

A Tactical Planning Model for Make-To-Order Environment under Demand Uncertainty

by

Teo Chee Chong

B.Eng. (Hons) in Mechanical & Production Eng., Nanyang Technological University, 2000
S.M. in IMST, SMA, Nanyang Technological University, 2001

SUBMITTED TO THE SINGAPORE-MIT ALLIANCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN INNOVATION IN MANUFACTURING SYSTEMS AND TECHNOLOGY (IMST)
AT THE
SINGAPORE-MIT ALLIANCE

2006

Signature of author: _____
IMST Programme

May 31, 2006

Certified by: _____
Associate Professor Rohit Bhatnagar
SMA Fellow, NTU
Thesis Advisor

Certified by: _____
Professor Stephen C. Graves
SMA Fellow, MIT
Thesis Advisor

Accepted by: _____
Professor Yue Chee Yoon
Programme Co-Chair
IMST Programme

Accepted by: _____
Professor David E. Hardt
Programme Co-Chair
IMST Programme

A Tactical Planning Model for Make-To-Order Environment under Demand Uncertainty

by

Teo Chee Chong

SUBMITTED TO THE SINGAPORE-MIT ALLIANCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN INNOVATION IN MANUFACTURING SYSTEMS AND TECHNOLOGY (IMST)

ABSTRACT

We consider production smoothing in a make-to-order (MTO) manufacturing environment that has a firm delivery lead time. We model the fundamentals of the tactical planning process in the MTO environment. In the model, a stochastic demand process represents a dynamic input into the master production schedule. Consequently the schedule gets converted into job release, and the jobs flow between the workstations in the production system. We model the production system as a job shop as it represents the most generic form of process structure. Our job shop model is based on the Tactical Planning Model (TPM) of Graves (1986), in that it is a discrete time, continuous-flow model and uses linear production functions for the output of the workstations.

The TPM is a discrete-time model in which all transitions occur at the start of each time period. The time period must be defined appropriately in order for the model to be meaningful. We extend the TPM in Graves (1986) to address the model's dependency on the choice of time period and to facilitate the application of the model to production planning. In particular, we extend the linear control rule in the TPM into a continuous-time production function that justifies the underlying assumptions of continuous workflow and Markovian workflow, and also to match the time buckets in typical planning systems.

We examine the roles of the tactical planning parameters, namely the planning windows and planned lead times, and analyze the key considerations in setting these tactical parameters. We first develop a model for the production of a single aggregate product. Our model determines the first two moments of production requirements as well as the expected queue lengths, which are the performance measures to assess the appropriate setting of the planning parameters. We then extend the single-family model to the more generic model for systems with multiple families. We apply our model to a manufacturing facility that builds oil-rigs with the intent to test our model and validate its relevance to real-life industrial setting. We share our experience on how we carry out this application and show that our model is able to capture the essentials of this real-life scenario.

Keywords: make-to-order, production planning, tactical planning model, planned lead times, planning windows, MRP, master production schedule, job shop.

Thesis Advisors:

- 1) Associate Professor Rohit Bhatnagar, SMA Fellow, Nanyang Technological University
- 2) Professor Stephen C. Graves, SMA Fellow, Massachusetts Institute of Technology.

ACKNOWLEDGEMENTS

I like to thank my thesis supervisors, Professor Steve Graves and Associate Professor Rohit Bhatnagar, for their valuable advice and guidance, without which the completion of this thesis would be impossible. Much credit also goes to the other professors in my committee, Professor Lam Yee Cheong and Assistant Professor Jeremie Gallien, for their helpful comments.

I am also thankful to the doctoral students at SMA, especially Michelle, Lip Pin, Chin Hock, Yexin, Guo Xun, Tran Duc Vi and Ricardo, for their help in many ways as well as the many enjoyable lunch chats and afternoon breaks in the canteen. I am also grateful to Kek Sei Wee for his support and valuable inputs in the industrial case study.

I wish to thank Theresa for her patience and understanding while I spent considerable time on my work. Finally, I owe a great deal to my parents for their encouragement and support throughout this journey. This thesis is dedicated to them.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION	8
1.1 PRODUCTION SMOOTHING IN MAKE-TO-ORDER MANUFACTURING	8
1.2 TACTICAL PLANNING IN MAKE-TO-ORDER ENVIRONMENT	10
1.3 RELATED WORK	13
1.4 OVERVIEW OF MODEL	17
1.5 THESIS OUTLINE	18

CHAPTER 2

A TACTICAL PLANNING MODEL WITH CONTINUOUS-TIME CONTROL FOR PRODUCTION PLANNING SYSTEMS.....

2.1 INTRODUCTION	20
2.2 RELATED WORK	21
2.3 REVIEW OF THE TACTICAL PLANNING MODEL	23
2.4 LIMITATION OF THE TPM.....	27
2.5 MODEL FORMULATION	32
2.6 DISCUSSION.....	37

CHAPTER 3

A MULTISTAGE TACTICAL PLANNING MODEL FOR MAKE- TO-ORDER ENVIRONMENT WITH MULTIPLE PRODUCTS...43

3.1 INTRODUCTION	43
3.2 PROBLEM SETTING	43
3.2.1 Smoothing of Master Production Schedule	44
3.2.2 Smoothing at the Workstations	46
3.2.3 Problems in Production Planning.....	47
3.3 SINGLE-FAMILY MODEL.....	48
3.3.1 Model of MPS Smoothing	49
3.3.2 Model of Smoothing at Workstations.....	51
3.3.3 Workflow Model	53
3.4 MULTI-FAMILY MODEL	57
3.5 OPTIMIZATION MODEL	63
3.6 SUMMARY	65

CHAPTER 4

NUMERICAL EXAMPLE	66
4.1 SCENARIO	66
4.2 MODEL INPUTS.....	68
4.3 SETTING OF PLANNING WINDOWS AND SPLTs	70
<i>Case 0: Base Case.....</i>	<i>70</i>
<i>Case 1: Smoothing of MPS.....</i>	<i>71</i>
<i>Case 2: More Smoothing of MPS for Thin Plates.....</i>	<i>75</i>
<i>Case 3: Assessing Further Smoothing of MPS.....</i>	<i>75</i>
<i>Case 4: "What-if" Analysis.....</i>	<i>77</i>

CHAPTER 5

A CASE STUDY ON TACTICAL PLANNING IN OIL-RIG BUILDING	86
5.1 CASE BACKGROUND	86
5.2 PRODUCT AND PROCESS DESCRIPTION	87
5.3 PROBLEMS IN PRODUCTION PLANNING	89
5.4 MODEL APPLICATION	91
5.5 DATA.....	93
5.6 OPTIMIZATION	96
5.7 RESULTS.....	98
5.8 VALIDATION	102
5.9 DISCUSSIONS AND CONCLUSIONS.....	105

CHAPTER 6

CONCLUSIONS AND DISCUSSION.....	107
6.1 CONTRIBUTIONS	107
6.2 DIRECTIONS FOR FUTURE RESEARCH	109
6.2.1 <i>Stability of Master Production Schedule</i>	<i>109</i>
6.2.2 <i>Limitations of TPM</i>	<i>109</i>
6.2.3 <i>Assembly of MTO Parts.....</i>	<i>110</i>
6.2.4 <i>Non-Stationary Demand.....</i>	<i>110</i>
APPENDIX A.....	111
APPENDIX B	113
REFERENCES	116

LIST OF FIGURES

Figure 1.1	Production smoothing: Inducing a less variable production output ..9
Figure 1.2	MRP II Planning Hierarchy 11
Figure 2.1	Actual system from start of period t to t + 230
Figure 2.2	TPM from start of period t to t + 230
Figure 2.3	Production levels at Station i and Station j in actual system and TPM31
Figure 2.4	Graph of β , γ , and $\text{Var}(P_t)$ as functions of α (with $\sigma = 1$).....38
Figure 2.5	Graph of $\text{Var}(P_t)$ and $E[Q_t]$ as functions of α (with $\mu = 1$, $\sigma = 1$).....38
Figure 2.6	Comparison of $\text{Var}(P_t)$ between TPM and continuous-time model..39
Figure 2.7	Comparison of $\text{Var}(Q_t)$ between TPM and continuous-time model.41
Figure 3.1	Production smoothing by planning window and SPLT46
Figure 3.2	Modeling of MPS smoothing by dummy station.....50
Figure 3.3	Matrix $\text{Var}(D_{kt} + \xi_{kt})$ for each product family k.....58
Figure 3.4	Workflow matrix Φ_k for product family k59
Figure 4.1	Process Flow Map67
Figure 5.1	High-level BOM for hull87
Figure 5.2	Process flow map for panel production88
Figure 5.3	Sources of variability in panel production90
Figure 5.5	Demand for Small Panels over three-month period..... 103

LIST OF TABLES

Table 3.1 Smoothing of the MPS	45
Table 4.1 Data for Workstations	68
Table 4.2 Case 0: Original Case.....	73
Table 4.3 Case 1: Smoothing the MPS.....	74
Table 4.4 Case 2: More Smoothing of MPS for Thin Plates	79
Table 4.5 Case 3A: Allocating <i>SPLT</i> of NC Gas Cut to planning window	80
Table 4.6 Case 3B: Reallocating <i>SPLT</i> of Blasting to planning window.....	81
Table 4.7 Case 3C: Further reallocation of <i>SPLT</i> of Blasting to planning window	82
Table 4.8 Optimal Solutions.....	83
Table 4.9 Case 4A: “What-if” analysis - Reducing DLT of both Product Families	84
Table 4.10 Case 4B: “What-if” analysis – Increasing nominal capacity at Blasting	85
Table 5.1 Mean and standard deviation of demand	95
Table 5.2 Data for Model Inputs	95
Table 5.3 Inputs for noise terms	96
Table 5.4 Optimal Solution	98
Table 5.5 Comparison of optimal solution with other settings	101
Table 5.6 Summary of sensitivity analysis for demand	104

CHAPTER 1

INTRODUCTION

1.1 PRODUCTION SMOOTHING IN MAKE-TO-ORDER MANUFACTURING

A make-to-order (MTO) manufacturer builds each customer order as needed and carries no uncommitted finished goods inventory to meet customer demand. A firm uses this form of production when it offers numerous possible product configurations due to product customization, which makes it difficult to anticipate the exact needs of the customers. Examples of MTO production range from traditional manufacturing, e.g. job shops that fabricate metal parts, to high-tech capital-intensive manufacturing, e.g. a semiconductor fabrication facility that produces application-specific integrated circuits (ASIC). The paradigm in manufacturing is increasingly shifting from mass-produced standardized products to highly customized products. This emerging trend underlines the significance of MTO manufacturing.

One major challenge faced by most capacity-constrained make-to-order manufacturers is how to adjust production and work-in-process (WIP) inventory levels in response to stochastic demand fluctuations. A MTO manufacturer, unlike a make-to-stock manufacturer, does not keep uncommitted finished goods inventory; thus a MTO manufacturer is unable to use finished goods inventory to buffer against the demand fluctuations. In many MTO industries, the customers expect a market-set delivery lead time. Hence the MTO firm cannot adjust its delivery lead times to buffer against the demand fluctuations. As a result, the primary tactic available to a MTO firm for dealing with demand fluctuations is to flex its capacity. In periods of high demand, the MTO manufacturer might incur substantial overtime or subcontracting cost whereas it would experience an under-utilization of its production capacity when the demand is low.

To mitigate the costs of flexing the production capacity, we consider *production smoothing* which is an important aspect of production planning in a MTO

environment. Production smoothing deals with the setting of production and WIP inventory levels to meet a fluctuating demand so as to achieve a more efficient utilization of production resources. The aim of production smoothing is to induce a less variable (smoother) production output, as depicted in Figure 1.1.

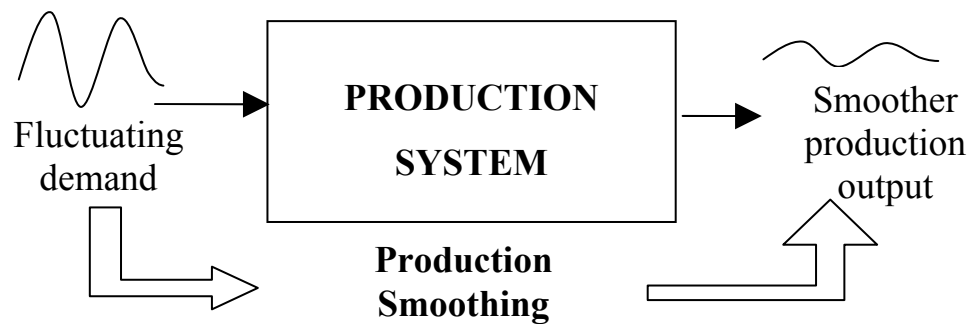


Figure 1.1 Production smoothing: Inducing a less variable production output

A common method used in MTO environments to smooth production is to maintain a constant production rate while quoting different delivery times to customers depending on the capacity load. The use of this method results in a longer delivery lead time in times of high demand when the capacity is heavily loaded. Therefore it is applicable only when the customers will tolerate the resulting longer delivery lead time. In today's increasingly customer oriented marketplace, time has become a crucial competitive dimension. The importance of delivery lead time in competitiveness has been well-documented (Thomas (1991), and Stalk and Hout (1990)).

Rather than varying the delivery lead time to accommodate varying demand, an alternative is to vary the production throughput. This requires some flexibility in its capacity, but will permit the MTO manufacturer to quote a fixed delivery lead time. Setting a firm delivery lead time might enable a MTO manufacturer to compete more effectively in the market. This firm lead time is made known to all potential customers, where the length of the lead time is set to be achievable with high likelihood by the manufacturer. The manufacturer makes every effort to quote the

firm lead time to every customer whenever its capacity allows, as quoting a longer lead time would result in a loss of goodwill or would lead to lost sales.

Furthermore in many markets, customers often do not tolerate long delivery lead times. It is common for industrial products, e.g. components, subassemblies, production equipment, to have firm delivery lead times; customers of such products often do not tolerate long lead times as it will lengthen their own production lead time. As a result, the MTO manufacturer has little or no flexibility to vary the delivery lead time. In addition, a non-variable delivery lead time for industrial products is beneficial to the customer as it adds significant predictability to the customer's own production planning, as opposed to a variable lead time which can be disruptive to the customer's schedule. Hence it is common for a MTO company to set a firm delivery lead time to facilitate the customer's production planning, which will in turn helps to establish good relationship with the customer.

In this research, we consider production smoothing in a MTO environment where the delivery lead time is firm. In particular, we look at production smoothing from the perspective of production planning at the tactical level, where our objective is to minimize the cost of flexing capacity and other related production costs. In the next section, we examine the key planning parameters that are vital to production smoothing in the MTO setting.

1.2 TACTICAL PLANNING IN MAKE-TO-ORDER ENVIRONMENT

Our research focuses on the tactical planning decisions to smooth production in a MTO environment. The tactical planning decisions involve a medium range planning horizon, e.g. allocation of production resources, aggregation of items into product families and setting of parameters that have impact on production in the medium range horizon. Typically, the two core planning modules at the tactical planning level are the *master production schedule* (MPS) and the *material requirements planning* (MRP).

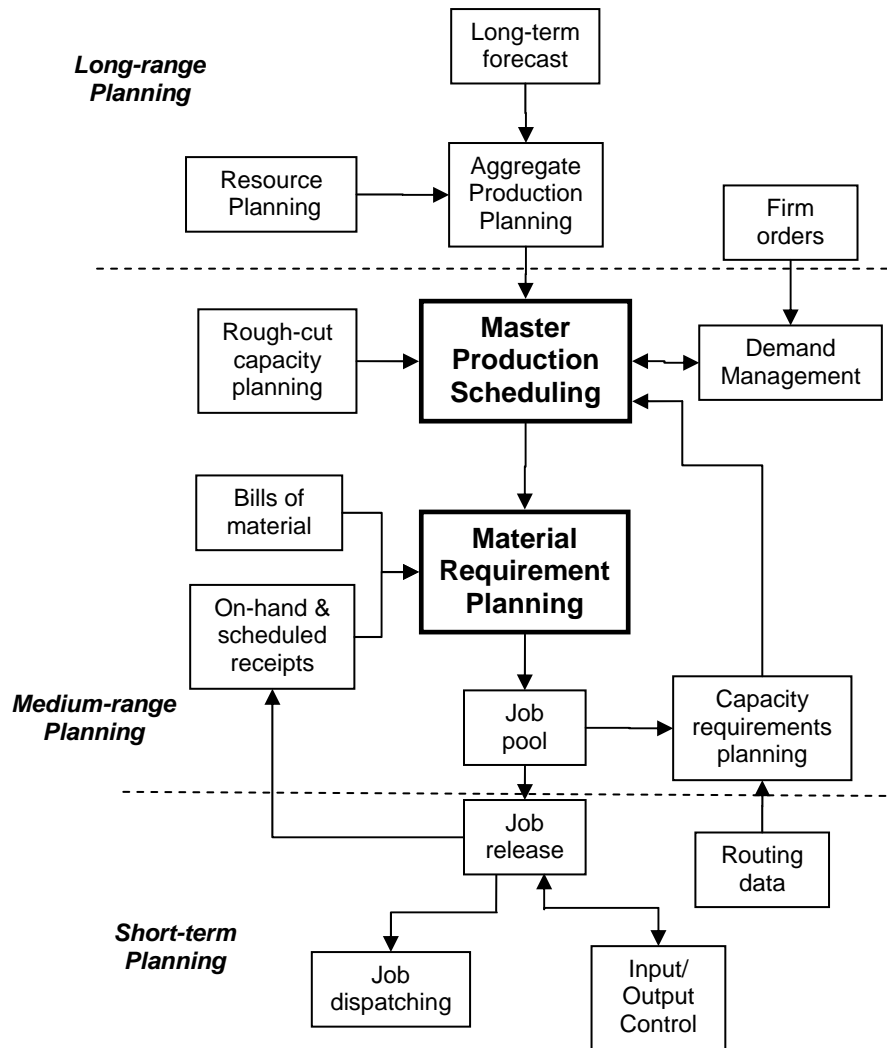


Figure 1.2 MRP II Planning Hierarchy

Figure 1.2 shows an example of the planning hierarchy for a *Manufacturing Resources Planning (MRP II)* system (Hopp and Spearman (2000)). In a MTO system, the MPS takes the firm customer orders from the demand management module as inputs. It then generates a specific manufacturing schedule of the necessary production outputs to meet the customer orders. In particular, the MPS provides production quantities for completion in each time bucket of the planning horizon. The rough-cut capacity planning and capacity requirements planning modules serve to check that the planned production quantities are within the aggregate capacity limitations. In MTO environments where there are numerous product configurations, the usual practice is to plan for aggregate items, i.e. groups of items that share similar

design and production process, instead of end items. The independent demands of the aggregate items from the MPS are inputs to the MRP module. Based on the bill of material for the aggregate item (bills of material module) as well as information on on-hand and on-order production quantities (on-hand and scheduled receipts), the MRP module in turn generates the time-phased planned production quantities and order release for downstream production stages. Vollmann et al. (2005) and Hopp and Spearman (2000) provide more detailed explanations on the individual modules of the planning hierarchy.

The main concerns in the tactical planning of the MPS and MRP include schedule stability, lot-sizing decisions, deciding the product aggregates for which to state the MPS, and setting the planned lead time at each production stage. In our context of MTO environments where production smoothing is crucial, we look at the two planning parameters that have a large influence on production smoothing, which are the planning windows (which we will discuss in greater detail in *Chapter 3*) and the planned lead time at each production stage.

The planned lead time (PLT) is the input parameter set by the production planner and is the key information required to perform the MRP processing. The PLT is used to determine job start times, which is equal to the due-date of the job at the stage minus the PLT. Thus, if the actual lead times are always exactly equal to the PLTs, MRP would result in jobs being ready exactly when needed. However, actual lead times vary unpredictably due to variability and uncertainties (e.g. processing times variability, fluctuating capacity loading, machine breakdowns). Therefore selecting an appropriate PLT for the MRP system can be a difficult task. There is a tradeoff involved in setting the PLT. On the one hand, if the PLT is short, it becomes difficult to smooth the production by leveling the workload over the lead time; consequently, production capacity might not be able to keep up with production in times of high customer demand. On the other hand, the longer the PLT is, the longer the jobs will wait for the next operation, and so the more WIP there will be in the system.

Besides the PLT at each production stage, we also need to consider the smoothing of the MPS. A smoother MPS implies a less fluctuating work release into the production system, which in turn results in a less varying load on capacity. In the MTO

manufacturing, the smoothness of the MPS depends on the planning window. The planning window is the difference between the delivery lead time and the planned lead time for a product. A longer planning window represents a greater degree of flexibility available to smooth the MPS. Thus the setting of the planning window of the MPS is an important tactical planning decision. In *Chapter 3*, we discuss in greater depth the importance of both the PLT and planning window on production smoothing in the MTO environment.

1.3 RELATED WORK

Aligned with our main objective of smoothing production in the MTO environment, we focus on literature that investigates production smoothing in the MTO environment as well as the setting of PLT in MRP systems.

Much of the literature in the area of production smoothing in MTO manufacturing is on the quoting of due-dates for customer orders placed against a known capacity. The research on due-date setting is closely related to the core function of the *demand management* module of MTO planning systems, which is to convert orders into promised dates and to balance demand with supply.

Most work considers either performance measures or penalty costs, e.g. tardiness of jobs and average delivery lead time, associated with the objective to maintain a certain service level (e.g. proportion of late jobs). Baker and Bertrand (1981) compare three due-date setting rules for a single-machine model using simulation. They conclude that setting due-dates on the basis of total workload gives the tightest average due-dates. There is a substantial amount of research on due-date setting that focuses on the combined problem of quoting due-dates and sequencing orders for a single machine. Seidmann and Smith (1981), Bertrand (1983) and Chand and Chhajed (1992) develop analytical models that select the optimal due-dates and processing sequence for a fixed number of jobs. Baker (1984) investigates the interaction between due-date assignment methods with sequencing rules, and also suggests the most effective combination for different operating objectives. Wein (1991) performs a simulation study with the objective to minimize average delivery lead times by

considering both due-date setting policies and sequencing rules; more importantly, the research suggests that proper due date setting offers a much larger improvement in performance than priority sequencing. His research highlights the importance of tactical planning; due-date setting policy is a tactical planning decision in contrast to the operational decision of job sequencing. Appropriate decisions made at the tactical level lead to well-designed systems that are much easier to control at the operational level than poorly designed systems. Tactical planning is precisely the focus of this research.

More recently, there is some work that explicitly models the tolerance of customers towards long delivery lead time. Most work assumes each customer (or customer class) has a known level of tolerance towards long lead time. The research in this area looks at selecting both due-dates and processing sequences for all potential customers with the objective to maximize profits. Duenyas and Hopp (1995) analyze a queuing model in which lead time quotation affects the demand rate. Duenyas (1995) develops an effective heuristic for quoting due-dates and sequencing orders. Keskinocak et al. (2001) propose a method that dynamically quotes due-dates with emphasis on meeting quoted due-dates. There is some recent work that considers the additional dimension of price. Here the customers are sensitive to both delivery lead time and price, and the manufacturer can use price incentives to generate orders with different lead times to maximize profits as well as to smooth production. So and Song (1998) develops a static model for setting both price and due-date, while Plambeck (2004) develops a model that sets static prices, and dynamically quotes due-dates and sequence orders.

The literature discussed above considers a business environment where the MTO manufacturer enjoys some flexibility (albeit with a service level constraint) to vary the delivery lead times depending on the capacity load. Our research is different in the way that we consider an exogenous firm delivery lead time that is set by the market. In addition the research on due-date setting focuses on the customer interface of quoting due-dates for a single-stage system, which is a key component of demand management. However our research looks into the production planning aspect of production smoothing that is subject to the constraint of a non-variable delivery lead

time. In particular, we look at the tactical planning of the MPS and MRP to smooth production for a multi-stage production system.

There is a limited amount of research that deals with production smoothing in MTO environments where customers require a firm delivery lead time. One effort is by Cruickshanks et al. (1984) in which they discuss production smoothing in a single-stage manufacturing process. The work is motivated by a case study at a manufacturing firm that produces electromagnetic instruments whose customers cannot tolerate late deliveries. The deliveries to customers are promised to be within a firm delivery lead time of L_d periods. Given that the production lead time is L_p where $L_d \geq L_p$, the value $L_d - L_p + I$ is called the *planning window*. They find that the variation in production declines as the planning window increases, and that a small increase in the planning window can give a large reduction in the production variation. However, a larger L_d leads to more inventory and longer delivery lead times for customers. Our model of the MPS smoothing is closely related to their work such that both smoothes production by means of a linear function of the order backlog.

Another related research topic to our work is the setting of PLTs in MRP systems. The literature on this topic consists of two main categories: one category encompasses studies (mostly simulation models) to investigate the relationship between the PLTs and other planning parameters; the other covers the analytical models that quantitatively determine the optimal PLTs.

In the first category, Whybark and Williams (1976) and Grasso and Taylor (1984) carry out simulation studies to compare the alternatives of safety stock and safety lead time in buffering against uncertainties in make-to-stock systems. Buzacott and Shanthikumar (1994) conclude that the safety time is preferred to safety stock if an accurate forecast is possible; otherwise safety stock is more robust in coping with changes in orders or forecasts. Other work involves investigating the relationship between the PLT and inventory carrying cost (e.g. Marlin (1986), Penlesky et. al (1989)), and between the PLT and lot-sizing rules (e.g. Melnk and Piper (1985)). Mohan and Ritzmans (1998) simulate a variety of different operating conditions and take into account many factors, including service level, inventory level, due-date tightness, lot-sizing and product structure.

In the category of analytical models to which our model belongs, Weeks (1978) proposes a one-period inventory model to set the optimal PLT and suggests that the PLT is a function of the cost of tardiness, delivery lead time and production lead time distribution. Kanet (1986) recommends that PLTs be set such that time allowances are set proportional to total work content. Yano (1987) determines the optimal PLTs in a serial production system. The objective is to minimize the expected sum of inventory holding costs and job tardiness costs given a specified due date. With identical objectives and considering a serial production line as well, Gong et al. (1994) determine the optimal PLTs by solving an equivalent serial multi-echelon inventory model. Adenso-Diaz and Laguna (2001) look into the challenge of identifying the best changes in an infeasible master production schedule in order to maintain a smoother load profile. They develop optimization models to support the planner's task of resolving capacity infeasibilities for a single-stage system.

Graves (1986) presents the Tactical Planning Model (TPM) for a job shop with multiple workstations. The TPM characterizes the interrelationship between PLT, work-in-process inventory and production requirements in a job shop. The model highlights the tradeoff between production smoothness and WIP by the selection of PLT at each workstation. Our model of workflow between processing stations is based on the TPM. Furthermore, similar to the TPM, our model is a discrete time, continuous-flow model and uses a linear production function (which is a function of the PLT) for the output of the workstations. We discuss other related work to the TPM in section 2.2.

All the analytical models that are discussed above for prescribing the PLTs assume an exogenous job arrival process into the production system and/or are restricted to single-stage or serial-flow systems. In contrast, we explicitly model the planning process of generating the MPS and work releases from the demand. Furthermore, we model the production system as a job shop that consists of a general network of workstations with no restriction on flow paths or process routes. Therefore we believe that our model is richer in its applicability to the tactical planning of MTO production.

