ABSTRACT

NOROUZI, AMIRHOSEIN. The Effect of Forecast Evolution on Production Planning with Resources Subject to Congestion. (Under the direction of Dr. Reha Uzsoy).

The value of demand forecast information in production-inventory planning systems has been the subject of numerous studies in the literature. While queuing models of production systems show that lead times increase nonlinearly with resource utilization, there are few models that analyze the effect of forecast evolution on production planning with resources subject to congestion. This dissertation consists of three papers, each addressing one or more aspects of these production systems. In the first paper, we show that the progressive realization of uncertain demands across successive discrete time periods through additive or multiplicative forecast updates results in the evolution of the conditional covariance of demand in addition to its conditional mean. A dynamic inventory model with forecast updates is used to illustrate the application of our method. We show how the optimal inventory policy depends on conditional covariances, and use a model without information updates to quantify the benefit of using the available forecast information in the presence of additive forecast updates. Our approach yields significant reductions in system costs, and is applicable to a wide range of production and inventory models. We also extend our approach to the case of multiplicative forecast updates, and discuss directions for future work.

In the second paper, we consider a single stage production-inventory system with work-load dependent lead times and uncertain demands. The dependency between workload and lead times is captured through clearing functions that take into account the nonlinear relationship between utilization and lead times. We propose a load dependent release policy leading to a tractable chance constrained optimization model. We compare the performance

of this approach with that of a multistage stochastic programming model by subjecting them to a simulation of uncertain demand realizations. Computational experiments show promising performance of our chance constrained model.

In the third paper we consider a dynamic production-inventory system with forecast updates based on the martingale model of forecast evolution (MMFE) used in our first paper. The nonlinear dependency between workload and lead times is again captured through clearing functions. In this setting, we first formulate a chance constrained model and show how information affects the performance of the system subject to congestion. We then evaluate the performance of this model compared to that of a multistage stochastic programming model, for which we propose a method to represent the forecast evolution in a set of discrete scenarios. Our computational study helps to quantify the value of forecast update information, and suggests that the chance-constrained models provide a good tradeoff between solution quality and model complexity.

The Effect of Forecast Evolution on Production Planning with Resources Subject to Congestion

by
Amirhosein Norouzi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Operations Research

Raleigh, North Carolina

2012

APPROVED BY:

_____	_____
Dr. Reha Uzsoy Chair of Advisory Committee	Dr. James Wilson
_____	_____
Dr. Yahya Fathi	Dr. Donald Warsing
_____	
Dr. Karl Kempf	

DEDICATION

This dissertation is dedicated to all those

whom I love, especially

My dear wife, Giti,

My father, Mohammadreza,

and

the memory of my mother, Minou

BIOGRAPHY

Amirhosein Norouzi is a Ph.D. student in the Operations Research Program at North Carolina State University. He received the Bachelor of Science degree in Industrial Engineering from Sharif University of Technology in 2005 and the Master's degree in Operations Research from North Carolina State University in 2009. During his years at NC State University, he served as a research assistant and obtained several semesters of teaching experience as a teaching assistant. His research interests include forecasting, production and inventory planning, and stochastic modeling.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Reha Uzsoy, for his utmost support and thoughtful guidance during the course of preparing this dissertation. My special thanks go to Dr. Wilson, Dr. Fathi, Dr. Warsing and Dr. Kempf for serving on my committee and for all constructive comments they have given me on this work. I also thank Dr. Denton who has provided valuable advice along the way. I would like to express my sincere gratitude to my wife for her continuous support and encouragement during my studies. This research was supported by the National Science Foundation under Grant No. 1029706, which is also gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

Chapter 1 Introduction 1

Chapter 2 Modeling the Evolution of Dependency between Demands, with Application to Inventory Planning 5

2.1 Introduction 6

2.2 Previous Related Work 7

2.3 Forecast Evolution Model 9

2.4 The Additive Model 10

2.4.1 Evolution of the Conditional Mean 11

2.4.2 Evolution of the Conditional Covariance 12

2.5 The Multiplicative Model 16

2.5.1 Evolution of the Conditional Mean 17

2.5.2 Evolution of the Conditional Covariance 18

2.6 Application to an Inventory Model 22

2.7 Numerical Experiments with the Additive Model 28

2.8 Extension to the Multiplicative Model 33

2.9 Conclusions and Future Directions 35

Chapter 3 Production Planning with Load Dependent Lead Time and Uncertain Demand 37

3.1 Introduction 38

3.2 Literature Review 40

3.3 Clearing Functions 44

3.4 Production Planning Models 48

3.4.1 Chance Constrained Model (CC) 48

3.4.2 Shortfall Chance Constrained Model (S-CC) 50

3.4.3 Stochastic Programming Model 57

3.5 Computational Experiments 60

3.5.1 Experimental Design 60
3.5.2 Results 62
3.6 Conclusions 67
Chapter 4 Value of Demand Forecast Information in a Production System with Load Dependent Lead Time 70
4.1 Introduction 71
4.2 Literature Review 73
4.3 The Model 77
4.3.1 The Forecast Update Model 77
4.3.2 The Production Planning Model 79
4.4 Analysis 82
4.4.1 Chance Constrained Model 83
4.4.2 Stochastic Programming Model 87
4.5 Computational Experiments 89
4.5.1 Experimental Design 89
4.5.2 Results 92
4.6 Conclusions 93
Chapter 5 Conclusions and Future Directions 96
References 100

LIST OF TABLES

Table 2.1 Impact of the lead time 33
Table 3.1 Two-segment clearing function 61
Table 3.2 Discrete approximation for scenarios 63
Table 3.3 Lead time variance 66
Table 3.4 Three-segment clearing function 67
Table 3.5 Lead time variance 68
Table 4.1 Five point approximation for standard normal distribution 90
Table 4.2 Clearing function parameters 90

LIST OF FIGURES

Figure 2.1 Unconditional pairwise covariances of demand showing the components of γ_i for different values of i 13
Figure 2.2 Known and unknown portions of D_t and D_{t+i} 14
Figure 2.3 Conditional covariance evolution 16
Figure 2.4 Comparison of expected cost in each period of the planning horizon T for different cost ratios $b/h + b$ 30
Figure 2.5 Impact of correlation on cost reduction k 32
Figure 2.6 Impact of correlation on cost reduction k: Early Uncertainty resolution 33
Figure 3.1 Different forms of clearing functions 45
Figure 3.2 Scenario tree 58
Figure 3.3 Rolling horizon 62
Figure 3.4 Impact of number of branches per node 64
Figure 3.5 Impact of system utilization 65
Figure 3.6 Impact of third segment 68
Figure 4.1 A typical clearing function and outer linearized segments 81
Figure 4.2 Rolling horizon 91
Figure 4.3 Impact of system utilization and service level 92
Figure 4.4 Impact of information 94

Chapter 1

Introduction

Demand forecasts over a planning horizon are a key input for most supply chain planning processes. Since these forecasts are revised over time as new information becomes available, production planning and inventory control models that use these forecasts should take into account the evolution of these updates over time. Since these updates depend on both revealed information about the uncertain demand and the forecasting mechanism in use, this requires a statistical characterization of the process of forecast evolution over time as demand information becomes available. With each forecast revision, ordering or production decisions can be updated on a rolling horizon basis in order to incorporate the most recent forecast information. Order releases determine work in process (WIP) inventory levels, which determine resource utilization and, in turn, the cycle time, the time between a product being released into the factory and its completion.

The cycle time is important determinant of global competitiveness. Long cycle times increase inventory costs due to high work in process (WIP) inventory levels as well as increased safety stocks caused by increased uncertainty about demand. They are thus important to effective production planning: the decision as to how to release work into

production facilities over time must take cycle times into account in some manner in order to match the output of the production facility with market demand in optimal way. We shall refer to the estimate of cycle time used in production planning as the lead time.

In most production planning and inventory models, the lead time is not generally modeled in detail. It is commonly treated as an exogenous parameter in deterministic optimization models and many inventory models (Hackman and Leachman 1989; Zipkin 2000, Chapter 6) or as an exogenously defined probability distribution when it is taken to be stochastic (Zipkin 2000, Chapter 7). Queuing models of production systems (Buzacott and Shanthikumar 1993) show that lead times increase nonlinearly with resource utilization. It is important for many manufacturing systems with uncertain demands, such as those encountered in semiconductor manufacturing, to run at high utilization to be profitable. Under these conditions, small fluctuations in utilization may cause large changes in lead times. However, there are no production planning models linking order releases and planning decisions to lead times in the presence of forecast evolution.

Our goal is to explore the following questions: How can forecast information be used within a production system with congestion? How much does the forecast information affect the performance of the system? While information is always beneficial, we want to investigate when this information is most beneficial and when it is only marginally useful. For this purpose, we develop models that adjust the manufacturer's production and inventory planning decisions according to the forecast updates in the face of production resources subject to congestion. This requires capturing the variation of forecast updates and the related correlation structure over time as well as the dependency between workload and cycle time

discussed above. The forecast evolution follows the Martingale Model of Forecast Evolution (MMFE), developed by Graves et al. (1986) and Heath and Jackson (1994). The forecasting process provides information about future demands. This information is used to determine order releases. The dependency between workload and cycle times is captured through the use of nonlinear clearing functions (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989) that relate the expected output of a capacitated resource in a planning period to the average WIP level at the resource over the period.

This dissertation contributes to the literature in three major ways. In Chapter 2, we show that the progressive resolution of demand uncertainty through forecast updates results in evolution of the conditional covariance of the demand in addition to the evolution of its conditional mean. This additional evolution yields a better characterization of demand behavior that leads to more effective decisions. The benefit of this approach is analyzed in a multi-period inventory planning model with additive forecast updates. The model indicates that it is the relative difference between unconditional and conditional demand variance over the lead time that determines the magnitude of the cost reduction. Our results can also be used to quantify the benefits of more advance information (earlier uncertainty resolution) that helps managers in their strategic decisions.

In Chapter 3, we study production planning models that consider workload dependent lead time and stochastic demand in an integrated manner using clearing functions. We propose a load dependent release policy leading to a chance constrained model that considers workload dependent lead times of clearing functions without requiring any external lead time parameters. The computational experiments indicate that the chance-constrained model

performs favorably relative to the stochastic programming models with significantly lower computation time requirements.

Finally, in Chapter 4, we extend production-inventory models by incorporating dynamic forecast evolution and workload dependent lead times into a unified production planning model. We first obtain a chance constrained formulation where the right hand side of the service level constraint depends on the evolution of demand forecast information. A numerical comparison with a model that does not use forecast evolution demonstrates when forecast evolution information would be most useful. The model indicates that the cost reduction increases with capacity, and there is almost no benefit at high utilization. It also shows that the service level does not affect the cost reduction. We also propose a method to generate discrete scenarios for use in multistage stochastic programming models that represent the correlated demand structure induced by the presence of forecast evolution.

Chapter 2

Modeling the Evolution of Dependency between Demands, with Application to Inventory Planning

Abstract

We show that the progressive realization of uncertain demands across successive discrete time periods through additive or multiplicative forecast updates results in the evolution of the conditional covariance of demand in addition to its conditional mean. A dynamic inventory model with forecast updates is used to illustrate the application of our method. We show how the optimal inventory policy depends on conditional covariances, and use a model without information updates to quantify the benefit of using the available forecast information in the presence of additive forecast updates. Our approach yields significant reductions in system costs, and is applicable to a wide range of production and inventory models. We also extend our approach to the case of multiplicative forecast updates, and discuss directions for future work.

2.1 Introduction

Demand forecasts over a planning horizon are a key input for most supply chain planning processes. Since these forecasts are revised over time as new information becomes available, production planning and inventory control models that use these forecasts should take into account the evolution of these updates over time. Since these updates depend on both revealed information about the uncertain demand and the forecasting mechanism in use, this requires a statistical characterization of the process of forecast evolution over time as demand information becomes available. With each forecast revision, decisions can be updated on a rolling horizon basis in order to incorporate the most recent forecast information.

Successive forecast updates for the demand of each time period describe the progressive realization of the demand uncertainty. Moreover, in each period s the newly generated forecast updates for future demand in time periods $t > s$ are rarely independent of each other, since these updates are all based on the same information history that is available in period s. This progressive revelation of demand uncertainty through successive forecast updates over time leads to an evolution in the structure of the dependency between demands in consecutive time periods.

Graves et al. (1986) and Heath and Jackson (1994) propose a model to describe the evolution of forecasts over time. This Martingale Model of Forecast Evolution (MMFE) assumes that forecasts represent the conditional expectations of demands given all available information. In this paper we propose a model that builds upon the MMFE to incorporate the evolution of dependency between demands in order to obtain a better characterization of

demand behavior. This is accomplished by modeling the evolution of the conditional covariances over time in addition to that of the conditional expectation. Our model is developed for both additive and multiplicative models of forecast updating.

We use a single item, finite-horizon, periodic-review inventory model with forecast updates to illustrate the application of our approach. We show how the optimal production policy depends on conditional covariances, and that under some assumptions a myopic policy is optimal with the optimal base stock levels being functions of both conditional expectations and conditional covariances. This enables us to address some theoretical and managerial questions: How can forecast information be used within inventory models? Which information is really important? How does lead time demand impact the value of information updates? Our computational study helps to quantify the value of this evolution-based model.

In the next section we briefly review previous related work. Section 2.3 reviews the forecast evolution model. Sections 2.4 and 2.5 develop the conditional mean and covariance for additive and multiplicative forecast updating, respectively. Section 2.6 describes the dynamic inventory planning model that serves as the application, and Section 2.7 contains the results of computational experiments for the additive model that can help managers in their strategic decisions. Section 2.8 extends our results to the case of multiplicative forecast updates. A summary of our conclusions and a discussion of future work conclude the paper.

2.2 Previous Related Work

Modeling the evolution of forecasts was first suggested by Hausman (1969), who models the evolution of the forecasts as a quasi-Markovian or Markovian system in which

the current forecast is used as the state variable of the system. His model is limited to a single selling season. To accommodate the simultaneous evolution of forecasts for demand in many periods, the Martingale Model of Forecast Evolution (MMFE) was independently developed by Graves et al. (1986) and Heath and Jackson (1994). The MMFE allows forecasts to evolve over time as new information becomes available in each period. Graves et al. (1986) construct a single item version of the MMFE to model the evolution of forecasts of independent and identically distributed stationary demand over discrete time periods. Heath and Jackson (1994) propose a multiproduct model that can capture correlations in demand forecasts both between products and across time periods. They use it to generate forecast updates in a simulation model in order to estimate safety stock levels in a production-distribution system, and show how the timing of forecast variability resolution could impact system performance. Using the MMFE, Güllü (1996) analyzes the value of this variability reduction in a capacitated single item system with uncorrelated demands. Graves et al. (1998) demonstrate this uncertainty resolution by defining the i period forecast error as the difference between the actual demand in a period and the forecast of this demand made i periods earlier. They show that the variance of the forecast error over the forecast horizon matches the variability of the demand process, and is equal to the trace of the covariance matrix of forecast errors. This variance decreases as time goes by and i becomes smaller.

This generalized demand model has proven useful in modeling several operational decisions. Toktay and Wein (2001) use the MMFE in a single item discrete-time continuous-state queue to analyze a stationary production-inventory system and develop a class of forecast corrected base stock policies. Aviv (2001) uses the MMFE to explore the effect of

collaborative forecasting in a two stage supply chain consisting of a retailer and a supplier. Gallego and Özer (2001) study an inventory model with advance demand information that can be viewed as a special case of the MMFE. Dong and Lee (2003) adopt the MMFE to show that the structure of the optimal stocking policy proposed by Clark & Scarf (1960) holds under time correlated demand processes. Iida and Zipkin (2006) develop a dynamic forecast-inventory model with forecast updates based on the MMFE and propose a technique to obtain approximately optimal inventory policies. Lu et al. (2006) study a periodic-review inventory system with the MMFE, develop a class of tractable bounds on the optimal base stock levels and use them to construct near-optimal policies. Chen and Lee (2009) show that many commonly used time-series demand models in the literature such as the general ARIMA model (Box and Jenkins 1970) and the linear state-space model (Aviv 2003) can be interpreted as special cases of the MMFE model.

Common to most of the above applications, however, is the use of demand information updates leading to the evolution of demand forecasts. This paper contributes by describing the evolution of the dependency between demands in consecutive periods as new information becomes available under both additive and multiplicative forecast updating.

2.3 Forecast Evolution Model

In this section we describe the concept of evolution of demand forecasts. In each period the demand forecasts are updated based on the most recent information. Eventually, this sequence of forecasts for a specific future period evolves into the realized demand in that period once the demand is observed. In order to capture the dynamic nature of the forecasts

and the underlying forecasting system, we need to describe this periodic modification activity with a probabilistic model.

We assume that demand forecasts are available for some number of periods in the future, which will be referred to as the forecast horizon. In each period these forecasts are updated on a rolling horizon basis. Let H be the number of periods in the forecast horizon and $D_{s,t}$ the demand forecast made at period s for period t. At the end of period s, forecasts $D_{s,t}$ are generated for periods $s \leq t \leq s + H$. The relation $D_{t,t} = D_t$ denotes the realized demand in period t since the forecast is made after the actual demand is revealed. The demand forecasts for periods $t > s + H$ are set equal to a constant μ. When time advances to the next period, period $s + 1$, additional information becomes available and new demand forecasts are generated. Heath and Jackson (1994) describe this forecast evolution by modeling the evolution of the forecast updates. They develop two classes of models for the behavior of forecast updates: the additive model and the multiplicative model.

In Sections 2.4 and 2.5, we first introduce these models of forecast updating as a model of the evolution of the conditional mean demand. We then propose our model of conditional covariance evolution to take into account the dependencies between consecutive demands over time.

2.4 The Additive Model

This model assumes that the size of the forecast updates is unrelated to the size of the forecasts. Let $\varepsilon_{s,t}$ be the random variable denoting the forecast update made at the end of period s for period t (where $s \leq t$) given by

$$\varepsilon_{s,t} = D_{s,t} - D_{s-1,t}.$$

Let $\varepsilon_s = (\varepsilon_{s,s}, \varepsilon_{s,s+1}, \ldots, \varepsilon_{s,s+H})$ denote the forecast update vector received at the end of period s. Therefore $\varepsilon_{s,s+H}$ is the first update made to μ to form the *H*-period ahead forecast and $\varepsilon_{s,s}$ the final update that determines the actual demand.

2.4.1 Evolution of the Conditional Mean

The MMFE assumes that forecasts represent the conditional expectations of demand given all available information at the time the forecast is made, i.e., $D_{s,t} = E[D_t|\mathcal{F}_s]$ where $\mathcal{F}_s$ is a σ-field describing the information available at the end of period s such that $\mathcal{F}_s \subseteq \mathcal{F}_{s+1}$. This implies that the successive forecasts of demand for period t, $\{D_{s,t}, s \leq t\}$, form a martingale such that for all s and t with $s \leq s' \leq t$ we have $E[D_{s',t}|\mathcal{F}_s] = D_{s,t}$. Thus the unconditional mean of the demand process $E[D]$ is constant and equal to μ. The conditional mean $D_{s,t}$ evolves over time following the expression

$$D_{s,t} = D_{s-1,t} + \varepsilon_{s,t}. \quad (1)$$

Heath and Jackson (1994) assume that the update vectors ε_s are independent, identically distributed, multivariate normal random vectors with mean 0 and covariance matrix $\sum = (\sigma_{ij})_{i,j=0,1,\ldots,H}$, where σ_{ij} represents the covariance of $\varepsilon_{s,s+i}$ and $\varepsilon_{s,s+j}$. The forecast updates thus are unbiased and uncorrelated with updates that occur in other periods, i.e., $E[\varepsilon_{s,t}\varepsilon_{s',t'}] = 0, \ \forall s \neq s', \forall t, t'$.

2.4.2 Evolution of the Conditional Covariance

In the previous section we noted that the progressive realization of the uncertain demands through successive forecast updates leads to evolution of the forecast, i.e., the conditional mean demand, given the information available at the time of the forecast's generation, for the corresponding periods over time as shown in Equation (1). If forecast updates over consecutive time periods are dependent, the progressive realization of these updates also results in an evolution of the dependency between consecutive demands over time.

To capture this evolution, we note the distinction between the unconditional covariance and the conditional covariance of a demand process. The unconditional covariance is constant, but the conditional covariance changes over time because it depends on the history of the process until the point in time for which the forecast is made.

Under the additive MMFE assumptions, demand in period t can be represented as the mean of the demand plus the sum of the forecast updates in the last H periods:

$$D_t = \mu + \sum_{j=0}^{H} \varepsilon_{t-H+j,t}. \tag{2}$$

Toktay and Wein (2001) note the following relationship between the covariance matrix of forecast updates and the unconditional covariance between demands:

$$\gamma_i = Cov(D_t, D_{t+i}) = \sum_{j=0}^{H-i} \sigma_{j,i+j}\,, \quad i = 0,1,\dots,H. \tag{3}$$

Figure 2.1 illustrates this concept for $H = 3$. This result states that the unconditional variance of demand equals the sum of the diagonal elements of the covariance matrix, and the unconditional lag i covariance is equal to the sum of the elements on the *i*th off diagonal.

