STAGGERED DELIVERIES IN PRODUCTION AND INVENTORY CONTROL

BY

Carl Philip T. Hedenstierna

A Thesis Submitted in Fulfilment of the Requirements for the Degree of Doctor of Philosophy of Cardiff University

Logistics Systems Dynamics Group, Logistics and Operations Management Section of Cardiff Business School, Cardiff University

February 2016

Acknowledgements

My thanks go to the faculty, staff, and Ph.D. students in the Logistics and Operations Management section at Cardiff Business School. In particular, I am grateful to

- My primary supervisor, Prof. Stephen Disney, who showed me that mathematical modelling is not mere technique, but an art that reveals the governing dynamics of the world.
- My secondary supervisor, Prof. Mohamed Naim, who provided much needed encouragement, and brought new perspectives on my research.
- The late Prof. Denis Towill, who taught me that good work must have both rigour and practical relevance.
- The examiners, Prof. Mohamed Zied Babai and Dr. Andrew Potter.

The phenomenon of staggered deliveries was brought to my attention by Atul Agarwal and Andy Birtwistle, with further insights provided by Dr. Leif Pehrsson. I hope that this research leads to a deeper understanding of the industrial reality they have described. Additional thanks go to Prof. Janet Godsell and Dr. Per Hilletofth. Finally, I wish to thank my family and Songmei Dong for their encouragement and support.

Abstract

This thesis investigates production-inventory systems where replenishments are received every period (for example every day or shift), but where production plans are determined less frequently (weekly, fortnightly, or monthly). Such systems are said to use *staggered deliveries*. This practice is common in industry, but the theoretical knowledge is limited to a small set of inventory models, none of which include capacity costs. This thesis uses time series analysis to expand our understanding of staggered deliveries from the perspectives of inventory and production-inventory control.

The contribution to inventory theory consists in the development of an optimal policy for autocorrelated demand and linear inventory costs, including exact expressions for costs, availability, and fill rate. In addition the thesis identifies a procedure for finding the optimal order cycle length, when a once-per-cycle audit cost is present. Notably, constant safety stocks are suboptimal, and cause both availability and fill rate to fluctuate over the cycle. Instead, the safety stocks should vary over time, causing the availability, but not the fill rate, to be constant.

The contribution to production-inventory theory comes from two perspectives: First, an optimal policy is derived for quadratic inventory and capacity costs; second, four pragmatic policies are tested, each affording a different approach to production smoothing and the allocation of overtime work (once per cycle, or an equal amount of overtime every period). Assuming independent and identically distributed demand, these models reveal that all overtime or idling should be allocated to the first period of each cycle. Furthermore, it is shown that the order cycle length provides a crude production smoothing mechanism. Should a company with long reorder cycles decide to plan more often, the capacity costs may increase. Therefore, supply chains should implement a replenishment policy capable of production smoothing before the order cycle length is reduced.

Contents

Chapter 1

Introduction

This thesis is concerned with production planning and inventory management, and seeks to find policies (decision rules) that result in low inventory and capacity costs. The particular topic of this thesis is the situation that appears when production plans are made on a weekly or monthly basis, but when production takes place every day. Such production systems are said to use *staggered deliveries*.

This chapter highlights the related concepts of order cycles and staggered deliveries (first used in this context by Flynn and Garstka, 1990). Their interaction with familiar production and inventory control (PIC) concepts is explained. Research questions are identified to provide a clear focus. Then, a path to the resolution of these questions is provided in an outline of this thesis. The main contributions to knowledge are also presented.

1.1 Background

In 1924, General Motors (GM) switched from making new production plans once every three months, to making them once every ten days. This was not the only supply chain modification made by GM in the 1920's, but it was perhaps the most important. By the end of the decade, GM's total inventory turnover had increased from about two, to seven and a half inventory turns per annum (Sloan, 1963, pp. 129–139).

Toyota also considered short order cycles as desirable. In the 1980's they operated with a ten-day order cycle, and considered reducing it to weekly or daily cycles (Shingo, 1989, p. 129). Ohno (1988, p. 51) illustrates that long reorder cycles may be problematic. First, the inventory level can drift away from its intended value, as the fixed production plan ignores recent changes

in demand. This inventory discrepancy will need to be corrected, requiring significant overtime work as the production plan is updated.

Order cycles are also central to Period Batch Control (PBC), a production concept used in the manufacture of the Spitfire aircraft. For PBC, Burbidge (1989, p. 159) states that the order cycle should be as short as capacity permits. Burbidge (1983) even includes this as one of his five golden rules to avoid bankruptcy, and states that weekly or biweekly order cycles are preferable to monthly cycles, if such short cycles can be implemented.

To this day, there is no definitive solution to selecting the best order cycle length. Informal enquiries with companies reveal that actual order cycle lengths span from a single shift, up to a month (Table 1.1). Similar results were obtained in a questionnaire survey of 292 Swedish companies; of those that used periodic review, 21% planned on a daily basis, 37% planned on a weekly basis, and 42% planned fortnightly or less frequently (Jonsson and Mattsson, 2013). We may speculate about the reasons for selecting a particular order cycle length. On the one hand, the order cycle should be short enough for us to respond to demand in a timely manner. But with a short order cycle follows the burden and the potential cost of releasing new plans frequently, not to forget the consequences on capacity utilization when production must respond to short-term demand fluctuations. On this basis, we seem to be dealing with a trade-off problem where demand uncertainty is involved.

Uncertainty in both demand and supply drives companies to use safety stocks, safety time, and safety capacity as buffers, to safeguard against lost sales. Nonetheless, buffering is expensive, and companies must not only identify the right amount of buffering, but also find ways to reduce the need for it. In

Table 1.1: Observed industrial planning cycles. (Hedenstierna and Disney, 2014)

Company	Industry	Location	Order cycle	When
Tesco	Retail	UK	Shift or daily	2005
Anonymous	Electronics	USA	Weekly	2010
Harman Kardon	Audio equipment	UK	Weekly	2001
P&G	Household goods	Worldwide	Weekly	2000
Princes	Beverages	UK	Weekly	2003
Anonymous	Coinage	Western Europe	Weekly	2015
TRW	Automotive	Worldwide	Weekly	1999
BAT[a]	Consumer goods	Worldwide	Monthly	2012
Renishaw	Measuring equip.	UK, India	Monthly	2014

[a] Reported in Hedenstierna (2009).

principle, this can be done by finding variables that drive cost, and then by reducing the variability they exhibit (Simchi-Levi, 2002). Many supply chain concepts work this way, including postponement (Pagh and Cooper, 1998), consolidation of stock-keeping units (Lee et al., 1993), warehouse centralization (Maister, 1976), vendor-managed inventory (Disney and Towill, 2002), capacity consolidation and production rate levelling (Hedenstierna and Disney, 2012). These concepts can be used either to reduce the total variability in the system, or to shift variability from one type of buffer to another (Burbidge, 1961).

Inventory centralization is an example of the first type, where the total variability is reduced according to the square root law for inventories (Maister, 1976), while the latter can be demonstrated with production smoothing, as levelling the production rate tends to decrease its variability, while increasing the variability of the finished goods inventory (John et al., 1994).

Before delving into the complexities of the order cycle in production and inventory systems, we require a firm understanding of both inventories, production, and the concept of staggered deliveries.

1.1.1 Inventory systems

Inventories store products for future consumption, often as a protective measure against supply or demand uncertainty. In itself, the term *inventory* refers to an accumulation of goods, whose stored quantity (the *inventory level*) increases with *receipts*, and decreases with *consumption*.