Finally, we note that our work complements the work by Graves et al. (1998), who develop a model for studying requirements planning in a multi-stage make-to-stock system. They study how the translation of an evolving forecast into the MPS affects the tradeoff between production smoothness and inventory. But unlike their work, we look at the MTO environment for which the smoothness-inventory tradeoff is determined by the PLTs.

1.4 OVERVIEW OF MODEL

In this thesis, we model the basics of the tactical planning process in a MTO environment for the purpose of studying production smoothing within the planning system. In the model, we translate the stochastic demand process into the master production schedule. Subsequently the schedule gets converted into job release, and the jobs flow between the workstations in the production system as governed by a flow matrix.

We model the production system as a job shop. The job shop consists of a network of workstations, produces a variety of products and has multiple process routes through the shop. We model the system as a job shop as it represents the most generic form of process structure and thus, we expect our model to cover a broad spectrum of production systems. Our job shop model is closely related to the Tactical Planning Model (TPM) of Graves (1986), in that it is a discrete time, continuous-flow model and uses linear production functions for the output of the workstations.

In this research, we extend the TPM in Graves (1986) to facilitate the application of the model to production planning. The TPM is a discrete-time model in which all transitions occur at the start of each time period. In the TPM, the time period must be defined appropriately in order for the model to be meaningful. Each period must be short enough so that a job is unlikely to travel through more than one station in one period. At the same time, the time period needs to be long enough to justify the assumptions of continuous workflow and Markovian job movements, and also to match the time buckets in planning systems (which are usually in days or weeks). We

extend the linear control rule in the TPM into a continuous-time production function to address its dependency on the choice of time period.

Given a fixed delivery lead time, our model determines the variance of the production requirements at each workstation, which is a measure of production smoothness. In addition, the model computes the expected WIP inventory level at each workstation, which is the performance measure for the inventory level. As such, we generate two performance measures that allow us to look into the tradeoff between production smoothness and inventory. We can make use of these performance measures to evaluate the two key planning parameters that affect production smoothness, i.e. the planning window in the MPS and the planned lead times at each workstation. We also embed the model in optimization procedures to determine the optimal planned lead times and planning windows. We show that the optimization model is a convex nonlinear program which can be solved easily using commercial optimization software. We carry out a case study at a large facility in the marine/offshore industry that builds oil-rigs, in which we test our model and validate its relevance to real-life industrial setting.

1.5 THESIS OUTLINE

The outline of the remaining chapters is as follows:

Chapter 2: A Tactical Planning Model with Continuous-Time Control for Production Planning Systems

We extend the Tactical Planning Model of Graves (1986) to improve its applicability to production planning. We first give a literature review of the work related to the TPM and then present a review of the model. Subsequently, we illustrate the limitation of the TPM due to the period sizing restriction that impedes the applicability of the model to production planning. We then derive a continuous-time linear control rule for a single-station system that relaxes this restriction, and also determine the first two moments of the production requirement and queue length.

Chapter 3: A Multistage Tactical Planning Model for Make-to-Order Environment with Multiple Products

We first define the problem setting and discuss the importance of the PLT and planning window for production smoothing in the MTO environment. We then model the smoothing of the MPS as well as the production at the workstations for a single product family. We employ the continuous-time control rule derived in *Chapter 2* for the model and extend the single-station model to a multi-station model. Next we explain how we extend the model to incorporate multiple product families. We conclude this chapter by setting up an optimization program to determine the optimal PLTs and planning windows with the objective to minimize expediting cost (e.g. overtime and subcontracting costs) and WIP inventory holding cost.

Chapter 4: Numerical Example

We demonstrate the use of our model by a numerical example, which is based on a real-life production shop that processes steel plates. We also illustrate the effects of the PLTs and planning windows on production smoothness and WIP inventory levels. We also show how the model can be applied in “what-if” analysis for tactical planning.

Chapter 5: A Case Study on Tactical Planning in Oil-Rig Building

We perform a case study on a MTO company that manufactures offshore oil-rigs. This study serves as a platform to validate our model and test its relevance to the real-world production setting. We first explain the problems faced by the company and illustrate how we apply our model to improve its tactical planning. We then discuss the benefits that our model brought upon to the company as well as the shortcomings of the model that we identified from this study.

Chapter 6: Conclusions and Future Research Directions

We give a review of the thesis and then discuss the opportunities for future research.

CHAPTER 2

A TACTICAL PLANNING MODEL WITH CONTINUOUS-TIME CONTROL FOR PRODUCTION PLANNING SYSTEMS

2.1 INTRODUCTION

In this research, we consider production smoothing in a make-to-order (MTO) manufacturing environment that has a firm delivery lead time. We consider the manufacturing system as a job shop that produces a variety of products and has multiple process routes. We model the MTO production system as a job shop that is based on the Tactical Planning Model (TPM) of Graves (1986). However, in order to employ the TPM for this research, we first need to remove the limitations of the TPM that hinder its applicability to production planning.

The TPM is a discrete-time model in which all transitions within the model are governed by an underlying time period. The model assumes that all movement of jobs occurs at the start of each time period. As such, one must set the time period to be short enough so that it is unlikely for one job to travel through two successive stations in one time period. This restriction of period sizing hinders the application of the TPM to model the MTO production planning. This is because parameters in most planning systems, such as demand requirements, are defined in daily or weekly time buckets, which can be much longer than the restricted time period. Thus the inability to set a longer time period impedes the use of the TPM in production planning.

The TPM does not explicitly model the flow of discrete jobs, but rather models the flow of work due to the jobs. Work completed in the current period flows to downstream stages in the next period. However, in discrete manufacturing, each job is only transferred to the downstream station upon completion. So in order to accurately model the job movement, the time period for the TPM should preferably be long to increase the “fluidity” of the flow of the discrete jobs.

Furthermore, the TPM assumes a Markovian workflow, i.e. the arrivals to a workstation do not depend on the history of the workflow. The validity of this assumption depends on whether each workstation in the shop produces a stable mix of jobs. If many jobs can be completed in one period, then it is more likely that there is a stable output; but this also argues for a longer time period. In addition, we can approximate the distribution of the production requirement to be normally distributed if the time period is sufficiently long for the Central Limit Theorem to apply. As such, we are able to characterize the distribution of the requirement since we can compute both the mean and variance of the production requirement using the TPM.

This chapter considers an extension to the tactical planning model of a job shop by Graves (1986). We extend the linear control rule in Graves (1986) into a continuous-time production function that facilitates the application of our model to production planning. We build an extension to the TPM that overcomes this restriction of period sizing by permitting production control over shorter time intervals. At the same time, the continuous-time production function permits jobs to complete processing at more than one station within a planning period. Thus we can set the planning period to be sufficiently long so as both to model the work flow as a continuous quantity, and to match the time buckets in most planning systems.

In the next section, we give a literature review of work related to the TPM. We then present a brief review of the TPM in section 2.3. Subsequently we illustrate in section 2.4 the limitation of the TPM due to the restriction of sizing the time periods. In section 2.5, we derive the continuous-time control rule that will relax this restriction for a single-station system, and also determine the first two moments of the production quantity and queue length. In section 2.6, we compare the continuous-time model with the TPM, and discuss the model and its key assumptions.

2.2 RELATED WORK

Graves (1986) develops the TPM as a tactical planning tool for job shop operations. The TPM is a discrete-time linear-system model that determines the first two

moments of the production and queue levels, given the planned lead times of the workstations. The model tracks the workload at each station rather than the individual jobs; the model assumes that the volume of work arrivals at a station are in fixed proportions of the work completed at upstream stations.

To date, there are several extensions to the TPM. Parrish (1987) proposes a framework for modeling work releases to meet the delivery due date for a finished product. In addition, he also shows how to adjust the control parameters of the TPM to change service measures in meeting demand. Graves (1988) presents three extensions to a single-station model of the TPM. First, he models a station that fails according to a Bernoulli process. Second, he incorporates variability due to lot-sizing, and finally, he presents the mathematical bounds on a single station with a capacity constraint. Mihara (1988) extends the work of Graves (1988) when he looks at an unreliable multi-station TPM. Similar to Graves' work, the stations fail according to a Bernoulli process. Hollywood (2001) demonstrates how to calculate approximations for the steady-state moments of TPM with general non-linear control rules. His model allows for the modeling of machine congestion due to capacity loading.

There is some work that applies the TPM to model actual manufacturing systems. Graves (1988) presents a model to provide a rough-cut assessment of both the staffing and component inventory levels of a repair depot. The repair rate of the depot has a production function that resembles the linear control rule of the TPM. Fine and Graves (1989) test the TPM on a real-life job shop when they apply the TPM to a shop that manufactures thermal conduction modules for mainframe computers. Here, the model is extended to allow consideration of features such as release policies. The model is then used to study the impact of various planning policies and the effect of changes in product mix.

Other efforts adapt the TPM to pull systems. Leong (1987) models a Kanban control system and other pull systems using the TPM in which work is produced at a station whenever there is a downstream inventory shortfall. Graves and Hollywood (2001) develop a constant-inventory TPM in which the release of work into the shop is regulated to maintain a constant inventory level. They then illustrate the use of the

model with an application and show the benefits of such a release policy with a computational experiment.

2.3 REVIEW OF THE TACTICAL PLANNING MODEL

The tactical planning model (TPM) is a discrete-time, continuous flow model. All transitions within the model occur at the start of each time period, and the jobs are modeled as workload measured in time units (e.g. hours). The workflow is assumed to have a Markov property: that is, the processing requirements at a station do not depend on how work got to the station. As such, each individual job has no identity.

Central to the TPM model is the linear control rule, which is stated as

$$P_{it} = \alpha_i Q_{it} \quad (2.1)$$

where P_{it} is the amount of production completed by work station i in time period t , Q_{it} is the queue level at the start of period t , and the parameter $\alpha_i, 0 < \alpha_i \leq 1$, is a smoothing parameter. This rule states that the production P_{it} at workstation i is a fixed portion (α_i) of the queue of work Q_{it} at the start of the period. In particular, $1/\alpha_i$ represents the number of periods, on average, the work requires to move through the work station. We interpret $1/\alpha_i$ to represent the planned lead time. We can view the control rule in (2.1) as a prescriptive equation, i.e. to preserve the integrity of the planned lead time, we must shift capacity to heavily loaded stations. But (2.1) can also be considered as a descriptive equation where production resources are naturally flexed to accommodate the varying workloads at the stations.

The queue level Q_{it} satisfies the standard inventory balance equation

$$Q_{it} = Q_{i,t-1} - P_{i,t-1} + A_{it} \quad (2.2)$$

where A_{it} is the amount of work that arrives at workstation i at the start of period t . By substituting (2.1) into (2.2), we obtain a first-order smoothing equation with α_i as the smoothing parameter:

$$P_{it} = (1 - \alpha_i)P_{i,t-1} + \alpha_i A_{it} \quad (2.3)$$

By repeated substitution and assuming an infinite history of arrivals, we find that

$$P_{it} = \sum_{s=0}^{\infty} \alpha_i (1 - \alpha_i)^s A_{i,t-s} \quad (2.4)$$

We assume that the arrival stream $\{A_{it}\}$ to station i are i.i.d. with mean μ and variance σ^2 . We then obtain

$$E[P_{it}] = \mu \quad (2.5)$$

and

$$Var(P_{it}) = \frac{\alpha_i \sigma^2}{2 - \alpha_i} \quad (2.6)$$

To obtain the first two moments of the queue level, we first substitute (2.4) into (2.1) to get

$$Q_{it} = \sum_{s=0}^{\infty} (1 - \alpha_i)^s A_{i,t-s} \quad (2.7)$$

From (2.7), we obtain

$$E[Q_{it}] = \frac{\mu}{\alpha_i} \quad (2.8)$$

and

$$\text{Var}(Q_{it}) = \frac{\sigma^2}{2\alpha_i - \alpha_i^2} \quad (2.9)$$

However, if we consider a network of workstations, the arrival stream to a workstation from upstream stations are generally not i.i.d. but are correlated over time. We now proceed to develop the model for a multi-station network that takes into account of correlated arrivals. Each workstation can receive two types of arrivals; one type of arrival consists of jobs that have their first processing step at the station, while the other type of arrival consists of in-process jobs that have just completed processing at an upstream station. We model the arrivals to station i from another station j by the equation:

$$A_{ijt} = \phi_{ij}P_{j,t-1} + \varepsilon_{ijt} \quad (2.10)$$

A_{ijt} is the flow of work arriving at station i from station j at the start of period t , ϕ_{ij} is a positive scalar and ε_{ijt} is a random variable. We assume that one unit (e.g. hour) of work at station j will trigger, on average, ϕ_{ij} time units of work at station i . The variable ε_{ijt} is a noise term that models uncertainty between production at j and arrivals to i , and is assumed to be an *i.i.d.* random variable with zero mean and a known variance.

The arrival to station i is given by

$$A_{it} = \sum_j A_{ijt} + N_{it} \quad (2.11)$$

where N_{it} is an i.i.d. random variable for the workload from new jobs that enter the shop at station i at time t . Substituting for A_{ijt} , we obtain

$$A_{it} = \sum_j \phi_{ij} P_{j,t-1} + \varepsilon_{it}, \text{ where } \varepsilon_{it} = N_{it} + \sum_j \varepsilon_{ijt} \quad (2.12)$$

The term ε_{it} represents arrivals that are not predictable from the production levels of the previous period, and consists of work from new jobs and noise in the flow. By assumption, the time series ε_{it} is independent and identically distributed over time. However the non-predictable arrivals for two different stations can be correlated.

We can restate the equations for production (2.3) and for arriving work (2.12) in matrix-vector form:

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{D})\mathbf{P}_{t-1} + \mathbf{D}\mathbf{A}_t, \quad (2.13)$$

$$\mathbf{A}_t = \mathbf{\Phi}\mathbf{P}_{t-1} + \boldsymbol{\varepsilon}_t \quad (2.14)$$

where $\mathbf{P}_t = \{P_{1t}, \dots, P_{mt}\}'$, $\mathbf{A}_t = \{A_{1t}, \dots, A_{mt}\}'$, and $\boldsymbol{\varepsilon}_t = \{\varepsilon_{1t}, \dots, \varepsilon_{mt}\}'$ are column vectors of random variables, m is the number of workstations, \mathbf{I} is the identity matrix, \mathbf{D} is a diagonal matrix with $\{\alpha_1, \dots, \alpha_m\}$ on the diagonal, and $\mathbf{\Phi}$ is an m -by- m matrix with elements ϕ_{ij} . By substituting equation (2.14) into equation (2.13), we find that

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{D} + \mathbf{D}\mathbf{\Phi})\mathbf{P}_{t-1} + \mathbf{D}\boldsymbol{\varepsilon}_t \quad (2.15)$$

By iterating this equation and assuming an infinite history of the system, we rewrite \mathbf{P}_t as an infinite series

$$\mathbf{P}_t = \sum_{s=0}^{\infty} (\mathbf{I} - \mathbf{D} + \mathbf{D}\mathbf{\Phi})^s \mathbf{D}\boldsymbol{\varepsilon}_{t-s} \quad (2.16)$$

The mean and the covariance for the noise vector $\boldsymbol{\varepsilon}_t$ are denoted by $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}'$, and $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$ respectively. The first two moments of \mathbf{P}_t are given by

$$\begin{aligned}
E[\mathbf{P}_t] &= \sum_{s=0}^{\infty} (\mathbf{I} - \mathbf{D} + \mathbf{D}\Phi)^s \mathbf{D}\boldsymbol{\mu} \\
&= (\mathbf{I} - \Phi)^{-1} \boldsymbol{\mu}
\end{aligned} \tag{2.17}$$

and

$$\mathbf{S} = \text{var}(\mathbf{P}_t) = \sum_{s=0}^{\infty} \mathbf{B}^s \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}\mathbf{B}'^s \tag{2.18}$$

where $\mathbf{B} = \mathbf{I} - \mathbf{D} + \mathbf{D}\Phi$

We note that \mathbf{S} provides the variance of the production requirements for each station, as well as the covariance for each pair of workstations. In addition, we determine the first two moments of the queue levels. From (2.1), we note that

$$\mathbf{Q}_t = \mathbf{D}^{-1} \mathbf{P}_t \tag{2.19}$$

Therefore we have

$$E[\mathbf{Q}_t] = \mathbf{D}^{-1} E[\mathbf{P}_t] \tag{2.20}$$

and

$$\text{Var}(\mathbf{Q}_t) = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}^{-1} \tag{2.21}$$

The infinite series in equations (2.17), (2.18), (2.20) and (2.21) converge provided that $\rho(\Phi) < 1$, where $\rho(\Phi)$ denotes the spectral radius of Φ (see Graves (1986)).

2.4 LIMITATION OF THE TPM

In the TPM, all job movements occur at the start of each time period. In order to model movement of jobs in the actual shop, we are restricted to set the period length to be short enough so that it is highly improbable for one job to travel through more than one station in a single time period.

However, due to the continuous-flow assumption, the time period should be long relative to the workload of the individual discrete jobs. This will increase the “fluidity” of the discrete jobs and will make the continuous-flow assumption more reasonable. Now we use an example to illustrate the discrepancies between the model and the actual system due to the above contradictory objectives in period sizing.

To simplify our illustration, we consider a simple system that consists of two stations in series, namely Station i and Station j . We further assume that the planned lead time of Station i is 2 hours, while that of Station j is 1 hour. We set the length of the time period to be 1 hour, as we assume that it is unlikely for a job in Station i to travel beyond Station j in 1 hour. Now consider a job that enters the empty system and arrives at Station i at the start of period t . We suppose that the job has a processing time of 2 hours at Station i , and 1 hour at Station j . We illustrate and compare the sequence of events for the actual system and the TPM from period t to $t + 2$.

Figure 2.1 shows the actual system from the start of period t to $t + 2$. The job arrives at Station i at the start of period t . Since the processing time is 2 periods, the job is processed at Station i till the end of $t + 1$. The job is then transferred to Station j at the start of $t + 2$. It is then processed at Station j till the end of period $t + 2$ since the processing time is 1 period.

Now we look at the same scenario in the context of the TPM. Suppose that both Stations i and j produce according to the TPM control rule in equation (2.1). In this case, given the planned lead time of each station, we have $\alpha_i = \frac{1}{2}$ and $\alpha_j = 1$. Job movements between the two stations are modeled by equation (2.10). We assume that the term $\phi_{ij} = 0.5$ given the processing times of the job, and $\varepsilon_{ijt} = 0$. Figure 2.2 illustrates the workflow in the TPM from the start of period t to $t + 2$. As shown in the figure, Station i processes half of the in-queue workload at the start of each period ($\alpha_i = \frac{1}{2}$), while Station j processes the entire workload ($\alpha_j = 1$). And the workload processed by Station i in each period generates half the workload at Station j at the start of the next period ($\phi_{ij} = 0.5$).

In the actual system, the discrete job moves to the downstream station only upon completion. However, in the TPM, work flows as a fluid to the downstream station even if the discrete job is not completed. In this example, the discrete job is “split” up and moved in parts to the downstream station. This is due to the fact that the period length is short (1 hour) compared to the workload of the job at Station i (2 hours). As a result, the model generates a workload from the job for Station j even though the job is still in-process at the upstream station.

Figure 2.3 illustrates the difference between the production levels in both the actual system and the TPM at Station i and Station j . At Station i in the actual system, the job completes its processing at the end of period $t + 1$ but in the TPM, the station continues to process the job after period $t + 1$. At Station j , the production level in the actual system consists of a “spike” of 1 hour in period $t + 2$. In the TPM, the production is “smoothed”, with production levels at 0.5 hour and 0.25 hour in period $t + 1$ and $t + 2$ respectively.

This example considers a system of only two stations. To model a complex job shop using the TPM, one must set the period length by considering the job movements between all stations. On the one hand, the period length should be such that it is unlikely that a job completes processing at more than one station in a time period. In a shop with many stations and where jobs move quickly between stations, this implies that the period length be set on the order of the average job workload.

On the other hand, the accuracy of the TPM depends on the assumption of a continuous workflow. We prefer to set a long time period relative to the workload of the jobs so that the discrete jobs will be “more fluid.” The restriction of period sizing may affect the use of the TPM for production planning. In some production systems, it takes only a short time for a job to travel through more than one station, and thus the discrete time period has to be short. However, the time buckets in most planning systems are defined in much longer time units. Thus the ability to set a longer time period will facilitate the application of the TPM to production planning.

Furthermore, the TPM assumes a Markovian workflow such that transitions do not depend on the history of the system. In essence, the model assumes that each station

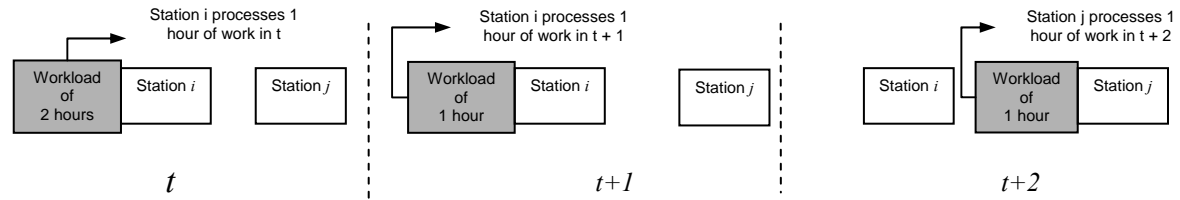


Figure 2.1 Actual system from start of period t to $t+2$

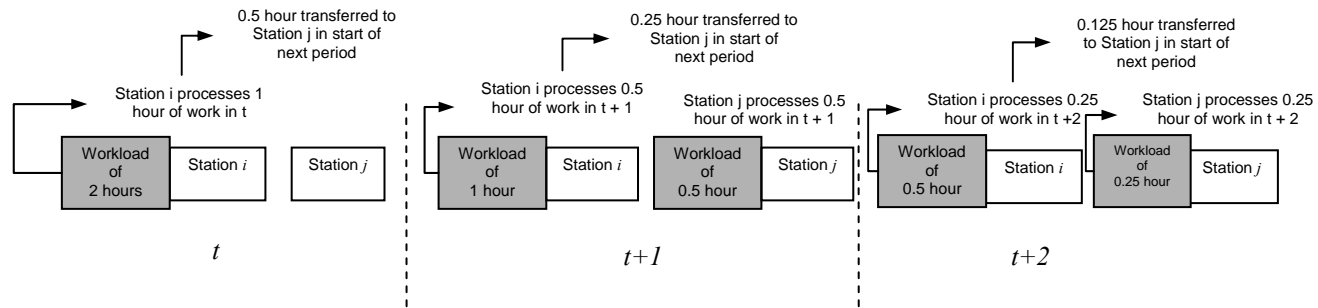
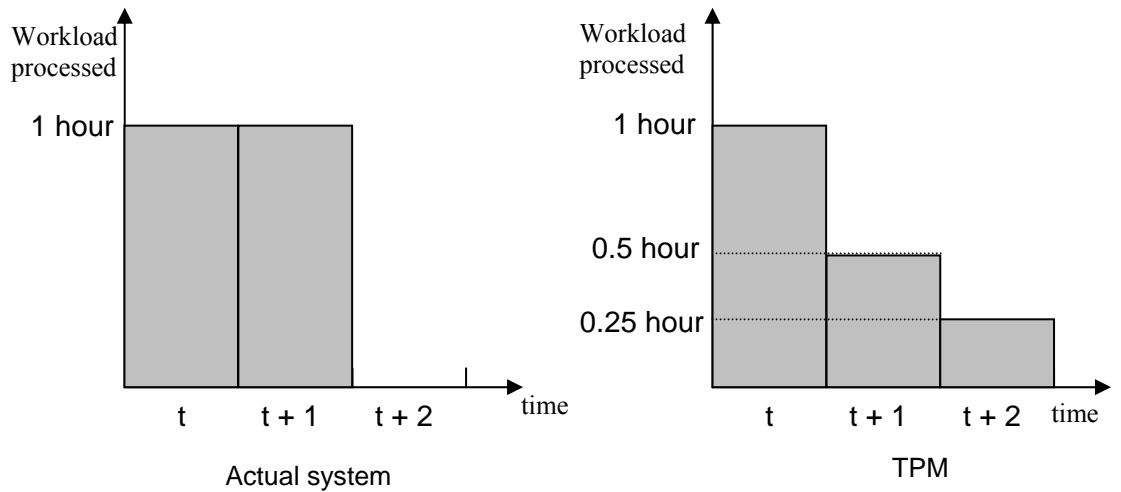
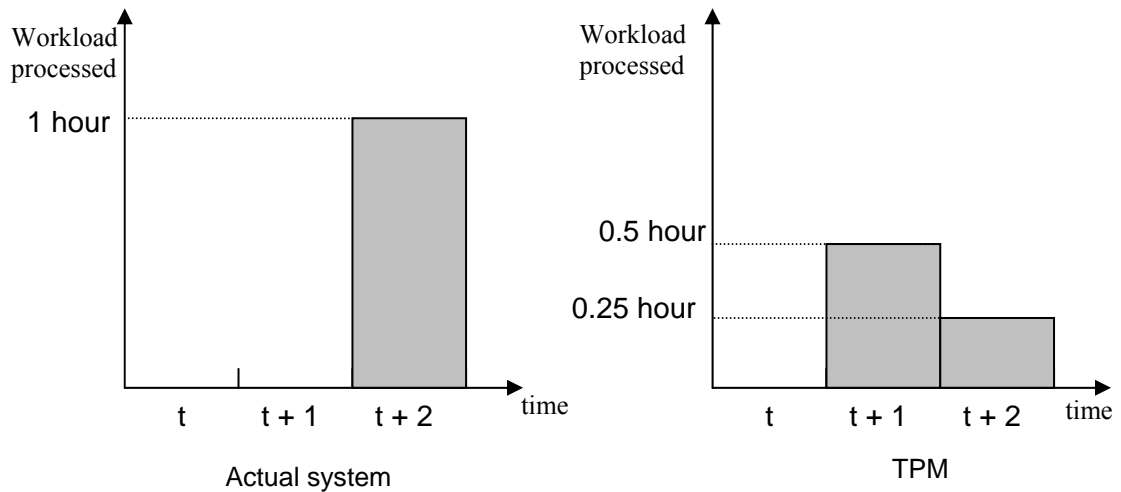


Figure 2.2 TPM from start of period t to $t+2$

processes a relatively stable mix of jobs in each time period, so that subsequent flow to downstream stations is stable as well. The validity of this assumption depends on the length of the time period. If only a few jobs are completed in each period, then it is unlikely that there is a very stable output. Therefore, this assumption will be more valid if we are able to set a longer time period.



Production levels at Station i from period t to $t+2$



Production levels at Station j from period t to $t+2$

Figure 2.3 Production levels at Station i and Station j in actual system and TPM

In addition, we can approximate the production requirement to be normally distributed if the time period is sufficiently long. This is because production requirements of successive periods will be relatively uncorrelated if we set an adequately long time period. As such, the Central Limit Theorem implies that the requirement approaches a normal distribution. If the normality assumption is valid and since the TPM computes the first two moments of the production requirements, we are able to characterize the distribution of the production requirements. Knowledge of the requirement distribution is useful for determining several performance measures, e.g. probability that production requirement will exceed capacity, and furthermore, it also enables us to incorporate our model into optimization procedures to determine the optimal planned lead times in *Chapter 5*.