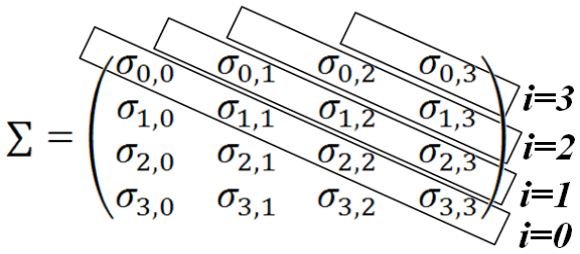

Figure 2.1 Unconditional pairwise covariances of demand showing the components of γ_i for different values of *i*

However, this expression for unconditional covariance of demands ignores the available demand information that forecast updates provide. We now incorporate this information using the concept of conditional covariance.

Proposition 1: *Under the additive MMFE assumptions, the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$, i.e., given all information available at time s, is*

$$Cov(D_t, D_{t+i}|\mathcal{F}_s) = \sum_{j=1}^{t-s} \sigma_{t-s-j,t+i-s-j}\,, \quad 1 \le t-s \le H-i+1.$$

Proof: Let s denote the current period and $D_m = g_m + a_m$ represent the demand for a future period m where the random variables g_m and a_m denote the known and unknown portions of D_m at time *s*, respectively. Thus we have

$$D_m = g_m + a_m = \left(\mu + \sum_{j=0}^{H-(m-s)} \varepsilon_{m-H+j,m}\right) + \left(\sum_{j=H-(m-s)+1}^{H} \varepsilon_{m-H+j,m}\right).$$

Figure 2.2 illustrates the known and unknown portions of D_t and D_{t+i}. The leftward solid arrow corresponds to g_m while a_m is represented by the rightward dashed arrows.

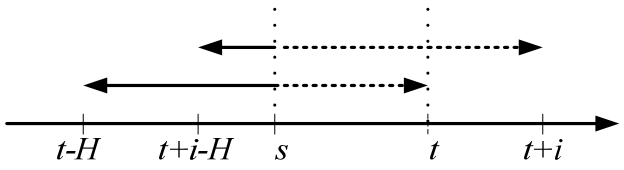

Figure 2.2 Known and unknown portions of D_t and D_{t+i}

Under the additive MMFE assumption, the forecast updates $\{\varepsilon_{m-H+j,m}: j = 1, \dots H\}$ are uncorrelated normal random variables. Thus, a_m and g_m are mutually independent normal random variables. Given these definitions, the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$, i.e., all information available at time *s*, is

$$Cov(D_t, D_{t+i}|\mathcal{F}_s) = \mathrm{C}ov\big((g_t + a_t), (g_{t+i} + a_{t+i})\big|\mathcal{F}_s\big).$$

Since a_m is independent of g_m and the information in $\mathcal{F}_s$, we have

$$Cov(D_t, D_{t+i}|\mathcal{F}_s) = \mathrm{C}ov(a_t, a_{t+i}) + \mathrm{C}ov(g_t, g_{t+i}|\mathcal{F}_s).$$

Because g_t and g_{t+i} are known at time s, the second term on the right is equal to zero; thus we have

$$\begin{aligned} Cov(D_t, D_{t+i}|\mathcal{F}_s) &= Cov(a_t, a_{t+i}) \\ &= E\left[\left(\sum_{j=H-(t-s)+1}^{H} \varepsilon_{t-H+j,t}\right)\left(\sum_{j=H-(t+i-s)+1}^{H} \varepsilon_{t+i-H+j,t+i}\right)\right] \\ &= E\left[\left(\sum_{j=1}^{t-s} \varepsilon_{s+j,t}\right)\left(\sum_{j=1}^{t+i-s} \varepsilon_{s+j,t+i}\right)\right] \end{aligned}$$

$$= \sum_{j=1}^{t-s} E\left[\varepsilon_{s+j,t}\varepsilon_{s+j,t+i}\right]$$

$$= \sum_{j=1}^{t-s} \sigma_{t-s-j,t-s+i-j} \,. \quad \blacksquare$$

This proposition states that the conditional variance of demand in a given future period equals the partial sum of the diagonal elements of the covariance matrix, and depends on how far in the future that period is relative to the current period. The further in the future the period of interest is, the more elements on the diagonal are considered. After passing the boundary of the forecast horizon, the partial sum is replaced by the full sum. This is also true for the conditional lag i covariances.

Let $\gamma_{s,(t,t+i)}$ denote the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$. Proposition 1 implies that $\gamma_{s,(t,t+i)}$ evolves over time following the relation

$$\gamma_{s,(t,t+i)} = \gamma_{s-1,(t,t+i)} - \sigma_{t-s,t+i-s}\,, \tag{4}$$

where $\sigma_{t-s,t+i-s}$ is the covariance between $\varepsilon_{s,t}$ and $\varepsilon_{s,t+i}$. This evolution for a forecast horizon of $H = 3$ periods is illustrated in Figure 2.3. The conditional covariances between D_t and D_{t+i} for $s < t + i - H$ are set equal to γ_i. When time advances to period s such that $s = t + i - H$, additional information $\varepsilon_{s,t}$ and $\varepsilon_{s,t+i}$ becomes available. Thus the covariances between these variables, the elements in the shaded area, are removed from the covariances between D_t and D_{t+i}. Since the elements of the covariance matrix can be positive or negative, the resolution of forecast updates may cause the conditional covariances to increase or decrease over time. When $s = t$, D_t is totally revealed; all covariance matrix elements are

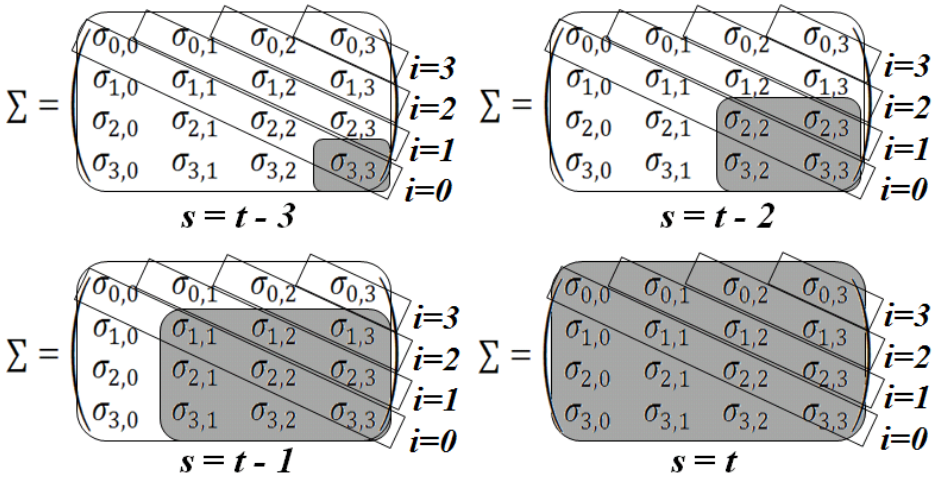

Figure 2.3 Conditional covariance evolution

removed and thus $\gamma_{s,(t,t+i)} = 0$.

This result implies the dependencies between demands are not constant over time, but rather depend on the point in time at which dependencies are computed. This additional evolution gives the decision maker a better characterization of demand that can lead to more effective decisions. The benefits of this approach will be illustrated in Section 2.6.

2.5 The Multiplicative Model

In practice forecast updates are likely to be related to the size of the demand forecasts (Hausman 1969; Heath and Jackson 1994). Thus the variance of updates becomes proportional to the size of the forecasts and it is appropriate to model forecast updates using the multiplicative model. Let $R_{s,t}$ be the random variable given by

$$R_{s,t} = \frac{D_{s,t}}{D_{s-1,t}},$$

where all demand forecasts are strictly positive and $R_s = \left(R_{s,s}, R_{s,s+1}, \ldots, R_{s,s+H}\right)$ denote the forecast ratio vector received at the end of period s. This expression does not satisfy the stationarity condition, but a log transformation can improve the stationarity (Heath and Jackson 1994). A multiplicative model thus can be represented by modeling the natural logarithm of the updates as differences in the logs of the demand forecasts. Thus, the forecast updates in the multiplicative model are given by $\varepsilon_{s,t} = ln\left(D_{s,t}\right) - ln\left(D_{s-1,t}\right)$.

2.5.1 Evolution of the Conditional Mean

If we assume that forecasts represent the conditional expectations of demand given the information available up to the point in time at which they are made, the conditional mean $D_{s,t}$ evolves over time following the expression

$$D_{s,t} = D_{s-1,t} R_{s,t}. \tag{5}$$

Heath and Jackson (1994) assume the successive forecasts of demand for period $t, \{D_{s,t}, s \leq t\}$, form a martingale. Thus the expected value of each future forecast is the same as the current forecast and the expected value of the ratio of these successive forecasts is approximately 1, i.e. $E\left[R_{s,t}\right] \approx 1$. The ratio vector R_s is uncorrelated with R_u for all $u \leq s - 1$. Finally, instead of assuming that forecast differences are jointly normally distributed, we assume that the components of ε_s are jointly normally distributed with mean $\mu_{s,t}$ and variance $\sigma_{s,t}^2$. Therefore, $E\left[exp(\varepsilon_{s,t})\right] = exp\left(\mu_{s,t} + \frac{\sigma_{s,t}^2}{2}\right)$. Note that $R_{s,t} = exp(\varepsilon_{s,t})$. Setting $E\left[exp\left(\varepsilon_{s,t}\right)\right] = 1$ yields $\mu_{s,t} = -\frac{\sigma_{s,t}^2}{2}$. It thus follows that for the multiplicative model, the

vectors ε_s are independent, identically distributed multivariate normal random vectors with the mean of each component $\mu_{s,t} = -\frac{\sigma_{s,t}^2}{2}$.

2.5.2 Evolution of the Conditional Covariance

The progressive realization of forecast updates results in an evolution of the dependency between consecutive demands over time as in the additive model. To model this evolution, we first find the unconditional covariance. Note that under the multiplicative model, demand in period t can be represented as the mean of the demand multiplied by the ratio updates coming from the periods of the forecasting horizon:

$$D_t = \mu \prod_{j=0}^{H} R_{t-H+j,t}$$

$$= exp\left(ln(\mu) + \sum_{j=0}^{H} \varepsilon_{t-H+j,t} \right).$$

Proposition 2: *Under the multiplicative MMFE assumptions, the unconditional covariance between D_t and D_{t+i} is given by*

$$\gamma_i = Cov(D_t, D_{t+i}) = \mu^2 \left(exp\left(\sum_{j=0}^{H-i} \sigma_{j,i+j} \right) - 1 \right), \quad 0 \le i \le H.$$

Proof: It is convenient to work with $D_t = exp(V_t)$ where $V_t = ln(\mu) + \sum_{j=0}^{H} \varepsilon_{t-H+j,t}$. By the assumptions of the multiplicative model, the variables $\{\varepsilon_{t-H+j,t} : j = 1, \dots H\}$ are uncorrelated normal random variables and $\mu_{t-H+j,t} = -\frac{\sigma_{t-H+j,t}^2}{2}$. Thus V_t is normally distributed and we have

$$E[V_t] = ln(\mu) - \frac{1}{2}Var(V_t),$$

yielding

$$E[exp(V_t)] = exp\left(E[V_t] + \frac{1}{2}Var(V_t)\right) = \mu.$$

Given these relations, the covariance between D_t and D_{t+i} can be calculated as

$$\begin{aligned} Cov(D_t, D_{t+i}) &= E[D_t D_{t+i}] - E[D_t]E[D_{t+i}] \\ &= E[exp(V_t + V_{t+i})] - E[exp(V_t)]E[exp(V_{t+i})] \\ &= exp\left(E[V_t + V_{t+i}] + \frac{1}{2}Var(V_t + V_{t+i})\right) - \mu^2 \\ &= exp(2ln(\mu) + Cov(V_t, V_{t+i})) - \mu^2. \quad \blacksquare \end{aligned}$$

By (3), we have

$$Cov(D_t, D_{t+i}) = \mu^2 \left(exp\left(\sum_{j=0}^{H-i} \sigma_{j,i+j}\right) - 1 \right).$$

This result shows that, as in the additive model, the unconditional pairwise covariances of demand depend on the diagonal and off-diagonal elements of the covariance matrix of the forecast updates, respectively. The next proposition derives the conditional covariance between demands to incorporate the time dependency of uncertainty.

Proposition 3: *Under the multiplicative MMFE assumptions, the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$ is*

$$Cov(D_t, D_{t+i}|\mathcal{F}_s) = \left(D_{s,t} D_{s,t+i}\right)\left(exp\left(\sum_{j=1}^{t-s} \sigma_{t-s-j,t-s-j+i}\right) - 1 \right), \quad 1 \le t - s \le H - i + 1$$

Proof: Let s denote the current period and $D_m = exp(g_m + a_m)$ represent the demand for a future period $m = s+1, s+2, ...,$ where the random variables g_m and a_m represent the known and unknown portions of D_m at time *s*, respectively. Thus we have

$$g_m = ln(\mu) + \sum_{j=0}^{H-(m-s)} \varepsilon_{m-H+j,m}$$

and

$$a_m = \sum_{j=H-(m-s)+1}^{H} \varepsilon_{m-H+j,m}.$$

Under the multiplicative model, $\{\varepsilon_{m-H+j,m}: j = 0, ... H\}$ are uncorrelated normal random variables. Thus, a_m and g_m are mutually independent normal random variables. Since a_m is normally distributed we have

$$E\big(exp(a_m)\big) = exp\left(E(a_m) + \frac{Var(a_m)}{2}\right).$$

Furthermore, the mean of each forecast update $\varepsilon_{m-H+j,m}$ equals to the negative of one half of its variance. Thus $E(a_m) = -Var(a_m)/2$ and

$$E\big(exp(a_m)\big) = 1.$$

Demand forecast $D_{s,m}$ can be written as

$$\begin{aligned} D_{s,m} &= E(D_m|\mathcal{F}_s) \\ &= E(exp(g_m + a_m)|\mathcal{F}_s), \end{aligned}$$

where a_m is independent of g_m and the information in $\mathcal{F}_s$ because $m > s$. Thus

$$D_{s,m} = E(exp(g_m)|\mathcal{F}_s)E\big(exp(a_m)\big)$$

$$= E(exp(g_m)|\mathcal{F}_s)$$

Given these relations, the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$ is

$$Cov(D_t, D_{t+i}|\mathcal{F}_s) = Cov(exp(g_t + a_t), exp(g_{t+i} + a_{t+i})|\mathcal{F}_s)$$

$$= E(exp(g_t + a_t).exp(g_{t+i} + a_{t+i})|\mathcal{F}_s)$$

$$- E(exp(g_t + a_t)|\mathcal{F}_s)E(exp(g_{t+i} + a_{t+i})|\mathcal{F}_s)$$

Given $\mathcal{F}_s$, g_t and g_{t+i} are fixed. Thus $E(exp(g_t)|\mathcal{F}_s) = exp(g_t)$, $E(exp(g_{t+i})|\mathcal{F}_s) = exp(g_{t+i})$ and $E(exp(g_t)exp(g_{t+i})|\mathcal{F}_s) = exp(g_t)exp(g_{t+i})$. Since a_m is independent of both g_m and the information in $\mathcal{F}_s$ we have

$$Cov(D_t, D_{t+i}|\mathcal{F}_s)$$

$$= E(exp(g_t)|\mathcal{F}_s).E(exp(g_{t+i})|\mathcal{F}_s).\Big(E\big(exp(a_t).exp(a_{t+i})\big)$$

$$- E\big(exp(a_t)\big)E\big(exp(a_{t+i})\big)\Big)$$

$$= D_{s,t}.D_{s,t+i}\big(E\big(exp(a_t + a_{t+i})\big) - 1\big)$$

$$= D_{s,t}.D_{s,t+i}.\left(exp\left(E(a_t + a_{t+i}) + \frac{1}{2}Var(a_t + a_{t+i})\right) - 1\right)$$

$$= D_{s,t}.D_{s,t+i}.\big(exp(Cov(a_t, a_{t+i}) - 1)\big)$$

By the expression for $Cov(a_t, a_{t+i})$ obtained in Proposition 1, we have

$$Cov(D_t, D_{t+i}|\mathcal{F}_s) = D_{s,t}.D_{s,t+i}.\left(exp\left(\sum_{j=1}^{t-s}\sigma_{t-s-j,t-s-j+i}\right) - 1\right). \quad \blacksquare$$

As expected, the conditional pairwise covariances depend on the partial sums of the diagonal or off-diagonal elements of the covariance matrix, and how far in the future that period is relative to the current period.

Let $D_t = exp(V_t)$ where $V_t = ln(\mu) + \sum_{j=0}^{H} \varepsilon_{t-H+j,t}$ and $\theta_{s,(t,t+i)}$ denote the conditional covariance between V_t and V_{t+i} given $\mathcal{F}_s$. As in the additive model, $\theta_{s,(t,t+i)}$ evolves over time following the relation

$$\theta_{s,(t,t+i)} = \theta_{s-1,(t,t+i)} - \sigma_{t-s,t+i-s} \tag{6}$$

Proposition 3 implies that the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$, i.e. $\gamma_{s,(t,t+i)}$, evolves over time. Using the evolutions of (5) and (6) we have

$$\gamma_{s,(t,t+i)} = D_{s,t}.D_{s,t+i}\left(exp\left(\theta_{s,(t,t+i)}\right) - 1\right).$$

This result shows that as new information becomes available the dependency between D_t and D_{t+i} evolves due to two effects: changes in the demand forecasts D_t and D_{t+i} and changes in the covariance between V_t and V_{t+i}. The effect of the demand forecasts on covariance between demands constitutes the major difference between the additive and multiplicative models.

2.6 Application to an Inventory Model

In this section, we illustrate the benefits of considering the evolution of conditional covariance in addition to that of the conditional mean in a single item multi-period inventory model with constant replenishment lead time. The planning horizon is divided into *T* discrete periods of equal length. Demand in each period is stochastic and may be correlated with demand in other periods. For each variable Z, we define $Z_{s,t}$ to be the prediction made in period s for the value of Z in period t while for $s \geq t$ the predictions equal to the actual value.

At the beginning of period t, demand D_t is observed and forecasts are updated. For simplicity of exposition, the demands D_t are assumed to be the only random variables in the problem; the replenishment process is assumed to be deterministic and uncapacitated within the lead times. Based on these forecast updates, planned orders $Q_{s,t}$ are determined representing the order quantity that is planned to be released into the system at time t based on information available at time s. Orders placed in a given period are completed and put into inventory after a delay of L periods, which represents a fixed positive integer lead time. Demands in each period are satisfied as much as possible. Thus the planned inventories $I_{s,t}$ based on information available at time *s* evolve with the relation

$$I_{s,t} = I_{s,t-1} + Q_{s,t-L} - D_{s,t}.$$

Unsatisfied demand is backlogged. Let h and b denote the unit inventory holding and backorder costs, respectively. To simplify the analysis, we assume that excess inventory can be returned without cost. Since there exists a positive constant lead time, we consider the inventory position as a state variable (Zipkin 2000). Let y_{st} and x_{st} be the planned inventory position at the end of period t after and before ordering, respectively:

$$y_{s,t} = x_{s,t} + Q_{s,t},$$

$$x_{s,t+1} = y_{s,t} - D_{s,t+1}.$$

Let $V_{s,t}(x)$ denote the optimal expected cost from period t to T, given preorder inventory position x and information $\mathcal{F}_s$. Then

$$V_{s,t}(x_t) = \min_{y_t}\{C_{s,t}(y_t) + E[V_{s,t+1}(y_t - D_t)]\},$$

where

$$C_{s,t}(y_t) = h \cdot E_s\left[\left(y_t - \sum_{i=1}^{L} D_{t+i}\right)^+\right] + b \cdot E_s\left[\left(\sum_{i=1}^{L} D_{t+i} - y_t\right)^+\right],$$

is the expected holding and backorder costs charged to period t, given all information available at time *s*. The operator E_s denotes the conditional expectation given information $\mathcal{F}_s$. Given the free return assumption, the myopic policy is optimal (Zipkin 2000). Since the function $C_{s,t}(y_t),\ t = s, \dots, T - s + 1$ is convex in y_t for any given information $\mathcal{F}_s$, a base stock policy is optimal. Setting $C'_{s,t}(y_t) = 0$ yields

$$y^*_{s,t} = G^{-1}_{s,t}\left(\frac{b}{h+b}\right), \tag{7}$$

where $G_{s,t}$ denotes the distribution function of $\sum_{i=1}^{L} D_{t+i}$ conditional on information $\mathcal{F}_s$. This result states that there exists an optimal threshold $y^*_{s,t}$ in each period, and orders should be placed to raise the inventory position up to this threshold. These base stock levels are a percentile of the conditional demand over the lead time for each period.

Proposition 4: *For the additive model, the optimal base stock levels* $y^*_{s,t}$ *and the optimal costs* $C^*_{s,t}$ *for* $t = s, \dots, T - s + 1$ *are*

$$y^*_{s,t} = \lambda_{s,t} + z^*\sqrt{\Delta_{s,t}},$$

$$C^*_{s,t} = (h+b)\phi(z^*)\sqrt{\Delta_{s,t}}$$

where $\lambda_{s,t} = \sum_{i=1}^{L} D_{s,t+i}$, $\Delta_{s,t} = \sum_{j=t+1}^{t+L} \gamma_{s,(j,j)} + 2\sum_{i=1}^{L-1}\sum_{j=t+1}^{t+L-i} \gamma_{s,(j,j+i)}$ *and* $z^* = \Phi^{-1}\left(\frac{b}{h+b}\right)$ *for the standard normal cumulative distribution* Φ *and the standard normal density* ϕ.

