag{A.17}$$

Then we will show that (A.10)-(A.11) remain valid at $u_{j,n}$ (thus the arrival preserves (A.10)-(A.11)). Clearly

$$x^c(u_{j,n}) = x^c(u_{j,n}^-) - 1 \leq x^{c+1}(u_{j,n}^-) - 1 = x^{c+1}(u_{j,n}),$$

so (A.11) is still valid for $u_{j,n}$. For $n^c(u_{j,n})$ we have, provided $n^c(u_{j,n}^-) > 0$,

$$n^c(u_{j,n}) = n^c(u_{j,n}^-) - 1_{[m^c(u_{i,n}^-) \geq c]} = n^c(u_{j,n}^-) - 1_{[S - x^c(u_{j,n}^-) + n^c(u_{j,n}^-) \geq c]}.$$

If $n^c(u_{j,n}^-) = 0$ or $n^c(u_{j,n}^-) < n^{c+1}(u_{j,n}^-)$, then clearly, (A.10) is valid for $u_{j,n}$ and if $0 < n^c(u_{j,n}^-) = n^{c+1}(u_{j,n}^-)$, then by (A.17):

$$n^c(u_{j,n}) = n^c(u_{j,n}^-) - 1_{[S - x^c(u_{j,n}^-) + n^c(u_{j,n}^-) \geq c]} \leq n^{c+1}(u_{j,n}^-) - 1_{[S - x^{c+1}(u_{j,n}^-) + n^{c+1}(u_{j,n}^-) \geq c+1]} = n^{c+1}(u_{j,n}).$$

Now suppose we only have an arrival of a replenishment order in the $c+1$ system as $\min(x^c(t_n), x^{c+1}(t_n)) = x^c(t_n) < j \leq x^{c+1}(t_n) = \max(x^c(t_n), x^{c+1}(t_n))$. This implies $x^c(u_{j,n}^-) < x^{c+1}(u_{j,n}^-)$ and thus (A.11) holds for $u_{j,n}$. Again, if $n^c(u_{j,n}^-) = 0$ or $n^c(u_{j,n}^-) < n^{c+1}(u_{j,n}^-)$, then clearly, (A.10) is valid for $u_{j,n}$. If $0 < n^c(u_{j,n}^-) = n^{c+1}(u_{j,n}^-)$, then, since $n^c(u_{j,n}) = n^c(u_{j,n}^-)$, we have to show that also $n^{c+1}(\cdot)$ does not decrease, i.e., $m^{c+1}(u_{j,n}^-) < c+1$. It holds

$$m^{c+1}(u_{j,n}^-) = S - x^{c+1}(u_{j,n}^-) + n^{c+1}(u_{j,n}^-) \leq S - x^c(u_{j,n}^-) + n^c(u_{j,n}^-) = m^c(u_{j,n}^-) \leq c,$$

where the last inequality follows from $n^c(u_{j,n}^-) > 0$. This completes the proof of (A.10)-(A.11) as, by induction, these relationships hold for all t.

(3.14) remains to be proven. As class 2 customers only get served when there are no backorders, i.e. $n^c(t) = 0$, and $c < m^c(t)$ we focus on times where $n^c(t) = 0$. By (A.11) and (A.12), we have that

$$m^c(t) = S - x^c(t) + 0 \geq S - x^{c+1}(t) + 0 = m^{c+1}(t)$$

Consider first the specific inventory levels

$$c + 1 \leq m^{c+1}(t) \leq m^c(t),$$

and note that when the first inequality holds at equality the system with critical level $c+1$ does not serve class 2 demand, while the system with critical level c does. When the first inequality is strict, both systems serve class 2 demand. Hence, the fraction of class 2 demand satisfied from inventory is larger under the system with critical level c. When $m^{c+1}(t) < c+1$ class 2 customers will not be served in the $c+1$ system but *may* be served in the c system and again the fraction of of class 2 demand satisfied from inventory is larger under the system with critical level c. □

Remark 2 *The relation between $m^c(t)$ and $m^{c+1}(t)$ is not monotonic in c as is illustrated by the following example with $S = 2$, $c = 0$, $t_1 = 1$, $i_1 = 2$, $t_2 = 2$, $i_2 = 2$, $t_3 = 3$ and $i_3 = 1$. Replenishments arrive at $t = 4$ and $t = 5$. This will lead to the following system as outlined in Table 2. This shows that $m^c(t) < m^{c+1}(t)$ (at $t = 2$) as well as $m^c(t) > m^{c+1}(t)$ (at $t = 5$) may happen.*

$t = 0$	$m^c(0) = 2 = m^{c+1}(0)$
	$n^c(0) = 0 = n^{c+1}(0)$
	$x^c(0) = 0 = x^{c+1}(0)$
$t = 1$	$m^c(1) = 1 = c + 1 = m^{c+1}(1)$
	$n^c(1) = 0 = n^{c+1}(1)$
	$x^c(1) = 1 = x^{c+1}(1)$
$t = 2$	$m^c(2) = 0 < 1 = m^{c+1}(2)$
	$n^c(2) = 0 < 1 = n^{c+1}(2)$
	$x^c(2) = 2 = x^{c+1}(2)$
$t = 3$	$m^c(3) = 0 = m^{c+1}(3)$
	$n^c(3) = 0 < 1 = n^{c+1}(3)$
	$x^c(3) = 2 < 3 = x^{c+1}(3)$
$t = 4$	$m^c(4) = 1 = m^{c+1}(4)$
	$n^c(4) = 0 < 1 = n^{c+1}(4)$
	$x^c(4) = 1 < 2 = x^{c+1}(4)$
$t = 5$	$m^c(5) = 2 > 1 = m^{c+1}(5)$
	$n^c(5) = 0 = n^{c+1}(5)$
	$x^c(5) = 0 < 1 = x^{c+1}(5)$

Table A.1: Example of potential system evolution for $S = 2$, $c = 0$

Degenerate Hyperexponential lead times

Definition 1 *A sample from a degenerate hyperexponential distribution is drawn from an exponential distribution with rate μ^* with probability p and is 0 with probability $1 - p$.*

We will now show that, with degenerate hyperexponential distributed lead times, for all $t \geq 0$, the performance measures depend on c in the following manner:

$$n^c(t) \leq n^{c+1}(t) \tag{A.18}$$

$$x^c(t) \leq x^{c+1}(t) \tag{A.19}$$

This proof follows along the same lines as the proof of (A.10)-(A.11). First we consider the points in time when a replenishment order is placed. Let $1_{[LT>0]}$ be 1 with probability p (i.e. the lead time is to be drawn from the exponential with rate μ^*), and 0 otherwise.

Then we modify (A.15) and (A.16) as follows:

$$
\begin{aligned}
x^c(t_n) = x^c(t_n^-) + 1_{[LT>0]}1_{[x^c(t_n^-)-n^c(t_n^-)<S]} &\leq x^{c+1}(t_n^-) + 1_{[LT>0]}1_{[x^{c+1}(t_n^-)-n^{c+1}(t_n^-)<S]} = x^{c+1}(t_n) \\
x^c(t_n) = x^c(t_n^-) + 1_{[LT>0]}1 &\leq x^{c+1}(t_n^-) + 1_{[LT>0]}1 = x^{c+1}(t_n),
\end{aligned}
$$

As the lead time that is drawn is the same in both system, we know (A.18)-(A.19) are valid at t_n (the argument for (A.18) did not change). To show that (A.18)-(A.19) remain valid on (t_n, t_{n+1}) we need to realize that only replenishment orders with non-zero lead times actually entered the pipeline. Focussing on these replenishment orders only the arguments from the proof of §A.2.3 are still valid. Which proves (A.18)-(A.19).

(3.14) for the degenerate hyperexponential case follows directly from the argument for (3.14) in §A.2.3. □

A.2.4 Proof of Lemma 3

The cost of a certain policy (see equation (3.1)) can be bounded from below as follows:

$$
\begin{aligned}
C(S,c) &= p_1\lambda_1(1-\beta_1(S,c)) + p_2\lambda_2(1-\beta_2(S,c)) + bB(S,c) + hI(S,c) \\
&= p_1\lambda_1(1-\beta_1(S,c)) + p_2\lambda_2(1-\beta_2(S,c)) + (b+h)B(S,c) + h\left(I(S,c) - B(S,c)\right) \\
&\geq p_2\lambda_2(1-\beta_2(S,c)) + (b+h)B(S,c) + h\left(I(S,c) - B(S,c)\right) \\
&= p_2\lambda_2(1-\beta_2(S,c)) + (b+h)B(S,c) + h\left(S - X(S,c)\right) \\
&\geq p_2\lambda_2(1-\beta_2(S,c)) + (b+h)B(S,c) + h\left(S - X(S,S)\right) \qquad \text{(A.20)} \\
&\geq p_2\lambda_2(1-\beta_2(S,c)) + (b+h)B(S,c) + h\left(S - \frac{\lambda}{\mu}\right), \qquad \text{(A.21)}
\end{aligned}
$$

where (A.21) follows from (A.20) by the monotonicity results in Theorem 1.

A.3. Markov Decision Process details

Since all events happen with exponentially distributed interarrival times it is sufficient to only look at the states of the system when events occur.

A.3.1 States, Events, Decisions, Transitions

The state of the system can be fully specified by:

- The amount of inventory on hand (I);
- The number of backorders (B, note that backorders may exist even if there is inventory).
- The number of items on order (DI, information about when an order is placed is not needed as long as exponential lead times are used);
- The last event that occurred (E, 1 if the event was such that in the current state orders can be placed, i.e. a demand was satisfied or backordered, 0 else, i.e. a demand was rejected or a replenishment arrived).

The state space is denoted by a four-tuple: (I, B, DI, E). In the MDP formulation we make three additional assumptions to bound the state space and thus also the maximum transition rate from any state. When solving the MDP we will ensure that these bounds have insignificant effect on the optimal solution by solving the MDP repeatedly with increasing bounds, as soon as the total probability mass in all boundary states drops below a threshold the effect of bounding becomes negligible. *i)* We assume there exists a maximum inventory level $\hat{I}$, *ii)* we assume there exists a maximum number of backorders $\hat{B}$, and *iii)* we assume that as soon as the number of backorders equals $\hat{B}$ no further class 2 demands will arrive.

Given the three bounding assumptions we get bounds: $0 \leq I \leq \hat{I}$, $0 \leq B \leq \hat{B}$, and subsequently $0 \leq DI \leq \hat{I} - I + \hat{B}$, effectively: $0 \leq DI \leq \hat{I} - I + B$. This second bound on the amount DI is because whenever all replenishment orders arrive before the next demand one must still be inside the state space.

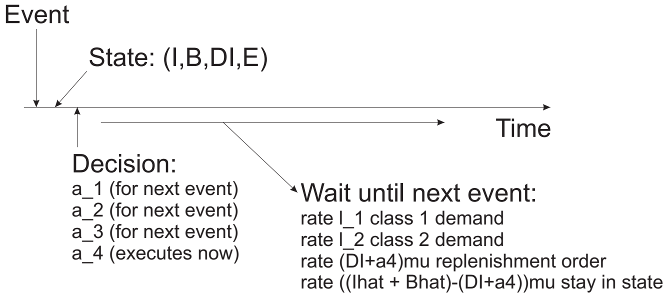


Figure A.1: Illustration of the sequence of events

We consider the following decisions at every event:

- $a_1 = 1$: Satisfy class 1 demand as long as inventory is available ($I = 0 \Rightarrow a_1 = 0$ since you have no inventory to act otherwise).
- $a_2 \in \{0, 1\}$: If a class 2 demand occurs whether to serve it or backorder it. If $a_2 = 0$ you would reject (i.e. backorder) the class 2 demand and if $a_2 = 1$ you serve the demand. ($I = 0 \Rightarrow a_2 = 0$ since you have no inventory to act otherwise).
- $a_3 \in \{0, 1\}$: How to use an incoming replenishment order. If $a_3 = 0$ any incoming item is added to inventory and if $a_3 = 1$ a backorder is cleared.
- $a_4 \in \{0, 1\}$: How much to order. This order immediately becomes effective and determines the current size of the pipeline.

We are aware of the fact that ordering at most 1 item at a time may lead to a suboptimal solution. Since we are mainly interested in how our critical level policy deals with class 2 demand we believe that this restrictions do not limit our insights in this repect.

