2006

Capacity expansion under a service level constraint for uncertain demand with lead times

Rahul Ratnakar Marathe
Iowa State University

Capacity expansion under a service level constraint for uncertain demand with lead times

by

Rahul Ratnakar Marathe

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:
Sarah M. Ryan, Major Professor
Mike Larsen
K. Jo Min
Sigurdur Olafsson
Ananda Weerasinghe

Iowa State University
Ames, Iowa
2006

Graduate College

Iowa State University

This is to certify that the doctoral dissertation of

Rahul R. Marathe

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I gratefully acknowledge my major professor, Dr. Sarah M. Ryan, for her invaluable suport and guidance in my doctoral research. Dr. Ryan is responsible for providing the stepping stones on which I began my work towards a doctoral degree. Her unique and convincing perceptions of the field proved to be a great source of motivation; her patience and understanding proved to be great source of strength. I am truly indebted to her for this.

I am also thankful to Dr. Min, Dr. Larsen, Dr. Olafsson, and Dr. Weerasinghe for not only serving as my committee members, but also helping me on various occasions with conceptual insights.

I also wish to thank all my friends who have made my stay at ISU very enjoyable and also have supported me every step of the way.

Lastly and most importantly, words cannot express the amount of affection, encouragement and support given by my parents, my brother and my other family members during my doctoral endeavor. I owe them my heartfelt gratitude.

CHAPTER I: INTRODUCTION

Capacity can be interpreted as some upper bounds on processing quantities. Capacity is the measure of processing abilities and limitations that stem from the scarcity of various processing resources and is represented as a vector of stocks for processing units (Van Mieghem, 2003). As seen from this definition, this capacity could be production or service capacity. Capacity expansion is the addition of facilities to serve some need. Capacity expansion decisions are made on daily basis by everybody - by various industries and businesses, governments, or by individuals. These decisions could require huge investments and time to finish. For example, in January 2005, the National Thermal Power Corporation of India decided to increase its electric power generation capacity by 5,600 MW in addition to the 7,790 MW expansion already underway (The Hindu, 2005).

Failure to understand the criticality of the capacity expansion decisions could lead to disastrous results. One of the most infamous examples of the effects of inadequate capacity is the power failure in the Northeast region in August 2003. It is believed that the reasons for that were deficient transmission capacity and bottlenecks in the region. The experts were quoted as saying that there was enough generating capacity in the upstate New York region; however, there was no way to get the power generated to the New York city along the existing transmission lines. As per the report by U.S - Canada Power Outage Task Force (2004), the failure of the transmission lines caused due to the power surges was one of the major cause of the blackout. Hence, one can say that insufficient transmission capacity played a role in this power failure.

1.1 General capacity expansion problem

The primary components of the capacity expansion problem are the *sizes* of the facilities to be added and the *times* of these additions. Often the *types* of capacity added, and the *location* of the capacity to be added are also important. In addition to these primary decisions, there could be some secondary

decision variables like capacity *utilization* to be considered (Freidenfelds, 1981). To find the optimal capacity expansion policy (that is, to find the optimal values for the components of the capacity expansion problem), it is imperative that we have some idea about the demand for the various resources. In today's complex environment, it is almost impossible to know the exact demand for the resources needed in future times; however, various forecasting techniques could be used to estimate the future demand for capacity. Thus we can obtain the probability distribution of the demand at different time instances in the future. The implicit assumption in this process is that the demand can be forecasted without knowledge of the future capacity levels.

Summing up, we can say that the basic capacity expansion problem is to find an optimal policy of expansion given a particular forecast demand pattern, assuming that the relevant cost and other expansion related factors are known. We use mathematical models for real life situations and then make simplifying assumptions to obtain models that can be analyzed readily. Likewise, although the real capacity expansion problems are invariably complex, we can develop mathematical models for these. In our case, this work analyzes one such model.

Various environments exhibit the demand for resources with different patterns. For each, the capacity expansion problem can be solved by taking into consideration special properties of that demand pattern. As seen from the literature discussed in the next chapter, various capacity expansion models have been developed for different scenarios of demand patterns. In our model, we assume that demand for the resource follows a geometric Brownian motion (GBM) process. Modeling demand as a GBM process is justified because of two reasons: in many applications the demand does follow a GBM process (again, we refer to the literature discussed in the next chapter as well as the Appendix); and also because modeling with a GBM process allows the capacity shortage potential to be estimated using various financial options techniques by drawing on an analogy between the demand and the stock price (refer to Chapter 4 for details).

1.2 Problem definition

We consider a manufacturer or a service provider with certain facilities having installed capacity to produce specific products or to provide certain services. We assume that this manufacturer or service provider has an obligation to meet a certain level of service. Also, we confine our attention to a single location capacity expansion problem for a single resource assuming that the demand for that resource follows a GBM process. The capacity added does not deteriorate; that is, once the capacity is installed, we assume that it is available for infinite time. This is in contrast to some models in the literature where the capacity is perishable. We also consider the effects of economies of scale while adding the capacities. This ensures that instead of adding capacities continuously, which may not be possible in some instances, we add capacities at discrete time epochs. In fact, discounting of the total cost of capacity expansion works against the economies of scale since discounting pushes larger expansions into the future and the presence of economies of scale calls for a bigger expansion now rather than having two separate expansion projects. Also in our model, there is a deterministic expansion lead time measured from the time the capacity expansion decision is made to the time when the added capacity is actually available to satisfy the demand. We keep the expansion lead time fixed so that the sole randomness in the model comes from the demand process.

The majority of the models in the past have concentrated on the scenario when the capacity expansion project starts before (or immediately when) the demand for the resource reaches the capacity position. By capacity position, we mean the capacity that will be available after any current expansion project is completed. However, we envision cases where the service provider may want to start the capacity expansion after the demand for resource reaches the capacity position. This delay could occur because of the specific problem parameter values observed in particular industries. This delay could arise because the required service level is not very high. Also, this delay could give the service provider more time to observe the demand before committing to the capacity expansion. Thus, the delay in the start of an expansion project gives the service provider another choice that would be

made according to the tradeoff between the total cost of taking this option and subsequent capacity shortages incurred because of this choice. Hence, depending on the total cost incurred and the level of service achieved, the service provider has to make a decision of whether to start the capacity expansion before or after the current capacity position has been reached. As mentioned earlier, the recent models on capacity expansion concentrate on the former case where the expansion project starts before, or immediately when the demand reaches the capacity position, and minimize the total cost involved in the expansion. Our model formulates a more general situation, which not only includes the former case but also allows for the latter case. We find the total cost of capacity expansion when the expansion starts with certain shortages and then investigate the conditions under which this delay is suitable and economically favorable. We define the service level to be maintained by the service provider as a proportion of demand over each expansion cycle that is satisfied with the available capacity. This permits us to use the concepts of service level used in the inventory theory for various production systems.

The goal, then, is to determine the optimal timing and sizes of the future capacity expansion projects that minimize the infinite time horizon total cost under the constraint that the service provider has to maintain a certain service level.

1.3 Applications of the capacity expansion problem

As mentioned earlier, decision makers in various fields are faced with the capacity expansion problem. As we see below, capacity expansion problems are encountered in the areas of personnel planning- staff hiring and training, where the resource for which we are expanding the capacity is the human resource; and in planning for natural resources like water. Capacity expansion issues are critical also in the high technology industry. In the succeeding sections we discuss some of these problems and present real life examples through published works in the respective areas. From the

discussion of these applications, we see that some of these problems are very similar to the one we are considering (Section 1.2).

1.3.1 Capacity troubles in the high-technology industry

Supply problems are not uncommon in the high-tech gear market. Numerous manufacturers are constantly vying for a limited number of components, such as display screens and memory. This makes it tough for manufacturers caught off guard by unexpected consumer demand to quickly increase production.

The beginning of the 21st century saw slowing profit growth of the companies like Sony, Motorola, and HP. This slowdown was, in fact, not because of the reduction in the consumer and business demand for the mobile phones, personal computers and printers. These products remained the market favorites keeping the demand levels high. However, the ability of the companies like Sony and Motorola to produce these products was not enough to meet the demand (Shameen, 2000).

A deep and prolonged chip shortage can push up production costs, which are eventually passed on to end-users. That could mean the end of the price advantage for the consumers. Examples of these could be seen in the 30% price increase by Japan's Sharp for the flash-memory chip that is used by the makers of the mobile phones, TV sets and personal digital assistants (PDAs). And these are not the only industries dependent on the chip: hand-held PCs, digital cameras, video-game consoles and other must-have digital products also need these flash-memory chips.

In fact, it is not just the supply shortages that hurt the bottom line of any company in this field. Sometimes, there are excess capacities in the market forcing the chipmaker to sell the product at a price below its production cost. In the 1980s and '90s, DRAM (Dynamic Random Access Memory) overproduction hit the industry every three or four years. DRAM chips are used in computers. "This is such a cyclical business that few people can accurately predict when there will be a supply-demand

imbalance," said Jonathan Dutton of UBS Warburg Securities in Seoul (Shameen, 2000). It is known that the tech market is plagued by vicious boom-and-bust cycles. And the result is that the prices of the memory devices have been known to plunge by 75 percent or surge three-fold in just months (Kanellos, 2004).

PalmOne's mobile phone model Treo 600 was facing shortages of liquid crystal displays (LCDs) (Shim, 2004). According to the analysts, there was enough glass supply in the market to meet demand, but supplies of components such as a backlight, color filters and drivers weren't as abundant. Also it was being predicted that this shortage could last as long as two years. Similar LCDs were used in the TV sets, which was much more lucrative market for the LCD screen manufacturer. This shortage in supply of the LCDs would result in shortages in supply of the Treo 600 leading to long waits for the customers of the Treo 600, which had received excellent reviews in the market. This also meant that there was a possibility of a competitor eating into the space of PalmOne.

1.3.2 Water resource planning

Similarly, the capacity expansion problem is encountered in the area of water management. Wollman (1976) presents a detailed review of various models dealing with the supply-demand problems of water resources. The models to which the paper refers not only account for the physical, biological and chemical requirements of the water management, but also focus on achieving economic optimization to the problem. In fact, various papers in the collection by Thrall (1976) give an interesting insight of mathematical models dealing the issue of water resources. Erlenkotter (1976) formulated a simple model for the water resource problem clearly demonstrating the close interrelationship between the scales and sequencing decisions for the water resource projects. The problem was solved using a simple formula for economies of scale and under the assumption that the water capacity could be added instantaneously.

In our capacity expansion model, we apply a similar approach, where we find the optimal size and timing factor for the expansion project that minimizes the total cost of capacity expansion.

1.3.3 Workforce planning

As stated earlier, capacity expansion means addition of resources to meet a particular need. That resource could be a physical product like the DRAM in the example from Section 1.3.1, or it could be human resource. Hence planning for human resources could be considered as a capacity expansion problem. There have been numerous models proposed for manpower planning. In their seminal book, Holt et al. (1963) presented the case for the importance of workforce planning in business decision-making. According to the authors, workforce planning was one of three ways of managing the stochastic nature of the demand; the other two being optimizing the production rate by changing hours of operations and optimizing inventories and backlogging. They also formulated a total cost minimization problem subject to the inventory balance constraints in each time period. Edwards and Morgan (1982) surveyed various manpower planning models- viz. Markov models, renewal models, etc., and applied optimal control theory to the general mathematical formulation for the problem. Young and Abodunde (1979) presented a linear programming model to investigate the consequences of controlling recruitment policies over fairly long periods of time. They assigned penalty costs for both under- and over-production to produce optimal long-term recruitment policies. Dellaert and de Kok (2004) considered a multi-stage periodic review made-to-stock assembly system with stationary stochastic demand. It was a capacity-planning problem where the capacity considered was the human capacity. They formulated a cost minimization problem where in addition to the inventory holding and penalty costs, relevant capacity costs were considered. The problem was solved under the constraint that the specified customer service level-- in terms of the probability of no shortage-- is to be achieved. The capacity was considered flexible in the sense of hiring temporary workers from an

external labor supply agency and/or subcontracting. In this model it was assumed that the temporary labor is available immediately. For both the push and pull types of production systems, two different approaches were tried. The paper first considered the resource and production separately and determined the order-up-to quantities and afterwards found the best mix of regular and temporary workforce. Also a second, integrated, approach was presented where these variables influence each other. We model the capacity expansion problem under a similarly broad perspective of minimizing the total cost of expansion under a service level constraint. Tan and Alp (2005) considered a similar periodic review made-to-stock production environment with non-stationary stochastic demand. A finite time horizon dynamic programming cost minimization problem was formulated where both the amount and capacity of production in each time period were decision variables. As with Dellaert and de Kok (2004), the capacity was considered in terms of human resources with the flexibility of temporary hiring. The authors note that changing the level of permanent capacity as a means of coping with demand fluctuations, such as hiring and firing of permanent workers, could be very costly. And in cases where the demand is highly volatile, it could have a very negative impact on the company. If the permanent capacity were increased following a number of high demand realizations, a stream of low demand realizations might result in a costly decrease in capacity. If this were followed by yet another stream of elevated demand, the result would be expanding the capacity that had been contracted. Hence, the authors consider flexible temporary capacity, in this case a temporary workforce, to meet the demand. The workforce shortage problem is encountered in the various areas, viz. airline pilot shortage, healthcare physician shortages, etc.

1.3.3.1 Workforce planning: Healthcare

Chan (2003) detailed a problem the Canadian healthcare medical establishment was facing. It was reported that the physician shortages were going to worsen because of the aging population and the retirement of 'baby-boomer' physicians. The report discussed the various assessment studies carried

out regarding the problem. One of the supply and demand studies included various approaches such as service level (productivity) achieved by the workforce, utilization analysis of the resource (physicians), etc. One of the solutions proposed was review of the International Medical Graduates (IMG) policy to meet the future demand of/need for physicians. Buchan and Edwards (2001) focused on the shortage of nurses in Britain. According to the authors, the main reasons of this shortage were not only demand increase, but also aging of the nurses and the dwindling pool of potential nurse 'returners' (former nurses returning to paid employment). The authors proposed integration of efforts at various levels to better plan the workforce. Integration is required between workforce planning and operational planning; also amongst various groups of workers (nurses, doctors, and other medical employees). The paper also stressed the need for a new pay system to attract more employment. O'Brien-Pallas et al. (2001) also addressed issues of integration in healthcare workforce planning. This report mainly focuses on two strategies. First was the Integrated Healthcare Human Resource Planning (IHHRP) that determines the number, mix and distribution of health providers that would be required to meet population health needs. This type of planning was for a long range of time. The second strategy was service planning. This short-term planning was aimed at ensuring that resources of healthcare are allocated and managed in an efficient manner, and was concerned with the number and type of health resources allocated amongst different sectors and between human and physical capital.

1.3.3.2 Workforce planning: Airline pilots

Hopkins (2001) presented a chronological account of the airline pilot problem. According to Hopkins (2001), the pilot shortages could be attributed to several reasons, principal among them being:

- The post-1993 economy boom, which caused huge growth in the air traffic,
- Also because of regulations in the industry, the retirement of old pilots was peaking,

- Amazing job growth in other sectors of economy siphoned off people who might otherwise have chosen flying careers,
- Military downsizing, which began in the late 1980s and accelerated after the Gulf War, meant restricted pilot 'production.'

Any shortage in the resources like pilots results in a response by the operators, which is to change the schedules, increase the ticket price, or increase block times between city-pairs (Donohue, 2000). And the results of these could be serious for passengers and industry in general-- reduced access, increased prices, reduced convenience, and somewhat increased delay. This critical shortage of airline pilots had an adverse effect on the air service in rural areas such as Alaska and parts of the upper Midwest (Woerth, 2000). The effects of pilot shortages on the rural air service were studied in detail by Barker (2000). The authors found that:

1. Emergency medical services use airplanes to fly doctors to some rural locations (like parts of Wyoming, and Colorado). Pilot shortage threatens this expansion of medical services to the rural parts.
2. Commerce and economic viability of communities are dependent on the access to air transportation. A pilot shortage severs this link.
3. Finally, the high value cargo, mail and express package services provided to the communities across the country are directly affected by the ability to have pilots able to safely operate the planes.

There have been many solutions proposed to tackle this problem. Some of them are on the policy level and others are local operator level. Examples of some of policy level solutions include relaxing the Age 60 Rule, which would increase the mandatory retirement age of air carrier pilots from 60 to 65; and changes in the flight and duty time regulations and reserve rest requirements, which would essentially reduce the time the pilots have for their rest (Woerth, 2000). As can be seen, there were inherent disadvantages associated with each of these suggestions.

Some of the local solutions proposed have been formulation of mathematical models at the operator level to reduce the risks of the pilot shortages, like the one proposed for Continental Airlines (Yu et al., 2004). Here, a new decision-support system called "Crew ResourceSolver" was developed to obtain an optimal solution for the large-scale pilot staffing and training problem. The system solves a mixed integer program for a total cost minimization problem under the constraints of capacity limitations on the training facility, pilot vacations, and maintaining pilot seniority.

As seen in the next chapter, the deseasonalized total airline passenger enplanements in the US over the 20-year period from 1981 to 2001 have followed a GBM process. Hence planning for resources to fulfill the demands of airline industry could be considered as a typical example for the problem defined in Section 1.2. One of the important resources in meeting the enplanement service level is the total number of airline pilots hired by the industry. Therefore, an airline company that wants to meet the service level target of providing for the certain percentage of the total enplanements, will have to solve a capacity expansion problem similar to the one mentioned in Section 1.2. This problem for the airline company will be more specific in the sense that capacity expansion means hiring of new pilots. The lead time could then be considered as the training period for the pilots. And since the time required to train pilots would be the same regardless of the number hired if the training sessions were in a classroom environment, they can be assumed to be constant. The problem we are considering (in Section 1.2) is more restrictive than the pilot shortage problem in the sense that we are considering only *hiring* of the pilot and not *firing* them when the demand for the airline seats (characterized by enplanements) dips. Here, we also note that union contracts frequently limit airlines ability to lay off excess employees. In this scenario, the airline pilot shortage problem would serve as an ideal example for the capacity expansion problem considered in this dissertation.

1.4 Research objectives

It is imperative that the optimization problems modeling critical issues of today be able to take care of the randomness of the today's market environment, because market volatility is going to affect the problem parameters and instead of ignoring this volatility, business decisions should be made considering this variability. In this dissertation, we are solving one such problem that deals with optimal planning of capacity where the demand for that capacity is stochastic. The presence of the expansion lead time further complicates our problem. The analysis of this capacity expansion problem demonstrates how the randomness of the demand process affects the optimal expansion policy for the service provider. The details of the problem, the assumptions and the constraints are described in Section 1.2 of this chapter. Our problem is broad enough that it can be applied to any industry as long as the demand for that industry follows a particular distribution, the expansion lead times are fixed and economies of scale exist. The decision variables of our problem are the timing and size of the future expansion projects. The research objectives then are as follows:

- Express the objective function and the service level constraint in terms of decision variables.
- Apply the financial option pricing theory to the service level expression to simplify the constraint and express it in terms of the timing and size variable.
- Formulate the capacity planning optimization problem.
- Find an optimization technique to solve the formulated optimization problem.
- Perform numerical analysis and draw managerial insights from the optimal solution.
- For given conditions, select the capacity planning policy parameters that minimize the total expansion cost.

Our problem provides optimal policy parameters for a service provider so that the total expansion cost for the service provider is minimized. Using the parameter values for any particular case, the service provider can then find out optimal starting times of the future capacity expansion

projects and also the sizes for each of those projects. Using our model the service provider will be able to determine conditions under which different expansion policies are appropriate.

1.5 Thesis organization

In the next chapter, we review the relevant literature for our model. The actual model is discussed in detail in Chapter 3. Mathematical analysis of the model is conducted and relevant expressions derived in Chapter 4, based on which some of numerical results are obtained in the subsequent chapter (Chapter 5). We propose the future work regarding the model in the final chapter of the dissertation (Chapter 6).

CHAPTER II: LITERATURE REVIEW

To begin with, we review the literature regarding the assumption of demand for capacity following a geometric Brownian motion (GBM) process. Later on in this chapter, we discuss various analytical models regarding the capacity expansion models in the literature and their relevance in our model. Since we use some concepts from financial option pricing and the production and inventory theory, we shall also discuss some of the important works from the literature regarding these respective fields. A cutting plane algorithm was used to numerically solve the capacity expansion optimization problem. We review the literature related to this important optimization technique. We conclude this chapter by summarizing the relationship between the literature discussed in this chapter and our model.

2.1 Validity of the GBM process assumption

Many recent engineering economic analyses have relied on an implicit or explicit assumption that some quantity that changes over time with uncertainty follows a GBM process. Below we briefly review a number of applications in different areas. The GBM process, also sometimes called a lognormal growth process, has gained wide acceptance as a valid model for the growth in the price of a stock over time. In fact, Hull (1999) refers to it as "the model for stock prices". Many recent examples of GBM models have arisen in real options analysis, in which the value of some "underlying asset" is assumed to evolve similarly to a stock price. In some cases, the GBM assumption is stated explicitly, while in others it is implicitly used when options are evaluated by the Black-Scholes formula. Nembhard et al. (2002) quantified the cost of applying quality control charts using real option pricing methods, where both the sales volume and the price of a product were assumed to follow GBM processes. Thorsen (1998) applied the real options theory to decisions of establishing a new forest stand and it is assumed that the future net prices of roundwood follow a

GBM process. The GBM model has also been used to represent future demand in capacity studies. Whitt (1981) studied capacity utilization over time assuming demand followed a GBM. An indirect validation of the assumption was provided by Lieberman (1989), which showed in an empirical study of the chemical industry that actual capacity utilization matched the predictions from the model proposed by Whitt (1981). Ryan (2004) assumed that the demand for services in rapidly growing industries follow a GBM and the expansion policy to minimize cost subject to a service level constraint was developed and analyzed. Marathe and Ryan (2005) verified empirically that the demand for electric utility in the US as well as the number of passenger enplanements in the airline industry follow a GBM process (also see the Appendix).

2.2 Capacity expansion problem

The capacity expansion problem is an extensively researched topic. As Van Mieghem (2003) mentions, there are over 15,000 articles with "capacity" in the title or keyword. Focusing primarily on the effects of resource scarcity and uncertainty over capacity decisions, Van Mieghem (2003) reviews the strategic capacity management literature concerned with determining the sizes, types, and timing of capacity investments and adjustments under uncertainty.

Most of the models found in the capacity expansion literature aim at minimizing the total expected discounted cost over a finite or infinite time horizon. One of the seminal work in the area of mathematical modeling of the capacity expansion problem was by Manne (1961). He proposed a model to decide the expansion sizes in cases where the demand follows a linear deterministic or random-walk pattern; also, the effects of economies of scale and penalties for demand not being satisfied were considered. He also showed that the stochastic problem is equivalent to a deterministic problem with just a small adjustment in the interest rate value. Smith (1979) analyzed the addition of capacity from a finite set of available possible additions for a case of exponential demand. A turnpike theorem was developed which gives the structural characteristics of the optimal policy. Smith (1980)

also considered the problem with exponential demand growth. Here the author developed a general formulation of the deterministic capacity expansion problem and proved that capacity models found in literature like Manne (1961) is a special case of this general model. Whitt (1981) analyzed the capacity expansion problem from the perspective of estimating the capacity utilization. Using results for stochastic clearing processes he obtained the stationary distribution function for utilization under a particular expansion policy when demand follows a GBM process. The long-term expected utilization depends on both the size and timing parameters, as will be discussed further in Chapter 4. The effect of the study horizon length on the solution to the capacity expansion problem was considered by Bean and Smith (1985). They developed an algorithm to determine the length of the horizon needed to identify an optimal first facility to install. A generalization of Brownian motion demand was considered by Bean et al. (1992), where demand was assumed to be either a nonlinear Brownian motion process or a semi-Markovian birth and death process. Like Manne (1961), they showed that the problem can be transformed into an equivalent deterministic problem and that the effect of uncertainty in the demand is to reduce the interest rate.

All of the preceding results relied on the absence of lead times to rule out unplanned shortages and obtain a regeneration point structure. In contrast, Davis et al. (1987) considered a capacity expansion problem to find optimal timing and sizes of future expansion where the demand was a random point process (that is, the demand increased by discrete amounts at random times). The lead time considered in this paper depended on the rate of investment. The authors applied stochastic control theory to find the optimal expansion policy. Buzacott and Chaouch (1988) examined the effects of demand plateaus on the capacity expansion problem. In their model, they assumed the demand process to be an alternating renewal process where a period of linear growth is interrupted by plateaus that occur at random times and last for an uncertain duration. However, they did not consider the effects of lead time. Chaouch and Buzacott (1994) examined the same problem including lead times and also considered two cases where the capacity addition started, respectively, before and after

the demand reached the current capacity. Assuming proportional shortage costs, they set up a total cost minimization problem resulting from an infinite horizon dynamic programming formulation. Our work is similar to theirs in the sense that we also consider initializing the capacity expansion after a certain deficit has been accumulated; however, the demand process considered in our model is different. Also we use a service level constraint rather than assigning a proportional shortage cost. Other work combining uncertain demand with lead times includes that of Cakanyildirim and Roundy (2002), who developed an expansion/contraction algorithm to compute the optimal capacity expansion and contraction times for situations when the demand first stochastically increases and later stochastically decreases. This type of model has applications in the semiconductor industry and the electric utility industry. Also, Angelus and Porteus (2003) considered the problem of deferring a capacity expansion project under the conditions of echelon capacity in a discrete time, finite horizon model and with multiple resources.

Ryan (2004) considered a fixed lead time for expansion when demand followed a GBM process. A timing policy was developed to provide a specified level of service. It was shown how the parameter of the timing policy could be obtained numerically using some concepts from financial options pricing. Our work extends this model. While Ryan (2004) assumed that the next expansion starts before the current capacity position is reached, in this model we consider a case where the next expansion can also be started after the accumulation of some shortages relative to the capacity position. Pak et al. (2004) considered a capacity expansion problem for exponential demand and studied the effect of technology improvement on the optimal timing and sizes of the capacity expansions. They also considered a fixed expansion lead time in their capacity expansion model.

2.3 Financial option pricing theory

Application of the financial options theory to problems such as stochastic optimization is a relatively new field of research. A very good reference in this area is Birge (2000). In this paper the author

applied the basic principles of risk-neutral valuation to general forms of constrained resource problems, such as capacity planning. Using the results from options pricing theory, he showed that risk could be effectively accounted for in a wide range of operational planning models, particularly the linear capacity planning models. In our model, the potential for capacity shortages can be compared to the barrier options in finance, in particular, the up-and-out call option. As defined by Musiela and Rutkowski (1997), the generic term "barrier options" refers to the class of options whose payoff depends on whether or not the underlying prices hit a pre-specified barrier during the option's life. The idea of these options was discussed as early as the 1970's by Merton (1973), and Goldman et al. (1979), who analyzed "path dependent options". Rubinstein (1991), and Rubinstein and Reiner (1991) arrived at analytical formulas for various types of barrier options as a limiting case of a discrete time model. A unified and intuitive mathematical foundation for the barrier option pricing formulas was given by Rich (1994). This work not only included key results to analyze the barrier options but also gave all the necessary derivations. Ritchken (1995) gave the equations for barrier option pricing using the binomial and trinomial lattice model. The paper also included cases where the barrier is an exponential function of time. Carr (1995) discussed two different extensions of barrier option pricing, and derived an expression for the price of the barrier option. The price of the barrier option was found from the joint distribution of a Brownian motion and its maximum in Chuang (1996). This paper found the equation for the joint distribution of the Brownian motion and its maximum when the time intervals considered for the Brownian motion and its maximum are different. This result was then used to find the price of 'partial' barrier options (Musiela and Rutkowski, 1997) – that is, barrier options in which the underlying price is monitored for barrier hits only during a prespecified portion of the option's lifetime. The paper also included some remarks about reducing the numerical computations by a clever change of variables. Similar results about the partial barrier option were obtained by Heynen and Kat (1997). They gave analytical expressions for all cases of barrier options viz. cash or nothing, asset or nothing, etc.

2.4 Measures of service level

In our model, we measure the service level in terms of the average proportion of demand that is satisfied over time. We aim to find the parameters for an assumed expansion policy that will achieve a specified level of service. This broad definition of 'service level', as we use it, has its roots in the continuous review inventory models. Hadley and Whitin (1963) presented an extensive work regarding the continuous review inventory models. All aspects of the various inventory models were discussed in detail and shortages were analyzed thoroughly. A study of service levels was carried out by Klemm (1971). Rigorous definitions for the three types of the service level, viz. α, β, and γ service levels were given for (s, S) and (r, Q) type inventory models. The α service level is defined as the probability of not being out of stock at an arbitrary time. This service level is more common in cases where the time is measured in discrete time periods. However, with this service level one doesn't know how large a part (quantity) of the demand is expected to be satisfied. The β service level is given as the fraction of demand not being satisfied per unit time. This definition of the service level is more suited for the capacity expansion problem we are considering. Lastly, we note that the definition of the γ service level is similar to that of the β service level. The only difference between the β and γ service levels is that instead of considering just the unsatisfied demand per unit time, the γ service level considers cumulative unsatisfied demand. Hence this service level is more relevant in case where backorders are being considered. Because services generally cannot be backordered, we use the β service level in our formulation of capacity expansion problem. Further mathematical calculations for each type of service level and its effect on the order points in the various inventory models was done by Schneider (1981). In this work, the analysis of various service levels was carried out with equations for each given under the assumption that the demand follows a Poisson process.

2.5 Cutting plane algorithm

Our capacity expansion problem is an infinite time horizon cost minimization problem under the service level constraint. In Chapter 3 we reduce it to a nonlinear optimization problem in two continuous variables. Because of the complexity of the service level constraint, we tried to solve the dual of the original capacity expansion problem. However, since this also proved to be difficult to solve, we solved the optimization problem using the cutting plane algorithm as an approximation to the dual problem. The cutting plane algorithms have proved to be computationally efficient and work under rather general assumptions (Kelley, 1960; Wolfe, 1961). Kelley (1960) is the premier reference on cutting plane algorithms applied to non-linear programming problems. Before this, Gomory (1963) had proposed a cutting plane method to solve integer programming problems. (We note that though Gomory's method originally appeared in 1959 as a Princeton-IBM Mathematical Research Project technical report, a formal paper in a technical journal appeared in 1963.) Wolfe (1961) proposed a different way of generating cuts than Kelley (1960), thereby improving the efficiency and ensuring faster convergence of the algorithm. The basic concepts and steps involved in implementing the cutting plane algorithm were described by Bazaraa et al. (1993). Zangwill (1969) proved the convergence of the cutting plane algorithm under various conditions. He also discussed the convergence of various versions of the cutting plane algorithm. Atlason et al. (2004) used the cutting plane algorithm to solve a call center staffing problem. As in our case, the optimization problem in their case was also constrained by the specified service level. They simulated the service level achieved by the different staffing plans. This paper proved the convergence of the cutting plane algorithm for a case where the constraint equation is obtained via simulation. Also, this paper proposed a convenient numerical method for checking convexity of a function.

2.6 Summary

From the literature discussed above, we found that there are very good mathematical models proposed for solving capacity planning problem under various scenarios. However, all of these models are different from the problem we are solving (described in Section 1.2 of Chapter 1). The demand process we are considering is similar to the one considered by Manne (1961) and Whitt (1981); where Manne (1961) considered a random-walk type of demand, Whitt (1981) considered exactly the same demand (GBM) process as in our model. However, neither of these models considered any expansion lead time. Moreover, Whitt (1981) assumed the demand to be following the GBM process under no explicitly stated reasons. We present cases where the demand for the capacity does follow the GBM process and also discuss a method to find whether the demand follows this assumption. Ryan (2004) proposed a capacity expansion model where the demand was assumed to follow a GBM process and where the effects of expansion lead time were considered. However, the new capacity expansion was initiated before (or immediately when) the demand hits the capacity position. Our model is broader in the sense that we keep the option of starting time of the expansion project open. Based on the trade-off between the service level achieved and the expansion cost incurred, the service provider in our model may initiate the new expansion project either before or after the demand reaches the capacity position. Chaouch and Buzacott (1994) considered both of these options in their capacity expansion model and proposed the conditions under which the service provider should adopt the policy of initiating the expansions either before or after the demand crosses the capacity position. However, in their model, the demand process considered was different than our assumption of the GBM process. They considered the linearly growing demand with plateaus occurring at, and for, random times. Like Ryan (2004), because of the assumption of the GBM process demand, the service level constraint in our model could be formulated using the financial option pricing theory. Where Ryan (2004) used the European call option price expression to formulate the service level constraint, we use the partial barrier call option price formula. In our study of the literature on applications of financial option

pricing theory to optimization problems, we could not find any other instance where the partial barrier option pricing concepts were used. Once the optimization problem for capacity planning has been formulated, it can be solved using various optimization techniques. Atlason et al. (2004) formulated a call center staffing problem under a service level constraint. As in our model they used the cutting plane algorithm to solve their problem. However, the service level expression in their model was evaluated by simulation of the demand process. In our case, we use financial option pricing theory to arrive at an analytical expression for our constraint.