Proof: In the additive model, the forecast update vectors are i.i.d. multivariate normal random vectors. The sum $\sum_{i=1}^{L} D_{t+i}$ conditional on information $\mathcal{F}_s$ is a linear combination of

the components of these update vectors. Thus it is a random variable with distribution $N(\lambda_{s,t}, \Delta_{s,t})$, where $\lambda_{s,t} = E(\sum_{i=1}^{L} D_{t+i} \,|\mathcal{F}_s) = \sum_{i=1}^{L} D_{s,t+i}$ and

$$\Delta_{s,t} = Var\left(\sum_{i=1}^{L} D_{t+i} \,\middle|\, \mathcal{F}_s\right)$$

$$= \sum_{i=1}^{L}\sum_{i=1}^{L} Cov\left(D_{t+i}, D_{t+j} \middle| \mathcal{F}_s\right)$$

$$= \sum_{j=t+1}^{t+L} \gamma_{s,(j,j)} + 2\sum_{i=1}^{L-1}\sum_{j=t+1}^{t+L-i} \gamma_{s,(j,j+i)}.$$

Use z to denote the standardized value of $y_{s,t}$; that is

$$z = \left(y_{s,t} - \lambda_{s,t}\right)/\sqrt{\Delta_{s,t}}.$$

From (7), $z^* = \Phi^{-1}\left(b/(h+b)\right)$ and thus $y^*_{s,t} = \lambda_{s,t} + z^*\sqrt{\Delta_{s,t}}$.

In order to find the expected cost, we first recall some basic facts about the normal distribution. Let ϕ denote the standard normal density function, Φ the standard normal cumulative distribution function, Φ^0 the standard normal complementary cumulative distribution function, and Φ^1 the standard normal loss function, that is,

$$\Phi^1(z) = \int_z^{\infty} (x - z)\phi(x)dx.$$

The following relations are useful:

$$\Phi^1(z) = -z\Phi^0(z) + \phi(z), \tag{8}$$

$$\Phi^1(-z) = z + \Phi^1(z). \tag{9}$$

The conditional expectation of backorder $\overline{B_{s,t}}$ and inventory $\overline{I_{s,t}}$ are

$$\overline{B_{s,t}} = E_s\left[\left(\sum_{i=1}^{L} D_{t+i} - y_t\right)^+\right] = \Phi^1(z)\sqrt{\Delta_{s,t}}\,,$$

$$\overline{I_{s,t}} = E_s\left[\left(y_t - \sum_{i=1}^{L} D_{t+i}\right)^+_t\right] = E_s\left[y_t - \sum_{i=1}^{L} D_{t+i} + \left(\sum_{i=1}^{L} D_{t+i} - y_t\right)^+\right]$$

$$= y_{s,t} - \lambda_{s,t} + \overline{B_{s,t}}$$

$$= \left[\frac{y_{s,t} - \lambda_{s,t}}{\sqrt{\Delta_{s,t}}} + \Phi^1(z)\right]\sqrt{\Delta_{s,t}}$$

$$= \Phi^1(-z)\sqrt{\Delta_{s,t}}.$$

The last equality uses (9). The conditional optimal cost in each period is

$$C^*_{s,t} = h\Phi^1(-z^*)\sqrt{\Delta_{s,t}} + b\Phi^1(z^*)\sqrt{\Delta_{s,t}}$$

$$= (h+b)\left[\Phi^1(z^*) + \left(\frac{h}{h+b}\right)z^*\right]\sqrt{\Delta_{s,t}}$$

$$= (h+b)\phi(z^*)\sqrt{\Delta_{s,t}}.$$

The second equality uses (9), while the last uses the definition of z^* and (8). ■

This proposition states that under additive forecast updating, for each period of the planning horizon, the order-up-to levels are the sum of two terms: the conditional expectation of the demand over the lead time, which captures the impact of the observed information, and a safety stock which shows the impact of the as yet unobserved information. Hence the base stock levels are a function of the conditional covariance between demands over the lead time, inventory holding and backorder costs.

Furthermore, under the additive model, we obtain the optimal costs $C^*_{s,t}$. Note that $C^*_{s,t}$ is the product of three terms. The first term $(h + b)$ measures the overall magnitude of the cost coefficients. The second term $\left(\phi(\, z^*)\right)$ depends on the relative costs of inventory and backorders. These two terms summarize the economics of the problem. The last term $\left(\sqrt{\Delta_{s,t}}\right)$ is again a function of the conditional variance of lead time demand. It depends on the dynamics of unobserved forecast updates over the lead time.

Notice that Proposition 4 holds for the additive case. When forecast updates are multiplicative, demands have a multiplicative lognormal distribution and the lognormal sum distribution $G_{s,t}$ does not have a closed-form expression. We discuss the multiplicative case further in Section 2.8.

Intuitively, the optimal safety stock for the *i*-th period in the planning horizon will be lower the closer period *i* is to the current period. The following proposition establishes this relationship.

Proposition 5: *For the additive model, the conditional variance of the lead time demand for period t observed in period s, i.e. $\Delta_{s,t}$ is monotonically non-increasing as s approaches t.*

Proof: Let $A(\alpha)$ denote the principal submatrix of a square matrix A whose entries are in the intersection of the rows and columns of A specified by α. When time advances to period s, new forecast updates for the demands in the forecast horizon become available and the covariances between these updates are removed from the variance of the lead time for period t. By Equation (4) and Proposition 1, these covariances are all entries of the principal submatrix $\sum(\alpha_{s,t})$ where $\alpha_{s,t} = \{i \in \{t + 1 - s, \dots, t + L - s\}; 1 \leq i \leq H + 1\}$. Since the

covariance matrix $\sum$ is positive semidefinite, any principal submatrix $\sum(\alpha_{s,t})$ is positive semidefinite and thus the sum of all entries of $\sum(\alpha_{s,t})$ is nonnegative. ■

This proposition demonstrates how the lead time demand uncertainty decreases as new information becomes available through forecast updates. In particular we have

Corollary 1: *The conditional variance over the lead time demand for period* t *is no greater than the unconditional variance.*

2.7 Numerical Experiments with the Additive Model

In this section, we illustrate the behavior of the inventory planning model of the previous section in order to show the benefit of considering information updates from forecast evolution. In particular, our goal is to compare the expected system cost of the inventory model that keeps track of the information provided by evolving forecasts with that of a comparable standard demand model.

When the model does not keep track of the forecasts, it does not utilize the information provided by the forecast evolution mechansim. The forecaster assumes that no advance information is available, i.e. the forecast horizon $H = 0$. The conditional terms are thus replaced with the unconditional ones in Proposition 4. Let $\Delta_0 = L\gamma_0 + 2\sum_{i=1}^{L-1}(L-i)\gamma_i$ denote the unconditional variance of the lead time demand. The base stock level $y_0^* = L\mu + z^*\sqrt{\Delta_0}$ is optimal for all periods of the planning horizon leading to the same expected cost $C_0^* = (h+b)\phi(z^*)\sqrt{\Delta_0}$. These results coincide with the classical dynamic inventory results with stationary costs and demands (Zipkin 2000).

By Proposition 5, the following relation holds between expected costs of periods of planning horizon *T:*

$$C^*_{s,s} \leq C^*_{s,s+1} \leq \cdots \leq C^*_{s,s+H} = C^*_{s,s+H+1} = \cdots = C^*_{s,T} = C^*_0.$$

The optimal expected costs of periods t for $t \geq s + H$ are set equal to C^*_0. When $t < s + H$, additional information is available leading to reduction of expected cost of period t. In order to illustrate the effect of uncertainty resolution on the expected cost of each period the following numerical example is considered. We use the additive MMFE demand model (2), and assume that the demand mean and standard deviation are 100 and 27.5 respectively. At each time s the forecasts are updated for the next two periods $(H = 2)$, and the lead time $L = 2$. The forecast updates follow a multivariate normal distribution. The covariance matrix of forecast updates is

$$\begin{pmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & \rho_{02}\sigma_0\sigma_2 \\ \rho_{01}\sigma_0\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_0\sigma_1 \\ \rho_{02}\sigma_0\sigma_2 & \rho_{12}\sigma_0\sigma_1 & \sigma_2^2 \end{pmatrix},$$

where $\sigma_0 = 18.8, \sigma_1 = 15.7,$ and $\sigma_2 = 12.5.$

An important factor is the relative magnitude of the inventory holding costs and the backorder costs. These costs are used to find the corresponding safety factor $z^* = \Phi^{-1}\big(b/(h+b)\big)$. The cost ratio $b/(h+b)$ can be interpreted as the service level for the system. Figure 2.4 illustrates the expected cost in each period of the planning horizon for different cost ratios when demands are independent, i.e. $\rho_{01} = \rho_{02} = \rho_{12} = 0$. The inventory cost is fixed at $h = 1$ and the backorder cost b varied to yield cost ratios between 0.50 and 0.98. As expected, the cost increases in the cost ratio. The differences between the

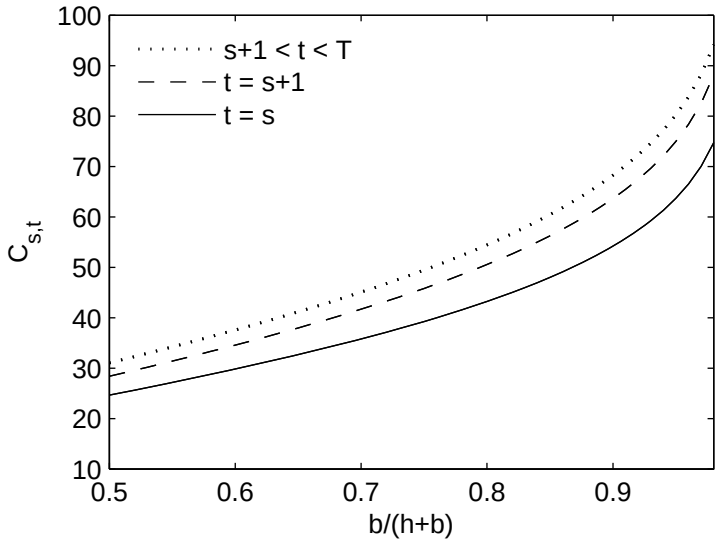

Figure 2.4 Comparison of expected cost in each period of the planning horizon *T* for different cost ratios $\boldsymbol{b/(h+b)}$

curves demonstrate the value of information updating in finding the base stock levels.

In particular, we are interested in the difference between the top and bottom curves. This difference gives us the cost reduction due to the keeping track of forecast updates. The relative first-period cost reduction is given by

$$k = \frac{C_0^* - C_{s,s}^*}{C_0^*} = \sqrt{\frac{\Delta_0 - \Delta_{s,s}}{\Delta_0}}$$

which suggests that the cost reduction is due to the fraction of the lead time variability resolved by using the evolution-based models. This reduction is always nonnegative as the conditional lead time variance $\Delta_{s,s}$ is no greater than the unconditional lead time variance Δ_0 by Corollary 1. At one extreme if demands can be predicted perfectly L periods in advance then the ratio k is one. On the other hand, in a situation where no prediction can be performed in advance the ratio k is zero as expected. The cost reduction for the different

service levels in Figure 2.4 is fixed at 20.6%. We are interested in understanding how the cost reduction is affected by different parameters of the model.

Correlation. Proposition 4 implies that positive (negative) correlations increase (decrease) the base-stock levels and the expected costs in each period as long as the correlation contributes to the corresponding conditional covariances. Furthermore, Propositions 1 and 4 imply that elements of the first $min\{H+1, L\}$ diagonals of covariance matrix contribute to the base stock level of the first period of the planning horizon. Thus management can ignore advance information beyond the lead time. The elements of the first $min\{H+1, L\}$ diagonals can be divided to two groups. The first group is the elements in the conditional covariances contributing to the expected cost. Since C_0^* and $C_{s,s}^*$ increases with them, the relative cost reduction is higher as those elements decreases. The second group is the elements that do not contribute to any of the conditional covariances that determine the optimal base stock level for the first period. While these elements do not impact $C_{s,s}^*$, they affect C_0^*. Thus the difference between the costs increases as the related correlations approach 1.

To illustrate the impact of correlation on the cost reduction, the previous example is considered. Since $L = 2$ the value of ρ_{02} does not affect the cost reduction. Figure 2.5 shows the cost reduction where ρ_{01} varies from -1 to 1 and ρ_{12} is kept at fixed values $-1, 0$, and 1. ρ_{01} contributes to the conditional covariances while ρ_{12} does not and thus the relative cost reduction k increases as ρ_{01} and ρ_{12} approaches -1 and 1, respectively. As we can see from the figure the value of using advance information varies quite widely. In one case, where $\rho_{01} = -1$ and $\rho_{12} = 1$ the cost reduction is as high as 47.4%. By switching the values of ρ_{01}

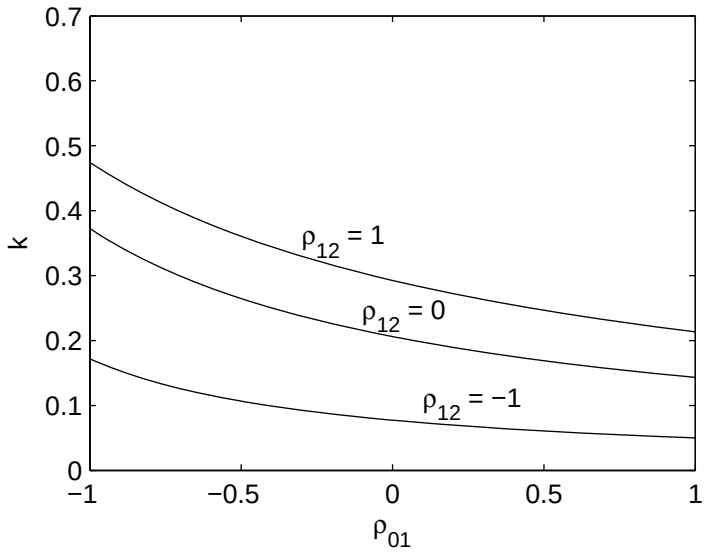

Figure 2.5 Impact of correlation on cost reduction *k*

and ρ_{12} this reduction goes down to 5.0%.

Timing of uncertainty resolution. The importance of incorporating information updating can be even larger when the decision maker obtains more advance information. Since $\sigma_0 > \sigma_2$, by reversing the order of the diagonal and off-diagonal elements in the covariance matrix, we construct an example where uncertainty is resolved earlier in the forecast horizon. Figure 2.6 shows the impact of the correlation on the early uncertainty resolution. The pattern is similar to that of Figure 2.5, but with higher cost reductions between 68.8% and 14.9%, compared to reductions between 47.4% and 5.0% where uncertainty was resolved later.

Lead time. We now explore the effects of the lead time on the cost reduction. Table 2.1 shows the cases where the lead time *L* varies from 1 to 4 and the values of all correlation coefficients are kept at fixed at 1. This computational study represents that as the lead time *L*

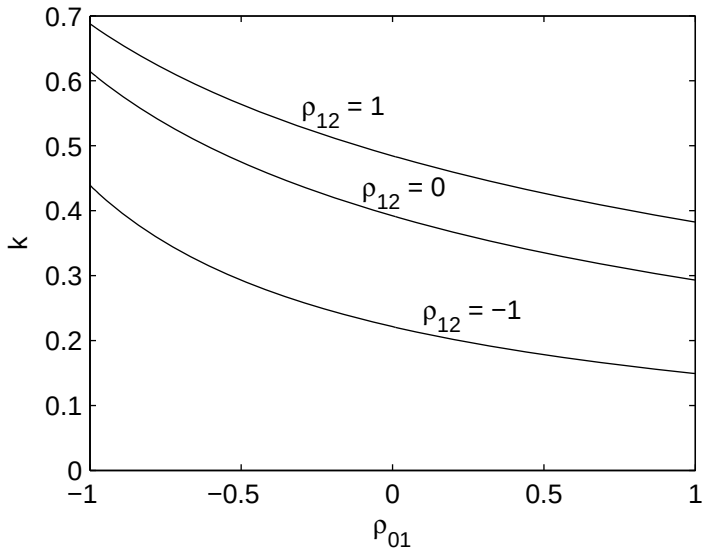

Figure 2.6 Impact of correlation on cost reduction *k*: Early Uncertainty resolution

Table 2.1 Impact of the lead time

L	C_0^*	$C_{s,s}^*$	k
1	45.51	66.57	31.6%
2	95.12	120.93	21.3%
3	148.30	166.04	10.7%
4	186.92	201.29	7.1%

increases, the impact of the information updates decreases due to the increase in the unknown portion of the lead time demand.

2.8 Extension to the Multiplicative Model

For the base stock policy explained in Section 2.6, we need to compute the percentile of the cumulative demand over the lead time. In the multiplicative model, since demands follow a multivariate lognormal distribution, the lead time demand consists of the sum of

lognormal random variables. The distribution of the sum of lognormals is known to have no closed-form. However, it has been recognized that the lognormal sum can be well approximated by a new lognormal variable. Several methods have been proposed to estimate the moments for the sum of independent (Fenton 1960; Beaulieu et al. 1995) and correlated random variables (Safak 1993; Mehta et al. 2007).

These methods are based on matching either the linear moments or the moments in the log-domain. Beaulieu et al. (1995) have studied the accuracy of several methods and shown that all the methods have their own advantages and disadvantages. The methods based on matching linear moments model the tail portion of the lognormal sum PDF more accurately, which is important in finding the percentile of demand over the lead time. More details have been explored in Mehta et al. (2007). Therefore, the approximation method based on matching the linear moments is used to find the percentile of the lead time demand.

First, the conditional sum of lognormal random variables $\sum_{i=1}^{L} D_{t+i} | \mathcal{F}_s$ is approximated by a conditional single lognormal random variable $e^{Z_t} | \mathcal{F}_s$, where Z_t is a normal random variable. We then match the mean $\lambda_{s,t} = \sum_{i=1}^{L} D_{s,t+i}$ and the variance $\Delta_{s,t} = \sum_{j=t+1}^{t+L} \gamma_{s,(j,j)} + 2 \sum_{i=1}^{L-1} \sum_{j=t+1}^{t+L-i} \gamma_{s,(j,j+i)}$ of the conditional sum to obtain the appropriate parameters for $e^{Z_t} | \mathcal{F}_s$, which are

$$\mu_{Z_t|\mathcal{F}_s} = ln\left(\lambda_{s,t}\right) - ln\left(1 + \Delta_{s,t}^2 / \lambda_{s,t}^2\right)/2,$$

$$\sigma_{Z_t|\mathcal{F}_s}^2 = ln\left(1 + \Delta_{s,t}^2 / \lambda_{s,t}^2\right).$$

Thus, from (7), for $t = s, \ldots, s + T - 1$ we have

$$y_{s,t}^* = exp\left(\mu_{Z_t|\mathcal{F}_s} + z^* \sigma_{Z_t|\mathcal{F}_s}^2\right)$$

$$= \lambda_{s,t} exp(z^*) \sqrt{1 + \Delta_{s,t}^2 / \lambda_{s,t}^2}.$$

Similar to the additive model, in the case of multiplicative forecast updating, the base stock levels are functions of conditional mean and covariance of demand over the lead time.

2.9 Conclusions and Future Directions

In this paper we show that the progressive resolution of demand uncertainty through forecast updates results in evolution of the conditional covariance of the demand in addition to the evolution of its conditional mean. We demonstrate that these conditional covariances fluctuate over time until they become zero at the time of full realization of demand. This additional evolution yields a better characterization of demand behavior that leads to more effective decisions. The benefit of this approach is analyzed in a multi-period inventory planning model with additive forecast updates. We demonstrate how the system costs are reduced when the information updates are incorporated explicitly in the planning model. The model indicates that it is the relative difference between unconditional and conditional demand variance over the lead time that determines the magnitude of the cost reduction. It also shows that management can ignore advance information beyond the lead time in this uncapacitated system. Our results can also be used to quantify the benefits of more advance information (earlier uncertainty resolution) that helps manager in their strategic decisions. We also show that how our results can be extended for the multiplicative forecast updates by a mild assumption.

A natural extension of this work would be to study the evolution of the conditional covariance for the multi-product model. Exploring this issue can reveal important insights

about the dependency evolution within demands of different products and across time periods. Another possible extension would be to implement the approach in dynamic multi-period models with load-dependent lead times and capacity constraints.

Chapter 3

Production Planning with Load Dependent Lead Time and Uncertain Demands

Abstract

While queuing models of production systems show that lead times increase nonlinearly with resource utilization, there are few models that analyze the relationship between resource utilization, lead times and service levels in the context of stochastic demand. We consider a single stage production-inventory system with workload dependent lead times and uncertain demands. The dependency between workload and lead times is captured through clearing functions that take into account the nonlinear relationship between utilization and lead times. We propose a load dependent release policy leading to a tractable chance constrained optimization model. We compare the performance of this approach to that of a multistage stochastic programming model by subjecting them to a simulation of uncertain demand realizations. Computational experiments show promising performance of our chance constrained model.