2.5 MODEL FORMULATION

In this section, we develop a single-station model to overcome the limitations of the TPM discussed in the last section. We will derive a linear control rule that accommodates more frequent arrivals to the work station; as a consequence, we can now permit work to flow through more than one workstation within a time period. We then find the first two moments of the production and queue length variables using the derived control rule.

Without loss of generality, we suppose that each discrete time period t has a length of one time unit. We sub-divide each time period t into p equal subintervals of sub-period s , where $s = 1, 2, \dots, p$. We define $p = 1/\Delta$, where Δ is the length of each sub-period.

We assume that workflow can arrive at the start of each sub-period. We also assume that we set the production in each sub-period according to the linear control rule (2.1). Thus, we control the production according to a finer time grid, and we allow for more fluid arrivals to the workstation.

We restate the control rule (2.1) for each sub-period s as

$$Y(\Delta, s) = \alpha\Delta X(\Delta, s) \quad \text{for } s = 1, 2, \dots, p \quad (2.22)$$

where $Y(\Delta, s)$ is the production level in sub-period s of length Δ , $X(\Delta, s)$ is the queue length at start of sub-period s of length Δ and α is the smoothing parameter. We interpret $1/\alpha$ as the planned lead time; however, we now permit α to assume any positive value, and thus, we permit the planned lead time to be less than one time period. Equation (2.22) is analogous to (2.1) such that the production $Y(\Delta, s)$ in each sub-period s is a fixed fraction $\alpha\Delta$ of the queue length $X(\Delta, s)$ at the start of each sub-period.

Now we proceed to develop the linear control rule for P_t in terms of Q_t and A_t . These variables have the same definition as in TPM: P_t is the production completed in period t , Q_t is the queue length at the start of period t , and A_t is the arrival of work to the station in period t . However, we now assume that A_t does not arrive at the start of the period, but rather arrives uniformly over period t . In particular, we assume that in each sub-period, the arrival amount is equal to A_t/p .

We have the following boundary condition for the queue length for the first sub-period:

$$X(\Delta, s = 1) = Q_t + A_t/p \quad (2.23)$$

For $s > 1$, we model the queue length in the sub-period s by the standard inventory equation

$$X(\Delta, s) = X(\Delta, s - 1) - Y(\Delta, s - 1) + A_t/p \quad (2.24)$$

By substituting (2.22) into (2.24), we obtain

$$X(\Delta, s) = (1 - \alpha\Delta) X(\Delta, s - 1) + A_t/p \quad (2.25)$$

To get an expression for P_t , we note that

$$P_t = \sum_{s=1}^p Y(\Delta, s) = \alpha\Delta \sum_{s=1}^p X(\Delta, s) \quad (2.26)$$

We sum the above expressions for $X(\Delta, s)$ to find

$$\sum_{s=1}^p X(\Delta, s) = Q_t + (1 - \alpha\Delta) \sum_{s=1}^{p-1} X(\Delta, s) + A_t \quad (2.27)$$

From (2.27), we observe that

$$\begin{aligned} \sum_{s=1}^p X(\Delta, s) &= Q_t + (1 - \alpha\Delta) \sum_{s=1}^p X(\Delta, s) + A_t - (1 - \alpha\Delta)X(\Delta, p) \\ \Rightarrow \alpha\Delta \sum_{s=1}^p X(\Delta, s) &= Q_t + A_t - (1 - \alpha\Delta)X(\Delta, p) \end{aligned} \quad (2.28)$$

We now combine (2.26) and (2.28) to get

$$P_t = Q_t + A_t - (1 - \alpha\Delta)X(\Delta, p) \quad (2.29)$$

From (2.29), in order to get an expression for P_t , we need to find $X(\Delta, p)$. From (2.25) and repeated substitution, we can then write

$$\begin{aligned} X(\Delta, p) &= (1 - \alpha\Delta)^{p-1} \times Q_t \\ &\quad + \left(1 + (1 - \alpha\Delta) + \dots + (1 - \alpha\Delta)^{p-1}\right) \times \frac{A_t}{p} \end{aligned} \quad (2.30)$$

We can use (2.30) to re-write (2.29) as

$$\begin{aligned} P_t &= \left(1 - (1 - \alpha\Delta)^p\right) \times Q_t \\ &\quad + \left(1 - \left(\frac{1 - \alpha\Delta}{\alpha}\right) \cdot \left(1 - (1 - \alpha\Delta)^p\right)\right) \times A_t \end{aligned} \quad (2.31)$$

Thus we can write (2.31) as

$$P_t = \beta(\Delta)Q_t + \gamma(\Delta)A_t \quad (2.32)$$

where

$$\beta(\Delta) = 1 - (1 - \alpha\Delta)^p$$

and

$$\begin{aligned} \gamma(\Delta) &= 1 - \left(\frac{1 - \alpha\Delta}{\alpha} \right) \left(1 - (1 - \alpha\Delta)^p \right) \\ &= 1 - \beta(\Delta) \left(\frac{1 - \alpha\Delta}{\alpha} \right) \end{aligned}$$

Now we proceed to determine the continuous-time limits for $\beta(\Delta)$ and $\gamma(\Delta)$ as the length of the sub-period goes to zero. This corresponds to a continuous-time control in which the production level will satisfy (2.22) at every instant in time. We use the

formula $\lim_{x \rightarrow 0} (1 - x)^{\frac{1}{x}} = e^{-1}$ to obtain the continuous-time limit of $\beta(\Delta)$:

$$\begin{aligned} \beta &= \lim_{\Delta \rightarrow 0} \beta(\Delta) = \lim_{\Delta \rightarrow 0} [1 - (1 - \alpha\Delta)^p] \\ &= \lim_{\Delta \rightarrow 0} [1 - (1 - \alpha\Delta)^{\frac{1}{\Delta}}] \\ &= 1 - e^{-\alpha} \end{aligned}$$

For $\gamma(\Delta)$, we find that

$$\begin{aligned} \gamma &= \lim_{\Delta \rightarrow 0} \gamma(\Delta) = \lim_{\Delta \rightarrow 0} \left[1 - (1 - (1 - \alpha\Delta)^p) \left(\frac{1 - \alpha\Delta}{\alpha} \right) \right] \\ &= \lim_{\Delta \rightarrow 0} \left\{ 1 - \left(\frac{1 - \alpha\Delta}{\alpha} \right) \left(1 - (1 - \alpha\Delta)^p \right) \right\} \\ &= 1 - \lim_{\Delta \rightarrow 0} \left(\frac{1 - \alpha\Delta}{\alpha} \right) + \lim_{\Delta \rightarrow 0} \frac{(1 - \alpha\Delta)(1 - \alpha\Delta)^p}{\alpha} = 1 - \frac{1}{\alpha} + \frac{1}{\alpha} e^{-\alpha} \\ &= 1 - \frac{1}{\alpha} (1 - e^{-\alpha}) = 1 - \frac{\beta}{\alpha} \end{aligned}$$

We can now restate (2.32) for the continuous-time control as:

$$P_t = \beta Q_t + \gamma A_t \quad (2.33)$$

where β and γ are given above.

The balance equation for the queue length for the single station is now given by:

$$Q_t = Q_{t-1} - P_{t-1} + A_{t-1}. \quad (2.34)$$

This balance equation differs from (2.2) in the TPM, due to the new assumption that arrivals occur continuously throughout a period. Hence, we define Q_t to be the queue length at the start of period t , prior to any arrivals in period t .

By substituting (2.33) into (2.34) and repeated substitution, we obtain:

$$Q_t = (1-\gamma) \sum_{i=1}^{\infty} (1-\beta)^{i-1} A_{t-i}. \quad (2.35)$$

If we assume that the arrivals are i.i.d. with mean μ and variance σ^2 , then we find the two moments for the queue length from (2.35):

$$E[Q_t] = \left(\frac{1-\gamma}{\beta} \right) \mu = \frac{\mu}{\alpha} \quad (2.36)$$

$$Var(Q_t) = \frac{(1-\gamma)^2 \sigma^2}{2\beta - \beta^2} \quad (2.37)$$

Similarly, we obtain the two moments for the production variable:

$$E[P_t] = \mu \quad (2.38)$$

$$\text{Var}(P_t) = \left(\frac{\beta}{2-\beta} (1-\gamma)^2 + \gamma^2 \right) \sigma^2 \quad (2.39)$$

Thus we have obtained the analytical expressions for the moments of these two variables. We note that the first moments for the production requirements and queue lengths are the same as the TPM, but the second moments differ.

2.6 DISCUSSION

We have expressed (2.33) in terms of the parameters β and γ , both of which are functions of α . These parameters represent the fraction of the starting queue and of the current period arrivals, respectively, that gets completed in each period. For small values of α , we see that β approaches α and γ approaches zero; hence, for small values of α , the continuous-time production rule in (2.33) becomes the TPM linear rule in (2.1). However, it is less clear how the rule behaves for large values of α . To get some insights into this, we graph β , γ , and $\text{Var}(P_t)$ as functions of the smoothing parameter α in Figure 2.4 (with $\sigma = 1$). We observe that the graph $\text{Var}(P_t)$ decreases (production requirement becomes less variable) as we perform more smoothing, i.e. smaller values of α . We illustrate the tradeoff between production smoothness and inventory in Figure 2.5, in which we graph both $\text{Var}(P_t)$ and $E[Q_t]$ (with $\mu = 1$, $\sigma = 1$). We observe that as we carry out more production smoothing (decrease in $\text{Var}(P_t)$), the expected queue length $E[Q_t]$ grows.

The variance of production requirement and the expected queue length as functions of the planned lead time provide a simple way to see the fundamental tradeoffs across the three elements of lead time, capacity and inventory. For given arrival variability, as we reduce the planned lead time, production becomes more variable and more capacity is required but it results in less WIP inventory; alternatively, as we smooth production by an increase in the planned lead time, we need less capacity but it leads to more WIP inventory. These insights are the same as for the TPM. But the model given here is more general in that we permit continuous arrivals to the work station and continuous-time production control.

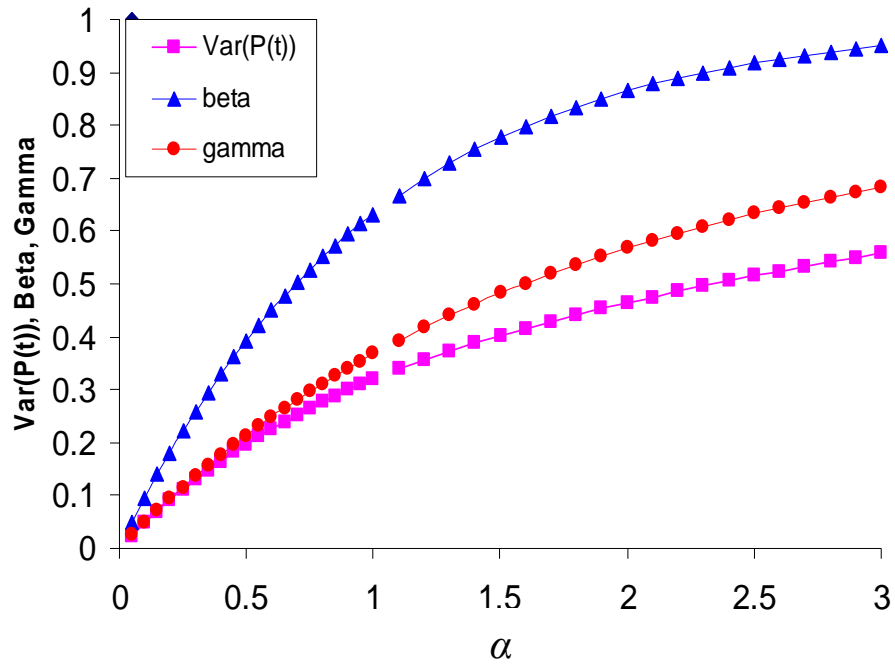


Figure 2.4 Graph of β , γ , and $\text{Var}(P_t)$ as functions of α (with $\sigma = 1$)

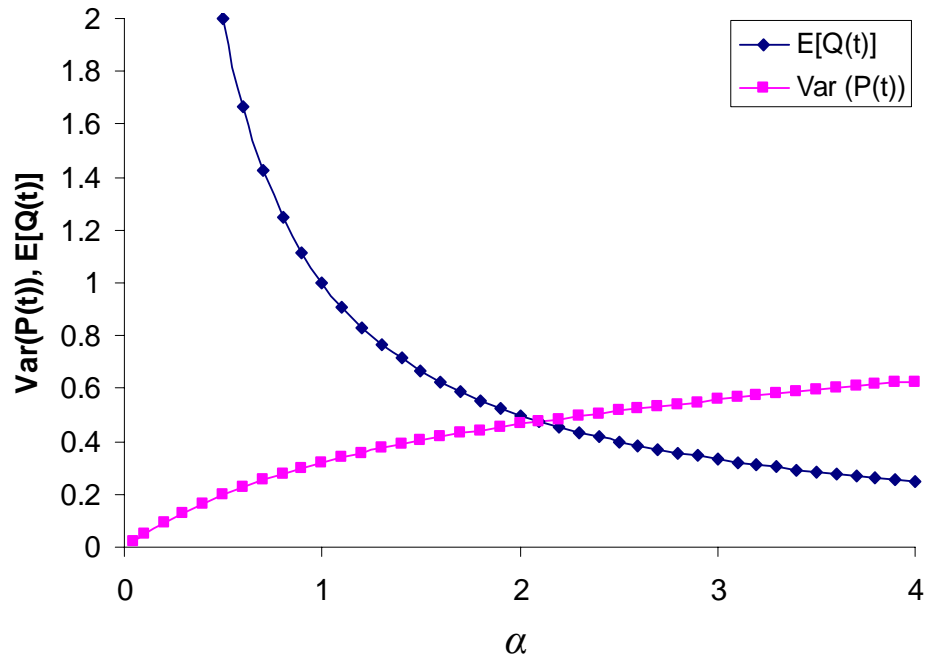


Figure 2.5 Graph of $\text{Var}(P_t)$ and $E[Q_t]$ as functions of α (with $\mu = 1$, $\sigma = 1$).

Now we compare the second moment of production requirement of our model with that of the TPM. Since $0 < \alpha \leq 1$ for the TPM, we only perform the comparison for this range of α . We determine the ratio of the production variance of the continuous-time model to that of the TPM, which we denote as $r(P)$. We divide (2.39) by (2.6) to obtain

$$\begin{aligned}
 r(P) &= \frac{\left\{ \frac{\beta}{2-\beta} \left((1-\gamma)^2 + \gamma^2 \right) \sigma^2 \right\}}{\left\{ \frac{\alpha}{2-\alpha} \sigma^2 \right\}} \\
 &= \frac{\beta}{2-\beta} \frac{2-\alpha}{\alpha} \left(\left(1 - \left(1 - \frac{\beta}{\alpha} \right) \right)^2 + \left(1 - \frac{\beta}{\alpha} \right)^2 \right) \\
 &= \left(\frac{\beta}{\alpha} \right) \left(\frac{2-\alpha}{2-\beta} \right) \left\{ 1 - 2 \left(\frac{\beta}{\alpha} \right) \left(1 - \frac{\beta}{\alpha} \right) \right\}
 \end{aligned}
 \tag{2.40}$$

We note that $\alpha > \beta$ (i.e. $\beta/\alpha < 1$)* and therefore we see from (2.40) that $r(P) < 1$, i.e. the production variance of the continuous-time model is smaller than that of the TPM. We graph $r(P)$ as well as $\text{Var}(P_t)$ of the two models in Figure 2.6 (with $\sigma = 1$).

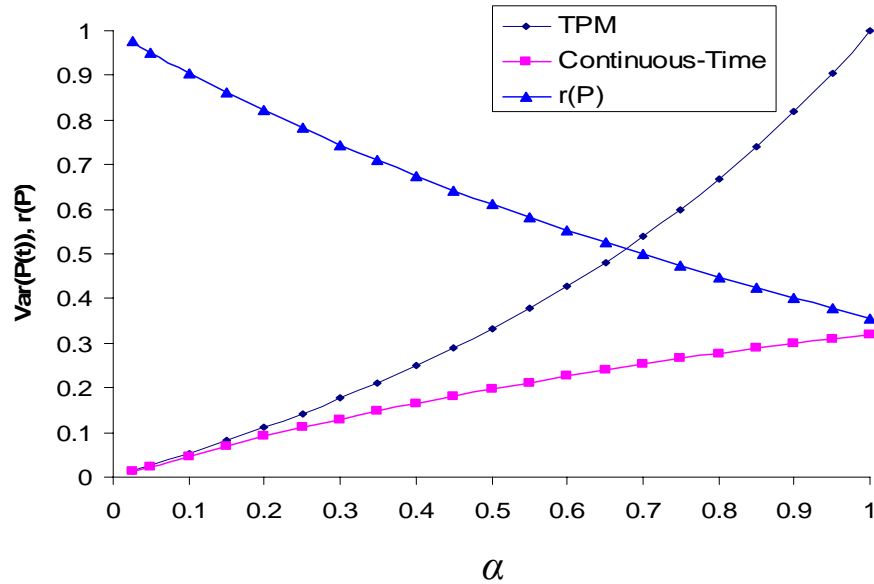


Figure 2.6 Comparison of $\text{Var}(P_t)$ between TPM and continuous-time model

*Using the formula for series $e^x = \sum_{v=0}^{\infty} \frac{x^v}{v!}$, we find that $1 - e^{-\alpha} = \alpha - \left(\frac{\alpha^2}{2!} - \frac{\alpha^3}{3!} + \frac{\alpha^4}{4!} - \frac{\alpha^5}{5!} + \dots \right)$.

Since the bracketed term is positive, we have $1 - e^{-\alpha} < \alpha \Rightarrow \beta < \alpha$

Why is $\text{Var}(P_t)$ of the continuous-time model smaller than that of the TPM? In the continuous-time model, we assume that work arrives uniformly within each period (although work arrivals vary between periods). In addition, since the instantaneous production requirement is a fixed proportion of the queue length (see (2.22)), the production requirement changes gradually and uniformly within the period in response to a change in the amount of work arrivals in that period. This is unlike the TPM in which the control rule defines that the production requirement in a period is a fixed proportion of the queue length *after* all work arrives at the start of each period. Thus the TPM is more sensitive to changes in work arrivals than the continuous-time model. This difference in variability of production requirement is especially greater when α is larger, which is illustrated in Figure 2.6; as α approaches zero, $r(P)$ approaches one but as we increase α , $r(P)$ gets smaller ($r(P) = 0.355$ at $\alpha = 1$). This is because the production requirement of the TPM is more sensitive to the variability of work arrival at larger values of α (shorter planned lead times). For example, if $\alpha = 1$ (i.e. planned lead time equal to one period) and according to the linear control rule of the TPM in (2.1), we have $P_t = Q_t$, i.e. the workstation has to produce all the work-in-queue after all work arrives at the start of each period. But for the continuous-time model with the same value of α , we have $P_t = 0.63Q_t + 0.37A_t$, which implies that the production level is set to a fraction of the work-in-queue and work arrivals. However, the actual expected lead time remains equal to the planned lead time of one period.

The above discussion is on the analysis of a single-station. In multi-station systems, as illustrated in the two-station example in section 2.4, the variability of workload at a station in the TPM is smoother than the actual workload variability due to the period sizing restriction. But for the continuous-time model, the workload variability approaches that of the actual workload as we set a longer period, which increases the fluidity of the workflow.

Now we compare the second moments of queue lengths between the two models for $0 < \alpha \leq 1$. Similarly, we determine the ratio of the queue variances of the continuous-time model to that of the TPM, which we denote as $r(Q)$. We divide (2.37) by (2.9) and find that

$$\begin{aligned}
r(Q) &= \frac{\left\{ \frac{(1-\gamma)^2}{2\beta - \beta^2} \sigma^2 \right\}}{\left\{ \frac{1}{2\alpha - \alpha^2} \sigma^2 \right\}} \\
&= \frac{2\alpha - \alpha^2}{2\beta - \beta^2} \left\{ 1 - \left(1 - \frac{\beta}{\alpha} \right) \right\}^2 \\
&= \frac{2\alpha - \alpha^2}{2\beta - \beta^2} \left(\frac{\beta}{\alpha} \right)^2 \\
&= \left(\frac{\beta}{\alpha} \right) \left(\frac{2 - \alpha}{2 - \beta} \right)
\end{aligned}
\tag{2.41}$$

We observe from (2.41) that since $\alpha > \beta$, we have $r(Q) < 1$. This implies that the continuous-time model has a smaller queue variance compared to the TPM. We graph $r(Q)$ and $\text{Var}(Q_t)$ of both models in Figure 2.7.

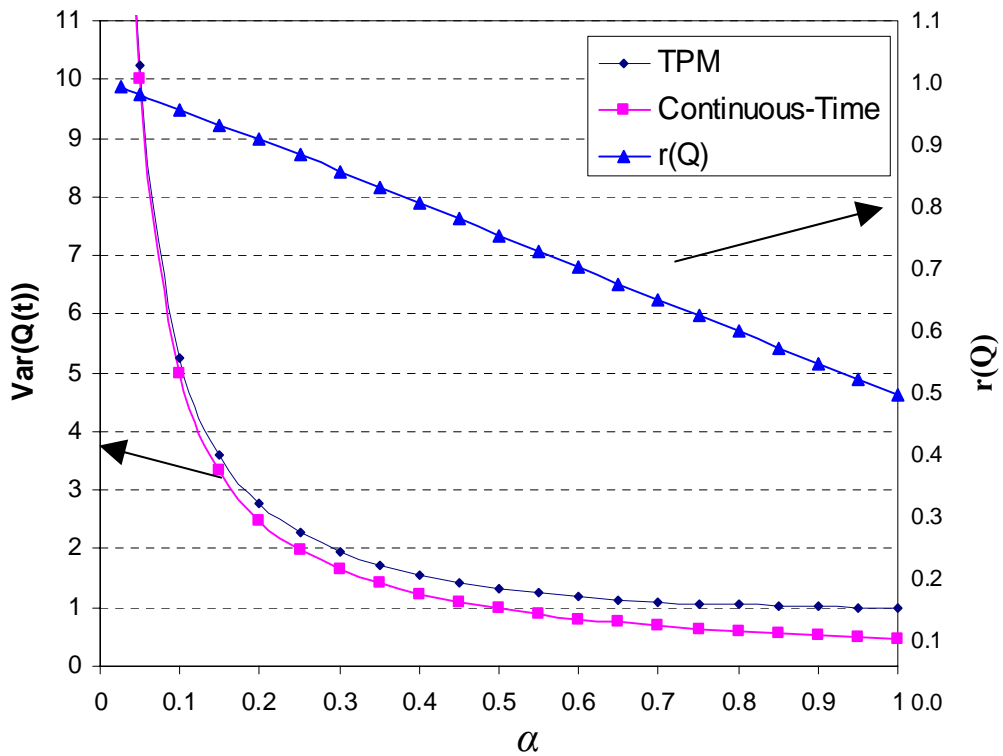


Figure 2.7 Comparison of $\text{Var}(Q_t)$ between TPM and continuous-time model

The reason for TPM having a smaller $\text{Var}(Q_t)$ is that in the TPM, Q_t is defined as the queue length at the start of each period after all work arrivals (see (2.2)). Although, in the continuous-time model, Q_t is also the queue length at the start of each period, but unlike the TPM, work arrives uniformly throughout the period and not only at the start of each period (see (2.34)). Therefore Q_t is more sensitive to the variability of work arrival compared to the continuous-time model.

Although the TPM has its limitations, the model is a reasonable representation of systems in which jobs move to the next processing station only at fixed intervals, e.g. at the start of work shifts. However, in using the continuous-time model to depict the workflow in such systems, we need to define the period interval suitably due to model's uniform work arrival assumption. In general, for systems that operate on large batch sizes which may entail very infrequent work arrivals within each period, we need to define a longer time period relative to the arrival interval. The purpose here is to increase the workflow fluidity that will improve the validity of the uniform flow assumption. But in attempting to do so, the resulting consideration is that we must also ensure that the length of the period remains consistent with the planning time bucket. A useful extension to the continuous-time model is to relax the uniform workflow assumption. One possible way to overcome this assumption is to introduce an additional random variable, e.g. a zero-mean and finite variance variable in (2.23) and (2.24), to model the varying work arrivals. However, one would imagine that such an extension would significantly increase the complexity of the model.

In this chapter, we have derived the continuous-time control rule that enables us to define a time period that corresponds to the time buckets in production planning systems. In the next chapter, we extend the single-station model of this chapter to a multi-station model for the tactical planning of MTO production. In the multi-station model, the fundamental assumptions for each individual station are the same as that of the single-station model presented in this chapter.

CHAPTER 3

A MULTISTAGE TACTICAL PLANNING MODEL FOR MAKE-TO-ORDER ENVIRONMENT WITH MULTIPLE PRODUCTS

3.1 INTRODUCTION

In this chapter, we develop the tactical planning model for the purpose of production smoothing in a MTO environment with firm delivery lead time. We model the smoothing of the MPS in which we translate the fluctuating demand into a smoother MPS. The smoother MPS is in turn converted into a job release schedule for the production system which we model as a job shop. We characterize the production smoothing at the workstations where we employ the continuous-time control rule derived in *Chapter 2*. Subsequently we model the workflow in the job shop and obtain the first two moments of production requirements as well as the expected queue lengths, which are the performance measures to assess the proper settings of the tactical planning parameters.

In the next section, we define the problem setting for our research and discuss the impact of the planning window and planned lead times on production smoothing. We present our single-family model in section 3.3, where we assume the manufacturer produces only one aggregate product family. Subsequently, we discuss how we extend the model for a system with multiple product families in section 3.4. We then embed our model in optimization procedures in section 3.5. We give a summary of the chapter in section 3.6.

3.2 PROBLEM SETTING

We consider a job shop that produces a wide variety of products. We assume that we can group the end items into product families; each product family consists of products that are similar in design and share the same or similar production process.

We define the unit used for stating the MPS at the product family level, and the production of each product family is managed according to its own MPS (see Vollmann et al. (2005) for a discussion of stating the MPS unit as groups of items instead of end items). In addition, each product family has a firm delivery lead time (*DLT*); that is, the manufacturer commits to meet demand for a product family by its *DLT*. For instance, for a product family with a firm delivery lead time of one week, the manufacturer would commit to deliver each order exactly one week upon the receipt of the order.

We assign to each workstation in the shop a fixed planned lead time, which we denote as the *station planned lead time (SPLT)*. The *SPLT* is the *intended* number of time periods that each job takes to get processed at the workstation. Each product family has its own predetermined route through the workstations of the shop. We associate with each product family a *product planned lead time (PPLT)*. The *PPLT* is the *planned* production lead time for the product and is equal to the sum of the *SPLTs* of all workstations in the route of the product family, i.e. the time taken for a job from a product family to move through the shop. In the MRP system, the *PPLT* is used to determine the release time of a job into the shop from the MPS; the release time is equal to the due date minus the *PPLT*.

In the rest of this section, we look at how production smoothing is performed in the MTO setting described above. First we look at how the MPS is smoothed before jobs are released into the shop and subsequently, we look at how production is smoothed at each workstation by the *SPLT*.

3.2.1 Smoothing of Master Production Schedule

We discuss how to convert the fluctuating demand for each product family into a smoother MPS and in turn a less variable job release. A smoother release is desirable as it results in a less fluctuating production requirement for the capacity. For each product family in an MTO operation, smoothing is possible only if the *DLT* is greater than the *PPLT*. Cruickshanks et al. (1984) term this time difference as the *planning window*. We can smooth the MPS over this planning window, where the extent of the smoothing depends on the length of the planning window.