In the next chapter, we discuss the details of our capacity expansion model and formulate the optimization problem.

CHAPTER III: GBM ASSUMPTION AND THE MODEL FORMULATION

The brief overview of the problem statement was presented in Chapter 1. In this chapter, we discuss the details of our capacity expansion model and the related assumptions. As discussed in Chapter 1, one of the major assumptions was that the demand for capacity faced by the service provider follows a GBM process. There are conditions under which this assumption could be justified. We begin this chapter with a discussion on the GBM process assumption. Later, with this assumption about the demand for the capacity, we detail the basic model environment and mathematical formulation. In this chapter, we formulate the service level constraint expression and the infinite time horizon expansion cost objective function for our optimization problem.

3.1 GBM assumption

Let $B(t)$ be a Brownian motion having drift μ and volatility σ^2 with $B(0) = 0$. The basic characteristics of a Brownian motion are discussed in details in Appendix I (Marathe and Ryan, 2005). Demand for the product or service is given by the GBM process $P(t) = P(0)e^{B(t)}$. We say that the variable $P(t)$, $0 \leq t < \infty$, follows a GBM (with drift parameter μ and volatility parameter σ) if, for all nonnegative values of k and t, the random variable $\frac{P(t+k)}{P(t)}$ is independent of all values of the variable up to time t and if in addition, the log ratio $w(k) \equiv \ln\left(\frac{P(t+k)}{P(t)}\right)$ has a normal distribution with mean μk and variance $\sigma^2 k$, independent of t, where μ and σ are constants. Define $\gamma \equiv \mu + \frac{\sigma^2}{2}$ as the mean (exponential) growth rate of the demand. This assumption that the demand for the services or product follows a GBM process may be checked using the procedure developed by Ross (1999) and discussed and applied by Marathe and Ryan (2005).

The assumption of a GBM process for demand may be reasonable in cases where (a) the demand growth during a period, as a percentage of total demand, has a stationary lognormal distribution over time, and (b) successive growth percentages are independent. That is to say that, referring to the definition above, there are two assumptions to be satisfied for any time series data to follow the GBM process (Ross, 1999):

-Normality of the log ratios ($w(k)$) with constant mean and variance,

-Independence from previous data (log ratios independent of their past values).

However, before we test the normality and independence of the log ratios, any seasonal variation should be removed from the data. Marathe and Ryan (2005) examined two ways of deseasonalizing the time series and found that the moving average method is an unbiased method. For details of the deseasonalization process and GBM process fit, see Appendix I, which includes the full Marathe and Ryan (2005) paper.

Marathe and Ryan (2005) found that historical usage of electric power and airline travel met both conditions – of normality and independence of log ratios – after seasonal effects were removed. On the other hand, although data availability limited the statistical tests that could be applied, the conditions were not met by time series that could serve as proxies for the demand for Internet and mobile telephone service due to their declining growth rates.

Summarizing, we can say that any demand data series needs to be tested before the GBM process assumption is made for the demand. Given a time series representing the demand for service or product, if the log ratios of the time series values are normally distributed and are independent of each other, then according to method prescribed by Ross (1999) and discussed by Marathe and Ryan (2005), the time series is consistent with observations of a GBM process. Our model can then be assumed to be applicable in such cases, where there is sufficient evidence of the demand being a GBM process.

3.2 Capacity expansion model and policy parameters

We assume that capacity additions occur at discrete time points and that a fixed lead time of L time units is required to install new capacity. The problem is to choose a sequence $\{(T_n, X_n), n \geq 1\}$, where T_n, the time when the n^{th} capacity expansion starts, is a stopping time with respect to the Brownian motion $B(t)$ and X_n is the n^{th} increase in capacity. For any realization ω of the Brownian motion $B(t)$, let $t_n \equiv T_n(\omega)$. Let K_n be the installed capacity after n additions are completed, where the initial capacity is K_0. Then,

$$K_n = K_0 + \sum_{j=1}^{n} X_j .$$

The installed capacity at time t is given by,

$$K(t) = \begin{cases} K_0, & 0 \leq t < t_1 + L \\ K_n, & t_n + L \leq t < t_{n+1} + L, \; n \geq 1. \end{cases}$$

The capacity position at time t is given by,

$$\Pi(t) = \begin{cases} K_0, & 0 \leq t < t_1 \\ K_n, & t_n \leq t < t_{n+1}, \; n \geq 1. \end{cases}$$

We assume that the policy proposed by Whitt and Luss (Whitt, 1981) for the same demand function is modified to account for the lead times and its parameters are adjusted to allow planned shortages to occur. Whitt (1981) showed that, without lead times, their policy results in a stationary distribution for the capacity utilization and provided a simple formula for its expected value. In the Whitt-Luss policy, each new expansion occurs when demand reaches some fixed proportion ($p < 1$) of current capacity, and after its instantaneous addition, the new capacity is a constant proportion of its previous value. Likewise, in our model, we assume that each expansion occurs when demand reaches some fixed proportion, p, of the capacity *position*, and that the new capacity is a certain proportion, v, of the old capacity: $K_n = vK_{n-1}$, where $v > 1$. Ryan (2004) showed that for $p \leq 1$ with

fixed lead times, the value of p to attain a specified service level can be found according to the Black-Scholes formula for pricing a European call option. Moreover, assuming this timing policy is followed, the expansion size policy minimizes the infinite horizon discounted cost under a widely used expansion cost function that reflects economies of scale. In our model, we consider the case of $p \leq 1$, and also allow for the case where $p > 1$. According to our model, the n^{th} capacity expansion starts when the demand reaches the level pK_{n-1} and the expansion takes the capacity position to the level of vK_{n-1}.

The policy assumes that ever-increasing increments of capacity can be installed within the same lead time to keep pace with exponentially growing demand. This assumption is most reasonable in industries where capacity bottlenecks are caused by facilities subject to continuous technological improvement, such as those that rely on information and communications technology. However, it may hold in more traditional situations as well. For example, in an empirical study of the chemical product industry, Lieberman (1989) found that the Whitt-Luss policy provided the closest fit among several alternatives to the capacity utilization. Over at least two decades, total output grew by an average of 6.2% per year, and the mean size of expansion increments translated to a value of $v = 1.09$ at the plant level.

Figures 1, 2, 3 and 4 illustrate the policy and potential shortages seen at the realized time t_n, when demand first reaches the level pK_{n-1}. The n^{th} capacity expansion has just started. With this expansion, the total installed capacity will reach level K_n after the lead time L. As stated earlier, we model the situation wherein the manufacturer has a choice of waiting until certain amounts of capacity shortages are accumulated before starting the next expansion project. This "certain amount of shortages" is represented by the variable p, $p>1$, as in Figures 3 and 4. However, when we consider the other case, $p \leq 1$, as in Figures 1 and 2, the service provider starts the new capacity expansion before (or immediately when) the demand reaches the current capacity position. In our model this

parameter 'p' is a decision variable. By allowing it to take values either less than or greater than or equal to 1, we are keeping both the options open and making our model broader. Depending on the other model parameter values, the optimal p value could be greater than or less than one, thus indicating whether it is optimal to start the expansion after a certain amount of shortages have been accumulated, or whether to start the expansion project before the demand hits the capacity position. In either case, since the new capacity position is K_n, the next expansion would start at the time when the demand $P(t)$ first reaches the position pK_n. Since the demand process is stochastic, this time for the start of the next expansion (T_{n+1}) is a random variable. The second decision parameter is the size of the expansion $v \equiv K_{n+1}/K_n$.

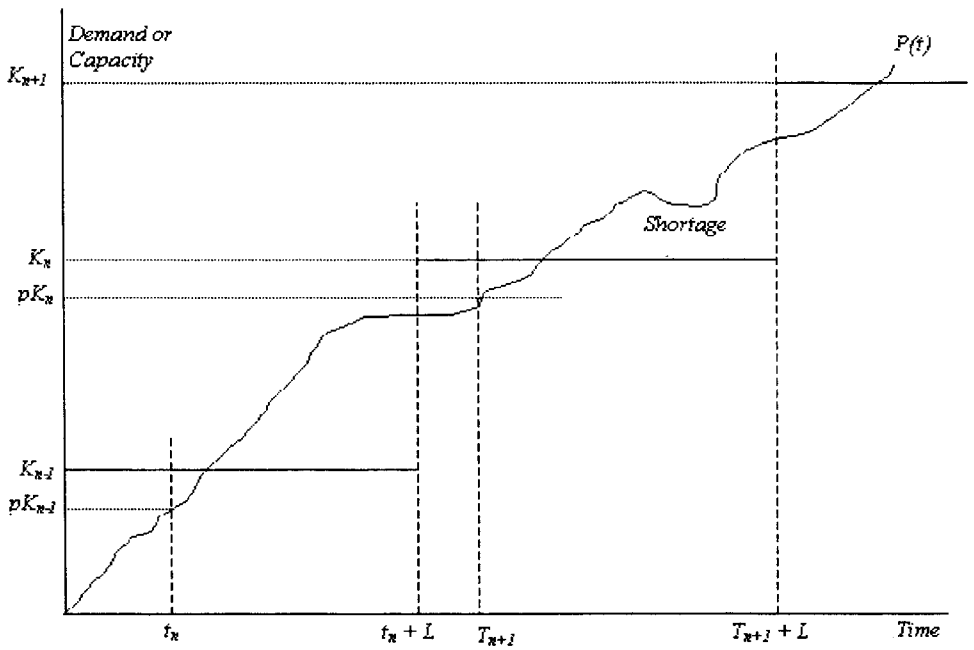

***Figure 1**. Capacity expansion policy when the expansion starts before the demand reaches the current capacity position ($p < 1$) but after the end of the current expansion cycle.*

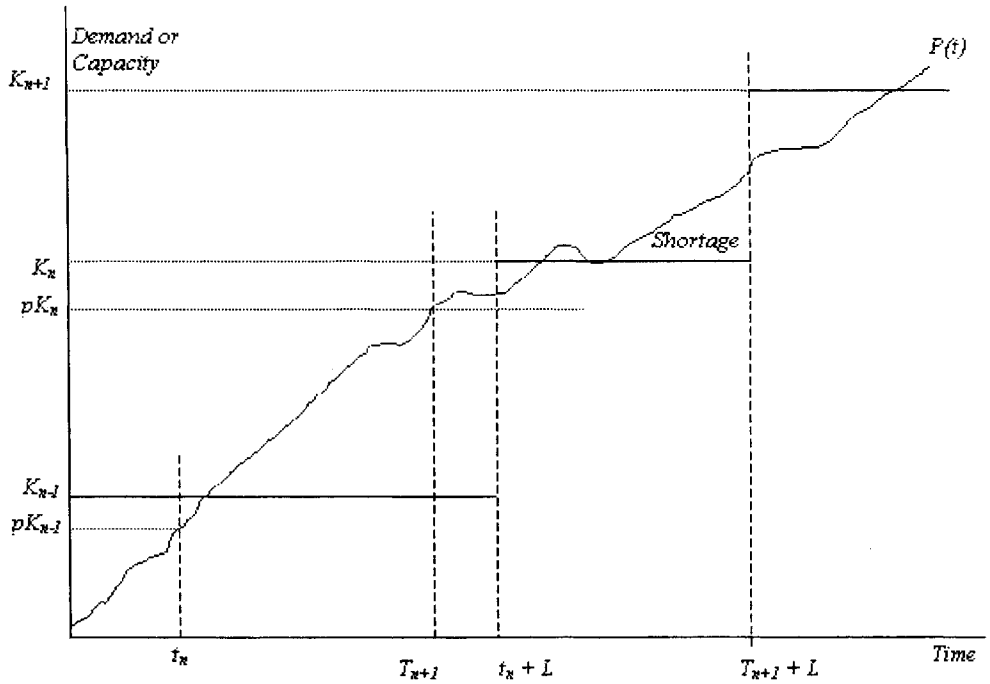

***Figure 2**. Capacity expansion policy with $p < 1$ when the expansion starts before the end of the current expansion cycle*

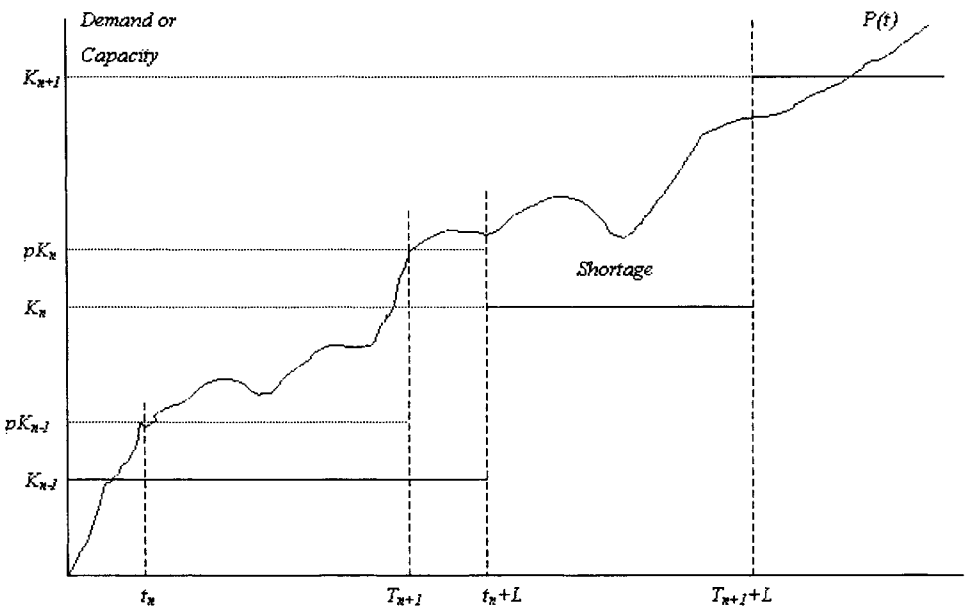

***Figure 3**. Capacity expansion policy when the expansion starts after some initial shortages ($p > 1$), but before the end of the current expansion cycle.*

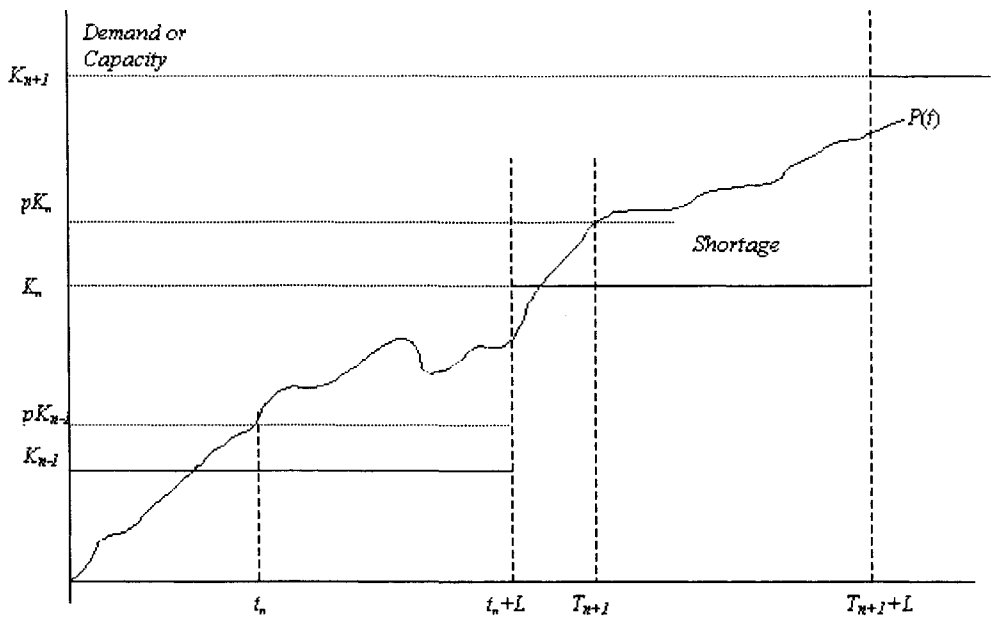

Figure 4*. Capacity expansion policy with p > 1 when the expansion starts after the end of the current expansion cycle.*

Having decided on the desired service level, then the problem is to find the values of the two decision variables that minimize the infinite horizon discounted cost of maintaining it.

3.3 Formulation of the service level expression

It is natural to define a cycle as the time interval from the end of one lead time to the end of the next, so that the actual capacity is constant over the cycle. For a generic cycle, we formulate a service measure akin to the fill rate used in periodic Sobel (2004) and continuous review (Hadley and Whitin, 1963; Klemm, 1971) inventory models. The cycle length may be longer than L if the current expansion is completed before the next one is needed, as in Figures 1 and 4, or less than L, if successive lead times overlap as in Figures 2 and 3. At a generic expansion epoch t_n, the decision

maker knows $P(t_n) = pK_{n-1}$ and wishes to predict, in order to control, the service level over the interval $[t_n + L, T_{n+1} + L)$. For inventory management, Schneider defines the β service level as the fraction of demand not being lost or backordered and identifies its relevance to lost sales or proportional backorder costs. In the capacity expansion problem, at time t_n, the proportion of demand that is satisfied during the next cycle is

$$\beta_n = \frac{\int_{t_n+L}^{T_{n+1}+L} \min\left[P(t), K_n\right] dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)\,dt} = 1 - \frac{\int_{t_n+L}^{T_{n+1}+L} \max\left[P(t) - K_n, 0\right] dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)\,dt}, \tag{1}$$

which is a random quantity. This expression for the service level leads to the constraint that must be satisfied in our problem. Typically, the service provider will decide the desired service level and the numerical value of the expression in Equation (1) should then be greater than or equal to the specified service level.

3.4 Formulation of the expansion cost expression

While Whitt (1981) simply formulated the expression for the capacity utilization in terms of the policy parameters without any considerations of the cost of installing that capacity, we do realize that the optimal values of the policy parameters would be the ones that minimize the cost the capacity expansion. To formulate the objective function of our problem, we consider the infinite horizon cost of expansion for installing these capacity units at different time instances in the future. The capacity expansion problem is very commonly modeled as an infinite horizon expansion cost minimization problem (for example, see Chaouch and Buzacott, 1994; Ryan, 2004 etc.).

We assume an economies of scale regime, under which the cost of installing capacity of size X is given by:

$$C_n(X) = kX^a, \tag{2}$$

where k is a constant and a (< 1) is the economies of scale parameter.

Similar to these models, for the given capacity level of K, let $V_t(K)$ be the minimum expected cost, discounted to time t, of expanding capacity over infinite horizon while satisfying the service level constraint. Then for $n \geq 1$ (see Figures 1-4 for chronology),

$$V_{t_n}(K_{n-1}) = \min_{X_n, T_{n+1}} \left[C_n(X_n) + E_{t_n}[e^{-r(T_{n+1}-t_n)}] V_{T_{n+1}}(K_n) \right]$$

$$\text{subject to } E[\beta_n] \geq \varepsilon$$

where ε is the desired service level, $K_n = K_{n-1} + X_n$, and $C_n(X)$ is the expansion cost for the n^{th} cycle.

Hence the problem is to find policy parameters that give the minimum valued infinite time horizon expansion cost discounted to time 0, given that the service level in each expansion cycle is met:

$$V_0(K_0) = \min\left[E\left[e^{-rT_1}\right] V_{T_1}(K_0)\right] \tag{3}$$

Therefore, the capacity expansion problem essentially reduces to solving the dynamic problem with the given service level constraint to find optimal values for the timing and size parameters for the expansion of the capacity. In the next chapter we simplify the expression for the service level using the concepts of the up-and-out barrier option; and the infinite horizon dynamic iterations to a simplified nonlinear expression for the expansion cost in terms of the policy parameters.

As shown in Appendix, the airline passenger enplanement data for a period of 20 years from 1981 to 2001 indicate that the demand for airline seats has followed a GBM process. Hence, if we consider an airline operator who has some restriction on the minimum service level achieved and who is planning on hiring pilots to increase the human resource capacity, then our capacity expansion model can be applicable in this scenario.

CHAPTER IV: MATHEMATICAL ANALYSIS

Having explained the model environment and discussed the policy parameters in chapter 3, we now analyze the mathematical model in detail. The service level expression arrived at in Equation (1) is expanded such that the expected shortage during an expansion cycle does not exceed the specified limit. We use the results from financial option pricing theory- particularly, the Up-and-Out barrier call option price equation-- to model the service level constraint. The infinite time horizon expansion cost equation (3) is also analyzed in this chapter so that we obtain a telescoping series, which is then simplified to obtain the total cost objective function in terms of the policy parameters.

4.1 Analysis of the service level constraint

At an expansion epoch, in order to meet a specified service level, we can control both the size of the current expansion and a criterion for choosing the time of the next expansion. Note that under the assumed expansion policy, this is just a question of finding p and v. In view of our policy, our first goal is to obtain an expression for the service level, in terms of our decision variables (the timing variable p and the size variable v), that is valid for any expansion epoch. This expression then can be solved to obtain the unknown; viz. obtaining the values for the decision variables to achieve a given service level, or estimating the service level for the given values of decision parameters.

From the previous chapter, Equation (1) gives the service level over the next cycle from the perspective of time t_n. So from Equation (1), the shortage during the n^{th} cycle, as a proportion of the total demand during the cycle, is a random quantity given by:

$$1-\beta_n = \frac{\int_{t_n+L}^{T_{n+1}+L} Max\left[P(t)-K_n,0\right]dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)dt}.$$

The service provider has an upper bound on the shortages. That is, the shortages in any expansion cycle have to be less than or equal to some specified limit. Let this specified shortage limit be δ (= 1-ε).

We define a shortage constraint violation function as a random quantity:

$$G(p,v) \equiv \int_{t_n+L}^{T_{n+1}+L} Max\left[P(t)-K_n,0\right]dt - \delta \int_{t_n+L}^{T_{n+1}+L} P(t)dt.$$

We note that although our decision variables (p, v) do not appear in the right hand side of this equation and the index n does, our goal in this section is to obtain an expression for the service level constraint in terms of the decision variables that is valid for each expansion cycle. Hence, denoting the shortage constraint violation function as $G(p, v)$ will seem logical by the end of this section. Taking expectations, the service level constraint requires that the expected shortage function is less than or equal to 0.

$$g(p,v) \equiv E[G(p,v)] = E\left[\int_{t_n+L}^{T_{n+1}+L} Max(P(t)-K_n,0)dt\right] - \delta E\left[\int_{t_n+L}^{T_{n+1}+L} P(t)dt\right] \le 0, \tag{4}$$

where the expectations are taken with respect to time t_n.

We now simplify each of the terms on the right hand side of Equation (4) to obtain the service level constraint expression in terms of the decision variables, p and v. The first of these terms is equal to:

$$I'_n = E_{t_n}\left[\int_{t_n+L}^{T_{n+1}+L} \left[P(t)-K_n\right]1\left\{P(t)\ge K_n\right\}dt\right] \tag{5}$$

where $1\{x\}$ is an indicator function such that $1\{x\}=1$ if x is true and 0 otherwise. In the above integration the upper limit of the integration $T_{n+1}+L$ is a random term because T_{n+1} is the time (unknown at time t_n) at which the demand will hit the value of pK_n for the first time.

To obtain deterministic integration limits in I'_n, we introduce an indicator function $1\{t \le T_{n+1} + L\}$ and remove the upper limit of integration. This step is justified because for $t \ge t_n + L$,

$$t \le T_{n+1} + L \Leftrightarrow t - L \le \min\{t \ge 0 : P(t) = pK_n\} \Leftrightarrow P(s) \le pK_n, \forall s \le t - L.$$

Therefore, $1\{t \le T_{n+1} + L\} = 1\{\max P(s) \le pK_n : 0 \le s \le t - L\}$, and

$$I'_n = \int_{t_n+L}^{\infty} E_{t_n}\left[[P(t) - K_n]1\{P(t) \ge K_n\}1\{\max P(s) \le pK_n; 0 \le s \le t - L\}\right]dt. \tag{6}$$

Next, given knowledge of events up to time t_n, using the Markov property we can shift the origin to time t_n and find the expected value in terms of a translated Brownian motion.

Writing Equation (6) in terms of the underlying standard Brownian motion,

$$I'_n = \int_{t_n+L}^{\infty} E_{t_n}\left[[P(t) - K_n]1\left\{B(t) \ge \ln\left(\frac{K_n}{P(0)}\right)\right\}1\left\{\max B(s) \le \ln\left(\frac{pK_n}{P(0)}\right); t_n \le s \le t - L\right\}\right]dt.$$

Let $A_1 \equiv \ln(K_n / P(0))$ and $A_2 \equiv \ln(pK_n / P(0))$. Then,

$$I'_n = \int_{t_n+L}^{\infty} E_{t_n}\left[[P(t) - K_n]1\{B(t - t_n + t_n) \ge A_1\}1\{\max B(s) \le A_2; t_n \le s \le t - L\}\right]dt. \tag{7}$$

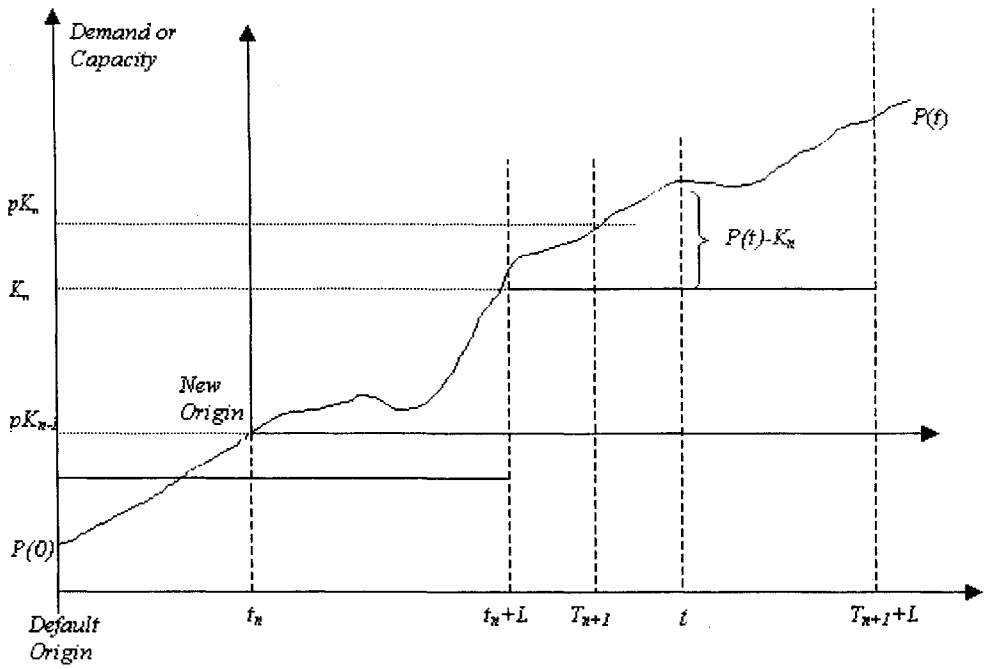

***Figure 5**. Important time instances of the P(t) process*

Define a new Brownian motion $W_n(t) \equiv B(t+t_n) - B(t_n)$, which has the same drift and volatility as $B(t)$ (Karlin and Taylor, 1975). In terms of the process $W_n(t)$, Equation (7) becomes:

$$I'_n = \int_{t_n+L}^{\infty} E_{t_n}\left[\left[P(t_n)e^{W_n(t-t_n)} - K_n\right]1\{W_n(t-t_n) \geq A_1 - B(t_n)\}1\{\max W_n(k) \leq A_2 - B(t_n) : 0 \leq k \leq t-L-t_n\}\right]dt,$$

since $P(t) = P(0)e^{B(t)} = P(t_n)e^{B(t)-B(t_n)}$ has the same distribution as $P(t_n)e^{W_n(t-t_n)}$ given $P(t_n)$.

Define $Q_n(t) \equiv P(t_n)e^{W_n(t)}$ as a GBM with respect to the Brownian motion $W_n(t)$. Also we define a new variable, $u \equiv t - t_n$, and finally after discounting all the shortage to the origin (at the time t_n), Equation (7) can be written as:

$$I'_n = \int_{L}^{\infty} e^{-ru} E\left[[Q_n(u) - K_n]1\{Q_n(u) \geq K_n\}1\{\max Q_n(s) \leq pK_n : 0 \leq s \leq u-L\}\right]du. \quad (8)$$

The integral in this equation can be evaluated by simplifying the joint probability of the Brownian motion and its maximum over different time periods. Chuang (1996) first presented this

joint probability distribution, and then used this distribution to value the knock-out barrier options, particularly the down-and-out call option. Appropriate changes could be made for the up-and-out call option.

4.1.1 Application of partial barrier call option value

A barrier option is a path dependent option where the payoff depends not only on the final price of the underlying asset but also on whether or not the underlying asset has reached some other "barrier" price during the life of the option (Rubinstein and Reiner, 1991). Barrier options are classified as *in options* or *out options* (Rich, 1994) where the *out* feature causes the option to terminate immediately if the underlying asset reaches the specified barrier price. In addition, if the initial price of the asset is below the barrier price, it is called an *up* option. Hence the up-and-out option is worthless if the asset price rises to the barrier. Heynen and Kat (1997) give an explicit analytical equation for the up-and-out call option value. Notations used by them are specified here. It can be shown that the results obtained by using Chuang (1996) are exactly the same as in Heynen and Kat (1997). For the up-and-out option, define

S_0: Initial price of the stock,

t_1: Arbitrary time before the expiration when the monitoring ends

T: Expiration time

K: Strike price

H: Barrier price

μ: Drift parameter

σ: Volatility parameter

γ: Growth rate $\left(\gamma \equiv \mu + \frac{\sigma^2}{2}\right)$

Then assuming the stock price follows a GBM process with drift μ and volatility σ, the price of the up-and-out call option is given by:

$$E\left[(S_T - K)1\{S_T \geq K,\ \max S_t \leq H, 0 < t < t_1\}\right]$$

$$= S_0\psi\left(d_1, -e_1, -\sqrt{\frac{t_1}{T}}\right) - \left(\frac{H}{S_0}\right)^{\frac{2\gamma}{\sigma^2}+1}\psi\left(f_1, -e'_1, -\sqrt{\frac{t_1}{T}}\right) - e^{-\gamma T}K\psi\left(d_2, -e_2, -\sqrt{\frac{t_1}{T}}\right)$$

$$+e^{-\gamma T}K\left(\frac{H}{S_0}\right)^{\frac{2\gamma}{\sigma^2}-1}\psi\left(f_2, -e'_2, -\sqrt{\frac{t_1}{T}}\right). \tag{9}$$

Here, $\psi(x, y, \rho)$ is the cumulative distribution function of the standard bivariate normal distribution with correlation coefficient ρ. And,

$$d_1 = \frac{-\ln\left(\frac{K}{S_0}\right) + (\gamma + \sigma^2/2)T}{\sigma\sqrt{T}}; \qquad d_2 = d_1 - \sigma\sqrt{T}$$

$$e_1 = \frac{-\ln\left(\frac{H}{S_0}\right) + (\gamma + \sigma^2/2)t_1}{\sigma\sqrt{t_1}}; \qquad e_2 = e_1 - \sigma\sqrt{t_1}$$

$$e'_1 = e_1 + \frac{2\ln\left(\frac{H}{S_0}\right)}{\sigma\sqrt{t_1}}; \qquad e'_2 = e'_1 - \sigma\sqrt{t_1}$$

$$f_1 = \frac{-\ln\left(\frac{K}{S_0}\right) + 2\ln\left(\frac{H}{S_0}\right) + (\gamma + \sigma^2/2)T}{\sigma\sqrt{T}}; \qquad f_2 = f_1 - \sigma\sqrt{T}$$

With respect to Equation (8), the terms defined by Heynen and Kat (1997) have following correspondence:

$$S_T \leftrightarrow Q_n(u);\quad K \leftrightarrow K_n;\quad H \leftrightarrow pK_n;\quad t_1 \leftrightarrow u - L;\quad T \leftrightarrow u;\quad S_0 \leftrightarrow P(t_n).$$

The expansion policy specifies $v = \frac{K_n}{K_{n-1}}$ as the ratio of successive capacity levels. We also know that $P(t_n) = pK_{n-1}$, because t_n is the expansion epoch determined by demand reaching the level of pK_{n-1}.