3.1 Introduction

Lead time, the time between an order being placed and being delivered to the customer or to final inventory, is an important determinant of global competitiveness. Long lead times increase inventory costs due to high work in process (WIP) inventory levels as well as increased safety stocks caused by increased uncertainty about demand. They are thus important to effective production planning: the decision as to how to release work into production facilities over time must take lead times into account in some manner in order to match the output of the production facility with market demand in some optimal way. However, in most production planning and inventory models, the lead time is not generally modeled in detail. It is commonly treated as an exogenous parameter in deterministic optimization models and many inventory models (Hackman and Leachman 1989; Zipkin 2000, Chapter 6) or as an exogenously defined probability distribution when it is taken to be stochastic (Zipkin 2000, Chapter 7). In particular, lead times can arise from congestion effects internal to the operation of a production system whose resources are governed by queuing behavior. When the production system operates at high levels of resource utilization the primary cause of delay may be queuing.

Queuing models of production systems (Buzacott and Shanthikumar 1993) show that lead times increase nonlinearly with resource utilization, which in turn is determined by the work release plan produced by the planning system. Additionally, uncertain demand is a fact of life in most production systems, affecting the work release decisions made by planning models and hence capacity allocations and lead times. It is important for many manufacturing systems with uncertain demands, such as those encountered in semiconductor manufacturing,

to run at high utilization to be profitable. Under these conditions, small fluctuations in utilization may cause large changes in lead times. However, there are few production planning models linking order releases and planning decisions to lead times under demand uncertainty.

The intent of this paper is to present models that consider both workload dependent lead times and stochastic demand in an integrated manner. We model a single item production-inventory system using the clearing function formulation, developed by (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989) to capture the dependency between workload and lead times.

This problem can be formulated as a dynamic program, but the size of the resulting state space renders an exact solution computationally prohibitive. Approximate methods must therefore be used to reduce the complexity of the decision process. We first separate the problem into several single periods. Then a production policy is specified that provides appropriate protection against demand uncertainty under the existence of a clearing function, leading to the determination of safety stocks. Once this is done, the problem is solved as a multi-period decision problem under uncertainty. We investigate the performance of this approach compared to that of a multistage stochastic programming model by subjecting them to simulations of uncertain demand realizations.

In the following section we provide a brief review of the relevant literature on production planning. Section 3.3 describes the basic concept of the clearing function used to represent congestion in the production system. In Section 3.4, we propose the new chance constrained model integrating considerations of stochastic demand with a clearing function to

capture workload dependent lead times. The stochastic programming model used as a benchmark in the computational experiments is also presented in this section. We show promising performance of our chance constrained model through computational experiments in Section 3.5. A discussion of the main conclusions and some directions for future research conclude the paper.

3.2 Literature Review

There have been several attempts in the literature to incorporate lead time into production systems. Missbauer and Uzsoy (2010) give a comprehensive review of this area. The simplest one is to treat lead time as a fixed, exogenous quantity independent of system workload. The Materials Requirements Planning (MRP) approach uses this fixed lead times in its backward scheduling step to determine job releases (Orlicky 1975). Linear programming (LP) models are another common approach to production planning under fixed lead times (Johnson and Montgomery 1974; Hackman and Leachman 1989). In these models, capacity is treated as a fixed upper bound on the amount of a resource that can be used in a time period. However, proposed plans of these models may be impossible to execute since the lead times are independent of the workload and thus the operational dynamics are not modeled explicitly.

A number of authors have proposed enhanced models that consider the dependency between lead times and resource utilization. Lautenschläger and Stadtler (1998) propose a model where lead times are captured by allowing the production in a given period to become available over several future periods. Voss and Woodruff (2003) suggest a nonlinear model

where the function linking lead time to workload is approximated using piecewise linearization. Ettl et al. (2000) follow a similar approach, adding a nonlinear term representing the cost of carrying WIP as a function of workload to the objective function. Graves (1986), Karmarkar (1989), Missbauer (2002), and Asmundsson et al. (2006) use nonlinear clearing functions to model the dependency between workload and lead times. We will discuss clearing functions, which are used in the models in this paper, more extensively in the next section. Reviews of production planning models with load dependent lead times are given by Pahl et al. (2005) and Missbauer and Uzsoy (2010).

Another approach to modeling the operational dynamics of the system is the use of detailed simulation or scheduling models in the planning process (Dauzere-Peres and Lasserre 1994; Pritsker and Snyder 1997). While these methods can capture the operational dynamics of the system correctly, they do not scale well, since simulation models of large systems are time-consuming to run and analyze. An innovative approach to integrating simulation and LP is that of Hung and Leachman (1996). Given initial lead-time estimates, an LP model for production planning is formulated and solved. The resulting plan is fed into a simulation model to estimate the lead-times the plan would impose on a real system. If these lead-times do not agree with those used in the LP, the LP is updated with the new lead-time estimates and resolved. This iteration is repeated until convergence. Similar models have been proposed by others (Byrne and Bakir 1999; Kim and Kim 2001; Riaño 2003; Byrne and Hossain 2005). However, the convergence of these methods is not well understood (Irdem et al. 2010; Kacar et al. 2010). The computational burden of the

simulation runs required is also a significant disadvantage for large systems such as those encountered in semiconductor manufacturing.

There is a vast literature on the production planning under uncertainty, as reviewed by Mula et al. (2006). One stream of research is stochastic inventory models seeking an optimal inventory policy (when to order, and how much to order) for individual items in the face of different environmental conditions and constraints. Much of the work in this area assumes that a supplier can supply any amount of material within the specified lead time, i.e., has unlimited capacity (Zipkin 2000). Federgruen and Zipkin (1986a; 1986b) consider the capacitated inventory problem with uncertain demand and explore the optimality of "modified" base stock policies when the cost for the single period is convex in the base stock level. Tayur (1993) extends this work by discussing the computation of the optimal base stock level. Ciarallo et al.(1994) describe the structure of optimal policies for problems with uncertain production capacity and a time-stationary demand distribution. Anupindi et al. (1996) provide bounds and heuristics for the problem with nonstationary demand and stochastic lead times, where the lead time distribution is stationary over time. However, these models use simple capacity constraints that ignore the dependency between load and lead times.

The idea of combining inventory and queuing models has attracted attention from many researchers (Buzacott and Shanthikumar 1993; Rao et al. 1998; Hopp and Spearman 2001). Zipkin(1986) develops a queuing framework to analyze supply chains facing a stationary demand distribution and where a *(Q,r)* policy is used to release units onto the shop floor. Ettl et al.(2000) develop an optimization model combining queuing and inventory

models to set lot sizes for a multi-item batch production system facing non-stationary demands. Liu et al.(2004) extend this approach.

Chance constrained programming was first introduced by Charnes and Cooper (1959) that allows constraints to be violated with a certain probability. However, the cost of violating the constraint is not considered in the models. Many authors have used chance constrained programming to analyze and solve production planning problems (Johnson and Montgomery 1974; Bookbinder and Tan 1988; Tarim and Kingsman 2004). None of these studies addresses workload dependent lead times. Ravindran et al. (2011) provide a production planning model with chance constraints that considers the lead time dependency though clearing functions. Their analysis depends on the assumption that inventory position is a percentile of the demand over the lead time that is variable due to the presence of the clearing function. This model requires the lead times used to establish inventory levels to be provided as an exogenous parameter. In this paper we present a new chance constrained model which considers workload dependent lead times without requiring any external lead time parameters.

One of the most popular frameworks for planning under uncertainty is stochastic programming (Kall and Wallace 1994; Prékopa 1995; Birge and Louveaux 1997). Uncertainty is represented by using a number of discrete scenarios to represent possible future states of nature at each decision epoch (stage), which allows stochastic linear programs to be modeled as large linear programming models. Several papers have explored the use of stochastic programming in production planning (Escudero et al. 1993; Huang and Ahmed 2009; Higle and Kempf 2010). However, the numbers of planning periods and possible

realizations of uncertain parameters make the use of stochastic programming computationally challenging as suggested by Sen and Higle (1999). Aouam and Uzsoy (2012) examine both two-stage and multi-stage stochastic programming models and compare their performance to the chance-constrained models of Ravindran et al. (2011). Their results show that the chance-constrained model performs quite favorably relative to the stochastic programming models with significantly lower computation time requirements.

In the next section we review the clearing function concepts that we use to develop a LP model that captures the load dependent lead time and demand uncertainty aspects simultaneously for a single-product production planning model.

3.3 Clearing Functions

Clearing functions (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989), express the expected output of a capacitated resource over a given period of time as a function of workload of the resource over that period, which in turn, is determined by the amount of work available for the resource to process in that period. We shall use the term “WIP” to denote any reasonable measure of the work in process inventory level, providing specifics when describing the implementation of our models. By applying Little’s Law the average lead time of the production resource were a given WIP level is given by the inverse slope of the clearing function.

Figure 3.1 depicts some possible forms of clearing functions discussed in the literature. The constant level clearing function represents an upper bound on output over the period. Without a lead time constraint, this implies instantaneous production since production

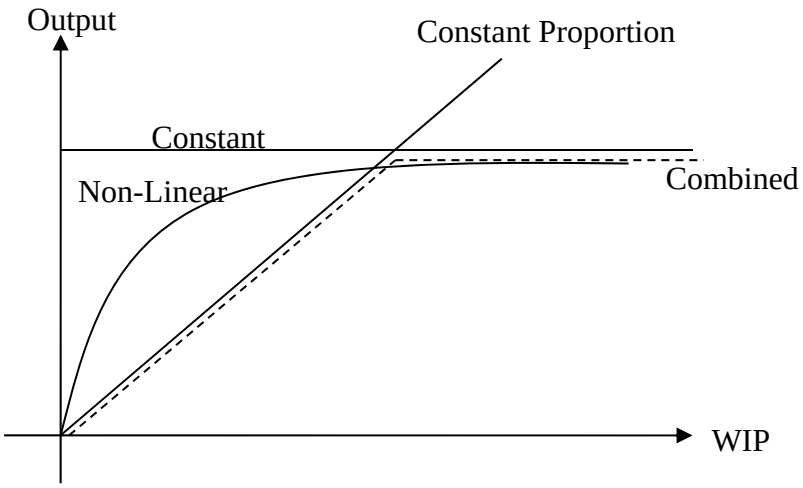

Figure 3.1 Different forms of clearing functions

can occur without any WIP in the system. The constant proportion clearing function represents the control rule given by Graves (1986) which implies infinite capacity with a fixed lead time. This clearing function differs from the workload-independent fixed lead time in most LP models in that the latter does not link output to WIP. Srinivasan et al. (1986) and Karmarkar (1989) independently extended this idea to a concave nonlinear clearing function which relates WIP levels to output such that the lead times are influenced by the workload of the production system and thus is able to capture the behavior of load dependent lead time. An extensive discussion of the derivation of clearing functions is given by Missbauer and Uzsoy (2010).

In this paper we focus on the behavior of these functions under demand uncertainty. Suppose we have a single stage production system producing a single product. The planning horizon is divided into discrete periods. The WIP can be interpreted as WIP at the end of the

period. The output becomes the amount of the production during the period. Let's define the following notation:

I_t Inventory level at the end of period t,

B_t Backorder level at the end of period t,

W_t WIP level at the end of period t,

R_t Release at the end of period t,

X_t Production during period t,

D_t Demand in period t.

The relationships that define the clearing function are given by the following equations which hold for all time periods t:

$$W_t = W_{t-1} + R_t - X_t,$$

$$I_t - B_t = I_{t-1} - B_{t-1} + X_t - D_t,$$

$$X_t = f(W_{t-1}).$$

The first two constraints represent flow conservation for WIP and finished goods inventory (FGI). The third constraint describes the clearing function. The output is the amount of production during the period. We follow Ravindran et al. (2011) in writing our clearing function as a function of W_{t-1}, the amount of work that is available for processing at the start of period *t*. The release variables R_t are defined such that releases are made at the end of period t and thus cannot contribute to production during period *t*. $f(W_{t-1})$ is a concave clearing function that is monotonically non-decreasing in W_{t-1}.

To transform this model into a tractable form Asmundsson et al. (2006) approximate the clearing function using outer linearization. The clearing function is thus given by

$$X_t = min\{a_k W_{t-1} + c_k\}, \ \ \forall k = 1, \dots, K,$$

where a_k is the slope and c_k the intercept of segment k of the piecewise linearized clearing function. In order to formulate the mathematical programming model we need to determine the costs of the system. The objective is to minimize the sum of the expected costs of holding FGI and WIP over the planning horizon. The deterministic LP model is stated as follows:

Minimize $\sum_{t=1}^{T} (wW_t + hI_t + bB_t)$

Subject to

$$W_t = W_{t-1} + R_t - X_t \qquad \forall t,$$

$$I_t - B_t = I_{t-1} - B_{t-1} + X_t - D_t \qquad \forall t,$$

$$X_t \leq a_k W_{t-1} + c_k \qquad \forall t, k,$$

$$W_t, R_t, X_t, I_t, B_t \geq 0 \qquad \forall t.$$

where w, h and b denote the WIP holding, FGI holding and backorder costs, respectively. Note that the clearing function appears as a bound rather than a definitional equality, since it is presumably always feasible to produce less than the maximum quantity possible. The effects of capacity loading are captured through the costs of WIP which, in turn, is required by the clearing model in order for the system to produce output.

An important advantage of clearing functions for our purpose is that lead times do not appear in the formulation; releases and lead times are jointly optimized, allowing the lead time to vary over the planning horizon. Extensive discussions of clearing function models for production planning are given by Asmundsson et al. (2006; 2009), Kacar et al.(2012) and Missbauer and Uzsoy(2010).

\

3.4 Production Planning Models

In this section we consider a single item multi-period production-inventory model under the effects of congestion and demand uncertainty. Assume that the system operates in the following manner. At the start of a period, planned production X_t takes place. Demand D_t is then observed and fulfilled as much as possible. Unsatisfied demand is backordered. At the end of the period inventory holding and backorder costs are charged and a release quantity R_t is determined based on that period's demand and the clearing function. The release quantity R_t, in turn, determines the WIP level W_t which, together with the clearing function, determines the production quantity X_{t+1} in the next period.

3.4.1 Chance Constrained Model (CC)

The chance constrained models assume knowledge of the demand distribution in each period and represent service level requirements with a set of probabilistic constraints that can be violated with a certain probability. While the cost of such violation is not explicitly considered in the model, there is a connection between the required service level and the relative magnitude of holding and shortage costs. We first introduce the chance constrained model of Ravindran et al. (2011) using the following notation:

α Service level,

$G_{[t,t+k]}$ Cumulative distribution function of total demand in periods t through $t+k$,

L_t Average lead time in period t.

Note that classical inventory theory assumes that when there is no fixed ordering cost and holding and shortage costs are linear, it is near-optimal to maintain the inventory

position at a critical percentile of the demand over the replenishment lead time. Here inventory position is defined as the sum of on hand and on-order inventory minus backorders. Assuming that the service level is high enough that backorders will be negligible, they ignore backorders in the formulation. Thus inventory position is given by $I_t + W_t$. This inventory analogy is equivalent to having a service level constraint requiring that inventory position is at least as great as the demand over the replenishment lead time with probability α (Graves 1988). The service level constraint can then be written as

$$P\left(I_t + W_t \geq \sum_{i=t+1}^{t+L_t} D_i\right) \geq \alpha$$

and its deterministic equivalent as

$$I_t + W_t \geq G^{-1}_{[t+1,t+L]}(\alpha),$$

where $G^{-1}_{[t+1,t+L]}(\alpha)$ is the α percentile of cumulative demand from period $t+1$ to period $t+L$. The LP model for this multi-period production planning problem can then be written as follows:

Minimize $\sum_{t=1}^{T}(w_t W_t + h_t I_t)$

Subject to

$W_t = W_{t-1} + R_t - X_t \qquad \forall t,$

$I_t = I_{t-1} + X_t - D_t \qquad \forall t,$

$I_t + W_t \geq G^{-1}_{[t+1,t+L_t]}(\alpha) \qquad \forall t,$

$X_t \leq a_k W_{t-1} + b_k \qquad \forall t, k,$

$W_t, R_t, X_t, I_t \geq 0 \qquad \forall t.$

In this model there are two different lead times at work. One is the estimated lead time L_t used on the right hand side of the chance constraint to determine the inventory position required for the desired service level. The other is the average lead time that is realized in the production system. This lead time is workload dependent and, as we discussed before, is explicitly represented by the clearing function. Ideally, these two lead times should be equal. Assuming the planning periods are long enough for Little's Law to be applicable, this yields a nonlinear constraint. In the interest of tractability, the authors treat the average lead time L_t on the right hand side of the chance constraint in period t as an exogenous parameter. Computational experiments show that although the realized lead time may deviate from the exogenous parameters, the results are still better than base stock models that do not consider the clearing function.

3.4.2 Shortfall Chance Constrained Model (S-CC)

In this section we develop a model to resolve the issue arising from the difference between the load dependent lead time of the clearing function and the exogenous estimated lead time in the right hand side of the service level constraints. Note that in production planning models using clearing functions, the decision emphasis falls on the release variables R_t rather than the production variables X_t. Once a release plan is determined, the output of the facility is determined. We thus need to propose a release policy capturing the clearing function properties.

3.4.2.1 Release Policy

To motivate the proposed release policy, let us first consider an uncapacitated production planning model when we have no forecast of future demand. Then a natural policy is to set the release in each period equal to the demand in that period, i.e. $R_t = D_t$ in order to produce the observed demand as soon as possible.

When the production output is determined by the clearing function, the release policy needs to be changed accordingly. This new policy must link order releases to the load dependent lead times implicit in the clearing function. By applying Little's Law the clearing function can be written in terms of lead times as

$$X_{t+1} = \frac{W_t}{L_t},$$

where L_t denotes the expected lead time for the release R_t determined after observing the demand D_t in the end of period t. To minimize the inventory holding and backorder costs, we seek a release policy R_t that provides enough W_t to produce the observed demand D_t. Thus we have

$$X_{t+1} = D_t.$$

Combining the above relations, we have

$$\frac{W_{t-1} + R_t - X_t}{L_t} = D_t,$$

$$\frac{L_{t-1}X_t + R_t - X_t}{L_t} = D_t,$$

$$R_t = L_t D_t - (L_{t-1} - 1)D_{t-1}.$$

The first term on the right hand side increases the WIP to produce the realized demand D_t while the second term corrects for the existing WIP that was released in the previous period. Note that releases can be negative. The probability of negative releases decreases as the demand variance decreases or the clearing function becomes smoother. In our analysis, we impose a constraint to ensure nonnegative release values.

3.4.2.2 Multi-Segment Clearing Function with no Upper Limit

When there is no upper limit for clearing functions, demand in a given period can always be met by production in the next period, although it may be very expensive to do so. Once we have a horizontal segment, the possibility of not being able to meet the current demand in the next period arises. In this section we assume the clearing function does not have an upper limit. The evolution of the system over time can be written as

$$W_t = W_{t-1} + R_t - X_t,$$

$$I_t - B_t = I_{t-1} - B_{t-1} + X_t - D_t,$$

where

$$R_t = L_t D_t - (L_{t-1} - 1)D_{t-1},$$

$$X_{t+1} = \frac{W_t}{L_t}.$$

Suppose $W_0 = 0$ and $I_0 = S$. If we substitute for R_t and X_t in the inventory balance constraints, we obtain

$$W_t = L_t D_t,$$

$$I_t - B_t = S - D_t,$$

$$S_t = I_t - B_t + \frac{W_t}{L_t}.$$

Our decision variable is S, the base stock level. To make the problem tractable, we assume that WIP and inventory holding cost are equal and thus WIP cost can be removed from the objective function, since the WIP level is determined by the demand D_t and the lead time L_t. We seek the value of S_t that minimizes the sum of the expected inventory holding and backorder costs. We have

$$C(S_t) = hE[I_t(S_t)] + bE[B_t(S_t)]$$

$$= h\int_{x=0}^{S_t}(S_t - x)f_D(x)dx + b\int_{x=S_t}^{\infty}(x - S_t)f_D(x)dx,$$

$$C'(S_t) = h\int_{x=0}^{S_t} f_D(x)dx - b\int_{x=S_t}^{\infty} f_D(x)dx$$

$$= hF_D(S_t) - b\big(1 - F_D(S_t)\big).$$

Setting $C'(S_t) = 0$ yields

$$S_t^* = F_D^{-1}\big(h/(h+b)\big).$$

This result states that the base stock level is a percentile of the distribution function of the demand. The above equations lead to the chance constraint

$$I_t - B_t + \frac{W_t}{L_t} \geq F_D^{-1}\big(h/(h+b)\big),$$

or

$$I_t - B_t + X_{t+1} \geq F_D^{-1}\big(h/(h+b)\big).$$

The left hand side of the new chance constraint represents the available inventory at the beginning of period $t+1$. This policy states that we should release enough material in

period t to have S_t units on-hand at the start of period $t+1$. The chance constrained formulation incorporating this policy is summarized below:

$$\text{Minimize} \quad \sum_{t=1}^{T}(hW_t + hI_t + bB_t)$$

$$\begin{aligned}
\text{Subject to} \quad & W_t = W_{t-1} + R_t - X_t && \forall t, \\
& X_t \le a_k W_{t-1} + c_k && \forall t, \forall k, \\
& I_t - B_t = I_{t-1} - B_{t-1} + X_t - D_t && \forall t, \\
& I_t - B_t + X_{t+1} \ge F_D^{-1}\big(h/(h+b)\big) && \forall t, \\
& W_t, R_t, X_t, I_t \ge 0 && \forall t.
\end{aligned}$$

3.4.2.3 Multi Segment Clearing Function with Upper Limit

We now consider the case where the clearing function has an upper bound on production in a given period. Thus production capacity in a period may not be sufficient to raise the on-hand inventory at the beginning of that period to S if the required production exceeds the upper bound C. The amount by which the target inventory S exceeds the actual on-hand inventory level is called the shortfall. Let the random variable Y_t represent the shortfall in period t, defined as

$$Y_t = S_t - (I_{t-1} - B_{t-1} + X_t),$$

or

$$Y_t = Y_{t-1} + D_{t-1} - X_t,$$

where $Y_{t-1} + D_{t-1}$ is the desired production for period t. Suppose $W_0 = 0$, $Y_0 = 0$ and $I_0 = S_t$. In this case, we obtain

$$W_t = L_t D_t + Y_t,$$

$$I_t - B_t = S_t - (D_t + Y_t).$$

Similar to the previous section we can show that

$$S_t{}^* = F_Q^{-1}\big(h/(h+b)\big),$$

where

$$Q_t = D_t + Y_t.$$

This result states that the base stock level is a percentile of the distribution function of sum of demand and shortfall. The above equations lead to the chance constraint. The above equations yield the chance constraint

$$I_t - B_t + X_{t+1} + Y_{t+1} \geq F_Q^{-1}\big(h/(h+b)\big).$$

The left hand side of the new chance constraint represents the sum of available inventory and shortfall after the production is realized at the beginning of period $t+1$. This new policy states that we should release enough material in period t to raise on-hand inventory and possible shortfall at the start of period $t+1$ to S_t units. The chance constrained formulation incorporating this constraint is as follows:

Minimize $\sum_{t=1}^{T}(hW_t + hI_t + bB_t)$

Subject to

$$W_t = W_{t-1} + R_t - X_t \qquad \forall t,$$

$$X_t \leq a_k W_{t-1} + c_k \qquad \forall t, \forall k,$$

$$I_t = I_{t-1} + X_t - D_t \qquad \forall t,$$

$$Y_t = Y_{t-1} + D_{t-1} - X_t \qquad \forall t,$$

$$I_t - B_t + X_{t+1} + Y_{t+1} \geq F_Q^{-1}\big(h/(h+b)\big) \qquad \forall t,$$

$$W_t, R_t, X_t, I_t \geq 0 \qquad \forall t.$$

The key to analyzing the above optimization problem is to characterize the distribution of $Q_t = D_t + Y_t$. Note that the lead time L_t does not appear directly in this distribution.