Inventories may appear straightforward, but once we take into account how inventory levels are measured, and how receipts are regulated, a more complicated picture emerges. Several aspects come into play:

- The *inspection period* provides the time between measurement of inventory levels. Delivery performance and inventory costs are also computed at this interval. This thesis assumes an inspection frequency of unity, so it takes place for every discrete time point $t \in \mathbb{Z}$. Periods with a negative time index are feasible, in the sense that the results are unaffected by the sign of the period. In practical terms, the timescale can be defined so that periods with a negative time index refer to periods that occurred before the system was operational, should we require information about past demand and other system states.
- The *order cycle length,* is a strictly positive integer, $P \in \mathbb{Z}^+$, that provides the time between occasions when order quantities are determined. We

assume a constant order cycle length.

- The *lot size* provides the number of units produced or procured in a single lot. It may vary over time. A lot with constant size is referred to as a *batch.* Small production lots are often considered as desirable, as they lead to the concept of a one-piece-flow.
- The *lot frequency* provides the number of lots produced per order cycle. Except for a brief discussion on the consequences of the lot size, this thesis assumes that at least one lot may be completed between two subsequent inventory inspections (P lots per order cycle).
- The *lead time,* $L \in \mathbb{Z}^*$, is a nonnegative integer that provides the time between the issuance of an order, and the receipt of the first lot into inventory. Taking into account the sequence-of-events delay that is inherent in discrete-time systems, a staggered order placed at time t results in the first lot being received and accounted for at time $t + L + 1$. The last lot will have been received and registered at time $t + L + P$. The lead time L may be constant or variable — In this thesis we assume it is constant.

It is often assumed that the lot frequency coincides with the inspection period, but this is not necessarily true. In Just-in-Time (JIT) systems, it is common to aim for small batch quantities in production, but this does not mean that inventory levels are tallied more than once, or perhaps a few times, per day.

In the same way, the order cycle length need not match the inspection period. We can choose to make a production plan only once per week, and at this point in time determine the orders to be produced in every single day throughout the week. Such a set-up is termed *staggered deliveries.* An illustration of this is presented in Figure 1.1; where orders are placed every seven periods ($P = 7$), as indicated by the tick marks.

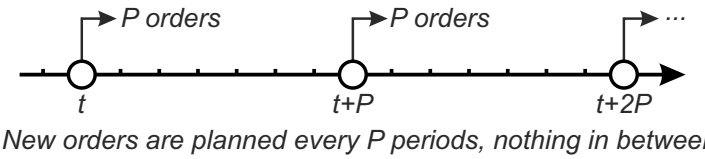

Figure 1.1: Illustration of a planning cycle with length P.

Inventory inspection

Inventory inspection refers to recording inventory levels, and the *inspection period* refers to the time between inspections. We inspect for three reasons:

1. To determine the timing or quantity of replenishment orders.
2. To calculate delivery performance, including availability, fill rate, and ready rate.
3. To calculate inventory costs that relate to the inventory level.

A single inventory system can use different inspection periods for each point above. This thesis assumes that inspections to calculate service and cost performance (2 and 3) occur at the same time. We use the term inspection period to denote the duration between two such events. The term *order cycle* refers to the duration between ordering events (1). The inspection period reveals little about an inventory system, unless one also knows the frequency of receipts. There are three cases of particular interest, illustrated in Figure 1.2.

The first case is when inventory inspections are more frequent than receipts. Then the measured inventory level will include some amount of cycle stock. A famous example is the classical economic order quantity (EOQ) model, where inspection is continuous (the inspection period approaches zero), but where receipts are discrete lots (Harris, 1990). The second case is when inventory inspections occur with the same frequency as receipts. Now we no longer measure cycle stock, but only the end-of-period inventory. It can be applied in continuous time (John et al., 1994) and in discrete-time (Dejonckheere et al.,

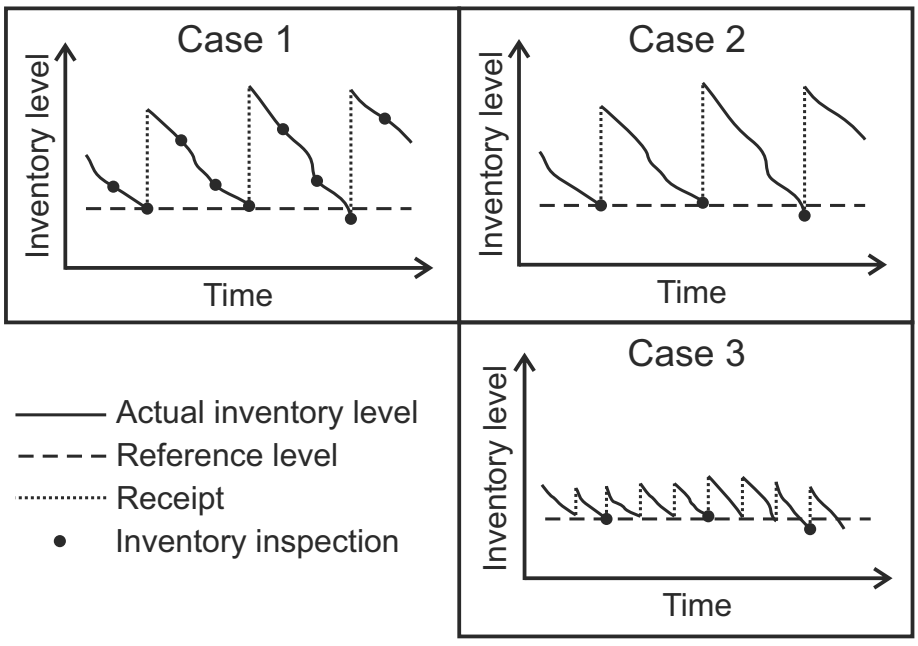

Figure 1.2: Relation between the lot frequency and the inspection interval.

2003). The third case has inventory inspections occurring less frequently than receipts. The measured inventory levels are identical to the second case if we do not account for temporary shortages that are filled in the same period as they occur. A JIT system with reasonably small batch sizes will be of this type.

This thesis investigates case 2, and by extension case 3, if service levels and inventory costs are applied only when the inventory level is inspected. In an industrial setting, we expect to see these cases realized for high-volume products that are produced almost every day. Such products tend to be called *runners* in the runners-repeaters-strangers classification (Slack, 2015).

1.1.2 Production and inventory control with staggered deliveries

Definition 1.1. The term *staggered deliveries* refers to production and inventory systems where the inventory receipts for multiple, subsequent inspection periods are determined at a single point in time (Flynn and Garstka, 1990).

Example 1.2 (Staggered deliveries)**.**

(a) Inventory is monitored at 18:00 every day, including weekends.

(b) Production plans for the following week (Monday to Sunday: 7 days, 21 available shifts) are made every Friday night.

(c) Output can be generated every day.

In this example, the inspection period is one day. The production plan determines all the orders to be produced from Monday morning to the following Sunday night. This cycle spans seven inventory inspections. Since we determine the production quantity, and hence the receipt quantity, for multiple inspection periods at once, we use staggered deliveries in our planning.