The timeline, seen in Figure A.1, is as follows: Suppose the state of the system is changed by the occurrence of an event (a class 1 demand occurs and it is served, a class 2 demand occurs, or a replenishment order arrives), at (after) the occurrence of an event (i.e. when one has the knowledge of which event happened) one *has* to make a decision consisting of an action tuple of size 4, $\{a_1, a_2, a_3, a_4\}$ [1]. The first three types of decisions become effective at the occurrence of the next event (i.e. they describe how to handle the next event). The fourth decision is the decision how much to order, which is immediately implemented and affects the number of items in the pipeline.

Now that we have specified the available actions we consider the transitions. These depend on both the event occurring and the action taken:

- At rate λ_1 a demand from class one arrives.
  - If $a_1 = 1$, i.e. the demand is served, the transition goes to $(I-1, B, DI+a_4, 1)$, and
  - if $a_1 = 0$, i.e. the demand is rejected, the transition goes to $(I, B, DI, 0)$, this is effectively a fake transition, happening at rate λ_1 if $a_1 = 0$ (only if $I = 0$).
- At rate λ_2 a demand from class two arrives (when $b < \hat{B}$).
  - If $a_2 = 1$, i.e. the demand is served, the transition goes to $(I-1, B, DI+a_4, 1)$, and
  - if $a_2 = 0$, i.e. the demand is backordered, the transition goes to $(I, B+1, DI+a_4, 1)$.
- At rate $(DI+a_4)\mu$ a replenishment order arrives. When a replenishment order arrives one has to decide how to use the incoming item. Depending on this decision (a_3) the transition goes to $(I+1-a_3, B-a_3, DI-1+a_4, 0)$.
- At rate $[(\hat{I}+\hat{B})-(DI+a_4)]\mu$= $(\hat{I}+\hat{B}-DI-a_4)\mu$ a "fake" transition will happen, this transition goes back to our current state (I, B, DI, E).

[1] When the event did not change the state of the system (e.g. reject a class 1 customer) the decision made at the previous event will remain optimal after the current event because events follow a (memoryless) exponential distribution.

A.3.2 The LP formulation

Now we have formulated the MDP we use linear programming to determine the optimal cost. At this point the bounding of the state space and the bounding of the transition rate out of any state enable us to uniformize (Ross 1997) using the "fake" transitions introduced in the previous section.

The LP formulation is based on Sections 8.8.1 and 11.4.4 from Puterman (1994) . Throughout the LP the variables are such that for each (state, action) combination there is a variable, denoted by $\pi(I, B, DI, E, a_1, a_2, a_3, a_4)$, which denotes the steady state distribution.

Objective function

The objective function consists of (A.22a) for the holding cost, (A.22b) for the penalty cost for rejecting a class 1 demand, backordering cost p_2 are incurred in those states summed in (A.22c), i.e. when a class 2 demand is backordered, and finally the cost b are incurred in all states where the number of backorders exceeds 0, see (A.22d).

$$min \quad \sum_{I=1}^{\hat{I}} \sum_{B=0}^{\hat{B}} \sum_{DI=0}^{\hat{I}-I+B} \sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} hI\pi(I, B, DI, E, a_1, a_2, a_3, a_4) \tag{A.22a}$$

$$+ \quad \sum_{I=0}^{\hat{I}} \sum_{B=0}^{\hat{B}} \sum_{DI=0}^{\hat{I}-I+B} \sum_{E=0}^{1} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} p_1\lambda_1\pi(I, B, DI, E, 0, a_2, a_3, a_4) \tag{A.22b}$$

$$+ \quad \sum_{I=0}^{\hat{I}} \sum_{B=0}^{\hat{B}} \sum_{DI=0}^{\hat{I}-I+B} \sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} p_2\lambda_2\pi(I, B, DI, E, a_1, 0, a_3, a_4) \tag{A.22c}$$

$$+ \quad \sum_{I=0}^{\hat{I}} \sum_{B=1}^{\hat{B}} \sum_{DI=0}^{\hat{I}-I+B} \sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} bB\pi(I, B, DI, E, a_1, a_2, a_3, a_4) \tag{A.22d}$$

Constraints

A constraint balance equation is needed for each state, (I, B, DI, E), and is of the form $OUT - IN = 0$. We use s and a to denote the state and action tuples for brevity. Here we describe the constraint for $s = (I, B, DI, E)$, assuming this is a state in the interior of the state space. Defining S as the set of states, now let $\Gamma(s, a) = \sum_{j \in S} \gamma(j|s, a)$; Γ denotes the total rate out of state s whenever action a is taken and $\gamma(j|s, a)$ is the probability of going from state s to state j whenever action a is taken. Due to the uniformization we know that $\Gamma(s, a) = \lambda_1 + \lambda_2 + (\hat{I} + \hat{B})\mu$.

The "OUT" part of the constraint would be as follows:

$$\sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} \Gamma(I, B, DI, E, a_1, a_2, a_3, a_4)\pi(I, B, DI, E, a_1, a_2, a_3, a_4) \tag{A.23}$$

Here $\Gamma(I, B, DI, E, a_1, a_2, a_3, a_4)$ consists of:

- λ_1 if $a_1 = 1$, and $I > 0$, since $I = 0 \Rightarrow a_1 = 0$;
- λ_2 if $I > 0$, and $B < \hat{B}$ (interior of state space), or $I > 0$, and $B = \hat{B}$, and $a_2 = 1$, or $I = 0$, and $B < \hat{B}$, where the later two are for the boundary states;
- $(DI + a_4)\mu$, the speed at which a replenishment will come in.
- $\lambda_1 + \lambda_2 + (\hat{I} + \hat{B})\mu - (\lambda_1 \mathcal{I}_{\{a_1=1 \wedge I>0\}} + \lambda_2 \mathcal{I}_{\{(I>0 \wedge B<\hat{B}) \vee (I>0 \wedge B=\hat{B} \wedge a_2=1) \vee (I=0 \wedge B<\hat{B})\}} + (DI + a_4)\mu)$, the fake transitions.

where $\mathcal{I}_{condition} = 1$ if the condition is true and 0 otherwise.

The "IN" constraints are more complicated and describe all possible ways to enter state (I, B, DI, E). They consist of 3 parts. Part (A.24a) deals with the arrival of a class 1 demand. Part (A.24b) handles the arrival of class 2 demand. Part (A.24c) refers to getting an item delivered and adding it to inventory ($a_3 = 0$), or using the incoming item to clear a backorder ($a_3 = 1$). The last term of each of the lines below denotes the rate at which you are going into our state s under consideration from the state in the summation, note that this rate is action independent. Recall that we are looking at states from which you will get **in**to state (I, B, DI, E), (note $0 \leq DI \leq \hat{I} - I + B$), then:

$$\sum_{E=0}^{1} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} \mathcal{I}_{(I+1,B,DI-a_4,E) \in S} \pi(I+1, B, DI-a_4, E, 1, a_2, a_3, a_4)\lambda_1 \tag{A.24a}$$

$$+ \sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} \mathcal{I}_{(I+a_2,B-1+a_2,DI-a_4,E) \in S} \pi(I+a_2, B-1+a_2, DI-a_4, E, a_1, a_2, a_3, a_4)\lambda_2 \tag{A.24b}$$

$$+ \sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} \mathcal{I}_{(I+a_3-1,B+a_3,DI+1-a_4,E) \in S} \pi(I+a_3-1, B+a_3, DI+1-a_4, E, a_1, a_2, a_3, a_4)(DI+1)\mu \tag{A.24c}$$

Furthermore, the fake transitions only lead into states $(I, B, DI, 0)$, as no ordering is allowed in a fake transition:

$$\sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} \mathcal{I}_{(I,B,DI-a_4,E) \in S} \pi(I, B, DI-a_4, E, 1, a_2, a_3, a_4)\Big(\lambda_1 + \lambda_2 + (\hat{I} + \hat{B})\mu - (\lambda_1 \mathcal{I}_{\{a_1=1 \wedge I>0\}} + \lambda_2 \mathcal{I}_{\{(I>0 \wedge B<\hat{B}) \vee (I>0 \wedge B=\hat{B} \wedge a_2=1) \vee (I=0 \wedge B<\hat{B})\}} + (DI+a_4)\mu)\mathcal{I}_{E=1}\Big) \tag{A.25}$$

This indicator function, $\mathcal{I}_{condition}$ is used to effectively truncate the state space. This leads to the constraint:

$$(\text{A.23}) - (\text{A.24a}) - (\text{A.24b}) - (\text{A.24c}) - (\text{A.25}) = 0 \tag{A.26}$$

Furthermore we need a "normalization" constraint:

$$\sum_{I=0}^{\hat{I}} \sum_{B=0}^{\hat{B}} \sum_{DI=0}^{\hat{I}-I+B} \sum_{E=0}^{1} \sum_{a_1 \in \{0,1\}} \sum_{a_2 \in \{0,1\}} \sum_{a_3 \in \{0,1\}} \sum_{a_4 \in \{0,1\}} \pi(I, B, DI, E, a_1, a_2, a_3) = 1. \qquad \text{(A.27)}$$

To find the optimal solution we solve the linear program outlined above using the CPLEX barrier method. To determine whether the truncation of our state space impacts our optimal solution, we track the probability mass in the boundary states. Whenever the probability mass exceeds 10^{-6} we increase the size of our state space. Since our problem is highly degenerate and the barrier method returns the solution with the largest support we implemented a second phase in which we minimize the probability mass in the boundary states while not moving away from the optimal cost.

A.4. Modifications for degenerate hyperexponential lead times

By setting p and μ^* as defined in Definition 1 (§A.2.3) properly, this distribution can match the first and second moment of any distribution with $C^2 \geq 1$: $p = 1 - \frac{C^2-1}{C^2+1}$ and $\mu^* = p\mu$.

A.4.1 Modifications in the evaluation of a given policy

In the behavior of the policy there are some changes when the lead times follow the degenerate hyperexponential distribution. The main difference is that some of the replenishment orders get a zero lead time and thus arrive immediately. The modifications to the Markov process in Figure 3.2 are the following:

- The rate of all transitions from (m, n) to $(m-1, n)$ gets multiplied by p,
- The transition from (m, n) to $(m, n+1)$ for $m < c$ gets replaced by two transitions, one to $(m, n+1)$ at rate $p\lambda_2$ and one to $(m+1, n+1)$ at rate $(1-p)\lambda_2$,
- The rate of the transitions from (c, n) to $(c, n+1)$ gets multiplied by p
- The replenishment rate μ gets replaced by μ^* throughout.

The above modifications change the matrices outlined in Appendix A.1 and thus the generator Q (3.6). Since our special structure is still maintained, the solution procedure remains the same. In the bounding procedure as outlined in Lemma 2 all λ values get multiplied by p and μ gets replaced by μ^*.

A.4.2 Modifications in the optimization

Since the monotonicity results do not depend on the specific structure of our Markov process (see Appendix A.2.3), these still hold and can be applied in the development of the bounds for the enumeration. Again λ should be replace by $p\lambda$ and μ by μ^*.