Y_t is the shortfall due to the limited production capacity. It can be viewed as the queue length at time *t* of a single server discrete time queue with D_t arrivals and C (potential) services in each period *t*. Glasserman (1997) shows that when demands are independent and the utilization is high, the distribution Y_t can be approximated by an exponential distribution. We have

$$P(Y_t > y) = e^{-\theta(y+\beta)},$$

where $\theta = 2(C - \mu_D)/\sigma_D^2$ and the correction term for normally distributed demand is $\beta = 0.583\sigma_D$. Hence, if demands are independent and normally distributed, we have

$$\begin{aligned} P(Q_t \le q) &= P(Y_t + D_t \le q) \\ &= \int_{x=-\infty}^{q} P(Y_t \le q - x) f_D(x) dx \\ &= \int_{x=-\infty}^{q} \left(1 - e^{-\theta(q-x+\beta)}\right) f_D(x) dx \\ &= \Phi\left(\frac{q - \mu_D}{\sigma_D}\right) - e^{-\theta(q+\beta)} \int_{x=-\infty}^{q} e^{\theta x} f_D(x) dx. \end{aligned}$$

Using integration by parts for the second term, we have

$$P(Q_t \le q) = \Phi\left(\frac{q - \mu_D}{\sigma_D}\right) - e^{-\theta(q+\beta-\mu_D-\frac{1}{2}\sigma_D^2\theta)} \Phi\left(\frac{q - \mu_D - \sigma_D^2\theta}{\sigma_D}\right).$$

Let us approximate the $\Phi(.)$ term by one for large q. Then, Q_t has an exponential distribution

$$P(Q_t \leq q) \approx 1 - e^{-\theta(q+\beta-\mu_D-\frac{1}{2}\sigma_D^2\theta)},$$

$$P(Q_t > q) \approx e^{-\theta(q+\beta-C)}.$$

Thus, for $b \gg h$, S_t^* is well approximated by

$$S_t^* = \frac{1}{\theta} ln(1 + b/h) - \beta + C.$$

This is the main result of this section. We have now obtained a chance-constrained formulation where the right hand side of the service level constraint can be calculated offline and does not depend on any external lead time parameter. We made a series of approximations to obtain this base stock level: the heuristic release policy, the shortfall approximation, and the assumption that $b \gg h$ in estimating $F_Q(q)$. To evaluate the performance of this model, we use develop a multistage stochastic programming model with recourse as a benchmark in the next section.

3.4.3 Stochastic Programming Model

Stochastic programming offers another approach to making decisions under uncertainty. In this approach, the emphasis is placed on the decision to be made at the current time, given present resources, future uncertainties and possible recourse actions that may be taken in the future as uncertainties are realized. It is essential to represent uncertainties in a form suitable for computation. If random variables are represented by continuous distributions, computation is difficult because the models require integration over such variables. To avoid this problem, stochastic programming approximates the probability distribution governing the sources of uncertainty with a small set of discrete scenarios.

3.4.3.1 Scenario Generation

The generation of discrete outcomes for the random variables is called scenario generation. A scenario in this model represents realizations of all random variables in all time periods. It is common to depict the possible scenarios in a form of scenario tree as illustrated in Figure 3.2. It has nodes organized in levels which correspond to decision stages (epochs) $1, \ldots, T$. Each stage denotes a time epoch when new information is available to the decision maker. The nodes n in the tree represent states of the world at a particular point in time. The root node of the tree represents the current state of the world. Each node n of the scenario tree, except the root node $n = 1$, has a unique parent $a(n)$, and each non-leaf node n has one or more children linked to the parent node by arcs that represent realizations of the uncertain variables.

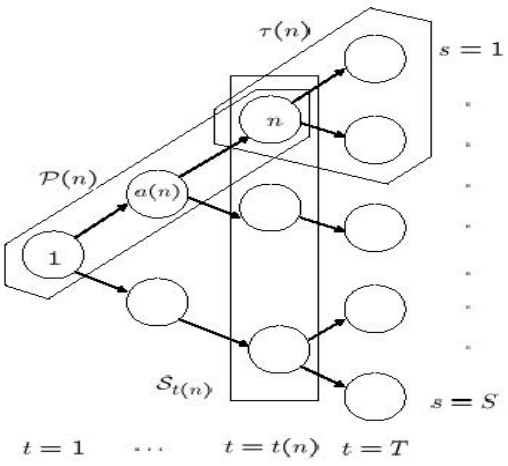

Figure 3.2 Scenario tree

The probability of realization associated with node n is denoted by p_n. The probabilities of the nodes in a time period sum to one and sum of all child nodes of a parent

node is equal to the probability of the parent node. A path from the root node to a node n describes one realization of the stochastic process from the present time to the period where node n appears. If n is a leaf node, the path corresponds to a scenario, and represents a joint realization of the demand over all periods.

3.4.3.2 Multi-stage Stochastic Programming Model

In this section we extend the deterministic model of Section 3.3 to a multi-stage stochastic programming problem (MSP) based on the scenario tree representing the uncertain demand. The decision variables of the deterministic model are the release quantities R_t and production quantities X_t. The inventory I_t, backorder B_t and WIP W_t variables are the consequences of those plans. In MSP, there are *T* stages in the tree, one corresponding to each time period. The decision maker can revise the release and production plans for different demand realizations at each stage of the demand scenario tree after observing the realizations of demand in previous periods. Thus all variables are indexed by the nodes, and the model is as follows:

$$\text{Minimize} \quad \sum_{n=1}^{N} p_n(hW_n + hI_n + bB_n)$$

$$\begin{aligned}
\text{Subject to} \quad & W_n = W_{a(n)} + R_n - X_n && \forall n, \\
& X_n \le a_k W_{a(n)} + b_k && \forall n, \forall k, \\
& I_n - B_n = I_{a(n)} - B_{a(n)} + X_n - D_n && \forall n, \\
& W_n, R_n, X_n, I_n, B_n \ge 0 && \forall n.
\end{aligned}$$

The objective function accounts for the expected WIP, inventory and backorder costs. The size of the scenario tree is clearly exponential in the number of periods *T*, and depends

on the number of possible demand realizations considered at each stage. For this reason the computational burden of this scenario based model rapidly becomes excessive. Therefore even for relatively small problem instances used to benchmark our heuristics, some means of reducing the computational effort must be devised.

3.5 Computational Experiments

In this section we will highlight key insights from our computational study.

3.5.1 Experimental Design

The computational experiments that evaluate the performance of the different formulations we have discussed are as follows:

Demand: Demand in each period is independent and normally distributed with mean 100 and standard deviation 25. In order to implement the multi-stage stochastic program, the scenario tree must be constructed based on the demand profile. We use the moment matching algorithm of Miller and Rice(1983) to generate the demand realizations and the corresponding probabilities for each node. The number of branches is equal for all periods of planning horizon T. We choose three values for the number of branches: 3, 4, and 5.

Clearing function: To study the performance of the solutions obtained from different formulations, we start with single segment and two-segment clearing functions whose parameters are given in Table 3.2. The upper limit C is one of the important parameters of the clearing function that directly impacts the utilization of the production system. When C is infinity, we have a single segment clearing function. Three other possible levels of C are also

considered: 117, 114, and 111 corresponding to utilization levels of 0.855, 0.877, and 0.90 respectively.

Table 3.1 Two-segment clearing function

Segment	Intercept	Slope
1	0	$1/2$
2	C	0

Costs: An important factor is the relative magnitude of the inventory holding and backorder costs. The cost ratio $b/(h+b)$ in the right hand side of the chance constraints in the CC and S-CC models can be interpreted as the desired service level for the system. The inventory cost is fixed at $h = 1$ and the shortage cost is varied to yield estimated service levels of 86%, 90%, 94%, and 98%.

Lead time: The values of L_t in the right hand side of the chance constraints for each time period in the model CC, based on Little's Law, are chosen to be

$$L_t = \frac{W_{t-1}}{X_t},$$

where W_{t-1} and X_t are the planned values at previous iteration of the rolling horizon.

Performance Evaluation: To compare the performance of the production planning models CC and MSP, a rolling horizon simulation is used (cf. Figure 3.3). In each period, the demand is realized. For CC, the model is solved and only the release and production plans for the first period are implemented. For the MSP model, the realized demand becomes the demand of the root node, the model is then solved and only the release and production plans

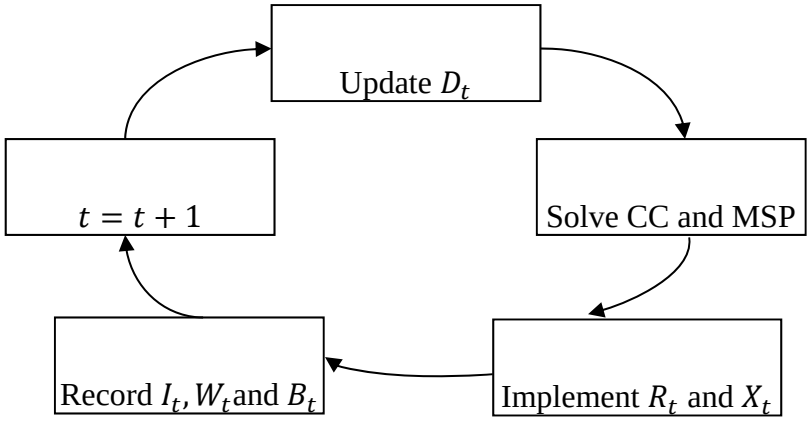

Figure 3.3 Rolling horizon

of the first stage are implemented. The occurrence of stockouts, the number of backorders, and the realized inventory levels are recorded in both cases and the planning horizon is shifted one period. This process continues until we reach the end of the rolling horizon. The average inventory holding and backorder costs along with the realized service level over the rolling horizon are computed. In the following sections we will highlight key insights from our computational study.

3.5.2 Results

In this section, we first present the performance of the MSP model under different scenario trees. Next, we compare the performance of CC, S-CC and MSP models considering various clearing function properties.

3.5.2.1 Performance of the MSP

The number of possible outcomes is an important parameter of multi-stage stochastic programming models. We first examine the performance of MSP under different values of

branches per node. The normal distribution is approximated by using the Gaussian quadrature method (Miller and Rice 1983). The values and probabilities of the approximated standard normal distribution for $N = 3, 4$ and 5 are shown in Table 3.2.

Table 3.2 Discrete approximation for scenarios

	Values and (Probabilities)				
$N = 3$	-1.73 (.16667)	0 (.66667)	1.73 (.16667)		
$N = 4$	-2.33 (.04588)	-0.74 (.45412)	0.74 (.45412)	2.33 (.04588)	
$N = 5$	-2.86 (.01126)	-1.36 (.22208)	0 (.53333)	1.36 (.22208)	2.86 (.01126)

Figure 3.4 summarizes our computational experiments. The graph in the top left corner shows the performance of the models for a single segment CF. The other graphs are all related to a two-segment clearing function with different upper limits. The cost of the system increases with service level and system utilization. However, different system costs result in the same service level in some cases. The reason is that while the required service level increases, the incomplete information about the demand distribution in the scenarios leads to a different service level. As the number of points used for the approximation increases, more information is included in the scenario tree and the performance of the stochastic programming model improves. The scenario tree based on a five point normal approximation is used for solving the MSP for the rest of this section.

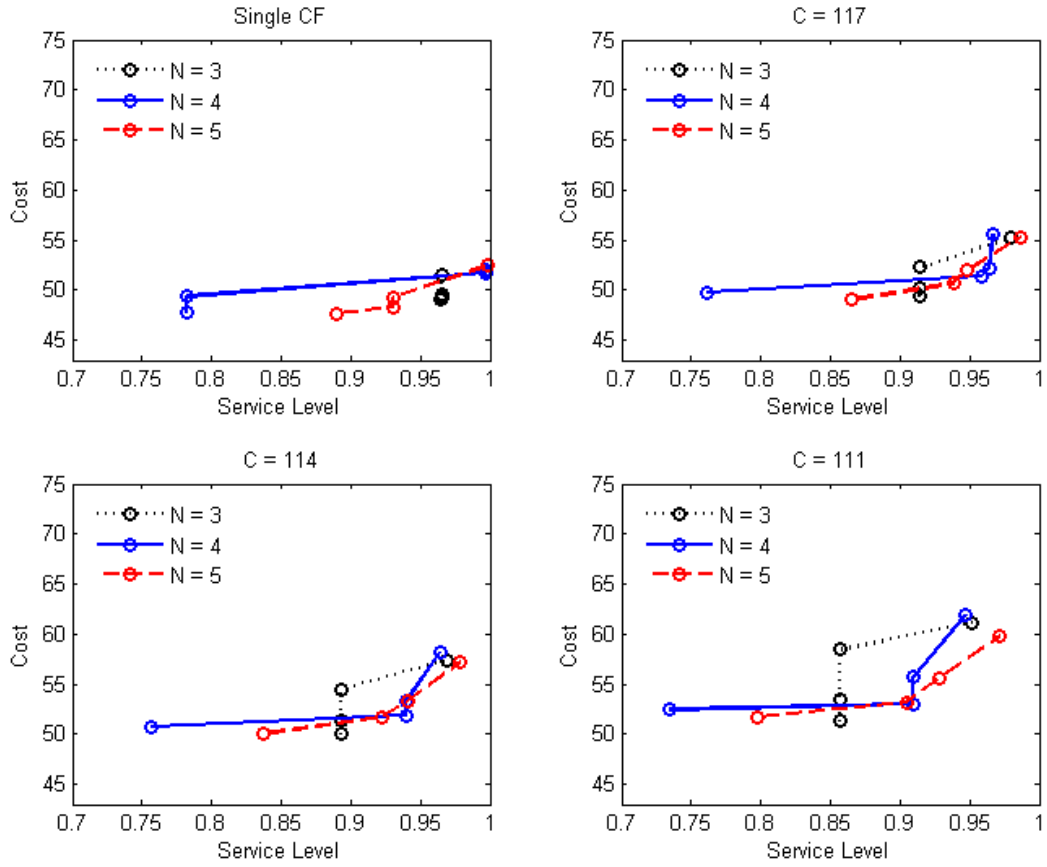

Figure 3.4 Impact of number of branches per node

3.5.2.2 Performance of the Three Models

In this section we examine the performance of production planning models of Section 3.4. The cost and service level for different required service levels are summarized in Figure 3.5. For the single segment CF, the efficient frontiers for the different models are almost the same. The cost of the system increases with system utilization. This increase is higher for the

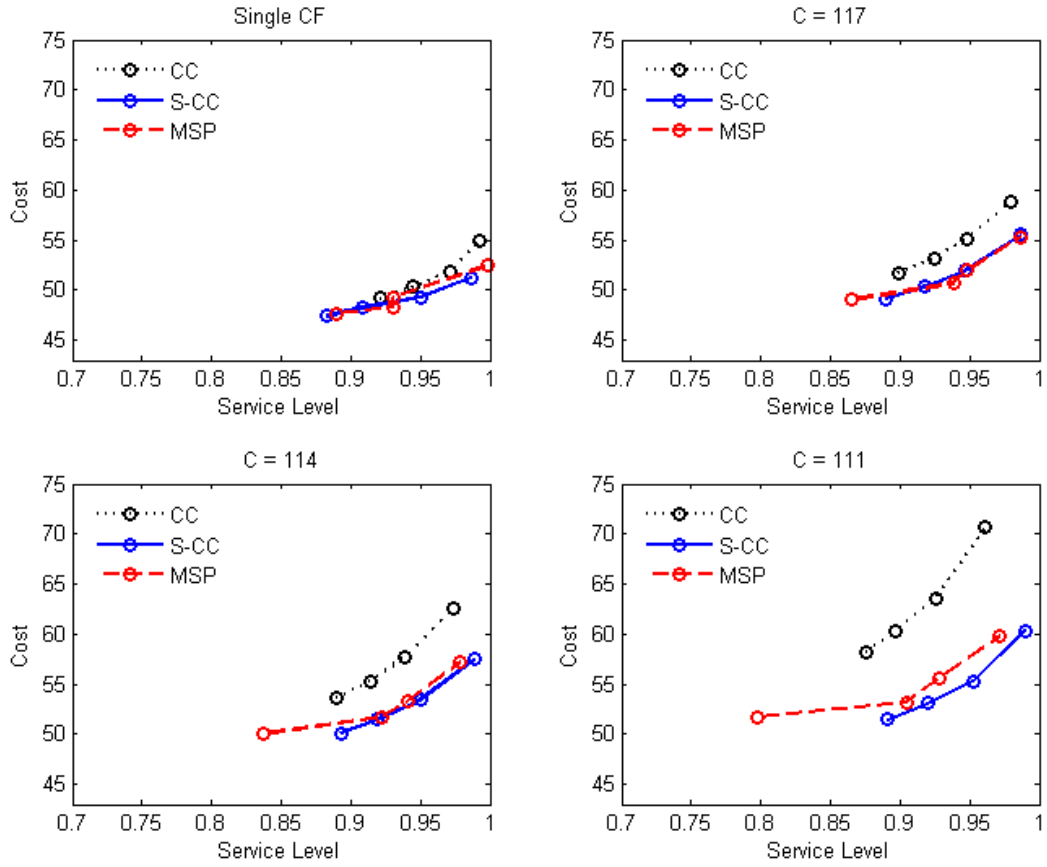

Figure 3.5 Impact of system utilization

CC model as the utilization approaches 0.90. One possibility is while the S-CC model and the MSP do not explicitly incorporate the lead time in the model, the CC uses lead time estimates in the right hand side of the chance constraints. Table 3.3 summarizes the lead time variability observed in the simulations for the CC model. The variability increases with the high utilization case, suggesting that the lead time estimation is not sufficient to characterize the right hand side of the chance constraints. The efficient frontiers for the S-CC and MSP are close to each other except when $C = 111$. There are two possible reasons for this

behavior. One is that the assumptions of the S-CC are more valid in high utilization system. Another possibility is that the MSP needs a scenario tree that approximates the tail of the demand distribution better.

Table 3.3 Lead time variance

	$\alpha = 86\%$	$\alpha = 90\%$	$\alpha = 94\%$	$\alpha = 98\%$
Single CF	0	0	0	0
C = 117	0.07	0.07	0.07	0.08
C = 114	0.15	0.16	0.17	0.19
C = 111	0.47	0.51	0.57	0.72

Another interesting result is the difference between the required and realized service level. For the CC model, while the realized service levels exceed the required ones under low utilization, they degrade as the utilization increases. There are two possible reasons for this behavior. One possibility is that linking order releases and planning decisions to lead times is not considered properly. Another possibility is that the shortfall due to congestion is not captured in this model. While the MSP implicitly considers the load dependent lead time, it characterizes the demand uncertainty through the use of approximated scenarios. Thus the realized service levels do not match the required ones. The success of S-CC is due to its incorporation of load dependent lead times in the release policy, which leads to the shortfall based chance constraint, while using a complete characterization of the demand distribution.