Ordering rules and costs under staggered deliveries

When staggering deliveries, we must consider any constraints on the size of orders from period to period. With linear holding and backlog costs, it is known that a staggered order-up-to (OUT) policy is optimal, where the receipt quantity in each period can be set freely (type 1 in Figure 1.3). There may also be other costs to consider. One of these may be nonlinear capacity costs, where a fixed rate of production commences at a low unit cost, while production

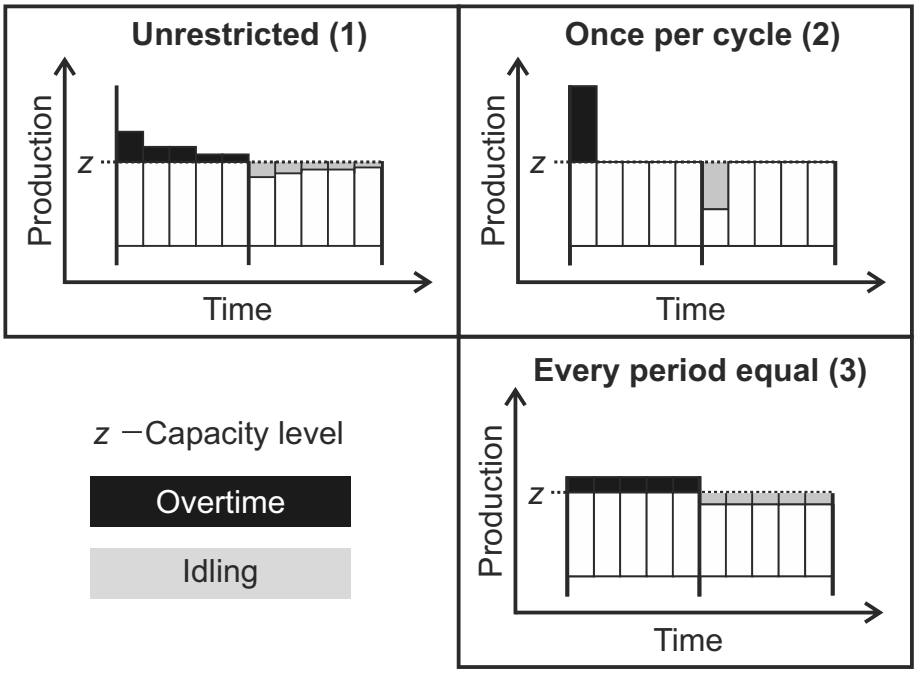

Figure 1.3: Approaches to restricting overtime work.

above this fixed quantity costs extra. This is the case with emergency orders (Rosenshine and Obee, 1976) or overtime costs (Holt et al., 1960). Another interesting aspect is that of allowing a single or only a few periods in which overtime can be worked (Chiang, 2009; Figure 1.3, type 2), or if overtime is distributed equally among all periods (type 3).

1.1.3 Performance: Costs and service levels

The concepts discussed earlier need some basis for comparison. We shall consider the cost of using any one policy, and the resulting service levels.

Cost types

Inventory costs are commonly assumed to be linear functions of on-hand inventory, the *holding cost*, and of backorders, the *backorder cost* (Axsäter, 2006, pp. 44–45). The usual justification for holding costs is that tied-up capital can be employed elsewhere, but it can also include storage costs, shrinkage, and insurance, if these costs are proportional to the on-hand inventory. Backorder costs are less intuitive. In systems where every backorder is a lost sale, the backorder cost is at least as large as the gross margin. If customers disfavour unreliable suppliers, the cost may be greater (Cachon and Terwiesch, 2009, p. 304). As it is difficult to quantify, the backorder cost can also be a numerical

representation of managers' aversion to backorders.

Payroll and overtime costs are of particular interest for systems with staggered deliveries, as there are different ways to manage overtime and idling (Figure 1.3). As an hour of overtime work tends to cost more than an hour of work during the working week, it is common to assume a piecewise-linear cost function (Holt et al., 1960, p. 54). Then, per period, there is a constant marginal cost of production up to the regular capacity limit, after which overtime work comes into effect. Passing this limit increases the marginal cost of production, as illustrated by the solid line in Figure 1.4.

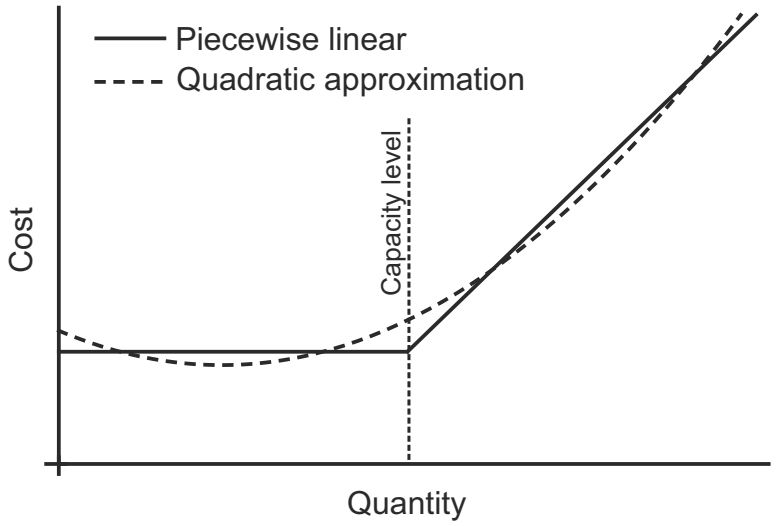

Figure 1.4: Piecewise-linear capacity costs and a quadratic approximation.

The inventory and capacity cost models are both piecewise linear. These costs can be approximated by quadratic cost functions (Holt et al., 1960, pp. 52–60), as this simplifies the analytical treatment. Furthermore, minimization of the quadratic cost is equivalent to the minimization of the variance of a system's output variables. Quadratic cost models are therefore helpful when we seek to model demand amplification (bullwhip), as this is proportional to the variance of orders (Lee et al., 1997). In Figure 1.4, the dotted line represents a quadratic function that estimates the piecewise-linear function. For details about estimating quadratic costs, see Holt et al. (1960).

The cost of making a new production plan / order decision is called the *audit cost*. It is incurred once per order cycle (Flynn and Garstka, 1990). Other setup costs are ignored, including a schedule change cost (Tang and Grubbström, 2002), changeover costs in production, and other specific costs. This delimitation is intentional, for it allows us to focus on the trade-off between inventory costs and capacity or audit costs.

Service levels

The backorder cost is one way to express the undesirability of shortages, but there is a more direct approach via service levels, which measure the ability to satisfy demand. For a single product, common service measures are the availability (S_1), also called cycle service level, which is the probability of not encountering a stockout in any period; the fill rate (S_2), which is the long run fraction of demand that can be satisfied immediately; and the ready rate (S_3), which is the fraction of time when the on-hand inventory is positive (Axsäter, 2006, p. 94). The characteristics and merits of these measures are discussued in Chapter 2.

1.2 Problem definition

This thesis investigates staggered deliveries and their consequences. We have already seen that the relation between the inspection interval and the lot size matters, and that cost assumptions can involve overtime work or emergency orders. We aim to develop a broad understanding of the subject, particularly regarding capacity costs and service levels.

1.2.1 Motivation

The author's interest in cyclical planning started with an investigation of the Western European supply chain at a global consumer goods company, where production volumes were determined once per month. Some products had so little demand that only one lot per month was produced, and some were produced in multiple lots per month, indicating a staggered system. The literature provided no adequate method for implementing this in a simulation or mathematical model.

Cyclical planning appeared once again when the author came into contact with a manager at a durable goods factory in Sweden. The production setting could be described as weekly production cycles with staggered deliveries. Products were called off daily by customers, which suggests a need for the daily monitoring of service levels. Due to the high labour content in assembly, labour costs, including overtime costs, were a major concern.

Although the research topic was inspired by industrial observations, no data was used from either of the mentioned companies.

1.2.2 Research questions

1. *What is the inventory-optimal policy under staggered deliveries and autocorrelated demand?* In many cases, demand has memory, in the sense that the present demand is influenced by the demand of the past. Lee et al. (2000) showed the industrial prevalence of this, by identifying the weekly sales of 150 out of 165 items at a supermarket as significantly autocorrelated. The literature on staggered deliveries ignores this, assuming independent and identically distributed (i.i.d.) demand, and in a single case independent, but not identically distributed demand (Lian et al., 2006). By identifying the optimal policy for autocorrelated demand, we learn how demand forecasts should be applied under staggered deliveries.