In any period t , the production planner sets the MPS (production level for completion) for period $t + PPLT$. This production level is also the job release into the shop in period t since the time of release is the due date offset by the $PPLT$. Thus a smoother MPS implies a less variable release into the shop. The planner sets this production level based on the knowledge of all contracted orders for delivery over the planning window of length $DLT - PPLT + 1$, which ranges from period $t + PPLT$ to period $t + DLT$. We illustrate the smoothing of the MPS with the following example.

Example: We consider a product family with $DLT = 4$, $PPLT = 2$ and planning window of length $= 4 - 2 + 1 = 3$. The current period is $t = 0$. The production planner has to decide on the MPS and the planned releases over the planning horizon that ranges from $t = 0$ to $t = 4$. To simplify the example, we assume that no orders will be received after period $t = 4$. Furthermore, we assume that there is no planned inventory at $t = 1$. We illustrate the smoothing of the MPS in Table 1.

Table 3.1 Smoothing of the MPS

Period t	0	1	2	3	4
Orders	-	-	10	20	30
MPS	-	-	20	20	20
Planned Inventory	-	0	10	10	0
Planned Release	20	20	20	-	-

The second row shows orders contracted to be delivered in each period, from $t = 2$ to 4. We do not state the order quantities for period $t = 0$ and 1, as they are already released into the shop and are thus inconsequential. In this example, the MPS is smoothed over the planning window, which ranges from $t = 2$ to 4. Smoothing is done by early production of some orders for period $t = 4$. Here the production output is 20 units in $t = 2$ instead of the order quantity of 10 units. As a result, the MPS is leveled at 20 units in each period of the planning window. However, due to the early

production, we need to hold a planned inventory of 10 units in period $t = 2$ which is carried over to period $t = 3$. The last row shows the planned releases into the shop at the end of each period. Since the $PPLT = 2$, we release jobs two periods beforehand to meet the corresponding MPS. For instance, we release 20 units at the end of $t = 0$ to meet the MPS in period $t = 2$.

3.2.2 Smoothing at the Workstations

Now we consider production smoothing at the workstations. Here the production requirements vary due to both the varying job arrivals and the inherent workload variability at each workstation as a result of e.g. the different processing requirements of jobs, machine breakdowns, setups and yields. The production smoothness of a workstation is determined by the length of the $SPLT$. A longer $SPLT$ allows us to achieve a relatively smoother production as we can level out the production output to a greater extent by early production of some orders. However, with a longer $SPLT$, jobs stay longer at the workstation and this leads to a higher work-in-process (WIP) inventory. Thus there is a tradeoff between production smoothness and inventory in the setting of the $SPLT$.

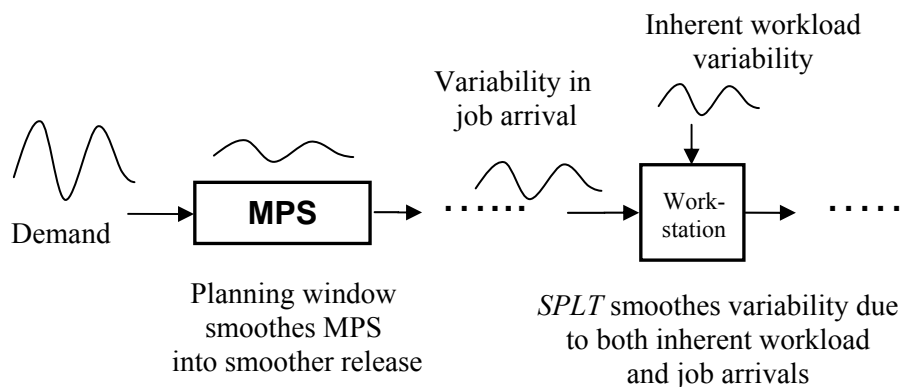


Figure 3.1 Production smoothing by planning window and SPLT

Figure 3.1 summarizes how we smooth production by the planning window and the $SPLT$. In summary, the planning window smooths the MPS while the $SPLT$ smooths the variability due to both job arrivals and inherent workload at the workstation. Thus a system that is subject to a highly variable demand but low

inherent workload variability at the workstations would prefer a longer planning window and shorter *SPLTs*. This is because the longer planning window results in a smoother release which leads to a less variable job arrival to the workstations; in addition, since the inherent workload is not highly variable, shorter *SPLTs* are sufficient to smooth production at the workstations. Furthermore, having shorter *SPLTs* has the added advantage of lower WIP inventory level. However if the system has high inherent workload variability, then long *SPLTs* are required to smooth production at the workstations.

3.2.3 Problems in Production Planning

The production planner in this MTO environment faces the following problems:

- 1) For each product family, how to assign the *SPLT* to each workstation and equivalently, how to set the *PPLT*? In setting the *SPLT*, the planner must take into account the tradeoff between capacity requirements (which depends on production smoothness) and WIP inventory level at each workstation.
- 2) The length of the *PPLT* of a product family is constrained by its *DLT* ($PPLT \leq DLT$). Since the *PPLT* is the sum of *SPLTs* of all stations along the route of the product family, the *DLT* imposes a constraint on the *SPLTs* of the workstations. Therefore the planner must satisfy this constraint when setting the *SPLTs* of the workstations.
- 3) The planner must consider the tradeoff between the smoothing of the MPS and that at the workstations. This is because a longer *PPLT* permits greater smoothing at the workstations but reduces the planning window, which in turn reduces the smoothing of the MPS. A longer *PPLT* implies more available *SPLTs* to smooth the inherent workload variability present at each workstation; but the ensuing shorter planning window leads to more variable job release and thus more variability in the job arrivals to the workstations.

The above problems are even more complicated if there is a large number of product families. Here each station can process jobs from many product families and the production planner must take into account the interactions between the various product families at each station. Our model helps the planner in assessing the setting

of the *SPLT* of each workstation, and equivalently the *PPLT* as well as the planning window given the *DLT*. More specifically, our model computes the performance measures of production smoothness and inventory, i.e. the variance of production requirement and the expected inventory level, respectively, based on the values of the *SPLTs*. In addition, our model can provide a “what-if” analysis. For example, the planner may wish to improve customer service by reducing the *DLT*. In this case, our model can quantify the resulting impact of a shorter *DLT* on production smoothness.

We note that our model does not explicitly take into account the inventory of finished products that result from the orders that are built ahead of time for the purpose of smoothing production. There is less risk involved in holding these committed inventories, as opposed to *uncommitted* finished goods inventory in make-to-stock environments, where there is a presence of risk due to the uncertainty in customer demand. In addition, building of orders in advance to smooth production enables some orders to be delivered earlier than the promised delivery dates, which would be beneficial to customer relations. Therefore we argue that the committed finished goods inventory level is not a crucial planning consideration as compared to the more critical task of production smoothing, which is the focus of this research.

3.3 Single-Family Model

In this section, we model the tactical planning process for a manufacturing system that produces one aggregate product family. We first model how we convert the stochastic demand into a smoother MPS before releasing into the job shop. Next we model the production smoothing at the workstations and how jobs move within the shop. We then obtain the first two moments of production requirements as well as the mean queue length at each workstation, which are the performance measures to assess the setting of the planning windows and the *SPLTs*.

We denote N and L as the *PPLT* and *DLT* respectively; we restrict these parameters to be positive integers. We define the length of the planning window as

$$W = L - N + 1 \quad (3.1)$$

Equation (3.1) implies that if $L = N$, we would have $W = 1$, which characterizes no smoothing of the MPS and jobs are released directly into the shop upon the receipt of orders. The $PPLT$ is the sum of the $SPLT$ s of all workstations along the processing route; thus, we have

$$N = \sum_i \omega_i n_i \quad (3.2)$$

where ω_i is the number of times each job visits workstation i and n_i is the $SPLT$ of workstation i , measured in time periods (time buckets in MRP planning systems). If there are numerous processing routes (e.g. a job shop), we can simplify the model by defining ω_i as the expected number of visits to workstation i and as such, the parameters W , L and N in (3.1) and (3.2) are the expected planning window, DLT and $PPLT$ respectively. In setting the planning window and $SPLT$ s, one must satisfy the relationship between the planning parameters in (3.1) and (3.2).

3.3.1 Model of MPS Smoothing

We assume that the demand D_t received in period t is i.i.d. (independent and identically distributed) with mean $E[D]$. As discussed in *section 3.2.1*, the production planner smoothes the MPS by considering the orders not yet released into production. We can view these orders that are not yet released as a queue of demand that forms outside the shop, waiting to be released into the shop.

Although the demand arrivals are *i.i.d.* the smoothed releases are generally not. To model the work release, we introduce a dummy station to the network of workstations. In this way, we can incorporate the MPS smoothing and job release into the analysis of the network flow, and as a result, we achieve a much more tractable model. The dummy station is located upstream of the first station that the jobs visit and it smoothes the demand before releasing the jobs into the shop. The order backlog over the planning window that is not yet released is equivalent to the queue in front of the dummy station. In addition, demand D_t is the job arrivals in period t to the dummy

station and the job release is the production output of the dummy station. This is illustrated in Figure 3.2.

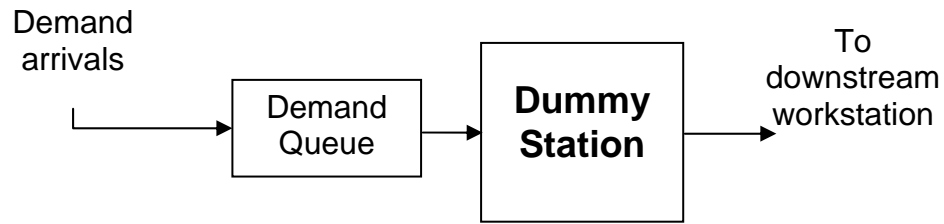


Figure 3.2 Modeling of MPS smoothing by dummy station

We model the production quantities of the dummy station, which we denote as station 0, by

$$P_{0t} = \frac{Q_{0t}}{W} \quad (3.3)$$

where P_{0t} is the production quantities (i.e. quantity of job release into the shop) in period t , Q_{0t} is the queue of work at the start of period t and W is the planning window ($W \geq 1$). In each period, given that the planning window is of length W , the shop must release an average of $1/W$ of the order backlog that is not yet released into the shop. We approximate the smoothing of the MPS by assuming that the shop releases *exactly* $1/W$ of the order backlog in each period, i.e. the dummy station produces $1/W$ of Q_{0t} in each period.

Equation (3.3) is analogous to the approximation in Cruickshanks et al. (1984) to their production smoothing model. They build an approximate model for their proposed smoothing model, which is intractable due to the requirement of no backorders. The approximate model achieves tractability by allowing backorders (our equation (3.3) as well). They demonstrate that the approximate model is indicative of the behavior of their proposed smoothing model. Equation (3.3) is also equivalent to the linear control rule in Graves (1986), in which the production output of each work center is a linear function of the queue length.

The standard balance equation for the dummy station is

$$Q_{0t} = Q_{0,t-1} - P_{0,t-1} + D_{t-1} \quad (3.4)$$

By substituting (3.3) into (3.4) to replace Q_{0t} in (3.4), we obtain

$$P_{0t} = (1 - 1/W)P_{0,t-1} + (1/W)D_{t-1} \quad (3.5)$$

which is a first-order exponential smoothing model with $1/W$ as the smoothing parameter. A larger value of W results in a smoother production at the dummy station, i.e. a smoother job release into the shop.

3.3.2 Model of Smoothing at Workstations

We proceed to look at the production requirement at the workstations. We assume that we set the production requirement according to the linear production rule stated in (2.22). We now explain (2.22) with respect to our current context of production planning for multiple workstations; we assume that each time bucket t has a length of one time unit (e.g. a day or a week) and we sub-divide each time bucket into p equal sub-period s of length Δ , where $s = 1, 2, \dots, p$. We express (2.22) as a function of n_i :

$$Y_i(\Delta, s) = (1/n_i)\Delta X_i(\Delta, s) \quad \text{for } s = 1, 2, \dots, p \quad (3.6)$$

where $Y_i(\Delta, s)$ is the production requirement at workstation i in sub-period s and $X_i(\Delta, s)$ is the queue length at workstation i at the start of sub-period s . Here in order to realize the *SPLT*, station i must process $(1/n_i)\Delta$ of the work in queue on average in each sub-period; this is approximated by (3.6) in which we assume that the production requirement is *precisely* $(1/n_i)\Delta$ of the queue. We note that equation (3.6) implies that the workstation has no capacity constraint. We assume that in times of high workload, the workstation takes expediting actions, e.g. overtime and subcontracting, to produce the jobs within the *SPLT*.

We now restate the continuous-time limits of (3.6) which we derive in *section 2.5* for workstation i as Δ goes to zero:

$$P_{it} = \beta_i Q_{it} + \gamma_i A_{it} \quad (3.7)$$

where we express the constants β_i and γ_i as functions of n_i

$$\beta_i = 1 - e^{-1/n_i} \quad (3.8)$$

$$\gamma_i = 1 - n_i \beta_i \quad (3.9)$$

P_{it} is the production requirement at workstation i in planning period t , Q_{it} is the queue length at the start of period t prior to any arrivals in period t and A_{it} is the arrival of work to workstation i in period t . Note that we assume the work arrives uniformly in each period, which follows the assumption for the continuous-time function in *Chapter 2*. The continuous-time production function in (3.7) enables us to set a time period long enough to correspond to the length of the time buckets, as discussed in *Chapter 2*.

The balance equation for workstation i at the start of each planning period is given by

$$Q_{it} = Q_{i,t-1} - P_{i,t-1} + A_{i,t-1} + \xi_{it} \quad (3.10)$$

The term ξ_{it} is a zero-mean noise term that represents any inherent workload variability in the arrival stream to station i . The noise signifies any deviation between the actual workload and the expected workload, based on the upstream production levels. We model the noise term in (3.10) rather than model it as part of the arrival process $\{A_{it}\}$ like in the TPM of Graves (1986) (see equation (2.12)). The reason for doing so is to reduce the complexity of the eventual expressions. In addition, we assume that the amount of noise in each period is independent of the work arrivals. The motive here is also to make the analysis more tractable.

We use the noise term ξ_{it} to model the various sources of variability e.g. processing times, machine failures, setups and yields. Here we adopt the concept of *effective processing time*. The effective processing time is the adjusted processing time that accounts for time that is unavailable for processing due to the sources of variability. For instance, a job that spends an hour on setup and another hour on processing would have an effective processing time of two hours. We can model the variability by the variance of ξ_{it} which is given by

$$\text{Var}(\xi_{it}) = u_i \text{Var}(T_i) \quad (3.11)$$

where we denote T_i as the effective processing time for a random job at workstation i , $\text{Var}(T_i)$ as its variance (which can be obtained through measurement in the actual system) and u_i is the expected number of job arrivals at workstation i in each period. We will see later in this section how we incorporate $\text{Var}(\xi_{it})$ into the computation of the variance of production requirement.

Now we proceed to model work arrivals to the workstation i , which is given by

$$A_{it} = \sum_j \phi_{ij} P_{jt} \quad (3.12)$$

where ϕ_{ij} is a positive scalar. We assume that every time unit of production at workstation j generates ϕ_{ij} time units of input to station i on average. When station j represents the dummy station, then we assume that every unit of order quantity triggers ϕ_{ij} time units of work, on average, for station i .

3.3.3 Workflow Model

Now in order to analyze the job shop model, we combine the production functions in (3.3) and (3.7), and express it in matrix notation,

$$P_t = \mathbf{G}Q_t + \mathbf{F}A_t \quad (3.13)$$

where $\mathbf{P}_t = \{P_{0t}, P_{1t}, \dots, P_{mt}\}'$, $\mathbf{Q}_t = \{Q_{0t}, Q_{1t}, \dots, Q_{mt}\}'$, $\mathbf{A}_t = \{A_{0t}, A_{1t}, \dots, A_{mt}\}'$ are column vectors and m is the number of workstations. Furthermore, \mathbf{G} is a diagonal matrix with its first diagonal element as $1/W$ and the remaining diagonal elements as β_i ($i = 1, 2, \dots, m$). \mathbf{F} is a diagonal matrix with zero as the top element and γ_i ($i = 1, 2, \dots, m$) as the next m diagonal elements.

We express the work arrivals to the workstations in (3.12) in matrix form:

$$\mathbf{A}_t = \mathbf{\Phi} \mathbf{P}_t \quad (3.14)$$

$\mathbf{\Phi}$ is a square matrix with elements ϕ_{ij} . We note that $\phi_{0j} = 0$ for all j and that ϕ_{j0} is the amount of work that starts at workstation j for each new job. By substituting (3.14) into (3.13), we obtain

$$\mathbf{Q}_t = \mathbf{G}^{-1} (\mathbf{I} - \mathbf{F} \mathbf{\Phi}) \mathbf{P}_t \quad (3.15)$$

where \mathbf{I} is an identity matrix. Note that \mathbf{G}^{-1} exists since it is a diagonal matrix, with each diagonal element being positive. We combine the inventory balance equations for the dummy stations in (3.4) and for the workstations in (3.10), and express in matrix form:

$$\mathbf{Q}_t = \mathbf{Q}_{t-1} - \mathbf{P}_{t-1} + \mathbf{A}_{t-1} + \boldsymbol{\xi}_t + \mathbf{D}_{t-1} \quad (3.16)$$

The first four terms in (3.16) models the workflow to the workstations. The vector $\boldsymbol{\xi}_t$ is a column vector with zero in the top row (as there is no noise arrival at the dummy stations) and elements $\{\xi_{1t}, \dots, \xi_{mt}\}$ in the next m rows. The last term in (3.16), \mathbf{D}_{t-1} , is the column vector that represents the demand arrival with D_{t-1} in the first row as the only non-zero element in the vector. Now by substituting (3.14) and (3.15) into (3.16), we have

$$\begin{aligned}
\mathbf{G}^{-1}(\mathbf{I} - \mathbf{F}\Phi)\mathbf{P}_t &= \mathbf{G}^{-1}(\mathbf{I} - \mathbf{F}\Phi)\mathbf{P}_{t-1} - \mathbf{P}_{t-1} + \Phi\mathbf{P}_{t-1} + \xi_t + \mathbf{D}_{t-1} \\
\Rightarrow \mathbf{P}_t &= \mathbf{P}_{t-1} - (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}\mathbf{P}_{t-1} + (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}\Phi\mathbf{P}_{t-1} \\
&\quad + (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}\xi_t + (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}\mathbf{D}_{t-1} \\
\Rightarrow \mathbf{P}_t &= (\mathbf{I} - \mathbf{F}\Phi)^{-1}(\mathbf{I} - \mathbf{G} + \mathbf{G}\Phi - \mathbf{F}\Phi)\mathbf{P}_{t-1} \\
&\quad + (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}(\mathbf{D}_{t-1} + \xi_t)
\end{aligned} \tag{3.17}$$

Now we let $\mathbf{H} = (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}$ and $\mathbf{B} = \mathbf{I} - \mathbf{H}(\mathbf{I} - \Phi)$, we obtain

$$\mathbf{P}_t = \mathbf{B}\mathbf{P}_{t-1} + \mathbf{H}(\mathbf{D}_{t-1} + \xi_t) \tag{3.18}$$

By repeatedly iterating \mathbf{P}_t and assuming an infinite history of the system, we get

$$\mathbf{P}_t = \sum_{s=0}^{\infty} \mathbf{B}^s \mathbf{H}(\mathbf{D}_{t-1-s} + \xi_{t-s}) \tag{3.19}$$

We proceed to determine the expectation of the production vector. The demand vector \mathbf{D}_t has a mean vector $\boldsymbol{\mu}$ which is a column vector with $E[D_i]$ in the first row as the only non-zero element. From the definition of the noise term ξ_{it} , the vector ξ_t has a zero mean.

From (3.19) we obtain

$$\begin{aligned}
\mathbf{E}[\mathbf{P}] &= (\mathbf{I} - \mathbf{B})^{-1} \mathbf{H}\boldsymbol{\mu} \\
&= (\mathbf{I} - (\mathbf{I} - \mathbf{H}(\mathbf{I} - \Phi)))^{-1} \mathbf{H}\boldsymbol{\mu} \\
&= (\mathbf{H}(\mathbf{I} - \Phi))^{-1} \mathbf{H}\boldsymbol{\mu} \\
&= (\mathbf{I} - \Phi)^{-1} \mathbf{H}^{-1} \mathbf{H}\boldsymbol{\mu} \\
&= (\mathbf{I} - \Phi)^{-1} \boldsymbol{\mu}
\end{aligned} \tag{3.20}$$

We note that the mean of the production vector in (3.20) does not depend on the planning parameters but depends only on the workflow matrix Φ and the mean

demand vector $\boldsymbol{\mu}$. Equation (3.20) is analogous to the equation for computing the arrival rate vector in queuing networks (e.g. Jackson network), where our workflow matrix $\boldsymbol{\Phi}$ corresponds to the probabilistic routing matrix and the mean demand vector $\boldsymbol{\mu}$ relates to the external arrival vector. The variance of the production vector is given by

$$\text{Var}(\mathbf{P}) = \sum_{s=0}^{\infty} \mathbf{B}^s \mathbf{H} \text{Var}(\mathbf{D}_t + \boldsymbol{\xi}_t) \mathbf{H}' \mathbf{B}'^s \quad (3.21)$$

where $\text{Var}(\mathbf{D}_t + \boldsymbol{\xi}_t)$ is a matrix that consists of the variance of \mathbf{D}_t and the covariances of the noise vector $\boldsymbol{\xi}_t$. We note that (3.21) provides the production variances for each station as well as the covariance for each pair of workstations. The knowledge of the covariances will be useful for understanding the interdependence among the noise terms.

The power series in (3.21) can be computed analytically or numerically using the methods discussed in Graves (1986). In *Appendix A*, we show that the series in (3.20) and (3.21) will converge if and only if the spectral radius of $\boldsymbol{\Phi}$ is less than one. This condition of convergence is identical to that for the TPM model in Graves (1986). This condition suggests that a unit of work processed at one station cannot eventually result in more than one unit of work for the same workstation, or else the system will not reach a steady state.

Now we compute the expected queue length at each workstation. We are only interested in queue lengths of the workstations, and not the dummy stations. Thus we let P'_{it} , Q'_{it} and A'_{it} be the production, queue and arrival random variables respectively, of workstation i . Now we compute the expected queue length at each workstation. By taking the expectation of (3.7) for workstation i (i.e. $P'_{it} = \beta_i Q'_{it} + \gamma_i A'_{it}$) and recognizing that we require $E[A'_{it}] = E[P'_{it}]$ in order for the system to be in steady-state, we have

$$\begin{aligned}
E[P'_{it}] &= \beta_i E[Q'_{it}] + \gamma_i E[A'_{it}] \\
E[P'_{it}] &= \beta_i E[Q'_{it}] + \gamma_i E[P'_{it}] \\
E[Q'_{it}] &= \left(\frac{1 - \gamma_i}{\beta_i} \right) E[P'_{it}] \\
E[Q'_{it}] &= n_i E[P'_{it}]
\end{aligned}
\tag{3.22}$$

We express (3.22) for the workstations in matrix form as

$$E[\mathbf{Q}'] = \mathbf{J}E[\mathbf{P}'] \tag{3.23}$$

where $E[\mathbf{P}']$ and $E[\mathbf{Q}']$ are the expected production and queue vectors respectively, of the workstations, while \mathbf{J} is the diagonal matrix with the *SPLT* of each station n_i on the diagonal. We note that (3.23) is consistent with the Little's Law.

3.4 MULTI-FAMILY MODEL

In the last section, we have developed a model for the production of a single aggregate product. Now we consider a manufacturing system with multiple product families and where each workstation processes jobs from one or more families. We are interested to determine the total production requirements and queue lengths for all product families at each workstation, and also the influence of each family on the production system.

We first analyze the production for each individual product family and then find the aggregate performance measures for all combined product families. We make use of the analysis in the previous section to develop the model for each product family k . We restate (3.1) for the planning window of product family k as

$$W_k = L_k - N_k + 1 \tag{3.24}$$

where N_k and L_k are the *PPLT* and *DLT* respectively for product family k . The *PPLT* of product family k is the sum of the *SPLTs* of all workstations along the processing route of family k ; we re-express (3.2) as

$$N_k = \sum_{i \in S_k} \omega_{ki} n_i \quad (3.25)$$

where S_k is the set of workstations along the processing route, ω_{ki} is the number of times each job from product family k visits workstation i , and n_i is the *SPLT* of workstation i . We must satisfy the constraints in (3.24) and (3.25) in setting the planning windows and *SPLTs* of each product family.

Suppose that there are f product families (i.e. we have f dummy stations) and m workstations. In this multi-family setting, we assume that the demand is independent between the product families. To begin, we first denote the demand for product family k in period t as \mathbf{D}_{kt} and the noise arrivals at workstation i in period t due to product family k as ξ_{kit} . We define the matrix for the demand variance as well as the noise terms for product family k , $\text{Var}(\mathbf{D}_{kt} + \xi_{kt})$ as a $(m + f) \times (m + f)$ matrix. Figure 3.3 shows the form of the matrix.

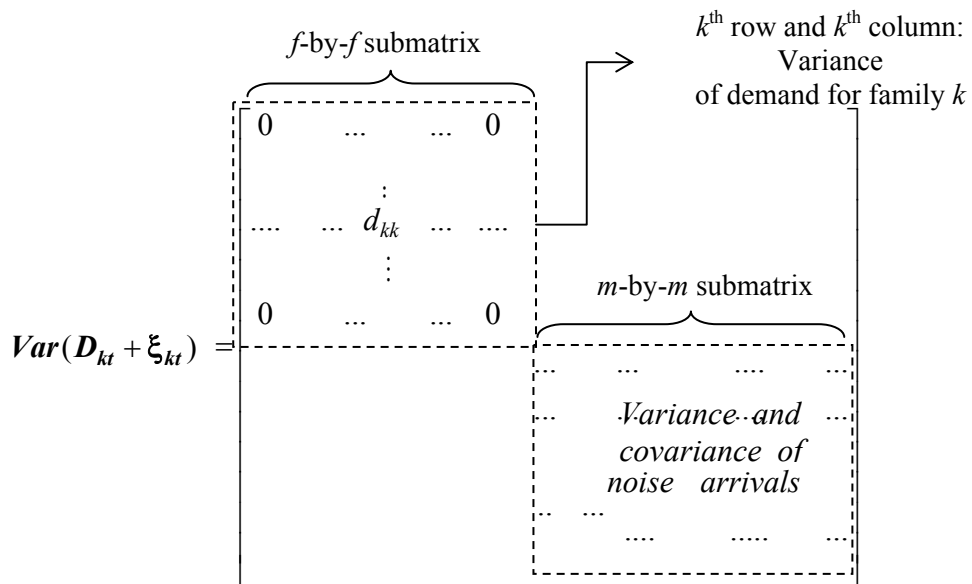


Figure 3.3 Matrix $\text{Var}(\mathbf{D}_{kt} + \xi_{kt})$ for each product family k

In Figure 3.3, the f -by- f submatrix characterizes the variance of demand for product family k . For convenience, we let the k^{th} row and column correspond to product family k . The only non-zero element in the submatrix is the variance of demand for product family k , which is denoted by d_{kk} . The other submatrix, which is of dimension m -by- m , represents the variances and covariances of the noise arrivals at the workstations due to family k .