Appendix B

Ambulance Traffic Coordination

B.1. Table of notation with chapter 4

B_{jED}	Number of regular ED beds in hospital j
B_{jSED}	Number of surge ED beds in hospital j
B_{jI}	Number of inpatient beds in hospital j
n_{jH} (n_{jL})	Number of high (low) acuity patients at hospital j
n_{jI}	Number of inpatients either at the backbone hospital or boarding at hospital j
n_{jB}	Number of patients boarding in the ED of hospital j
$\bar{n}_j$	Waiting room size at hospital j
λ_{jA}	Ambulance arrival rate to hospital j
λ_{jW}	Walk In arrival rate to hospital j
λ_{jI}	Arrival rate to the ID of hospital j
λ_{jAH} (λ_{jAL})	Ambulance arrival rate of high (low) acuity patients to hospital j
λ_{jWH} (λ_{jWL})	Walk-In arrival rate of high (low) acuity patients to hospital j
Λ_j^H (Λ_j^L)	Net arrival rate of high (low) acuity patients to hospital j
P_{jAH}	Probability that an ambulance arrival to hospital j is triaged as high acuity patient
P_{jWH}	Probability that a walk-in arrival to hospital j is triaged as high acuity patient
λ_{jH} (λ_{jL})	Arrival rate of high (low) acuity patients to hospital j
μ_{jH} (μ_{jL})	Treatment rate of high (low) acuity patients at hospital j
μ_{jI}	Treatment rate of inpatients at hospital j
μ_{jHD}	Rate at which high acuity patients that are waiting for service at hospital j die
p_{jHR} (p_{jLR})	Probability that a high (low) acuity patient is discharged after treatment at hospital j
p_{jHA} (p_{jLA})	Probability that a high (low) acuity patient is admitted after treatment at hospital j
p_{jdiv}	A ratio of patients willing to be re-routed from hospital j to the suggested hospital
p_{div}	A ratio of patients willing to be re-routed independent of the original hospital choice
c_{jnE}	Cost of using a normal ED bed at hospital j
c_{jnS}	Cost of using a surge ED bed at hospital j
c_{jnB}	Cost of using an ID bed at hospital j
c_{jBB}	Cost of re-scheduling an inpatient if the ID of hospital j is full
c_{jDT}	Cost of having a patient die while undergoing treatment at hospital j
c_{jDW}	Cost of having a patient die while waiting at hospital j
c_{jW}	Cost of making a patient wait at hospital j
r_{jnH} (r_{jnL})	Revenue generated by a high (low) acuity patient at hospital j
r_{jnB}	Revenue generated by an inpatient at hospital j
p_{ij}	Fraction of patients re-directed from hospital i to hospital j
D_j	Diversion indicator for hospital j
$\hat{n}_j$	Diversion threshold used at hospital j
$\vec{n}_{jI}$	Number of ID patients currently receiving treatment at hospital j
$\vec{n}_{jH}$ ($\vec{n}_{jL}$)	Number of high (low) acuity patients currently receiving treatment at hospital j
π_S	Steady state probability of being in state S
α	Weight given to the most recent observation in EWMA estimation
d_R	Average distance that an ambulance has to travel to a patient
t_R	Epected response time to pick up a patient
w_a	Width of a square area that is serviced by one ambulance crew
v_a	Average speed of an ambulance traveling with signals
w_{All}	Width of Allegheny county
N_A	Number of ambulances ”in service” in Allegheny county

B.2. Details of test instances for our queueing model

This table lists the instances used to illustrate the queueing model in Section 4.5.1:

	Hospital 1				Hospital 2			
Instance	1	2	3	4	1	2	3	4
B_{jED}	8	8	8	8	7	7	5	6
B_{jSED}	3	3	3	3	0	0	0	0
B_{jI}	12	12	12	12	7	7	7	7
$\bar{n}_j$	20	20	20	20	8	8	8	8
λ_{jA}	15	15	15	15	15	15	15	15
λ_{jW}	60	60	60	60	60	60	60	60
λ_{jI}	100	100	100	100	100	100	100	100
μ_{jH}	5	5	5	5	5	5	5	5
μ_{jL}	45	45	45	45	45	45	45	45
μ_{jI}	10	10	10	10	10	10	10	10
μ_{jHD}	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
c_{jnE}	1000	1000	1000	1000	1000	1000	1000	1000
c_{jnS}	1200	1200	1200	1200	1200	1200	1200	1200
c_{jnB}	500	500	500	500	500	500	500	500
c_{jBB}	20	20	20	20	20	20	20	20
c_{jDT}	500	500	500	500	500	500	500	500
c_{jDW}	10^4	10^7	10^4	10^4	10^4	10^7	10^4	10^4
c_{jW}	10	1000	10	10	10	1000	10	10
r_{jnH}	1500	1500	1500	1500	1500	1500	1500	1500
r_{jnL}	1150	1150	1150	1150	1150	1150	1150	1150
r_{jnB}	700	700	700	700	700	700	700	700
P_{jAH}	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
P_{jwH}	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
p_{jHR}	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
p_{jHA}	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
p_{jLR}	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
p_{jLA}	0.199	0.199	0.199	0.199	0.199	0.199	0.199	0.199
p_{jdiv}	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.8

B.3. Distribution of population by ZipCode

This table lists the distance from each ZipCode to the center of the 2-mile radius circle in which all hospitals are located:

Zip	Name	Pop.	Dist.	Prob.
15213	Pittsburgh	28,320	0.000	0.0223
15224	Pittsburgh	12,095	1.519	0.0095
15232	Pittsburgh	11,792	1.713	0.0093
15219	Pittsburgh	19,204	1.781	0.0151
15201	Pittsburgh	14,326	1.953	0.0113
15203	Pittsburgh	9,613	1.998	0.0076
15217	Pittsburgh	26,425	2.186	0.0208
15222	Pittsburgh	1,999	2.653	0.0016
15206	Pittsburgh	32,482	3.065	0.0256
15210	Pittsburgh	31,216	3.350	0.0246
15207	Pittsburgh	13,203	3.351	0.0104
15209	Pittsburgh	12,891	3.667	0.0102
15208	Pittsburgh	13,352	4.070	0.0105
15212	Pittsburgh	31,850	4.084	0.0251
15211	Pittsburgh	12,477	4.120	0.0098
15223	Pittsburgh	7,991	4.180	0.0063
15215	Pittsburgh	13,212	4.627	0.0104
15214	Pittsburgh	17,519	4.637	0.0138
15120	Homestead	20,437	4.706	0.0161
15227	Pittsburgh	29,621	4.725	0.0234
15218	Pittsburgh	14,956	4.845	0.0118
15233	Pittsburgh	4,876	4.988	0.0038
15226	Pittsburgh	14,648	5.447	0.0116
15221	Pittsburgh	36,387	5.735	0.0287
15220	Pittsburgh	19,693	6.246	0.0155
15116	Glenshaw	14,921	6.272	0.0118
15216	Pittsburgh	24,691	6.313	0.0195
15234	Pittsburgh	14,911	6.707	0.0118
15122	West Mifflin	21,861	6.770	0.0172
15104	Braddock	11,434	6.785	0.0090
15236	Pittsburgh	30,630	7.130	0.0242
15204	Pittsburgh	9,502	7.305	0.0075
15229	Pittsburgh	13,677	7.510	0.0108
15034	Dravosburg	2,015	7.806	0.0016
15228	Pittsburgh	17,723	7.883	0.0140
15238	Pittsburgh	13,571	7.968	0.0107
15205	Pittsburgh	22,586	8.157	0.0178
15235	Pittsburgh	39,126	8.499	0.0309
15112	E. Pittsburgh	3,616	8.616	0.0029
15136	Mc Kees Rocks	22,537	8.648	0.0178
15110	Duquesne	7,332	8.774	0.0058
15101	Allison Park	24,323	8.816	0.0192
15202	Pittsburgh	21,022	8.972	0.0166
15147	Verona	18,442	9.145	0.0145
15045	Glassport	4,993	9.236	0.0039
15237	Pittsburgh	42,597	9.253	0.0336
15145	Turtle Creek	7,974	9.389	0.0063
15243	Pittsburgh	13,660	9.389	0.0108
15139	Oakmont	6,911	9.651	0.0055
15106	Carnegie	19,074	9.717	0.0150
15133	McKeesport	6,816	9.859	0.0054
15102	Bethel Park	30,825	10.027	0.0243
15132	McKeesport	26,131	10.167	0.0206
15051	Indianola	628	10.763	0.0005
15137	North Versailles	11,053	10.891	0.0087
15035	E. Mc Keesport	2,354	10.927	0.0019
15025	Clairton	17,341	10.998	0.0137
15148	Wilmerding	2,907	11.020	0.0023
15129	South Park	11,458	11.199	0.0090
15241	Pittsburgh	20,616	11.343	0.0163
15024	Cheswick	8,484	12.145	0.0067
15131	McKeesport	9,132	12.158	0.0072
15142	Presto	713	12.504	0.0006
15140	Pitcairn	3,695	12.581	0.0029
15017	Bridgeville	14,530	12.745	0.0115
15049	Harwick	982	12.847	0.0008
15044	Gibsonia	23,196	12.956	0.0183
15146	Monroeville	29,394	13.110	0.0232
15225	Pittsburgh	1,232	13.393	0.0010
15135	McKeesport	5,623	13.507	0.0044
15144	Springdale	4,648	13.759	0.0037
15090	Wexford	18,252	13.982	0.0144
15037	Elizabeth	11,676	14.352	0.0092
15076	Russellton	931	14.502	0.0007
15007	Bakerstown	345	14.502	0.0003
15064	Morgan	340	14.543	0.0003
15239	Pittsburgh	21,108	14.884	0.0166
15030	Creighton	1,053	15.483	0.0008
15015	Bradfordwoods	1,174	15.947	0.0009
15031	Cuddy	576	15.962	0.0005
15143	Sewickley	16,518	16.056	0.0130
15071	Oakdale	9,287	16.090	0.0073
15108	Coraopolis	37,804	16.341	0.0298
15018	Buena Vista	708	16.534	0.0006
15084	Tarentum	10,542	16.795	0.0083
15086	Warrendale	284	17.818	0.0002
15014	Brackenridge	3,543	18.571	0.0028
15056	Leetsdale	1,215	19.502	0.0010
15046	Crescent	2,242	20.334	0.0018
15065	Natrona Heights	11,996	20.335	0.0095
15126	Imperial	6,743	21.291	0.0053

B.4. Results of statistical analysis

B.4.1 Arrival process

This Appendix describes how the significant factors driving the arrival processes have been determined. We use unbalanced ANOVA to test which factors are statistically significant in determining the mean number of arrivals. Table B.1 and Figure B.1 display the results for the model that included all five factors. Day of week is determined not to be significant as this factor has a p-value of 0.370.

Table B.1: Model fit results for arrival process (NHAMCS).

Factor	Type	Levels	Values
Month	fixed	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
DayOfWeek	fixed	7	Fri, Mon, Sat, Sun, Thu, Tue, Wed
PeriodOfDay	fixed	3	1, 2, 3
Acuity	fixed	5	1, 2, 3, 4, 5
ArrivalMode	fixed	2	Ambulance, Walk-In

(a) Factors included in the model.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Month	11	14.01	19.20	1.75	4.72	0.000
DayOfWeek	6	1.68	2.40	0.40	1.08	0.370
PeriodOfDay	2	484.91	568.58	284.29	769.06	0.000
Acuity	4	1654.41	1743.68	435.92	1179.25	0.000
ArrivalMode	1	2318.41	2318.41	2318.41	6271.81	0.000
Error	4630	1711.51	1711.51	0.37		
Total	4654	6184.94				

S = 0.607994 R-Sq = 72.33% R-Sq(adj) = 72.18%

(b) Analysis of variance for LN(# Arrivals), using Adjusted SS for Tests.

In Table B.2 and Figure B.2 we display the model fit results for a simplified model, eliminating the Month variable. We observe that the R^2 only decreases from 72.18% to 71.94%, leading us to the conclusion that a model without the Month variable is almost as appropriate for our purposes.

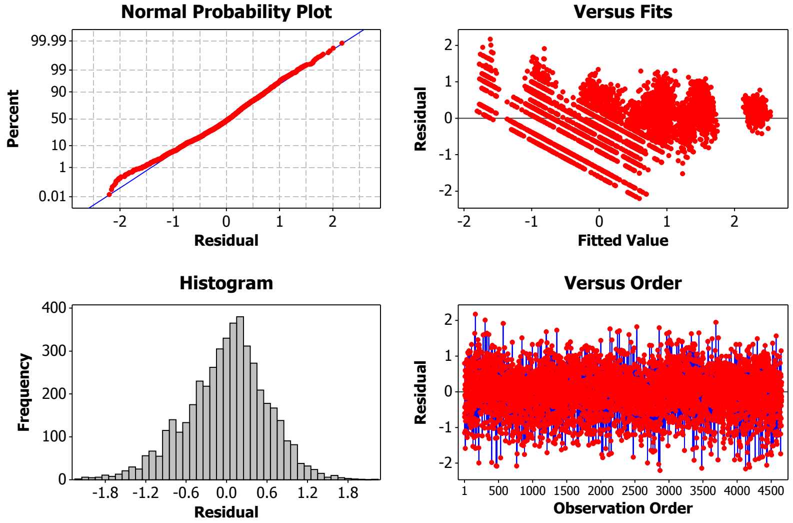


Figure B.1: ANOVA residual plots for the number of arrivals.

Table B.2: Model fit results for arrival process (NHAMCS), simplified model.

Factor	Type	Levels	Values
DayOfWeek	fixed	7	Fri, Mon, Sat, Sun, Thu, Tue, Wed
PeriodOfDay	fixed	3	1, 2, 3
Acuity	fixed	5	1, 2, 3, 4, 5
ArrivalMode	fixed	2	Ambulance, Walk-In

(a) Factors included in the model.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
DayOfWeek	6	1.66	2.37	0.28	0.74	0.615
PeriodOfDay	2	484.40	567.69	242.20	649.48	0.000
Acuity	4	1651.93	1740.76	412.98	1107.44	0.000
ArrivalMode	1	2316.24	2316.24	2316.24	6211.14	0.000
Error	4641	1730.71	1730.71	0.37		
Total	4654	6184.94				

S = 0.610670 R-Sq = 72.02% R-Sq(adj) = 71.94%

(b) Analysis of variance for LN(# Arrivals), using Adjusted SS for Tests.