2. *How do costs and service levels develop under staggered deliveries and autocorrelated demand?* So far Flynn and Garstka (1990) showed the cost implications for i.i.d. demand, and Lian et al. (2006) provided the average availability under the same conditions. We seek to expand this to autocorrelated demand, and to provide the exact fill rate under staggered deliveries. This will provide us with deeper insights about the consequences of staggering deliveries.

3. *Under inventory costs and audit costs, can an optimal order cycle length be identified when demand is autocorrelated?* Flynn and Garstka (1997) identified the optimal order cycle length when demand was i.i.d. Autocorrelated demand is known to influence inventory costs in non-staggered systems, and we may anticipate that the optimal length of the order cycle is affected by demand autocorrelation.

4. *How can a linear production smoothing policy be applied under staggered deliveries and i.i.d. demand?* Linear production smoothing policies are used industrially, and their theoretical properties are well documented. Staggered deliveries have been justified as a way for creating smooth production plans (Chiang, 2009), but the consequences on bullwhip and capacity costs have not yet been explored, although both concepts, as will become apparent, are intertwined.

5. *How do overtime work rules affect the performance of systems with staggered deliveries?* Overtime may be collected into single shifts, or distributed evenly over a production cycle. Understanding how this affects

costs and service is integral to understanding production systems with staggered deliveries.

6. *Can an optimum order cycle length be identified when capacity and inventory costs are present?* How does the order cycle length affect capacity costs? This question is important for companies considering to modify their order cycles, as it may turn out that costs increase when ordering too frequently.

7. *How does the order cycle length interact with production smoothing?* Policies with non-unit order cycle lengths intrinsically contain a smoothing mechanism, as there is a temporal pooling effect. We are interested to understand how a given order cycle length affects the additional amount of smoothing required.

1.3 Thesis outline

This chapter has presented the research area and some important concepts. A literature review follows, where the current knowledge is presented, and research gaps are identified. Chapter 3 presents the methodology. This includes the philosophical underpinnings of the research, as well as the step-by-step approach to this research. The model development is divided into two parts, the first being Chapter 4, which provides the inventory-optimal policy under autocorrelated demand, and evaluates the performance of this policy. This addresses research questions (RQ) 1 and 2 directly, which leads to the resolution of RQ 3, also appearing in this chapter. The second model, developed in Chapter 5, investigates the case of production smoothing via a proportional policy for i.i.d. demand. This includes the optimal policy under quadratic cost, as well as policies with pragmatic overtime strategies, addressing RQ 4 and RQ 5. The answer these questions leads to the resolution of RQ 6 and RQ 7, presented in the same chapter. Chapter 6 presents simulation results to validate the analytical results, and Chapter 7 concludes this dissertation by reviewing and answering the research questions, and by identifying the managerial implications of this investigation. Future research opportunities are also noted.

To make the thesis self-contained, proofs for the optimal safety stock and capacity levels are presented in Appendix A. Appendix B contains some large proofs, which are related to the development of policies for staggered deliveries in Chapters 4 and 5. The existing theory on the fill rate was not strong enough

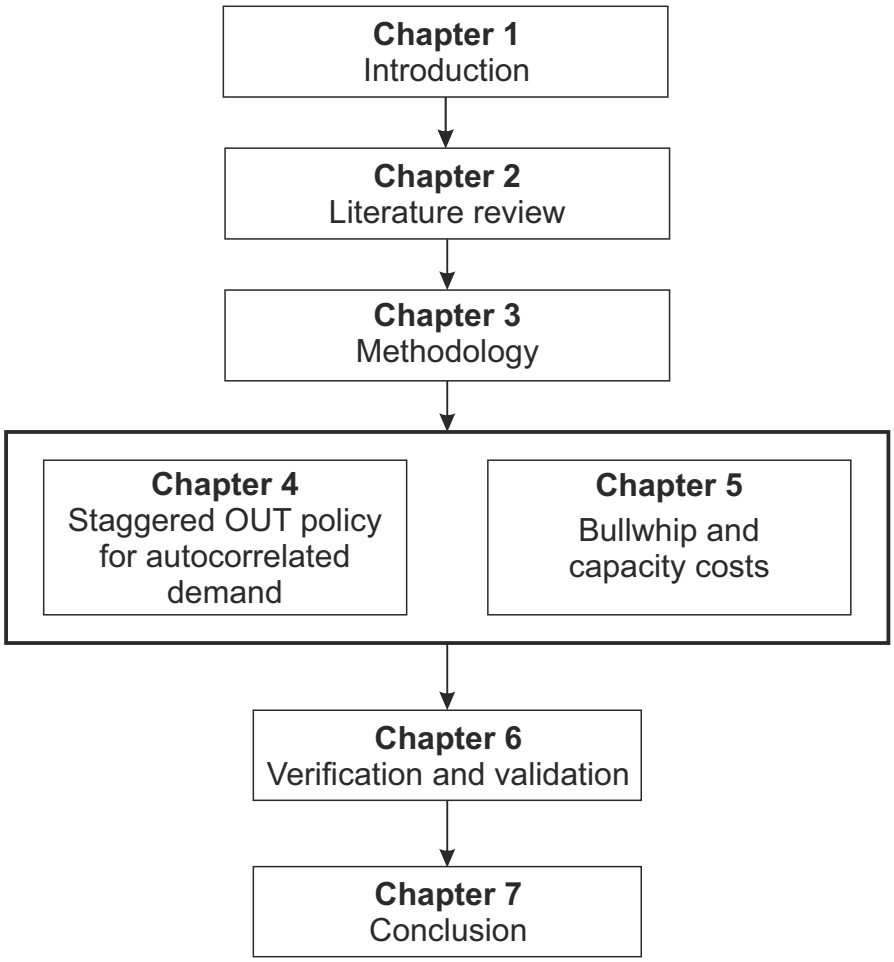

Figure 1.5: The chapters of this thesis and their interdependencies.

to accommodate autocorrelated demand, if negative demand is interpreted as returns. Therefore, a new expression for calculating the fill rate had to be developed. This is presented in Appendix C.

1.4 Contribution to knowledge

This work expands on knowledge by developing new insights about PIC systems with staggered deliveries. Important developments are:

1. The optimal staggered policy is identified for autocorrelated demand, when holding, backlog, and audit costs are present (RQ 1). This is accompanied by exact expressions for availability and fill rate for each period of the cycle (RQ 2), as well as an exact procedure for identifying the optimal order cycle length (RQ 3). Several important properties of

this system are also identified, as well as the consequences of maintaining a fixed safety stock, rather than an optimal time-varying safety stock.

2. The optimal staggered policy that minimizes the quadratic costs of inventory and production (i.e. the bullwhip-optimal policy) under i.i.d. demand is identified (RQ 4).

3. An investigation of four pragmatic policies operating under inventory costs and capacity costs is conducted (RQ 4). The approach extends in principle to any linear control policy. It is shown that staggered systems operating under a myopic policy may have an optimal order cycle greater than unity (RQ 6), due to a pooling effect that reduces capacity costs (RQ 7). Furthermore, this investigation shows that under piecewise-linear costs, it is best to perform all the required overtime work in the first period of each cycle (RQ 5). However, when considering inventory and capacity costs only, without an audit cost, it is better to use a policy capable of smoothing, and not to stagger deliveries (RQ 7).

4. A more specific definition of the fill rate is developed, providing reasonable results when negative demand represents returns. This definition is realized for the optimal staggered policy under autocorrelated, normally distributed demand. Although necessary for RQ 3, it is also useful in a wider context, as Disney et al. (2015) show.