The workflow matrix Φ_k characterizes the workflow for product family k . We define the matrix as a $(m + f) \times (m + f)$ matrix. Φ_k represents the workflow only for product family k whose MPS smoothing is represented by only one dummy station. Figure 3.4 shows the form of the Φ_k matrix.

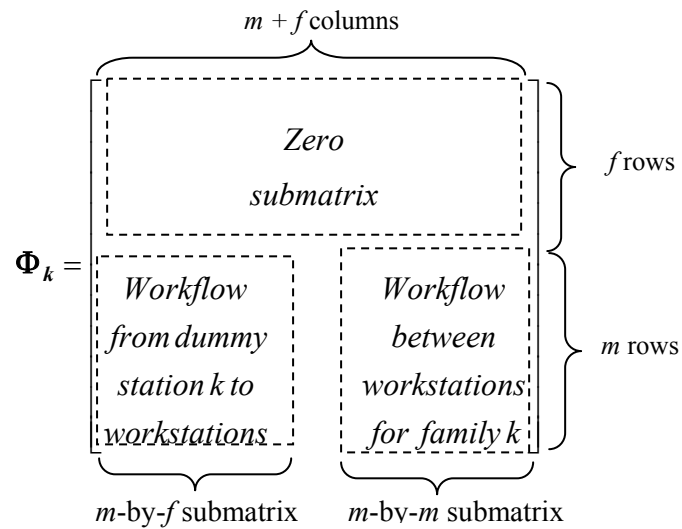


Figure 3.4 Workflow matrix Φ_k for product family k

As shown in Figure 3.4, the first f rows consist of a zero submatrix and the remaining m rows of the Φ_k matrix consist of two submatrices. The m -by- f submatrix models the workflow from the dummy station k , as was the case for the single-family model. The other submatrix, which is a m -by- m submatrix, models the workflow between the workstations in processing jobs for family k .

In order to be dimensionally consistent with the $\text{Var}(\mathbf{D}_{kt} + \xi_{kt})$ and Φ_k matrix, we need to re-define the dimensions of the matrices in the single-family model. The

vectors for \mathbf{P}_{kt} , \mathbf{Q}_{kt} and \mathbf{A}_{kt} , which are the production, queue and arrival vectors, respectively for product family k are:

$$\mathbf{P}_{kt} = \begin{pmatrix} P_{0kt} \\ P_{1kt} \\ P_{2kt} \\ \vdots \\ P_{mkt} \end{pmatrix}, \quad \mathbf{Q}_{kt} = \begin{pmatrix} Q_{0kt} \\ Q_{1kt} \\ Q_{2kt} \\ \vdots \\ Q_{mkt} \end{pmatrix}, \quad \mathbf{A}_{kt} = \begin{pmatrix} A_{0kt} \\ A_{1kt} \\ A_{2kt} \\ \vdots \\ A_{mkt} \end{pmatrix}$$

Each of the column vectors for \mathbf{P}_{kt} , \mathbf{Q}_{kt} and \mathbf{A}_{kt} consists of $m + f$ rows, instead of $m + 1$ rows in the single-family model. For ease of presentation, we let the top f rows of each column vectors to correspond to the dummy stations and the next m rows to represent the workstations. However we note that each product family requires solely one dummy station, which are denoted by P_{0kt} , Q_{0kt} and A_{0kt} in the k^{th} row; hence only one out of the f elements that represents the dummy stations is non-zero.

We restate the continuous-time production function in (3.13) for product family k as

$$\mathbf{P}_{kt} = \mathbf{G}\mathbf{Q}_{kt} + \mathbf{F}\mathbf{A}_{kt} \quad (3.26)$$

where matrix \mathbf{G} and \mathbf{F} follows the definitions in the single-family model, except that they are now square matrices of dimension $m + f$. In parallel to the \mathbf{H} and \mathbf{B} matrices in the single-family model, we let $\mathbf{H}_k = (\mathbf{I} - \mathbf{F}\mathbf{\Phi}_k)^{-1}\mathbf{G}$ and $\mathbf{B}_k = \mathbf{I} - \mathbf{H}_k(\mathbf{I} - \mathbf{\Phi}_k)$.

We let $\boldsymbol{\mu}_k$ be the vector of expected demand for product family k which is a column vector with the expected demand in the k^{th} row as the only non-zero element. Corresponding to (3.20) and (3.21) in the single-family model, we obtain the mean and variance of the production vector for product family k , which we restate as

$$\mathbf{E}[\mathbf{P}_k] = (\mathbf{I} - \mathbf{\Phi}_k)^{-1}\boldsymbol{\mu}_k \quad (3.26)$$

$$\text{Var}(\mathbf{P}_k) = \sum_{s=0}^{\infty} \mathbf{B}_k^s \mathbf{H}_k^s \text{Var}(\mathbf{D}_{kt} + \boldsymbol{\xi}_{kt}) \mathbf{H}_k'^s \mathbf{B}_k'^s \quad (3.27)$$

We let matrix \mathbf{J} to take on the same definition as in the single-family model, i.e. the diagonal matrix with the *SPLT* of each station n_i on the diagonal. Corresponding to (3.23), the expected queue length for product family k is:

$$\mathbf{E}[\mathbf{Q}_k'] = \mathbf{J}\mathbf{E}[\mathbf{P}_k'] \quad (3.28)$$

We then obtain the mean and variance of the total production requirements and the total queue lengths by aggregating the production across all product families.

$$\mathbf{E}[\mathbf{P}_{total}] = \sum_k \mathbf{E}[\mathbf{P}_k] \quad (3.29)$$

$$\text{Var}(\mathbf{P}_{total}) = \sum_k \text{Var}(\mathbf{P}_k) \quad (3.30)$$

$$\mathbf{E}[\mathbf{Q}_{Total}'] = \sum_k \mathbf{J}\mathbf{E}[\mathbf{P}_k'] \quad (3.31)$$

Thus we have obtained vectors for the first two moments of the total production requirements as well as the mean total queue lengths, as functions of the planning windows and the *SPLTs*. The production planner can assess the setting of the planning windows and *SPLTs* by utilizing these vectors. In the next section, we embed our model into optimization procedures to determine the optimal planning windows and *SPLTs*.

We have assumed in our model above that the demands are independent between each pair of product families. However, in the multi-family setting, an important consideration would be the covariances between the demands of the product families. Here we discuss how we can relax this assumption. One way to take into account the demand correlations is to model the production network as a single linear system; this is in contrast to the previous model in which we model each product family separately

and then aggregate the resulting model output. Here we model each actual workstation as f individual sub-workstations, where each sub-workstation is dedicated to a single product family. The motivation here is to model the workflow of all the product families within a single model so as to capture the effects of the correlated demands on the workstations. Therefore we would have a total of fm sub-workstations, in addition to the f dummy stations to smooth the MPS. For each sub-workstation, we model the production requirement where P_{ikt} denotes the production of family k (where $k = 1, 2, \dots, f$) at workstation i in time period t ; similarly, we model the queue length Q_{ikt} . The workflow would be captured by a single $(f + fm) \times (f + fm)$ workflow matrix Φ . The f -by- f submatrix in the $\text{Var}(\mathbf{D}_t + \boldsymbol{\xi}_t)$ matrix (which is also a $(f + fm) \times (f + fm)$ matrix) would characterize both the variances and covariances of demand for all the product families. The resulting $\text{Var}(\mathbf{P}_t)$ matrix for this system would give both the variances and covariances of the production requirements, which we can then utilize to compute the total production variance for workstation i across all the product families. From the covariance matrix $\text{Var}(\mathbf{P}_t)$, we find that

$$\text{Var}(P_{it}) = \sum_{k=1}^f \text{Var}(P_{ikt}) + 2 \sum_{k=1}^f \sum_{v=1}^{f-k} \text{Cov}(P_{ikt}, P_{i,k+v,t}) \quad (3.32)$$

To determine the total expected production requirement for each workstation, we can simply sum the expected production requirement at all the sub-workstations for workstation i , i.e.

$$E[P_{it}] = \sum_k E[P_{ikt}] \quad (3.33)$$

Corresponding to (3.28) and (3.31)), the expected queue length for product family k at workstation i and the total queue length at workstation i are, respectively

$$E[Q'_{ikt}] = n_i E[P'_{ikt}] \quad (3.34)$$

$$E[Q'_{it}] = \sum_k E[Q'_{ikt}] \quad (3.35)$$

Thus we obtain the first two moments of production requirements and queue lengths for the multi-family setting with correlated demands.

3.5 OPTIMIZATION MODEL

We present the nonlinear optimization program of which our objective is to minimize the total expected expediting cost (e.g. overtime and subcontracting) plus total WIP inventory holding cost, with the planning windows and *SPLTs* as the decision variables.

$$\begin{aligned}
 & \text{Min } \sum_i c_i E[P_{it} - M_i]^+ + h_i E[Q_{it}] \\
 & \text{s.t.} \\
 & \sum_i \omega_{ki} n_i + W_k - 1 = L_k, \quad \forall k \\
 & W_k \geq a_k, \quad \forall k \\
 & n_i \geq b_i, \quad \forall i
 \end{aligned}$$

We denote c_i as the expediting cost per hour for workstation i , M_i as the nominal capacity for workstation i (e.g. hours per period) and h_i as the WIP inventory holding cost in terms of workload, e.g. per workhour per day. The other notations follow the definitions stated earlier in this chapter, where P_{it} is the total production requirements across all product families at workstation i , Q_{it} is the sum of queue lengths for all families at workstation i , ω_{ki} is the number of times a job from product family k visits workstation i , n_i is the *SPLT* of workstation i , W_k is the planning window for product family k and L_k is the *DLT* for product family k .

In the objective function, as usual, x^+ implies $\max(x,0)$ and therefore $E[P_{it} - M_i]^+$ represents the expected amount of expediting work per period. The objective function is the total expected expediting cost plus total WIP inventory holding costs per period for all the workstations. We assume that the production requirement P_{it} is normally distributed; as discussed in Chapter 2, the validity of this assumption improves if we set a longer discrete time period relative to the job arrival interval in order for the

Central Limit Theorem to apply. In addition, we note that P_{it} is normally distributed if we assume the demands and noise terms to be normal. The normality assumption allows us to solve the objective function by the normal linear loss integral. To compute the loss integral, we need to find the mean and variance of P_{it} , which we obtain using (3.29) and (3.30). The first set of constraints defines the relationship between the planning windows, *SPLTs* and the *DLT* (see (3.24) and (3.25)). The second and third sets of constraints assure that the planning windows and *SPLT* of each workstation i are at least a_k and b_i respectively, which are the minimum durations that are suitable for production control in keeping up with the dynamics of the system. If the arrival rate to a workstation is high and greatly fluctuating, a short *SPLT* would imply a need for frequent monitoring at the workstation to track the job progress. Likewise, a short planning window for a highly varying demand would require a large amount effort in monitoring the MPS. Hence both a_k and b_i are managerial inputs that must be set appropriately in order to avoid complications in the production control.

We now look at the convexity of the optimization model. We conjecture that the objective function is convex and produce a plausible argument here for this conjecture. The expected amount of expediting work at workstation i , $E[P_{it} - M_i]^+$ is smaller with smoother production requirements, which in turn depends on the planning windows as well as the *SPLTs* of workstation i and all its upstream workstations. Longer planning windows and *SPLTs* result in smoother production requirements. Furthermore, as we increase the amount of smoothing, we get decreasing returns in terms of less variable production requirements. Therefore, the expected amount of expediting work decreases as we increase the planning windows and *SPLTs*; we expect that the rate of decrease declines with larger values of the planning windows and *SPLTs*. Thus, we argue that the expected amount of expediting work at each workstation is strictly convex.

To establish the convexity of the objective function, we would need to consider the Hessian matrix $U(\mathbf{n})$ associated with the expected expediting work, which is given by

$$U(\mathbf{n}) = \partial^2 f(\mathbf{n}) / \partial n_i \partial n_j \quad (3.36)$$

where $f(\mathbf{n}) = \sum_i E[P_{it} - M_i]^+$ and \mathbf{n} is the vector for the *SPLTs* (every element of $\mathbf{n} > 0$). We have not explicitly analyzed (3.36) for our model. But our above argument for convexity, in terms of the relationship between the *SPLTs* and the expediting cost, implies that the second partial derivatives of $f(\mathbf{n})$ with respect to \mathbf{n} are positive, i.e. each element of $\mathbf{U}(\mathbf{n}) > 0$. This entails that $\mathbf{n}'\mathbf{U}\mathbf{n} > 0$, which is the sufficient condition for $\mathbf{U}(\mathbf{n})$ to be positive definite and $f(\mathbf{n})$ strictly convex. Since the expected WIP inventory holding cost $E[Q_{it}]$ at workstation i is linear with the *SPLT* (see (3.31)), we have a convex objective function. Furthermore, as the constraints are linear as well, we have a convex optimization program which ensures a global optimum solution. In *Chapter 5*, we solve an optimization model of similar structure by using the *Sequential Quadratic Programming* method in the *MATLAB* optimization toolbox.

3.6 SUMMARY

In this chapter, we have discussed the roles of the planning windows and the planned lead times (in terms of *PPLT* and *SPLT*) for production smoothing in the MTO environment. We have also explained the key considerations in setting these parameters: the tradeoff between the WIP inventory and the production smoothness in setting the *SPLT*, the *DLT* constraint that limits the parameter values, and the allocation between the planning window and *PPLT* within the *DLT*. We model the smoothing of the MPS by inserting a dummy station into the network of workstations which we model as a job shop. The dummy station and the workstations produce according to the linear production rule and the continuous-time production function respectively. We first obtain the first two moments of production requirements and the expected queue length for systems that produce only one aggregate product family. We then extend the single-family model to systems with multiple product families. Eventually, we build an optimization model that minimizes total expediting cost plus WIP inventory holding cost. In the next chapter, we illustrate the use of the model by a numerical example.

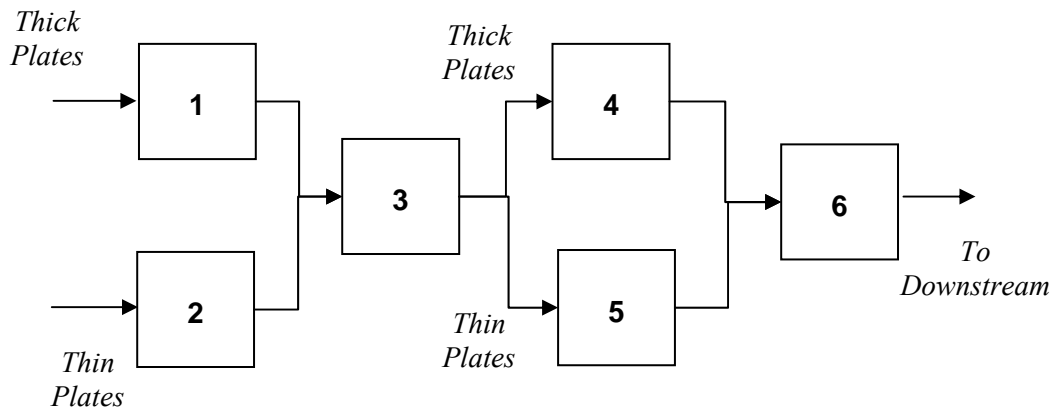
CHAPTER 4

NUMERICAL EXAMPLE

4.1 SCENARIO

In this chapter, we use a simple example based on a real-world setting to illustrate the roles of the planning window and *SPLT* in production smoothing, and also the use of our model. We look at a small production shop that is part of a manufacturing facility of an oil-rig builder. (In *Chapter 5*, we present a case study on another relatively larger facility of the same oil-rig builder). The data presented in this chapter has been altered to protect proprietary information. However the resulting qualitative relationships and insights drawn from this example are the same as they would be from using the actual data.

The shop processes and cuts steel plates into required dimensions, which are then used for downstream production stages to construct the necessary steel parts. The shop operates in a MTO environment because the internal customer orders are highly customized, due to the numerous possible cutting dimensions and plate thicknesses. As part of the ongoing improvement efforts, the management is examining how to classify the numerous production options into product families as well as how to set the planned lead times appropriately. In this example, we analyze the situation in which we categorize the customer orders according to the plate thickness. We classify the products into two product families; those whose thicknesses exceed a specified dimension are classified as *Thick Plates* and those less than the dimension as *Thin Plates*. The shop consists of 4 types of processing workstations, namely Blasting, NC (Numerical Control) Gas Cutting, NC Plasma Cutting and Manual Cut. The process flow map is shown in Figure 4.1, in which we also show the dummy stations that smooth the job release of each product family.



1. Dummy station for Thick Plates
2. Dummy station for Thin Plates
3. Blasting (Both)
4. NC Gas Cutting (Thick Plates)
5. NC Plasma Cutting (Thin Plates)
6. Manual Cut (Both)

Figure 4.1 Process Flow Map

The details of the production processes are as follows. Raw steel plates (both *Thick Plates* and *Thin Plates*) are first sent to the Blasting station where ball grids are splashed onto the plate surfaces to remove impurities, followed by an application of a primary paint layer that protects the steel against corrosion. Subsequently, the *Thick Plates* and *Thin Plates* are sent to different downstream workstations. The blasted *Thick Plates* are transferred to the NC Gas Cutting station, which is capable of cutting the *Thick Plates*. The blasted *Thin Plates* are moved to the NC Plasma Cutting station, which can only cut the *Thin Plates*. After cutting at the NC machines, the steel plates are then sent to the Manual Cut station where the final cutting is done by manual labor using hand-held tools.

The daily internal demand for the *Thick Plates* has a mean of 20 plates and standard deviation of 10 plates, while the demand for *Thin Plates* has a mean of 26 plates and standard deviation of 12 plates. The demands for the product families are uncorrelated. Furthermore, we approximate the demand processes for both product families to be of normal distribution. The firm delivery lead times for the *Thick Plates* and *Thin Plates* are 9 days and 8 days respectively. In Table 4.1, we show the

expectation and standard deviation of the effective processing times at each workstation, which we assume to be normally distributed. The standard deviation of the effective processing time is attributed to the different processing requirements of each order. We compute the resulting noise arrival due to the variability in the effective processing times using (3.11). In the same table, we also show the nominal capacity available per day in workhours. Each workstation has flexibility to expand its capacity in each day by subcontracting its production to nearby local shops. We also show the subcontracting cost per hour for outstanding work as well as holding cost per hour of work at each workstation. For the holding cost, the available data is the average holding cost per plate but we require the expected holding cost per hour for our model; we obtain the required data by dividing the average holding cost per steel plate by the expected processing time at each workstation.

Table 4.1 Data for Workstations

	Thick Plates		Thin Plates		Capacity (hours/day)	Subcontract Cost (\$ per hour)	Holding Cost (\$ per workhour per day)
	E[hours per plate]	SD (hours per plate)	E[hours per plate]	SD (hours per plate)			
Blasting ¹	0.55	0.35	0.55	0.35	28	550	0.72
NC Gas Cutting	1.69	1.96	-	-	43	368	0.61
NC Plasma Cutting	-	-	1.34	1.46	49	441	0.77
Manual Cut	3.50	2.55	1.07	0.86	128	788	0.74

¹The expectation and standard deviation of processing times at the Blasting station are identical for both product families

4.2 MODEL INPUTS

We present and explain the inputs for the model as follow:

Workflow matrix: The workflow matrices for *Thick Plates* and *Thin Plates*, denoted by Φ_{Thick} and Φ_{Thin} respectively, are given as:

$$\Phi_{Thick} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.55 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.08 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.07 & 0 & 0 \end{bmatrix} \quad \Phi_{Thin} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.55 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.43 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.80 & 0 \end{bmatrix}$$

For example, each demand unit for *Thick Plates* (each unit of work at Dummy Station for *Thick Plates*) generates 0.55 hour of work at the Blasting station and thus, $\phi_{31} = 0.55$; in addition, each hour of work at Blasting generates on average $1.69/0.55 = 3.08$ hours of work at the NC Gas Cutting station, i.e. $\phi_{43} = 3.08$.

Variance of Demand and Noise Vector: This model input is needed to compute the variance of production requirements in (3.21) and (3.30). Our assumption of no correlation between the demands as well as between the noises at the workstations implies that the off-diagonal terms of the matrix are all equal zero. We set the diagonal elements for the dummy stations (first two rows) to be the variance of the daily demand; the diagonal elements for the workstations (bottom four rows) represent the noise arrivals, which is equal to the variance of the workload for the mean number of jobs processed at the station per day (equation (3.11)). For example, at the Blasting station, the mean number of *Thick Plates* processed per day is 20 and the standard deviation of effective processing time is 0.35 hour, which gives a workload variance equals to $20(0.35)^2 = 2.45$ hours. The matrices for the variances of demand and noise vector for the two product families are:

$$\text{Var}(D_{Thick,t} + \xi_{Thick,t}) = \begin{bmatrix} 100 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.45 & 0 & 0 & 0 \\ 0 & 0 & 0 & 75.83 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 129.71 \end{bmatrix}$$

$$\begin{aligned} & \text{Var}(D_{Thin,t} + \xi_{Thin,t}) \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 144 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.19 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 55.04 & 0 \\ 0 & 0 & 0 & 0 & 0 & 19.16 \end{bmatrix} \end{aligned}$$

4.3 SETTING OF PLANNING WINDOWS AND *SPLTs*

Now we are ready to consider several scenarios of production smoothing at the workstations. We specify each scenario by setting the planning window for each product family and the *SPLT* for each workstation. To simplify our illustrations, we set the planning windows and the *SPLTs* to take on only integer values.

We show how the setting of the planning windows and *SPLTs* (given the *DLT* constraint) affect the subcontracting and inventory holding costs. We also illustrate the smoothness-inventory tradeoff in setting the *SPLTs*, as well as the tradeoff between MPS smoothing (by planning window) and workload smoothing at the workstations (by *PPLT*). We then demonstrate the value of our model in performing “what-if” analysis.

Case 0: Base Case

First we look at the base scenario of the shop in which the raw plates are “pushed” into production with no smoothing of the MPS. Furthermore, the management allocates the *DLT* roughly equally among the *SPLTs* of the stations. Table 4.2 gives the values of the planning parameters as well as the characterization of the workstations’ behaviors. We report the mean and standard deviation of the release for each product family as well as the production requirement at each workstation, which are computed using (3.26), (3.27), (3.29) and (3.30). We state the production requirement due to each of the product families as well as the aggregated production requirement at each workstation. In addition, we give the expected queue lengths that

are calculated by (3.31) and also compute the probability that the total production requirement exceeds the nominal capacity. We also determine the expected subcontracting cost per day and holding cost per day for each workstation as well as the expected total cost for the shop. Due to the normality assumption of the internal demands and the effective processing times, the production requirements at the stations are also normally distributed. We solve the expected subcontracting cost using the normal linear loss integral. To compute the loss integral, we find the mean and variance of the production requirements, which we obtain using (3.29) and (3.30). It is relatively straightforward to calculate the expected inventory holding cost using (3.28) and (3.31).

From Table 4.2, we see that the planning window for each product family is equal to one period and thus there is no smoothing of the MPS. As a result, the release into the production shop is highly variable and the standard deviation of the release equal that of the demand process. At the Blasting station, there is a relatively high probability of 0.21 that the production requirement would exceed the nominal capacity, in which case the shop has to subcontract some work in order to complete the production within the planned lead time. The expected subcontracting cost of the blasting station is \$223.36 per day. The NC Gas Cutting has a probability of 0.07 that its nominal capacity is inadequate and its expected subcontracting cost is high at \$68.31. The other two workstations have much lower probability of insufficient nominal capacity. However the queue length at the Manual Cut station is the highest, with 293.8 hours of work in queue. The total cost for the shop per day is \$712.84.

Case 1: Smoothing of MPS

Now we attempt to smooth the MPS for both product families by increasing the planning windows of each product family to 3 days. In order to satisfy the delivery lead time requirement, this increase in the planning window must be compensated by a reduction of the *SPLTs* at the workstations. We observe that the Manual Cut station has the lowest expected subcontracting cost and the highest inventory holding cost. Hence we choose to reduce the *SPLT* of the Manual Cut station from 3 days to 1 day. The results of these changes are shown in Table 4.3.

We see from Table 4.3 that the releases for both product families become less variable. For instance, the standard deviation of the release for *Thick* Plates falls from 10 plates to 4.47 plates. The smoother release results in less variable job arrivals at the workstations which in turn causes smoother production requirements. For example, the standard deviation of production requirement for the Blasting station decreases from 3.38 hours in the last case to 2.76 hours. This leads to a reduction in the probability of insufficient nominal capacity at the Blasting station from 0.21 to 0.16, with a corresponding fall in expected subcontracting cost from \$223.36 to \$119.64. The NC Gas Cutting and the NC Plasma Cutting also undergo a reduction in variability of production requirements.

The Manual Cut station experiences two opposite effects on its production variability. On the one hand, it now has a shorter *SPLT* to smooth the arrival variability but on the other hand, the resulting longer planning windows leads to smoother job arrivals to the workstation. However the combined effect causes an increase in production variability at the workstation because of the relatively higher variability of noise than work arrivals. We note that its expected subcontracting cost rises from \$21.37 to \$77.14. But its queue length decreases from 293.79 hours to 97.9 hours with a decline in expected holding cost from \$216.53 to \$72.12. As a result of the MPS smoothing, the expected total cost for the shop falls significantly from the previous case of \$712.84 to \$495.50.

Table 4.2 Case 0: Original Case

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	1	20.00	10.00	-	-	20.00	10.00	-	-	-	-	-
Release (Thin Plates)	1	-	-	26.00	12.00	26.00	12.00	-	-	-	-	-
Blasting	3	11.00	2.17	14.30	2.59	25.30	3.38	75.90	28	0.21	223.36	55.20
NC Gas Cutting	3	33.84	6.14	-	-	33.84	6.14	101.6	43	0.07	68.31	62.49
NC Plasma Cutting	2	-	-	34.88	6.33	34.88	6.33	69.5	49	0.01	11.72	53.87
Manual Cut	3	70.12	11.40	27.73	4.33	97.85	12.19	293.8	128	0.01	21.37	216.53
Total Cost (\$) =										712.84		

Table 4.3 Case 1: Smoothing the MPS

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	3	20.00	4.47	-	-	20.00	4.47	-	-	-	-	-
Release (Thin Plates)	3	-	-	26.00	5.37	26.00	5.37	-	-	-	-	-
Blasting	3	11.00	1.77	14.30	2.11	25.30	2.76	75.90	28	0.16	119.64	55.20
NC Gas Cutting	3	33.88	5.77	-	-	33.84	5.77	101.5	43	0.06	48.36	62.41
NC Plasma Cutting	2	-	-	34.75	5.87	34.88	5.87	69.8	49	0.01	6.55	54.07
Manual Cut	1	70.13	13.60	27.80	5.22	97.85	14.58	97.9	128	0.02	77.14	72.12
Total Cost (\$) =										495.50		

Case 2: More Smoothing of MPS for Thin Plates

We note in the previous case that the NC Plasma Cutting station has a low probability of 0.01 that its production requirement would exceed nominal capacity with a corresponding low expected subcontracting cost of \$6.55. Thus we reduce the *SPLT* of the NC Plasma Cutting station by one day (i.e. from 2 days to 1 day) and reallocate this one day (i.e. from 3 days to 4 days) to the planning window for the *Thin Plates* to further smooth its MPS.