Hence exploiting the correspondence between the Equations (8) and (9), we have that:

$$I^1_n = K_n \int_L^\infty e^{-ru} \left\{ e^{\gamma u} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + (\gamma + \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{\ln(v) - (\gamma + \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right) \right.$$

$$- v^{\frac{2\gamma}{\sigma^2}+1} e^{\gamma u} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln(v) + (\gamma + \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{-\ln(v) - (\gamma + \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right)$$

$$- \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + (\gamma - \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{\ln(v) - (\gamma - \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right)$$

$$\left. + v^{\frac{2\gamma}{\sigma^2}-1} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln(v) + (\gamma - \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{-\ln(v) - (\gamma - \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right) \right\} du. \qquad (10)$$

Now, going back to Equation (4), we evaluate the second term of the right hand side of Equation (4) in a way similar to above. We have that,

$$I^2_n = E\left[\int_{t_n+L}^{T_{n+1}+L} P(t)dt \right]$$

After similar steps as for I^1_n, we have that,

$$I^2_n = E_{t_n}\left[\int_L^\infty e^{-ru} Q_n(u) 1\{max\ Q_n(s) \le pK_n, 0 \le s \le u-L\} du \right]$$

$$= \int_L^\infty e^{-ru} E\left[Q_n(u) 1\{max\ Q_n(s) \le pK_n, 0 \le s \le u-L\} \right] du$$

Once again we use Heynen and Kat (1997), and find that the above expression is equal to:

$$I^2_n = K_n \int_L^\infty e^{(\gamma-r)u}\left(\frac{p}{v}\right)\varphi\left(\frac{\ln(v)-(\gamma+\frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}\right)du$$

$$-K_n \int_L^\infty e^{(\gamma-r)u} v^{\frac{2\gamma}{\sigma^2}+1}\left(\frac{p}{v}\right)\varphi\left(\frac{-\ln(v)-(\gamma+\frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}\right)du. \quad (11)$$

where $\varphi(\cdot)$ is the standard normal distribution function.

The final service level constraint in Equation (4) is now expressed as:

$$g(p,v) = (\text{exp}ected\ unmet\ demand\ during\ expansion\ cycle) - \delta(\text{exp}ected\ total\ demand\ during\ expansion\ cycle) \le 0 \quad (12)$$

where the first term is I^1_n which is evaluated using Equation (10) and the second term (total demand during the expansion cycle) is I^2_n which is evaluated using Equation (11).

As seen from Equations (10) and (11), the value for the service level does not depend on the value of n. Hence the expression for the service level constraint is the same for all the expansion cycles. The expected shortage in our model depends on both the decision variables. In fact, this is consistent with the expression for the capacity utilization without lead times Whitt (1981), which involves both the timing and the size parameter of the expansion policy.

4.2 Analysis of the infinite time horizon expansion cost

Now that the expression for the service level constraint is obtained in terms of the policy parameters (i.e., the timing and size parameters), in this section, we will derive the expression for the infinite time horizon expansion cost in terms of the same decision variables. Starting from the problem defined in Equation (3), we will analyze the expression for the expansion cost. In finding this expansion cost, the implicit assumption made is that the total cost of expansion is incurred at the beginning of the expansion project. Hence if the n^{th} capacity expansion starts at time t_n, then the total

cost of this n^{th} expansion occurs at time t_n, and since our assumption is that the expansion lead time is fixed, this is not a critical assumption.

Now at time T_1, the total costs (TC) incurred are the actual cost of expansion (from the initial capacity position of K_0 to the new capacity position of K_1), because of the start of the expansion project; and the total cost of all the future expansion discounted to time T_1.

$$\begin{aligned} TC &\equiv E[e^{-rT_1}]TC_{T_1} \\ &= E[e^{-rT_1}]\left\{kX_1^a + E_{t_1}[e^{-r(T_2-t_1)}]TC_{T_2}\right\}, \end{aligned}$$

where TC_{T_1} is the total cost from the first expansion onwards with the future costs discounted to time T_1; X_1 is the size of the capacity expansion; and the cost of second capacity expansion is first discounted to time T_1 and then the total cost at time T_1 including the cost of expansion is discounted to time 0 (zero). Also, $X_1 = K_1 - K_0 = K_0(\nu - 1)$.

Now the infinite cost from time T_2 can also be expanded in terms of cost of expansion at time T_2 and the further infinite horizon cost, so that $TC_{T_2} = kX_2^a + E_{t_2}[e^{-r(T_3-t_2)}]TC_{T_3}$. Then the expression for the expansion cost can be written as an infinite telescoping series of costs:

$$TC = E[e^{-rT_1}]\left\{kX_1^a + E_{t_1}[e^{-r(T_2-t_1)}]\left\{kX_2^a + E_{t_2}[e^{-r(T_3-t_2)}]\left\{kX_3^a + E_{t_3}[e^{-r(T_4-t_3)}]\left\{kX_4^a + \ldots\right\}\right\}\right\}\right\}. \qquad (13)$$

4.2.1 The discount factor

The stochastic component in Equation (13) is the Laplace transform for the hitting time of the GBM. However, for each expansion project, the time basis for the expected value of the hitting time for the GBM is shifted to the (known) time when the current expansion project started. For each of these expected values of the exponential of the hitting time, the time 0 (zero) value of the GBM process is pK_{n-1} and the hitting time is the time for the process to hit the value of pK_n for the first time. This

expected value of the exponential of the hitting time can be simplified using a similar technique used during the analysis of the service level constraint. There, whenever we shifted the time basis (the origin) of the GBM process to a new time, we defined a new Brownian motion, which had the same drift and volatility as the original Brownian motion.

For example, consider the start of the first capacity expansion project (refer to Figures 1- 4). We are at time t_1, where the value of the GBM has just hit pK_0 and has triggered the first capacity expansion. From t_1, the time until the next capacity expansion starts is equal to T_2-t_1, and hence the discount factor for the second capacity expansion cost has this time factor. Here T_2 is the start of the next capacity expansion- it is the time when the GBM process hits the value pK_1. In effect, the time difference between the successive starting points of expansion projects is the time it takes for the GBM process to go from pK_0 to pK_1.

However, for discounting the cost of second expansion, our time origin is t_1. So we define a new Brownian motion process that starts at time t_1 with initial value such that the corresponding GBM has the value of pK_0. For the new GBM (corresponding to the new Brownian motion), the time to hit pK_1 is a hitting time which is same as T_2-t_1. And for this new Brownian motion, the expression for the Laplace transform of the hitting time is given by (Karlin and Taylor, 1975; and Borodin and Salminen, 2002):

$$E\left[e^{-rT(pK_1)} \mid P(t_1) = P'(0) = pK_0\right] = \left(\frac{pK_0}{pK_1}\right)^{\lambda} = \left(\frac{1}{v}\right)^{\lambda},$$

$$\text{where } \lambda = \sqrt{\frac{\mu^2}{\sigma^4} + \frac{2r}{\sigma^2}} - \frac{\mu}{\sigma^2}.$$

Similarly, consider the cost of the third expansion being discounted to time t_2. The discount factor for that could be calculated using the method described above. At time t_2, the GBM process has a value of pK_1. We want to find the discount factor for the cost of third expansion that will be incurred at time T_3, the start of the third expansion. Once again we define a new Brownian motion

starting at time t_2, with an initial value such that the corresponding GBM process has a value of pK_1, and find the Laplace transform of the hitting time: the time required for the new GBM (which is derived from the newly defined Brownian motion) to reach the value of pK_2. Once again, we have:

$$E\left[e^{-rT(pK_2)} \mid P(t_2) = P''(0) = pK_1\right] = \left(\frac{pK_1}{pK_2}\right)^{\lambda} = \left(\frac{1}{v}\right)^{\lambda}.$$

In fact, using the same technique we can see that all these expected values of the exponential of the difference between the hitting time and the known time, which are the discount factors for the successive capacity expansions, are in fact the same. That is to say that,

$$E_{t_1}[e^{-r(T_2-t_1)}] = E_{t_2}[e^{-r(T_3-t_2)}] = E_{t_3}[e^{-r(T_4-t_3)}] \ldots = \left(\frac{1}{v}\right)^{\lambda}. \tag{14}$$

4.2.2 The expansion cost

Now that we have the expression for the discount factor in terms of the policy parameters (Equation (14)), we attempt the same for the actual cost of expansion. As described earlier, because of economies of scale costing, the cost of expansion is given by Equation (2). We also know that $X_1 = K_1 - K_0 = K_0(v-1)$. Similarly the sizes of successive expansions are given by:

$$\begin{aligned}
X_2 &= K_2 - K_1 = vK_1 - K_1 = v^2K_0 - vK_0 = K_0v(v-1).\\
X_3 &= K_3 - K_2 = K_0v^2(v-1).\\
X_4 &= K_4 - K_3 = K_0v^3(v-1).\\
&\ldots\\
X_n &= K_n - K_{n-1} = K_0v^{n-1}(v-1); n \geq 1
\end{aligned}$$

Hence, the cost of the n^{th} expansion will be

$$C(X_n) = k\left(K_0v^{n-1}(v-1)\right)^a = kK_0{}^a(v-1)^a v^{a(n-1)}.$$

And the total cost from the first expansion onwards, given by Equation (13), discounted to time T_1 can then be written as:

$$TC_{T_1} = kX_1^a + E_{t_1}[e^{-r(T_2-t_1)}]\left\{kX_2^a + E_{t_2}[e^{-r(T_3-t_2)}]\left\{kX_3^a + E_{t_3}[e^{-r(T_4-t_3)}]\left\{kX_4^a + E_{t_4}[e^{-r(T_5-t_4)}]...\right\}\right\}\right\}$$

$$= k\left(K_0 v^{1-1}(v-1)\right)^a +$$
$$E_{t_1}[e^{-r(T_2-t_1)}]\{k\left(K_0 v^{2-1}(v-1)\right)^a +$$
$$E_{t_2}[e^{-r(T_3-t_2)}]\{k\left(K_0 v^{3-1}(v-1)\right)^a +$$
$$E_{t_3}[e^{-r(T_4-t_3)}]\{k\left(K_0 v^{4-1}(v-1)\right)^a +$$
$$.....+$$
$$E_{t_n}[e^{-r(T_{n+1}-t_n)}]...+.....\}\}\}.....\}.$$

And applying Equation (14), the above equation becomes:

$$TC_{T_1} = k\left(K_0 v^{1-1}(v-1)\right)^a + \left(\frac{1}{v}\right)^\lambda \{k\left(K_0 v^{2-1}(v-1)\right)^a + \left(\frac{1}{v}\right)^\lambda \{k\left(K_0 v^{3-1}(v-1)\right)^a +$$
$$\left(\frac{1}{v}\right)^\lambda \{k\left(K_0 v^{4-1}(v-1)\right)^a + ...+\left(\frac{1}{v}\right)^\lambda \{k\left(K_0 v^{n-1}(v-1)\right)^a ...+.....\}\}\}.....\}$$

$$= \sum_{n=0}^{\infty} k(K_0(v-1))^a \left(v^{a-\lambda}\right)^n.$$

We can see that the above expression converges only if $a < \lambda$, which is always satisfied (for any $r>0$, $\lambda>1$; *also* $a<1$, so the condition $a < \lambda$ is always true*)*. Here we note that for $\lambda>1$, we must have $r>\gamma$(Karlin and Taylor, 1975). So now we have a geometric series. Hence the above expected total cost equation becomes:

$$TC_{T_1} = \frac{k(K_0)^a (v-1)^a}{1-v^{a-\lambda}}.$$

In fact, this equation of the expected total cost is same as the infinite horizon cost obtained by Bean et al. (1992), and also by Ryan (2004), even though the former model did not consider the lead time and the latter model restricted $p \le 1$. Our equation was derived independently of these two

models. And since we consider the case of $p > 1$ also, we have in effect generalized the cost equation to the unrestricted p case.

Lastly, the simplified expression for the term $E[e^{-rT_1}]$ could be found using the fact that this is just the Laplace transform of the hitting time for a GBM process starting at time *0* (zero) with value K_0, and is the time needed for the process to hit the value pK_0 for the first time (this is the time for the first capacity expansion to start, T_1). Hence this expression is given by:

$$E[e^{-rT_1}] = E[e^{-rT(pK_0)} \mid P(0) = K_0] = \left(\frac{K_0}{pK_0}\right)^{\lambda} = p^{-\lambda}$$

Going back to the Equation (13), the final expression, in terms of our policy parameters, is:

$$TC = \frac{k(K_0)^a (v-1)^a p^{-\lambda}}{1 - v^{a-\lambda}} \equiv f(p,v). \tag{15}$$

As discussed in the Chapter 1, this expression of the infinite time horizon total cost of expansion is our objective function for the non-linear program in terms of the policy parameters, p and v. The optimal values of the policy parameters must minimize this total cost of expansion and also must satisfy the constraint of maintaining the service level. The expression for the service level in terms of the policy parameters was found in Equation (12). Therefore, the optimization problem for our capacity expansion policy becomes:

$$\begin{aligned} &\textit{Min} && f(p,v) \\ &\text{subject to} && \\ & && g(p,v) \le 0 \\ & && v \ge 1, p \ge 0. \end{aligned} \tag{16}$$

Here, the expression for the infinite time horizon expansion cost objective function is obtained from Equation (15) and the service level constraint expression is obtained from Equation (12). As explained earlier, the constraint $p \ge 0$ is more general in the sense that when $p>1$, then it means that the service provider has to start the next expansion project with some initial shortages;

however when $p \leq 1$, then the next expansion project is started no later than when the demand hits the current capacity position. Also, since we are considering only the capacity 'expansions', we allow values of size parameter v such that $v \geq 1$.

CHAPTER V: SOLUTION METHODOLOGY AND NUMERICAL RESULTS

In the last section of the previous chapter, we formulated our optimization problem. Equation (16) mathematically states the capacity expansion problem we are solving. The objective function is the infinite time horizon cost of expansion (indicated by the function $f(p, v)$). The main constraint of the problem is the service level limit in each expansion cycle. We want the shortages in each expansion cycle to be less than or equal to some specified value (which was finally converted to the expression $g(p, v)$). Our decision variables are the timing variable (p), which indicates when to initiate the next expansion project (as described in Chapter 3, the expansion project is initiated when the demand hits some proportion 'p' of the current capacity position) and the size variable (v), which indicates by how much to increase the capacity (recall that the new capacity will be some proportion 'v' (> 1) of the current capacity position). Moreover, as seen from Equations (15) and (12), both the objective function and the constraint equation are functions of the decision variables. The complexity of the problem in Equation (16) is evident from the fact that it is a non-linear optimization problem with a rather difficult constraint expression. In this chapter, we discuss the solution methodology used to solve the problem. We look into the steps involved in optimally finding the values of our decision variables. Also, we numerically solve this optimization problem under various conditions of the other problem parameters and discuss the results.

5.1 Optimization technique- Cutting plane algorithm

We used the well-known cutting plane algorithm to solve the optimization problem in Equation (16). As seen from Equation (12), the constraint equation involves integrals of bivariate normal distribution functions, and hence finding the partial derivatives of the constraint equation is difficult. Since we are using the financial option pricing theory to model the constraint equation, we might have used what are called 'Greeks' (Rubinstein and Reiner, 1991) to find the gradients of the constraint equations.

'Greeks' are the partial derivatives of the financial option price with respect to option parameters such as the stock price, volatility etc. These 'Greek' letters are defined and their expression found to a great detail for the regular 'vanilla' options. However, for exotic financial option like the partial barrier option we are using in our model, we could not find any published work about the Greek letters. And since the gradient of the constraint equation could not be found readily, the usual gradient-based optimization methods could not be used for our problem. We then tried solving the Lagrangean dual of the original optimization problem of Equation (16). However, once again because of the complexity of the constraint equation, the dual problem could not be solved. Hence to approximate the Lagrangean dual problem, we used the cutting plane algorithm (which Zangwill (1969) calls the 'dual cutting plane algorithm'), which bypasses finding feasible directions at each step of the problem (Bazaraa et al., 1993). This algorithm just cuts off infeasible solutions in each cut and converges to the optimal solution (Kelley, 1960). Via the proof by Zangwill (1969) of the convergence of this algorithm, which assumes convexity, the optimality of the solution found by this algorithm is guaranteed.

5.1.1 Convexity

One of the primary requirements for the convergence of the cutting plane algorithm for a minimization problem is the convexity of the objective function and the constraint expression. We were able to find evidence only of pseudo-convexity of the objective function. The details are included here:

Definition: Let S be a non-empty open set in E_n, and let f: S E_1, be differentiable on S. The function f is said to be pseudoconvex if for each x_1, x_2 in S, with $\nabla f(x_1)'(x_2 - x_1) \geq 0$, we have $f(x_2) \geq f(x_1)$; or equivalently, if $f(x_2) < f(x_1)$, then $\nabla f(x_1)'(x_2 - x_1) < 0$ (Bazaraa et al., 1993).

For the objective function given in Equation (15) we have:

$$\frac{\delta f}{\delta v} = p^{-\lambda}\left(\frac{(a-\lambda)(v-1)^{a-\lambda-1}}{(1-v^{a-\lambda})^2} - \frac{a(v-1)^{a-1}}{1-v^{a-\lambda}}\right)$$

$$\frac{\delta f}{\delta p} = \frac{-\lambda p^{-\lambda-1}(v-1)^a}{1-v^{a-\lambda}}.$$

Let point $x_1 = (p_1, v_1)$ and $x_2 = (p_2, v_2)$ and suppose that $f(x_2) < f(x_1)$.

$$\nabla f(p,v)' = \begin{pmatrix} \frac{\delta f}{\delta p} \\ \frac{\delta f}{\delta v} \end{pmatrix} \Rightarrow \nabla f(x_1)' = \begin{pmatrix} \frac{\delta f}{\delta p} \\ \frac{\delta f}{\delta v} \end{pmatrix}_{x_1}$$

$$\text{Then, } \nabla f(x_1)'(x_2 - x_1) = \left[\frac{\delta f}{\delta p}\right]_{x_1}(p_2 - p_1) + \left[\frac{\delta f}{\delta v}\right]_{x_1}(v_2 - v_1)$$

$$= p_1^{-\lambda}\frac{(v_1-1)^a}{(1-v_1^{a-\lambda})}\left[\frac{-\lambda(p_2-p_1)}{p_1} + \left(\frac{a}{(v_1-1)} - \frac{(\lambda-a)}{(1-v_1^{a-\lambda})(v_1-1)^{\lambda+1}}\right)(v_2-v_1)\right]$$

This expression is negative if

$$\frac{1}{(v_1-1)}\left(a - \frac{(\lambda-a)}{(1-v_1^{a-\lambda})(v_1-1)^{\lambda}}\right)(v_2-v_1) \le \frac{\lambda}{p_1}(p_2-p_1) \qquad (17)$$

Consider the case where $\frac{\delta f}{\delta v} > 0$, that is : $a - \frac{(\lambda-a)}{(1-v_1^{a-\lambda})(v_1-1)} > 0$. If $v_2 > v_1$ (this means that the left hand side of the inequality is positive), then $f(x_2) < f(x_1)$ implies that $p_2 >> p_1$ and the inequality (17) holds. Even with $v_1 > v_2$ (the left hand side of the inequality is then negative), to satisfy the condition that $f(x_2) < f(x_1)$, although we need $p_1 > p_2$, numerically, for the values of parameters tested, the inequality (17) holds. In this numerical analysis, we sampled some numerical values of the parameters and tested the inequality (17) for each set of values. We found that the inequality (17) holds for each of the sampled numerical values. The range of parameter values from which this sample was taken: the economies of scale parameter (a) varied between 0.7 to 1, λ from 1.01 to 3, p from 0.001 to 5 and v from 1 to 10.

Now consider $\frac{\delta f}{\delta v} < 0$, so for $p_1 = p_2$, $v_2 - v_1 > 0$, and the inequality holds. For the case where the p values are not the same for the two points, we once again sampled numerical values for the parameters from the range mentioned above and found, numerically, that the inequality holds for all the parameter values tested. Hence we have some evidence of the pseudoconvexity of the objective function. We note that the partial derivatives of the objective function with respect to the decision variables are complex and hence no conclusions could be drawn analytically.

Once again, owing to the complexity of the constraint equation, analytical proof of convexity is difficult. Atlason et al. (2004) discussed a numerical method for checking whether a function is concave. Via Theorem 9 of their work, they proposed solving a relatively simple linear program (LP) to check for concavity of any function. So this method can be used to check convexity of a function by just a change of sign.

The idea behind this method is that if a one-dimensional function is concave then it is possible to set a ruler above each point and rotate until the function completely lies below the ruler. This can also be done when dealing with functions of higher dimensions- then the ruler takes the form of a plane (for two dimensions) or hyperplane (for higher dimensions). This idea is illustrated in Figure 6 below.

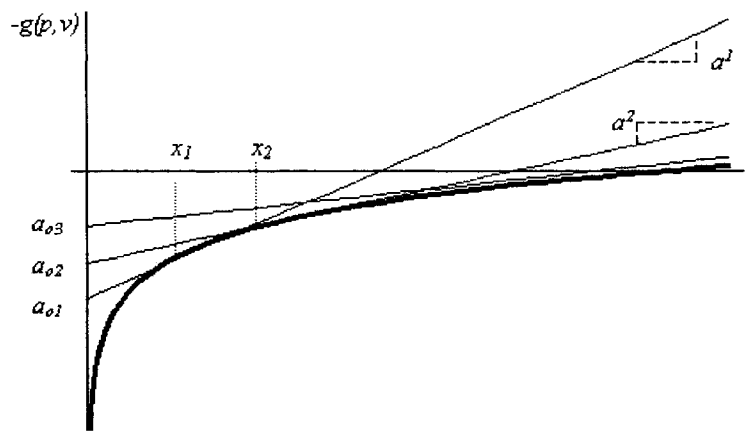

Figure 6. Hyperplanes below which a concave function lies (here x = (p, v)).

The LP proposed by Atlason et al. (2004) changes given function values so that a supporting hyperplane for the convex hull of the points can be fitted through each point. The objective of this LP is to minimize the change in the function values that needs to be made to accomplish this goal. The LP to test the convexity of the service level constraint expression of our problem is formulated as (Atlason et al., 2004):

$$\min \sum_{i=1}^{k} |b_i|$$

subject to

$$a_{0i} + (a^i)[p^i \; v^i]^T = -g(p^i, v^i) + b_i \quad \forall\, i \in \{1,...k\}$$

$$a_{0i} + (a^i)[p^j \; v^j]^T = -g(p^j, v^j) + b_j \quad \forall\, i \in \{1,...k\}, \forall\, j \in \{1,...k\}, j \neq i$$

Here, k is the number of sampled points. To linearize the objective function, the standard trick of writing $b_i = b_i^+ - b_i^-$ can be adopted and then we have $|b_i| = b_i^+ + b_i^-$, where b_i^+ and b_i^- are non negative. The decision variables are:

$a_{0i} \in R, i \in \{1,...k\}$: intercepts of the hyperplane,
$a^i \in R^2, i \in \{1,...k\}$: slopes of the hyperplane and
$b_i^+, b_i^+ \in R, i \in \{1,...k\}$: change in the function values.

Atlason et al. (2004) also proved that when the optimal objective value of the LP is 0, then there exists a concave function such that it has the same value as the function in question at all the points sampled.

Atlason et al. (2004) solved a call center staffing problem using cutting plane algorithm. The authors proposed solving the linear program after each of the iterations of the cutting plane algorithm to check for concavity. We applied this method to our objective function and constraint equation after changing sign of the value of the equations. As proposed by Atlason et al., after each iteration of the cutting plane algorithm the solution of that iteration was included in the set of points at which the linear program was tested. The solution of that linear program was zero, so according to Theorem 9

and the succeeding corollary of Atlason et al. (2004), there exists a concave function that has the same value as the function in question at all the sampled points. At every instance, the constraint function and the objective function for our problem passed this test of convexity. A sample of the linear program used to prove this concavity is included in Appendix IV. The linear program verified the concavity of the function based on five sampled points for one instance of the problem.

5.1.2 Steps involved in the cutting plane algorithm

Bazaraa et al. (1993) discussed the dual cutting plane algorithm for non-linear convex programming problems and proposed the following steps involved in the same. Following Bazaraa et al. (1993), the steps of the dual cutting plane algorithm as it applies to our problem are as follows:

Initialization step: Select an initial feasible point (p_0, v_0).

For each iteration k, solve the Master Problem for z and u, which is given as:

$$\begin{aligned} &\textit{Maximize } z \\ &s.t.\ z \le f(p_j, v_j) + ug(p_j, v_j) \text{ for } j = 0...k-1 \\ &u \ge 0 \end{aligned}$$

Let (z_k, u_k) be the optimal solution.

Now using the optimal value of the penalty variable u_k, solve the Sub Problem:

$$\begin{aligned} &\textit{Minimize } f(p,v) + u_k g(p,v) \\ &s.t. \qquad p \ge 0, v \ge 1. \end{aligned}$$

Let (p_k, v_k) be the optimal solution for the sub problem.

Let $\theta(u_k) = f(p_k, v_k) + u_k g(p_k, v_k)$.

If $z_k = \theta(u_k)$ then stop. Otherwise continue with the Master Problem with added constraint:

$z \leq f(p_k, v_k) + ug(p_k, v_k)$.

Figure 7. Cutting plane algorithm steps.

As with our problem, the feasible set of a nonlinear program may sometimes be difficult to handle. The cutting plane algorithm, instead of attacking the feasible set directly, start with a simpler set that approximates the feasible set. From this feasible set, a point is selected (which, in our case, is the optimal solution to the master problem). If this point lies in the original feasible set, the algorithm stops and we have found the optimal solution to the original problem. To test whether the point lies in the feasible region, a solution test is performed, which corresponds to checking the equality of Sub Problem solution and the Master Problem solution of that iteration ($z_k = \theta(u_k)$). If the solution test fails, indicating that the current solution does not lie in the feasible region, this point is 'cut off' from the set that approximates the feasible region (hence the name, cutting plane algorithm). This is achieved by adding the constraint $z \leq f(p_k, v_k) + ug(p_k, v_k)$ to the Master Problem. This gives a new approximation of the feasible region, which does not contain the previous infeasible solution. The algorithm continues until a point is found which passes the solution test.

From Figure 7, we notice that the Master Problem is a linear program the solution for which gives an upper bound for the solution to the Sub Problems. Moreover, the Sub Problem constraints are linear with a non-linear objective function. Hence the total computation time to solve these Sub Problems is less than that for the original problem. Finally, Zangwill (1969) provides a proof of the convergence of the cutting plane algorithm, which means that the optimal solution to the original problem in Equation (16) will eventually be found, provided that the problem is feasible.

Using the cutting plane algorithm described in Figure 7, we solved the non-linear program of Equation (16). In the next section, we present some of the numerical results based on various problem parameter values.

5.2 Numerical results

We start with a numerical analysis of the service level constraint formulated in Equation (12). Later on we apply the cutting plane algorithm described in the previous section to the problem defined in Equation (16). We show how the cutting plane algorithm converges for our problem instance.

The software package *Mathematica 5.1* (Wolfram Research Inc, 2004) was used to obtain numerical results for the model. Evaluating Equation (12) involves integrating the bivariate normal distribution functions over infinite regions. Initially troubles were encountered when the built-in functions for evaluating the bivariate normal distribution were used. Also, the method of directly evaluating the integrals involved in the bivariate normal distribution failed for our model. Finally the method prescribed by Rose and Smith (1996) was used. In this method, the authors define the multivariate normal distribution function in terms of matrices for mean and variance. This method is more general than using the built-in bivariate normal distribution function because using this we can in fact model the multivariate case. The authors define a function '*MVN*' which constructs the probability density function of a multivariate standard normal distribution using the mean vector and variance-covariance matrix. The authors proceed by setting up a function *MVN[x, μ, var]* to calculate the *n*-dimensional multivariate normal distribution for the vector $\mathbf{x} = [x_1, \ldots x_n]$ defined everywhere in the *n*-dimensional real space, the mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]$, and a symmetric positive-definite variance-covariance matrix **var**:

$$MVN[\mathbf{x}, \boldsymbol{\mu}, \mathbf{var}] = 2\pi^{(-n/2)}\sqrt{Det[\mathbf{var}^{-1}]} * e^{[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}).\mathbf{var}^{-1}.(\mathbf{x}-\boldsymbol{\mu})]},$$

Then for a standard normal bivariate normal density function ($\mathbf{x} = [x_1, x_2]$, $n = 2$), with zero mean vector, variance elements unity, and correlation coefficient η, the MVN function returns:

$$MVN[\mathbf{x}, \mathbf{mu}, \mathbf{var}] = \frac{e^{(x_1^2 - 2\eta x_1 x_2 + x_2^2)/(2(\eta^2 - 1))} \sqrt{\frac{1}{1-\eta^2}}}{2\pi}$$

This density function can then be integrated to find the probability distribution function. This method was found to be more efficient than the built-in procedure in *Mathematica 5.1* (Wolfram Research Inc, 2004), specifically when integration of the multivariate normal distribution function is involved. While solving the iterations of the cutting plane algorithm, the Master Problem was solved using the software package *LINDO*. We note that since the Master Problem is a linear program with finite number of constraints, it can be solved using any software package available.

5.2.1 Results regarding the constraint equation

We first examine the effect of the timing and size parameters on the service level constraint. We know that, as the value of timing parameter p increases, we are in effect delaying the start of expansion project further and further. Hence intuitively, the shortage violation function $g(p, v)$ should become less and less negative and in fact become positive corresponding to constraint violation for large values of the timing parameter. Similarly, as the value of size parameter v increases, since we are having larger expansions each time, we expect the $g(p, v)$ to become more and more negative (hence, more and more favorable). From Figure 8 (*a* and *b*) below, we see the quantitative behavior of our service level constraint equation.

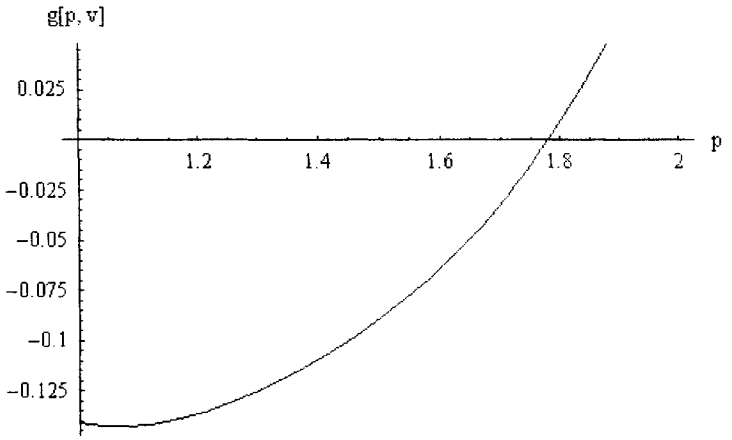

Figure 8a: Relationship between timing parameter and the amount of constraint violation. Parameter values: size factor (v) = 6, drift (μ) = 8%, volatility (σ) = 20%, lead time (L) = 2 years, interest rate (r) = 11%.

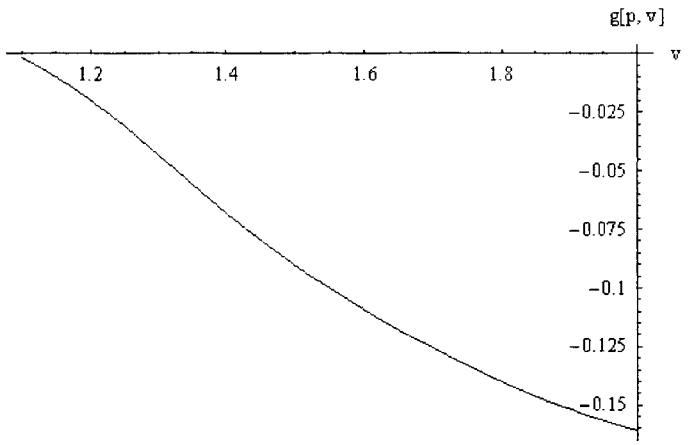

Figure 8b: Relationship between size parameter and the amount of constraint violation. Parameter values: time factor (p) = 1.001, drift (μ) = 8%, volatility (σ) = 20%, lead time (L) = 2 years, interest rate (r) = 11%

As seen from these two figures, numerically the constraint equation behaves the way it is expected to. Hence, we have that the constraint equation becomes more negative with decreasing values of timing parameter and/or with increasing values of size parameter. However, from Equation (15), we see that the infinite time horizon expansion cost decreases with higher values of timing parameter and lower values of size parameter. This is shown in Figures 9 (a and b). The pull in

opposite direction for the values of timing and size parameter is what sets up an interesting optimization problem.

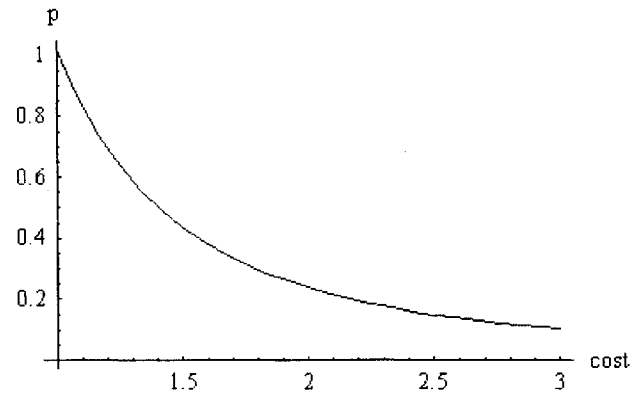

Figure 9a: Relationship between timing parameter and the infinite time horizon cost of expansion. Parameter values: size factor (v) = 1.5, drift (μ) = 8%, volatility (σ) = 20%, lead time (L) = 2 years, interest rate (r) = 11%.