To further understand the impact of the clearing function on production planning models, we add another segment to the clearing function with upper limit $C = 114$. The parameters of the new set of clearing functions are shown in Table 3.4. When $0 < c_3 < 38$,

this gives us a three-segment clearing function. The performance of the production system is evaluated for $c_3 = 38$, 32, 26, and 20.

Table 3.4 Three-segment clearing function

Segment	Intercept	Slope
1	0.0	$1/2$
2	114	0
3	c_3	$1/3$

Figure 3.6 illustrates the results of this analysis. As the intercept of the third segment decreases, the cost of the system increases. While this increase is higher for the CC model as c_3 changes from 38 to 32 and 32 to 26, it is lower when c_3 changes from 26 to 20. The behavior of the CC model again corresponds to variance of the lead time estimation used to set the safety stock level summarized in Table 3.5.

3.6 Conclusions

In this paper we examine production planning models that consider workload dependent lead time and stochastic demand in an integrated manner using clearing functions. We propose a load dependent release policy leading to a tractable chance constrained model. The significance of our work lies in this analytical characterization, which allows a simple way to find and implement a near optimal policy.

We study the CC and MSP models as benchmarks. A numerical comparison with these models quantifies the S-CC model's impact on the expected cost and service level under different business environment. While the S-CC model explicitly links the releases to

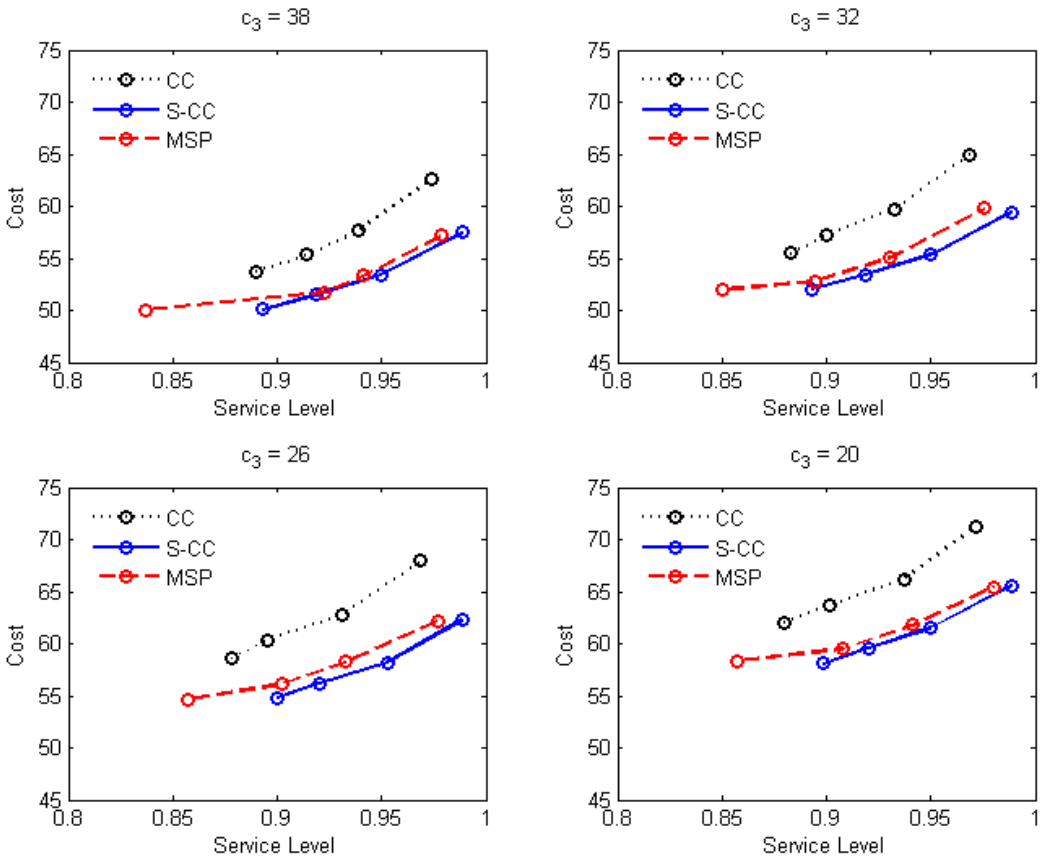

Figure 3.6 Impact of third segment

Table 3.5 Lead time variance

	$\alpha = 86\%$	$\alpha = 90\%$	$\alpha = 94\%$	$\alpha = 98\%$
$c_3 = 38$	0.15	0.16	0.17	0.19
$c_3 = 32$	0.18	0.19	0.20	0.23
$c_3 = 26$	0.18	0.19	0.21	0.24
$c_3 = 20$	0.17	0.18	0.20	0.23

the load dependent lead times and captures shortfall due to congestion, the CC model does not consider these properties and the MSP model implicitly includes them through the scenarios. The CC model is thus completely dominated as the system utilization or the lead time variability increases. On the other hand the MSP model is sensitive to the number and choice of scenarios. This approach appears to produce good solutions if enough scenarios are considered in the scenarios tree, which results in very large models relative to the CC and S-CC models. In our case, the S-CC model even outperforms the five point normal approximation.

A natural extension of this work is to consider correlated demands, particularly those demands arising from forecast evolution models. Heath and Jackson (1994) develop a general forecast evolution model and we build upon this model in Chapter 2 to capture dependency evolution between demands. However, care must be taken in the development of the conditional chanced constrained and methods to generate scenario trees for correlated demand. An interesting question in this context is the value of forecast updates in the presence of workload dependent lead time. Extension to multiple product and multiple stage production-inventory systems are also a natural direction for future work.

Chapter 4

Value of Demand Forecast Information in a Production System with Load Dependent Lead Time

Abstract

The value of demand forecast information in production-inventory planning systems has been the subject of numerous studies in the literature. Most of these studies used different forms of lead time models, assuming that there is no dependency between workload and lead times. In this paper we consider a dynamic production-inventory system with forecast updates based on the martingale model of forecast evolution (MMFE). The nonlinear dependency between workload and lead times is captured through clearing functions. Under such settings, we first obtain a chance constrained model and show how information affects the performance of the system subject to congestion. We then evaluate the performance of this model compared to that of a multistage stochastic programming model, for which we propose a method to represent the forecast evolution in a set of discrete scenarios. Our computational study helps to quantify the value of forecast update information.

4.1 Introduction

Demand forecasts over a planning horizon are a key input for production and inventory planning systems that try to better match supply with demand. Since forecasts are revised over time as new information becomes available, production planning and inventory control models that use these forecasts should take into account the evolution of these updates over time. With each forecast revision, ordering or production decisions can be updated on a rolling horizon basis in order to incorporate the most recent forecast information. Order releases determine work in process (WIP) inventory levels, which determine resource utilization and, in turn, the cycle time, the time between a product being released into the factory and its completion. These cycle times must be considered in order to match the output of the production facility with market demand in some optimal way. We shall refer to the estimate of cycle time used in production planning as the lead time.

Our goal is to develop models that adjust the manufacturer's production and inventory planning decisions according to the forecast updates in the face of production resources subject to congestion. This requires capturing the variation of forecast updates and the related correlation structure over time as well as the dependency between workload and cycle time discussed above. In particular, we explore the following questions: How can forecast information be used within a production system with congestion? How much does the forecast information affect the performance of the system? While information is always beneficial, we want to investigate when this value is most beneficial and when it is only marginally useful.

We consider a single item, multi-period production-inventory system with demand forecast updates. The forecast evolution follows the Martingale Model of Forecast Evolution (MMFE), developed by Graves et al. (1986) and Heath and Jackson (1994) and extended in Chapter 2 of this thesis, in order to model the evolution of the conditional covariances over time. The dependency between workload and cycle times is captured through the use of nonlinear clearing functions (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989) that relate the expected output of a capacitated resource in a planning period to the average WIP level at the resource over the period.

We first develop a chance constrained model that represents service level requirements with a set of probabilistic constraints to protect against demand uncertainty under the existence of forecast information and congestion. We use this model to show how the production policy depends on the forecast information. We then investigate the performance of this model compared to that of a multistage stochastic programming model, for which we develop a method to represent the forecast evolution in a set of discrete scenarios. Our computational study helps to quantify the value of forecast update information.

Section 4.2 reviews the related literature. Section 4.3 presents the forecast update model and describes the production planning model with the clearing function. Section 4.4 analyzes the production planning problem with two models of decision making, chance constrained model and stochastic programming model. Section 4.5 describes the results of computational experiments. Section 4.6 concludes the paper with a summary of our principal conclusions and a discussion of future work.

4.2 Literature Review

Most production planning models establish order releases over time using a multi-period model to meet the demands. Methods for cycle time prediction must be integrated into this model to optimize the performance of the production system. The simplest one is to treat lead time as a fixed quantity that is independent of workload (Hackman and Leachman 1989). However, queuing models of production systems (Buzacott and Shanthikumar 1993) show that cycle times depend on the workload or capacity utilization of the production system, which, in turn, is determined by the release decisions made by the planning system. Hence the lead times used to estimate these cycle times should be adjusted accordingly.

Two main approaches have proposed to consider this dependency between lead times and resource utilization. The first approach uses lead time based models. Estimated lead times for the order releases are generated that depend on the order release plan. These estimated lead times can be derived from a simulation of the production system, which requires iterative adjustment of order releases and lead time prediction (Hung and Leachman 1996; Byrne and Bakir 1999), or from operational characteristics that represent long term averages (Voss and Woodruff 2003). The second approach uses WIP based models that represent the flow of WIP through the production system by inventory balance equations and a clearing function. While lead times are not represented explicitly in models using clearing functions, they can be computed from the WIP level at a certain time and the output in the future periods. Graves (1986) introduce a proportional clearing function that is consistent with constant lead times. This model states that in each planning period a constant proportion of the available WIP is processed. Srinivasan et al. (1988) and Karmarkar (1989)

independently extended this idea to a concave nonlinear clearing function which relates WIP levels to output such that the lead times are influenced by the workload of the production system and thus is able to capture the congestion behavior of capacity constrained production systems. Missbauer and Uzsoy (2010) give a comprehensive review of this area.

Several studies have recently explored the use of the clearing functions within models for production planning with uncertainty. Ravindran et al. (2011) provide a production planning model with chance constraints and the clearing functions. Their analysis depends on the assumption that inventory position is a percentile of the demand over the lead time that is variable due to the presence of the clearing function. This model requires the lead times used to establish inventory levels to be provided as an exogenous parameter. Aouam and Uzsoy (2012) examine both two-stage and multi-stage stochastic programming models and compare their performance to the chance-constrained models of Ravindran et al. (2011). Their results show that the chance-constrained model shows quite promising performance relative to the stochastic programming models with significantly lower computation time requirements. Chapter 3 of this thesis proposes a load dependent release policy leading to a chance constrained model which considers workload dependent lead times of clearing functions without requiring any external lead time parameters. The computational experiments indicate that their chance-constrained model performs favorably relative to that of Ravindran et al. (2011) and the stochastic programming models with significantly lower computation time requirements. None of these studies, however, considers evolution of demand forecast information – all assume that demand in each period is independent of demand in other periods.

Modeling the evolution of forecasts was first suggested by Hausman (1969). He models the evolution of the forecasts as a quasi-Markovian or Markovian process. His model is limited to a single selling season. To accommodate the simultaneous evolution of forecasts for demand in many periods, the Martingale Model of Forecast (MMFE) was independently developed by Graves et al. (1986) and Heath and Jackson (1994). The MMFE allows forecasts to evolve over time as new information becomes available in each period. Graves et al. (1986) construct a single item version of the MMFE to model the evolution of forecasts of independent and identically distributed stationary demand over discrete time periods. Heath and Jackson (1994) propose a multiproduct model that can capture correlations in demand forecasts both between products and across time periods. Using the MMFE, Güllü (1996) analyzes the value of this variability reduction in a capacitated single item system with uncorrelated demands. Graves et al. (1998) demonstrate this uncertainty resolution by defining the i period forecast error as the difference between the actual demand in a period and the forecast of this demand made i periods earlier. They show that the variance of the forecast error over the forecast horizon matches the variability of the demand process, and is equal to the trace of the covariance matrix of forecast errors. This variance decreases as time goes by and i becomes smaller.

This generalized demand model has been already proven useful in modeling several operational decisions. Toktay and Wein (2001) use the MMFE in a single item discrete-time continuous-state queue to analyze a stationary production-inventory system and develop a class of forecast corrected base stock policies. Aviv (2001) uses the MMFE to explore the effect of collaborative forecasting in a two stage supply chain consisting of a retailer and a

supplier. Gallego and Özer (2001) study an inventory model with advance demand information that can be viewed as a special case of the MMFE. Dong and Lee (2003) adopt the MMFE to show that the structure of the optimal stocking policy proposed by Clark & Scarf (1960) holds under time correlated demand processes. Iida and Zipkin (2006) develop a dynamic forecast-inventory model with forecast updates based on the MMFE and propose a technique to obtain approximately optimal inventory policies. Lu et al. (2006) study a periodic-review inventory system with the MMFE, develop a class of tractable bounds on the optimal base stock levels and use them to construct near-optimal policies. Chen and Lee (2009) show that many commonly used time-series demand models in the literature such as the general ARIMA model (Box and Jenkins 1970) and the linear state-space model (Aviv 2003) can be interpreted as special cases of the MMFE model. Common to most of the above applications, however, is the use of demand information updates leading to the evolution of demand forecasts. Chapter 2 of this thesis contributes by describing the evolution of the dependency between demands in consecutive periods as new information becomes available under both additive and multiplicative forecast updating. None of these papers considers load dependent lead times.

In sum, the literature on production planning with load dependent lead times does not consider evolution of demand forecast information, while the literature on the forecast evolution models ignores the load dependent lead times. It seems reasonable that the production congestion should have a major impact on the value of demand forecast information. To our knowledge, this paper is the first to consider these interactions.

4.3 The Model

In this section we present a single item multi-period production-inventory model with stochastic demand and load dependent lead time. We first describe the forecast update model and state our assumptions about how the forecast evolves over time. We then develop the production model that captures the load dependent lead time and the forecast process.

4.3.1 The Forecast Update Model

We assume that demand forecasts are available for some number of periods in the future. We call this number the forecast horizon and denote it H. Let $D_{s,t}$ be the demand forecast made in period s for period $t, s \leq t \leq s + H$. $D_{t,t}$ denotes the realized demand in period t since the forecast is made after the actual demand is revealed. Beyond the forecast horizon, there is no specific information about demands. The demand forecasts for periods $t > s + H$ are thus set equal to a constant μ. When time advances to the next period, period $s + 1$, additional information becomes available and new demand forecasts are generated. Eventually, this sequence of forecasts for a specific future period evolves into the realized demand in that period once the demand is observed. Heath and Jackson (1994) describe this forecast evolution by modeling the evolution of the forecast updates. They define

$$\varepsilon_{s,t} = D_{s,t} - D_{s-1,t}$$

as the forecast update made in period s for period $t, s \leq t$. Let $\varepsilon_s = (\varepsilon_{s,s}, \varepsilon_{s,s+1}, \dots, \varepsilon_{s,s+H})$ denote the forecast update vector received at period s. The MMFE assumes that forecasts represent the conditional expectations of demand given all available information at the time

the forecast is made, i.e., $D_{s,t} = E[D_t|\mathcal{F}_s]$ where $\mathcal{F}_s$ is a σ-field describing the information available at the end of period s such that $\mathcal{F}_s \subseteq \mathcal{F}_{s+1}$. This implies that the forecast updates are unbiased and independent over s but not necessarily over t. Under these relatively mild assumptions, the MMFE speculates that the update vectors ε_s are independent, identically distributed, multivariate normal random vectors with $E[\varepsilon_s] = 0$ and $Var[\varepsilon_s] = \sum$, the covariance matrix.

The MMFE is not a forecasting technique, but rather a framework for representing how forecast updates evolve over time. It is compatible with a wide range of forecasting techniques, including statistical and judgment-based methods. It only needs historical forecast data and can be used to model the forecast evolution without specifying the particular forecasting methods used to generate the forecasts.

In Chapter 2 we show that the progressive realization of uncertain demands across successive discrete time periods through additive or multiplicative forecast updates results in the evolution of the conditional covariance of demand in addition to its conditional mean. Let $\gamma_{s,(t,t+i)}$ denote the conditional covariance between D_t and D_{t+i} given $\mathcal{F}_s$. We then have

$$\gamma_{s,(t,t+i)} = Cov(D_t, D_{t+i}|\mathcal{F}_s) = \sum_{j=1}^{t-s} \sigma_{t-s-j,t+i-s-j}\,, \quad 1 \le t-s \le H-i+1,$$

which implies that $\gamma_{s,(t,t+i)}$evolves over time following the relation

$$\gamma_{s,(t,t+i)} = \gamma_{s-1,(t,t+i)} - \sigma_{t-s,t+i-s},$$

where $\sigma_{t-s,t+i-s}$ is the covariance between $\varepsilon_{s,t}$ and $\varepsilon_{s,t+i}$. The conditional covariances between D_t and D_{t+i} for $s < t+i-H$ are set equal to the unconditional covariance γ_i.

When time advances to period s such that $s = t + i - H$, additional information $\varepsilon_{s,t}$ and $\varepsilon_{s,t+i}$ becomes available. Thus, the covariances between these variables are removed from the covariances between D_t and D_{t+i}. Since the elements of the covariance matrix can be positive or negative, the resolution of forecast updates may cause the conditional covariances to increase or decrease over time. When $s = t$, D_t is totally revealed; all covariance matrix elements are removed and thus $\gamma_{s,(t,t+i)} = 0$. This result implies the dependencies between demands are not constant over time, but rather depend on the point in time at which dependencies are computed. This additional evolution gives the decision maker a better characterization of demand that can lead to more effective decisions.

4.3.2 The Production Planning Model

We consider a production system that produces a single product with stochastic demand. Unsatisfied demand is backlogged. The production manager observes the output of the forecasting activity in the form of a stream of forecast updates and must convert these updates into a production policy. The planning is done in a periodic review setting for a certain planning horizon and at each period the method of rolling horizons is applied to plan according to the demand forecast revisions. To make things precise we need the following notation: For each variable Z, we define $Z_{s,t}$ to be the prediction made in period s for the value of Z in period t while for $s \geq t$ the predictions equal to the actual value.

The following sequence of events occurs during a single period. At the beginning of period s, demand D_s is observed and forecasts are updated. Based on these forecast updates, planned releases $R_{s,t}$ are determined representing the amount of work that is planned to be

released into the production system at time *t* based on information available at time *s*. A fraction of work that is in the production system is processed and put into inventory. Let $X_{s,t}$ be the planned amount of production for period t as of period s. The planned work in process inventory at time t is the expected level of WIP in a future period given the information available at time s:

$$W_{s,t} = W_{s,t-1} + R_{s,t} - X_{s,t}.$$

Demands in each period are satisfied using on-hand inventory and the production completed during that period. Thus the planned inventories $I_{s,t}$ and planned backorders $B_{s,t}$ evolve with the relation

$$I_{s,t} - B_{s,t} = I_{s,t-1} - B_{s,t-1} + X_{s,t} - D_{s,t}.$$

In order to incorporate the lead time dependency, the flow of WIP that performs the production is modeled by means of inventory balance equations, accompanied with a clearing function. A clearing function can be defined as the expected output of a capacitated resource over a given period of time as a function of some measure of WIP in that period. In this paper we use the planned amount of available work at the start of that period as the WIP measure. Hence the clearing function for period *t* as of period *s* is a functional relationship of the form

$$X_{s,t} = f(W_{s,t-1}).$$

Following Asmundsson et al. (2009), we approximate the clearing function using outer linearization, yielding the constraints

$$X_{s,t} \leq min\{\alpha_k W_{s,t-1} + \beta_k\}, \ \ \forall k = 1, \dots, K,$$

where α_k is the slope and β_k is the intercept of segment k of the piecewise linearized clearing function. To capture the concavity of the clearing function, we assume that successive segments have monotonically increasing intercepts ($\beta_1 < \beta_2 < \cdots < \beta_k$) and monotonically decreasing slopes ($\alpha_1 > \alpha_2 > \cdots > \alpha_k$) as seen in Figure 4.1. The intercept of the first segment is set to zero ($\beta_1 = 0$), since production cannot take place without some WIP being present. The last segment has a slope of zero ($\alpha_k = 0$) since at very high workloads increasing the workload by releasing additional work will not increase output.