From Table 4.4, as a result of these changes, we observe that the release for *Thin Plates* becomes less variable with a reduction in standard deviation from 5.37 to 4.54 plates. The less variable release leads to a smoother production requirement at the Blasting station, which has its expected subcontracting cost reduced from \$119.64 to \$105.98. Due to the reduction in the *SPLT* at the NC Plasma Cutting station, its queue length decreases and its expected holding cost reduces from \$54.07 to \$27.04. However the production variability increases and its subcontracting cost raises from \$6.55 to \$19.90. The production variability at the Manual Cut also increases due to the more variable work arrival from the NC Plasma Cutting station. The overall effect causes the expected total cost to decrease to \$471.33.

Case 3: Assessing Further Smoothing of MPS

In this case, we assess whether it is possible to further smooth the MPS by reallocating the *SPLTs* of the Blasting and the NC Gas Cutting station to the planning windows.

Case 3A: Allocating *SPLT* of NC Gas Cut to planning window

First we consider the reduction of the *SPLT* of the NC Gas Cutting station from 3 days to 2 days and subsequently increase the planning window of the *Thick Plates* from 3 days to 4 days. The results of these changes are shown in Table 4.5. Although the increase in the planning window for *Thick Plates* leads to smoother job arrivals at the NC Gas Cutting station but it also results in a shorter *SPLT* to smooth the noise arrivals. The overall effect leads to a more variable production requirement at the NC Gas Cutting station as there is comparatively more noise arrival than variability in job arrivals. As a result, its expected subcontracting cost increases from \$48.70 to \$71.33.

However its queue length decreases and the expected holding cost falls from \$62.41 to \$41.61. The increase in production variability at the NC Gas Cutting station results in a more variable production requirement to the downstream Manual Cut station, which has its subcontracting cost increases from \$79.98 to \$96.56. The increase in the planning window of the *Thick Plates* leads to a smoother release into the shop and consequently, the expected subcontracting cost for the Blasting station decreases from \$105.98 to \$96.81. These changes increases the expected total cost from \$471.32 in Case 2 to \$480.34, and thus we should not adopt this change.

Case 3B: Reallocating *SPLT* of Blasting to planning window

Now we consider adjusting the parameters in Case 2 by reducing the *SPLT* of the Blasting station by one day (from 3 days to 2 days) and allocate this one day to the planning windows of both the *Thick Plates* and *Thin Plates*. The results are shown in Table 4.6. The resulting effect is an increase in the variability of production requirement at the Blasting station and consequentially the production variability of the downstream workstations. Hence there is an increase in the expected subcontracting cost at all workstations. However the decrease in the *SPLT* at the blasting station brought about a reduction in the expected queue length and the resulting expected holding cost falls from \$55.20 to \$36.80. This increase in expected holding cost more than offset the reduction in the expected subcontracting cost at the Blasting station. The expected total cost decreases slightly from \$471.33 to \$465.27 which implies that we should adopt this adjustment of parameters.

Case 3C: Further reallocation of *SPLT* of Blasting to planning window

Now we attempt to further smooth the MPS by reallocating another day of the Blasting station's *SPLT* to the planning windows. We decrease the *SPLT* of the Blasting from 2 days in *Case 3A* to 1 day and in turn increase the planning windows of the *Thick Plates* and *Thin Plates* to 6 and 5 days respectively. We show the results in Table 4.7. In this case, the increase in expected subcontracting costs of all the stations due to the decrease in the *SPLT* is greater than the resulting decrease in the expected holding cost at the Blasting station. As a result, the expected total cost increases from \$465.27 to \$472.35.

We obtain the optimal planning windows and *SPLTs* using the optimization program that we develop in *section 3.5*. We reproduce the program below:

$$\begin{aligned}
 & \text{Min } \sum_i c_i E[P_{it} - M_i]^+ + h_i E[Q_{it}] \\
 & \text{s.t.} \\
 & \sum_i \omega_{ki} n_i + W_k - 1 = L_k, \quad \forall k \\
 & W_k \geq a_k, \quad \forall k \\
 & n_i \geq b_i, \quad \forall i
 \end{aligned}$$

The notations follow that described in *section 3.5*. We assume $a_k = 1$ for all k and $b_i = 1$ for all i , and we do not restrict the planning windows and *SPLTs* to be integers. The optimal solution and the associated results are shown in Table 4.8. We observe that the minimum expected total cost is \$464.10, which is quite close to that in Case 3B. Now we compare the expected total cost in *Case 0* and our final optimal cost. The total expected cost falls from \$712.84 to \$464.10, which is a 34.9% reduction.

Case 4: “What-if” Analysis

In this case, we illustrate the capability of our model for “what-if” analysis with two examples: what if we shorten the DLT? What if we increase the nominal capacity to meet this change?

Case 4A: Shortening of DLTs for both product families

Suppose that the management of the shop considers reducing the delivery lead times of both product families to 5 days. The management wants to determine the consequential impact on its production resources. In Table 4.9, we obtain the optimal planning windows and the *SPLTs* in order to meet this shorter DLT. We note that the variability of production requirements at the workstations increases due to the more variable release and also because of the shorter *SPLT* at each workstation. The total expected cost in this case is \$797.95.

Case 4B: Increasing nominal capacity at Blasting

In face of the reduction in production smoothness in Case 4A, the management might also consider increasing the nominal capacity at some workstations. For example, if the management increases the nominal capacity of the Blasting station to 30 hours per day, the optimal total expected cost would be \$638.02. This is illustrated in Table 4.10.

We have illustrated only two “what-if” scenarios above although there are several other possible scenarios for which we can utilize the model to determine the impact of both planned and unplanned changes. For instance, what if the demand increases? What if the subcontractors raise the subcontracting costs, should we acquire more capacity? What if we standardize the cutting dimension and how would the resulting decrease in processing time variability affect the total cost? The management should be able to employ the model to analyze the impact of these changes and make appropriate decisions.

Table 4.4 Case 2: More Smoothing of MPS for Thin Plates

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	3	20.00	4.47	-	-	20.00	4.47	-	-	-	-	-
Release (Thin Plates)	4	-	-	26.00	4.54	26.00	4.54	-	-	-	-	-
Blasting	3	11.00	1.77	14.30	1.96	25.30	2.65	75.90	28	0.15	105.98	55.20
NC Gas Cutting	3	33.88	5.77	-	-	33.84	5.78	101.5	43	0.06	48.70	62.41
NC Plasma Cutting	1	-	-	34.75	6.80	34.88	6.81	34.88	49	0.02	19.90	27.04
Manual Cut	1	70.12	13.60	27.80	5.44	97.85	14.66	97.85	128	0.02	79.98	72.12
Total Cost (\$) =										471.33		

Table 4.5 Case 3A: Allocating *SPLT* of NC Gas Cut to planning window

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	4	20.00	3.78	-	-	20.00	3.78	-	-	-	-	-
Release (Thin Plates)	4	-	-	26.00	4.54	26.00	4.54	-	-	-	-	-
Blasting	3	11.00	1.65	14.30	1.96	25.30	2.57	75.90	28	0.15	96.81	55.20
NC Gas Cutting	3	33.88	6.29	-	-	33.84	6.29	67.68	43	0.07	71.33	41.61
NC Plasma Cutting	1	-	-	34.75	6.80	34.88	6.80	34.88	49	0.02	19.68	27.04
Manual Cut	1	70.12	14.08	27.80	5.44	97.85	15.09	97.85	128	0.02	96.56	72.12
Total Cost (\$) =										480.34		

Table 4.6 Case 3B: Reallocating *SPLT* of Blasting to planning window

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	4	20.00	3.78	-	-	20.00	3.78	-	-	-	-	-
Release (Thin Plates)	5	-	-	26.00	4.00	26.00	4.00	-	-	-	-	-
Blasting	2	11.00	1.82	14.30	2.03	25.30	2.73	50.60	28	0.16	115.84	36.80
NC Gas Cutting	3	33.88	5.79	-	-	33.84	5.79	101.5	43	0.06	49.19	62.41
NC Plasma Cutting	1	-	-	34.75	6.84	34.88	6.85	34.88	49	0.02	20.75	27.04
Manual Cut	1	70.12	13.62	27.80	5.46	97.85	14.69	97.85	128	0.02	81.13	72.12
Total Cost (\$) =										465.27		

Table 4.7 Case 3C: Further reallocation of *SPLT* of Blasting to planning window

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	5	20.00	3.33	-	-	20.00	3.33	-	-	-	-	-
Release (Thin Plates)	6	-	-	26.00	3.62	26.00	3.62	-	-	-	-	-
Blasting	1	11.00	1.96	14.30	2.19	25.30	2.94	25.30	28	0.18	142.43	18.40
NC Gas Cutting	3	33.88	5.78	-	-	33.84	5.78	101.5	43	0.06	48.84	62.41
NC Plasma Cutting	1	-	-	34.75	6.91	34.88	6.90	34.88	49	0.02	21.70	27.04
Manual Cut	1	70.12	13.58	27.80	5.47	97.85	14.64	97.85	128	0.02	79.42	72.12
Total Cost (\$) =										472.35		

Table 4.8 Optimal Solutions

	Optimal W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	4.16	20.00	3.70	-	-	20.00	3.70	-	-	-	-	-
Release (Thin Plates)	5.06	-	-	26.00	3.97	26.00	3.97	-	-	-	-	-
Blasting	1.94	11.00	1.82	14.30	2.03	25.30	2.73	49.02	28	0.16	115.61	35.65
NC Gas Cutting	2.90	33.88	5.83	-	-	33.84	5.83	98.16	43	0.06	50.78	60.35
NC Plasma Cutting	1	-	-	34.75	6.84	34.88	6.84	34.88	49	0.02	20.59	27.04
Manual Cut	1	70.12	13.66	27.80	5.46	97.85	14.71	97.85	128	0.02	81.96	72.12
Total Cost (\$) =										464.10		

Table 4.9 Case 4A: “What-if” analysis - Reducing DLT of both Product Families

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	1	20.00	10.00	-	-	20.00	10.00	-	-	-	-	-
Release (Thin Plates)	2	-	-	26.00	6.93	26.00	6.93	-	-	-	-	-
Blasting	2	11.00	8.22	14.30	2.63	25.30	3.67	50.57	28	0.23	247.34	36.78
NC Gas Cutting	2	33.88	7.43	-	-	33.84	7.43	67.72	43	0.11	136.36	41.63
NC Plasma Cutting	1	-	-	34.75	7.72	34.88	7.72	34.88	49	0.03	42.89	27.04
Manual Cut	1	70.12	15.88	27.80	6.04	97.85	16.99	97.85	128	0.04	193.79	72.12
Total Cost (\$) =										797.95		

Table 4.10 Case 4B: “What-if” analysis – Increasing nominal capacity at Blasting

	W_i or n_i	Thick Plates		Thin Plates		Total			Nominal Capacity	Prob {P>Capacity}	Subcontract Cost (\$)	Holding Cost per day (\$)
		E[P]	σ (P)	E[P]	σ (P)	E[P]	σ (P)	E[Q]				
Release (Thick Plates)	1	20.00	10.00	-	-	20.00	10.00	-	-	-	-	-
Release (Thin Plates)	2	-	-	26.00	6.93	26.00	6.93	-	-	-	-	-
Blasting	2	11.00	8.22	14.30	2.63	25.30	3.67	50.57	28	0.10	87.41	36.78
NC Gas Cutting	2	33.88	7.43	-	-	33.84	7.43	67.72	43	0.11	136.36	41.63
NC Plasma Cutting	1	-	-	34.80	7.72	34.88	7.72	34.88	49	0.03	42.89	27.04
Manual Cut	1	70.12	15.88	27.75	6.04	97.85	16.99	97.85	128	0.04	193.79	72.12
Total Cost (\$) =										638.02		

CHAPTER 5

A CASE STUDY ON TACTICAL PLANNING IN OIL-RIG BUILDING

5.1 CASE BACKGROUND

In this chapter, we describe a case study in which we apply our model to a producer of jack-up oil-rigs. Jack-up oil-rigs are offshore rigs that are mobile on waters, and can anchor themselves by deploying jack-like legs. These oil-rigs are built to customers' specifications.

Recently, the company experienced a demand upturn due to the increase in global demand for energy. However, in face of this demand surge, the management does not want to invest heavily on expanding its capacity so as to avoid over-capacity in times of demand downturn. Instead, the company's strategy is to outsource part of its production and at the same time, focuses on improving its production efficiencies. This case study is undertaken as part of the initiatives to improve the company's production planning.

In this chapter, we present our experience in applying our model in this MTO environment. Here we study a production facility that is larger and more complex than the production shop described in *Chapter 4*. In the next section, we describe the products as well as the production processes. In section 5.3, we discuss the challenges that exist in the production system. We explain in section 5.4 how our model can improve the company's production planning and state the objectives for this case study. Subsequently, we describe how we obtain the data needed for the model inputs in section 5.5. In section 5.6, we present the optimization program that finds the optimal planning windows and *SPLTs* for the production system. Next we give some insights on the optimization results in section 5.7. We validate the results and also identify some weaknesses of our model in section 5.8. We discuss in section 5.9 the benefits that this case study has brought upon the oil-rig manufacturer.

5.2 PRODUCT AND PROCESS DESCRIPTION

The fundamental structure of a jack-up oil-rig is the hull as shown in Figure 5.1. The hull is the platform on which most of the facilities of the rig are built. The high-level bill-of-material (BOM) of a typical hull is shown in the same figure.

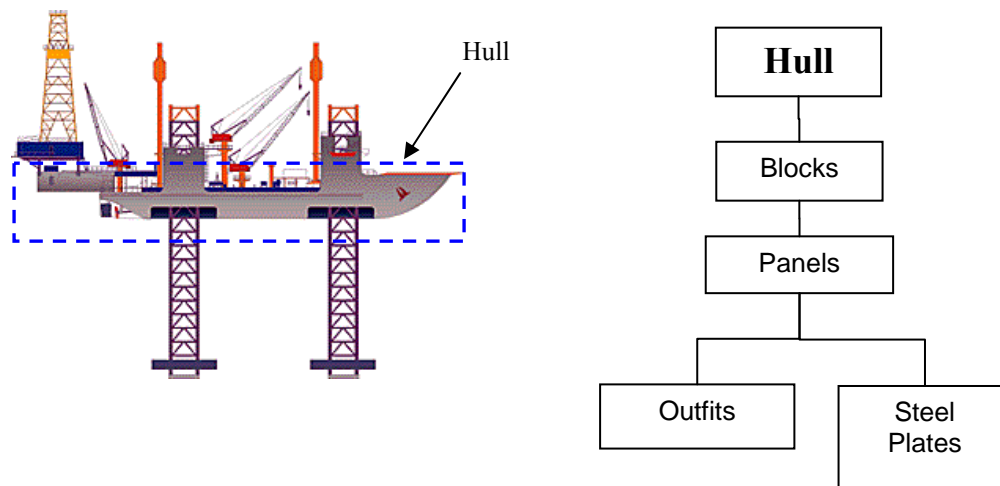
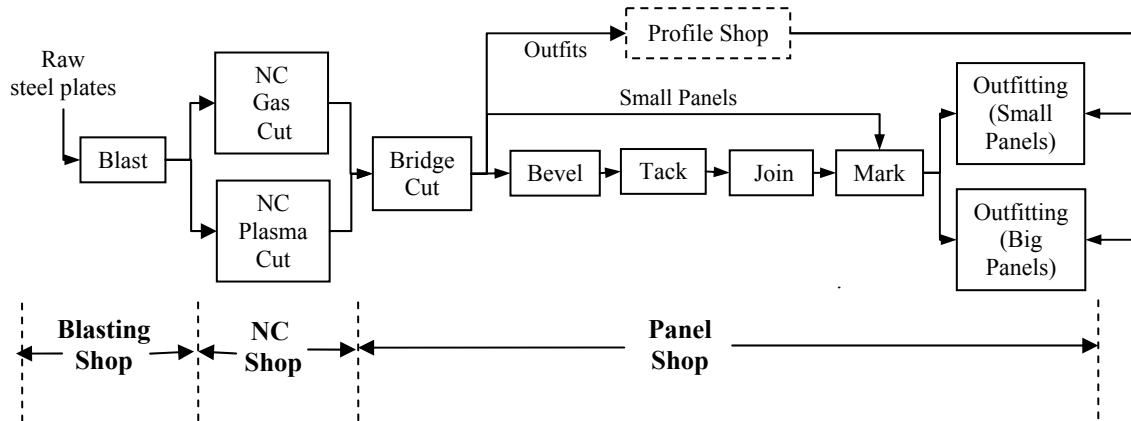


Figure 5.1 High-level BOM for hull

The next level of the BOM is the steel blocks. A typical hull is made up of about fifteen steel blocks. Each steel block is constructed by joining the steel structures called panels. A block commonly consists of ten to more than sixty panels. The steel panel represents the most elementary structure in the hull construction. A panel is built by first welding the steel plates together and then by fitting outfitting components (or outfits) to them to strengthen the structure. But before the welding process, the steel plates have to go through some preparatory processes and be cut according to specified dimensions.

In this case study, we study the production planning for the panels. Figure 5.2 shows the process flow map for the panel production. There are three process routes for the production of *Big Panels*, *Small Panels* and *Outfits*. We define *Big Panels* as panels built by joining two or more steel plates. The *Small Panels* are built on a single steel plate and therefore requires no joining of plates. As described earlier, the *Outfits* are

components that are welded onto the panels (both *Big Panels* and *Small Panels*) to improve their structural strength. In the context of our model, we classify the *Big Panels*, *Small Panels* and *Outfits* as our product families (although strictly speaking, each of them is not a complete product by itself).



Big Panels: Blast→NC Gas→Bridge Cut→Bevel→Tack→Join→Mark→Outfitting (Big Panels)

Small Panels: Blast→ NC Gas→ Bridge Cut→ Mark→ Outfitting (Small Panels)

Outfits: Blast→ NC Plasma or NC Gas→ Bridge Cut→Profile Shop→Outfits (both Big and Small Panels)

Figure 5.2 Process flow map for panel production

Now we give some details of the production process for each product family. We first explain the production of the *Big Panels*. At the *Blast* station, raw plates are “blasted” by ball grids to remove impurities on the steel surface, followed by an application of a coating to protect the plates against corrosion. Next the plates are sent to downstream NC (numerical control) machines to be cut into the required dimensions. There are two stations for NC cutting, namely the *NC Gas Cut* and *NC Plasma Cut*. The *NC Gas Cut* station is capable of cutting thick plates while the *NC Plasma Cut* station can only cut thin plates. The *Big Panels* are constructed using thick plates and are thus sent to the *NC Gas Cut* station. After cutting, the cut parts remain joined to the steel plate by connectors called “bridges”. At the *Bridge Cut* station, the cut parts are split from the steel plates by manually cutting off the bridges. The cut steel plates are then

beveled at the *Bevel* station. Beveling is done only for the thick plates and thus this operation is essential for the *Big Panels*. Subsequently, the plates are transferred to the *Tack* station where the plates are tack welded (sometimes called spot welding) along the join lines to keep the plates in position before the final welding. The tack welded plates then go through the welding process to form the steel panel at the *Join* station. At the *Mark* station, the panels are paint-marked to indicate the locations of the outfits to be fitted. Eventually, the panel is moved to the *Outfitting (Big Panels)* to perform the welding operation of fitting the outfits onto the panel.

Now we proceed to explain the production process for the *Small Panels* and *Outfits*. The *Small Panels* are typically manufactured using thin steel plates. However, after blasting, the steel plates are processed at the *NC Gas Cut* station (which are capable of cutting thick plates) instead of the *NC Plasma Cut*; the motive here is to balance the workload between the two stations. Since the *Small Panels* are built using a single plate, no welding to join the steel plates is needed. Therefore, once the steel plates complete processing at the *Bridge Cut* station, they are moved directly to the *Mark* station and subsequently to the *Outfitting (Small Panels)* station where outfits are welded to complete the panels. Similar to the *Small Panels*, the *Outfits* are also constructed using thin plates. To balance the workload between the two NC stations, about 27% of the *Outfits* are cut in the *NC Gas Cut* station while the rest of the 73% are processed at the *NC Plasma Cut* station. After completing processing at the *Bridge Cut*, the *Outfits* are sent to the *Profile Shop* where the cut steel parts undergo some simple metal forming operations to produce the outfitting components. Finally the *Outfits* are transferred to the *Outfitting (Big Panels)* and *Outfitting (Small Panels)* to be fitted onto the panels. The completed panels are then ready for quality checks and upon satisfying the quality standards, they are moved downstream to construct the steel blocks

5.3 PROBLEMS IN PRODUCTION PLANNING

Production control in the panel production is based on the planned lead times at the shop level. The panel production described in the last section is physically housed in three separate shops, namely the *Blasting Shop*, *NC Shop* and *Panel Shop*, that are

physically located next to each other. The *Blasting Shop* consists of the *Blast* station, the *NC Shop* comprises of the two NC stations and the *Panel Shop* includes the rest of the processing stations. This is indicated in Figure 5.2. The planned lead times for the *Blasting Shop*, *NC Shop* and *Panel Shop* are 5, 8 and 13 days respectively, which gives a total delivery lead time of 26 days. All jobs are expected to complete processing within this duration (from the release of the raw steel plates to completed panels). Therefore all jobs are released 26 days before due-dates, regardless of the product family that the job is from; this implies that the *Small Panels* and *Outfits* are released the same number of days beforehand as the *Big Panels*, even though they require fewer processing steps. Hence there are opportunities to smooth the MPS by creating planning windows for the two product families.

Variable MPS and release due to

1. Lack of coordination between project schedules
2. Unavailable plates
3. Variable picking time for plates

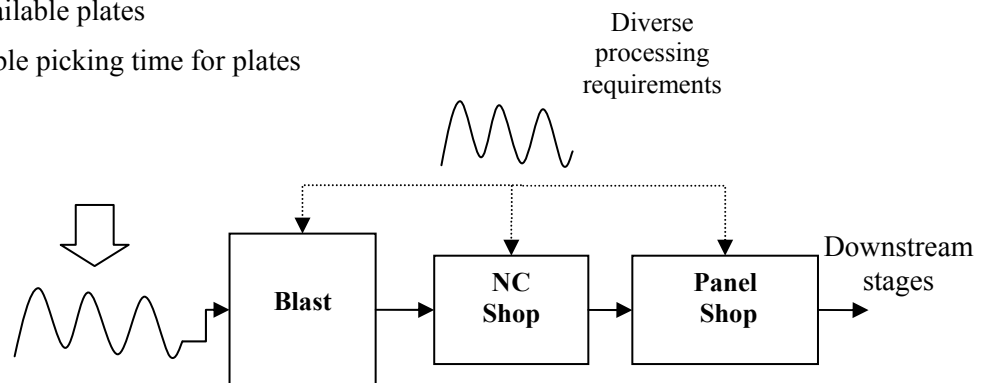


Figure 5.3 Sources of variability in panel production

The main challenge in the panel production is the large amount of variability that exists in the system. We illustrate the sources of variability in Figure 5.3. First, there is a lack of coordination between production schedules of panels for the different oil-rig projects. As a result, the total internal demand for panels is highly fluctuating. Second, raw steel plates of required thickness and grades are frequently unavailable and this delays the release of some panels. Third, the raw steel plates are stacked to conserve space and therefore picking a required plate at the bottom of the stack can take substantial time; this adds considerable variability to the picking time of the raw plates. As a result, the final MPS that is adjusted to these sources of variability is

highly fluctuating and this leads to a variable job release. Furthermore, within the production system, there are diverse processing requirements at the stations due to, e.g. different cut lengths of steel plates, and different number of plates and outfitting components for each panel. Thus the workload at the production stations is highly varying.

The company's main concern is the timely completion of the panels so as to avoid any delay in the production for the downstream stages. In face of this varying workload, there is a frequent shortfall in capacity. Jobs are subcontracted out to vendors (which are also located within the facility) if they are anticipated to be late based on the planned lead times. The management is examining ways to better utilize the capacity and reduce the subcontracting cost.

5.4 MODEL APPLICATION

We identified this as an opportunity to apply our model as we recognized a good fit between the model and the problem features. We believe that our model is able to capture the essence of the problem, namely the production control using planned lead times, the various sources of variability and the workflow of the product families. We presented the underlying concepts of production smoothing as well as our model's capability to the company's managers. They agreed that the model was a suitable tool to study their problems.

The management had also considered the alternative of using simulation to look into the problem, especially so as they already own a simulation model that is built on the platform of commercial simulation software. However, they found that simulation is not suitable for the extensive "what-if" analysis they would perform because it would be slow to make the numerous simulation runs, especially if the runs require optimization. Furthermore, there are constraints in the modeling environment of their existing software that prevents the modeling of the production control using the planned lead times as well as the workload smoothing of the MPS and at the workstations. In contrast, our model can be conveniently formulated and solved in *MATLAB* due to the software's ease of working with matrices. In addition, *MATLAB*

provides an optimization toolbox which we can conveniently utilize to find the optimal solution. More importantly, the project team from the company is familiar with the software which greatly reduces the project lead time.

For this case study, we aimed to achieve the following objectives:

- 1) Separate the schedules of the product families (*Big Panels*, *Small Panels* and *Outfits*) into three individual MPS. This would enable us to smooth the MPS of the individual product family and not release all jobs the same number of days before due-dates. This would also lead to a less variable release into the production system and smoother production requirements at the stations.

- 2) Determine the optimal planning windows and *SPLTs* of the individual stations with the objective of minimizing the total subcontracting cost. At this point of time, the management does not wish to change the current total *DLT* of 26 days because they do not want to affect the production schedule for the downstream stages. Thus our task was to find the optimal allocation of the 26 days to the planning windows and the *SPLTs*. We note that the production control within the *Panel Shop* is not according to the planned lead times, i.e. the subcontracting decisions at each individual station in the *Panel Shop* is not based on its *SPLT*. Rather, the decision at each station is based on predicting whether the job can be completed within the aggregate planned lead time of 13 days. Here we also determine the optimal planned lead times for each station in the *Panel Shop*, so that the company can track the progress of the jobs using the *SPLTs* of the individual station rather than the aggregate planned lead time of the shop. As such, better subcontracting decisions can be made to reduce the subcontracting cost.

There are a few characteristics of the problem that need to be mentioned. We do not consider the work-in-process (WIP) inventories at the workstations. To explain the reasons for this, we need to consider the two main components of inventory holding cost, i.e. the cost of material and the value added to the jobs through processing. The raw plates intended for an entire oil-rig project are typically purchased and stored in the warehouse before production commences. As a result, the cost of material is a

sunk cost and is thus inconsequential to how the plates are scheduled through the production system. In addition, we performed a rough-cut calculation of the value added to the jobs at each production station. We found that the value added is relatively much lower than the subcontracting cost. Hence we decided to ignore the WIP holding cost in our analysis. Accordingly in this case, we do not consider the tradeoff between WIP and production smoothness in setting the *SPLTs*, but how to set the planning windows and *SPLTs* given the *DLT* constraint.

In addition, we assume that the *SPLT* of the *Profile Shop* which produces the *Outfits* is fixed at 2 days. We observed that the *Profile Shop* has a steady production lead time. The processing and queuing time at the station is relatively short due to its simple processing operations and its considerable capacity slackness. Other tasks performed at the station, i.e. packing the completed parts, organizing them for specific panels and transporting them to the *Panel Shop*, take a relatively stable lead time that is reasonably independent of the workload.