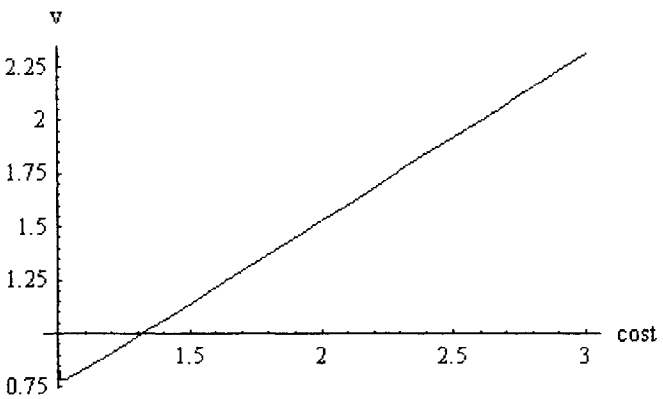

Figure 9b: Relationship between size parameter and the infinite time horizon cost of expansion. Parameter values: time factor (p) = 1.5, drift (μ) = 8%, volatility (σ) = 20%, lead time (L) = 2 years, interest rate (r) = 11%.

Because of the complexity of the constraint equation, we could not obtain the contour plots for the relationship between the two decision variables for a given value of the constraint violation. In the next section, we discuss the application of the cutting plane algorithm to the problem defined in Equation (16).

5.2.2 Optimization results

We applied the dual cutting plane algorithm described in Section 5.1 to our capacity expansion problem (Equation (16)). While solving the Sub Problems of Figure 7, we added a dummy constraint of $p \leq 2$ for each of the iterations. This was done to reduce the number of iterations required for convergence of the cutting plane algorithm. As seen from the Table 1 below, the first iteration of algorithm minimizes the total cost (Equation (15)) subject to the constraints on the decision variables ($p \leq 2, p \geq 0, v \geq 1.$). Here, with the dummy constraint, the optimal timing variable value of this iteration is limited to 2 and hence faster convergence is achieved. The other parameter values used for this instance of the problem were: drift (μ) = 0.08, volatility (σ) = 0.2, lead time (L) = 2 years, interest rate (r) = 0.13 and economies of scale parameter (a) = 0.99. These values were selected hypothetically. The first two parameters described the demand process. It means that the demand for the capacity has a drift rate of 8%, and the volatility of that process is 20%. To maintain the condition that the interest rate be larger than the growth rate (see Chapter 4), we have chosen an interest rate of 13%. Lastly, the economies of scale parameter value of 0.99 means that there is little cost incentive in having larger size capacity additions (refer Equation (2)). The initial feasible point was (p_0, v_0) = (1.4, 3.4). The successive iterations and the convergence of the cutting plane algorithm are summarized in Table 1. The level of accuracy used for all the numerical studies was up to 3 decimal places. Since the Master Problem is a simple linear program, the optimal solution to that is obtained nearly instantaneously. However, as the Sub Problem at each iteration involves the expression $g(p, v)$, which includes integration of bivariate normal distribution functions, the average computational time required to solve each Sub Problem is approximately 2 hours on a Intel© Pentium IV personal computer with Windows XP operating system and 1gigabyte of memory. And as seen from Table 1, we solved 8 Sub Problems for this instance of the problem.

Table 1. Results of the cutting plane algorithm applied to the capacity expansion problem.

Iteration	Constraint Added	Master problem solution (z, u)	Sub-problem solution (p, v), θ
1	$z \le 5.94 - 0.064u$	(5.94, 0)	(2, 1.009), 1.782
2	$z \le 1.782 + 1.127u$	(5.74, 3.1)	(1.23, 1.008), 3.85
3	$z \le 3.24 + 0.196u$	(5.27, 10.348)	(0.98, 1.01), 4.21
4	$z \le 4.27 - 0.0054u$	(4.29, 5.12)	(1.1, 1.009), 4.1
5	$z \le 3.74 + 0.07u$	(4.23, 7.017)	(1.04, 1.009), 4.19
6	$z \le 3.99 + 0.027u$	(4.228, 8.35)	(1.018, 1.007), 4.21
7	$z \le 4.12 + 0.011u$	(4.223, 9.017)	(1.06, 1.01), 4.216
8	$z \le 4.196 + 0.002u$	**(4.218**, 10)	(0.989, 1.016), **4.218**

As seen from Table 1, at the end of the 8th iteration, the optimal objective function value of the Master Problem is equal to that of the Sub Problem; hence, we stop and say that the cutting plane algorithm has converged. The optimal solution in this instance is the solution to the Sub Problem in the last iteration that is, $p^* = 0.989$, $v^* = 1.01$. This means that the service provider should start the new capacity expansion project when the demand hits 98.9% of the current capacity position and the expansion should be such that after the project is completed, the capacity available is 101% of the current capacity position. The optimal value of the objective function $f(p^*, v^*)$ is found to be 4.26 and the value of shortage violation expression $g(p^*, v^*)$ is –0.0004.

In a similar fashion the cutting plane algorithm was implemented for different values of the problem parameters and the results of these tests are summarized below. Here, we discuss the effects

of problem parameters on the optimal timing variable (p) and not on the optimal size variable (v). This is so because there was no clear trend for the optimal size parameter as we varied other problem parameters one at a time. We observe that at the convergence of the cutting plane algorithm, the shortage violation equation is not satisfied to the same extent in each problem instance. In some cases the numerical value of the $g(p, v)$ is –0.02, and in other case it is –4.5e-5, the reason for which is unclear to us at this point.

Effect of the drift parameter on optimal timing factor

To test how the decision regarding the timing of the new expansions is affected due to change in the drift parameter of the demand process, the default parameters values were: volatility (σ) = 20%, lead time (L) = 2 years, interest rate (r) = 13%, and economies of scale parameter (a) = 0.99. The service level was assumed to be 95%, meaning that the shortages were limited to 5% of the total demand during the expansion cycle (δ = 0.05). Then values of drift parameter were tested in the cutting plane algorithm and optimal values of timing and size parameters were obtained.

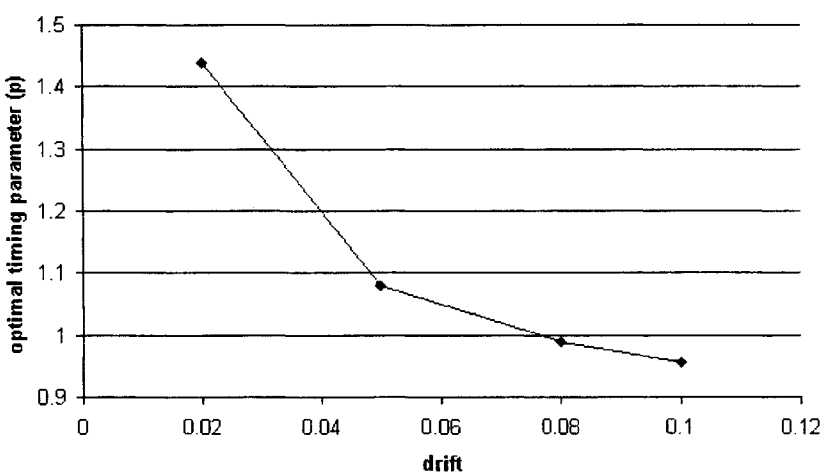

Figure 10: Effect of demand drift on optimal timing parameter

Table 2: Effect of demand drift on optimal decision variables

μ	Optimal p^*	Optimal v^*	$f(p^*, v^*)$	$g(p^*, v^*)$
0.02	1.44	2.05	0.892	-0.002
0.05	1.08	1.363	1.9056	-0.037
0.08	0.989	1.01	4.218	-0.0004
0.1	0.956	1.632	17.23	-0.087

From Figure 10, we can see that as the drift for the demand process increases, it prompts earlier initiation of next expansion project. From Table 2, we also observe that the value of the objective function, which is the total cost of expansion, also increases as the drift parameter increases. This is because of the fact that the infinite time horizon cost of expansion increases with decreasing timing variable (p). Higher drift values for the demand process implies a high growth industry. Hence, we can see that for high growth industries the optimal timing parameter values get smaller and smaller. Here, also we see that it is more and more expensive to meet the service level constraint for a demand that is growing fast.

Effect of demand volatility on the optimal timing parameter

In the real world, demand fluctuations are common and hence it is critical to study the effects of this demand volatility on the decision variables of our capacity expansion problem. Here, the numerical values of the parameters used were: drift (μ) = 2%, lead time (L) = 2 years, interest rate (r) = 13%, and economies of scale parameter (a) = 0.99. The results are indicated in Figure 11.

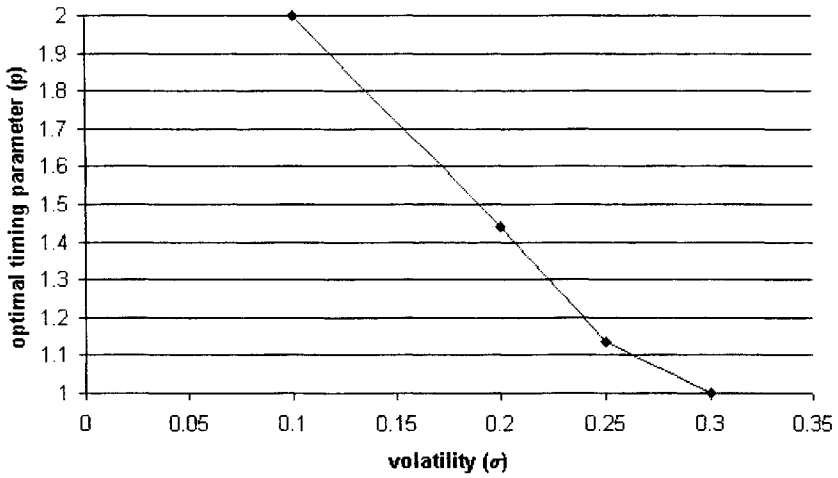

Figure 11: Effect of demand volatility on optimal timing parameter

Table 3: Effect of demand volatility on optimal decision variables

σ	Optimal p^*	Optimal v^*	$f(p^*, v^*)$	$g(p^*, v^*)$
0.1	2	3.32	0.217	-0.0129
0.2	1.44	2.05	0.892	-0.002
0.25	1.135	1.347	1.3968	-0.0093
0.3	1	1.032	2.114	-0.005

From Figure 11 and Table 3, we can see that as the volatility of the demand process increases, it forces the service provider to initiate the expansion earlier and earlier and also the optimal size of the future expansions gets smaller and smaller. Hence for an industry where the demand experienced is highly fluctuating, it is optimal to start the newer capacity expansions earlier and not wait for initial shortages to accumulate. We also observe that the optimal expansion cost grows along with the demand volatility.

Effect of lead time length on the optimal timing parameter

We also studied the effects of the length of the expansion lead time on the optimal starting time of the expansion project. Once again, the parameter values used to test this relationship are: drift (μ) = 2%, volatility (σ) = 20%, interest rate (r) = 13%, and economies of scale parameter (a) = 0.99.

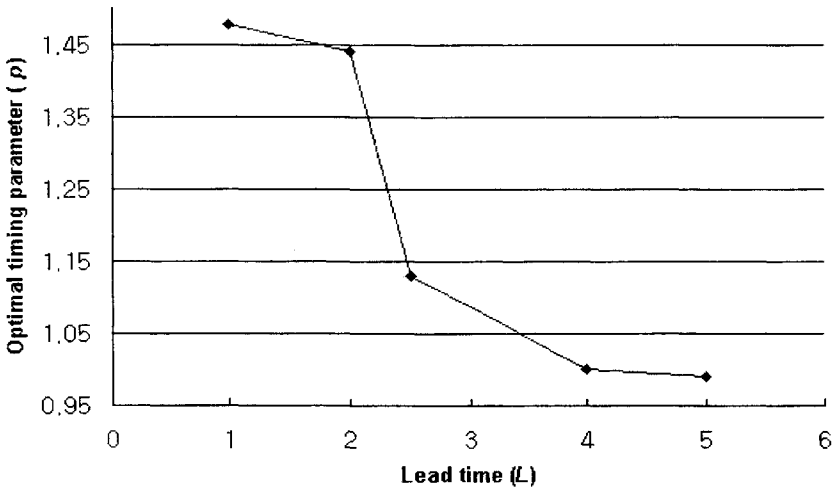

Figure 12: Effect of expansion lead time on optimal timing parameter

Table 4: Effect of expansion lead time on optimal decision variables

L	Optimal p^*	Optimal v^*	$f(p^*, v^*)$	$g(p^*, v^*)$
1	1.477	1.82	0.747	-0.02
2	1.44	2.05	0.892	-0.002
2.5	1.13	1.41	1.006	-0.025
4	1.0	1.008	0.963	-0.0013
5	0.99	1.004	0.962	-4.5e-5

From Figure 12, we can see that as it takes longer and longer to complete a given expansion project, it is optimal to initiate the expansion project with smaller and smaller initial shortage, in order to maintain the given service level.

Effect of the allowed shortage on the optimal timing variable

Lastly, we observe the effect of allowed shortage value (δ) on the optimal values of decision variables, more particularly, the optimal value of the timing variable. The numerical values for the other problem parameters used to study this case were: drift (μ) = 5%, volatility (σ) = 20%, interest rate (r) = 10%, lead time (L) = 0.5, and economies of scale parameter (a) = 0.9. These numerical values were selected from the Ryan (2004) numerical analysis. We find that as we allow more and more shortage during the expansion cycle, the optimal timing variable value increases. Results are indicated in Figure 13. A similar trend was observed in Ryan (2004) though we note that the service level equation in Ryan (2004) is different than ours. In Ryan (2004), the service level was defined as the total unsatisfied demand per unit of capacity over an expansion cycle. With notation similar to our model, for $t_n + L \le t \le t_{n+1} + L$, the shortage at time t as a proportion of installed capacity was defined as:

$$S^n(t) = \frac{\max[P(t) - K_n, 0]}{K_n}.$$

And from this definition, the service level constraint was formulated as:

$$E_{t_n}\left[\int_{\max(t_n+L, T_{n+1})}^{T_{n+1}+L} S^n(u)du\right] \le \varepsilon,$$

where ε was the specified shortage limit. We also note that because of this definition of the service level, the timing policy could be separated from the increment policy. That is, the timing parameter p could be optimally evaluated for, independent of the size parameter v, via setting up correspondence between this expression for the service level and that for the price of a European call option.

In our case, we define the service level as the ratio of the total unsatisfied demand over the total demand during the expansion cycle (refer Equation (1)).

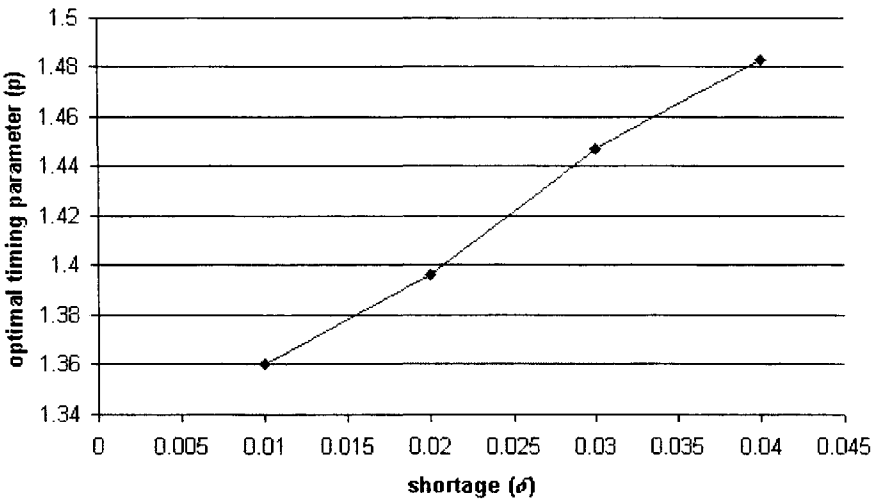

Figure 13: Effect of allowable shortage on optimal timing parameter

Table 5: Effect of allowable shortage on optimal decision variables

δ	Optimal p^*	Optimal v^*	$f(p^*, v^*)$	$g(p^*, v^*)$
0.01	1.36	1.77	2.52	-0.023
0.02	1.44	2.05	2.517	-0.0635
0.03	1.447	1.856	2.381	-0.0327
0.04	1.483	1.8486	2.30	-0.0318

We note that the convergence of the cutting plane algorithm to an optimal solution is dependent on the modeling of the service level equation. We refer to Marathe and Ryan (2006), where the service level equation (Equation (1)) was modeled using some approximations. In this approach, starting from Equation (1), the numerator was evaluated using the same approach mentioned in Chapter 4 for expressing I^l_n in terms of decision variables. Hence, after taking expectation we have the service level as,

$$\beta = E\left[\frac{\int_{t_n+L}^{t_{n+1}+L} \min\left[P(t), K_n\right]dt}{\int_{t_n+L}^{t_{n+1}+L} P(t)\,dt}\right] = 1 - E\left[\frac{\int_{t_n+L}^{t_{n+1}+L} \max\left[P(t) - K_n, 0\right]dt}{\int_{t_n+L}^{t_{n+1}+L} P(t)\,dt}\right].$$

However, an approximation was made for the total demand during the expansion cycle. We first approximated the expectation of the ratio as the ratio of expectation and then approximated the total demand during an expansion cycle by a term $pK_nE[T_{n+1} - t_n]$. The service level, after these approximations was:

$$\beta_n \approx 1 - \frac{E\left[\int_{t_n+L}^{t_{n+1}+L} \left(\max\left[P(t) - K_n, 0\right]/K_n\right)dt\right]}{pE\left[T_{n+1} - t_n\right]}$$

Then the service level constraint was set up directly such the calculated shortage is less than or equal to some specified limit. So instead of having the shortage violation equation $g(p, v)$, we directly had the *approximated* shortage equation. The results from the optimization problem using this direct shortage constraint indicated an unbounded solution. By unbounded solution we mean that minimum expansion cost was achieved with acceptable service level with unlimited initial shortages, which seems unrealistic. Simulation of the GBM demand process was also tried. In this approach, instead of analytically finding the service level value via the financial option pricing method, it was found through simulation of the demand process. With that scenario, too, we encountered unbounded solutions. A complete copy of the said paper (Marathe and Ryan, 2006) is included in Appendix II.

Finally, having known the trend of optimal solutions vis-à-vis the values of the other parameters, we are in a position to draw rich managerial insights from our model. These conclusions are discussed in the next chapter. Also in the next chapter we discuss future extensions of our capacity expansion model.

CHAPTER VI: CONCLUSION AND FUTURE WORK

Based on the work discussed in the previous chapters, now are ready to draw conclusions for our capacity expansion model. The problem parameters and the basic conditions under which we formulated the problem were discussed in Chapter 1. Our initial task was to find examples of industries where the demand for the capacity follows a GBM process assumption. Towards this, a method for checking the GBM fit was discussed in Chapter 2, where we also discussed the case where the data series for the observed demand over time has seasonal effects. From the description of the problem conditions, we mathematically formulated the infinite time horizon expression cost expression (Equation (15)), which was used as the objective function in the optimization problem. The service provider for the capacity we are considering in our problem has an obligation to maintain a certain level of service. This means that the shortages during any expansion cycle cannot be greater than the allowable limit. This service level constraint was formulated using the financial option pricing theory (specifically, the Up-and-Out partial barrier call option price) in Equation (12). This optimization problem (defined in Equation (16)) was found to be difficult to solve, hence an approximation of its dual problem using the cutting plane algorithm was solved, which was discussed in Chapter 5. Using the cutting plane algorithm, optimal solutions to our problem were found for some instances of the parameter values. The trends observed in these optimal solutions were discussed in the last section of Chapter 5. In this chapter, we provide managerial insights based on these results.

6.1 GBM assumption

Given a data series representing demand values over a period of time for any particular capacity, we can check whether this data series satisfies the assumptions of the GBM process. The procedure discussed in Ross (1999) achieves this by checking for normality and independence of the log ratios

(refer Section 3.1). However, this process is complicated when the demand data series has seasonal variation. In this case, the seasonal effect needs to be removed before we check the data series for the GBM fit. We found that the moving average method for removing the seasonal effects is well suited when the data series is to be checked for the assumptions of the GBM process (Marathe and Ryan, 2005). We then applied this method of deseasonalization and subsequent checking of the GBM fit to some actual data series. However, since obtaining data values for the 'actual' demand for capacity is difficult, we surrogated that with data values for actual usage (or sale). For example, to check the demand data for the electric utility industry, we worked with the monthly consumption data. Out of the four data series analyzed, we found that the one characterizing demand values for electric utilities in the USA and also the one representing the demand for airline seats (airline passenger enplanements) followed the GBM process. The two data series corresponding to the number of the registered Internet hubs (to indicate the growth of Internet) and the revenue from the cellular phone industry over a period of 15 years (from 1986 to 2001) failed on either the normality test or the independence test and hence we could not conclude that these data series' agreed with the GBM assumptions. One important reason for the last two data series failing the normality of independence test was an insufficient number of data points. Since there were not enough data values the statistical tests to confirm the GBM assumptions were not statistically significant. Based on this work, we now have a method to check the GBM fit, which works in cases where the demand data series has seasonal variation. With deseasonalization of the data series, we smooth out the data series and take away the seasonal changes in the values. While planning for capacity over a longer time horizon, this deseasonalization is a better technique because the irregularities because of seasonal changes are ironed out. This is true because over a longer time horizon (and we are considering an infinite time horizon for capacity planning), we don't want our capacity expansion policy to fluctuate with every seasonal variation. We are planning for strategic decisions where the seasonal effects from the demand has been removed.

6.2 Capacity expansion problem

Focusing on the numerical results obtained in the previous chapter regarding the actual capacity expansion problem in this section we discuss the trends in the optimal solution to the problem. In the last chapter, we studied the behavior of the optimal solution as we change the values of the other parameters one after the other.

First, we studied the effect of drift parameter of the demand process on the optimal solution to the capacity expansion problem. From Figure 10 and Table 2, we found that as the drift parameter increases, the optimal value of the timing variable (p) decreases. We recall that the decision variable p represents the initiation of the expansion project in the sense that we initiate the $(n+1)^{st}$ expansion project when the demand process hits the value pK_n, where K_n was the capacity position after n expansions. This means that the increase in drift is prompting earlier initiation of the new expansion project. In fact, at higher values of the drift parameter, the optimal value of the timing variable falls below one, indicating that the new expansion is to be initiated before the demand hits the capacity position. A higher drift value for the demand indicates a high growth industry. Hence, we can conclude that for a high growth industry, the service provider should not wait till there is an initial accumulation of shortage before the next expansion project is started. On the other hand, for a low growth industry the service provider has more time before the next expansion project is started and the expansion could optimally be delayed such that by the time when the next project starts, there already is some shortage accumulated. From the data observed for the passenger enplanements in the airline industry, we found that the drift parameter for the observed demand was 3.3%; hence the airline industry, in terms of passenger enplanements, can be characterized as a low growth industry. So for an airline operator following the capacity expansion policy similar to our model, it may be optimal to delay the expansion. We note that a similar trend was not observed for the size of the expansions.

The random demand process also implies that demand volatility and its effect on the expansion policy are critical. We found that (refer Figure 11 and Table 3), for a highly volatile demand the optimal value for the timing variable (p) is smaller than that for the less volatile demand. Hence, similar to the discussion above, we can say that for a highly volatile industry (industry where the demand for the capacity fluctuates to a greater degree), it is optimal to initiate the next expansions even before the demand reaches the current capacity position. A service provider in a more stable industry, however, has the luxury of delaying the expansion and can in fact tolerate initial shortages before the start of the expansion project. This trend of the optimal solution, though, is the opposite of what is observed in the financial options theory where the volatility of the stock price is considered favorable and can be exploited by the writer of the option.

Also from the values of expansion costs optimally incurred (from Table 2 and 3), we can conclude that it is more expensive for the service provider to maintain the service level above some specified limit in cases where the demand is growing at a faster rate or where the demand is highly volatile.

We also observed the effect of the expansion lead time durations on the expansion policy decision variables (Figure 12 and Table 4). It was observed that for industries where it takes a longer time to finish expansion projects, the future expansion projects should be started at earlier times without any initial shortages. However, for industries where the expansion projects can be completed in less time duration, the service provider can afford some initial shortages.

Lastly we quantified the effects of the service level limit itself on the expansion policy and found that more tighter service level goal requires the future expansions to be started early, before the time the demand hits the capacity position; moreover, it also incurs higher expansion costs. On the other hand, when the service provider can allow larger total shortages during the expansion cycle, then the newer expansions could be started later. For a more relaxed service level goal, the expansion cost optimally needed is less than the one for the higher service level case.

We can clearly see that all these results are intuitively logical. We expected that when the lead times are larger, and the probability of higher shortages during the expansion cycle is bigger, the service provider would want to start the expansions earlier. And this is what we found from the numerical study. Similarly, when the service provider can tolerate lower values of achieved service level, we expected the timing variable value to be larger indicating a delayed expansion. Delaying this expansion would reduce the total expansion cost because of the discounting effect and although this delay would reduce the service level, and this would not be a big concern because the service provider can afford that. Once again, our numerical studies confirm this behavior.

Hence we can see that using our capacity expansion model, the service provider can optimally choose an expansion policy, which would be based on the numerical values of the parameters observed in the industry in which the service provider operates. This model takes into effect the uncertainty caused because of the stochastic nature of the demand and still optimally finds the policy parameters for the concerned planner of the capacity.

In the next section, we discuss some of extensions to our current model.

6.3 Future extensions

The current formulation of the capacity planning model works under the conditions mentioned in this dissertation. During these stated assumptions, this model helps the service provider in developing an optimal capacity expansion policy that will minimize the total cost of expansion. However, we can envision multiple directions in which this basic model can be extended to make it more realistic. Some of these directions are discussed in this section.

In our model, we assumed that the demand for the capacity follows a GBM process where the drift and volatility parameter of the process remain constant throughout the period of analysis. In many industries, this assumption may not be satisfied. Returning to the case of the airline industry, we saw that during the initial years of this century, the demand for airline seats actually was very low

after the incidents following 9/11 terrorist attack. Only during recent years has the demand started to rise. To realistically consider a situation like this, one may consider a demand process that can model certain *jumps* during the time of analysis. One example of such a demand process would be a GBM process with jumps. The capacity expansion model then can be modified to consider properties of this different process. Considering a different distribution function for demand may mean that we cannot use the financial option pricing theory to model the service level and have relatively simpler expression for the infinite time horizon expansion cost. However, since many authors have formulated capacity expansion models using different demand processes, we expect this extension to be a relatively easy one.

Secondly, in the current model, we are considering only capacity *expansions*. This means for the airline industry where we are planning for the number of airline pilots, we are currently considering only hiring of new pilots. In the current formulation we are deciding when to hire these new pilots and how many to hire. However, for a situation described above, where the demand for the airline seats have a downward trend for considerable time, the service provider may want to cut costs by laying off some of the pilots and reducing the manpower. This cannot be achieved with the current problem. Hence in the future the definition of size parameter (v) can be modified to reflect *firing* of pilots as an admissible policy and/or consider attrition due to retirement of pilots. Retiring part of capacity over a period of time has been considered in some capacity expansion models. In fact, this could mean that we have to move away from the stationary expansion policy considered in our current model, because the service provider may want to hire pilots during some expansion cycles and may want to lay off pilots during some expansion cycles when the demand has been low. This would make the model considerably more difficult to formulate.

In our model, we have ignored the impact of price elasticity through our assumption that demand grows independently of capacity. This is true for some industries such as electric utility. However, for some industries, the demand is dependent on the available capacity and hence the price.

In the future, this price elasticity could be included in the model. Inclusion of this will change the demand distribution to an extent. And since we use the properties of the demand distribution to arrive at analytical expressions for the service level and the expansion costs, this extension to the model could lead to substantial complications in the current analysis.

Lastly, for the model described in this dissertation, we considered that the expansion lead time remains constant. This was done to transfer all the uncertainty to the demand process. In the future, to include the uncertainty effects due to the lead time, a random lead time could be considered. Specifically, an exponentially distributed expansion lead time that is independent of the demand process could be considered for the capacity expansion model. We think that this could be easily accomplished. Also, in reality, because of production limitations or other reasons, the lead time may depend on the size of expansion. The current expansion problem does not consider this case. Making the lead time dependent on the size of the expansion can be interesting and more realistic. This would change the expression for the service level because of the change in the integration limits. We think that this change would not be very difficult to incorporate in the model. However, this would be harder than making the lead time exponentially distributed.

REFERENCES

1. Angelus, A. and E. L. Porteus, (2003)."On capacity expansions and deferrals." *Working Paper,* Graduate School of Business, Stanford University.

2. Atlason, J., M. A. Epelman and S. G. Henderson, (2004). "Call center staffing with simulation and cutting plane methods." *Annals of Operations Research* 127, 333-358.

3. Barker, L., (2000). "Pilot shortages: How to reduce their Impact on rural and smaller market." *Committee on Commerce, Science and Transportation, Subcommittee on Aviation and Aeronautics* (Viewed June, 2005) http://commerce.senate/gov/hearings/0725bar.pdf.

4. Bazaraa, M. S., H. D. Sherali and C. M. Shetty (1993). *Nonlinear programming: Theory and algorithm.* New York, John Wiley and Sons.

5. Bean, J. C., J. L. Higle and R. L. Smith, (1992). "Capacity expansion under stochastic demands." *Operations Research* 40(2), S210-S216.

6. Bean, J. C. and R. L. Smith, (1985). "Optimal capacity expansion over an infinite horizon." *Management Science* 31(12), 1523-1532.

7. Birge, J. R., (2000). "Options methods for incorporating risk into linear capacity planning models." *Manufacturing and Service Operations Management* 2(1), 19-31.

8. Borodin, A. N. and P. Salminen (2002). *Handbook of Brownian motion: facts and formulae.* Basel, Boston, Birkhäuser.

9. Buchan, J. and N. Edwards, (2001). "Nursing numbers in Britain: The argument for workforce planning." *British Medical Journal* 320, 1067-1070.

10. Buzacott, J. A. and A. B. Chaouch, (1988). "Capacity expansion with interrupted demand growth." *European Journal of Operational Research* 34, 19-26.

11. Cakanyildirim, M. and R. O. Roundy, (2002)."Capacity expansion and contraction under demand uncertainty." *Technical paper,* School of Management, University of Texas at Dallas.

12. Carr, P., (1995). "Two extensions to barrier options pricing." *Applied Mathematical Finance* 2(4), 1-39.

13. Chan, B., (2003). "Physician workforce planning: What have we learned? Lessons for planning medical school capacity and IMG policies. The Canadian perspective". *International Medical Workforce Conference*, Oxford, UK.

14. Chaouch, A. B. and J. A. Buzacott, (1994). "The effects of lead time on plant timing and size." *Production and Operations Management* 3(1), 38-54.

15. Chuang, C.-S., (1996). "Joint distribution of Brownian motion and its maximum, with a generalization to correlated BM and applications to barrier options." *Statistics and Probability Letters* 28, 81-90.

16. Davis, M. H. A., M. A. H. Dempster, S. P. Sethi and D. Vermes, (1987). "Optimal capacity expansion under uncertainty." *Advances in Applied Probability* 19, 156- 176.

17. Dellaert, N. and T. de Kok, (2004). "Integrating resource and production decisions in a simple multi-stage assembly system." *International Journal of Production Economics* 90, 281-294.

18. Donohue, G. L., (2000). "Testimony before the House of Representatives." *Committee on Science, Subcommittee on Space and Aeronautics* (Viewed June, 2005) http://house.gov/science/donohue.

19. Edwards, J. S. and R. W. Morgan, (1982). "Chapter 4: Optimal control models in manpower planning". *Optimisation and Control of Dynamic Operational Research Models*. Ed. Tzafestas. North--Holland Amsterdam.

20. Erlenkotter, D., (1976). "Coordinating scale and sequencing decisions for water resources projects". *Economic Modeling for Water Policy Evaluation*. Ed. R. M. Thrall. North-Holland Amsterdam.

21. Freidenfelds, J. (1981). *Capacity expansion: analysis of simple models with application*. New York, North Holland.

22. Goldman, B. M., H. B. Sosin and M. A. Gatto, (1979). "Path dependent options: buy at the low; sell at the high." *The Journal of Finance* XXXIV(5), 1111-1127.

23. Gomory, R. E., (1963). "An algorithm for integer solutions to linear programs". *Recent advances in mathematical programming*. Ed. R. L. Graves and P. Wolfe. McGraw- Hill Book Company, Inc. New York.

24. Hadley, G. and T. M. Whitin (1963). *Analysis of inventory systems*. New Jersey, Prentice-Hall Inc Englewood Cliffs.

25. Heynen, R. C. and H. M. Kat, (1997). "Barrier options". *Exotic Options*. Ed. L. Clewlow and C. Strikland. International Thompson Business Press: 125-138.

26. Holt, C. C., F. Mondigliani, J. F. Muth and H. A. Simon (1963). *Planning production, inventories, and work force*. Englewood Cliffs, NJ, Prentice--Hall Inc.