The plots of residuals vs. fitted values in Figures B.1 and B.2 shows that the residuals have different scattering patterns around zero at different factor levels, which indicates that

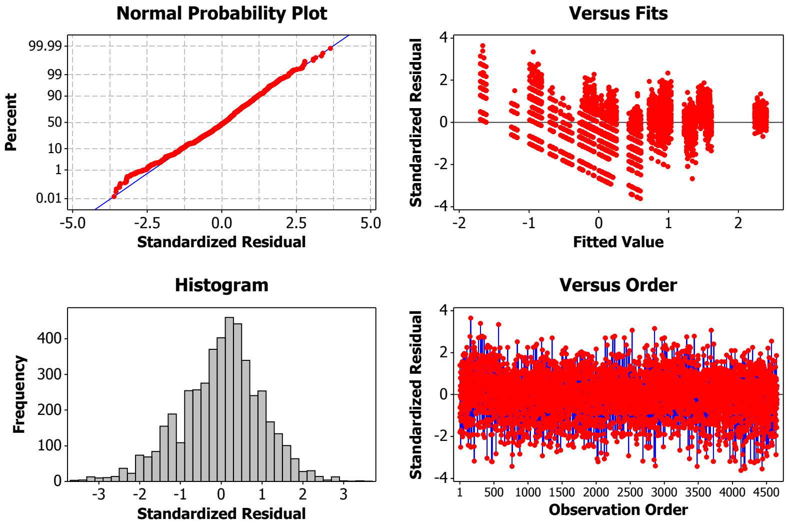


Figure B.2: ANOVA residual plots for the simplified model for the number of arrivals.

the variances between subgroups may be unequal. As our sample sizes are unequal we are likely to find more factors to be significant than when this phenomenon would not be present (Neter et al. 1996). ANOVA is generally believed to be quite robust for violation of the "equal variances" assumption (Neter et al. 1996). In addition, we used several procedures (both Tukey (reported) and Bonferroni) to run the ANOVA and found the same significance results. Therefore, we are confident that our model with Period of day, Acuity, and Arrival mode captures the relevant dynamics of the arrival process of our system.

B.4.2 Treatment time

This Appendix describes how the significant factors driving the treatment time in the ED have been determined. Table B.3 and Figure B.3 display the results for the model that included all five factors.

Table B.3: Model fit results for treatment times (NHAMCS).

Factor	Type	Levels	Values
Acuity	fixed	6	1, 2, 3, 4, 5, 6
ArrivalMode	fixed	2	Ambulance, Walk-In
PeriodOfDay	fixed	3	1, 2, 3
Month	fixed	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
DayOfWeek	fixed	7	Fri, Mon, Sat, Sun, Thu, Tue, Wed

(a) Factors included in the model.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Acuity	5	3951.15	2741.97	548.39	506.82	0.0000
ArrivalMode	1	2290.65	2231.87	2231.87	2062.68	0.0000
PeriodOfDay	2	96.03	97.25	48.62	44.94	0.0000
Month	11	314.78	314.06	28.55	26.39	0.0000
DayOfWeek	6	76.97	76.97	12.83	11.86	0.0000
Error	54673	59157.58	59157.58	1.08		
Total	54698	65887.15				

S = 1.04020 R-Sq = 10.21% R-Sq(adj) = 10.17%

(b) Analysis of variance for LN(Treatment time), using Adjusted SS for Tests.

Although all factors are significant, we explored options to simplify the model. In Table B.4 and Figure B.4 we display the estimation results for a simplified model, keeping only the Acuity and Arrival Mode variables. We observe that the R^2 only decreases from 10.17% to 9.64%, leading us to the conclusion that the simplified model still captures the main features among our factors that drive treatment times.

Inspecting Figures B.3 and B.4 we notice that there are numerous observations with high standardized residuals and/or high leverage. This may indicate that these observations are influential outliers. However, Cook's distance measure, a combined measure of residuals and

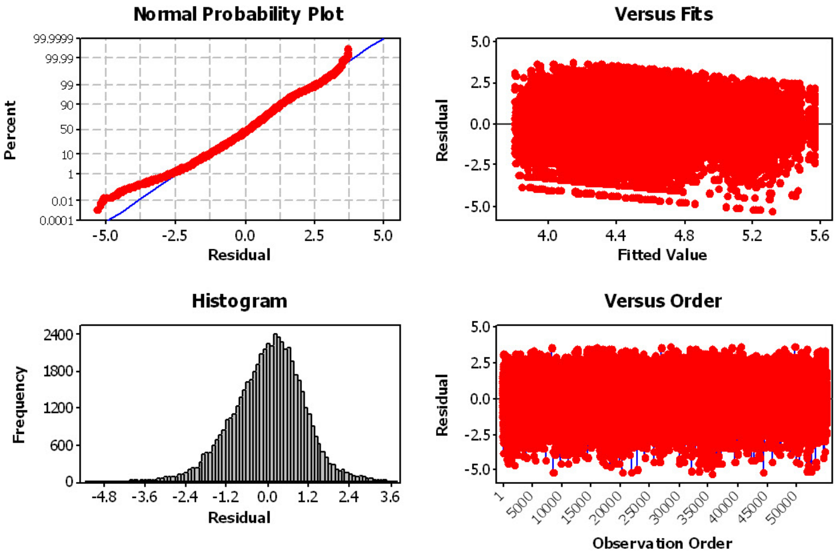


Figure B.3: ANOVA residual plots for treatment times.

Table B.4: Model fit results results for treatment times (NHAMCS), simplified model.

Factor	Type	Levels	Values
Acuity	fixed	6	1, 2, 3, 4, 5, 6
ArrivalMode	fixed	2	Ambulance, Walk-In

(a) Factors included in the model.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Acuity	5	3951.1	2774.2	554.8	508.76	0.0000
ArrivalMode	1	2290.6	2290.6	2290.6	2100.42	0.0000
Error	54692	59645.4	59645.4	1.1		
Total	54698	65887.2				

S = 1.04430 R-Sq = 9.47% R-Sq(adj) = 9.46%

(b) Analysis of variance for LN(treatment time), using Adjusted SS for Tests.

leverage, does not show any abnormalities (all values close to 0), hence no action is required. Moreover, inclusion of influential outliers would reduce F-values, making significant factors look insignificant which has not occurred as all factors are still significant.

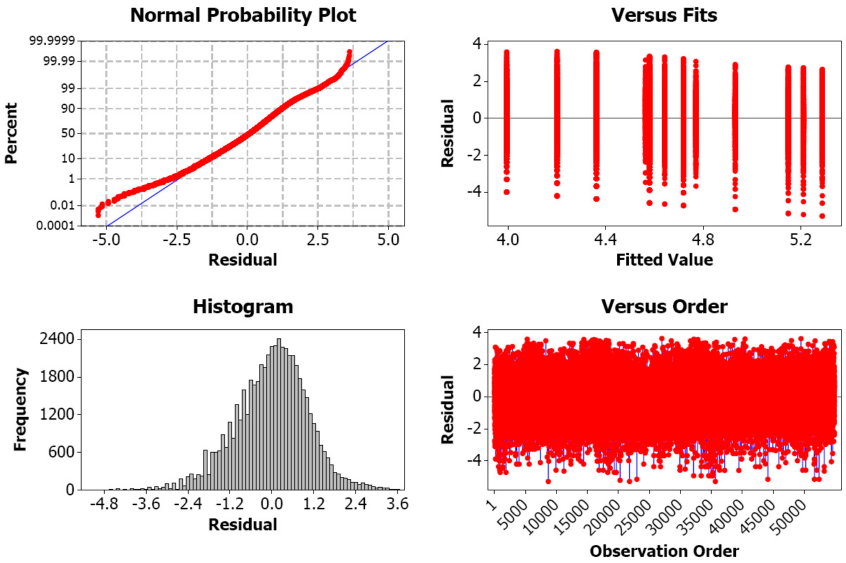


Figure B.4: ANOVA residual plots for the simplified model of treatment times.

B.4.3 Length of stay - Nonurgent

We have attempted to use ANOVA to determine which factors have a significant impact on the LOS for admitted patients. However, regardless of different remedial measures tried (such as various transformations and a model with weighted least squares) we could not fit a model that would have normally distributed error terms and would capture a reasonably large percentage of variability. This indicates that the LOS is impacted by factors that we are not considering. If we split our data set by Acuity level (as suggested by Guyette, 2009) we can use ANOVA for the Nonurgent arrivals (see Table B.5 and Figure B.5). In this model none of the factors considered (Month, Day of week, Period of day, Mode of arrival) were significant. Hence, we decided to split the LOS by Acuity only and estimate a distribution based on this factor.

Table B.5: Model fit results for Length of stay (NHAMCS).

Factor	Type	Levels	Values
ArrivalMode	fixed	2	Ambulance, Walk-In
PeriodOfDay	fixed	3	1, 2, 3
Month	fixed	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
DayOfWeek	fixed	7	Fri, Mon, Sat, Sun, Thu, Tue, Wed

(a) Factors included in the model.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
DayOfWeek	6	5.1687	6.0008	1.0001	2.20	0.044
ArrivalMode	1	0.6433	0.4939	0.4939	1.09	0.298
PeriodOfDay	2	2.4722	2.3972	1.1986	2.64	0.074
Month	11	5.9058	5.9058	0.5369	1.18	0.300
Error	219	99.3611	99.3611	0.4537		
Total	239	113.5511				

S = 0.673575 R-Sq = 12.50% R-Sq(adj) = 4.51%

(b) Analysis of variance for LN(LOS), using Adjusted SS for Tests.

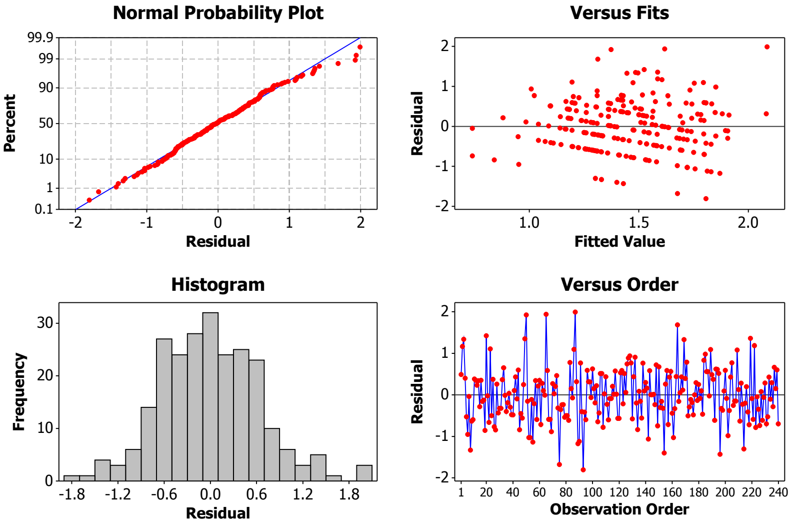


Figure B.5: ANOVA residual plots for the LOS for Nonurgent ED arrivals.

B.5. OptQuest progress

Figure B.6 provides insight into the progress of an OptQuest optimization run. This figure displays the best objective value obtained up to each point in the optimization, simulation sequence.