5.5 DATA

To apply the model, we required extensive amount of data to parameterize the model inputs. In this section, we discuss the model inputs as well as how we obtained the relevant data. We define the discrete time period of our model to be one day in order to be consistent with the time buckets used in the current planning system. Hence we define the demand in terms of units per day and capacity in hours per day. The data presented in this chapter has been disguised to protect the company's confidential data but the insights drawn are identical to the conclusions based on the actual data.

- *Demand*: We obtained the mean and standard deviation of the internal demand for each product family from an eight-month demand record. We carried out regression analysis to determine whether there is any correlation between the families. We found a low correlation of 0.13 between the demands of *Big Panels* and the *Small Panels*. Furthermore, the correlations between the *Big Panels* and *Outfits*, as well as between the *Small Panels* and *Outfits* are 0.28 and 0.22

respectively. Given the low correlations, we assume that the demand of each product family is independent of each other.

- Effective Processing Times: We acquired the mean and variance of the effective processing times (in hours) at each workstation from data collected in a two-month period. We needed the mean processing times to define the workflow matrix and the variances to model the noise arrivals $Var(\xi_{it})$. The noise arrival term at each station equals the variance of the workload for the mean number of jobs processed at the station per day (equation (3.11)). We note that the stations from *Blast* to *Bevel* process steel plates while the rest of the stations process panels and outfits. Thus for the stations that process plates, we need to convert the time taken per plate to per panel and per outfit.
- Capacity: We observed the throughput rate in periods of high demand when most stations are operating at full capacity with many jobs being subcontracted out. We approximate the nominal capacity levels by the mean throughput (in hours) per day observed in these periods. The *Blasting Shop* and *NC Shop* consist of processing equipment and are thus machine-constrained. The *Panel Shop* operates with skilled workers and is therefore labor-constrained. Some workers are cross-trained to work at more than one station. Thus there is some flexibility in deploying more workers at stations with higher workload. But we assume that the nominal capacity level at each station is fixed.
- Subcontract Cost: The vendors quote the subcontracting costs in terms of cost per metric ton. Since workload in our model is measured in hours, we had to convert the subcontracting cost at each workstation into average cost per hour. We approximated this by taking into account the average weight of the jobs at the workstation (in metric tons) and the mean processing time of the jobs (in hours).

The (disguised) data is shown in Table 5.1, 5.2 and 5.3.

Table 5.1 Mean and standard deviation of demand

	Demand Per Day	
	Mean Number of Panels or Outfits	Standard Deviation of Panels or Outfits
<i>Big Panels</i>	2.67	1.91
<i>Small Panels</i>	2.05	1.32
<i>Outfits</i>	283	151

Table 5.2 Data for Model Inputs

STATION	Processing Time (hours)						Capacity (hours / day)	Subcontract Cost (\$ / hour)
	<i>Big Panels</i>		<i>Small Panels</i>		<i>Outfits</i>			
	μ	σ	μ	σ	μ	σ		
Blast	0.76	0.23	0.17	0.125	0.03	0.06	13	550
NC Gas Cut	3.87	2.33	0.86	1.35	0.17	0.60	28	420
NC Plasma Cut	-	-	-	-	0.13	0.46	33	520
Bridge	3.44	2.05	0.77	1.19	0.15	0.53	64	1200
Bevel	4.58	2.37	-	-	-	-	28	1100
Tack	4.70	3.50	-	-	-	-	33	1100
Join	40.9	31.1	-	-	-	-	120	130
Mark	2.94	2.05	2.94	2.05	-	-	32	1390
Outfitting (Big Panels)	32.7	32.1	-	-	-	-	96	150
Outfitting (Small Panels)	-	-	31.2	27.8	-	-	84	80

Table 5.3 Inputs for noise terms

STATION	Noise Arrivals, $Var(\xi_{it})$ (hour ²)		
	<i>Big Panels</i>	<i>Small Panels</i>	<i>Outfits</i>
Blast	0.14	0.03	1.02
NC Gas Cut	14.50	3.74	27.51
NC Plasma Cut	-	-	43.71
Bridge	11.22	2.90	79.50
Bevel	15.00	-	-
Tack	32.71	-	-
Join	2589	-	-
Mark	11.22	8.62	-
Outfitting (Big Panels)	2744	-	-
Outfitting (Small Panels)	-	1588	-

5.6 OPTIMIZATION

We build our model for a network of 3 dummy stations (one for each product family) and 10 workstations. In order to characterize the production random variables, we assume that the production requirements are normally distributed. To validate this assumption, we constructed normal probability plots for the daily production output from a two-month data and the plots showed that this assumption is reasonable (see *Appendix B*).

We now discuss the nonlinear optimization program for which our objective is to minimize the total expected subcontracting cost with the planning windows and *SPLTs* as decision variables.

$$\begin{aligned}
& \text{Min } \sum_i c_i E(P_{it} - M_i)^+ \\
& \text{s.t.} \\
& \sum_i \omega_{ki} n_i + W_k - 1 = L_k, \quad \forall k \\
& W_k, n_i \geq 1, \quad \forall i, k
\end{aligned}$$

The notations follow the definitions in *section 3.5*, which we restate here: c_i is the subcontracting cost per hour for workstation i , M_i is the nominal capacity for workstation i (hours per day), P_{it} is the production requirement at workstation i , ω_{ki} is the number of times a job from product family k visits workstation i , n_i is the *SPLT* of workstation i , W_k is the planning window for product family k and L_k is the *DLT* for product family k (which equals 26 for all k).

As described in *section 3.5*, the first set of constraints defines the relationship between the planning windows, *SPLTs* and the *DLT* (see (3.24) and (3.25)). The second set of constraints assures that the planning windows and *SPLTs* are at least the planning time bucket of one day, which is also the minimum time duration that the management perceives to be appropriate to keep up with the highly dynamic system. We do not restrict the decision variables to take on only integer values as the management finds non-integer values of planning windows and *SPLTs* acceptable for production control.

We have contended in *section 3.5* that the expected subcontracting cost $E[P_{it} - M_i]^+$ is convex with n_i , i.e. as we increase the *SPLT*, we have a decreasing rate at which the subcontracting cost declines. Therefore we expect the objective function to be convex. In addition, since the constraints are linear, the optimization program is convex. We solve the convex nonlinear program by using the optimization toolbox in *MATLAB* which employs the *Sequential Quadratic Programming* method (see Fletcher (1980) for an overview of the method). The computer runtime for this problem is less than 3 seconds on a Pentium PC with a 2.53 Giga-Hertz Intel processor.

5.7 RESULTS

In this section, we first discuss the key insights gained from the optimal solution. We then compare our optimal cost with other reasonable alternatives that have been suggested by the shop planners and managers.

Table 5.4 Optimal Solution

Station	Optimal w_i or n_i	Expected Subcontracting Work (hours)	Expected Subcontracting Cost (\$)
Dummy for Big Panels	1.0	-	-
Dummy for Small Panels	10.8	-	-
Dummy for Outfits	9.8	-	-
Blast	1.0	0.13	72.77
NC Gas Cut	4.4	0.37	155.06
NC Plasma Cut	4.4	0.12	60.48
Bridge	3.1	0.12	143.40
Bevel	1.0	<0.01	<0.01
Tack	1.0	<0.01	0.11
Join	7.8	5.96	775.35
Mark	1.0	<0.01	<0.01
Outfitting (Big Panels)	6.6	5.83	874.14
Outfitting (Small Panels)	6.6	4.39	351.34
Total (\$)			2,432.60

The optimal solution, as well as the corresponding expected subcontracting work and costs, are shown in Table 5.4. The optimal solution suggests that there should be a substantial amount of smoothing for the release of *Small Panels* and *Outfits*, with planning windows equal to 10.8 and 9.8 days respectively. Recall that jobs for these two families are currently released the same number of days before their due-dates as the *Big Panels*, even though they are processed by fewer stations. Our solution recommends that jobs for these product families should not be released immediately. Instead we should smooth the release so as to generate smoother production requirements at the workstations. However the planning window for *Big Panel* equals

to 1.0 which indicates that there should be no smoothing of its MPS, and more days should be allocated to the *SPLTs* to smooth production at the workstations.

We note that the optimal *SPLT* of the *Blast* station is 1.0 (binding to the second set of constraint) and is much shorter than the current *SPLT* of 5 days. A large proportion of its present *SPLT* acts as safety lead time to buffer against the uncertainties of unavailable steel plates and long picking times. The solution suggests that we should reduce the *SPLT* of *Blast* to 1.0 day. For the *Small Panels* and *Outfits*, the reduction of the *SPLT* is mainly reallocated to the planning windows. Here the planning windows would also act as the safety lead times besides smoothing the MPS. For the *Big Panels*, the *SPLT* of the *Blast* station is mainly reallocated to the other workstations.

At the *Join* station, the optimal solution for the *SPLT* is the longest among the workstations at 7.8 days. We observed that the utilization at the *Join* station is high at more than 90% and the variability of its effective processing time is also high with a coefficient of variation of 0.76. Hence the *Join* station requires a relatively longer *SPLT* to smooth its production. In addition, the optimal *SPLTs* of the *Outfitting (Big Panels)* and *Outfitting (Small Panels)* are also relatively long with each equals to 6.6 days. On the other hand, the *Bevel*, *Tack* and *Mark* stations have the shortest *SPLT* of 1.0 day due to their comparatively lower utilization. In fact, even at the *SPLT* of 1.0 day, we observed that the total expected subcontracting costs are significantly low.

The optimal total expected subcontracting cost is \$2,432.60. We attempted to quantify the potential cost savings that would result from these changes. We compared the optimal total expected subcontracting costs with the actual average total subcontracting cost. In view of the presence of noise in the actual data, we estimated that the recommendations would result in a 20% to 30% cost reduction.

In another effort to estimate the potential cost savings, we input the current *SPLTs* into our model and compared the resulting total expected subcontracting cost with our optimal cost. However presently, since the individual stations in the *Panel Shop* are not assigned the *SPLTs*, we were unable to compute the current subcontracting cost using our model. Instead we estimated the current cost by first setting the *SPLTs* of the stations in the *Panel Shop* proportional to the station's utilization rate while

satisfying the *Panel Shop*'s planned lead time of 13 days. The basis for such a method is that the more heavily loaded stations would require longer *SPLTs* to smooth production. However in this method, unlike our model, we note that we do not explicitly take into account the variability of workload as well as the subcontracting cost of the stations. Table 5.5 shows the comparison between the optimal solutions and our estimates of the current setting, which we termed as *Setting One*. The total expected subcontracting cost is \$3110.40. Hence we would reduce the cost by about 21.8% if we adopt the optimal solution. We regard this potential savings as a conservative estimate of the actual savings. This is because we assume that the production control within the *Panel Shop* in *Setting One* is according to the *SPLTs* of the individual stations, which would be better than the actual setting where the production control is according to the aggregate planned lead time of the shop.

Now we compare our optimal cost with that of another alternative suggested by the management. Here they suggested that the *SPLTs* of *all stations* are set proportional to each station's utilization rate while satisfying the *DLT* constraints. The rationale behind this alternative is identical to that for *Setting One*, i.e. the more heavily utilized stations would need longer *SPLTs*. Table 5.5 shows the *SPLTs* of this alternative setting (which we termed as *Setting Two*) and its associated expected subcontracting work and costs. The total expected subcontracting cost of *Setting Two* is \$2,842.05 which is 14.4% higher than that of the optimal cost. Thus we have shown that our model is a valuable tactical planning tool that would lead to substantial cost savings.

Table 5.5 Comparison of optimal solution with other settings

Station	<i>Optimal Solutions</i>			<i>Setting One</i>			<i>Setting Two</i>		
	Optimal w_i or n_i	Expected Subcont' Work (hours)	Expected Subcontract' Cost (\$)	w_i or n_i	Expected Subcont' Work (hours)	Expected Subcontract' Cost (\$)	w_i or n_i	Expected Subcont' Work (hours)	Expected Subcontract' Cost (\$)
Dummy for Big Panels	1.0	-	-	1.0	-	-	1.0	-	-
Dummy for Small Panels	10.8	-	-	1.0	-	-	1.0	-	-
Dummy for Outfits	9.8	-	-	1.0	-	-	1.0	-	-
Blast	1.0	0.13	72.77	5.0	0.14	79.48	4.1	0.18	101.09
NC Gas Cut	4.4	0.37	155.06	8.0	0.17	69.30	4.1	0.37	153.55
NC Plasma Cut	4.4	0.12	60.48	8.0	0.08	41.34	8.5	0.08	42.22
Bridge	3.1	0.12	143.40	2.8	0.11	137.76	3.8	0.10	115.56
Bevel	1.0	<0.01	<0.01	1.5	<0.01	<0.01	2.0	<0.01	<0.01
Tack	1.0	<0.01	0.11	1.3	<0.01	<0.01	1.8	<0.01	<0.01
Join	7.8	5.96	775.35	3.0	8.86	1151	4.2	7.59	986.69
Mark	1.0	<0.01	<0.01	1.4	<0.01	<0.01	2.0	<0.01	<0.01
Outfitting (Big Panels)	6.6	5.83	874.14	3.0	9.08	1362	4.1	7.58	1138
Outfitting (Small Panels)	6.6	4.39	351.34	8.8	3.37	269.53	7.6	3.82	305.30
Total (\$)			2,432.60			3,110.40			2,842.05

5.8 VALIDATION

We needed to develop confidence in the optimization results. We felt that the strongest assumption in our model for this case study is that the internal demands for the product families are stationary. We observed that the actual demands are generally non-stationary over time. However, the demands are stationary if we consider the time horizon to be divided into segments. Each segment in the time horizon consists of a constant number of projects in production, with each project in its stable project phase. We noticed that the demand is somewhat stationary within each segment. The demand process is stationary within each segment but the demand parameters vary from segment to segment. We illustrate the demand processes by showing the demand data (in units of panels) for the *Big Panels* and *Small Panels* over a three-month period in Figure 5.4 and Figure 5.5 respectively. As shown in both figures, we subdivide the time horizon into segments. In each segment, the demand is reasonably stationary as it fluctuates about a fairly fixed mean with roughly the same amount of fluctuation within the segment.

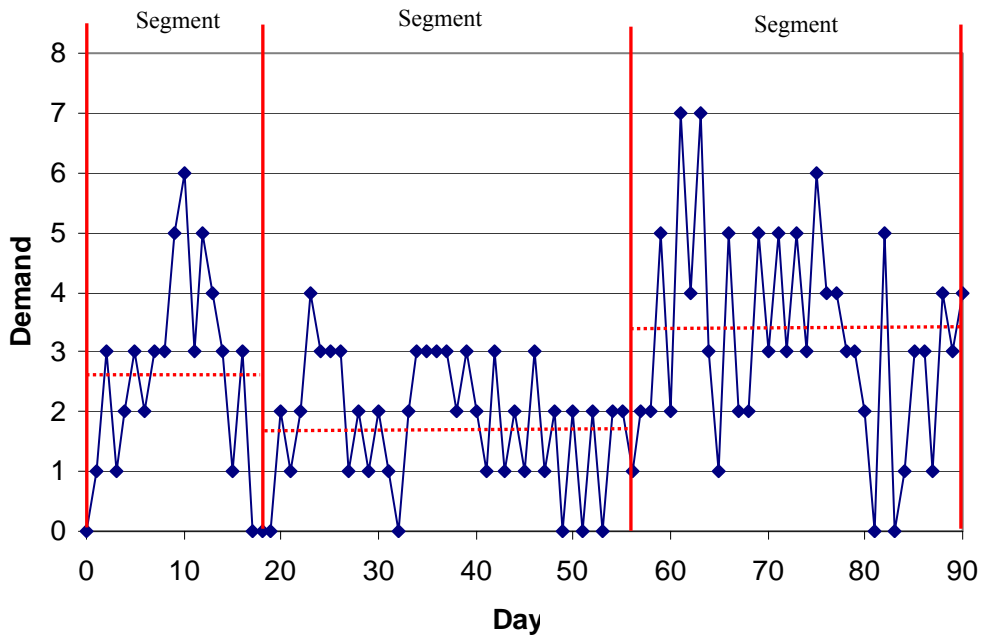


Figure 5.4 Demand for Big Panels over three-month period

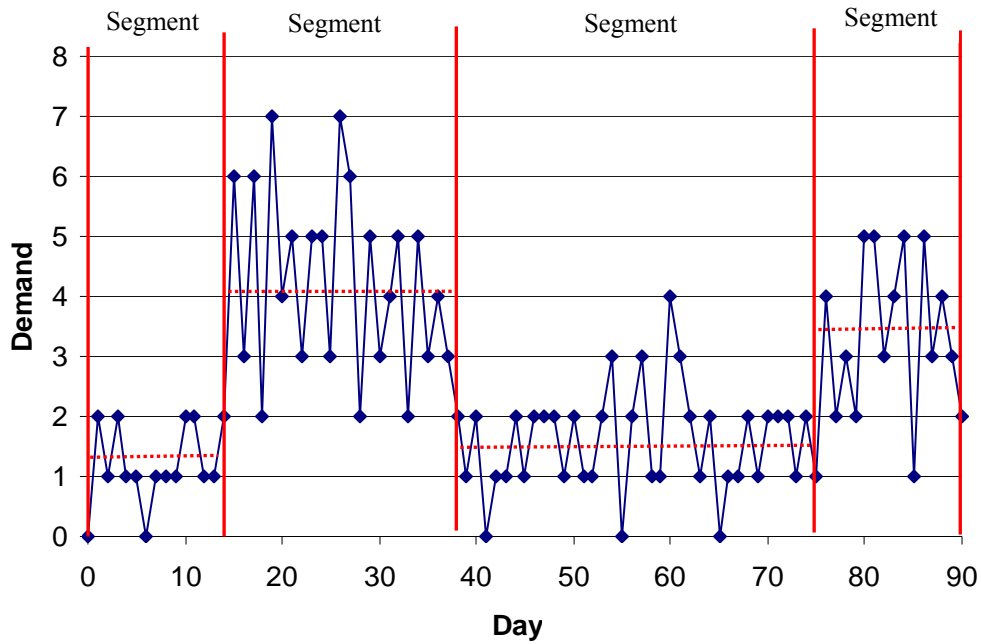


Figure 5.5 Demand for Small Panels over three-month period

Furthermore, the actual demand data shows that although the mean and variance of demand changes between the segments, the coefficient of variation (CV) remains fairly constant. The range of CV for the *Big Panels*, *Small Panels* and *Outfits* are 0.48 to 0.71, 0.55 to 0.79 and 0.38 to 0.59 respectively.

To validate our results against the stationary demand assumption, we test the sensitivity of our solution to both the mean and variance of demand. Here we are concerned with the higher mean demand levels as these would result in higher subcontracting cost. To simulate the demand processes in the different time segments, we subject each product family to either the original demand level or 125% of the original mean demand and at the same time, keep the coefficient of variation constant. Table 5.6 summarizes the results of the sensitivity analysis for the 8 combinations of demand levels, in which we compare the expected total subcontracting cost in using the optimal solution for the original demand with the optimized cost in each demand scenario.

Table 5.6 Summary of sensitivity analysis for demand

Station	Demand Scenario ⁺							
	Original	B	S	O	B, S	B, O	S,O	B,S,O
Dummy for Big Panels	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Dummy for Small Panels	10.8	10.1	10.3	9.6	9.8	9.6	9.4	9.4
Dummy for Outfits	9.8	9.1	9.3	8.6	8.8	8.6	8.4	8.4
Blast	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
NC Gas Cut	4.4	5.2	4.5	5.6	5.1	5.4	5.5	5.2
NC Plasma Cut	4.4	5.2	4.5	5.6	5.1	5.4	5.5	5.2
Bridge	3.1	3.8	3.3	5.7	4.1	5.7	5.8	5.9
Bevel	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Tack	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Join	7.8	7.1	7.3	6.6	6.8	6.6	6.4	6.4
Mark	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Outfitting (Big Panels)	6.6	5.8	6.9	4.1	6.0	4.3	4.3	4.5
Outfitting (Small Panels)	6.6	5.8	6.9	4.1	6.0	4.3	4.3	4.5
Optimal Cost (\$)		7,100	3,115	7,778	7,862	14,328	8,773	15,433
Cost of using optimal w_i and n_i for original demand (\$)		7,129	3,118	8,060	7,899	14,597	9,045	15,685
% Difference		0.41	0.08	3.50	0.47	1.84	3.01	1.61

+ We denote the product family at 125% of the original mean demand as B: *Big Panels*, S: *Small Panels* and O: *Outfits*. E.g. B, S implies both *Big Panels* and *Small Panels* are at 125% of the original mean demand (while *Outfits* is at the original mean demand).

Our analysis showed that the optimal solution is rather insensitive to all 8 cases of demand scenarios, with the percentage difference in subcontracting cost differ by less than 4% for all cases.

We faced difficulties in validating the predictive capability of our model. The most useful validation would be to compare the actual amount of work subcontracted out with that predicted by the model. However, the subcontracting decisions in the current *Panel Shop* are based on the aggregate planned lead time of the shop but in our

model, we consider the production control using the *SPLT* of the individual workstation. Thus we are unable to make such a comparison for the *Panel Shop* and unfortunately, this is where most of the workstations are located. That left us with only data from the *Blasting Shop* and *NC Shop* for a partial validation. We found that the amount of subcontracted work predicted by our model is about 26% and 21% less than the actual data for the *Blasting Shop* and *NC Shop* respectively. We consider this as a reasonable validation given the presence of high variability in the system; furthermore, the data for the subcontracted work contains some noise as we were unable to filter subcontracted jobs from other affiliated facilities.

We identified two weaknesses of our model from this case study. First, our model does not take into account the rush orders for urgent panels. These panels are crucial components for the downstream steel blocks and whose late completion would lead to serious delays in the steel block production. But in our model, we do not distinguish between normal and rush orders. Our model is unable to capture these orders because the workload for these orders often does not allowed for smoothing as they are processed immediately upon arrival. Furthermore, these rush orders are frequently subcontracted to expedite their completion. Each order is subcontracted at some selected stations depending on the order's processing requirement; therefore the rush orders do not have distinct flow paths but have numerous processing routes. As a result, with our existing shop data, we faced difficulties in accurately capturing the rush orders by defining them as separate product families. Second, we assume that the nominal capacity levels are fixed and we do not model the workforce flexibility that allows workers to be directed to more heavily loaded stations. Nevertheless, the management find that the collected data, model assumptions and results are reasonable and that our model is able to capture the core features of the problem.

5.9 DISCUSSIONS AND CONCLUSIONS

In this case study, we applied our model to improve the panel production by minimizing the total subcontracting cost. We carried out the project in stages: we first understood the scenario and its associated problems. Next, we determined whether our model fitted closely with the problem. We then collected and validated the

required data. Subsequently, we set up the optimization program to determine the optimal solution. Eventually, we made our recommendations upon validating the results.

However at the present, we are unable to implement our recommendations. This is because the production facility as well as its planning system is currently undergoing major changes. For instance, the company has recently purchased an automated welding machine for the *Join* station which will significantly change its capacity level. Furthermore the company is also undertaking various projects to improve its production efficiencies, e.g. to improve the steel plate availability and also to better coordinate the production schedules of the various oil-rig projects. As a result of these ongoing changes, our model output based on our collected data would have little predictive values. Once these changes are implemented and the system becomes steady again, we would collect new data to re-generate a set of recommendations.

Even though the recommendations are not yet implemented, this project has already brought several benefits to the company. It has led to a greater awareness of the importance of production smoothing within the organization. One outgrowth from this case study is a project that looks into the estimation of workload for every panel so as to better regulate the job release. Our project has also emphasized to the management the importance of steel plate availability and coordination between the production schedules for different oil-rigs. There are now projects carried out to induce these improvements. Furthermore, our model that we created in *MATLAB* provides a valuable tactical planning tool for the company, especially for “what-if” analysis, as it is straightforward to change the model parameters in the software.

CHAPTER 6

CONCLUSIONS AND DISCUSSION

6.1 CONTRIBUTIONS

In this final chapter, we review the contributions of this thesis and also present some ideas on the future research opportunities.

In this research, we consider production smoothing in a make-to-order (MTO) manufacturing environment that has a firm delivery lead time. Most previous research on production smoothing in the MTO environment is on the quoting of due-dates for a single-stage system, with the emphasis on managing customer demand. In contrast, our research studies production smoothing from the standpoint of production planning, subject to the constraint of a non-variable delivery lead time. In particular, we look into the setting of the planning windows and planned lead times, which are the two key tactical planning parameters for production smoothing. We explicitly model the planning process of generating the MPS and work releases from a stochastic demand process. Furthermore, we model the production system as a multi-station job shop, which is the most generic process structure. Therefore we believe our model has a broader applicability as compared to other existing models, which typically assume an exogenous job arrival process and/or are restricted to single-stage or serial-flow systems.

In *Chapter 2*, we extend the TPM of Graves (1986) to improve its applicability to production planning by overcoming its restriction of period sizing. To be specific, we extend the linear control rule in Graves (1986) into a continuous-time production function that allows jobs to complete processing at more than one station within a planning period. Thus we can set the planning period to be sufficiently long so as both to model the work flow as a continuous quantity, and to match the time buckets in most planning systems. In addition, the ability to set a longer time period improves the validity of the Markovian workflow assumption of the TPM. This extension also enables us to characterize the production random variable to be normally distributed

as we are able to set a long enough time period for the Central Limit Theorem to take effect. We believe that this extension greatly improves the applicability of the TPM and also strengthens the validity of the fundamental assumptions of the TPM.

In *Chapter 3*, we develop the tactical planning model for the purpose of production smoothing in a MTO environment. We analyze the roles of the planning windows and the planned lead times (in terms of *PPLT* and *SPLT*) for production smoothing. We provide insights on how to set these parameters with respect to the relative amount of demand variability and inherent workload variability at the workstations. We also examine the key considerations in prescribing these parameters, i.e. the WIP-smoothness tradeoff, the *DLT* constraint that limits the planning parameters, and the allocation of the planning window and *PPLT* given the *DLT*.

We then develop the model for the production of a single aggregate product. We model the smoothing of the MPS by introducing a dummy station to the production network. We characterize the production smoothing at the workstations where we employ the continuous-time control rule derived in *Chapter 2*. Subsequently we model the workflow in the job shop and then obtain the first two moments of production requirements as well as the expected queue lengths, which are the performance measures to assess the appropriate setting of the planning windows and *SPLTs*. We then extend the single-family model to the more generic model for systems with multiple families. We illustrate the use of the model in *Chapter 4*, and demonstrate how to set the planning windows and *SPLTs* appropriately.

In *Chapter 5*, we present a case study in which we apply our model to the manufacturing of steel panels for jack-up oil-rigs. We show that our model is capable of capturing the essence of this real-life scenario and is useful for studying the problems that arise in this context. We share our experience on how we perform the important tasks in this application, i.e. problem understanding, determining the appropriateness of model for the problem, data collection and validation, model validation and recommendations. This case study has led to a greater awareness of the importance of production smoothing within the organization; it has also helped to identify the important areas where efforts should be channeled to produce significant improvements in the production system.

6.2 DIRECTIONS FOR FUTURE RESEARCH

Now we discuss the research areas that we believe would improve and enrich the model.