27. Hopkins, G. E., (2001). "A short history of pilot shortages." *Air Line Pilot* February 2001.

28. Hull, J. C. (1999). *Futures, and other derivatives*. New Jersey, Prentice Hall.

29. Kanellos, M., (2004) "Take 2 for PC memory." CNET News.com, (Viewed, June 2006) http://news.com.com/Take+2+for+PC+memory/2100-1004_3-5190234.html

30. Karlin, S. and H. M. Taylor (1975). *A first course in stochastic processes*. New York, Academic Press.

31. Kelley, J. E., (1960). "The cutting plane method for solving convex programs." *Journal of Society of Industrial Applications of Mathematics* 8(4), 703-712.

32. Kelley Jr., J. E., (1960). "The cutting plane method for solving convex programs." *Journal of Society of Industrial Applications of Mathematics* 8(4), 703-712.

33. Klemm, H., (1971). "On the operating characteristic 'service level'". *Inventory Control and Water Storage*. North-Holland Publishing Company Amsterdam: 169-178.

34. Lieberman, M. B., (1989). "Capacity utilization: theoretical models and empirical tests." *European Journal of Operations Research* 40, 155-168.

35. Manne, A. S., (1961). "Capacity expansion and probabilistic growth." *Econometrica* 29(4), 632-649.

36. Marathe, R. R. and S. M. Ryan, (2005). "On the validity of geometric Brownian motion assumption." *The Engineering Economist* 50(2), 159-192.

37. Marathe, R. R. and S. M. Ryan, (2006). "Optimal solution to a capacity expansion problem". *IIE Annual Research Conference*, Orlando, FL.

38. Merton, R., (1973). "Theory of rational pricing." *Bell Journal of Economics and Management Science* 4(1), 141-183.

39. Musiela, M. and M. Rutkowski (1997). *Martingale methods in financial modeling*, Springer.

40. Nembhard, H. B., L. Shi and M. Aktan, (2002). "A Real options design for quality control charts." *The Engineering Economist* 47(1), 28-50.

41. O'Brien- Pallas, L., S. Birch, A. Baumann and G. T. Murphy, (2001). "Integrating workforce planning, human resources, and service planning". *World Health Organization, Department of Health Services Delivery, Workshop on Global Health Workforce Strategy*, Annecy, France.

42. Pak, D., N. Pornsalnuwat and S. M. Ryan, (2004). "The effect of technological improvement on capacity expansion for uncertain exponential demand with lead time." *The Engineering Economist* 49(2), 95- 118.

43. Rich, D. R., (1994). "The mathematical foundations of barrier option-pricing theory." *Advances in Futures and Operations Research* 7, 267-311.

44. Ritchken, P., (1995). "On pricing barrier options." *The Journal of Derivatives* 3, 19-28.

45. Rose, C. and M. D. Smith, (1996). "The multivariate normal distribution." *The Mathematica Journal* 6(1), 32-37.

46. Ross, S. M. (1999). *An introduction to mathematical finance*. Cambridge, UK, New York, Cambridge University Press.

47. Rubinstein, M., (1991)."Exotic options." *Unpublished Manuscript,* University of California at Berkeley.

48. Rubinstein, M. and E. Reiner (1991). "Breaking down the barriers." *Risk* 4, 28-35.

49. Ryan, S. M., (2004). "Capacity expansion for random exponential demand growth with lead time." *Management Science* 50(6), 740-748.

50. Schneider, H., (1981). "Effect of service-levels on order-points or order-levels in inventory models." *International Journal of Production Research* 19(6), 615-631.

51. Shameen, A., (2000). "Too much of a good thing? Red-hot demand raises fears of chip shortages." 26(20). (Retrieved July, 2006), from http://www.asiaweek.com/asiaweek/magazine/2000/0526/biz.semi.html.

52. Shim, R., (2004) "Parts shortages could hang up Treo 600 sales." CNET News.com, (Viewed July, 2006) http://news.com.com/Parts+shortage+could+hang+up+Treo+600+sales/2100-1041_3-5187602.html

53. Smith, R. L., (1979). "Turnpike results for single location capacity expansion." *Management Science* 25(5), 474-484.

54. Smith, R. L., (1980). "Optimal expansion policies for the deterministic capacity problem." *Engineering Economist* 25(3), 149-160.

55. Sobel, J. M., (2004). "Fill rates of single-stage and multistage supply system." *Manufacturing and Service Operations Management* 6(1), 41-52.

56. Tan, T. and O. Alp, (2005). "An integrated approach to inventory and flexible capacity management under non-stationary stochastic demand and set-up cost". *The Fifth International Conference on Analysis of Manufacturing Systems- Production Management*, Zakynthos Island, Greece.

57. The Hindu, (2005). "NTPC revises capacity expansion target upwards for the Eleventh Plan." (Retrieved July, 2006), from http://www.thehindubusinessline.com/2005/01/26/stories/2005012602340200.htm.

58. Thorsen, B. J., (1998). "Afforestation as a real option: some policy implications." *Forest Science* 45(2), 171-178.

59. Thrall, R. M., (1976). "Economic modeling for water policy evaluation. Ed. R. M. Thrall. North-Holland Amsterdam.

60. U.S - Canada Power Outage Task Force, (2004). "Final Report on the August 14, 2003 Blackout in the United States and Canada- Causes and Recommendations " (Viewed July, 2006) https://reports.energy.gov.

61. Van Mieghem, J. A., (2003). "Capacity management, investment, and hedging: review and recent developments." *Manufacturing and Service Operations Management* 5(4), 269-302.

62. Whitt, W., (1981). "The stationary distribution of a stochastic clearing process." *Operations Research* 29(2), 294-308.

63. Woerth, D. E., (2000). "Statement of Captain Duane E. Woerth, President, Air Line Pilots Association." *Subcommittee on Aviation, Committee on Commerce, Science, and Transportation, United State Senate, On the Pilot Shortages and Effects on Rural Air Service* (Viewed July, 2006) http://cf.alpa.org/Internet/TM/tm072500.htm.

64. Wolfe, P., (1961). "Accelerating the cutting plane method for nonlinear programming." *Journal of Society of Industrial Applications of Mathematics* 9(3), 481-488.

65. Wolfram Research Inc, (2004). *Mathematica 5.1*, http://www.wolfram.com/

66. Wollman, N., (1976). "Water resource models: A historical summary". *Economic Modeling for Water Policy Evaluation* Ed. R. M. Thrall. North-Holland Amsterdam.

67. Young, A. and T. Abodunde, (1979). "Personnel recruitment policies and long-term production planning." *Journal of the Operational Research Society* 30(3), 225-236.

68. Yu, G., J. Pachon, B. Thengvall, D. Chandler and A. Wilson, (2004). "Optimizing pilot planning and training for continental airlines." *Interfaces* 34(4), 253-264.

69. Zangwill, W. I. (1969). *Nonlinear programming: a unified approach.* Englewood Cliffs, New Jersey, Prentice Hall.

APPENDICES

Appendix I: Marathe, R.R., and S.M. Ryan, (2005). "On the validity of the GBM assumption," *The Engineering Economist*, 50(2), 159- 192.

Appendix II: Marathe, R.R., and S.M. Ryan, (2006). "Optimal solution to a capacity expansion problem," *Proceedings of the 15th Annual IIE Research Conference, May 20- 24, 2006, Orlando Florida.*

Appendix III: Marathe, R.R., and S.M. Ryan, (2005), "Capacity expansion for uncertain demand with initial shortages," *Proceedings of the 14th Annual IIE Research Conference, May 14-18, 2005, Atlanta, Georgia.*

Appendix IV: *Mathematica* Code

4A- LP for checking the concavity of a function.

4B- Optimization code representing the service level expression.

Appendix I: On the Validity of the Geometric Brownian Motion Assumption

Rahul R. Marathe
Department of Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, IA 50011-2164

Sarah M. Ryan*
Department of Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, IA 50011-2164

Corresponding author: smryan@iastate.edu Phone: 515-294-4347

On the Validity of the Geometric Brownian Motion Assumption

Abstract

The geometric Brownian motion (GBM) process is frequently invoked as a model for such diverse quantities as stock prices, natural resource prices, and the growth in demand for products or services. We discuss a process for checking whether a given time series follows the GBM process. Methods to remove seasonal variation from such a time series are also analyzed. Of four industries studied, the historical time series for usage of established services meet the criteria for a GBM, however the data for growth of emergent services do not.

I. Introduction

Many recent engineering economic analyses have relied on an implicit or explicit assumption that some quantity that changes over time with uncertainty follows a geometric Brownian motion (GBM) process. Below we briefly review a number of applications in different areas. The GBM process, also sometimes called a lognormal growth process, has gained wide acceptance as a valid model for the growth in the price of a stock over time. In fact, [9] refers to it as "the model for stock prices". Under this model, the Black-Scholes formulas for pricing European call and put options, as well as their variations for a few of the more complex derivatives, provide relatively simple analytical evaluation of asymmetric risks. The increasingly numerous and varied applications of the GBM model to processes other than the stock price motivate this paper: to review the assumptions underlying the GBM model, outline established statistical procedures for checking these assumptions, and illustrate their applications to actual data series.

Many recent examples of GBM models have arisen in real options analysis, in which the value of some "underlying asset" is assumed to evolve similarly to a stock price. In some cases, the GBM assumption is stated explicitly, while in others it is implicitly used when options are evaluated by the Black-Scholes formula. In [17], the cost of applying quality control charts was quantified using real option pricing methods, where both the sales volume and the price of a product were assumed to follow GBM processes. The same authors discussed the problem of product outsourcing as a real options problem in [18]. Here, three variables are supposed to follow the GBM process; viz. the unit cost of internally producing the item, the unit outsourcing cost of the item, and the unit delivery cost of outsourced items during the time interval. The GBM process has been also assumed in problems related to natural resources. In [25] the real options theory is applied to decisions of establishing a new forest stand and it is assumed that the future net prices of roundwood follow a GBM process. In [2], the Black-Scholes option pricing formula is applied to the capital allocation for investment. For the machine replacement problem considered in the paper, the present value of the machine cash flows is modeled as a GBM process. The options value of expansion flexibility in evaluating manufacturing investment is studied in [13], wherein the authors use sequential exchange options to value expansion flexibility in justifying the investment. In valuing flexibility an initial investment is considered as being analogous to purchasing an option to exchange one risky asset (subsequent investment, called the delivery asset) for another risky asset (returns

accruing from the subsequent asset, called the optioned asset) within a time period from the initial investment. The prices for both of the assets are assumed to follow the GBM.

The GBM model has also been used to represent future demand in capacity studies. In [28], the authors studied capacity utilization over time assuming demand followed a GBM. An indirect validation of the assumption was provided by [14], which showed in an empirical study of the chemical industry that actual capacity utilization matched the predictions from the model in [28]. In [22] demand for services in rapidly growing industries was assumed to follow a GBM and the expansion policy to minimize cost subject to a service level constraint was developed and analyzed. In this paper, we analyze data to test whether the GBM model is valid for demand; however, the methods we employ are applicable to any data series. Our approach is motivated by Ross [20], who analyzed data for crude oil prices and found that they were inconsistent with a key assumption of the GBM model. In this paper we also consider seasonal effects and show how to remove them before testing for the GBM characteristics. The effect of seasonality may be overcome in financial markets: Samuelson [23] proved that, even when spot prices have systematic seasonal variation, futures prices will not. However, in the demand series we examine in this paper, there are no quantities analogous to futures prices. As pointed out in [25], the GBM process assumption must be subject to test. Where significant financial impacts may result from the decision, it is of utmost importance to verify that a time series follows the GBM process, before relying on the result of such an assumption.

The next section discusses the theory of the GBM process and the parameters involved. The definitions and concepts of the Brownian motion used in the paper are explained in this section. Some data series may contain seasonal variation in addition to exponential growth with uncertainty. Hence before testing the GBM assumption the data series must be deseasonalized. In Section III, two methods of removing the seasonal indices are studied and the unbiased method is selected. Finally, the theory and methods developed in sections II and III are applied to real-life time series in Section IV. The data analyzed in this paper are from varied industries. As the cellular phone industry has been growing multi-fold over short recent intervals, it makes an interesting case to be considered as a GBM process. Also analyzed are airline passenger enplanements, electric power consumption and the growth of the Internet. Finally the results obtained from the data are discussed and summarized. We have our concluding remarks and plan for future work in the last section.

II. Geometric Brownian Motion (GBM) Process

2.1 Preliminaries

A Markov process is a particular type of stochastic process where only the present value of a variable is relevant for predicting the future. The past history of the variable and the way that the present has emerged from the past are irrelevant. A Wiener process is a type of Markov stochastic process in which the mean change in the value of the variable is zero with the variance of change equal to one per unit time. The Wiener process was first applied in physics to describe the motion of a particle that is subject to a large number of small molecular shocks and was called Brownian motion [9]. The mathematical description of the process was later developed by Wiener [20].

If a stochastic process $\{z(t), t \geq 0\}$ follows a Brownian motion process, it exhibits the following two properties.

- Property I: The change in the value of z, Δz, over a time interval of length Δt is proportional to the square root of Δt where the multiplier is random; specifically, $\Delta z = z(t+\Delta t) - z(t) = \varepsilon\sqrt{\Delta t}$, where ε is a standard normal random variable. Hence values of Δz follow a normal distribution with mean 0 and variance equal to the change in time (Δt) over which Δz is measured.
- Property II: The changes in the value of $z(t)$ for any two non-overlapping intervals of time are independent.

Using the principles of ordinary calculus where it is usual to proceed from small changes to the limit as the small changes becomes closer to zero, the Wiener process is the limit as $\Delta t \to 0$ of the process described above for $z(t)$.

A Wiener process is not differentiable with respect to time [15] as seen from the fact that:

$$E\left[\frac{z(s)-z(t)}{s-t}\right]^2 = \frac{s-t}{(s-t)^2} = \frac{1}{s-t} \to \infty, \text{ as } s-t \to 0.$$

However, it is useful to define a term for the expression dz/dt. A term commonly used in engineering to denote this quantity is white noise. The white noise process is the derivative of the Brownian motion process, which does not exist in the normal sense.

The standard Brownian motion process has a drift rate of zero and a variance of one. The drift rate of zero means that the expected value of z at any future time is equal to the current value. The variance of one means that variance of the change in z in a time interval of length T is equal to T. The Brownian motion process is the basis for a collection of more general processes. These generalizations are obtained by inserting white noise in an ordinary differential equation.

A generalized Brownian motion process is of the type: $dx = a\,dt + b\,dz$, where a and b are constants and z is a Brownian motion process. To understand the equation, each of the components is considered separately. The first term implies that x has an expected drift rate of a per time unit, whereas the second term involving dz can be regarded as adding noise or variability to the path followed by x. The amount of this noise is b times the differential of the Brownian motion process. Hence for a small interval of time, the change in the value of x, Δx, is given by

$$\Delta x = a\Delta t + b\varepsilon\sqrt{\Delta t}\ .$$

Thus Δx has a normal distribution with mean $a\Delta t$ and variance $b^2\Delta t$.

Further generalization of the Wiener process yields the Ito process, where the constants a and b may depend on the values of x and t. The Ito process is of the form [9]:

$$dx = a(x,t)dt + b(x,t)dz\ .$$

2.2 Definition of Geometric Brownian Motion Process

The case of stock prices is slightly different from the generalized Brownian motion process. In the case of the Brownian motion process, a constant drift rate was assumed. However, in the case of stock prices, it is not the drift rate that is constant. For stock prices, the return on investment is assumed to be constant, where the rate of return at a given time is the ratio of the drift rate to the value of the stock at that time. Hence the constant expected drift-rate assumption in the case of Brownian motion process is inappropriate and needs to be replaced by an assumption of constant expected rate of return [9].

Let Y be the price of the stock at time t and assume the expected drift rate is μY for some constant μ. This means that in a short interval of time Δt, the expected increase in Y is $\mu Y\Delta t$. The constant parameter μ is

the expected rate of return. If the volatility of the stock price is zero, then the model implies that $\Delta Y = \mu Y \Delta t$, and when the limit is taken as $\Delta t \to 0$, the expected stock price at time T finally becomes $E[Y_T] = Y_0 e^{\mu T}$, where Y_0 is the original value.

However, the stock prices do have volatility. Hence taking that into consideration, the above model can be written as

$$dY = \mu Y dt + \sigma Y dz, \text{ or } \frac{\Delta Y}{Y} = \mu \Delta t + \sigma \varepsilon \sqrt{\Delta t}.$$

The first term of the second equation above is the expected value of the return provided by the stock for a time period of Δt and the second term is the stochastic component of the return. Here σ is the volatility rate. Taking limits as $\Delta t \to 0$, we have

$$E[Y_T] = Y_0 e^{(\mu + \frac{\sigma^2}{2})T}.$$

Geometric Brownian motion is useful in modeling stock prices over time when one believes that the percentage changes over equal length, non-overlapping intervals are independent and identically distributed. For example, if Y_n is the price of the stock at time n = 0, 1, 2... then it is reasonable to suppose that the ratios Y_{n+1}/Y_n, $n \geq 1$, are independent and identically distributed [21]. Let $u_n = \frac{Y_{n+1}}{Y_n}$. After taking the log of both sides and rearranging, we have, $\ln(Y_{n+1}) = \ln(Y_n) + \ln(u_n)$. Now let $w(n) = \ln(u_n)$, that is $w(n) = \ln(Y_{n+1}) - \ln(Y_n)$.

If $w(n)$ for n $\geq$ 1 are independent and are identically distributed normal random variables with mean μ and variance σ^2, it can be said that the variable u_n will have a lognormal distribution [15]. The successive prices can be found to be [15] $Y_t = u_{t-1} u_{t-2} \cdots u_0 Y_0$. Taking the natural log of this equation, we have

$$\ln[Y_t] = \ln[Y_0] + \sum_{i=0}^{k-1} \ln[u_i] = \ln[Y_0] + \sum_{i=0}^{k-1} w(i).$$

The term $ln[Y_0]$ is constant, and the $w(i)$'s are each normal random variables. Since the sum of normal variables is a normal random variable, it follows that $ln[Y_t]$ is a normal random variable. Hence the stock price Y_t has a lognormal distribution, with

$$E\left[\ln\left(\frac{Y_t}{Y_0}\right)\right] = \mu t$$
$$Var\left(\ln\left(\frac{Y_t}{Y_0}\right)\right) = \sigma^2 t\ .$$

Thus, it can be seen that the ratio $\ln\left(\frac{Y_{k+t}}{Y_k}\right)$ has distribution approaching that of a normal random variable with mean μt and variance $\sigma^2 t$.

The Geometric Brownian Motion process can formally be defined as follows [20]:

We say that the variable Y_k, $0 \leq k < \infty$, follows a GBM (with drift parameter μ and volatility parameter σ) if, for all nonnegative values of k and t, the random variable $\frac{Y_{k+t}}{Y_k}$ is independent of all values of the variable up to time k and if in addition, $\ln\left(\frac{Y_{k+t}}{Y_k}\right)$ has a normal distribution with mean μt and variance $\sigma^2 t$, independent of k, where μ and σ are constants.

2.3 Checking for GBM Process Fit

After any seasonal variation is removed from the data, the data can be tested for the GBM process. Referring to the analysis above, there are two assumptions to be satisfied [20]:

1. Normality of the log ratios ($w(k)$) with constant mean and variance
2. Independence from previous data (log ratios independent of their past values).

2.3.1 Normality:

The simplest (however not very accurate) way to check for normality is to plot a histogram of the log ratios and compare it to a normal distribution plot. Another graphical method of testing the normality assumption is to examine the **normal probability plot**. A normal probability plot, also known as a normal Q-Q plot or normal quantile-quantile plot, is the plot of the ordered data values against the associated quantiles of the normal distribution. For data from a normal distribution, the points of the plot should lie close to a straight line.

The statistical tests of normality can be conducted in many ways by using any of the goodness-of-fit tests on the *w(k)* values. One way is to run a chi-square test for goodness-of-fit. Another goodness-of-fit test is the **Shapiro-Wilk W** Test [24]. This is the test used in the statistical package JMP for $n \leq 2000$ [12]. In this test, the hypothesis set is:

H_0: The distribution is normal, against

H_a: The distribution is not normal.

The test gives the value of the statistic 'W' and the corresponding p-value. The p-value is compared to the specified level of significance α. If the observed p-value is greater than the level of significance the test statistic is not in the rejection region and the null hypothesis of a normal distribution cannot be rejected. Note that a large p-value does not definitively identify the data as normally distributed; it only means that the data could plausibly have been generated by a normal distribution.

2.3.2 Independence from the past data:

To test the serial independence of the *w(k)*, the **chi-square test on two-way tables** [3] can be used. The chi-square test provides a method for testing the association between the row and column variables in a two-way table. The null hypothesis is

H_0: There is no association between the variables (in other words, one variable does not vary according to the other variable), while the alternative hypothesis is

H_a: Some association does exist. (The alternative hypothesis does not specify the type of association; close attention to the data is required to interpret the information provided by the test.)

The chi-square test is based on a test statistic that measures the divergence of the observed data from the values that would be expected under the null hypothesis of no association.

To test serial independence of the *w(k)* values, the two variables in the chi-square test are *w(k)* and *w(k+1)* for each *k*. To carry out the test the log ratios are segregated into different groups (or intervals) depending on the number of data points and the range of data values. These groups or intervals of the log ratios are formed in such a way that number of observed values in each of the intervals is approximately equal. The two way table is formulated on the concept that the probability of *w(k)* being in state *j* (interval *j*) now after

being in state i (interval i) in the last period is equal for all j. Equivalently, if a daily data series follows a GBM process, then tomorrow's state will not depend on today's state. One way to verify that is to see the proportion of time that an observation in state i is followed by a state j observation [20]. Thus the two way table is constructed with rows of state i and columns of state j. Under the GBM process model, tomorrow's change would be unaffected by today's change and so the theoretically expected percentages in the two-way table would be same for all rows. The expected value for each cell in a two-way table is equal to $\frac{(row\,total)*(col.total)}{n}$, where n is the total number of observations included in the table, row total is the total number of data points in state i, and column total is the total number of data points in state j. Once the expected values have been computed, the chi-square test statistic is computed as $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected frequency}}$ where the square of the differences between the observed and expected values in each cell, divided by the expected value, are added across all of the cells in the table.

The distribution of the statistic X^2 is chi-square with $(r\text{-}1)(c\text{-}1)$ degrees of freedom, where r represents the number of rows in the two-way table and c represents the number of columns. The p-value for the chi-square test is $P(\chi^2 \geq X^2)$, the probability of observing a value at least as extreme as the test statistic for a chi-square distribution with $(r\text{-}1)(c\text{-}1)$ degrees of freedom. A small p-value indicates support for the alternative hypothesis; in our case suggesting that successive log ratios are not independent. Note once again that a p-value greater than the chosen level of significance does not positively confirm that the log ratios are serially independent, but it indicates that the data do not contradict that assumption.

III. Seasonality

As mentioned above, the data from various industries were considered for their fit to the GBM process. In some cases the time series exhibited trend and/or seasonal patterns. The usual assumption is that four separate components – trend, cyclical, seasonal and irregular – combine to provide specific values for the time series data [1]. The GBM process can account for exponential trend via the drift term and irregularity in terms of the white noise process; however, it does not include cyclical or seasonal effects. In this paper, we neglect the cyclic variation and consider the component of the time series that represents the variability in the data due to

seasonal influences. It is usual to consider the seasonal movement to be occurring annually, however it should be noted that the season could also be different from a year.

Two common models for decomposing a time series, which aim to isolate each component of the series as accurately as possible, are the additive model and the multiplicative model.

Suppose X_t is the time series value at period t, S_t is the seasonal index at period t, T_t is the trend-cycle component at period t, and E_t is the irregular component at period t,

The additive model has the form $X_t = S_t + T_t + E_t$. That is, the seasonal, trend, and irregular components are added together to give the observed series. In the additive model, the seasonal indices over the periods of a particular season add up to zero [4].

Alternatively, the multiplicative decomposition has the form $X_t = S_t T_t E_t$. Here, the seasonal, trend-cycle and irregular components are multiplied to give the observed series [16]. In multiplicative model, the average seasonal index for a season is unity [1].

An additive model is used if the magnitude of the seasonal fluctuations does not vary with the level of the series. However, if the seasonal fluctuation increases or decreases in the level of the series, then a multiplicative model is more appropriate. As seen from the data analysis, for the data series considered in this paper, the magnitude of seasonal variation increases with time (please refer to Figures 1 and 6 in Section IV). Hence a multiplicative model is used. Often the transformed series can be modeled additively, when the original data are not additive. Logarithms, in particular, turn a multiplicative relationship into an additive relationship [16], since

$$X_t = S_t T_t E_t \Rightarrow \ln[X_t] = \ln[S_t] + \ln[T_t] + \ln[E_t]$$

Suppose we have observations $X_1, X_2 \ldots X_T$ of a process; in particular, for our model, $X_t = Y_t S_t$ or

$$\ln[X_t] = \ln[Y_t] + \ln[S_t] \tag{1}$$

where Y_t are observations at discrete time points of a GBM process and S_t is the seasonal factor.

The observations of the process can also be recorded in terms of the seasons and periods. Let X_{ij} be the observation corresponding to the j th period of the i th season, where $i = 1 \ldots m$ and $j = 1 \ldots p$, that is, we have data for m seasons, with each season having p periods in it and $T = mp$. Correspondingly, let Y_{ij} be the

observation of the j th period of the i th season of a GBM process; and let S_{ij} be the seasonal index for period j of season i, where $S_{ij} = S_j$, for each i. Then, $X_{ij} = S_j Y_{ij}$. Hence Equation (1) can be written as:

$$\ln[X_{ij}] = \ln[Y_{ij}] + \ln[S_j] \tag{2}$$

where in Equation (1), $t = (i-1)p + j$.

In the following, we treat this as a usual additive model, the only difference being that we are using log values, instead of the actual values.

Our goal is to remove the seasonal effects from the time series. This process is referred to as deseasonalization [1]. The first step in deseasonalization is to estimate the values of the seasonal indices for each period. And once the estimates $\hat{S}_j$ of the seasonal variation for each period j are found out, the lognormal variable Y can be estimated using the equation

$$\ln[\hat{Y}_{ij}] = \ln[X_{ij}] - \ln[\hat{S}_j] \tag{3}$$

In the additive model, two estimation methods have been proposed. The analysis of both the methods with respect to the GBM model that is to be tested is included in the following sections. The two methods are compared on the basis of bias and the unbiased one selected. The first method uses moving averages of the consecutive data values [4], [16] and the second one uses averages of all the data values corresponding to each period of the season in turn [8]. In this paper we examine the estimates of the seasonal indices obtained from the series in Equation (1) or (2) using both the methods to see whether the estimates add up to zero and are free of bias.

3.1 Method I [3], [4], [16]

Here, we use the arithmetic centered moving average. The arithmetic moving average of $(2t+1)$ data points centered at k is calculated by $\frac{(L_{k-t} + L_{k-t+1} + \ldots + L_k \ldots + L_{k+t})}{2t+1}$ where $\{L_i: i = 1, 2, 3\ldots\}$ is the series of data points. The moving averages isolate the seasonal components, which then can be estimated in the case of the additive model by subtracting the moving average from the corresponding data series value. The values thus found are estimates of the seasonal indices for those periods. Hence, first we make sure that these estimated

seasonal indices for any season add up to zero in case of additive model, that is $\sum_{j=1}^{p} E[\hat{S}_j] = 0$. Secondly, we prove that these calculated values are unbiased estimators of the actual seasonal index. That is, $E[\hat{S}_j] = S_j$.

Now, since Y_t follows a GBM process, continuing from the Section 2.2, from the properties of lognormal distribution [15], given Y_1, we have,

$$E[\ln Y_t] = E[\ln Y_1] + \left(\mu + \frac{\sigma^2}{2}\right)(t-1) = E[\ln Y_1] + \delta(t-1), \text{ where } \delta = \mu + \frac{\sigma^2}{2}.$$

So, if we let Y_{11} be the first value of the series,

$$E[\ln Y_{ij}] = E[\ln Y_{11}] + \delta(t-1), \text{ where } t = (i-1)p + j \tag{4}$$

Let P_{ij} represent the arithmetic moving average for the j th period of the i th season (year, for example). Let $\lceil x \rceil$ denote the smallest integer that is greater than or equal to x, and $\lfloor x \rfloor$ denote the largest integer that is less than or equal to x.

From [4], [16], for our model, the centered moving average P_{ij} when p is odd will be given by

$$P_{ij} = \frac{(\ln X_{t-\lfloor p/2 \rfloor} + \ldots \ln X_t + \ldots \ln X_{t+\lfloor p/2 \rfloor})}{p}, \text{where } t = (i-1)p + j.$$

And when the number of periods p is even, the centered moving average is calculated as [16]:

$$P_{ij} = \frac{1}{p}\left(0.5\ln X_{t-\frac{p}{2}} + \ln X_{t-\frac{p}{2}+1} + \ldots + \ln X_{t+\frac{p}{2}-1} + 0.5\ln X_{t+\frac{p}{2}}\right), \text{where } t = (i-1)p + j. \tag{5}$$

After calculating the values of centered moving average, we compute the deviation, $\ln \hat{S}_{ij} = \ln X_{ij} - P_{ij}$, to estimate the log of the seasonal index for the period j based on season i. The log of the estimated seasonal index for a period is calculated as a simple arithmetic average of log of all the seasonal indices for that particular period from all the seasons. That is,

$$E[\ln \hat{S}_j] = \frac{1}{m}\sum_{i=1}^{m} E[\ln \hat{S}_{ij}]. \tag{6}$$

In particular, for an odd number of periods p,

$$\ln \hat{S}_j = \frac{1}{m}\sum_{i=1}^{m} \ln \hat{S}_{ij}, \text{ for } j = \lceil p/2 \rceil$$

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=2}^{m} \ln \hat{S}_{ij}, \text{ for } j = 1 \text{ to } \lfloor p/2 \rfloor$$

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=1}^{m-1} \ln \hat{S}_{ij}, \text{ for } j = \lceil p/2 \rceil + 1 \text{ to } p$$

When p is even, the corresponding equations are given as;

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=2}^{m} \ln \hat{S}_{ij} \text{ for } j = 1 \text{ to } \frac{p}{2}$$

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=1}^{m-1} \ln \hat{S}_{ij} \text{ for } j = (\frac{p}{2}+1) \text{ to } p$$

Lemma 1: For the model in Equation (2), using Method I, the estimates of the logs of the seasonal indices add to 0. That is, $\sum_{j=1}^{p} E[\ln \hat{S}_j] = 0$.

Proof: See Appendix A.

Lemma 2: For Method I, the expected value of the log of the estimate of the seasonal index for a particular period is equal to the log of the seasonal index for the period. That is, $E[\ln \hat{S}_j] = \ln S_j$.

Proof: See Appendix B.

Theorem 1: The expected value of the log of a variable following the GBM process for a particular period, obtained from subtracting the corresponding expected log of the seasonal factor from the log of the observation, is an unbiased estimator of the actual log of that variable. That is $E[\ln \hat{Y}_{ij}] = E[\ln Y_{ij}]$.

Proof: From Equation (3), $\ln[\hat{Y}_{ij}] = \ln[X_{ij}] - \ln[\hat{S}_j]$, and taking expectation of both sides,

$$E\left[\ln \hat{Y}_{ij}\right] = E\left[\ln X_{ij}\right] - E\left[\ln \hat{S}_j\right]$$

However, from Lemma 2, we have $E[\ln \hat{S}_j] = \ln S_j$.

The above equation becomes, $E\left[\ln \hat{Y}_{ij}\right] = E\left[\ln X_{ij}\right] - \ln S_j$. Hence from Equation (2) we can see that, $E[\ln \hat{Y}_{ij}] = E[\ln Y_{ij}]$.

3.2 Method II [8]

In the previous method, the moving average was used. Here the simple arithmetic average of the log values across seasons [8] is examined as an alternative method of deseasonalization. Let P_j denote the average value for the j th period, that is, the average of all the period j values over all m seasons.

$$P_j = \frac{1}{m}\sum_{i=1}^{m} \ln\left[X_{ij}\right] = \frac{1}{m}\sum_{i=1}^{m} \ln\left[Y_{ij}\right] + \frac{1}{m}\sum_{i=1}^{m} \ln\left[S_{ij}\right].$$

The overall average for the season is the average of all the periods of the season, $\overline{P} = \frac{1}{p}\sum_{k=1}^{p} P_k$.

Now the seasonal indices for each period can be calculated by the equation $\ln\left[\hat{S}_j\right] = P_j - \overline{P}, \ \forall\, j$.

Using this method, the sum of the estimated logs of the seasonal indices of all the periods of the season is zero. That is, $\sum_{j=1}^{p} E[\ln \hat{S}_j] = 0$. However, the expected value of the ratio of the X variable and the estimated seasonal index for the particular period is not the actual Y variable for the period, as obtained in the previous method. This method can be shown to overestimate the Y_{ij} values by the factor $e^{(\frac{p-(2j-1)}{2})\delta}$ (Note that this factor is greater than 1 for $j < \frac{p+1}{2}$ and less than 1 for $j > \frac{p+1}{2}$; see Figure 2).