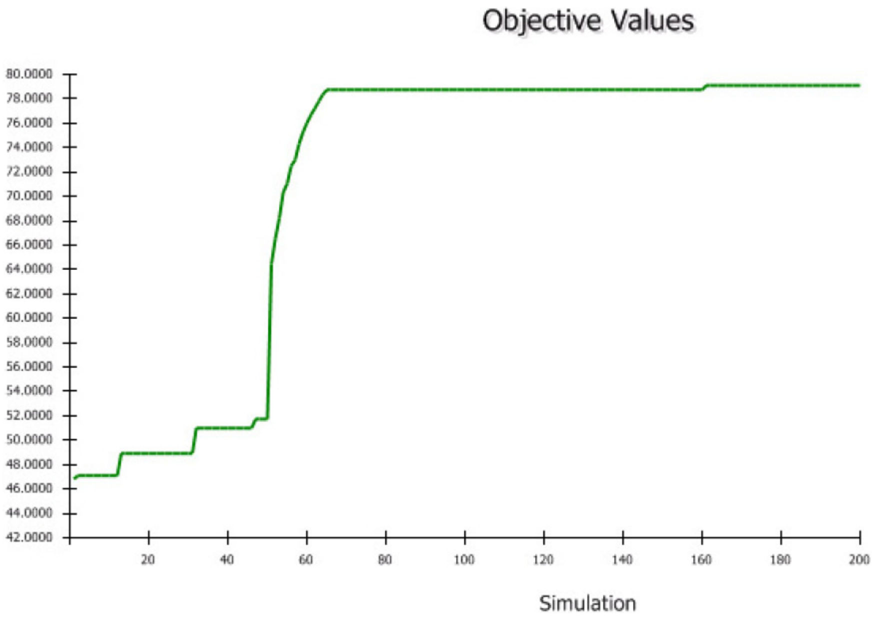


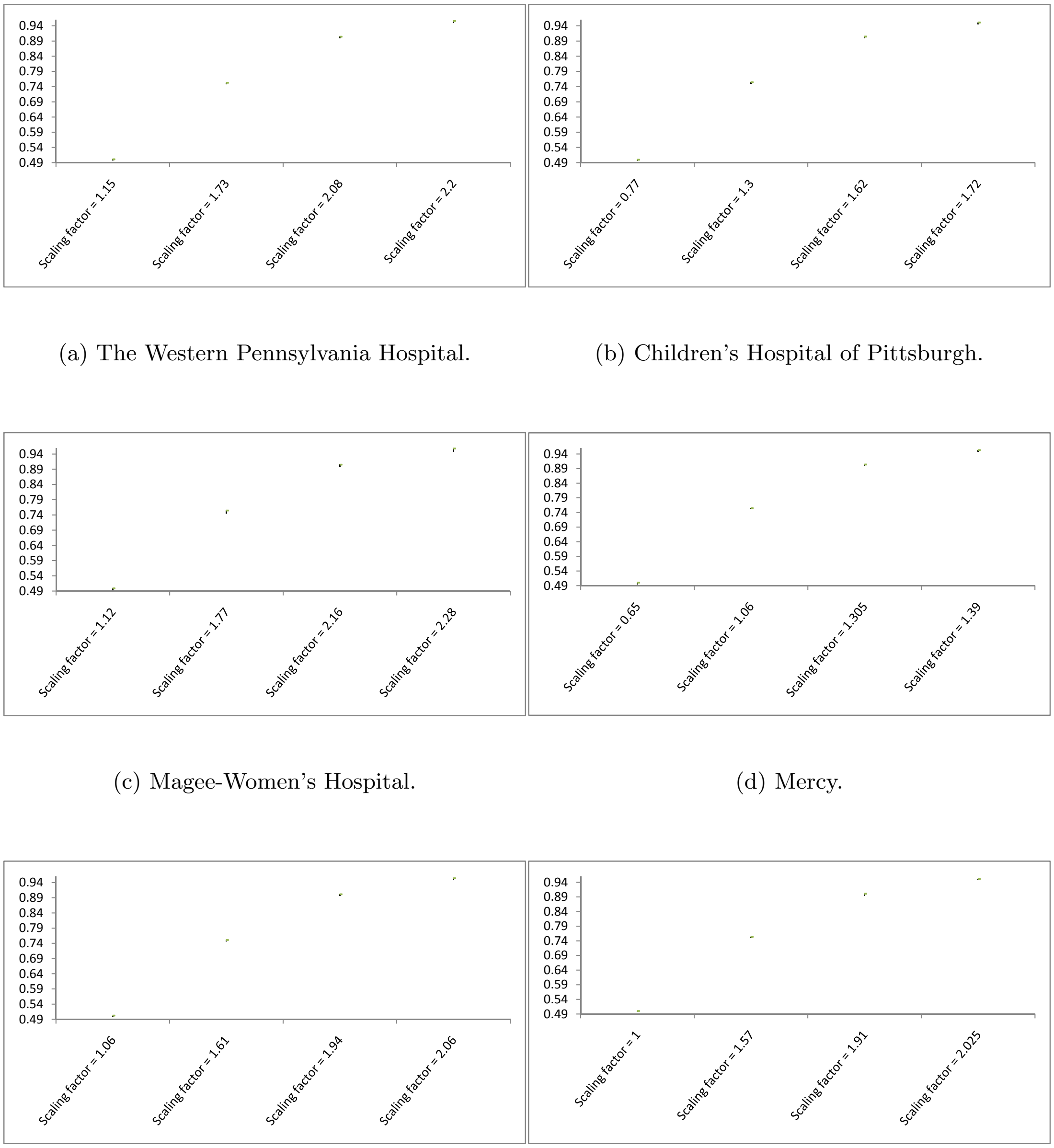
Figure B.6: Progress of an OptQuest optimization run.

B.6. Optimal diversion levels

The optimal diversion levels for the diversion level policy analyzed in Section 4.5.2 are as follows:

Utilization	50%			75%			90%			95%		
p_{div}	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75
Average discounted quality												
Allegheny General Hospital	1	8	5	8	8	9	3	2	6	8	4	11
Western Pennsylvania Hospital	1	1	1	8	25	25	0	1	1	0	1	1
Children's Hospital of Pittsburgh	16	22	31	15	21	26	17	18	15	15	15	17
Magee-Women's Hospital	0	3	9	0	1	2	10	5	5	5	5	6
Mercy	1	3	6	3	5	7	9	10	9	4	5	7
Presbyterian	11	15	15	8	3	15	4	5	9	5	8	7
Shadyside	2	3	8	6	5	10	11	6	3	1	9	3
Hospital revenue												
Allegheny General Hospital	12	16	18	8	14	13	12	14	14	10	12	12
Western Pennsylvania Hospital	1	1	1	8	25	25	1	1	1	1	1	1
Children's Hospital of Pittsburgh	22	25	28	25	20	25	21	19	19	20	24	21
Magee-Women's Hospital	9	9	9	4	6	3	4	10	5	9	7	5
Mercy	17	20	16	15	19	20	13	21	17	15	16	15
Presbyterian	20	20	16	22	8	22	14	14	16	16	16	16
Shadyside	17	21	20	14	14	15	17	17	18	11	12	11

B.7. Confidence intervals for the utilization of EDs



(a) The Western Pennsylvania Hospital.

(b) Children's Hospital of Pittsburgh.

(c) Magee-Women's Hospital.

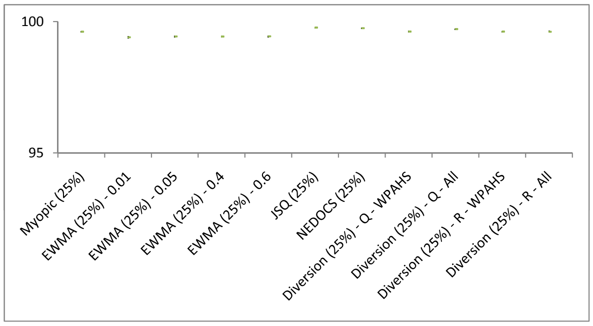
(d) Mercy.

(e) Presbyterian.

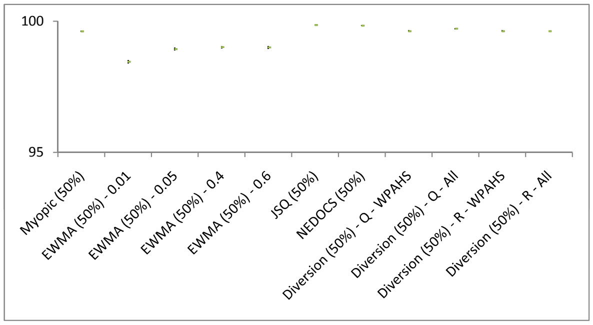
(f) Shadyside.

Figure B.7: 95% Confidence intervals for the utilization of specific EDs, for each scaling factor assuming $p_{div} = 0$.

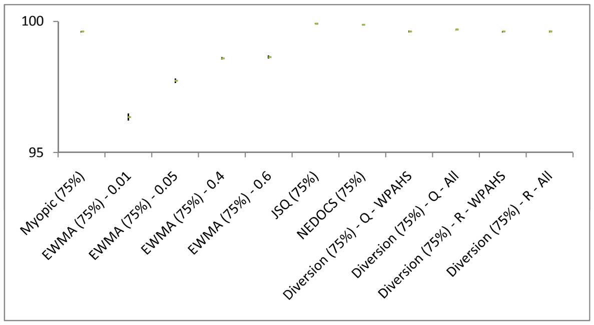
B.8. Confidence intervals for the average discounted quality



(a) $p_{div} = 25\%$.

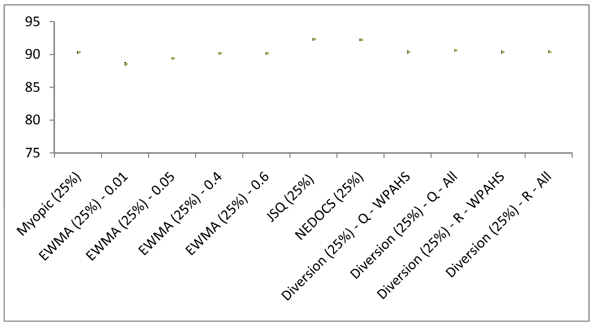


(b) $p_{div} = 50\%$.

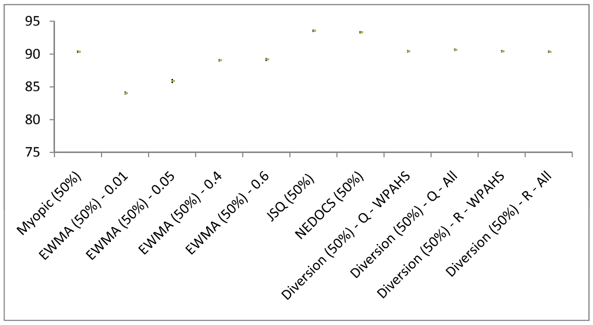


(c) $p_{div} = 75\%$.

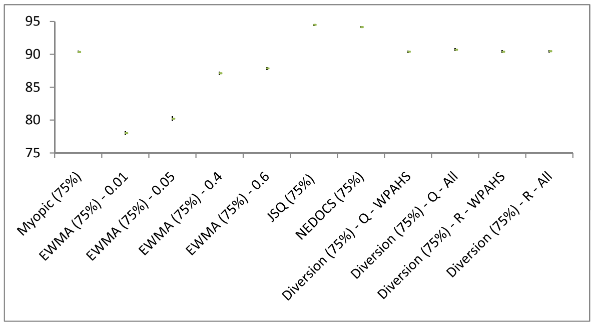
Figure B.8: Average discounted quality for the 50% utilization level



(a) $p_{div} = 25\%$.

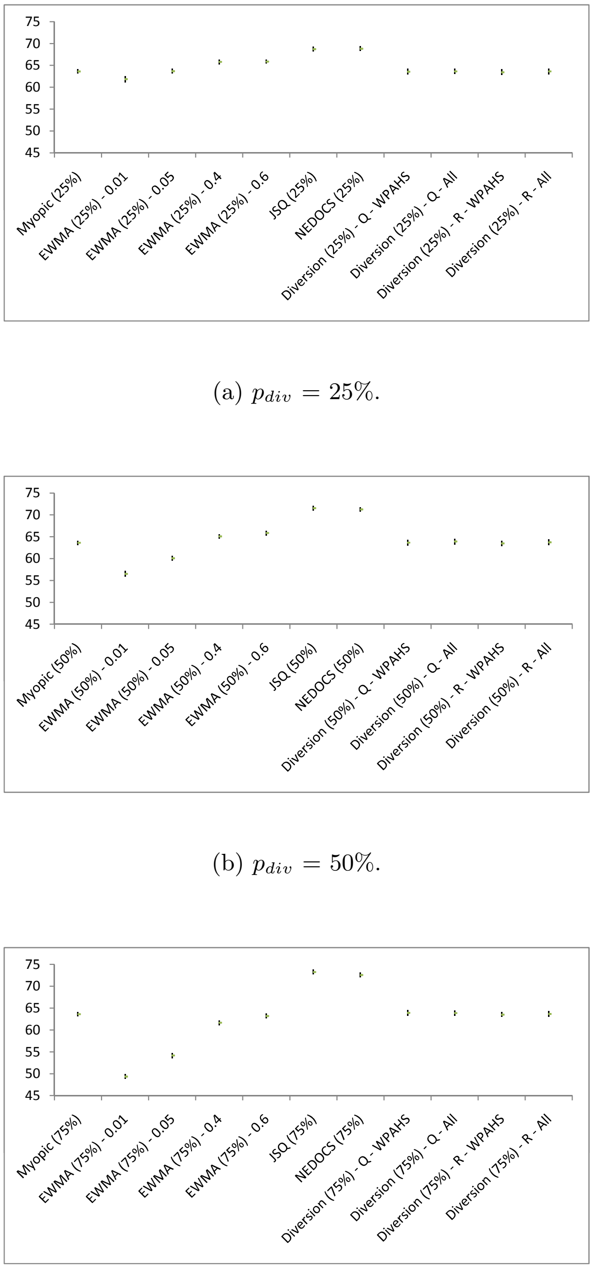


(b) $p_{div} = 50\%$.