1.5 Summary

Companies have order cycles ranging from a shift, up to a month or more. In many cases, we expect there to be several deliveries and inventory inspection per cycle, in which case we have staggered deliveries. Even so, the consequences of staggered deliveries are not well documented. It is the purpose of this thesis to shed light upon this phenomenon, so that we can understand the consequences of our production policies and find ways for improvement.

Chapter 2

Literature review

This chapter presents a literature review of the field of production and inventory control, first consisting of a general introduction to the field based on historically significant or well-cited sources; second, presenting a structured literature review of the narrow area which concerns cyclical ordering in inventory systems; and third, providing a brief overview of the literature on delivery performance.

To ensure a good match with the research questions, this literature review is focused on aggregate planning and control, i.e. determining the production requirements in the master production schedule (MPS). Other contributions are noted more briefly. Following the literature review, we reflect upon the state of the art for staggered delivery systems, and identify relevant research gaps, by comparing the literature on staggered inventory systems to that of non-staggered systems.

2.1 Theoretical overview

The following overview attempts to capture major developments in inventory control, and in PIC. As this field is broad, the goal is to identify different branches of research in this area, and to highlight important contributions within each branch. To classify as important, a work should be historically relevant, provide a foundational result, introduce new concepts or policies, or be particularly well-cited (as per Harzing's *Publish or Perish*). The focus lies on quantitative studies, with the occasional case study interspersed.

2.1.1 Inventory control

Fundamentals

The EOQ model of 1913 is likely to be the first step towards a quantitative theory of inventory control (Harris, 1990). Under some assumptions, it shows that the order quantity can be used to balance the cost of holding inventory and the fixed cost of placing an order. This was later extended by Wagner and Whitin (1958) to consider (known) time-varying demands and setup costs. The assumptions may not reflect a realistic industrial setting, but it demonstated that the ordering decision can be used to obtain an economic optimum by balancing conflicting costs. The EOQ assumptions can be compared with an argument by Lansburgh (1928, p. 398), who mentions that companies may tolerate a temporary inventory build-up to prevent idle capacity. In contrast to Harris, Lansburgh suggests that inventory levels (for each stock-keeping unit) are kept between a maximum and a minimum, with specific order quantities determined by judgment.

A limitation of the EOQ model is the assumption of constant demand. When seeking the optimal policy under stochastic demand, it was first found that the base stock policy (also known as the OUT policy) is optimal when the order cost is zero (Bellman, 2003, p. 153-182). The OUT policy observes the inventory position, and then places an order such that the inventory position is raised to a predetermined target level. When a fixed order cost is added, the optimal policy is of the (s,S) type (Scarf, 1959; Veinott, 1966). This policy operates by not ordering until the inventory position falls down to, or below, s; when this occurs, an order is placed to raise the inventory position to S, where $s \leq S$.

Forecasting for inventory control

To a large extent, the performance of inventory systems depends on the quality of forecasts. Gardner (1990) demonstrates this, and presents an example where a damped trend provides the best cost / service trade-off, outperforming exponential smoothing, linear regression, and naïve forecasts. The damped trend forecast is again connected to inventory control in Li et al. (2014), where the stability boundaries for this forecast are identified. Babai and Dallery (2009) present three dynamic inventory policies (variations of the reorder-point and OUT policies) that take forecasts into account. Under nonstationary demand, and with reliable forecasts, the dynamic policies outperformed the

(conventional) static policies; however when forecasts were unreliable, the static policies performed better (Babai and Dallery, 2009). In the same vein, Sethi et al. (2003) show that the optimal policy under fixed order costs and forecast updates is of the (*s,S*) type, with an adaption for the forecast updates. If the demand observations are sampled less frequently than the inventory system operates, the resulting forecast may be *temporally aggregated*. Rostami-Tabar et al. (2013) shows that a temporally aggregated exponential smoothing forecast outperforms (by mean squared error) a non-aggregated forecast of the same kind if demand is MA(1), or AR(1) with a low or moderate positive autocorrelation. As mentioned in Petropoulos et al. (2014), one must also consider the possibility of cross-sectional aggregation, i.e. aggregation across products, and to select an appropriate forecasting algorithm for the chosen level of aggregation.

Multi-product and multi-echelon systems

In this section, we have so far considered single-product inventories, but the theory also expands to simultaneous control of multiple products. With linear holding and backorder costs, but no fixed ordering cost, Veinott (1965) shows that the OUT policy is optimal. A well-known heuristic for multi-product ordering is given by Roundy (1986). A pragmatic way to manage multi-product systems is via ABC classification, where products are grouped into categories (usually three), based on some criteria, for example the expected revenue per product (Teunter et al., 2010). However, it may be better to classify products not by volume, but by a ratio involving holding costs, backorder costs, and the frequency of replenishment, as Teunter et al. (2010) demonstrate.

A supply chain may comprise several linked inventory installations, and is then termed a *multi-echelon* supply chain. The simplest case represents two serially linked inventories, but complex networks with multiple nodes of consumption, and of origin, are also tenable. Clark and Scarf (1960) identified the optimal policy for serially linked systems under similar assumptions as Scarf (1959), and characterized the optimal policy for each echelon: It is an (*s,S*)-type policy which, apart from the conventional inventory position, requires full knowledge of all inventory and WIP levels downstream of the echelon of interest. This result also implies that each echelon should operate without regard to any upstream installations — a result that agrees with Bellman's (2003) *principle of optimality* (discussed in the following chapter). As with single-echelon models, multi-echelon supply chains can also be configured for target service levels; Diks et al. (1996) present a review. Owing to their complexity, multi-echelon systems

are prone to disruptions in supply and demand; the effects of such disruptions can be mitigated by appropriate dimensioning of safety stocks (Schmitt and Singh, 2012).

When different echelons have conflicting economic interests, full information sharing may not be tenable. The consequences of withholding information is investigated by Lee et al. (2000), who find that the upstream echelon of a two-tier supply chain suffers when information is lacking, particularly when demand is significantly autocorrelated and when lead times are long. Not all multi-echelon supply chains are centrally controlled, as the previous references imply. The decision makers in each echelon may have conflicting objectives, encouraging policies that are locally optimal, but detrimental to the supply chain as a whole. To alleviate these effects, Lee and Whang (1999) propose transfer pricing (retailer pays only the variable cost of goods), consignment (retailer inventory is owned by supplier), an additional backlog penalty (self-explanatory), and shortage reimbursement (supplier penalized for inadequate deliveries). The second of these points is a half-way step towards *vendor-managed inventory* (VMI), which refers to arrangements where the supplier places replenishment orders on behalf of its customers. VMI does not refer to any specific policy or arrangement, but commonly, there will be an agreement on permissible minimum and maximum inventory levels (Jonsson and Mattsson, 2005, pp. 455–458). Another possible implementation uses reorder points and minimum order quantities (Holmström, 1998). VMI also appears in multiproduct settings, but due to complexity, heuristic policies are used (see e.g. Cárdenas-Barrón et al., 2012).

2.1.2 Delivery performance

There is more to inventory management than cost balancing. Hadley and Whitin (1963, p. 217) introduce the ratio of incurred backorders to average demand (over a year). The complement of this ratio (i.e. the long-run fraction of filled orders) would later be known as the fill rate. Inventories are often maintained so that demand may be satisfied from stock. There are several ways to measure how well demand is satisfied:

- As the fraction of items delivered immediately
- As the fraction of periods in which all orders are filled immediately
- As the fraction of customer orders which are completed immediately
- As the fraction of order lines filled immediately

- As the fraction of products, by dollar value, filled immediately.