We conclude that the method of using moving average (with additive model of log of parameters) is better than the one using simple average. Hence we use Method I to analyze the numerical data.

IV. Data Analysis

The purpose of fitting a model to historical data is to help predict the future, assuming that past and current trends will continue. In trying to fit and forecast demand for services, two difficulties immediately arise. First, the demand will depend on price to varying extents depending on the level of necessity of service and the availability of alternatives for meeting the same need. Secondly, without extensive consumer surveys, the only

way to measure past demand is by actual usage, which was limited by the available supply of the service. As a surrogate for the actual demand data, we collected usage data for publicly available sources in the energy, transportation and telecommunication sectors, the analysis for which is given in the succeeding sections.

4.1 Electric Power Consumption

The data were collected from the U.S. Department of Energy's Office of Scientific and Technical Information, which provides access to energy, science, and technology research and development information [7]. The data represent the total monthly sales by electric utilities to all the sectors (namely, residential, commercial, industrial and others). The monthly consumption (in million kilo-watt-hours) for electric power was recorded for each month for 8 years (from 1993 to 2002). Hence the total of 120 data points were used for the analysis of the electric power consumption.

First, the seasonal variation was removed from the data. For this the two methods described in Section III were tested. The results are shown in Table 1, which gives the value for the seasonal index for each month using each of the methods.

Table 1 here

Figure 1 here

The difference between the two methods of evaluating of seasonal variation is seen in Table 1 and Figure 1. For the detailed difference, Figure 2 compares the values before and after deseasonalization by both methods for a representative year, 1998. The deseasonalized values obtained by Method I are seen to have more of an upward trend over the year.

Figure 2 here

The deseasonalized data obtained from Method I were analyzed to check the normality of the log ratios and also their independence.

Even before the normality test, as a visual check for the independence of the log ratios, we observe a scatter plot of log ratios in Figure 3. As there is no apparent pattern to the $w(k)$ values for the data points, we may tentatively say that the $w(k)$ values are independent, which will be examined analytically in the chi-square test of independence. The plot also indicates the plausibility of a constant mean and variance of the $w(k)$ values.

Figure 3 here

Figure 4 shows the histogram and normal probability plot of the log ratios with fitted mean and variance. Since the Shapiro-Wilk test statistic is 0.9844 and the corresponding p value is 0.768, we fail to reject the null hypothesis that the distribution of the log ratios is normal. Hence we can conclude that the data are consistent with the lognormal aspect of GBM.

Figure 4 here

The remaining key characteristic of the GBM process is independent increments. Figure 5 plots the deseasonalized log ratios for years 1994, 1997, 1999, and 2001. The lack of any visible pattern in values for any given year indicates the independence of the successive ratios.

Figure 5 here

Next the independence of the log ratios is checked using a two-way chi-square test. The $w(k)$ values were divided into 4 categories as shown in Table 2 and the two way table chi-square test resulted in a p-value for the test of 0.319. The null hypothesis that the variables are independent cannot be rejected.

Table 2 here

Hence we conclude that overall the data are consistent with the periodic observations from a GBM process. The mean log ratio was 0.0025 with a standard deviation of 0.02, indicating the mean growth rate of 3% per annum.

The importance of removing the seasonal variation prior to checking for the normality and independence is stressed from the fact that, for the original time series (before the deseasonalization) the normality test for the log ratios failed (with p-value 0.0004, rejecting the null hypothesis that the distribution for log ratios is normal); also these log ratios were not found to be independent. In fact, the two-way chi-square test on these log ratios gave a p-value of 0.001, indicating that we reject the null hypothesis of independence of the variables. The same fact could also be observed from the scatter plot of log ratios with respect to the prior values (see Figure 6). If the log ratios had been independent, the points of the scatter plot would not have had any trend.

Figure 6 here

4.2 Airline Passenger Enplanement

We collected the historical monthly data on U.S. Revenue Passenger Enplanements for the years 1981 through 2001 from the U.S. Aeronautical Board [26]. Revenue Passenger Enplanement can be defined as the number of paying passengers boarding a flight, including origination, stopovers and connections. It should be noted that each connecting flight between origination point and destination counts as one enplanement.

While analyzing the passenger data, a seasonal trend was observed for which the moving average (Method I) was applied to deseasonalize the log ratios. The final seasonal indices were as given in Table 3.

Table 3 here

The variation in the data values with respect to time is given in Figure 7. It can be seen that as the time increases, the amount of seasonal variation increases (observing the original data); motivating the use of the multiplicative model described in Section 3.1.

Figure 7 here

Figure 8 plots the corresponding log ratios over time. From the plot, it can be seen that there is no visible pattern in the values of log ratios, which indicates their distribution is stationary, and suggests serial independence.

Figure 8 here

The histogram and normal probability plot for the normality test for the passenger data are given in Figure 9. As the p-value of the Shapiro-Wilk test is 0.4416 (greater than 0.05), we cannot reject the hypothesis that the log ratios are normally distributed.

Figure 9 here

Again, as with the electric utility data, the random nature of the deseasonalized $w(k)$ values can be visually inspected by the graphs of $w(k)$ values given in Figure 10. Here, the changes in $w(k)$ values appear to be independent over time, as seen from the randomness of the $w(k)$ values for various years. Hence it can be tentatively concluded that the $w(k)$ values are independent of each other.

Figure 10 here

To more rigorously test independence by the chi-square test, four intervals of $w(k)$ values were selected as shown in Table 4.

Table 4 here

The p-value for the test was found to be 0.058, so we cannot reject the null hypothesis that the *w(k)* values are independent at a 5% significance level.

Once again, as done in the electric demand case, the importance of removing the seasonality factors before checking normality and independence of log ratios is confirmed by performing similar tests with the original log ratios (obtained from the time series without deseasonalization). The normality of the log ratios could not be confirmed (the p-value of the normality test is 0.0001); also the chi-square test gives a p-value, which is very close to zero, forcing the rejection of null hypothesis of the independence test. Hence prior to the deseasonalization, the log ratios are not independent. The same fact could be observed by inspecting the scatter plot of these log ratios with respect to the prior values (Figure 11), which indicates clear trend in the values.

Figure 11 here

Thus, we can conclude that the lognormal ratios after deseasonalization are independent; however, a higher significance level could lead to the opposite conclusion. Hence, the independence test is not as convincing as for the electric power data. The mean log ratio was found to be 0.00271 with a standard deviation of 0.029, and hence the average growth rate was calculated to be 3.3% per annum.

4.3 Cell Phone Revenue

Usage of mobile phone service might be measured by minutes of usage, total connections made, or even the number of handsets sold. Because of the lack of available data on these quantities, the revenue collected from the cellular phone subscribers was analyzed for the period of January 1985 to June 2002, with data collected every 6 months [5].

First of all, the plot given in Figure 12 of log ratios over time was observed. It is seen that there is a decreasing trend in both the mean and the variance of log ratios. Hence visual inspection reveals that the *w(k)* variable may be neither stationary nor independent. Note that, since revenue is the product of sales volume and price, the downward trend could be attributed to price drops rather than flattening growth in demand.

Figure 12 here

The normality test, which includes the histogram of the log ratios and the normal probability plot, is given in Figure 13. From the Shapiro-Wilk test, the p-value is 0.0003, proving that the log ratios are not normally distributed.

Figure 13 here

The result could be influenced by the fact that the number of data points available was only 35. However, the Chi-square test did show independence of the $w(k)$ values. The p-value for the independence test is 0.3735. Hence the null hypothesis that the log ratios are independent cannot be rejected. For this independence test the intervals of $w(k)$ values used are given in Table 5.

Table 5 here

4.4 Internet Hosts

Internet growth can be measured by changes in either the number of users or number of hosts connected to the network. A host used to be a single machine on the net. However, the definition of a host has changed in recent years due to virtual hosting, where a single machine acts like multiple systems (and has multiple domain names and IP addresses) [11]. Typically, multiple users are connected to a host and the hosts are connected to the network. Since there is no central mechanism for tracking the number of users connected to the network [19], we use number of hosts as a measure of Internet size. In an attempt to gauge the growth of the Internet over the years, The Internet Software Consortium [11] conducted a survey called 'The Domain Survey' and measured the number of hosts. This survey was used in conjunction with the data in [19] to obtain a time series of the number of Internet hosts from 1982 to 2003 with data points recorded every six months. As before, $w(k)$ values for the data are calculated and tested for normality and independence.

Figure 14 indicates the values of log ratios over time. It is seen again that the values do not appear to be random. There is visible downward trend in the values of $w(k)$, indicating that the values may not be stationary. One can also observe possible cyclical behavior.

Figure 14 here

The plots for the test of normality are given in Figure 15. The p-value for the Shapiro-Wilk test of normality is less than 0.001; hence the null hypothesis that the log ratios are normal is rejected.

Figure 15 here

To test the independence of $w(k)$ values, the Chi-square test cannot be used as before, because the number of data points is too small to create cells such that each holds a positive number of observed values as required by the chi-square test. Hence the $w(k)$ scatter plot is analyzed visually to determine the independence of $w(k)$.

From the plot in Figure 16, it can be seen that the $w(k)$ values are not random, but rather large (small) log ratios tend to be immediately followed by other large (small) values. Hence we can say that the $w(k)$ values are not independent of each other.

Figure 16 here

4.5 Summary of Results

The results of the data analysis for different industries are summarized in Table 6.

Table 6 here

Hence it can be seen that data related to service consumption from different sectors of industry may or may not meet the criteria for the GBM process. Among the services examined, the ones that fail one test or another are in newer industries that perhaps can still be considered emergent. Data on the usage of these services are also less direct and more difficult to obtain. The older and more established services of electric power and airline travel exhibit a better fit to the GBM assumption after deseasonalization. Having ascertained the model's fit to the deseasonalized data, a forecast of future demand can be obtained from the GBM model with the fitted parameters by re-inserting the seasonal factors. How the seasonal patterns would affect decision-making depends on the application, for example, capacity decisions typically consider the peak demand in a season.

We caution that, even when a model appears to closely fit historical data, extrapolation into the future does not carry any guarantee of accuracy. In 1995, logistic growth models showed a very good fit to historical data on the number of cell phone subscribers [27]. Extrapolation suggested that the number of U.S. subscribers would level out close to 80 million early in the 21st century. As of December 2004, however, the Cellular

Telecommunications and Internet Association reported over 173 million current U.S. wireless subscribers [5]. The fit of a model to historical data is a necessary but not sufficient condition for the credibility of its forecasts.

V. Conclusion

From the theory of the Brownian motion discussed in the paper and the subsequent analysis, it can be concluded that the structure for the analysis to check whether a particular time series data follows a Geometric Brownian motion process or not can be applied to varied data types. The result may be different for different data types; for some of the data sets, the GBM process may be appropriate, based on the criteria of normality and independence (for example, electric utility data and passenger data); however for some of the data sets, the assumption of GBM process distribution may not be appropriate (example, cell-phone revenue data and Internet host data). Hence in any given model, caution should be taken before assuming that the particular data set follows the GBM process. It was observed during the analysis of Cellular phone data and the Internet host data that the number of data points may affect the analysis results. Hence attempts need to be made to collect more data points for the given example type.

For cell phone revenue data and Internet hosts' data, it was observed that the log ratios decrease over time. It could be possible that the drift for those time series is dependent on time and the level of the time series. Hence the criteria for the GBM (with assumption of constant drift and volatility) were not being followed in these cases. For these data not following the GBM process, the data can be analyzed for other stochastic diffusion processes [6]. Also to incorporate the dynamic nature of drift (and possibly volatility) parameter, the Ito process for the stock prices can be used. More generalized models can also be studied. In [10], authors discuss some of the extended one-state-variable interest-rate models that involve time dependent parameters. The data for the cell phone revenue and Internet hosts might be analyzed using models similar to the ones given in that paper.

References

1. Anderson, D. R.; Sweeney, D. J.; Williams, T. A. (1994). *An Introduction to Management Science*, 7th ed. West Publishing Company, St. Paul, Minn.
2. Benninga, S.; Tolkowsky, E. (2002). "Real Options- An Introduction and An Application to R&D Evaluation". *The Engineering Economist*, 47(2). 151-168.
3. Blair, M. M. (1952). *Elementary Statistics, with General Applications*, Henry Holt and Company, New York.
4. Brockwell, P. J.; Davis, R. A. (2002). *Introduction to Time Series and Forecasting*, Springer Texts in Statistics, New York.
5. Cellular Telecommunications and Internet Association, *Wireless Industry Survey* (Viewed December 2004) http://www.ctia.org
6. Dixit, A. K.; Pindyck, R. S. (1993). *Investment under Uncertainty*, Princeton University Press. Princeton, N.J.
7. Energy Information Administration. *Electric Power Monthly*. (Viewed December 2003) http://www.osti.gov/servlets/purl/212513-0AAD20/webviewable/212513.pdf (for data from 1993 to 1995)
   http://www.osti.gov/servlets/purl/584882-h5Z86P/webviewable/584882.pdf (for data from 1995 to 1997)
   http://www.eia.doe.gov/cneaf/electricity/epm/epmt44p1.html (for data from 1998 to 2000)
8. Hillier, F. S.; Lieberman, G. J. (2001), *Introduction to Operations Research*, 7th ed. McGraw Hill. Boston.
9. Hull, J. C. (2000). *Options, Futures, and Other Derivatives*, 4th ed. Prentice Hall, NJ.
10. Hull, J. C.; White A. (1990). "Pricing Interest-Rate-Derivatives Securities". *The Review of Financial Studies*. 3(4). 573-592.
11. Internet Software Consortium, *Internet Domain Survey* (Viewed December 2003) http://www.isc.org/ds/host-count-history.html
12. JMP (2003). Statistical Discovery Software; http://www.jmp.com/index.html

13. Karsak, E. E.; Ozogul, O. C. (2002). "An Options Approach to Valuing Expansion Flexibility In Flexible Manufacturing Systems Investment". *The Engineering Economist*, 47(2). 169-193.

14. Lieberman, B. M. (1989). "Capacity Utilization: Theoretical Models and Empirical Tests." *European Journal of Operational Research* 40, 155-168.

15. Luenberger, D. (1995). *Investment Science*, Oxford University Press, New York.

16. Makridakis S.; Wheelwright S. C.; Hyndman R. J. (1998). *Forecasting: Methods and Applications*, 3rd ed., John Wiley and Sons Inc. New York.

17. Nembhard, H. B.; Shi, L.; Aktan, M. (2002). "A Real Options design for Quality Control Charts." *The Engineering Economist*, 47(1). 28-50.

18. Nembhard, H. B.; Shi L.; Aktan M. (2003). "A Real Options design for Product Outsourcing." *The Engineering Economist*, 48(3). 199-217.

19. Rai, A.; Ravichandran, T.; Samaddar, S. (1998). "How to anticipate the Internet's global diffusion." *Communications to the ACM*, 41, 97-106

20. Ross, S. (1999). *An Introduction to Mathematical Finance*, Cambridge University Press, Cambridge, U.K., New York.

21. Ross, S. (2000). *Introduction to Probability Models*, 7th ed., Harcourt Academic press, New York.

22. Ryan, S. M. (2004). "Capacity Expansion for Random Exponential Demand Growth with Lead Times." *Management Science*, 50(6). 740-748

23. Samuelson, P. A. (1965). "Proof that Properly Anticipated Prices Fluctuate Randomly," *Industrial Management Review*, 7, 41-49.

24. Shapiro S. S.; Wilk M. B., (1965). "An Analysis of Variance Test for Normality." *Biometrika*, 52, 3/4, 591-611.

25. Thorsen, B. J. (1998). "Afforestation as a Real Option: Some Policy Implications." *Forest Science*, 45(2). 171-178.

26. U.S. Aeronautical Board, *Origin and Destination Survey of Airline Passenger Traffic* (Viewed December 2003) http://www.bts.gov/oai/indicators/airtraffic/annual/1981-2001.html

27. Wang, M.; Kettinger, W. J. (1995). "Projecting the Growth of Cellular Communication." *Communications of the ACM*, 38, No. 10. 119-122

28. Whitt, W. (1981). "The Stationary Distribution of a Stochastic Clearing Process." *Operations Research* 29(2), 294-308.

Biographical Sketches

Rahul R. Marathe is a doctoral student in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University, Ames, IA. His research interests are in the area of stochastic processes and their applications. His email address is rahulm@iastate.edu.

Sarah M. Ryan is an associate professor of Industrial & Manufacturing Systems Engineering at Iowa State University. She teaches courses in optimization, stochastic modeling and engineering economic analysis. Her research uses stochastic models to study long-term investment decision problems as well as resource allocation problems in manufacturing. She is the recipient of a Faculty Early Career Development (CAREER) Award from the National Science Foundation. She serves on the editorial board of *IIE Transactions* and as an Area Editor for *The Engineering Economist.*

Appendix A: Proof of Lemma 1

We consider the case where the number of seasons m is even and number of periods p is odd.

$$\sum_{j=1}^{p} E\left[\ln \hat{S}_j\right] = E\left[\ln \hat{S}_{\lceil p/2 \rceil}\right] + \sum_{j=1}^{\lfloor p/2 \rfloor} E\left[\ln \hat{S}_j\right] + \sum_{j=1+\lceil p/2 \rceil}^{p} E\left[\ln \hat{S}_j\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} \ln \hat{S}_{i\lceil p/2 \rceil}\right] + \sum_{j=1}^{\lfloor p/2 \rfloor} E\left[\frac{1}{m-1}\sum_{i=2}^{m} \ln \hat{S}_{ij}\right] + \sum_{j=1+\lceil p/2 \rceil}^{p} E\left[\frac{1}{m-1}\sum_{i=1}^{m-1} \ln \hat{S}_{ij}\right]$$

$$= \sum_{i=1}^{m} E\left[\frac{1}{m} \ln \hat{S}_{i\lceil p/2 \rceil}\right] + \sum_{j=1}^{\lfloor p/2 \rfloor}\sum_{i=2}^{m} E\left[\frac{1}{m-1} \ln \hat{S}_{ij}\right] + \sum_{j=1+\lceil p/2 \rceil}^{p}\sum_{i=1}^{m-1} E\left[\frac{1}{m-1} \ln \hat{S}_{ij}\right]$$

$$\sum_{j=1}^{p} E\left[\ln \hat{S}_j\right] = \sum_{i=1}^{m} E\left[\frac{1}{m}\left(\ln X_{i\lceil p/2 \rceil} - P_{i\lceil p/2 \rceil}\right)\right] + \sum_{j=1}^{\lfloor p/2 \rfloor}\sum_{i=2}^{m} E\left[\frac{1}{m-1}\left(\ln X_{ij} - P_{ij}\right)\right] + \sum_{j=1+\lceil p/2 \rceil}^{p}\sum_{i=1}^{m-1} E\left[\frac{1}{m-1}\left(\ln X_{ij} - P_{ij}\right)\right],$$

where P_{ij} is the arithmetic moving average, as defined earlier.

Substituting values of P_{ij}, we have that

$$\sum_{j=1}^{p} E\left[\ln \hat{S}_j\right] = \sum_{j=1}^{\lfloor p/2 \rfloor}\left(\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right)E\left[\ln X_{1j}\right] + E\left[\ln X_{\lceil p/2 \rceil} + \ln X_{m\lceil p/2 \rceil}\right]\left(\frac{p-1}{pm} - \frac{\lfloor p/2 \rfloor}{p(m-1)}\right) + \sum_{j=\lceil p/2 \rceil+1}^{p}\left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right)E\left[\ln X_{1j}\right]$$

$$+ \sum_{\substack{j=1 \\ j\neq\lceil p/2 \rceil}}^{p}\sum_{i=2}^{m-1} E\left[\ln X_{ij}\right]\left(\frac{1}{p(m-1)} - \frac{1}{mp}\right) + \sum_{i=2}^{m-1} E\left[\ln X_{i\lceil p/2 \rceil}\right]\left(\frac{p-1}{mp} - \frac{p-1}{p(m-1)}\right) + \sum_{j=\lceil p/2 \rceil}^{p}\left(\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right)E\left[\ln X_{mj}\right]$$

$$+ \sum_{j=1}^{\lfloor p/2 \rfloor}\left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right)E\left[\ln X_{mj}\right]$$

Now we have from Equation (2), $\ln[X_{ij}] = \ln[Y_{ij}] + \ln[S_j]$. Substituting this for each of the X_{ij}, we cancel out the seasonal indices S_j from the above equation. To evaluate the Y_{ij} terms, we use Equation (4) and write all the Y_{ij} in terms of Y_{11} and solve the equation. We get

$$\sum_{j=1}^{p} E[\ln \hat{S}_j] = 0$$

For the case where the number of season m is odd, and the periods p is also odd, the condition can be found out as:

$$\sum_{j=1}^{p} E\left[\ln \hat{S}_j\right] = \sum_{j=1}^{\lfloor p/2 \rfloor}\left(\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right)E\left[\ln X_{1j}\right] + E\left[\ln X_{\lceil p/2 \rceil} + \ln X_{m\lceil p/2 \rceil}\right]\left(\frac{p-1}{pm} - \frac{\lfloor p/2 \rfloor}{p(m-1)}\right)$$

$$+ \sum_{j=\lceil p/2 \rceil+1}^{p}\left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right)E\left[\ln X_{1j}\right]$$

$$+ \sum_{j=1}^{p}\sum_{i=2}^{m-1} E\left[\ln X_{ij}\right]\left(\frac{1}{p(m-1)} - \frac{1}{mp}\right) + \sum_{j=\lceil p/2 \rceil}^{p}\left(\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right)E\left[\ln X_{mj}\right]$$

$$+ \sum_{j=1}^{\lfloor p/2 \rfloor}\left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right)E\left[\ln X_{mj}\right]$$

The case when the number of periods p is even is also similar to the one formulated above, using Equation (5), which had to be centered because of the even number of periods [5].

Hence, the sum of estimated log of seasonal indices for a season is:

$$E[\ln \hat{S}_j] = \sum_{j=1}^{p}\left(\frac{u_j}{m-1} - \frac{2j-1}{2(m-1)p}\right)E\left[\ln X_{1j}\right] + \sum_{j=1}^{m}\left(\frac{w_j}{m-1} - \frac{1}{2(m-1)p} - \frac{p-j}{(m-1)p}\right)E\left[\ln X_{mj}\right],$$

where $u_j = \begin{cases} 0 & for\ j \le p/2 \\ 1 & for\ j > p/2 \end{cases}$

And $w_j = \begin{cases} 1 & for\ j \le p/2 \\ 0 & for\ j > p/2 \end{cases}$

which, when we use Equation (2) for $\ln[X_{ij}]$ and subsequently Equation (4) for $\ln[Y_{ij}]$ as before, comes to zero.

Appendix B: Proof of Lemma 2

To show $E\left[\ln \hat{S}_j\right] = \ln S_j$, for any j

We have,

$$E\left[\ln \hat{S}_{ij}\right] = E\left[\ln X_{ij} - P_{ij}\right]$$

$$= E\left[\ln X_{ij} - \frac{1}{p}\sum_{k=t-\lfloor p/2 \rfloor}^{t+\lfloor p/2 \rfloor} \ln X_k\right], \text{ where } t = (i-1)p + j, \text{ therefore}$$

$$E\left[\ln \hat{S}_{ij}\right] = E\left[\frac{p-1}{p}\ln X_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln X_k\right]$$

Since $\ln X_{ij} = \ln S_{ij} + \ln Y_{ij}$, we have

$$E\left[\ln \hat{S}_{ij}\right] = E\left[\frac{p-1}{p}\ln S_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln S_k\right] + E\left[\frac{p-1}{p}\ln Y_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln Y_k\right],$$

We know that $\sum_{j=1}^{p} \ln S_{ij} = 0$, hence $\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln S_k = -\ln S_{ij}$

$$E\left[\ln \hat{S}_{ij}\right] = E\left[\frac{p-1}{p}\ln S_{ij} - \frac{1}{p}\left(-\ln S_{ij}\right)\right] + E\left[\frac{p-1}{p}\ln Y_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln Y_k\right]$$

$$= E\left[\frac{p-1}{p}\ln S_{ij} - \frac{1}{p}\left(-\ln S_{ij}\right)\right] + \frac{p-1}{p}E\left[\ln Y_{ij}\right] - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} E\left[\ln Y_k\right]$$

$$= \ln S_{ij} + \frac{p-1}{p}\left\{E\left[\ln Y_{11}\right] + \delta(t-1)\right\} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} E\left[\ln Y_{11}\right] + \delta(k-1), \text{where } t = (i-1)\,p + j$$

which gives,

$$E\left[\ln \hat{S}_{ij}\right] = \ln S_{ij} + 0 = \ln S_{ij} \ .$$

From the above equation:

$$\frac{1}{m}\sum_{i=1}^{m} E\left[\ln \hat{S}_{ij}\right] = \frac{1}{m}\sum_{i=1}^{m} \ln S_{ij}$$

From Equation (6),

$$E\left[\ln \hat{S}_j\right] = \frac{1}{m}\sum_{i=1}^{m} \ln S_{ij} = \frac{1}{m}\left[m \cdot (\ln S_j)\right]$$

because we know that, $\ln S_{ij} = \ln S_j$, $\forall j$

Hence $E\left[\ln \hat{S}_j\right] = \ln S_j$.

In the case where the number of periods p is even, the calculations are similar except for the fact that again Equation (4) for the centered moving average is used instead of the simple moving average. Hence, the values obtained after substituting respective values in the equation $E\left[\ln \hat{S}_{ij}\right] = E\left[\ln X_{ij} - P_{ij}\right]$ change accordingly, however the concept is similar; and it gives a similar result.

List of Figures

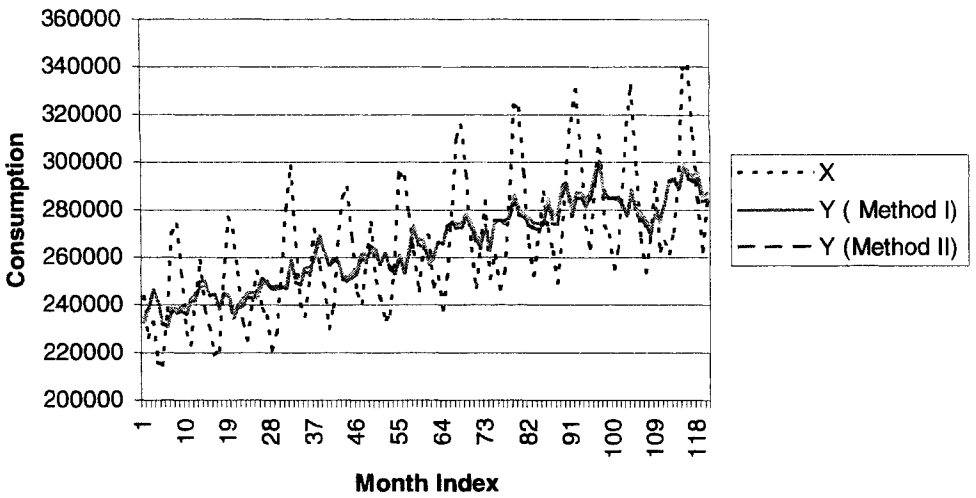

Figure 1. Original and deseasonalized demand data obtained from two different methods (* units: million kWh)

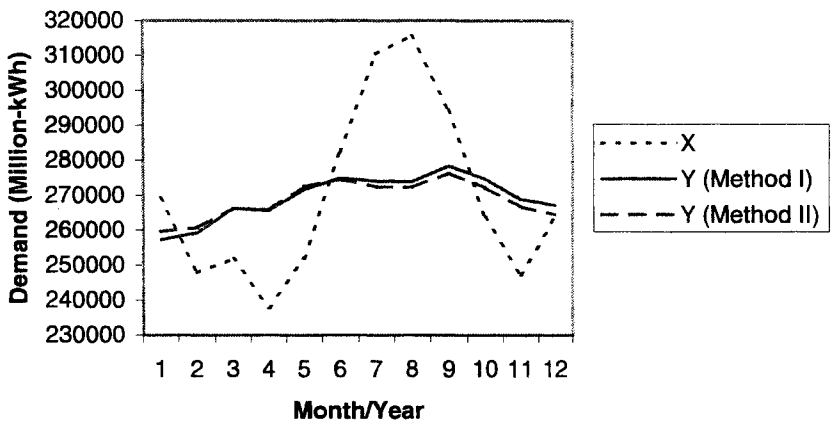

Figure 2. Original and deseasonalized data (obtained from each of the methods) for each month of the year

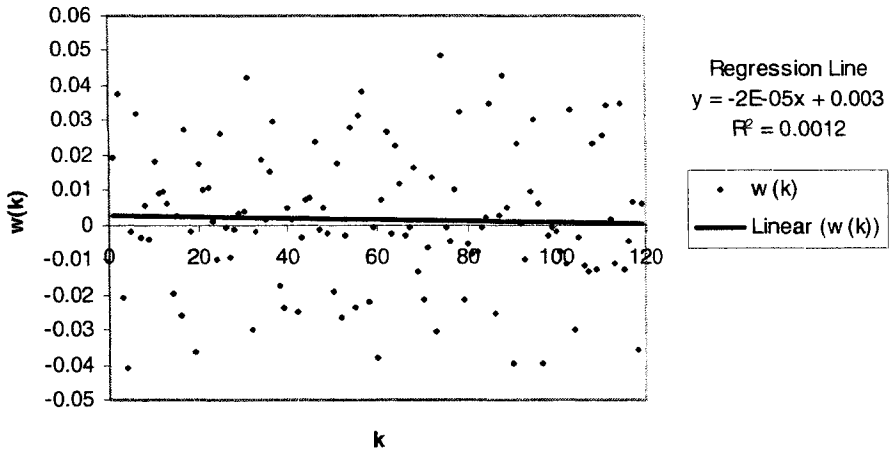

Figure 3. Log ratios ($w(k)$) for all 120 data points indicating the randomness of the $w(k)$ values

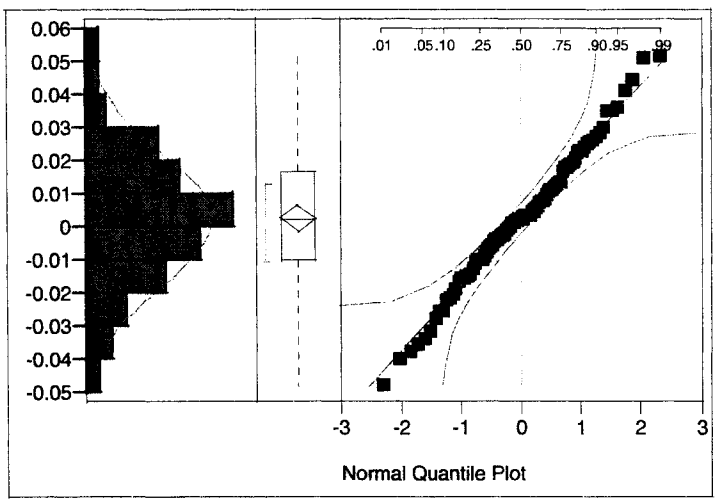

Figure 4. Distributions for the $w(k)$ values obtained from the moving average method.