(c) $p_{div} = 75\%$.

Figure B.9: Average discounted quality for the 75% utilization level



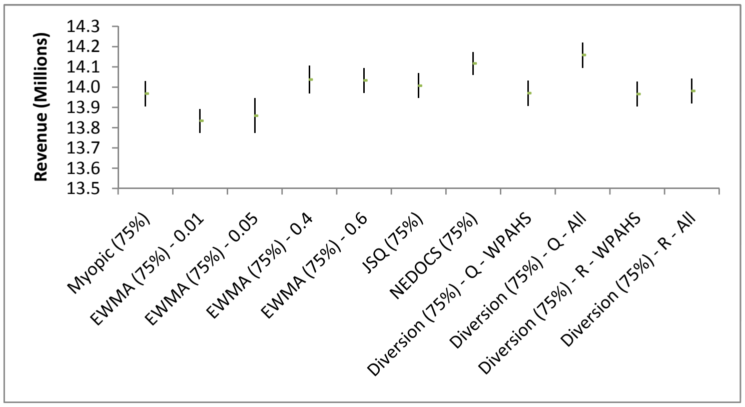
(a) $p_{div} = 25\%$.

(b) $p_{div} = 50\%$.

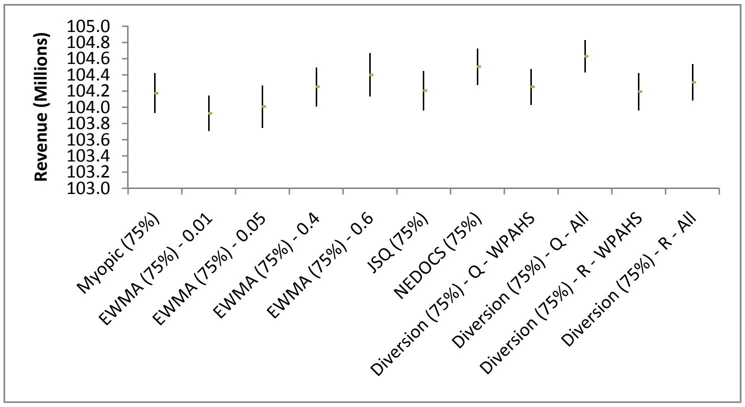
(c) $p_{div} = 75\%$.

Figure B.10: Average discounted quality for the 90% utilization level

B.9. Hospital and ED revenue confidence intervals

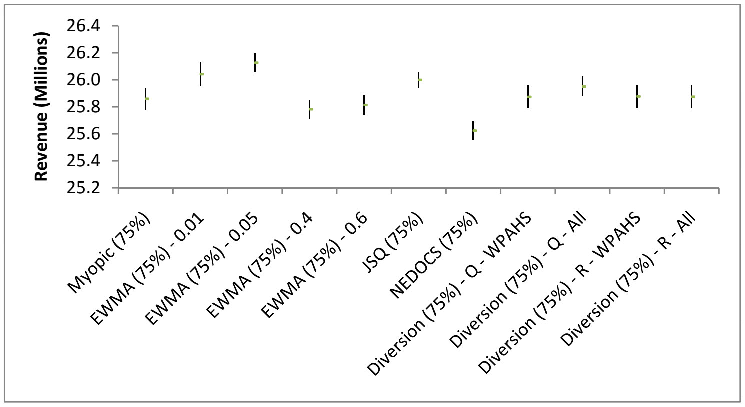


(a) ED.

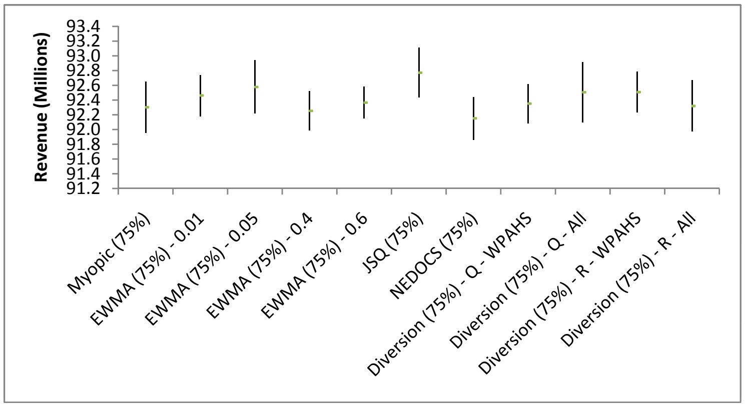


(b) ED & ID.

Figure B.11: The Western Pennsylvania Hospital revenues for a utilization of 95% and $p_{div} = 75\%$.

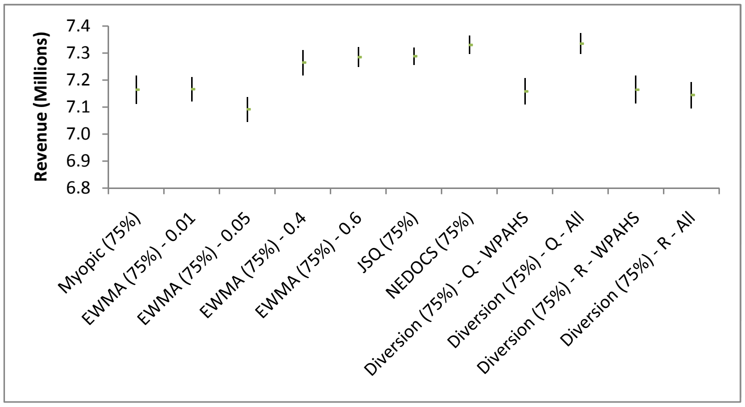


(a) ED.

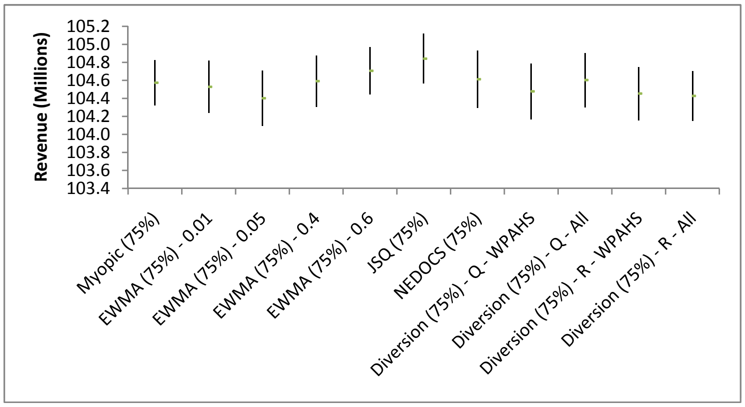


(b) ED & ID.

Figure B.12: Children's Hospital of Pittsburgh revenues for a utilization of 95% and $p_{div} = 75\%$.

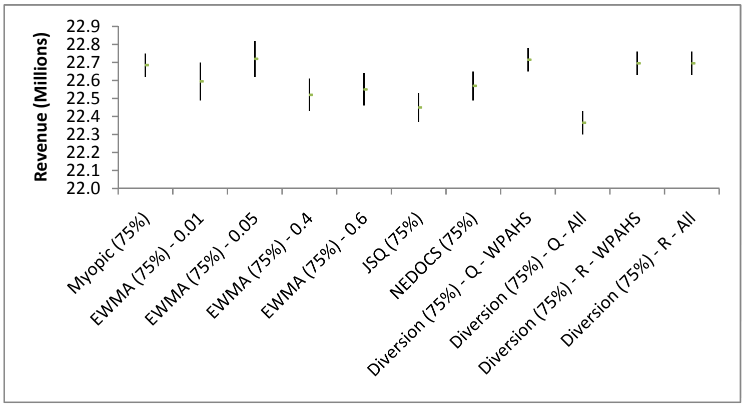


(a) ED.

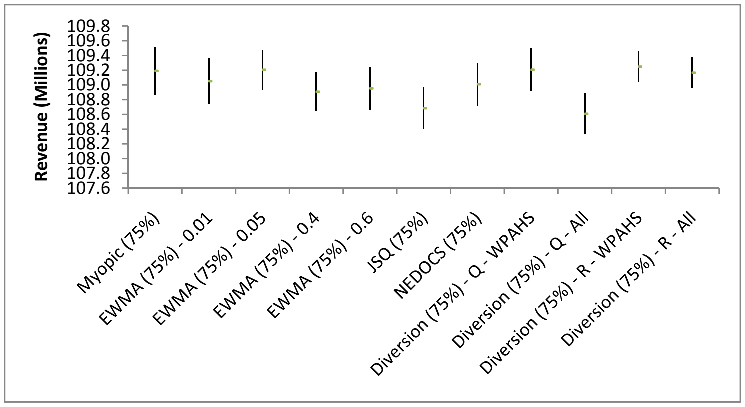


(b) ED & ID.

Figure B.13: Magee Women's Hospital Hospital revenues for a utilization of 95% and $p_{div} = 75\%$.

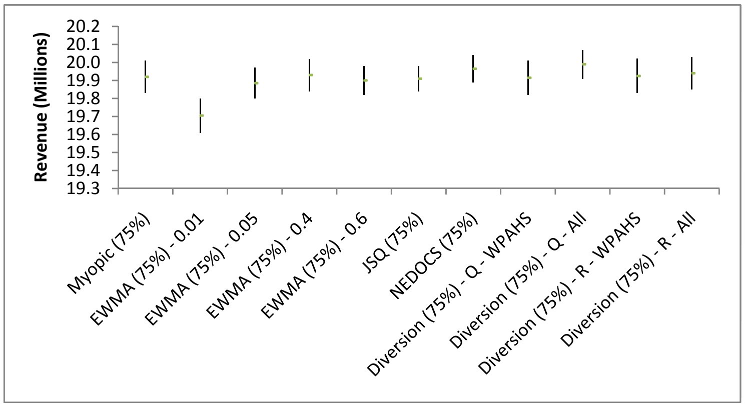


(a) ED.

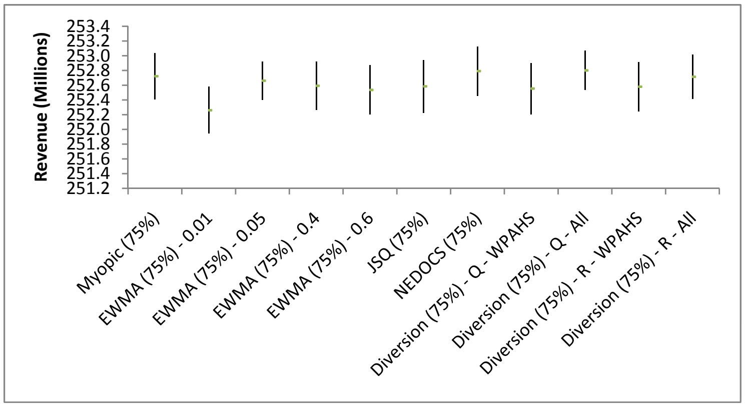


(b) ED & ID.

Figure B.14: Mercy revenues for a utilization of 95% and $p_{div} = 75\%$.