These service measures can provide quite different results, and it is important to select an appropriate one. In a single-product setting we have only

1. *Availability,* the probability of satisfying all demand within an order cycle
2. *Fill rate,* the fraction of demand filled immediately
3. *Ready rate,* the fraction of time when there is no shortage.

In periodic inventory systems, availability tends to be straightforward to calculate.

$$S_1 = \mathbb{P}(\kappa \geq 0), \tag{2.1}$$

where κ is the inventory level at the end of the inventory cycle (See Figure 2.1), and $\mathbb{P}$ is the probability.

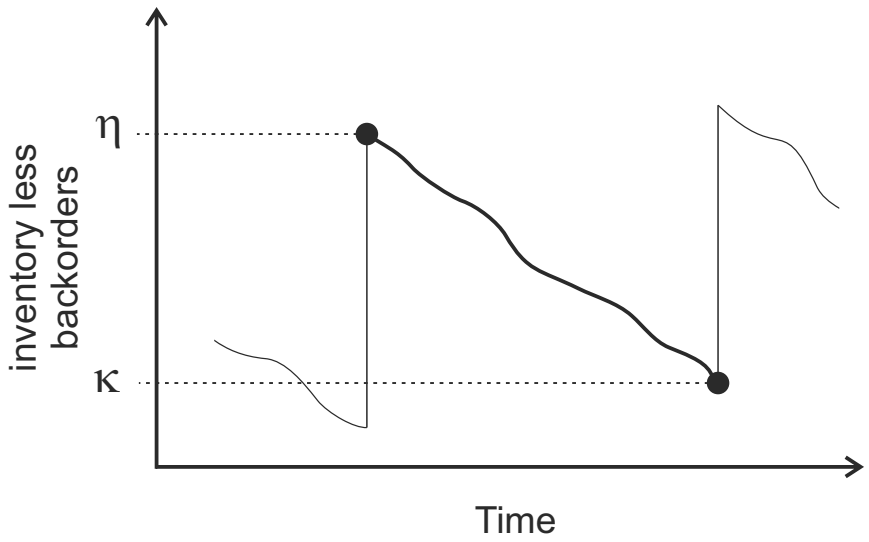

Figure 2.1: Starting and ending inventory levels in an inventory cycle.

The fill rate is a popular service measure among practitioners (Guijarro et al., 2012), and is considered as a more appropriate service measure than availability (Axsäter, 2006, p. 95). The formulae for calculating the exact fill rate often make restrictive assumptions, even in general cases. For his general fill rate formula, Axsäter (2006) requires inventory cycles to be periodic, but does not require lead times to be multiples of the inspection period. Axsäter's fill rate expression is only exact for non-negative demand. It has the form

$$S_2 = 1 - \frac{\mathbb{E}\left[(\kappa)^-\right] - \mathbb{E}\left[(\eta)^-\right]}{\mu}, \tag{2.2}$$

where η is the inventory level at the start of the inventory cycle, $\mu = \mathbb{E}[\eta - \kappa]$ is the expected demand over the inventory cycle, $\mathbb{E}$ is the expectation operator, and $(x)^- = \max(-x, 0)$. Teunter (2009) provides an equivalent expression that is more concise,

$$S_2 = \frac{\mathbb{E}\left[(\eta)^+\right] - \mathbb{E}\left[(\kappa)^+\right]}{\mu}, \tag{2.3}$$

where $(x)^+ = \max(x, 0)$. For the OUT policy, an exact fill rate expression is given by Sobel (2004). It was later refined by Zhang and Zhang (2007), to accomodate for lead times that are not a multiple of the review period. Both approaches assume non-negative demand that is independent and identically distributed, and a constant lead time. Both of these are special cases of the fill rate expression used by Axsäter (2006) and Teunter (2009).

When fill rates are high, it is common to use the approximation

$$S_2 \approx 1 - \frac{(\kappa)^-}{\mu} = \frac{(\kappa)^+}{\mu}, \tag{2.4}$$

Note that $\mu = \eta - \kappa$. When demand is normally distributed and the replenishment policy is specified, some specific fill rate expressions can be obtained.

Axsäter (2006) provides approximations for both the continuous-time and the discrete-time reorder-point policy. Teunter (2009) provides the discrete-time case as well. Silver and Bischak (2011) investigate the discrete-time case under normally distributed demand, identifying an approximate fill rate expression for demand with a low coefficient of variation. The exact fill rate under normally distributed demand was first given by Johnson et al. (1995), where the effect of negative demand is taken into account. As with the other fill rates, it assumes i.i.d. demand and an order-up-to policy.

2.1.3 Production and inventory control

Genesis

With Production and Inventory Control (PIC) we shall refer to systems whose performance is not only determined by the variations in inventory levels, but also by the variations of the order rate, which tend to be an important consideration when finite capacity or overtime costs are in place.

The first documented approach to this is likely to be Simon (1952), who defined the problem using differential equations and the Laplace transform. Simon showed that there was a trade-off between the amplitude of the order

rate and the inventory level, when sinusoidal demand is assumed. This work is extended to a practical application in Simon and Holt (1954), who claim that the Eastman Kodak Company effectively managed to remove seasonal variations from its production rates, and also provided a solution for how this can be done. With monthly ordering, the suggested order rate was

$$o_{t+k+L} = \hat{d} + 0.04(i^* - i_t), \tag{2.5}$$

where o_{t+k+L} denotes the orders to be received in month $t + k + L$, $\hat{d}$, is a 12-month moving average forecast, i^* is the desired safety stock level, and i_t is the current inventory. Note the strong damping of the inventory error. Simon's approach to PIC is developed and documented in Holt et al. (1960), where evidence of successful industrial applications is documented.

The policy above can be compared to the OUT policy described in Arrow et al. (1951) and refined in Bellman (2003). The OUT policy is optimal when there are convex inventory costs and no other costs. In every period there is a target inventory position, x_t^* and an observed inventory position x_t. The optimal order decision is then to raise the inventory position to the target value, i.e. $x_t^* = o_t + x_t$. At roughly the same time as Simon's initial paper, Tustin (1953) published *The Mechanism of Economic Systems*, wherein a PIC system with lead times was investigated. A conclusion reached by Tustin was

> "... it is possible for a trading system to show phenomena of an oscillatory kind, quite apart from price phenomena, due to no more than the efforts of traders to maintain stocks in the face of delays in delivery. The possibility is indeed an obvious one." (Tustin, 1953, p. 98)

This phenomenon gained notoriety following Forrester (1958), who proposed Industrial Dynamics (a type of discrete-time simulation, often with nonlinear components), to model and understand excessive order rate fluctuations in industry. The example was a multi-echelon supply chain, in which demand, backlogs, capacity, cash flow, information, and orders were accounted for.

Howard (1963) investigated another policy, where the order quantity was equal to a proportion of the sum of the last two recorded inventory levels,

$$o_t = \alpha \left[(i^* - i_t) + (i^* - i_{t-1}) \right]. \tag{2.6}$$

Using Z-transform techniques and the impulse response, Howard showed that the special case $\alpha = 1/2$ produces an underdamped system.

Deziel and Eilon (1967) present a discrete-time model, analogous to Simon (1952), which took an inventory-optimal ordering rule (Vassian, 1955) and added a proportional feedback controller to allow for damping of the order rate. Simon (1956) showed that a linear feedback policy is optimal under quadratic costs. Another optimal policy was characterized in Sobel (1970) under the assumption of convex inventory costs, and convex, piecewise-linear production costs. This result is related to Beckmann (1961), who investigated a system without overtime, but with a specific cost for changing the capacity level.