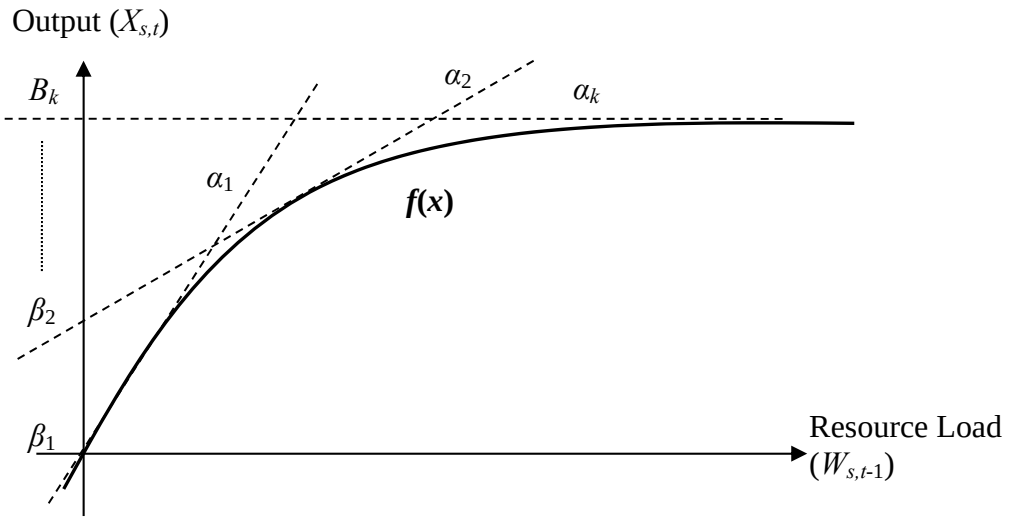

Figure 4.1 A typical clearing function and outer linearized segments

An important advantage of clearing functions for our purpose is that lead times do not appear in the formulation; releases and lead times are jointly optimized, allowing the lead time to vary over the planning horizon. Extensive discussions of clearing function models for production planning are given by Asmundsson et al. (2006; 2009), Kacar et al. (Kacar et al. 2012), Missbauer (2002) and Purgstaller and Missbauer (in press). Extensive computational

experiments with deterministic demand have shown that when properly parameterized these models yield superior production plans to those generated by conventional LP models with fixed exogenous lead times, even when used in iterative procedures in combination with simulation models.

Our objective is to determine releases to minimize the total planned inventory holding and backorder costs over the planning horizon,

$$\sum_{t=s}^{s+T-1} \left(hW_{s,t} + hI_{s,t} + bB_{s,t}\right),$$

where h and b are the unit inventory holding and backorder costs, respectively.

4.4 Analysis

In this section we develop two different production planning models that differ in the way in which they model uncertainty. Model 1 is a chance constrained model that assumes knowledge of the demand distribution in each period, and represents service level requirements with a set of probabilistic constraints that can be violated with a certain probability. While the cost of such violation is not explicitly considered in the model, there is a connection between the required service level and the relative magnitude of holding and shortage costs. Model 2 is a stochastic programming model that represents the uncertain nature of demand as a set of discrete scenarios, distinct sample paths drawn from the underlying demand distributions specifying a specific demand realization for each period in the planning horizon. As the number of scenarios tends to infinity, this model will yield the

optimal expected cost solution by determining a set of decision variables specifying the action to be taken in each time period for each scenario.

4.4.1 Chance Constrained Model

In this section we develop a chance constrained model (CC). Assume the production system operates under a base-stock policy. Production is set in each period to restore the on-hand inventory at the beginning of that period to a target level S_t while not exceeding the maximum possible output which is the intercept of the last segment of the clearing function. As long as there is a positive probability that demand in a period can exceed the upper limit of the clearing function C, then there is a positive probability that the production capacity in a period will not be sufficient to raise the on-hand inventory to S_t. The amount by which S_t exceeds the actual on-hand inventory level is called the shortfall (Tayur 1993; Glasserman 1997).

Let the random variable $Y_{s,t}$ represent the planned shortfall in period t as of period s, defined as

$$Y_{s,t} = S_t - \left(I_{s,t-1} - B_{s,t-1} + X_{s,t}\right),$$

where $I_{s,t-1} - B_{s,t-1} + X_{s,t}$ is the desired planned-on-hand inventory at the beginning of period t. When there is no advance forecast information, we have shown in Chapter 3 that the inventory level at the end of period t has to be raised to protect against the sum of the demand and shortfall of period $t+1$:

$$I_t - B_t = S_t - Q_{t+1},$$

where

$$Q_{t+1} = D_{t+1} + Y_{t+1}.$$

In our case, we need to protect against the as yet unobserved portion of Q_{t+1}. The WIP and FGI holding cost are equal and thus WIP cost can be removed from the objective function, since the WIP level is determined by the demand D_t and the lead time L_t. The objective is to find the value of S_t that minimizes the sum of the expected inventory holding and backorder costs charged to period t, given all information available at time s, given by

$$C_{s,t}(S_t) = hE_s[(S_t - Q_{t+1})^+] + bE_s[(S_t - Q_{t+1})^-],$$

where the operator E_s denotes the conditional expectation given information $\mathcal{F}_s$. The derivative of $C_{s,t}(S_t)$ with respect to S_t is

$$C'_{s,t}(S_t) = hG_{s,t+1}(S_t) - b\left(1 - G_{s,t+1}(S_t)\right),$$

where $G_{s,t}$ denotes the cumulative distribution function of Q_{t+1} conditional on information $\mathcal{F}_s$. Setting $C'(S) = 0$ yields

$$S^*_{s,t} = G^{-1}_{s,t+1}\big(h/(h+b)\big).$$

This result states that for each period t, the base stock level is a percentile of the conditional distribution function of sum of the demand and shortfall of period $t+1$ given the information $\mathcal{F}_s$. The above equations lead to the chance constraint

$$I_{s,t} - B_{s,t} + X_{s,t+1} + Y_{s,t+1} \geq G^{-1}_{s,t+1}\big(h/(h+b)\big).$$

The left hand side of the chance constraint represents the sum of planned inventory and shortfall after the planned production is realized at the beginning of period $t+1$ as of period s. This new policy states that the planned order in period t should be placed to raise this planned inventory up to the percentile of the conditional sum of the demand and shortfall

of period $t+1$. The chance constrained formulation incorporating this constraint is as follows:

$$\text{Minimize} \quad \sum_{t=1}^{T}\left(hW_{s,t} + hI_{s,t} + bB_{s,t}\right)$$

$$\begin{aligned}
\text{Subject to} \quad & W_{s,t} = W_{s,t-1} + R_{s,t} - X_{s,t} && \forall t = s, \dots, T-s+1, \\
& X_{s,t} \le \alpha_k W_{t-1} + \beta_k && \forall t = s, \dots, T-s+1, \forall k, \\
& I_{s,t} - B_{s,t} = I_{s,t-1} - B_{s,t} + X_{s,t} - D_{s,t} && \forall t = s, \dots, T-s+1, \\
& Y_{s,t} = Y_{s,t-1} + D_{s,t-1} - X_{s,t} && \forall t = s, \dots, T-s+1, \\
& I_{s,t} - B_{s,t} + X_{s,t+1} + Y_{s,t+1} \ge G_{s,t+1}^{-1}\left(h/(h+b)\right) && \forall t = s, \dots, T-s+1, \\
& W_{s,t}, R_{s,t}, X_{s,t}, I_{s,t} \ge 0 && \forall t = s, \dots, T-s+1.
\end{aligned}$$

We now develop a closed form approximation for the right hand side of the chance constraints. Toktay and Wein (2001) shows when demands are correlated and the utilization is high, the distribution Y_t can be approximated by an exponential distribution such that

$$P(Y_t > y) = e^{-\theta(y+\beta)},$$

where $\theta = 2(C-\mu_D)/(\gamma_0 + 2\sum_{i=1}^{H}\gamma_i)$ and the correction term for normally distributed demand is $\beta = 0.583\sqrt{\gamma_0 + 2\sum_{i=1}^{H}\gamma_i}$.

There are two sources of variability in the distribution of Q_{t+1}. The first is the demand uncertainty of D_{t+1} and the second the variability of the shortfall Y_{t+1}. Since the second variability is mainly due to the limited capacity of the production resource as expressed in the clearing function, we assume that the information at time s can only affect the uncertainty of D_{t+1}. Hence

$$P(Q_{t+1} \le q \,|\, \mathcal{F}_s) = P(Y_{t+1} + D_{t+1} \le q \,|\, \mathcal{F}_s)$$

$$= \int_{x=-\infty}^{q} P(Y_{t+1} \leq q - x) f_{s,t+1}(x) dx$$

$$= \int_{x=-\infty}^{q} \left(1 - e^{-\theta(q-x+\beta)}\right) f_{s,t+1}(x) dx,$$

where $f_{s,t+1}(.)$ denotes the distribution of D_{t+1} conditional on information $\mathcal{F}_s$. Since the forecast update vectors are assumed to follow a multivariate normal distribution, $f_{s,t+1}(.)$ follows a normal distribution with mean $D_{s,t+1}$ and variance $\gamma_{s,(t+1,t+1)}$. Therefore,

$$P(Q_{t+1} \leq q | \mathcal{F}_s) = \Phi\left(\frac{q - D_{s,t+1}}{\sqrt{\gamma_{s,(t+1,t+1)}}}\right) - e^{-\theta(q+\beta)} \int_{x=-\infty}^{q} e^{\theta x} f_{s,t+1}(x) dx.$$

Using integration by parts for the second term, we have

$P(Q_{t+1} \leq q | \mathcal{F}_s)$

$$= \Phi\left(\frac{q - D_{s,t+1}}{\sqrt{\gamma_{s,(t+1,t+1)}}}\right)$$

$$- e^{-\theta(q+\beta-D_{s,t+1}-\frac{1}{2}\gamma_{s,(t+1,t+1)}\theta)} \Phi\left(\frac{q - D_{s,t+1} - \gamma_{s,(t+1,t+1)}\theta}{\sqrt{\gamma_{s,(t+1,t+1)}}}\right).$$

Let us approximate the $\Phi(.)$ term by one for large q. Then

$$P(Q_{t+1} \leq q | \mathcal{F}_s) \approx 1 - e^{-\theta(q+\beta-D_{s,t+1}-\frac{1}{2}\gamma_{s,(t+1,t+1)}\theta)}.$$

Thus, for $b \gg h$, ${S_{s,t}}^*$ is well approximated by

$${S_{s,t}}^* = \frac{1}{\theta} ln(1 + b/h) - \beta + D_{s,t+1} + \frac{1}{2}\gamma_{s,(t+1,t+1)}\theta.$$

By means of these approximations, we have now obtained a chance-constrained formulation where the right hand side of the service level constraint can be calculated offline

and does not depend on any external lead time parameter. In order to understand the quality of the approximations, we also consider a stochastic programming model in the next section.

4.4.2 Stochastic Programming Model

In this section we develop a multi-stage stochastic programming problem (MSP) that uses scenarios to characterize the uncertainty in demand. Solutions are obtained for each scenario and then individual scenario solutions are aggregated to yield a final solution. The possible scenarios can be described as a scenario tree where nodes organized in levels which correspond to decision stages $s, \dots, T - s + 1$. The nodes n of the tree represent states of the world at a particular point in time. The root node of the tree represents the current state of the world. Each node n of the scenario tree, except the root node $n = 1$, has a unique parent $a(n)$. The probability of realization associated with node n is denoted by p_n. Let t_n denote the time stage corresponding to node n. A path from the root node to a node n describes one realization of the stochastic process from the present time $t_1 = s$ to t_n. If n is a leaf node, the path corresponds to a scenario, and represents a joint realization of the demand over all periods.

The decision maker can revise the release and production plans for different demand realizations at each stage of the scenario tree after observing the realizations of demand in previous periods. Thus all variables are indexed by nodes *n* as well as by periods *s*. For each variable Z, we define $Z_{s,n}$ to be the prediction made in period s for the value of Z in node n while for $s \geq t_n$ the predictions equal to the actual value. At the beginning of period *s*, demand D_s is observed and forecasts are updated. Based on these forecast updates, the

demand realizations at each node of the scenario tree $D_{s,n}$ are revised. A multi-stage stochastic programming model can be formulated as follows:

$$\text{Minimize} \quad \sum_{n=1}^{N} p_n\left(hW_{s,n} + hI_{s,n} + bB_{s,n}\right)$$

$$\begin{aligned}
\text{Subjec to} \quad & W_{s,n} = W_{s,a(n)} + R_{s,n} - X_{s,n} && \forall n, \\
& X_{s,n} \le a_k W_{s,a(n)} + b_k && \forall n, \forall k, \\
& I_{s,n} - B_{s,n} = I_{s,a(n)} - B_{s,a(n)} + X_{s,n} - D_{s,n} && \forall n, \\
& W_{s,n}, R_{s,n}, X_{s,n}, I_{s,n}, B_{s,n} \ge 0 && \forall n.
\end{aligned}$$

The size of the scenario tree is clearly exponential in the number of periods *T*, and depends on the number of possible demand realizations considered at each stage. In the rest of this section we show how to integrate the forecast updates model in the stochastic programming formulation through scenario trees. We generate scenario trees that take into account the evolution structure included by the MMFE.

Since the forecast update vectors are independent multivariate normal random vectors, demands $\{D_{s+i}: i = 1, \dots, T-1\}$ conditional on information $\mathcal{F}_s$ follow a multivariate normal distribution with mean $\mu = \left(D_{s,s+1}, \cdots, D_{s,s+T-1}\right)^T$ and covariance matrix M where $M_{ij} = \gamma_{s,(s+i,s+j)}$ represents the conditional covariance between D_t and D_{t+i} given information $\mathcal{F}_s$. Any multivariate normal random vector with mean μ and covariance matrix M can be written as $UZ + \mu$ such that Z is a standard multivariate normal random vector and $M = UU^T$ (Tong 1990). Therefore, we generate a standard multivariate normal random vector and then apply a transformation so that the resulting random vector is

consistent with a given mean and covariance matrix. This is the basic philosophy of the scenario generation method proposed by Høyland et al. (2003).

The tree structure is chosen so that the number of branches per node, starting with the root node, is N for all levels of the scenario tree. We start by sampling from a discrete univariate random variable to approximate the standard normal distribution using the Gaussian quadrature method (Miller and Rice 1983). The procedure generates an N-point distribution that matches the first $2N-1$ moments of the continuous distribution. We then create matrix $\mathbb{X}$ where $\mathbb{X}_{t,j}$ represents the standard normal approximated value for period t of scenario j. The last step is the matrix transformation

$$\mathbb{Y} = U\mathbb{X} + \mu,$$

where $\mathbb{Y}_{t,j}$ denotes the demand realization in period t for scenario j. The jth column of the matrix $\mathbb{Y}$ thus includes the demand realizations over the planning horizon for scenario j.

4.5 Computational Experiments

In this section, we evaluate the performance of the different formulations for the production planning with the clearing functions and forecast updates. Our goal is to understand the trade-offs between inventories, capacities, and information.

4.5.1 Experimental Design

The computational experiments that evaluate the performance of the different formulations we have discussed are as follows:

We assume that the demand mean and standard deviation are 100 and 25 respectively. At each time s the forecasts are updated. The forecast updates follow a multivariate normal distribution. The covariance matrix of forecast updates is

$$\begin{pmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & \rho_{02}\sigma_0\sigma_2 \\ \rho_{01}\sigma_0\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_0\sigma_1 \\ \rho_{02}\sigma_0\sigma_2 & \rho_{12}\sigma_0\sigma_1 & \sigma_2^2 \end{pmatrix},$$

where $\sigma_0 = 18, \sigma_1 = 14$, and $\sigma_2 = 10.25$.

For generating the scenario tree, the standard normal distribution is approximated by five point using the Gaussian quadrature method (Miller and Rice 1983). The values and probabilities of the approximated standard normal distribution are shown in Table 4.1.

Table 4.1 Five point approximation for standard normal distribution

Values	-2.86	-1.36	0	1.36	2.86
Probabilities	.01126	.22208	.53333	.22208	.01126

The parameters of the clearing functions are given in Table 4.2 Clearing function parameters. The upper limit *C* is an important parameter of the clearing function that directly impacts the utilization of the production system. Four possible levels of *C* are also considered: 120, 117, 114, and 111 corresponding to utilization levels of 0.80, 0.855, 0.877, and 0.90 respectively.

Table 4.2 Clearing function parameters

Segment	Intercept	Slope
1	0.0	$1/2$
2	*C*	0.0

An important factor is the relative magnitude of the inventory holding and backorder costs. The cost ratio $b/(h+b)$ in the right hand side of the chance constraints can be interpreted as the desired service level for the system. The inventory cost is fixed at $h = 1$ and the shortage cost is varied to yield estimated service levels of 86%, 90%, 94%, and 98%.

To compare the performance of the production planning models, a rolling horizon simulation is used (cf. Figure 4.2). In each period, the demand forecasts are realized. For the CC, the model is solved and only the release and production plans for the first period are implemented. For MSP, the realized demand becomes the demand of the root node and the other forecast updates revise the scenario tree. The model is then solved and only the release and production plans of the first stage are implemented. The occurrence of stockouts, the number of backorders, and the realized inventory levels are recorded in both cases and the planning horizon is shifted one period. This process continues until we reach the end of the rolling horizon. The average inventory holding and backorder costs along with the realized service level over the rolling horizon are computed. In the following sections we will highlight key insights from our computational study.

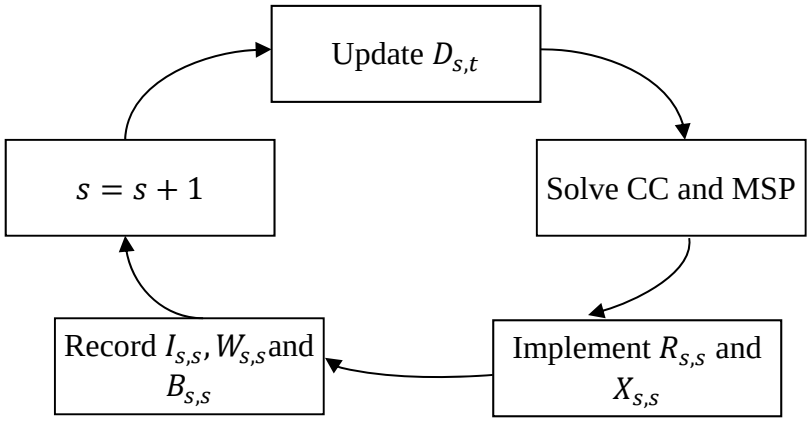

Figure 4.2 Rolling horizon

4.5.2 Results

We first compare the performance of the CC and MSP models. The average cost and service level for different required service levels are summarized in Figure 4.3.

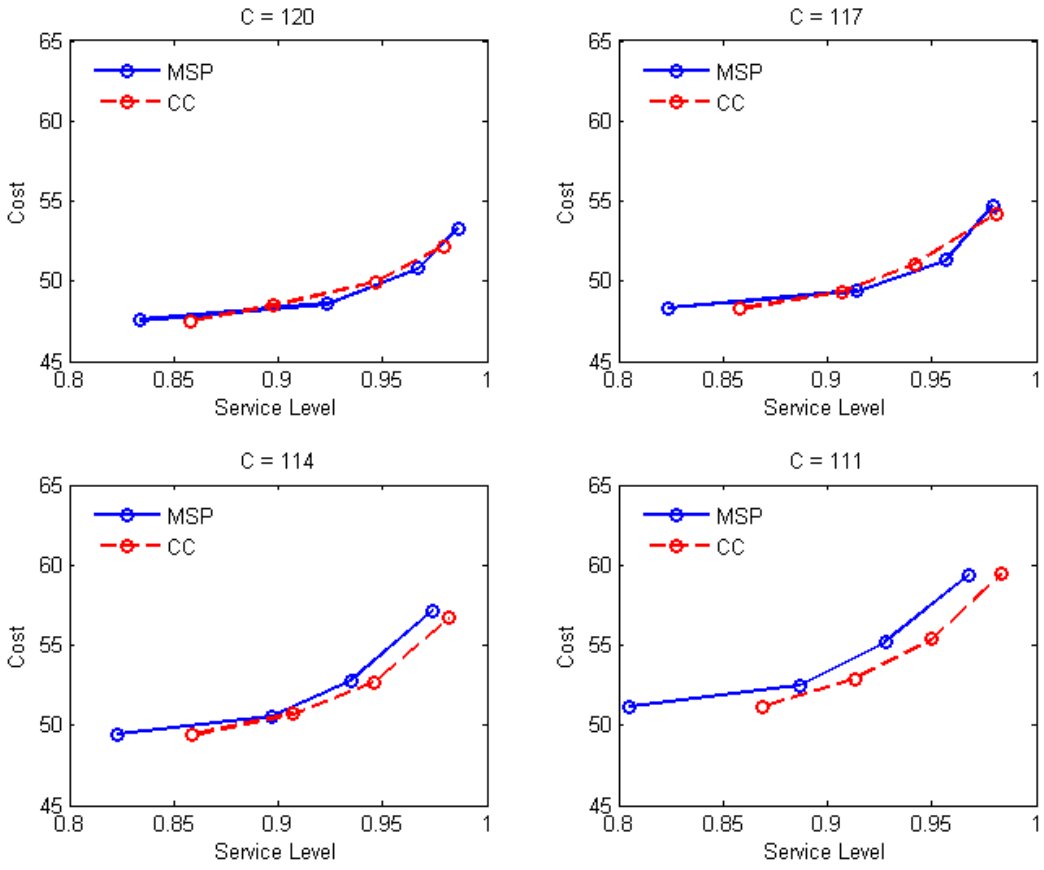

Figure 4.3 Impact of system utilization and service level

For both models the cost of the system increases with system utilization and required service level. However, this increase is lower for the CC model as the utilization and service level approach 0.90 and 98%, respectively. One possibility is that the assumptions used to obtain the chance constraints are more valid in high utilization and service levels. Another

possibility is that the MSP needs a scenario tree that approximates the tail of the demand distribution better. Another important issue is that while the efficient frontiers for the two models are almost the same; the realized service levels do not match the required ones for the MSP model. That is due to the use of approximated scenarios.