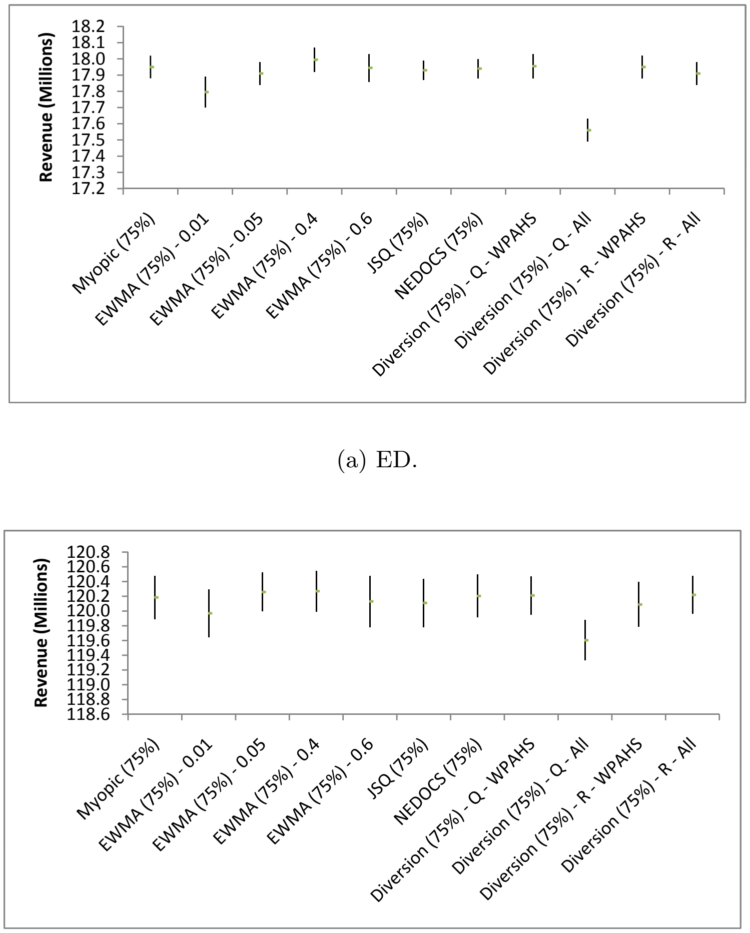


(a) ED.



(b) ED & ID.

Figure B.15: Presbyterian revenues for a utilization of 95% and $p_{div} = 75\%$.



(a) ED.

(b) ED & ID.

Figure B.16: Shadyside revenues for a utilization of 95% and $p_{div} = 75\%$.

Bibliography

Al-Harthy, M.H. 2007. Stochastic oil price models: comparison and impact. *The Engineering Economist* **52**(3) 269–284.

Alfredsson, P., J. Verrijdt. 1999. Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science* **45**(10) 1416–1431.

Allegheny County. 2010. Map of municipalities. `http://www.alleghenycounty.us/munimap/index.asp`. Accessed: March 31 2010.

Allon, G., S. Deo, W. Lin. 2009. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Kellogg working paper* .

American College of Emergency Physicians. 2006. Crowding. *Annals of Emergency Medicine* **47**.

Andrulis, D.P., A. Kellermann, E.A. Hintz, B.B. Hackman, V.B. Weslowski. 1991. Emergency departments and crowding in United States teaching hospitals. *Annals of Emergency Medicine* **20**(9) 980–986.

Anonymous. 1993. Emergency departments: unevenly affected by growth and change in patient use. *General Accounting Office* .

Asplin, B.R., D.J. Magid, K.V. Rhodes, L.I. Solberg, N. Lurie, C.A. Camargo Jr. 2003. A conceptual model of emergency department crowding. *Annals of Emergency Medicine* **42**(2) 173–180.

Barthell, E.N., S.L. Foldy, K.R. Pemble, C.W. Felton, P.J. Greischar, R.G. Pirrallo, W.J. Bazan. 2003. Assuring community emergency care capacity with collaborative Internet tools: the Milwaukee experience. *Journal of Public Health Management and Practice* **9**(1) 35.

Benjaafar, S., M. ElHafsi, C.Y. Lee, W. Zhou. 2006. Optimal control of assembly systems with multiple stages and multiple demand classes. *Working paper* University of Minnesota.

Blomkalns, A.L., W.B. Gibler. 2004. Emergency department crowding: Emergency physicians and cardiac risk stratification as part of the solution. *Annals of Emergency Medicine* **43**(1) 77–78.

Bradley, V.M. 2005. Placing emergency department crowding on the decision agenda. *Journal of Emergency Nursing* **31**(3) 247–258.

Brennan, M.J., E.S. Schwartz. 1985. Evaluating natural resource investments. *Journal of Business* **58**(2) 135–157.

Bright, L., P.G. Taylor. 1995. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models* **11**(3) 497–525.

Burt, C.W., L.F. McCaig, R.H. Valverde. 2006. Analysis of ambulance transports and diversions among US emergency departments. *Annals of Emergency Medicine* **47**(4) 317–326.

Cameron, PA. 2006. Hospital overcrowding: a threat to patient safety? *The Medical Journal of Australia* **184**(5) 203.

Carter, A.J.E., R. Grierson. 2007. The impact of ambulance diversion on EMS resource availability. *Prehospital Emergency Care* **11**(4) 421–426.

Cattani, K.D., G.C. Souza. 2002. Inventory rationing and shipment flexibility alternatives for direct market firms. *Production and Operations Management* **11**(4) 441–457.

Cheung, N.K., J.H.H. Yeung, J.T.S. Chan, P.A. Cameron, C.A. Graham, T.H. Rainer. 2006. Primary trauma diversion: initial experience in Hong Kong. *The Journal of Trauma* **61**(4) 954.

Clarke, H.R., W.J. Reed. 1990. Oil-well valuation and abandonment with price and extraction rate uncertain. *Resources and Energy* **12**(4) 361–382.

Clegg, J.D., S.M. Bucaram, N.W. Hein. 1993. Recommendations and comparisons for selecting artificial-lift methods. *Journal of Petroleum Technology* **45**(12) 1128–1131, 1163–1167.

Clewlow, L., C. Strickland. 2000. *Energy Derivatives: Pricing and Risk Management*. Lacima Publications, London, UK.

Coşkuner, G., T. Strocen. 2003. Production optimization of liquid loading gas condensate wells: A case study. *Journal of Canadian Petroleum Technology* **42**(11) 40–44.

Cohen, J.W. 1976. *On Regenerative Processes in Queueing Theory*. Lecture Notes in Economics and Mathematical Systems 121, Springer, Berlin.

Coleman, S.B., H.B. Clay, D.G. McCurdy, H.L. Norris III. 1991a. Applying gas-well load-up technology. *Journal of Petroleum Technology* **43**(3) 344–349.

Coleman, S.B., H.B. Clay, D.G. McCurdy, H.L. Norris III. 1991b. A new look at predicting gas-well load-up. *Journal of Petroleum Technology* **43**(3) 329–333.

Coleman, S.B., H.B. Clay, D.G. McCurdy, H.L. Norris III. 1991c. Understanding gas-well load-up behavior. *Journal of Petroleum Technology* **43**(3) 334–338.

Cortazar, G., E.S. Schwartz, J. Casassus. 2001. Optimal exploration investments under price and geological-technical uncertainty: A real options model. *R&D Management* **31**(2) 181–189.

De Véricourt, F., F. Karaesmen, Y. Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Science* **48**(11) 1486–1501.

Dekker, R., R.M. Hill, M.J. Kleijn, R.H. Teunter. 2002. On the $(s-1, s)$ lost sales inventory model with priority demand classes. *Naval Research Logistics* **49** 593–610.

Dekker, R., M.J. Kleijn, P.J. de Rooij. 1998. A spare parts stocking policy based on equipment criticality. *International Journal of Production Economics* **56–57** 66–77.

Derlet, R.W., J.R. Richards. 2000. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Annals of Emergency Medicine* **35**(1) 63–68.

Deshpande, V., M.A. Cohen, K. Donohue. 2003a. An empirical study of service differentiation for weapon system service parts. *Operations Research* **51**(4) 518–530.

Deshpande, V., M.A. Cohen, K. Donohue. 2003b. A threshold inventory rationing policy for service-differentiated demand classes. *Management Science* **49**(6) 683–703.

Dixit, A.K., R.S. Pindyck. 1994. *Investment under Uncertainty*. Princeton University Press, Princeton, New Jersey.

Dugan, T. 2006. Personal communication.

Duran, S., T. Liu, D. Simchi-Levi, J. Swann. 2008. Policies utilizing tactical inventory for service-differentiated customers. *OR Letters* **36**(2) 259–264.

Durrer, E.J., G.E. Slater. 1977. Optimization of petroleum and natural gas production - A survey. *Management Science* **24**(1) 35–43.

Eckstein, M., L.S. Chan. 2004. The effect of emergency department crowding on paramedic ambulance availability. *Annals of Emergency Medicine* **43**(1) 100–105.

Eckstein, M., S.M. Isaacs, C.M. Slovis, B.J. Kaufman, J.R. Loflin, R.E. O'Connor, P.E. Pepe. 2005. Facilitating EMS turnaround intervals at hospitals in the face of receiving facility overcrowding. *Prehospital Emergency Care* **9**(3) 267–275.

Enders, P., A.A. Scheller-Wolf, N. Secomandi. 2010. Interaction between scaling options in natural gas production. *IIE Transactions* **42**(9).

Evans, R.V. 1968. Sales and restocking policies in a single item inventory system. *Management Science* **14**(7) 463–472.

Falvo, T., L. Grove, R. Stachura, W. Zirkin. 2007. The financial impact of ambulance diversions and patient elopements. *Academic Emergency Medicine* **14**(1) 58–62.

Fisher, M. 2007. Strengthening the empirical base of Operations Management. *Manufacturing & Service Operations Management* **9**(4) 368–382.

Frank, K., R.Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Operations Research* **51**(6) 993–1002.

Gans, N., S. Savin. 2007. Pricing and capacity rationing for rentals with uncertain durations. *Management Science* **53**(3) 390–407.

Garb, F.A., T.A. Larson. 1989. Valuation of oil and gas reserves. H.B. Bradley, ed., *Petroleum Engineering Handbook*, chap. 41. Society of Petroleum Engineers.

Glover, F. 1989. Scatter search and path relinking. D. Corne, M. Dorigo, F. Glover, eds., *New Methods in Optimization*. McGraw-Hill.

Glover, F., JP Kelly, M. Laguna. 1999. New advances for wedding optimization and simulation. *Simulation Conference Proceedings, 1999 Winter*, vol. 1.

Glover, F., M. Laguna. 1997. *Tabu Search*. Kluwer Academic Publishers.

Grayson, C.J. Jr. 1960. *Decisions under Uncertainty: Drilling Decisions by Oil and Gas Operators*. Harvard Business School, Boston, Massachusetts.

Green, LV, V. Nguyen. 2001. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* **36**(2) 421.

Guyette, F.X. (MD MPH). 2009. Personal communication.

Ha, A.Y. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science* **43**(8) 1093–1103.

Ha, A.Y. 1997b. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics* **44** 457–472.

Ha, A.Y. 2000. Stock rationing in an $m/e_k/1$ make-to-stock queue. *Management Science* **46**(1) 77–87.

Hahn, W.J., J.S. Dyer. 2008. Discrete time modeling of mean-reverting stochastic processes for real option valuation. *European Journal of Operational Research* **184**(2) 534–548.

Hoot, N.R., D. Aronsky. 2008. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of Emergency Medicine* **52**(2) 126–136.

Hostler, D. (PhD). 2009. Personal communication.

Hostler, D. (PhD). 2010. Personal communication.

Inoue, H., S. Yanagisawa, I. Kamae. 2006. Computer-simulated assessment of methods of transporting severely injured individuals in disaster–Case study of an airport accident. *Computer Methods and Programs in Biomedicine* **81**(3) 256–265.

Jaillet, P., E.I. Ronn, S. Tompaidis. 2004. Valuation of commodity-based swing options. *Management Science* **50**(7) 909–921.

Kamrad, B., R. Ernst. 2001. An economic model for evaluating mining and manufacturing ventures with output yield uncertainty. *Operations Research* **49**(5) 690–699.

Kamrad, B., S. Lele. 1998. Production, operating risk and market uncertainty: A valuation perspective on controlled policies. *IIE Transactions* **30**(5) 455–468.

Kaplan, A. 1969. Stock rationing. *Management Science* **15**(5) 260–267.

Kekre, S., N. Secomandi, E. S

"onmez, K. West. 2008. Balancing risk and efficiency at a major commercial bank. *Manufacturing and Service Operations Management. Articles in Advance* 1–14.

Kelton, W.D., R.P. Sadowski, D.A. Sadowski. 1998. *Simulation with ARENA*. McGraw-Hill USA.

Kranenburg, A.A., G.J. Van Houtum. 2007. Cost optimization in the (s-1,s) lost sales inventory model with multiple demand classes. *OR Letters* **35** 493–502.

Kranenburg, A.A., G.J. Van Houtum. 2008. Service differentiation in spare parts inventory management. *Journal of the Operational Research Society* **59** 946–955.

Krochmal, P., T.A. Riley. 1994. Increased health care costs associated with ED overcrowding. *The American Journal of Emergency Medicine* **12**(3) 265–266.

Lagoe, R.J., J.C. Kohlbrenner, Hall L.D., M. Roizen, P.A. Nadle, R.C. Hunt. 2003. Reducing ambulance diversion: a multihospital approach. *Prehospital Emergency Care* **7**(1) 99–108.