The modern era

The proportional feedback rule was again investigated by Towill (1982), who identified the variance amplification ratio of the order rate and inventory to white noise demand for continuous PIC systems with an exponential smoothing forecast and no work-in-progress (WIP) feedback. In a development, John et al. (1994) presented a generalization of this, the Automated Pipeline and Inventory Order Based Production Control System (APIOBPCS), with separate proportional feedback controllers for WIP, and inventory errors. APIOBPCS variants have been implemented at Lexmark (Disney et al., 2013), and at Tesco (Potter and Disney, 2010).

An investigation of the effect of various parameter setting can be found in Disney and Towill (2003) and Zhou et al. (2010); one important insight is that the Deziel-Eilon rule, where the feedback parameters are equal, has easily recognizable stability limits and also provides good performance without unnecessary oscillations in the order rate. Further work on APIOBPCS includes the application of demand signals using the Fourier Transform (Dejonckheere et al., 2003), the performance of systems with VMI (Disney and Towill, 2002), and the application of costs to both inventory and orders (Disney and Grubbström, 2004; Hosoda and Disney, 2012). In the last reference, the optimal capacity from labour appears to have a newsvendor-type solution (we confirm this for arbitrary distributions in Appendix A). When the future production requirements are known, the optimal staffing rule is obtainable from an algorithm by Rao (1990), resembling the well-known Wagner-Whitin algorithm (Wagner and Whitin, 1958).

The problem of excessive order rate fluctuations gained notoriety as the *bullwhip effect* (for brevity: bullwhip) following Lee et al. (1997). Bullwhip is present when the variance of orders exceeds the variance of demand, and the

condition is usually stated

$$\frac{\text{var}(o_t)}{\text{var}(d_t)} > 1. \tag{2.7}$$

This formula is used as a test to see if the ordering policy amplifies demand fluctuations (a possiblility Tustin [1953] predicted), or if the policy attenuates them (when the ratio is less than unity). Lee et al. (1997) identifies four potential causes of bullwhip:

- Demand signal processing
- Rationing gaming
- Order batching
- Price variations

Bullwhip is not always detrimental. The causes illustrated by Lee et al. (1997) use policies where there is a cost of inventory, but no cost related to bullwhip, or to the production decision. Therefore, we should ignore the bullwhip effect, when there is no penalty associated with it. However, when capacity costs are present the bullwhip effect should be considered — In an example, Metters (1997) finds that bullwhip mitigation can reduce relevant costs by 10%–30%. The potential benefits of targeting bullwhip are the greatest under quadratic cost models, like the ones used in Holt et al. (1960). For this cost setting, a full-state policy is optimal for ARMA(p,q) demand, as shown by Gaalman (2006). For a comprehensive review of bullwhip research, see Wang and Disney (2016).

The multiechelon aspect also appears in PIC systems. For example, Disney and Towill (2002) develop a VMI policy capable of production smoothing by expanding APIOBPCS to a VMI setting, identifying transfer function for orders and inventory levels, as well as the stability criteria for the policy.

2.2 Structured review

One research question is centred on the implications of cyclical ordering, and the following section pinpoints the current literature on this specific topic via a structured literature review, which was done according to Table 2.1. Apart from this search, the reference lists of relevant papers in the search were investigated for additional contributions, in particular by searching reference lists for authors of other germane papers, and for relevant paper titles. Scopus returned 351 hits, and Web of Knowledge returned nine. Following an inspection of the

Table 2.1: Specification of literature search

Search engines:	Scopus, Web of Knowledge
Time period:	Published before 2015-07-20
Publication type:	Journal papers
Language:	English
Search parameters:	Scopus: Title and keywords Web of Knowledge: Topic
Search keywords:	(inventory AND "*order period") OR (inventory AND "*order cycle") OR (inventory AND "*replenishment cycle") (inventory AND "staggered deliveries")

abstracts, a majority of the papers did not distinguish between order cycles and inspection periods or batch frequency. After a detailed inspection of the papers, fifteen were relevant — Out of these, six were of moderate relevance, and nine addressed staggered deliveries explicitly.

The moderately relevant papers are termed so, because they contain models that are either not staggered, or because the staggering is in response to a known production requirement, which has more to do with detailed scheduling than with aggregate planning. Bradley and Conway (2003), Kaku and Krajewski (1995) and Nathan and Venkataraman (1998) are closest to shop floor control, as they consider scheduling and set-up times. The most essential division of these papers pertains to the level of aggregation which each paper investigates. Modigliani and Hohn (1955) is also related to these as it assumes known future demand over a fixed planning horizon. These papers also make very specific assumptions about the production system, limiting the insights gained to a narrow set of production systems. Other relevant papers operating on an aggregate level (MPS) include Tang and Grubbström (2002), where there is a conditional schedule change cost that may be incurred once per period; and Zhao and Lee (1993), presenting a simulation model of multi-level MRP systems where the MPS may be frozen. Zhao and Lee (1993) finds that longer order cycles can stabilize production, and that forecast errors degrade performance.

The nine remaining papers take an aggregate approach to staggered deliveries, where the production volume per period is the variable of interest. Despite investigating the same concept, the terminology varies: Five papers with James Flynn as an author refer to *staggered deliveries*, Chiang (2009) calls it a *periodic review replenishment model*, Lian et al. (2006) talks about the *frozen period*, Prak et al. (2015) uses *periodic review and continuous ordering*, and Chiang (2001) uses the term *order splitting*. The last term is more commonly used in

the context of splitting orders not across time, but over multiple suppliers, as in Kelle and Silver (1990) (not a staggered deliveries paper). The papers identified from the literature search make it clear that the effects of the order cycle are compared either to a fixed or binary cost per cycle or delivery, or to a capacity cost or constraint. It is worth noting that non-trivial order cycle lengths are investigated in very few papers, with vastly different assumptions and modelling approaches, even though several notable industrialists (see Chapter 1) have mentioned it as a pertinent problem. The following section summarizes the main insights from the papers that treat staggered deliveries on the level of aggregated planning.

2.2.1 Staggered deliveries

Only a handful of the papers from the structured literature review addressed staggered deliveries, where multiple consecutive orders are planned at a single point in time.

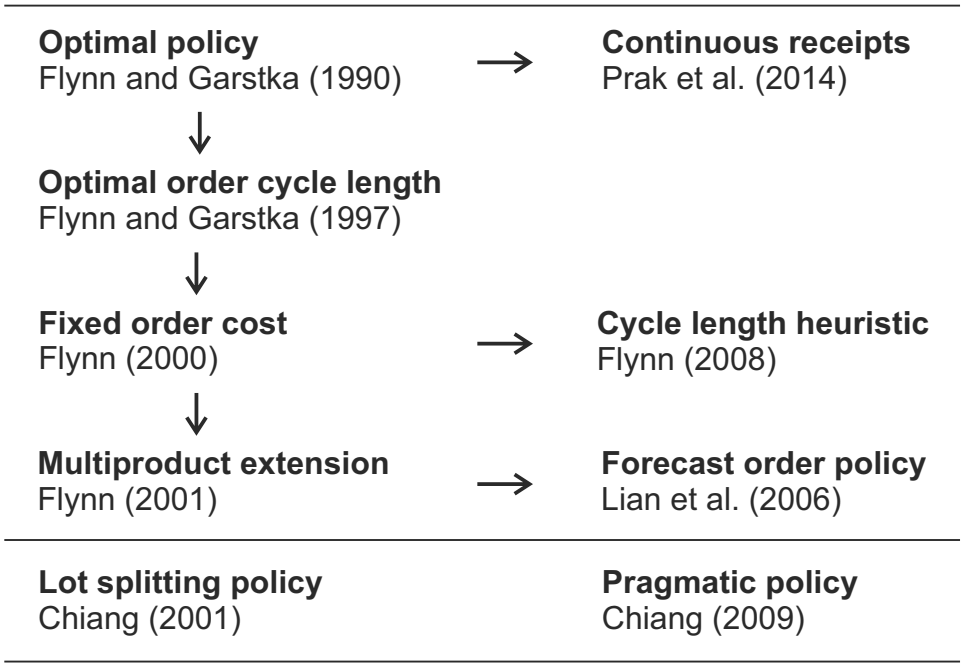

Figure 2.2: Interdependencies between the papers with staggered deliveries

The first of these is Flynn and Garstka (1990), which proves a staggered variation of the OUT policy to be optimal under convex inventory costs, i.i.d. demand, and a non-negative ordering constraint. Apart from the inventory cost, Flynn and Garstka considered an audit cost, incurred once per order cycle. The audit cost reflects the cost of inspecting the inventory level and determining the

order quantity. When the model is constrained to piecewise-linear inventory costs, the optimal OUT level for each period corresponds to a critical fractile, i.e. optimal OUT levels produce a certain probability of ending *each period* with positive inventory. This holds for both finite and infinite time horizons. Notably, all order cycles but the terminal one (in finite-horizon cases) have the same set of OUT levels.