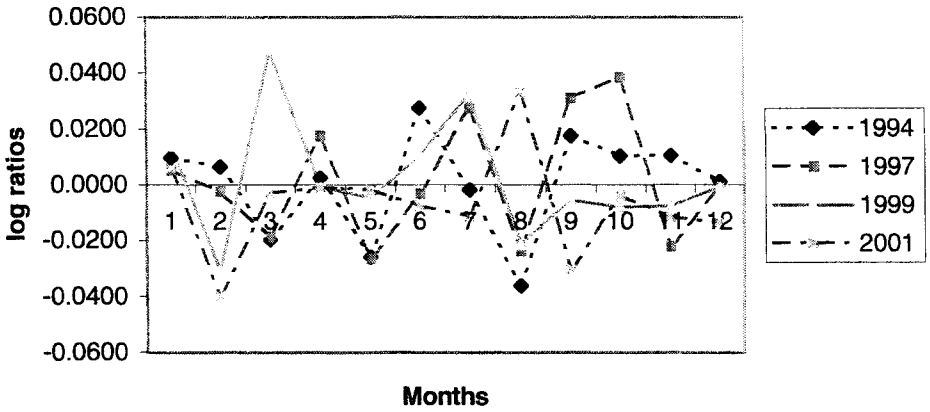

Figure 5. Log ratios for each month of year plotted for 4 representative years indicating randomness

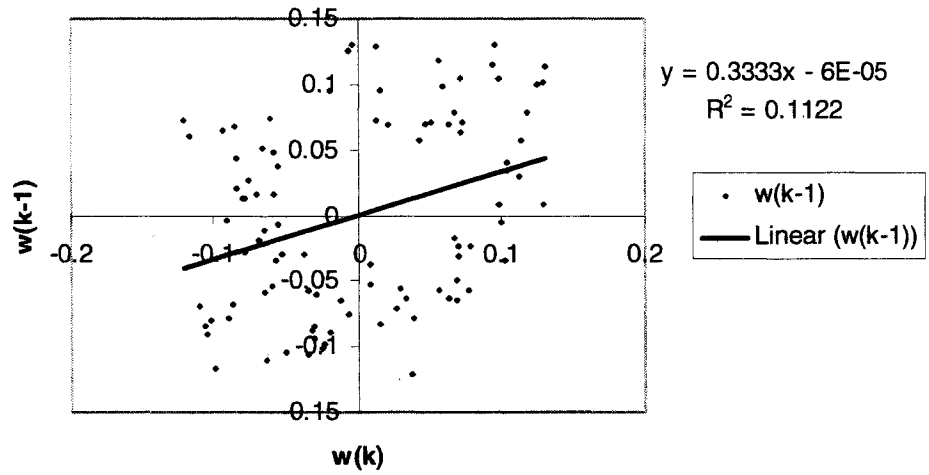

Figure 6. Scatter plot of *w(k)* values for electric demand before deseasonalization

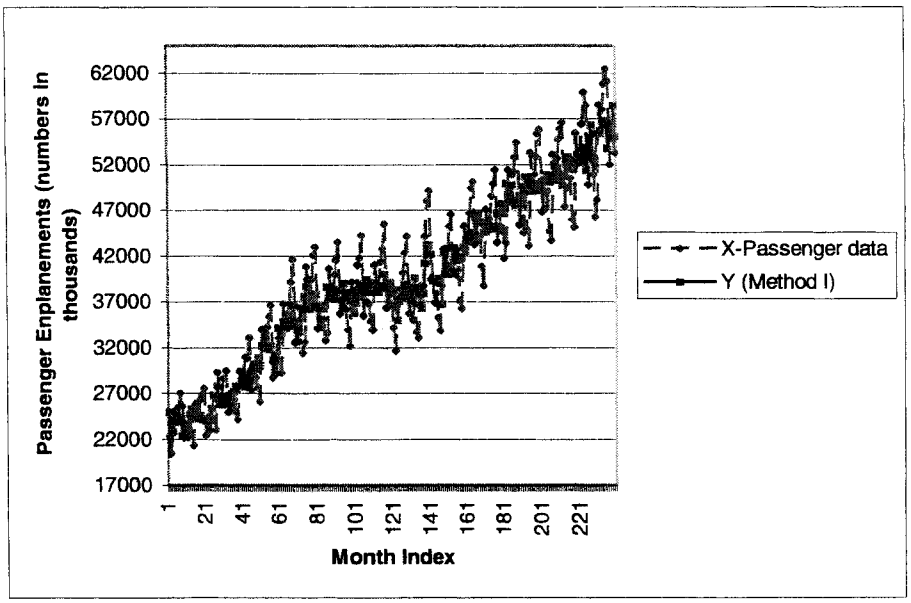

Figure 7. Original and deseasonalized enplanement data obtained from deseasonalization Method I of Section 3.1

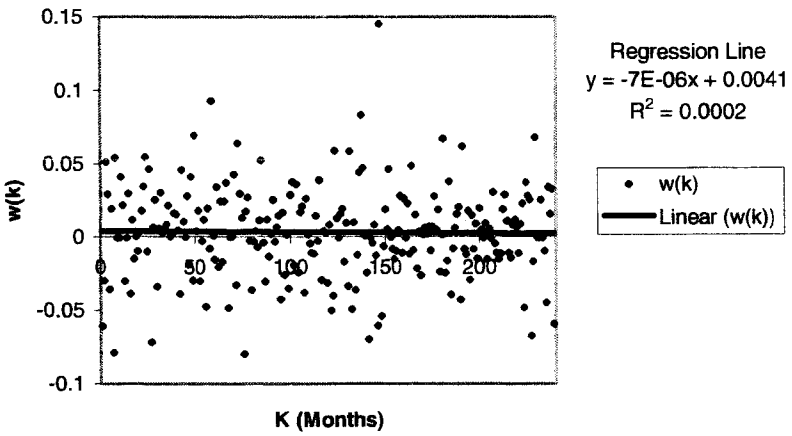

Figure 8. *w(k)* values for all 238 data points indicating the randomness of the *w(k)* values

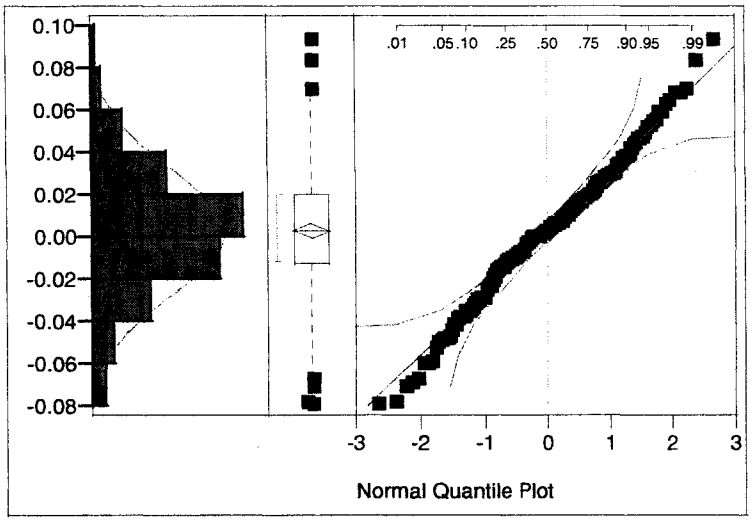

Figure 9. Distributions of *w(k)* for airline passenger enplanement data

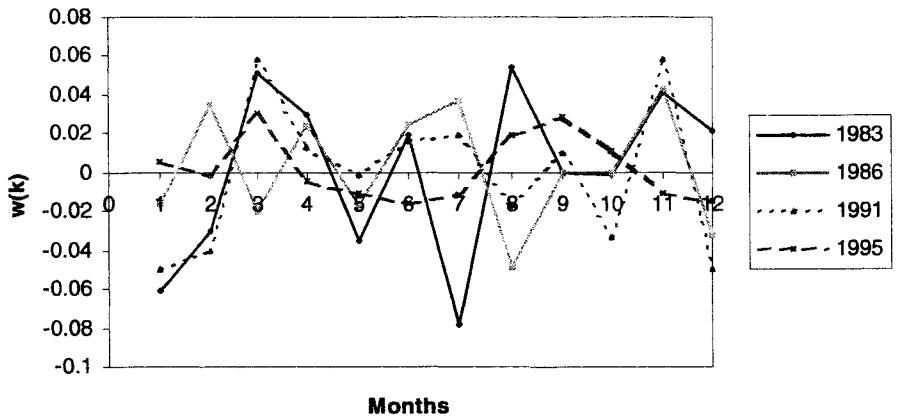

Figure 10. *w(k)* values for each month of year plotted for 4 years indicating randomness

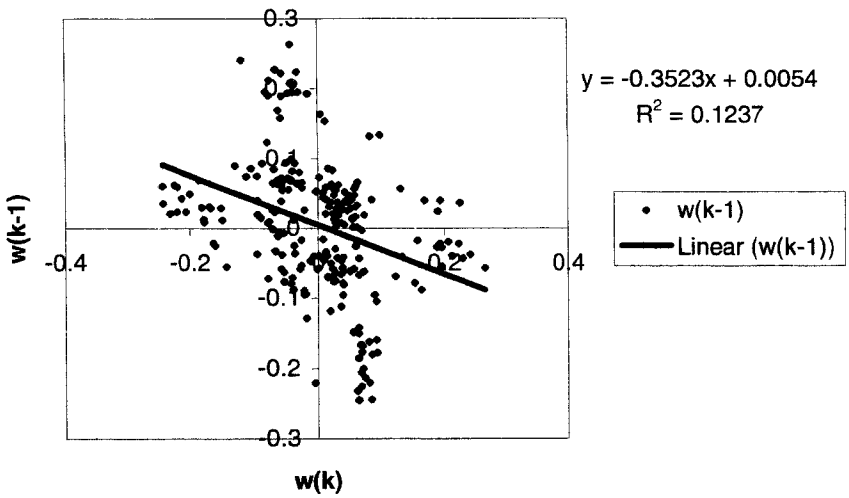

Figure 11. Scatter plot of *w(k)* values for airline passenger data before deseasonalization

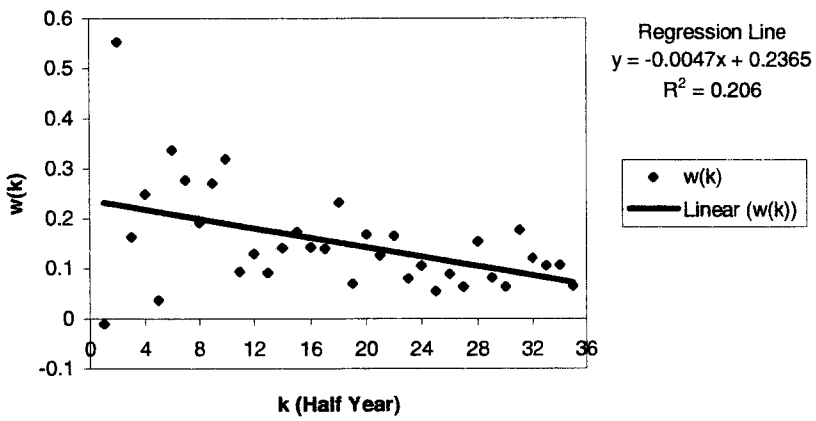

Figure 12. *w(k)* values for data points of cell phone data indicating the randomness of the *w(k)* values

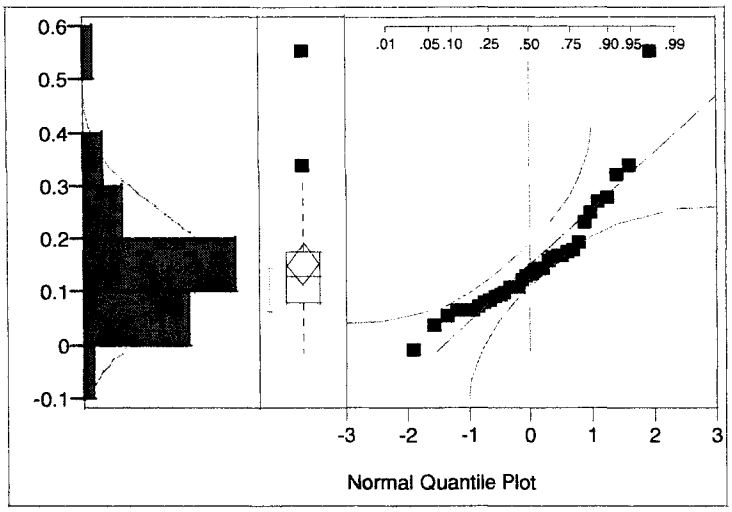

Figure 13. Distributions of *w(k)* for cell phone revenue data

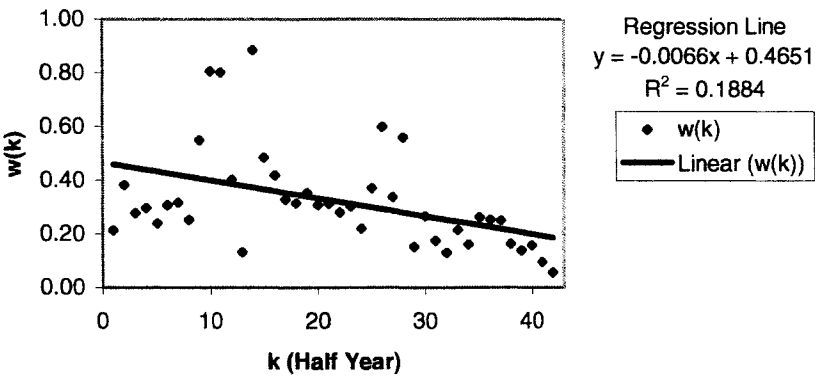

Figure 14. *w(k)* values for data points for internet host data

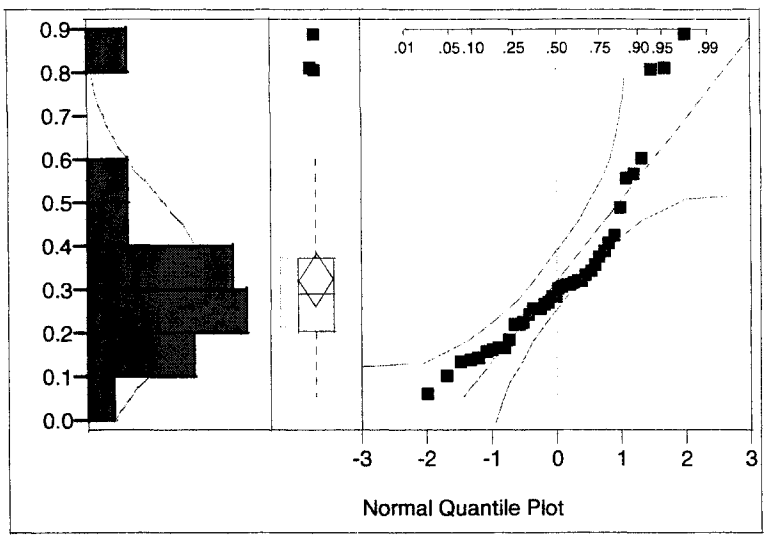

Figure 15. Distributions of *w(k)* for Internet host data

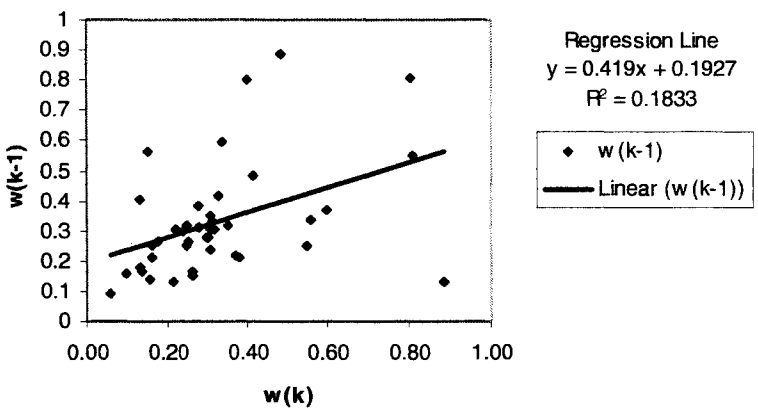

Figure16. *w(k)* scatter plot for Internet host data

List of Tables

Table 1. Seasonal Indices for electric power consumption data from two different methods

Month	Method I	Method II
January	1.0469	1.0372
February	0.9560	0.9507
March	0.9459	0.9458
April	0.8957	0.8952
May	0.9274	0.9244
June	1.0275	1.0288
July	1.1328	1.1396
August	1.1524	1.1594
September	1.0576	1.0649
October	0.9597	0.9686
November	0.9202	0.9269
December	0.9872	0.9971
Sum of Log	-0.00232	0

Table 2. Categories of $w(k)$ values for the independence test

Categories	$w(k)$ ranges
1	From -0.05 to -0.02
2	From -0.02 to 0
3	From 0 to 0.02
4	From 0 to 0.05

Table 3. Seasonal indices for airline passenger data using moving average method

Month	Sea. Index
January	0.8928
February	0.8686
March	1.0545
April	1.0073
May	1.0203
June	1.0704
July	1.1095
August	1.1361
September	0.9335
October	0.9962
November	0.9368
December	0.9664
Sum of Log	-0.00381

Table 4. Categories of $w(k)$ used for airline passenger data

Categories	$w(k)$ ranges
1	$w(k) > 0.04$
2	$w(k)$ from 0.04 to 0.01
3	$w(k)$ from 0.01 to -0.03
4	$w(k)$ from -0.08 to -0.03

Table 5. Categories of $w(k)$ values used for cell-phone revenue data

Categories	$w(k)$ range
1	from -0.02 to 0.085
2	from 0.085 to 0.16
3	from 0.16 to 0.25
4	from 0.25 to 0.6

Table 6. Summary of results of the various data sets

Data Set	Time Series	Normality	Independence	Remarks
Electric Utility	Electric Consumption data	Yes $p = 0.768$	Yes $p = 0.319$	Log ratios stationary and independent
Airline	Revenue Passenger Enplanement	Yes $p = 0.4416$	Yes $p = 0.058$	Log ratios stationary and independent
Cell phone	Revenue from Consumer Subscription	No $p = 0.0003$	Yes $p = 0.3735$	Independence test not credible because of fewer data points (Downward trend in log ratios over time)
Internet Industry	Number of Internet Hosts	No $p < 0.001$	No (No chi-square, just scatter plot)	Few Data Points, hence Chi-square independence test cannot be carried out (Downward trend in log ratios over time)

Appendix II: Optimal Solution to a Capacity Expansion Problem

Rahul R. Marathe and Sarah M. Ryan; Department of Industrial & Manufacturing Systems Engineering; Iowa State University; Ames, IA 50011-2164, USA.

Abstract

For a service provider, stochastic demand growth along with expansion lead times and economies of scale may encourage delaying the start of expansion until after some shortages have been accumulated. Assuming demand follows a geometric Brownian motion, we define the service level in terms of the proportion of demand satisfied, which is then analytically evaluated using financial option pricing theory. Under a stationary expansion policy, an infinite time horizon discounted expansion cost is minimized under the service level constraint, where the expansion timing and size parameters are the decision variables. With the current formulation, the problem seems to be unbounded.

Keywords

Capacity expansion, service level, barrier option pricing, cutting plane algorithm

1. Introduction

Capacity expansion problems arise in numerous applications varying from communications networks to manufacturing facilities. The problem is to find an optimal policy of expansion given a particular forecasted demand pattern, assuming that the costs and lead times of expansion are known.

We consider a service provider having certain facilities with installed capacity to provide certain services. We consider a single location and single resource assuming that the demand for that resource follows a geometric Brownian motion (GBM) process. The capacity added does not deteriorate; that is, once the capacity is installed, we assume that it is available forever. Expansion costs exhibit economies of scale and there is a deterministic expansion lead time from the time the capacity expansion decision is made to the time when the added capacity is actually available to satisfy the demand.

Modeling demand as a GBM process may be justified when empirical data show that demand growth in a period is on average a constant percentage of demand at the beginning of the period, and periods of higher or lower than average demand occur at random. Marathe and Ryan [1] verified empirically that the historical usage of electric power in the US as well as the number of passenger enplanements in the airline industry each followed a GBM process.

The capacity expansion literature is richly stocked. Manne [2] considered a random-walk pattern demand and proposed the optimal size of the capacity expansion when there were economies of scale available. Whitt [3] considered the utilization aspect of the capacity expansion and found the stationary distribution for the capacity utilization under a simple policy that we adapt in this paper. Chaouch and Buzacott [4] considered the demand with plateaus and formulated the capacity expansion for two cases, viz., when the expansion starts with some initial shortages and when it starts before the demand reaches the current capacity. Our model is similar to this case, with our demand being GBM process driven. Bean et al. [5] considered demand to be following either a transformed Brownian motion process or a semi-Markovian birth and death process. They showed that the problem can be transformed into an equivalent deterministic problem and that the effect of the probabilistic nature of demand is to reduce the interest rate. This result was used by Ryan [6] wherein the effect of a fixed lead time was also considered. Financial option pricing theory was used to develop a stationary expansion policy so that the specified service level is met when the expansion started before the demand reaches the current capacity position. Our model is further extension of Ryan [6] in the sense that we consider the case where the expansion starts when the demand has already crossed the current capacity position.

Our parameter definitions are similar to Ryan [6]; and most of the details may be found in Marathe and Ryan [7, 8]. We briefly summarize the model and definitions in section 2; the service level and the expansion cost analyses are carried out in section 3. We present a numerical example in section 4 and conclude the paper with section 5.

2. Model

Demand for the service is given by the GBM process $P(t) = P(0)e^{B(t)}$, where $B(t)$ is a Brownian motion with drift μ and variance σ^2. Define $\gamma \equiv \mu + \frac{\sigma^2}{2}$ as the mean (exponential) growth rate of the demand.

We assume that capacity additions occur at discrete time points and that a fixed lead time of L time units is required to install new capacity. The problem is to choose a sequence $\{(T_n, X_n), n \geq 1\}$, where T_n, the time when the n^{th} capacity expansion starts, is a stopping time with respect to the Brownian motion $B(t)$ and X_n is the n^{th} increase in capacity. For any realization ω of the Brownian motion $B(t)$, let $t_n \equiv T_n(\omega)$. Let K_n be the installed capacity after n additions are completed, where the initial capacity is K_0. The capacity position includes capacity on order (being constructed or installed) in addition to the installed capacity.

We describe the model by quoting directly from Marathe and Ryan [7, 8]. We follow the Whitt-Luss policy from Whitt [3], where each new expansion starts when demand reaches some fixed proportion (say, 'p') of current capacity position, and after its addition at the end of the lead-time, the new capacity is a constant proportion of its previous value. That is $K_n = vK_{n-1}$, where $v > 1$. For the case of $p < 1$ Ryan [6] used financial option pricing theory to find optimal stationary expansion policy (that is, the values of parameters p and v). In this paper, we consider the case where $p \geq 1$.

Figure 1 illustrates the policy and potential shortages seen at the realized time t_n, when demand first equals pK_{n-1}. The n^{th} capacity expansion has just started. With this expansion, the total installed capacity will reach level K_n after the lead time L. As stated earlier, we model the situation wherein the service provider waits until certain amounts of initial capacity shortages are accumulated before starting the next expansion project. Hence, since the new capacity position is K_n, the next expansion would start at the time when the demand $P(t)$ first reaches the position pK_n. Since the demand process is stochastic, this time for starting the next expansion (T_{n+1}) is a random variable. The goal, then, is to find the optimal initial shortage that will trigger the start of capacity expansion (the parameter p), and the optimal size of each expansion (the parameter v). Figure 1 shows a non-overlapping expansion cycle where the capacity being built is already available before we begin the next expansion. It is also possible for expansion cycles to overlap.

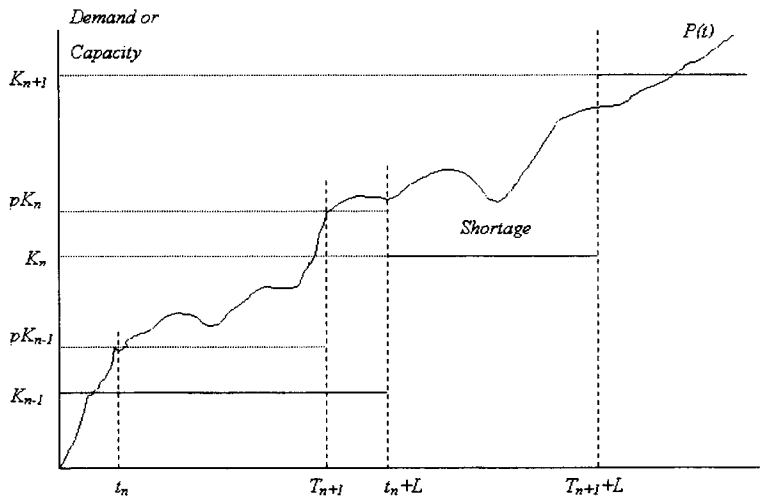

Figure 1. Capacity expansion policy when the expansion starts after the end of the current expansion cycle.

As explained in Marathe and Ryan [7], the service level in the expansion cycle $[t_n + L, T_{n+1} + L)$ is defined as:

$$\beta = E\left[\frac{\int_{t_n+L}^{T_{n+1}+L} \min\left[P(t), K_n\right]dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)dt}\right] = 1 - E\left[\frac{\int_{t_n+L}^{T_{n+1}+L} \max\left[P(t) - K_n, 0\right]dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)dt}\right]. \quad (1)$$

After approximations, the above equation becomes:

$$\beta(p,v) \approx 1 - \frac{E\left[\int_{t_n+L}^{T_{n+1}+L} \left(\max\left[P(t) - K_n, 0\right]/K_n\right)dt\right]}{pE[T_{n+1} - t_n]}. \quad (2)$$

We note that the service level is the same for each expansion cycle and a function only of the policy parameters p and v.

3. Analysis of Service Level and Expansion Cost

By comparing the numerator of Equation (2) with the up-and-out barrier option and also simplifying the denominator, we have the capacity shortage equation as:

$$1 - \beta(p,v) = \frac{I}{pE[T_{n+1} - t_n]} = \frac{I\mu}{p\ln[v]}, \text{ where}$$

$$I = \int_L^{\infty} e^{(\gamma - r)u} \left\{ \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + (\gamma + \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{\ln(v) - (\gamma + \frac{\sigma^2}{2})(u - L)}{\sigma\sqrt{u - L}}, -\sqrt{\frac{u - L}{u}} \right) \right.$$

$$- v^{\frac{2\gamma}{\sigma^2}+1} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln(v) + (\gamma + \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{-\ln(v) + (\gamma + \frac{\sigma^2}{2})(u - L)}{\sigma\sqrt{u - L}}, -\sqrt{\frac{u - L}{u}} \right)$$

$$- e^{-\gamma u} \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + (\gamma - \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{\ln(v) - (\gamma - \frac{\sigma^2}{2})(u - L)}{\sigma\sqrt{u - L}}, -\sqrt{\frac{u - L}{u}} \right)$$

$$\left. + e^{-\gamma u} v^{\frac{2\gamma}{\sigma^2}-1} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln(v) + (\gamma - \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{-\ln(v) + (\gamma - \frac{\sigma^2}{2})(u - L)}{\sigma\sqrt{u - L}}, -\sqrt{\frac{u - L}{u}} \right) \right\} du, \quad (3)$$

and $\psi(x, y, \rho)$ is the bivariate normal distribution function evaluated at (x, y) with correlation coefficient ρ.

To evaluate the infinite time horizon total cost of expansion, let $V_t(K)$ be the minimum expected cost, at time t with capacity position K, of expanding capacity over infinite horizon while satisfying the service level constraint. Let the rate at which future costs are discounted be r. Referring to Figure 1, at time t_n, when the expansion has just been initiated, our goal is to find the timing (p) and size (v) parameters for the next expansion. We assume an economies of scale regime, under which cost of installing capacity of size X is given by $C(X) = kX^a$, where k is a constant and a (< 1) is the economies of scale parameter. Hence, $C_n \equiv kX_n^a$ is the cost of expansion of size X_n, and, for the nth expansion,

$$V_{t_n}(K_{n-1}) = C_n + E_{t_n}[e^{-r(T_{n+1} - t_n)}]V_{T(pK_{n-1})}(vK_{n-1}) \quad (4)$$

Now at time T_1, the total costs (TC) incurred are the actual cost of expansion (from initial capacity position of K_0 to the new capacity position of K_1), because of start of the expansion project; and the total cost of all the future expansion discounted at time T_1.

$$\begin{aligned} TC &= E[e^{-rT_1}]V_{T_1}(K_0) \\ &= E[e^{-rT_1}]\left\{kX_1^a + E_{t_1}[e^{-r(T_2-t_1)}]V_{T_2}(K_1)\right\}, \end{aligned} \tag{5}$$

where $X_1 = K_1 - K_0 = K_0(v-1)$ is the size of the first capacity expansion; and cost of continuing from the second capacity expansion is first discounted to time T_1 and then the total cost at time T_1 including the cost of expansion is discounted to time zero. In Equation (5), we note that the expected discount factor can be evaluated independently of $V_1(K)$ is possible because of the underlying independent increments in the demand model. Now, if we keep expanding the $V_1(K)$ term in Equation (5) using Equation (4), then the expression for the expansion cost can be written as a telescoping infinite series of costs:

$$TC = E[e^{-rT_1}]\left\{kX_1^a + E_{t_1}[e^{-r(T_2-t_1)}]\left\{kX_2^a + E_{t_2}[e^{-r(T_3-t_2)}]\left\{kX_3^a + E_{t_3}[e^{-r(T_4-t_3)}]\left\{kX_4^a + E_{t_4}[e^{-r(T_5-t_4)}]...\right\}\right\}\right\}\right\}$$

Now it can be shown that the total cost equation is equivalent to:

$$f(p,v) \equiv TC = \frac{k(K_0)^a (v-1)^a p^{-\rho}}{1-v^{a-\rho}}, \text{ where } \rho = \sqrt{\frac{\mu^2}{\sigma^4} + \frac{2r}{\sigma^2}} - \frac{\mu}{\sigma^2}. \tag{6}$$

Hence, the optimization problem is to find the minimum infinite horizon cost of expansion given by Equation (6), under the constraint that the capacity shortages (from Equation (3)) in the expansion cycle cannot exceed a certain specified limit. The decision variables are the timing and the size factors of expansion, as explained earlier. The problem is formulated as:

$$\begin{aligned} &\min_{p,v\geq 1} f(p,v) = \frac{k(K_0)^a (v-1)^a p^{-\rho}}{1-v^{a-\rho}}; \\ &\text{subject to:} \\ &g(p,v) \equiv 1 - \beta(p,v) \leq \varepsilon \end{aligned} \tag{7}$$

4. Solution Methodology

The non-linear program (7) is inherently difficult because of the complex constraint expression. Since the constraint inequality involves integration of bivariate normal density functions, it is very difficult to apply the commonly used gradient-based solution methods. Hence, a derivative-free cutting plane algorithm (Bazaraa et al. [9]) was used for the problem. Important steps of the cutting plane algorithm as it applies to our problem instance are described here:

Initialization step: Select an initial feasible point $x_0 = (p_0, v_0)$.

For each iteration, solve the Master Problem, which is given as

$$\begin{aligned} &\textit{Maximize } z \\ &\textit{s.t. } z \leq f(p_j, v_j) + ug(p_j, v_j) \text{ for } j = 0...k-1 \\ &u \geq 0 \end{aligned}$$

Let (z_k, u_k) be the optimal solution. Now using the optimal value of the penalty variable u_k, solve the sub-problem:

$$\textit{Minimize } f(p,v) + u_k g(p,v) : p, v \geq 1.$$

Let $x_k = (p_k, v_k)$ be the optimal solution for the sub problem. Let $\theta(u_k) = f(p_k, v_k) + u_k g(p_k, v_k)$.

If $z_k = \theta(u_k)$ then stop. Otherwise continue with the master problem with added constraint: $z \leq f(p_k, v_k) + ug(p_k, v_k)$.

Figure 2. Cutting plane algorithm steps.

Zangwill [10] proved the convergence of this algorithm in a finite number of steps.

5. Numerical Results

The capacity expansion problem (CEP) in Equation (7) was solved using the cutting plane method with the parameter values as: drift of 8%, volatility of 20%, discount rate of 15%, economies of scale parameter of 0.9, and lead-time of 1 year, with the specified service level of 95%. The initial feasible point was taken to be (1.01, 1.01984). Initial numerical runs of the algorithm indicated an unbounded solution. Hence to test convergence of the algorithm in Figure 2, we added an artificial constraint $p \le 2$ to each sub-problem. The successive iterations and the convergence of the cutting plane algorithm are summarized in Table 1.

Table 1. Results of the cutting plane algorithm applied to CEP.

Iteration	Constraint Added	Master problem solution (z, u)	Sub-problem solution (p, v)	Sub-problem optimal value θ
1	$Z \le 3.0156 - 0.001U$	(3.0156, 0)	(2, 1.15272)	1.04207
2	$Z \le 1.04272 + 3.579U$	(2.88, 5.1455)	(2, 2.08893)	1.39851
3	$Z \le 1.35628 + 0.0083U$	(1.762, 48.94)	(1, 1.29867)	0.0331
4	$Z \le 2.7977 - 0.0564U$	(1.539, 22.32)	(2, 3.177)	1.0751
5	$Z \le 1.7725 - 0.0367U$	(1.431, 9.307)	(2, 2.34)	1.40
6	$Z \le 1.4543 - 0.0073U$	(1.4075, 6.45)	(2, 2.22)	1.403
7	$Z \le 1.407 - 0.00072U$	(1.403, 5.974)	(2, 2.17)	1.4027
8	$Z \le 1.387 + 0.00265U$	(1.4031, 6.16)	(2, 2.186)	1.4031

As seen from Table 1, the minimum cost for the CEP is achieved at decision variable values $(p, v) = (2, 2.186)$. Since our feasible region was $1 \le p \le 2$, $1 \le v$, we see that the optimal solution is reached at the artificially imposed boundary level of one of the decision variables-- indicating an *unbounded solution* to the CEP.