6.2.1 Stability of Master Production Schedule

One important extension to the model would be to model the stability of the MPS when it is subject to changes in customer orders. In our model, we assume that there is no change to the order quantity between the time at which a customer order is received to the time at which the order is delivered on the due-date. However in many MTO firms, it is common for customers to adjust their order quantities. Very often, there exist rush orders, like some jobs in the panel production in our case study, which need to be rushed through the production system or subcontracted out to speed up its completion. Frequent changes in the MPS are costly as it affects the production smoothness of the production system. On the other hand, not responding to these changes results in poor customer service and a loss of goodwill. Thus it would be of interest to model the dynamic stability of the MPS and its impact on production smoothness, especially in light of the techniques that are used to achieve a stable MPS, e.g. firm planned orders, frozen time periods and time fencing (see Vollmann et al. (2005) for a discussion).

6.2.2 Limitations of TPM

Our model is based on the job shop model in Graves (1986) and accordingly, both have similar assumptions and limitations (see Graves (1986) for further discussion). One common limitation is the assumption that production is unconstrained; if production exceeds the nominal capacity, we assume that we have sufficient flexibility (by expediting actions, e.g. overtime or subcontracting) to complete the production within the planned lead time. This assumption is fairly reasonable for the case study on the panel production, as the facility does rely on a number of subcontracting options to handle its excess production requirements. However, in many instances, one might not have this flexibility and it would certainly be useful to look into ways to model the constrained production outputs. In capacity-constrained systems, highly fluctuating production requirements would lead to frequent late jobs.

It would also be worthwhile to study the effects of production smoothness on the tardiness of jobs in such systems.

Another common limitation is that both our model and the TPM assume that the amount of noise arrival is independent of both the work arrivals and production output. But one might expect that the noise amount should be positively correlated with the work arrivals and production output, i.e. greater production volume should lead to greater variability in the workflow. However, one can imagine that an extension to address this limitation might significantly increase the complexity of the model. Nevertheless, it would be of interest to look into ways to overcome this limitation without resorting to a much more complex model.

6.2.3 Assembly of MTO Parts

We have considered the MTO environment where jobs get processed as they progress through the workstations. However, there are some MTO products that require the assembly of two or more customized parts. We believe that it is useful to extend our model to include the assembly of MTO parts. To model such an assembly system, we would have to consider the coordination of the parts completion when setting the planned lead times. In addition, we would have to introduce more than one dummy station for each product family, as demand for a family can trigger the release of jobs for more than one part. Although we have some ideas for this extension, we need to further understand the practical operations of such systems before we can effectively explore how we can extend our existing model.

6.2.4 Non-Stationary Demand

We assume a stationary demand process in this research. However, a stationary demand process is not a good depiction of the demand for many products. For instance, some products have seasonal demand while most short life-cycle products experience different demand phase in its product lifespan. In our case study on the panel production, we are fortunate that our results are rather insensitive to the stationary demand assumption. However, this may not be true for other situations. Therefore it would be useful to model the non-stationary demand in some ways to widen the model's applicability.

APPENDIX A

Now we look at the conditions of convergence for the first two moments of the production vectors in (3.20) and (3.21). For notational convenience, we drop the subscript k for the product families.

For the moments to converge, we need the spectral radius of \mathbf{B} to be less than one, i.e. $\mathbf{B}^s \rightarrow \mathbf{0}$ as $s \rightarrow \infty$ where $\mathbf{0}$ is the matrix of zeroes. We now prove that this condition is satisfied if and only if the spectral radius of Φ is less than one.

Let $\rho(\cdot)$ denotes the spectral radius of a matrix. Suppose that $\rho(\mathbf{B}) \geq 1$, and we let λ and \mathbf{x} be the maximal absolute eigenvalue and corresponding eigenvector for \mathbf{B} . Assume that $\rho(\Phi) < 1$.

$$\begin{aligned} \mathbf{B}\mathbf{x} &= \lambda\mathbf{x} \\ \{\mathbf{I} - \mathbf{H}(\mathbf{I} - \Phi)\}\mathbf{x} &= \lambda\mathbf{x} \\ (\mathbf{I} - (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}(\mathbf{I} - \Phi))\mathbf{x} &= \lambda\mathbf{x} \\ \{\mathbf{I} - (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G} + (\mathbf{I} - \mathbf{F}\Phi)^{-1}\mathbf{G}\Phi\}\mathbf{x} &= \lambda\mathbf{x} \end{aligned}$$

Multiplying by $(\mathbf{I} - \mathbf{F}\Phi)$ throughout,

$$\begin{aligned} (\mathbf{I} - \mathbf{F}\Phi - \mathbf{G} + \mathbf{G}\Phi)\mathbf{x} &= \lambda(\mathbf{I} - \mathbf{F}\Phi)\mathbf{x} \\ \lambda\mathbf{F}\Phi - \mathbf{F}\Phi\mathbf{x} + \mathbf{G}\Phi\mathbf{x} &= \lambda\mathbf{x} - \mathbf{x} + \mathbf{G}\mathbf{x} \\ ((\lambda - 1)\mathbf{F} + \mathbf{G})\Phi\mathbf{x} &= ((\lambda - 1)\mathbf{I} + \mathbf{G})\mathbf{x} \\ \Phi\mathbf{x} &= ((\lambda - 1)\mathbf{F} + \mathbf{G})^{-1}((\lambda - 1)\mathbf{I} + \mathbf{G})\mathbf{x} \end{aligned} \tag{A1}$$

We note that all elements of matrix \mathbf{F} are less than or equal one. Furthermore, the terms $\{(\lambda - 1)\mathbf{I} + \mathbf{G}\}$ and $\{(\lambda - 1)\mathbf{F} + \mathbf{G}\}$ on the right-hand side of (A1) are diagonal matrices. Hence, for $\lambda \geq 1$, we have

$$\begin{aligned}
& \{(\lambda - 1)\mathbf{F} + \mathbf{G}\}\mathbf{x} \leq \{(\lambda - 1)\mathbf{I} + \mathbf{G}\}\mathbf{x} \\
& \{(\lambda - 1)\mathbf{F} + \mathbf{G}\}^{-1}\mathbf{x} \geq \{(\lambda - 1)\mathbf{I} + \mathbf{G}\}^{-1}\mathbf{x} \\
& \{(\lambda - 1)\mathbf{I} + \mathbf{G}\}\{(\lambda - 1)\mathbf{F} + \mathbf{G}\}^{-1}\mathbf{x} \geq \mathbf{x} \\
& \{(\lambda - 1)\mathbf{F} + \mathbf{G}\}^{-1}\{(\lambda - 1)\mathbf{I} + \mathbf{G}\}\mathbf{x} \geq \mathbf{x}
\end{aligned}$$

and so from (A1) we obtain

$$\mathbf{\Phi}\mathbf{x} \geq \mathbf{x}$$

This result contradicts to our assumption that $\rho(\mathbf{\Phi}) < 1$. Thus if $\rho(\mathbf{\Phi}) < 1$, we must have $\rho(\mathbf{B}) < 1$.

Now suppose that $\rho(\mathbf{\Phi}) \geq 1$ and let λ and \mathbf{x} be the maximal absolute eigenvalue and corresponding vector for $\mathbf{\Phi}$. But first we note that for $\lambda \geq 1$ and since \mathbf{G} is a positive matrix, we have

$$(\lambda - 1)\mathbf{G}\mathbf{x} \geq 0$$

By adding $(\mathbf{I} - \lambda\mathbf{F})\mathbf{x}$ to both sides of the above inequality, we obtain

$$\begin{aligned}
& (\mathbf{I} - \lambda\mathbf{F})\mathbf{x} + (\lambda - 1)\mathbf{G}\mathbf{x} \geq (\mathbf{I} - \lambda\mathbf{F})\mathbf{x} \\
& (\mathbf{I} - \lambda\mathbf{F} + \lambda\mathbf{G} - \mathbf{G})\mathbf{x} \geq (\mathbf{I} - \lambda\mathbf{F})\mathbf{x} \tag{A2}
\end{aligned}$$

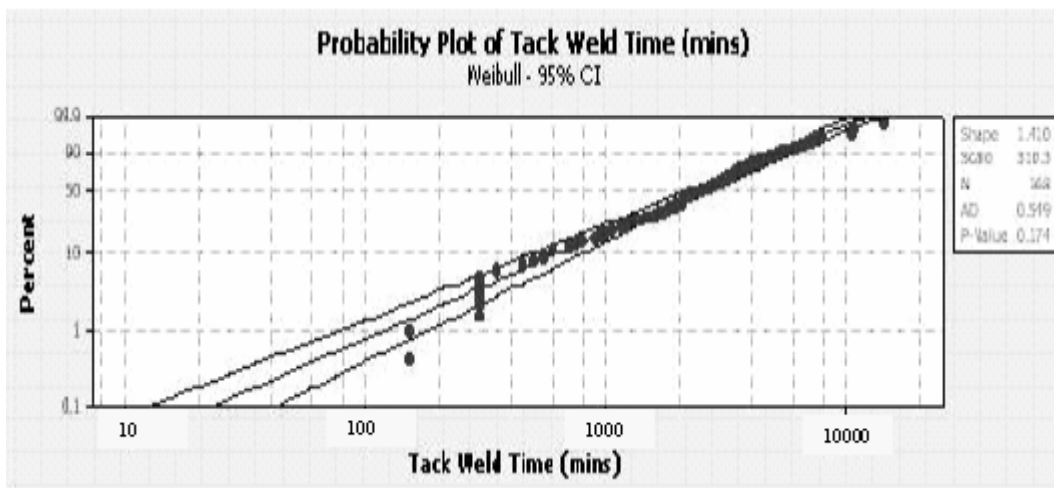
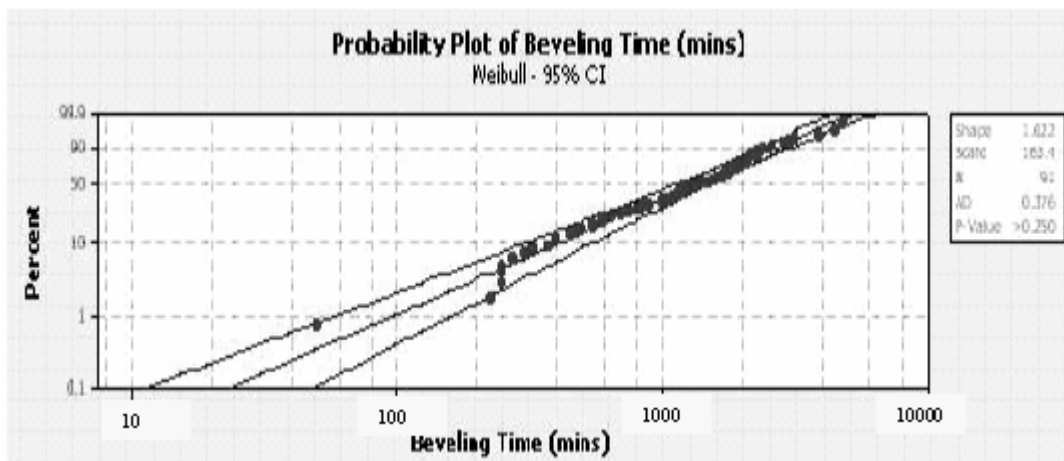
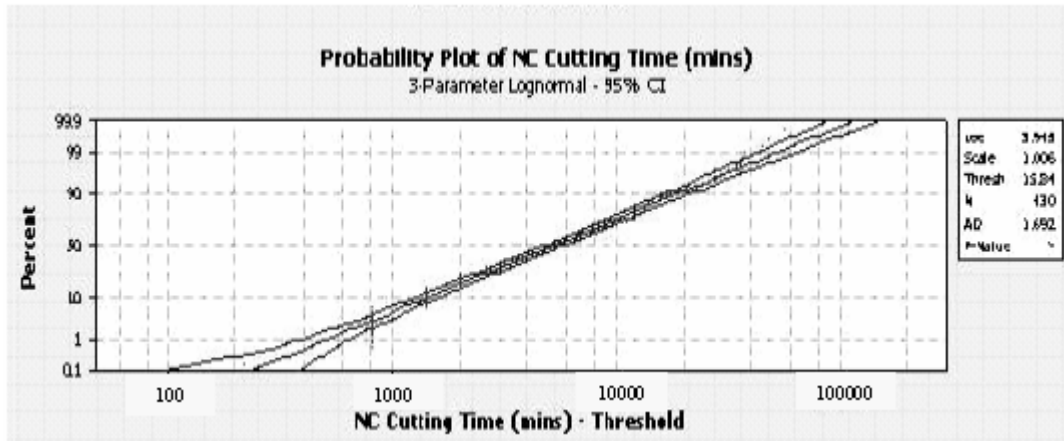
By substituting $\lambda\mathbf{x} = \mathbf{\Phi}\mathbf{x}$ into (A2), we find that

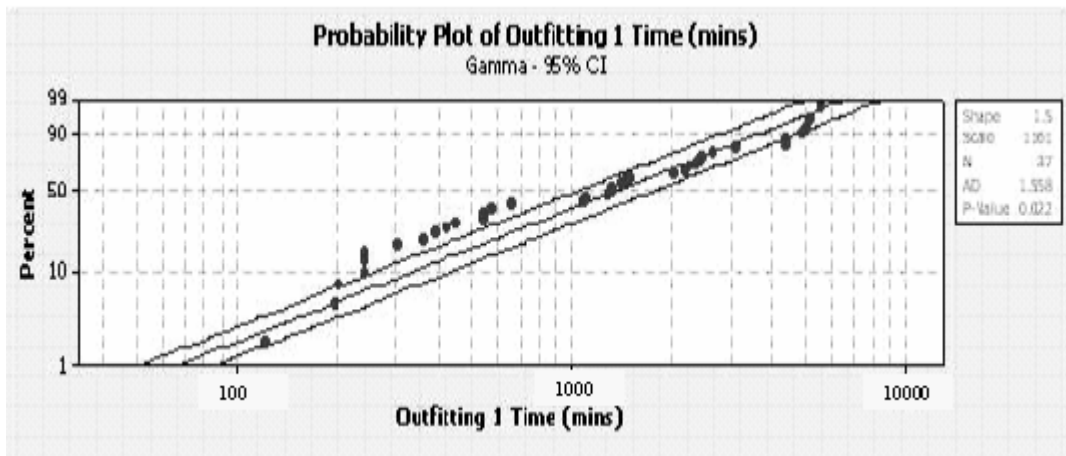
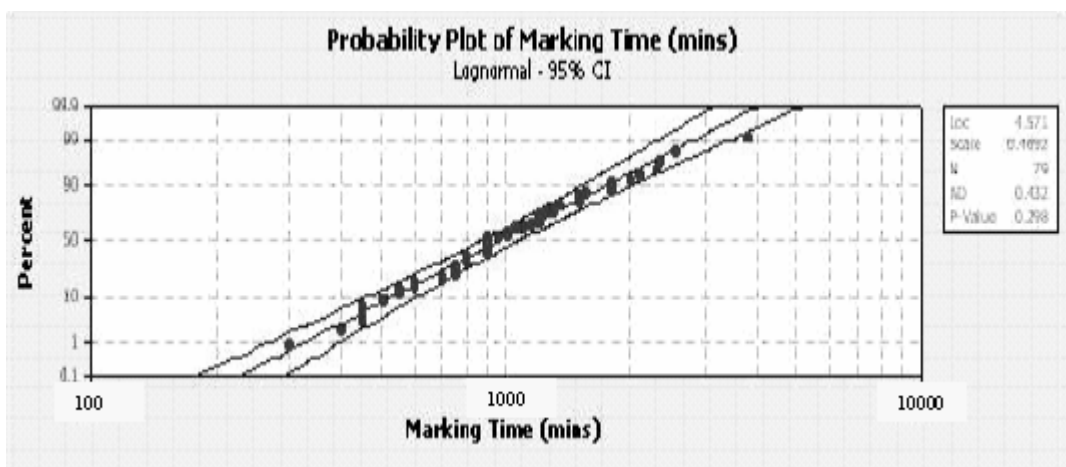
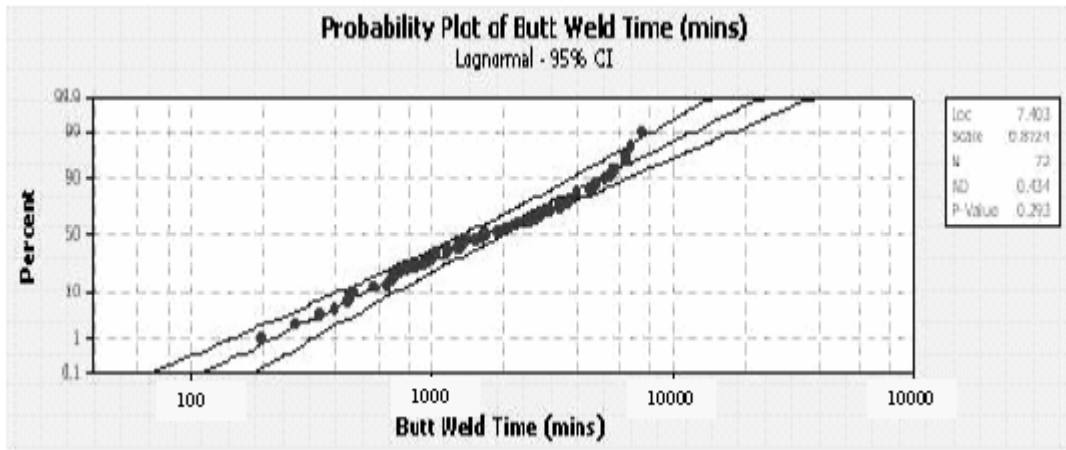
$$\begin{aligned}
& (\mathbf{I} - \mathbf{F}\mathbf{\Phi} + \mathbf{G}\mathbf{\Phi} - \mathbf{G})\mathbf{x} \geq (\mathbf{I} - \mathbf{F}\mathbf{\Phi})\mathbf{x} \\
& (\mathbf{I} - \mathbf{F}\mathbf{\Phi})^{-1}(\mathbf{I} - \mathbf{F}\mathbf{\Phi} + \mathbf{G}\mathbf{\Phi} - \mathbf{G})\mathbf{x} \geq \mathbf{x} \\
& \left(\mathbf{I} - (\mathbf{I} - \mathbf{F}\mathbf{\Phi})^{-1}\mathbf{G}(\mathbf{I} - \mathbf{\Phi})\right) \geq \mathbf{x} \\
& \mathbf{B}\mathbf{x} \geq \mathbf{x}
\end{aligned}$$

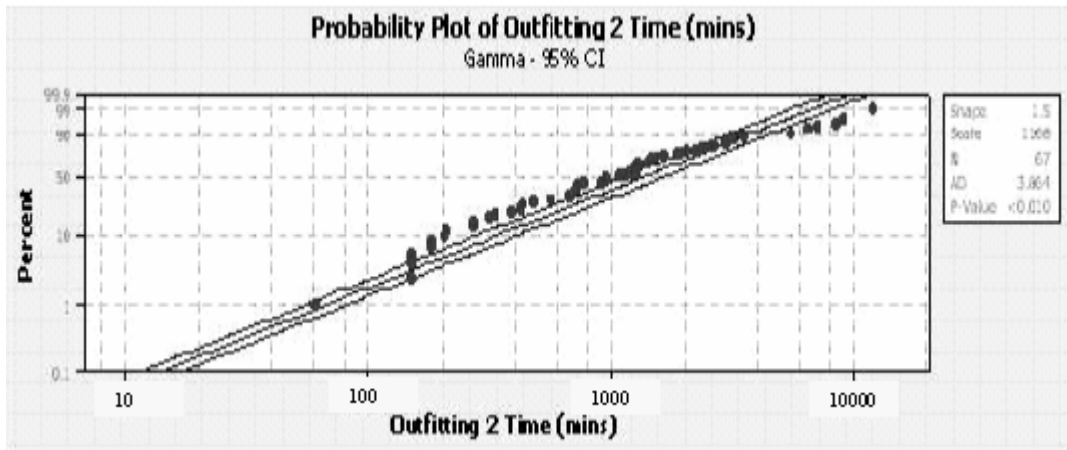
Therefore we have $\mathbf{B}\mathbf{x} \geq \mathbf{x}$ if $\lambda \geq 1$. Thus if $\rho(\mathbf{B}) \geq 1$, we must have $\rho(\mathbf{\Phi}) \geq 1$. Hence we have proven that (3.20) and (3.21) converges if and only if $\rho(\mathbf{\Phi}) < 1$.

APPENDIX B

We validate that the production requirements at the stations in our case study in *Chapter 5* to be normally distributed. We construct the following probability plots for the daily total production output in minutes.







REFERENCES

- Adenso-Diaz. B. and M. Laguna. 1996. Modeling the Load Leveling Problem in Master Production Scheduling for MRP Systems. *Int. J. Prod. Res.*, **34**, 483 - 493.
- Baker, K. R. and J. W. M. Bertrand. 1981. A Comparison of Due-Date Selection Rules. *AIIE Trans.*, **13**, 123 – 131.
- Baker, K. R. 1984. Sequencing Rules and Due-Date Assignments in a Job Shop. *Management Sci.*, **30**, 1093 – 1104.
- Bertrand J., W. M. 1983. The Effect of Workload Dependent Due-Dates on Job Shop Performance. *Management Sci.*, **29**, 799 - 816.
- Buzacott, J. A. and J. G. Shanthikumar. 1994. Safety Stock versus Safety Time in MRP Controlled Production Systems. *Management Sci.*, **40**, 12, 1678-1689.
- Chand, S., D. Chhajed. 1992. A Single-Machine Model for Determination of Optimal Due Dates and Sequence. *Oper. Res.*, **40**, 596 – 602.
- Cruickshanks, A. B., R. D. Drescher and S. C. Graves. 1984. A Study of Production Smoothing in a Job Shop Environment. *Management Sci.*, **30**, 36-42.
- Duenyas, I., W. J. Hopp. 1995. Quoting customer lead times. *Management Sci.*, **41**, 43 – 57.
- Duenyas, I. 1995. Single facility due date setting with multiple customer classes. *Management Sci.* **41**, 608 - 619.
- Enns. S. T. 2001. MRP Performance Effects Due to Lot Size and Planned Lead Time Settings. *Int. J. Prod. Res.*, **39**, 3, 461 - 480.

- Fine, C. H. and S. C. Graves. 1989. A Tactical Planning Model for Manufacturing Subcomponents in Mainframe Computers. *J. Manuf. and Opns. Mgmt.*, **2**, 1, pp 4-34.
- Fletcher, R. 1980. *Practical Methods of Optimization: Vol. 2 - Constrained Optimization*, John Wiley and Sons.
- Gong, L., T. d. Kok., and J. Ding. 1995. Optimal Leadtimes Planning in a Serial Production System. *Management Sci.*, **40**, 5, 629-632.
- Grasso, E. T. and Taylor, B. W. 1984. A Simulation-Based Investigation of Supply/Timing Uncertainty in MRP Systems. *Int. J. Prod. Res.*, **22**, 3, 485-497.
- Graves, S. C. 1986. A Tactical Planning Model for a Job Shop. *Oper. Res.*, **34**, 4, pp 522-533.
- Graves, S. C. 1988a. Safety Stocks in Manufacturing Systems. *J. Manuf. and Opns. Mgmt.* **1**, 1, 67-101.
- Graves, S. C. 1988b. Determining the Spares and Staffing Levels for a Repair Depot. *J. Manuf. and Opns. Mgmt.* **1**, 2, 227-241.
- Graves, S. C. 1988c. Extensions to a Tactical Planning Model for a Job Shop. *Proceedings of the 27th IEEE Conference on Decision and Control*, Austin, Texas.
- Graves, S. C., D. B. Kletter, and W. B. Hetzel. 1998. A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization. *Oper. Res.*, **46**, 3, 35 – 49.
- Graves, S. C. and J. S. Hollywood. 2001, revised March 2004. A Constant-Inventory Tactical Planning Model for a Job Shop. *Working paper*, 35 pp.
- Hollywood, J. S.. 2001. An Approximate Planning Model for a Job Shop with Nonlinear Production Rules. Working Paper, 32 pp.

- Hopp, W. J., M. L. Spearman. 1996. *Factory Physics: Foundations of Manufacturing Management*, 2nd edn, McGraw-Hill.
- Kanet, J. J. 1986. Toward a better understanding of lead times in MRP systems. *J. of Oper. Mgt.*, **6**, 3, 305 – 315.
- Keskinocak, P., R. Ravi, S. Tayur. 2001. Scheduling and reliable lead time quotation for orders with availability intervals and lead time sensitive revenues. *Management Sci.*, **47**, 264–279.
- Leong, T. Y. 1987. A Tactical Planning Model for a mixed push and pull system. *Ph.D. program second year paper*, Sloan School of Management, MIT, Cambridge MA.
- Marlin, P. G. 1986. Manufacturing lead time accuracy. *J. of Oper. Mgt.*, **6**, 2, 179 – 202.
- Melnyk, S. A., and Piper, C. J. 1985. Lead time errors in MRP: The lot-sizing effect. *Int. J. of Prod. Res.*, **23**, 2, 253 – 264.
- Mihara, S. 1988. A Tactical Planning Model for a job shop with unreliable work stations and capacity constraints, *S.M. Thesis, Operations Research Center*, MIT, Cambridge MA.
- Mohan, R. P. and L. P. Ritzman. 1998. Planned Lead Times in Multistage Systems. *Decision Sci.*, **29**, 1, 163 - 191.
- Parrish, S. H. 1987. Extensions to a model for tactical planning in a job shop Environment. *S.M. Thesis, Operations Research Center*, MIT, Cambridge MA.
- Penlesky, R. J., Berry, W. L., and Wemmerlöv, U. 1989. Open order due date maintenance in MRP systems., *Mgt Sci.*, **35**, 5, 571 – 584.

- Plambeck, E. L. 2004. Optimal Leadtime Differentiation via Diffusion Approximations. *Oper. Res.* **52**, 213 - 228.
- Seidmann, A., and M. L. Smith. 1981. Due date assignment for production systems. *Management Sci.*, **27**, 571 – 581.
- So, K. C., J.-S. Song. 1998. Price, delivery time guarantees and capacity selection. *Eur. J. Oper. Res.*, **111**, 28 – 49.
- Stalk, G. Jr., T.M. Hout. 1990. *Competing Against Time: How Time-Based Competition is Reshaping Global Markets*. Free Press, New York.
- Thomas, P. R. 1990. *Competitiveness Through Total Cycle Time: An Overview for CEOs*, McGraw-Hill, NY.
- Vollmann, T. E., W. L. Berry, D. C. Whybark, and F. R. Jacobs. 2005. *Manufacturing Planning and Control for Supply Chain Management*, 5th edn, McGraw-Hill.
- Weeks, J. K. 1978. Optimizing planned lead times and delivery date. *Proceedings of the 21st American Production and Inventory Control Society Annual Meeting*, 177 – 188.
- Wein, L. M. 1991. Due-Date Setting and Priority Sequencing in a Multiclass M/G/1 Queue. *Management Sci.*, **37**, 834 – 850.
- Whybark, D. C. and J. G. Williams. 1976. Material Requirement Planning under Uncertainty. *Decision Sci.*, **7**, 4, 595 – 606.
- Yano, C. A. 1987. Setting Planning Leadtimes in Serial Production Systems with Earliness Costs. *Management Sci.*, **33**, 1, 95-106.