Next, we compare the average system cost of the inventory model that keeps track of the information provided by evolving forecasts with that of a comparable standard demand model. When the model does not keep track of the forecasts, it does not utilize the information provided by the forecast evolution mechansim. The forecaster thinks that no advance information is available, i.e. the forecast horizon $H = 0$. The conditional terms are thus replaced with the unconditional ones. Figure 4.4 illustrates the impact of information. The cost of the model with information updates is always smaller. This leads us to conclude that information is always beneficial.

The cost reduction increases with capacity limit and there is almost no benefit at high utilization. When the capacity limit is high, the model has flexibility and thus can use the information to modify the production. However, when the capacity limit is low, the model does not have much ability to revise production plans since the system is producing at its maximum capacity in any case. Another interesting result is that the service level does not affect the cost reduction. This could be due to the fact that the service level only impacts the safety stock level.

4.6 Conclusions

We extend production-inventory models by incorporating dynamic forecast evolution

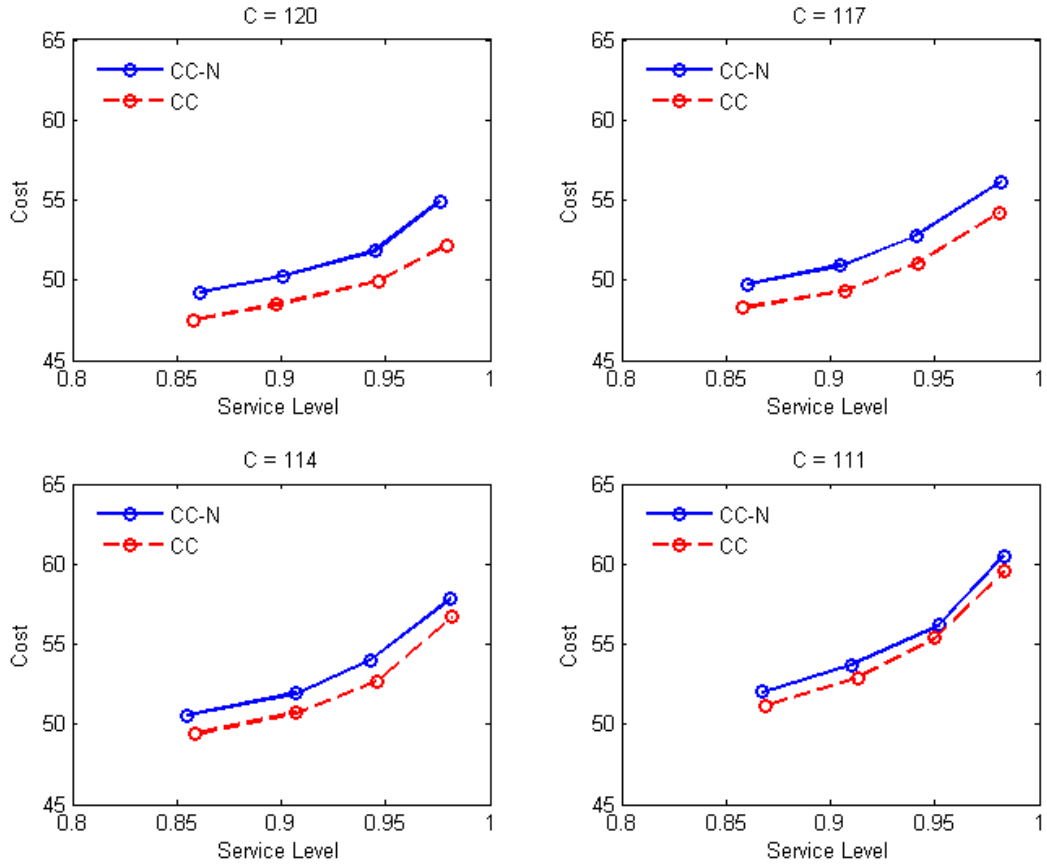

Figure 4.4 Impact of information

and workload dependent lead time. We first obtain a chance constrained formulation where the right hand side of the service level constraint depends on the evolution of demand forecast information. A numerical comparison with a model that does not use forecast evolution demonstrates when information would be most useful. The model indicates that the cost reduction increases with capacity limit and there is almost no benefit at high utilization. It also shows that the service level does not affect the cost reduction.

We then investigate the performance of the chance constrained model compared with that of a multistage stochastic programming model, for which we develop a method that produces a discrete joint distribution consistent with the MMFE. While the chance constrained model explicitly links the order releases to the load dependent lead times and capture shortfall due to congestion, the MSP model implicitly includes them through the scenarios. This approach appears to produce good solutions if enough scenarios are considered in the scenarios tree, which results in very large models relative to the CC model. Our experiments show the CC model even outperforms the MSP with five point normal approximation leading to 5^{T-1} scenarios. The accuracy of the chance constrained model increases with the system utilization and service levels.

The MSP model used in this work highlights the primary issue with multistage stochastic programming when applied to production planning: the size of the scenario tree grows very rapidly, resulting in very large formulations even when a very limited number of different demand realizations are considered in each period. There needs to be possible solution methods that will allow solving these problems in the presence of forecast evolution. Another extension of this work would be to study the multi-product model. Exploring this issue can reveal important insights about the impact of forecast evolution within demands of different products and across time periods.

Chapter 5

Conclusions and Future Directions

We conducted this research in three steps. In the first step, we assumed that there is no congestion in the production system and thus the lead time is fixed. The focus was on the forecast evolution. We showed that the progressive resolution of demand uncertainty through forecast updates results in evolution of the conditional covariance of the demand in addition to the evolution of its conditional mean. We demonstrated that these conditional covariances fluctuate over time until they become zero at the time of full realization of demand. This additional evolution provides a better characterization of demand behavior over time that leads to more effective decisions. The benefit of this approach is analyzed using a multi-period inventory planning model with additive forecast updates. We demonstrated how the system costs are reduced when the information updates are incorporated explicitly in the planning model. The model indicates that it is the relative difference between unconditional and conditional demand variance over the lead time that determines the magnitude of the cost reduction. It also shows that management can ignore advance information beyond the lead time in this uncapacitated system. Our results can also be used to quantify the benefits of

more advance information (earlier uncertainty resolution) that helps manager in their strategic decisions. We also showed that how our results can be extended for the multiplicative forecast updates by a mild assumption.

In the second step, we ignored the effect of forecast evolution on the production system. We examined production planning models that consider workload dependent lead time and stochastic demand in an integrated manner using clearing functions. We proposed a load dependent release policy leading to a tractable chance constrained model. The significance of our work lies in this analytical characterization, which provides a simple way to find and implement a near optimal policy. We studied a previously propose chance constrained model and a multistage stochastic programming model as benchmarks. A numerical comparison with these models quantifies the S-CC model's impact on the expected cost and service level under different business environment. While the S-CC model explicitly links the order releases to the load dependent lead times and capture shortfall due to congestion, the CC model does not consider these properties and the MSP model implicitly includes them through the scenarios. The CC model is thus completely dominated as the system utilization or the lead time variability increases. On the other hand the MSP model is sensitive to the number and choice of scenarios. The MSP approach appears to produce good solutions if enough scenarios are considered in the scenarios tree, which results in very large models relative to the CC and S-CC models. In our case, the S-CC model outperforms even the five point normal approximation.

In the third step, we incorporated the results of the first two steps to study the effect of forecast evolution on production planning with resources subject to congestion. We first

obtained a chance constrained formulation where the right hand side of the service level constraint depends on the evolution of demand forecast information. A numerical comparison with a model that does not use forecast evolution demonstrates when information would be most useful. The model indicates that the cost reduction increases with capacity limit and there is almost no benefit to considering forecast evolution at high utilization. It also shows that the service level does not affect the cost reduction. We then investigated the performance of the chance constrained model compared with that of a multistage stochastic programming model, for which we developed a method that produces a discrete joint distribution consistent with the MMFE. While the chance constrained model explicitly links the order releases to the load dependent lead times and capture shortfall due to congestion, the MSP model implicitly includes them through the scenarios. This approach appears to produce good solutions if enough scenarios are considered in the scenarios tree, which results in very large models relative to the CC model. Our experiments show the CC model even outperforms the MSP with five point normal approximation leading to 5^{T-1} scenarios. The accuracy of the chance constrained model increases with the system utilization and service levels.

Several aspects of this work merit further research. A natural extension would be to study the effect of multiplicative MMFE model on the production system subject to congestion. We developed the dependency evolution between demands for the multiplicative forecast updates in Chapter 2. However, care must be taken in the development of the conditional chance constraints and a method to generate scenario trees that are consistent with the multiplicative model. Exploring this issue can reveal important insights about the impact of forecast updates in a non-stationary demand model.

Another possible extension would be to study the dependency evolution for the multi-product model and then examine the impact of demand forecast information. Heath and Jackson (1994) developed a multi-product MMFE model. It is required to provide a model to show the dependency evolution within demands of different products and across time periods. Then one tractable approach is to use the multi-product production planning model of Asmundsson et al. (2009) and obtain the corresponding conditional chance constraints. An interesting question in this context is the value of forecast evolution within demands of different products and across time periods.

Finally, the MSP model used in this work highlights the primary issue with multistage stochastic programming when applied to production planning: the size of the scenario tree grows very rapidly, resulting in very large formulations even when a very limited number of different demand realizations are considered in each period. It is required to provide possible solution methods that will allow a scaling up of these approaches to problems of industrial size. The scenario-based structure of the MSP model makes decomposition methods attractive. Commonly used methods include Dantzig-Wolfe decomposition and Benders decomposition, which decompose the large scale problem into a master problem and several independent subproblems. However, there needs to be a systematic investigation of how many scenarios need to be considered to provide a reasonably good solution.

References

Anupindi, R., et al. (1996). "The nonstationary stochastic lead-time inventory problem: near-myopic bounds, heuristics, and testing." Management Science **42**(1): 124-129.

Aouam, T. and R. Uzsoy (2012). An Exploratory Analysis of Production Planning in the face of Stochastic demand and Workload-Dependent Lead Times. Decision Policies for Production Networks. K. G. Kempf and D. Armbruster, Springer.

Asmundsson, J. M., et al. (2009). "Production Planning Models with Resources Subject to Congestion." Naval Research Logistics **56**: 142-157.

Asmundsson, J. M., et al. (2006). "Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities." IEEE Transactions on Semiconductor Manufacturing **19**: 95-111.

Aviv, Y. (2001). "The Effect of Collaborative Forecasting on Supply Chain Performance." Management Science **47**(10): 1326-1343.

Aviv, Y. (2003). "A Time-Series Framework for Supply-Chain Inventory Management." Operations Research **51**(2): 210-227.

Beaulieu, N. C., et al. (1995). "Estimating the distribution of a sum of independent lognormal random variables." Communications, IEEE Transactions on **43**(12): 2869.

Birge, J. R. and F. Louveaux (1997). Introduction to Stochastic Programming. New York, Springer.

Bookbinder, J. H. and J. Y. Tan (1988). "Strategies for the Probabilistic Lot Sizing Problem with Service Level Constraints." Management Science **34**(9): 1096-1108.

Box, G. E. P. and G. M. Jenkins (1970). Time series analysis: forecasting and control. San Francisco, Holden-Day.

Buzacott, J. A. and J. G. Shanthikumar (1993). Stochastic Models of Manufacturing Systems. Englewood Cliffs, NJ, Prentice-Hall.

Byrne, M. D. and M. A. Bakir (1999). "Production Planning Using a Hybrid Simulation-Analytical Approach." International Journal of Production Economics **59**: 305-311.

Byrne, M. D. and M. M. Hossain (2005). "Production Planning: An Improved Hybrid Approach." International Journal of Production Economics **93-94**: 225-229.

Charnes, A. and W. W. Cooper (1959). "Chance-Constrained Programming." Management Science **6**(1): 73-79.

Chen, L. and H. L. Lee (2009). "Information Sharing and Order Variability Control Under a Generalized Demand Model." Manage. Sci. **55**(5): 781-797.

Ciarallo, F. W., et al. (1994). "A Periodic Review, Production Planning-Model with Uncertain Capacity and Uncertain Demand - Optimality of Extended Myopic Policies." Management Science **40**(3): 320-332.

Clark, A. J. and H. Scarf (1960). "Optimal Policies for a Multi-Echelon Inventory Problem." Management Science **6**(4): 475-490.

Dauzere-Peres, S. and J. B. Lasserre (1994). An Integrated Approach in Production Planning and Scheduling. Berlin, Springer-Verlag.

Dong, L. and H. L. Lee (2003). "Optimal Policies and Approximations for a Serial Multiechelon Inventory System with Time-Correlated Demand." Operations Research **51**(6): 969-980.

Escudero, L. F., et al. (1993). "Production Planning via Scenario Modelling." Annals of Operations Research **43**: 311-335.

Ettl, M., et al. (2000). "A Supply Chain Network Model with Base-Stock Control and Service Requirements." Operations Research **48**: 216-232.

Federgruen, A. and P. Zipkin (1986a). "An Inventory Model with Limited Production Capacity and Uncertain Demands I: The Average Cost Criterion." Mathematics of Operations Research **11**(2): 193-207.

Federgruen, A. and P. Zipkin (1986b). "An Inventory Model with Limited Production Capacity and Uncertain Demands II: The Discounted Cost Criterion." Mathematics of Operations Research **11**(2): 208-215.

Fenton, L. (1960). "The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems." Communications Systems, IRE Transactions on **8**(1): 57-67.

Gallego, G. and Ö. Özer (2001). "Integrating Replenishment Decisions with Advance Demand Information." Management Science **47**(10): 1344-1360.

Glasserman, P. (1997). "Bounds and asymptotics for planning critical safety stocks." Operations Research **45**(2): 244-257.

Graves, S. C. (1986). "A Tactical Planning Model for a Job Shop." Operations Research **34**(4): 522-533.

Graves, S. C. (1988). "Safety Stocks in Manufacturing Systems." Journal of Manufacturing and Operations Management **1**: 67-101.

Graves, S. C., et al. (1998). "A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization." Operations Research **46**(3): S35-S49.

Graves, S. C., et al. (1986). Two-Stage Production Planning in a Dynamic Environment. Multi-Stage Inventory Planning and Control. S. Axsater, C. Schneeweiss and E. Silver. Berlin, Springer-Verlag.

Güllü, R. (1996). "On the value of information in dynamic production/inventory problems under forecast evolution." Naval Research Logistics (NRL) **43**(2): 289-303.

Hackman, S. T. and R. C. Leachman (1989). "A General Framework for Modeling Production." Management Science **35**: 478-495.

Hausman, W. H. (1969). "Sequential decision problems - model to exploit existing forecasters." Management Science Series B-Application **16**(2): B93-B111.

Heath, D. C. and P. L. Jackson (1994). "Modeling the Evolution of Demand Forecasts with Application to Safety Stock Analysis in Production/Distribution Systems." IIE Transactions **26**(3): 17- 30.

Higle, J. L. and K. G. Kempf (2010). Production Planning under Supply and Demand Uncertainty: A Stochastic Programming Approach. Stochastic Programming: The State of the Art. G. Infanger. Berlin, Springer.

Hopp, W. J. and M. L. Spearman (2001). Factory Physics : Foundations of Manufacturing Management. Boston, Irwin/McGraw-Hill.

Høyland, K., et al. (2003). "A Heuristic for Moment-Matching Scenario Generation." Computational Optimization and Applications **24**(2): 169-185.

Huang, K. and S. Ahmed (2009). "The Value of Multistage Stochastic Programming in Capacity Planning Under Uncertainty." Oper. Res. **57**(4): 893-904.

Hung, Y. F. and R. C. Leachman (1996). "A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations." IEEE Transactions on Semiconductor Manufacturing **9**(2): 257-269.

Iida, T. and P. H. Zipkin (2006). "Approximate Solutions of a Dynamic Forecast-Inventory Model." Manufacturing \& Service Operations Management **8**(4): 407-425.

Irdem, D. F., et al. (2010). "An Exploratory Analysis of Two Iterative Linear Programming-Simulation Approaches for Production Planning." IEEE Transactions on Semiconductor Manufacturing.

Johnson, L. A. and D. C. Montgomery (1974). Operations Research in Production Planning, Scheduling and Inventory Control. New York, John Wiley.

Kacar, N. B., et al. (2010). An Experimental Comparison of Production Planning using Clearing Functions and Iterative Linear Programming-Simulation Algorithms. Research Report, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University.

Kacar, N. B., et al. (2012). "An Experimental Comparison of Production Planning using Clearing Functions and Iterative Linear Programming-Simulation Algorithms." IEEE Transactions on Semiconductor Manufacturing **25**(1): 104-117.

Kall, P. and S. W. Wallace (1994). Stochastic programming. Chichester ; New York, Wiley.

Karmarkar, U. S. (1989). "Capacity Loading and Release Planning with Work-in-Progress (WIP) and Lead-times." Journal of Manufacturing and Operations Management **2**(105-123).

Kim, B. and S. Kim (2001). "Extended Model for a Hybrid Production Planning Approach." International Journal of Production Economics **73**: 165-173.

Lautenschläger, M. and H. Stadtler (1998). Modelling Lead Times Depending on Capacity Utilization. Research Report, Technische Universitat Darmstadt.

Liu, L., et al. (2004). "Analysis and Optimization of Multi-stage Inventory Queues." Management Science **50**: 365-380.

Lu, X., et al. (2006). "Inventory Planning with Forecast Updates: Approximate Solutions and Cost Error Bounds." Operations Research **54**(6): 1079-1097.

Mehta, N. B., et al. (2007). "Approximating a Sum of Random Variables with a Lognormal." Wireless Communications, IEEE Transactions on **6**(7): 2690-2699.

Miller, A. C., III and T. R. Rice (1983). "Discrete Approximations of Probability Distributions." Management Science **29**(3): 352-362.

Missbauer, H. (2002). "Aggregate Order Release Planning for Time-Varying Demand." International Journal of Production Research **40**: 688-718.

Missbauer, H. and R. Uzsoy (2010). Optimization Models for Production Planning. Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook. K. G. Kempf, P. Keskinocak and R. Uzsoy. Norwell, MA, Springer.

Mula, J., et al. (2006). "Models for Production Planning Under Uncertainty: A Review." International Journal of Production Economics **103**: 271-285.

Orlicky, J. (1975). Material Requirements Planning: the New Way of Life in Production and Inventory Management. New York, McGraw-Hill.

Pahl, J., et al. (2005). "Production Planning with Load Dependent Lead Times." 4OR: A Quarterly Journal of Operations Research **3**: 257-302.

Prékopa, A. (1995). Stochastic programming. Dordrecht ; Boston, Kluwer Academic Publishers.

Pritsker, A. A. B. and K. Snyder (1997). Production Scheduling Using FACTOR. The Planning and Scheduling of Production Systems. A. Artiba and S. E. Elmaghraby, Chapman and Hall.

Purgstaller, P. and H. Missbauer (in press). "Rule-based vs. optimization-based order release in workload control: A simulation study of an MTO manufacturer." International Journal of Production Economics.

Rao, S. S., et al. (1998). "Waiting line model applications in manufacturing." International Journal of Production Economics **54**(1): 1-28.

Ravindran, A., et al. (2011). Production Planning with Load-Dependent Lead Times and Safety Stocks. Research Report, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University.

Riaño, G. (2003). Transient Behavior of Stochastic Networks: Application to Production Planning with Load-Dependent Lead Times. School of Industrial and Systems Engineering. Atlanta, GA, Georgia Institute of Technoogy.

Safak, A. (1993). "Statistical analysis of the power sum of multiple correlated log-normal components." Vehicular Technology, IEEE Transactions on **42**(1): 58-61.

Sen, S. and J. L. Higle (1999). "An introductory tutorial on stochastic linear programming models." Interfaces **29**(2): 33-61.

Srinivasan, A., et al. (1988). Resource Pricing and Aggregate Scheduling in Manufacturing Systems. Graduate School of Industrial Administration, Carnegie-Mellon University. Pittsburgh, PA.

Tarim, S. A. and B. G. Kingsman (2004). "The Stochastic Dynamic Production/Inventory Lot Sizing Problem with Service Level Constraints." International Joural of Production Economics **88**: 105-119.

Tayur, S. R. (1993). "Computing the Optimal Policy for Capacitated Inventory Models." Communications in Statistics: Stochastic Models **9**(4): 585 – 598.

Toktay, L. B. and L. M. Wein (2001). "Analysis of a Forecasting-Production-Inventory System with Stationary Demand." Management Science **47**(9): 1268-1281.

Tong, Y. L. (1990). "Multivariate normal distribution."

Voss, S. and D. L. Woodruff (2003). Introduction to Computational Optimization Models for Production Planning in a Supply Chain. Berlin ; New York, Springer.

Zipkin, P. H. (1986). "Models for Design and Control of Stochastic, Multi-Item Batch Production Systems." Operations Research **34**(1): 91-104.

Zipkin, P. H. (2000). Foundations of Inventory Management. Burr Ridge, IL, Irwin.