To explore whether the unboundedness was caused by the approximations used to transform Equation (1) to Equation (2), the same capacity expansion problem was solved using simulation to evaluate the constraint in Equation (1). That is, instead of using the analytical expression of the service level developed through the approximations and use of barrier option valuation formulas, the service level was directly computed by simulation of the GBM process using Matlab. The expression for the infinite time horizon expansion cost was the same as that was used in the analytical solution (Equation (6)). The graphical solution for the CEP is given below:

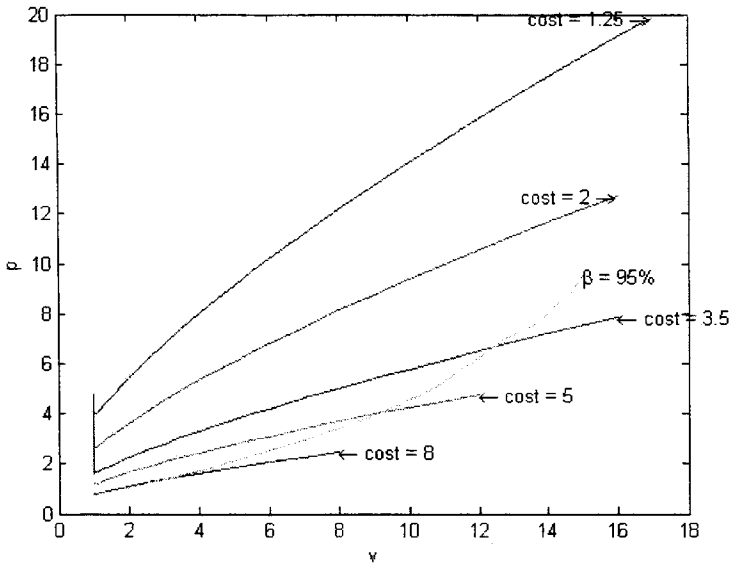

Figure 3. Graphical solution for the CEP via simulation for a 95% service level.

The feasible region is the area below the curve for the service level. The *p-v* cost contours are also plotted. As we see from Figure 3, we can go on decreasing the cost as we move away from the origin along the *p-v* curve for the service level. The simulation result was tested over wide range of parameter values, and every time we obtained similar plots. This leads us to believe that the problem under the current formulation is unbounded.

6. Conclusions and Future Work

From the formulation of the CEP and the numerical solution in section 5, we can conclude that if the service provider wants to start the capacity expansion when the demand has already crossed the current capacity position, he can do so by starting with any amount of initial shortage, provided that at each expansion the capacity increment is subsequently matched to satisfy the service level. In other words, if the service provider wants to start with a higher level of initial shortage, all that needs to be adjusted is the size of the capacity expansion so that the specified level of service is achieved (which can be accomplished by manipulating the constraint equation); moreover, this can be done without losing the minimum cost advantage (because, although the expansion cost increases in expansion size, the higher initial shortage will always pull it down). Simulation of the service level, instead of using analytical expectation, also leads to similar conclusions.

The possibility of achieving minimal cost with an acceptable service level by an unbounded initial shortage seems counter-intuitive. Hence, currently a reworking of the service level constraint is underway. Instead of considering the total shortages during the expansion cycle of random length, the distribution of instantaneous capacity shortage (in the same expansion cycle) is being considered.

7. References

1. Marathe, R.R. and Ryan, S.M., 2005a, "On the Validity of the Geometric Brownian Motion Processes," The Engineering Economist, 50(2), 159-192.
2. Manne, A.S., 1961, "Capacity Expansion and Probabilistic Growth," Econometrica, 29(4), 632-649.
3. Whitt, W., 1981, "The Stationary Distribution of a Stochastic Clearing Process," Operations Research, 29(2), 294-308.
4. Chaouch, A.B. and Buzacott, J.A., 1994, "The Effects of Lead Time on Plant Timing and Size," Production and Operations Management, 3(1), 38-54.
5. Bean, J.C., Higle, J.L. and Smith, R.L., 1992, "Capacity Expansion under Stochastic Demands," Operations Research, 40(2), S210-S216.

6. Ryan, S.M., 2004, "Capacity Expansion for Random Exponential Demand Growth with Lead Time," Management Science, 50(6), 740-748.
7. Marathe, R.R. and Ryan, S.M., 2005b, "Undercapacity as a Barrier Option: Evaluating the Service Level Constraint in Capacity Expansion," Under review with Naval Research Logistics (special issue on Applications of Financial Engineering).
8. Marathe, R.R. and Ryan, S.M., 2005c, "Capacity Expansion for Uncertain Demand with Initial Shortages," Proceedings of the 14th IIE Research Conference, May 14-18, 2004, Atlanta, Georgia.
9. Bazaraa, M.S., Sherali, H.D. and Shetty, C.M., 1993, "Nonlinear Programming: Theory and Algorithms," 2nd edition, John Wiley and Sons, New York.
10. Zangwill, W.I., 1969, "Nonlinear Programming: A Unified Approach," Prentice Hall, Englewood Cliffs, NJ.

Appendix III: Capacity Expansion for Uncertain Demand with Initial Shortages

Rahul R. Marathe and Sarah M. Ryan; Department of Industrial & Manufacturing Systems Engineering; Iowa State University; Ames, IA 50011-2164, USA

Abstract

For service providers, uncertain demand for capacity and expansion lead time may create unavoidable capacity shortages, which may be allowed to accumulate before initiating an expansion. For the demand following a geometric Brownian motion process, we assume a stationary expansion policy where the timing and size of expansion are determined as fixed proportions of the capacity position. We define the service level in terms of the capacity shortages, which can be evaluated by applying pricing formulae for barrier options in finance. We observe the relationship between the two policy parameters at different specified service levels and for other model parameters.

Keywords
Capacity expansion, Service level, Barrier options

1. Introduction
Capacity expansion is the addition of facilities to keep up with the increasing demand. The goal is to find the optimal sizes and times of expansion under given conditions. The problem is complicated in cases where the demand is stochastic and where capacity cannot be added instantaneously, meaning there is some lead-time present. This paper formulates a model for such a case in which we assume a fixed lead-time and a random demand.

The capacity expansion problem has been widely researched. Manne [1] proposed a model to decide the size of each expansion in the case where the demand follows random-walk pattern; also the effects of economies of scale and penalties for demand not being satisfied were considered in the model. Whitt [2] developed the Whitt-Luss utilization formula for the capacity expansion problem where the demand is stochastic. In the current paper, we extend this analysis for the case where the capacity expansion starts with initial shortages and there is a fixed lead time for expansion. Chaouch and Buzacott [3] examined the same problem as in Buzacott and Chaouch [4] with consideration of lead time. They also considered two cases, where the capacity addition started before and after the current capacity is reached. Our paper is similar to the work of Chaouch and Buzacott [3] in the sense that we also consider initializing the capacity expansion after certain deficit has been accumulated; only the demand process considered in our model is different. A generalization of Brownian motion demand was considered by Bean Higle and Smith [5], where demand was assumed to be following either a transformed Brownian motion process or a semi-Markovian birth and death process. They showed that the problem can be transformed into an equivalent deterministic problem and that the effect of the probabilistic nature of demand is to reduce the interest rate. This result was extended in Ryan [6] with consideration of fixed lead-time. In this model, the demand was assumed to be following a geometric Brownian motion process and a timing policy was developed to provide a specified level of service. It was showed how the parameters of the timing policy could be obtained numerically using some of concepts of financial options pricing. Our paper is a further extension of this model. While Ryan's model assumed that the next capacity expansion starts before the installed capacity level is reached, in this paper we consider a case where the next expansion is started only after accumulation of some shortages.

This situation can be compared to the barrier options in the world of finance. In particular, the value of an up-and-out call option is mathematically similar to the expected shortage considered in this paper. Heynen and Kat [7] discuss some of the important results about barrier options when the Brownian motion and its maximum are tracked over different time intervals.

In this paper, we record the service level in terms of the average capacity shortages per unit time. A detailed study of service levels for inventory models was carried out by Klemm [8]. Rigorous definitions for the three types of the service level viz. α, β, γ service levels were given for (s, S) and (r, Q) type inventory models. Further mathematical calculations for each type of service level and its effect on the order points in the various inventory models was done by Schneider [9].

As stated earlier, this paper builds on the variables and environment analyzed in Ryan. We start the paper by describing all the variables and notations used in Section 2. Here we also describe the basic model for this paper. Section 3 discusses the expression for the shortages in terms of the policy parameters. We present our numerical analyses in Section 4 and concluding remarks in Section 5.

2. Model

As our model is similar to Ryan [6], we will use consistent notation. Let $B(t)$ be a Brownian motion having drift μ and volatility σ^2 with $B(0)=0$. Demand for the product or service is given by the geometric Brownian motion (GBM) process $P(t) = P(0)e^{B(t)}$. As $P(t)$ is a GBM process, for any values of k and t given $P(t)$, the ratio $\frac{P(k+t)}{P(t)}$ is a random variable independent of all the values of the process up to t and in addition, its logarithm $\ln\left(\frac{P(t+k)}{P(t)}\right)$ has a normal distribution with mean μk and variance $\sigma^2 k$. And hence, given $P(t)$, the logarithmic growth in demand over a short interval of time Δt is given by $\ln\left(\frac{P(t+\Delta t)}{P(t)}\right) = \mu\Delta t + \sigma\sqrt{\Delta t}z$, where z is a standard normal random variable. Define $\gamma \equiv \mu + \frac{\sigma^2}{2}$ as the mean (exponential) growth rate of the demand. Marathe and Ryan [10] empirically verified the fit of the GBM process to historical data series for usage of airline and electric power capacity.

We assume that capacity additions occur at discrete time points and that a fixed lead time of L time units is required to install new capacity. The problem is to choose a sequence $\{(T_n, X_n), n \geq 1\}$, where T_n, the time when the n^{th} capacity expansion starts, is a stopping time with respect to the Brownian motion $B(t)$ and X_n is the n^{th} increase in capacity. For any realization ω of the Brownian motion $B(t)$, let $t_n \equiv T_n(\omega)$. Let K_n be the installed capacity after n additions are completed, where the initial capacity is K_0. Then,

$$K_n = K_0 + \sum_{j=1}^{n} X_j \cdot$$

The installed capacity at time t is given by,

$$K(t) = \begin{cases} K_0, & 0 \leq t < t_1 + L \\ K_n, & t_n + L \leq t < t_{n+1} + L, \; n \geq 1. \end{cases}$$

The capacity position at time t is given by,

$$\Pi(t) = \begin{cases} K_0, & 0 \leq t < t_1 \\ K_n, & t_n \leq t < t_{n+1}, \; n \geq 1. \end{cases}$$

We assume that the policy proposed by Whitt and Luss for the same demand function is modified to account for the lead times and to allow planned shortages to occur. Whitt [2] showed that, without lead times, their policy results in a stationary distribution for the capacity utilization and provided a simple formula for its expected value. In the Whitt-Luss policy, each new expansion occurs when demand reaches some fixed proportion (< 1) of current capacity, and after its instantaneous addition, the new capacity is a constant proportion of its previous value. In this paper, we assume that each expansion occurs when demand reaches some fixed proportion, p, of the capacity *position*, and $K_n = vK_{n-1}$, where $v > 1$. Ryan [6] showed that for $p < 1$ with fixed lead times, the value of p to attain a specified service level can be found by using the Black-Scholes formula for pricing a European call option. Moreover, assuming this timing policy is followed, the expansion size policy minimizes

the infinite horizon discounted cost under a widely used expansion cost function that reflects economies of scale. In this paper, we consider the case where $p \geq 1$.

The policy assumes that ever-increasing increments of capacity can be installed within the same lead time to keep pace with exponentially growing demand. This assumption is most reasonable in industries where capacity bottlenecks are caused by facilities subject to continuous technological improvement, such as those that rely on information and communications technology. However, it may hold in more traditional situations as well. For example, Lieberman [11] found that the Whitt-Luss policy provided the closest fit among several alternatives to the capacity utilization in an empirical study of the chemical product industry. Over at least two decades, total output grew by an average of 6.2% per year, and the mean size of expansion increments translated to a value of $v = 1.09$ at the plant level.

Figures 1 and 2 illustrate the policy and potential shortages seen at the realized time t_n, when demand first equals pK_{n-1}. The n^{th} capacity expansion has just started. With this expansion, the total installed capacity will reach level K_n after the lead time L. As stated earlier, we model the situation wherein the manufacturer waits until certain amounts of capacity shortages are accumulated before starting the next expansion project. This "certain amount of shortages" is represented by the decision variable $p, p \geq 1$. Hence, since the new capacity position is K_n, the next expansion would start at the time when the demand $P(t)$ first reaches the position pK_n. Since the demand process is stochastic, this time for starting the next expansion (T_{n+1}) is a random variable. The second decision parameter is the size of each expansion $v = K_{n+1}/K_n$.

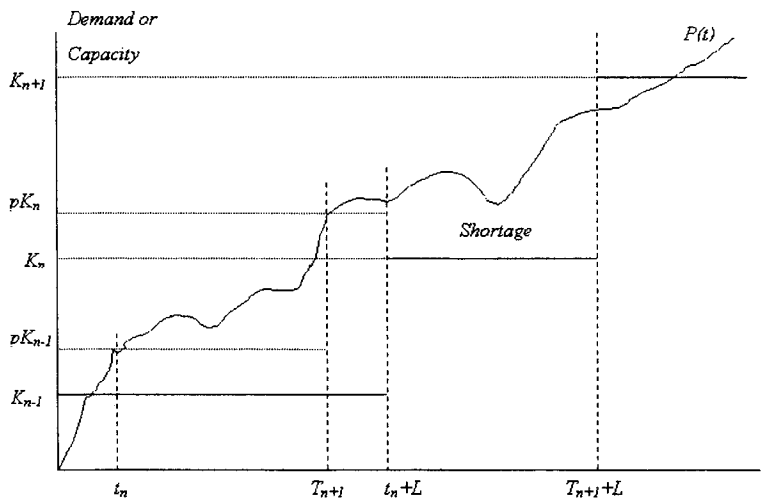

Figure 1. Capacity expansion policy when the expansion starts after the end of the current expansion cycle.

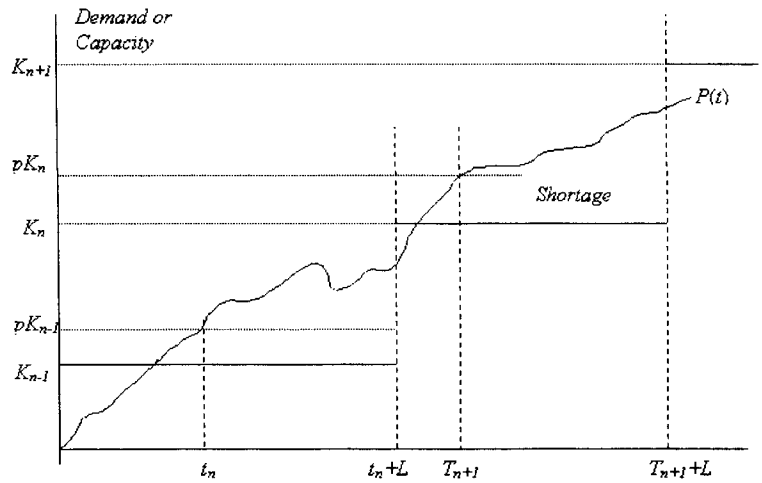

Figure 2. Capacity expansion policy when the expansion starts before the end of the current expansion cycle.

It is natural to define a cycle as the time interval from the end of one lead time to the end of the next, so that the actual capacity is constant over the cycle. For a generic cycle, we formulate a service measure akin to the fill rate used in periodic [12] and continuous review [8, 13], inventory models. The cycle length may be less than, equal to, or greater than L, depending on whether successive lead times overlap. At a generic expansion epoch t_n, the decision maker knows $P(t_n) = pK_{n-1}$ and wishes to predict the service level over the interval $[t_n + L, T_{n+1} + L)$. Schneider [9] defines the β service level as the fraction of demand not being lost or backordered, which is relevant for lost sales or proportional backorder costs. Here, the proportion of demand that is satisfied is

$$\beta = E\left[\frac{\int_{t_n+L}^{T_{n+1}+L} \min\left[P(t), K_n\right] dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)\,dt}\right] = 1 - E\left[\frac{\int_{t_n+L}^{T_{n+1}+L} \max\left[P(t) - K_n, 0\right] dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)\,dt}\right],$$

where the expectation is taken with respect to time t_n. As is commonly done in inventory models [13, 14], we employ a series of approximations to obtain a tractable service measure.

First, since the closest known value for demand during the cycle is $P(T_{n+1}) = pK_n$, we approximate the denominator as $\int_{t_n+L}^{T_{n+1}+L} P(t)\,dt \approx pK_n(T_{n+1} - t_n)$. Second, we approximate the expected value of the ratio as the ratio of expected values:

$$\beta \approx 1 - \frac{E\left[\int_{t_n+L}^{T_{n+1}+L} \left(\max\left[P(t) - K_n, 0\right]/K_n\right) dt\right]}{pE\left[T_{n+1} - t_n\right]}. \tag{8}$$

3. Mathematical Analysis

As done in [2] for capacity utilization, our goal here is to express the average shortages in terms of the decision variables viz. the timing and size parameter. The expression can be used to obtain the values for the decision variables that achieve a given service level, or estimate the service level for given values of the decision parameters.

From equation (8), the total expected shortage for the next cycle assuming that we know the demand and capacity at time t_n is (numerator of the equation (8)):

$$\mathrm{I} = E_{t_n}\left[\int_{t_n+L}^{T_{n+1}+L} \frac{1}{K_n}(P(t) - K_n)1_{(P(t)\geq K_n)}\,dt\right], \tag{9}$$

where $1_{(x)}$ is indicator function such that $1_{(x)} = 1$ if x is true and

$= 0$ otherwise.

The above integration can be solved by comparing the shortages to the barrier option scenario in the finance world – particularly, the up-and-out call option. Heynen and Kat [7] give the analytical equation for Up-and-Out Call option. After simplifications, the integral I of equation (9) becomes:

$$I = \int_{L}^{\infty} e^{\gamma u} \left\{ \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + (\gamma + \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{\ln(v) - (\gamma + \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right) \right.$$

$$- v^{\frac{2\gamma}{\sigma^2}+1} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln(v) + (\gamma + \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{-\ln(v) + (\gamma + \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right)$$

$$- e^{-\gamma u} \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + (\gamma - \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{\ln(v) - (\gamma - \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right)$$

$$\left. + e^{-\gamma u} v^{\frac{2\gamma}{\sigma^2}-1} \left(\frac{p}{v}\right) \psi\left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln(v) + (\gamma - \frac{\sigma^2}{2})u}{\sigma\sqrt{u}}, \frac{-\ln(v) + (\gamma - \frac{\sigma^2}{2})(u-L)}{\sigma\sqrt{u-L}}, -\sqrt{\frac{u-L}{u}} \right) \right\} du.$$

where $\psi(x, y, \rho)$ is the bivariate normal distribution function for variables X and Y with coefficient of correlation ρ.
Hence,

$$1 - \beta = \frac{I}{pE[T_{n+1} - t_n]} = \frac{I\mu}{p \ln[v]}.$$

For complete mathematical treatment of the above equation please refer to the full version of this paper [15].

Unlike the timing policy in Theorem 1 of [6], the average shortage in our model depends on both decision variables, which is similar to the case in [2] where the capacity utilization was dependent on the timing and size parameters.

4. Results

In addition to the timing and size parameters, the average shortage (1-β) also is affected by other parameters in the model, viz. length of the lead time, the drift and volatility factor of the demand process, etc. The following plots show the effect of each parameter on the average shortage. While analyzing the effect of any particular parameter on the shortages, the values of other parameters were kept constant.

Figure 3 depicts the effect of the parameter p on the values of the average shortage, i.e., given the value of the variable p on the x-axis; the plot gives the value of the corresponding shortages, for a given value of the size parameter (v). The values of the other parameters are $v = 1.1$, $\sigma = 15\%$, $\mu = 8\%$, and $L = 2$ years. As expected, to achieve the target of low average shortage during the expansion cycle, the manufacturer should start the expansion project with low initial shortages. Also the average shortages are reduced by increase in the volatility of the demand process. In contrast, an increase in the drift parameter of the demand process causes the average shortage to increase. It was also found that the average shortages increase with the length of the lead time.

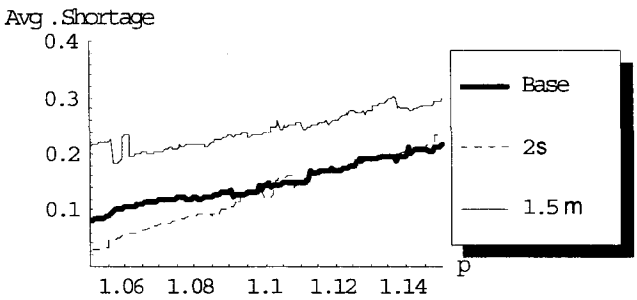

Figure 3. Effects of p and other variables on the average shortage.

Figure 4 tells the story of the second decision variable v and its relationship with our other decision variable p. In Figure 4, the values for the decision variable v, is plotted for different scenarios. For the base case, we assume that p = 1.05, σ = 0.2, μ = 0.08, L = 2 years and the average shortages are held at 5% of the current capacity level. If we delay the start of capacity expansion project (increase the value of p), then to maintain the same average shortage we have to increase the size of each expansion. Also, if the allowable shortages are increased, as expected, the required size of each expansion reduces. It was also seen that for increase in the volatility, the size parameter decreases for the same level of average shortage. Similarly, increase in lead time would force us to increase the size parameter to achieve the same target average shortage.

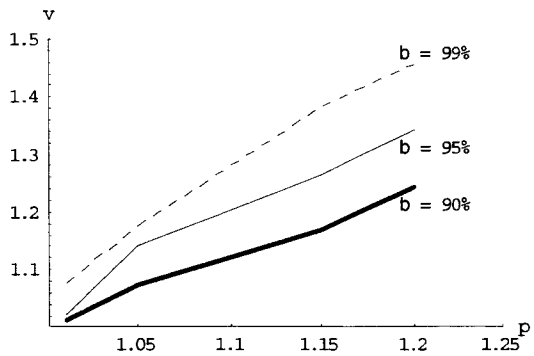

Figure 4. Relationship between size and timing parameters.

5. Conclusion

Random demand combined with the presence of expansion lead times increase the criticality of the capacity expansion problem and this may lead to unavoidable initial capacity shortages. Another motivation to delay the expansion could be to allow additional observation of the uncertain demand before initiating the expansion. We have modeled one such case when capacity shortages were defined as a proportion of the existing capacity. Using the concepts from financial option pricing, an analytical expression for the capacity shortages was found in terms of the timing and size parameters of the expansion. We found out that as we allow more shortages, the size of each expansion project decreases; whereas if we delay the expansion project, the resulting size of expansion to maintain the specified level of shortages increases.

References

[1] Manne, A. S., 1961, "Capacity Expansion and Probabilistic Growth," Econometrica, 29(4), 632-649.

[2] Whitt, W., 1981, "The Stationary Distribution of a Stochastic Clearing Process," Operations Research, 29(2), 294-308.

[3] Chaouch, A. B. and Buzacott, J. A., 1994, "The Effects of Lead Time on Plant Timing and Size," Production and Operations Management, 3(1), 38-54.

[4] Buzacott, J. A. and Chaouch, A. B., 1988, "Capacity Expansion with Interrupted Demand Growth," European Journal of Operational Research, 34(19-26.

[5] Bean, J. C., Higle, J. L. and Smith, R. L., 1992, "Capacity Expansion under Stochastic Demands," Operations Research, 40(2), S210-S216.

[6] Ryan, S. M., 2004, "Capacity Expansion for Random Exponential Demand Growth with Lead Time," Management Science, 50(6), 740-748.

[7] Heynen, R. C. and Kat, H. M., 1997, "Chapter 6: Barrier Options," appears in Exotic Options, Clewlow, L. and Strikland, C. (eds.), International Thompson Business Press, 125-138.

[8] Klemm, H., 1971, "On the Operating Characteristic 'Service Level'," appears in Inventory Control and Water Storage, Prekopa, A. (eds.), North-Holland Publishing Company, Amsterdam, 169-178.

[9] Schneider, H., 1981, "Effect of Service-Levels on Order-Points or Order-Levels in Inventory Models," International Journal of Production Research, 19(6), 615-631.

[10] Marathe, R. R. and Ryan, S. M., 2005, "On the Validity of Geometric Brownian Motion Assumption," Forthcoming in The Engineering Economist.

[11] Lieberman, M. B., 1989, "Capacity Utilization: Theoretical Models and Empirical Tests," European Journal of Operations Research, 40(155-168.

[12] Sobel, J. M., 2004, "Fill Rates of Single-Stage and Multistage Supply System," Manufacturing and Service Operations Management, 6(1), 41-52.

[13] Hadley, G. and Whitin, T. M., 1963, Analysis of Inventory Systems, Prentice-Hall Inc Englewood Cliffs, New Jersey.

[14] Janssen, F., Heuts, R. and de Kok, T., 1999, "The Impact of Data Collection on Fill Rate Performance in the (R, s, Q) Inventory Model," The Journal of the Operational Research Society, 50(1), 75-84.

[15] Marathe, R. R. and Ryan, S. M., 2005, "Undercapacity as a Barrier Option: Evaluation of Service Level Constraint in Capacity Expansion," Working paper, Iowa State University Ames, Iowa USA.

Appendix IV: Mathematica 5.1 code

Appendix 4A

```
p1 = 1.227; v1 = 1.0077; f1 = -0.482;

p2 = 1.0076; v2 = 1.01; f2 = -0.6893;

p3 = 0.929; v3 = 1.029; f3 = -0.7603;

p4 = 0.966; v4 = 1.0136; f4 = -0.7056;

p5 = 0.9971; v5 = 1.0111; f5 = -0.67067;

NMinimize[{x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10,
  a1 + p1 a11 + v1 a12 - (x1 - x2) == f1,
  a2 + p2 a21 + v2 a22 - (x3 - x4) == f2,
  a3 + p3 a31 + v3 a32 - (x5 - x6) == f3,
  a4 + p4 a41 + v4 a42 - (x7 - x8) == f4,
  a5 + p5 a51 + v5 a52 - (x9 - x10) == f5,
  -a1 - p2 a11 - v2 a12 + (x3 - x4) ≤ -f2,
  -a1 - p3 a11 - v3 a12 + (x5 - x6) ≤ -f3,
  -a1 - p4 a11 - v4 a12 + (x7 - x8) ≤ -f4,
  -a1 - p5 a11 - v5 a12 + (x9 - x10) ≤ -f5,
  -a2 - p1 a21 - v1 a22 + (x1 - x2) ≤ -f1,
  -a2 - p3 a21 - v3 a22 + (x5 - x6) ≤ -f3,
  -a2 - p4 a21 - v4 a22 + (x7 - x8) ≤ -f4,
  -a2 - p5 a21 - v5 a22 + (x9 - x10) ≤ -f5,
  -a3 - p1 a31 - v1 a32 + (x1 - x2) ≤ -f1,
  -a3 - p2 a31 - v2 a32 + (x3 - x4) ≤ -f2,
  -a3 - p4 a31 - v4 a32 + (x7 - x8) ≤ -f4,
  -a3 - p5 a31 - v5 a32 + (x9 - x10) ≤ -f5,
  -a4 - p1 a41 - v1 a42 + (x1 - x2) ≤ -f1,
  -a4 - p2 a41 - v2 a42 + (x3 - x4) ≤ -f2,
  -a4 - p3 a41 - v3 a42 + (x5 - x6) ≤ -f3,
  -a4 - p5 a41 - v5 a42 + (x9 - x10) ≤ -f5,
  -a5 - p1 a51 - v1 a52 + (x1 - x2) ≤ -f1,
  -a5 - p2 a51 - v2 a52 + (x3 - x4) ≤ -f2,
  -a5 - p3 a51 - v3 a52 + (x5 - x6) ≤ -f3,
  -a5 - p4 a51 - v4 a52 + (x7 - x8) ≤ -f4, x1 ≥ 0, x2 ≥ 0,
  x3 ≥ 0, x4 ≥ 0, x5 ≥ 0, x6 ≥ 0, x7 ≥ 0, x8 ≥ 0, x9 ≥ 0, x10 ≥ 0},
 {x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, a1, a2, a3,
  a4, a5, a11, a12, a21, a22, a31, a32, a41, a42, a51, a52}]
```

```
{0., {a1 → 0., a11 → -0.392828, a12 → 0.,
  a2 → -31.0758, a21 → 1.2472, a22 → 28.8414, a3 → 0.,
  a31 → 0.906762, a32 → -1.55751, a4 → -0.636283,
  a41 → 1.03812, a42 → -1.05776, a5 → 0., a51 → 0.906762,
  a52 → -1.55751, x1 → 0., x10 → 0., x2 → 0., x3 → 0.,
  x4 → 0., x5 → 0., x6 → 0., x7 → 0., x8 → 0., x9 → 0.}}
```

Appendix 4B:

```
<< Statistics`ContinuousDistributions`

MVN[x_, mu_, var_] :=
 Module[{SSS = Inverse[var]},
  (2 π)^(-Length[x]/2) Det[SSS]^(1/2) e^((-1/2) (x-mu).SSS.(x-mu))];

xvec = {x1, x2};

muvec = {0, 0};

varcov = {{1, ρ}, {ρ, 1}};

nord := NormalDistribution[0, 1];

cdfunc[x_] := CDF[nord, x];

f = MVN[xvec, muvec, varcov] // Simplify;

r = μ + σ^2/2;

h1[p_, v_, σ_, r_, u_] := (-Log[v/p] + (r + σ^2/2) u)/(σ Sqrt[u]);

t1[v_, σ_, r_, u_] := (Log[v] - (r + σ^2/2) (u - L))/(σ Sqrt[u - L]);

h2[p_, v_, σ_, r_, u_] := h1[p, v, σ, r, u] + 2 Log[v]/(σ Sqrt[u]);

t2[v_, σ_, r_, u_] := (-Log[v] - (r + σ^2/2) (u - L))/(σ Sqrt[u - L]);

h3[p_, v_, σ_, r_, u_] := (-Log[v/p] + (r - σ^2/2) u)/(σ Sqrt[u]);

t3[v_, σ_, r_, u_] := (Log[v] - (r - σ^2/2) (u - L))/(σ Sqrt[u - L]);
```

```
h4[p_, v_, σ_, r_, u_] := h3[p, v, σ, r, u] + 2 Log[v] / (σ √u);

t4[v_, σ_, r_, u_] := (-Log[v] - (r - σ^2/2) (u - L)) / (σ √(u - L));

ρ = -√((u - L)/u);

L = 2;

μ = 0.02;

σ = 0.3;

spec = 0.05;

a = 0.99;

b = 1.49192;

γ = 0.13;

F1[p_, v_, σ_, r_, u_] :=
 NIntegrate[f, {x1, -∞, h1[p, v, σ, r, u]},
  {x2, -∞, t1[v, σ, r, u]}, AccuracyGoal → 4,
  PrecisionGoal → 4, SingularityDepth → 25,
  MaxRecursion → 30];

F2[p_, v_, σ_, r_, u_] :=
 NIntegrate[f, {x1, -∞, h2[p, v, σ, r, u]},
  {x2, -∞, t2[v, σ, r, u]}, AccuracyGoal → 4,
  PrecisionGoal → 4, SingularityDepth → 25,
  MaxRecursion → 30];

F3[p_, v_, σ_, r_, u_] :=
 NIntegrate[f, {x1, -∞, h3[p, v, σ, r, u]},
  {x2, -∞, t3[v, σ, r, u]}, AccuracyGoal → 4,
  PrecisionGoal → 4, SingularityDepth → 25,
  MaxRecursion → 30];

F4[p_, v_, σ_, r_, u_] :=
 NIntegrate[f, {x1, -∞, h4[p, v, σ, r, u]},
  {x2, -∞, t4[v, σ, r, u]}, AccuracyGoal → 4,
  PrecisionGoal → 4, SingularityDepth → 25,
  MaxRecursion → 30];

F5[v_, σ_, r_, u_] := cdfunc[t1[v, σ, r, u]];

F6[v_, σ_, r_, u_] := cdfunc[t2[v, σ, r, u]];
```

```
ShortFun[p_, v_, σ_, r_, γ_] :=
 NIntegrate[
  e^((r-γ) u)
   ((p/v) F1[p, v, σ, r, u] - v^(2r/σ^2+1) (p/v) F2[p, v, σ, r, u] -
     e^(-r u) F3[p, v, σ, r, u] +
     e^(-r u) v^(2r/σ^2-1) (p/v) F4[p, v, σ, r, u] -
     spec (p/v) F5[v, σ, r, u] +
     spec (p/v) v^(2r/σ^2+1) F6[v, σ, r, u]), {u, L, ∞},
  AccuracyGoal → 4, PrecisionGoal → 4,
  SingularityDepth → 25, MaxRecursion → 30];

TotC[p_, v_, a_, b_] := (v - 1)^a p^(-b) / (1 - v^(a-b));

ShortFun[0.99938, 1.0327, σ, r, γ]

TotC[0.99938, 1.0327, a, b]
```

```
NIntegrate::nintp : Encountered the non-number
    3.33333 (-«20» + «1»)/√u at {x1, x2} = {x1, x2}. More…

NIntegrate::nintp :
 Encountered the non-number 0.214512/√u + 3.33333 «1» («1» «1»)/√u
   at {x1, x2} = {x1, x2}. More…

NIntegrate::nintp :
 Encountered the non-number 0.214512/√u + 3.33333 «1» («1» «1»)/√u
   at {x1, x2} = {x1, x2}. More…

General::stop : Further output of NIntegrate::nintp will
   be suppressed during this calculation. More…

-0.00547128

2.11412
```