Lai, G., M.X. Wang, S. Kekre, A. Scheller-Wolf, N. Secomandi. 2009. *Valuation of the real option to store liquefied natrual gas at a regasification terminal*. Working paper 2006-E99, Tepper School of Business, Carnegie Mellon University.

Lambe, S., D.L. Washington, A. Fink, K. Herbst, H. Liu, JS Fosse, S.M. Asch. 2002. Trends in the use and capacity of California's emergency departments, 1990-1999. *Annals of Emergency Medicine* **39**(4) 389–396.

Larson, G. 2008. Ambulance destination determination system for ambulance distribution as an alternative to ambulance diversion. *Journal of Emergency Nursing* **34**(4) 357–358.

Lautouche, G., V. Ramaswami. 1987. *Introduction to Matrix Analytic Methods in Stochastic Modeling (ASA-SIAM Series on Statistics and Applied Probability)*. Society for Industrial Mathematics, Philadelphia, PA.

Law, A.M., W.D. Kelton. 1999. Simulation modeling and analysis .

Luenberger, D.G. 1998. *Investment Science*. Oxford University Press, New York, New York.

Lumley, R.R., M. Zervos. 2001. A model for investments in the natural resource industry with switching cost. *Mathematics of Operations Research* **26**(4) 637–653.

Lund, M.W. 2000. Valuing flexibility in offshore petroleum projects. *Annals of Operations Research* **99**(1) 325–349.

Medical Advisory Committee, Pennsylvania. 2004. Joint position statement: guidelines for hospital ambulance diversion policies. `http://www.pitt.edu/~kconover/ftp/diversion.pdf`. Accessed: April 8 2010.

Melchiors, P. 2003. Restricted time-remembering policies for the inventory rationing problem. *International Journal of Production Economics* **81–82** 461–468.

Melchiors, P., R. Dekker, M.J. Kleijn. 2000. Inventory rationing in an (s,q) inventory model with lost sales and two demand classes. *Journal of the Operational Research Society* **51** 111–122.

Moel, A., P. Tufano. 2002. When are real options exercised? An empirical study of mine closings. *The Review of Financial Studies* **15**(1) 35–64.

Möllering, K.T., U.W. Thonemann. 2008. An optimal critical level policy for inventory systems with two demand classes. *Naval Research Logistics* **55** 632–642.

Montgomery, D.C., G.C. Runger. 1999. *Applied statistics and probability for engineers*. John Wiley & Sons.

Moskop, J.C., D.P. Sklar, J.M. Geiderman, R.M. Schears, K.J. Bookman. 2009. Emergency department crowding, part 1-concept, causes, and moral consequences. *Annals of Emergency Medicine* **53**(5) 605–611.

Nahmias, S., W.S Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Science* **27**(11) 1236–1245.

Nawar, E.W., R.W. Niska, J. Xu. 2007. National hospital ambulatory medical care survey: 2005 emergency department summary. *National Health Statistics Report* 1–39.

Neely, KW, RL Norton, GP Young. 1994. The effect of hospital resource unavailability and ambulance diversions on the EMS system. *Prehospital and Disaster Medicine* **9**(3) 172.

Neter, J., M.H. Kutner, C.J. Nachtsheim, W. Wasserman. 1996. *Applied linear statistical models*. McGraw-Hill, Burr Ridge, Illinois.

Neuts, M.F. 1981. *Matrix-geometric solutions in stochastic models*. John Hopkins Univerity Press, Baltimore.

Olshaker, J.S. 2009. Managing emergency department overcrowding. *Emergency Medicine Clinics of North America* **27**(4) 593–603.

Olshaker, J.S., N.K. Rathlev. 2006. Emergency department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the emergency department. *Journal of Emergency Medicine* **30**(3) 351–356.

Olson, T.E., G. Stensland. 1988. Optimal shutdown decisions in resource extraction. *Economics Letters* **26**(3) 215–218.

Patel, P.B., R.W. Derlet, D.R. Vinson, M. Williams, J. Wills. 2006. Ambulance diversion reduction: the Sacramento solution. *The American Journal of Emergency Medicine* **24**(2) 206–213.

Paterson, C., G. Kiesmuller, R. Teunter, K. Glazebrook. 2009. Inventory models with lateral transshipments: A review. *Beta Working paper #287* .

Patterson, MPH EMT-B), P.D. (PhD. 2010. Personal communication.

Pines, J.M., J.E. Hollander. 2008. Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine* **51**(1) 1–5.

Pitts, S.R., R.W. Niska, J. Xu, CW Burt. 2008. National hospital ambulatory medical care survey: 2006 emergency department summary. *National Health Statistics Report* **7** 1–39.

Portsmouth City Council. 2006. Policy & review topic panel a. `http://www.portsmouth.gov.uk/media/pra20060912m.pdf`. Accessed: April 9 2010.

Puterman, M.L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, New York, NY.

Reid, P.P. 2005. *Building a better delivery system: a new engineering/health care partnership*. National Academy Press.

Richards, J.R., M.L. Navarro, R.W. Derlet. 2000. Survey of directors of emergency departments in California on overcrowding. *Western Journal of Medicine* **172**(6) 385.

Ross, S.M. 1997. *Introduction to probability models*. Academic Press.

Routledge, B., D. Seppi, C. Spatt. 2000. Equilibrium forward curves for commodities. *Journal of Finance* **55**(3) 1297–1338.

Ruffner, M. (MD PhD). 2010. Personal communication.

Schafermeyer, R.W., B.R. Asplin. 2003. Hospital and emergency department crowding in the United States. *Emergency Medicine* **15** 22–27.

Schneider, S.M., M.E. Gallery, R. Schafermeyer, F.L. Zwemer. 2003. Emergency department crowding: a point in time. *Annals of Emergency Medicine* **42**(2) 167–172.

Schull, M.J., K. Lazier, M. Vermeulen, S. Mawhinney, L.J. Morrison. 2003. Emergency department contributors to ambulance diversion: a quantitative analysis. *Annals of Emergency Medicine* **41**(4) 467–476.

Schull, M.J., L.J. Morrison, M. Vermeulen, D.A. Redelmeier. 2002a. Emergency department overcrowding and ambulance transport delays for patients with chest pain. *Canadian Medical Association Journal* **168**(3) 277.

Schull, M.J., P.M. Slaughter, DA Redelmeier. 2002b. Urban emergency department overcrowding: defining the problem and eliminating misconceptions. *Canadian Journal of Emergency Medicine* **4**(2) 76–83.

Schwartz, E., J.E. Smith. 2000. Short-term variations and long-term dynamics in commodity prices. *Management Science* **46**(7) 893–911.

Schwartz, E.S. 1997. The stochastic behavior of commodity prices: Implications for valuation and hedging. *Journal of Finance* **52**(3) 923–973.

Sciullo, F.C. 2006. Personal communication.

Seppi, D.J. 2003. Risk-neutral stochastic processes for commodity derivative pricing: An introduction and survey. E.I. Ronn, ed., *Real Options and Energy Management: Using Options Methodology to Enhance Capital Budgeting Decisions*, chap. 1. Risk Books.

Shah, M.N., R.J. Fairbanks, C.L. Maddow, E.B. Lerner, J.I. Syrett, E.A. Davis, S.M. Schneider. 2006. Description and evaluation of a pilot physician-directed emergency medical services diversion control program. *Academic Emergency Medicine* **13**(1) 54–60.

Simchi-Levi, D., P. Kaminsky, E. Simchi-Levi. 2008. *Designing and managing the supply chain*. McGraw-Hill.

Slade, M.E. 2001. Valuing managerial flexibility: An application of real-option theory to mining investments. *Journal of Environmental Economics and Management* **41**(2) 193–233.

Smith, J.E. 2005. Alternative approaches for solving real-options problems. *Decision Analysis* **2**(2) 89–102.

Smith, J.E., K.F. McCardle. 1998. Valuing oil properties: Integrating option pricing and decision analysis approaches. *Operations Research* **46**(2) 198–217.

Smith, J.E., K.F. McCardle. 1999. Options in the real world: Lessons learned in evaluating oil and gas investments. *Operations Research* **47**(1) 1–15.

Smith, J.E., R.F. Nau. 1995. Valuing risky projects: Option pricing theory and decision analysis. *Management Science* **41**(5) 795–816.

Spaite, D.W., T.D. Valenzuela, H.W. Meislin, E.A. Criss, P. Hinsberg. 1993. Prospective validation of a new model for evaluating emergency medical services systems by in-field observation of specific time intervals in prehospital care. *Annals of Emergency Medicine* **22**(4) 638–645.

Sprivulis, P., B. Gerrard. 2005. Internet-accessible emergency department workload information reduces ambulance diversion. *Prehospital Emergency Care* **9**(3) 285–291.

Swaminathan, J.M., S.R. Tayur. 2003. Models for supply chains in e-business. *Management Science* **49**(10) 1387–1406.

Tabachnick, B.G., L.S. Fidell. 2001. *Computer-assisted research design and analysis*. Allyn and Bacon Boston:.

Talluri, K.T., G.J. Van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Springer.

Tarek, A.H. 2006. *Reservoir Engineering Hanbook*. Gulf Professional Publishing.

Teunter, R.H., W.K. Klein Haneveld. 2008. Dynamic inventory rationing strategies for inventory systems with two demand classes, poisson demand and backordering. *European Journal of Operational Research* **190** 156–178.

Thomson Reuters. 2009. Profiles of U.S. hospitals.

Topkis, D.M. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science* **15**(8) 160–176.

Toregas, C., R. Swain, C. ReVelle, L. Bergman. 1971. The location of emergency service facilities. *Operations Research* **19**(6) 1363–1373.

Trigeorgis, L. 1999. *Real Options: Managerial Flexibility and Strategy in Resource Allocation*. MIT Press, Cambridge, Massachusetts.

UPMC website. 2010. Facts & Figures - UPMC, Pittsburgh, Pa, USA. `http://www.upmc.com/ABOUTUPMC/FAST-FACTS/Pages/FactsandFigures.aspx`. Accessed: April 10 2010.

U.S. Census. 2000a. Gazetteer. `http://www.census.gov/geo/www/gazetteer/places2k.html`. Accessed: March 29 2010.

U.S. Census. 2000b. Quickfacts. `http://quickfacts.census.gov/qfd/states/42/42003.html`. Accessed: April 1 2010.

Van Dijk, N.M., E. Van der Sluis. 2004. To pool or not to pool in call centers. *Production and Operations Management* **17** 296–305.

Van Houtum, G.J., W.H.M. Zijm, I.J.B.F. Adan, J. Wessels. 1998. Bounds for performance characteristics: a systematic approach via cost structures. *Stochastic Models* **14** 205–224.

Van Wijk, A.C.C., I.J.B.F. Adan, G.J. Van Houtum. 2009. Optimal lateral transshipment policy for a two location inventory problem. *Eurandom report* $\#2009 - 027$.

Veinott, A.F. 1965. Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. *Operations Research* **13** 761–778.

Vilke, G.M., L. Brown, P. Skogland, C. Simmons, D.A. Guss. 2004a. Approach to decreasing emergency department ambulance diversion hours. *Journal of Emergency Medicine* **26**(2) 189–192.

Vilke, G.M., E.M. Castillo, M.A. Metz, L. Upledger Ray, P.A. Murrin, R. Lev, T.C. Chan. 2004b. Community trial to decrease ambulance diversion hours: the San Diego county patient destination trial. *Annals of Emergency Medicine* **44**(4) 295–303.

Wang, Toktay B.L., T. 2008. Inventory management with advance demand information and flexible delivery. *Management Science* **54**(4) 716–732.

Weiss, S.J., R. Derlet, J. Arndahl, A.A. Ernst, J. Richards, M. Fernandez-Frankelton, R. Schwab, T.O. Stair, P. Vicellio, D. Levy, et al. 2004. Estimating the degree of emergency department overcrowding in academic medical centers: results of the National ED Overcrowding Study (NEDOCS). *Academic Emergency Medicine* **11**(1) 38–50.

Williams, R.M. 2006. Ambulance diversion: economic and policy considerations. *Annals of Emergency Medicine* **48**(6) 711–712.

Wolff, R.W. 1982. Poisson arrivals see time averages. *Operations Research* **30**(2) 223–231.

Wong, H., G.J. Van Houtum, D. Cattrysse, D. Van Oudheusden. 2006. Multi-item spare parts systems with lateral transshipments and waiting time constraints. *European Journal of Operational Research* **171**(3) 1071–1093.