Building on these results, Flynn and Garstka (1997) present a procedure for identifying the optimal order cycle length, which can be greater than one period if the audit cost is sufficiently high — More specifically, they show that the optimal order cycle length increases with the audit cost. In addition, Flynn and Garstka identify a lower bound for the average cost per period, and suggest two heuristics for finding the optimal order cycle length. The main benefit of these heuristics is the ease of implementation compared to the exact procedure.

A fixed cost for ordering a sequence of P orders (that is not incurred if all order quantities in a cycle are zero) is introduced in Flynn (2000), who identifies the optimal policy under this setting as a staggered (s,S)-type policy. Flynn also proves that an optimal order cycle length exists, and provides an algorithm for its computation. This exact procedure is tested against four heuristics, which perform well when demand is normally distributed, with a coefficient of variation no greater than 50%. Flynn (2008) develops another heuristic, which is asymptotically optimal for normally distributed demand.

The model in Flynn and Garstka (1990) is also expanded to a multi-product scenario in Flynn (2001), where each product has a specific order cycle length. Flynn (2001) identifies the OUT policy as optimal regardless of the products' order cycle lengths, proves that an optimal set of order cycle lengths exists, and presents sufficient optimality conditions for the set of order cycle lengths. To identify the parameter settings of the optimal policy, Flynn (2001) uses a branch-and-bound algorithm, which performs well when the number of products is no greater than ten. When there are more products, Flynn suggests two heuristics which are demonstrated to compute faster, but with an error in the estimated cost of up to 2.59%.

Staggered deliveries are studied in relation to lot splitting in Chiang (2001), where inventory costs are incurred periodically, and a single-lot approach is compared with two lots or more. A fixed order cost, a cost per lot, and holding costs are considered. Backlog costs are not included, but their equivalent can be found in a service level (availability) constraint. The lot-splitting policy is essentially an OUT policy (applied to an entire order cycle) which is divided

into P receipts that need not be equal, as each receipt quantity is weighted differently between the periods in a cycle. Chiang (2001) demonstrates that lot-splitting can be beneficial when the audit cost is significant, presenting numerical examples with savings of 7%–12%.

Another staggered policy appears in Chiang (2009), where the order quantity in the beginning of the cycle is arbitrary (until a target inventory position has been reached), and constant in the remaining periods of each cycle. This ensures that the production rate is constant in all but a few periods. Chiang (2009) studies this policy with and without a non-negative ordering constraint. As one would expect, the unconstrained policy achieves a lower cost, and also has the benefit of only having a variable production quantity in the first period of the cycle, and constant production in the remaining; this does not hold for the constrained policy, as production may be reduced when necessary (but not increased), in the periods following the first. Chiang's numerical examples investigate several values for the audit cost, indicating that for either of the two policies, the optimal order cycle length increases with the audit cost.

Real demand data and forecasts are applied to staggered deliveries in Lian et al. (2006), who present a simplified policy based on a single safety stock for the cycle, and on the forecasts of individual periods. Holding, backlog and audit costs are considered, implying that a staggered OUT policy is optimal. The performance of the simplified policy is compared with the optimal policy (which adds time–varying safety stocks), indicating that the policy performs reasonably well. For their numerical example, the savings potential of switching to the optimal policy is less than 2%. Lian et al. (2006) also investigate the optimal order cycle length, and identify this by enumeration. Notably, the optimal length in their numerical example depends on the policy used: For the optimal policy the optimal order cycle length is three periods, whereas it is five periods for the simplified policy with forecasts.

Prak et al. (2015) investigate a system with discrete audits and continuously staggered receipts, finding that a staggered OUT policy is optimal. In a numerical example, they compare continuous ordering with periodic ordering, and identify a savings potential of $30\% - 60\%$ under continuous ordering. The current state of research on staggered deliveries is summarized in Table 2.2. Notably, almost all models use some variation of the OUT policy, and they differ by how the OUT levels are calculated.

Table 2.2: Classification of the papers on staggered deliveries.

Paper	Policy	Type	P^* treatment	Notes
Flynn and Garstka (1990)	OUT	Optimal	—	—
Flynn and Garstka (1997)	OUT	Optimal	Analytical	—
Flynn (2000)	(S,s)	Optimal	Analytical	Fixed order cost
Flynn (2001)	OUT	Optimal	Analytical	Multi-product
Chiang (2001)	OUT[a]	Heuristic	—	Lot splitting
Lian et al. (2006)	OUT	Heuristic	Numerical	Forecasting, S_1 service [b]
Flynn (2008)	OUT	Optimal	Analytical	—
Chiang (2009)	OUT	Heuristic	Numerical	—
Prak et al. (2015)	OUT	Optimal	—	Continuous time

[a] OUT policy over the cycle; all receipt quantities within each cycle are identical.

[b] Average long-run availability calculated over all periods simultaneously.

2.2.2 Identifying research gaps

To get an understanding of the state of staggered delivery research, the papers about staggered deliveries were compared with major themes in the non-staggered literature. As the research on staggered deliveries is limited, many results which are fundamental to the non-staggered literature do not exist (e.g. fill rates, capacity costs, multiechelon systems). Figure 2.3 presents the state of the staggered delivery research based on eight classes of criteria that are derived from our general review of inventory control and PIC. The figure mentions only the first papers to expand staggered deliveries in a particular direction, as these papers reflect the recognition and treatment of research gaps.

Figure 2.3 reveals the following about the state of staggered research: All models up until now have been single-echelon, starting with Flynn and Garstka (1990), which also constitutes an optimal policy under inventory costs and an audit cost. As there is still much to be explored in this setting, all research questions, RQ 1–7, concern single-echelon systems. The first suboptimal, but pragmatic, policy was the order-splitting policy of Chiang (2001), later to be followed by Lian et al. (2006) and Chiang (2009). We consider new pragmatic policies as a partial solution the capacity cost problem of RQ 4. The work on RQ 4 also provides an optimal policy under quadratic costs, while RQ 1

provides the optimal policy under autocorrelated demand and piecewise-linear inventory costs. In all cases demand has been assumed to be i.i.d., except for Lian et al. (2006), where demand was merely assumed to be independent over time. Our contribution to staggered delivery models with i.i.d. demand concerns service levels (RQ 2) as well as the study of systems with capacity costs, covered by RQ 4–7. No work on autocorrelated demand has been done within the context of staggered deliveries; this gap is filled by answering RQ 1–3. Since Flynn and Garstka (1990), the typical costs considered are audit costs and inventory costs (treated by RQ 1–3, and 6) — Capacity costs have not been modelled before, but this research gap is addressed by RQ 4–7. Only Lian

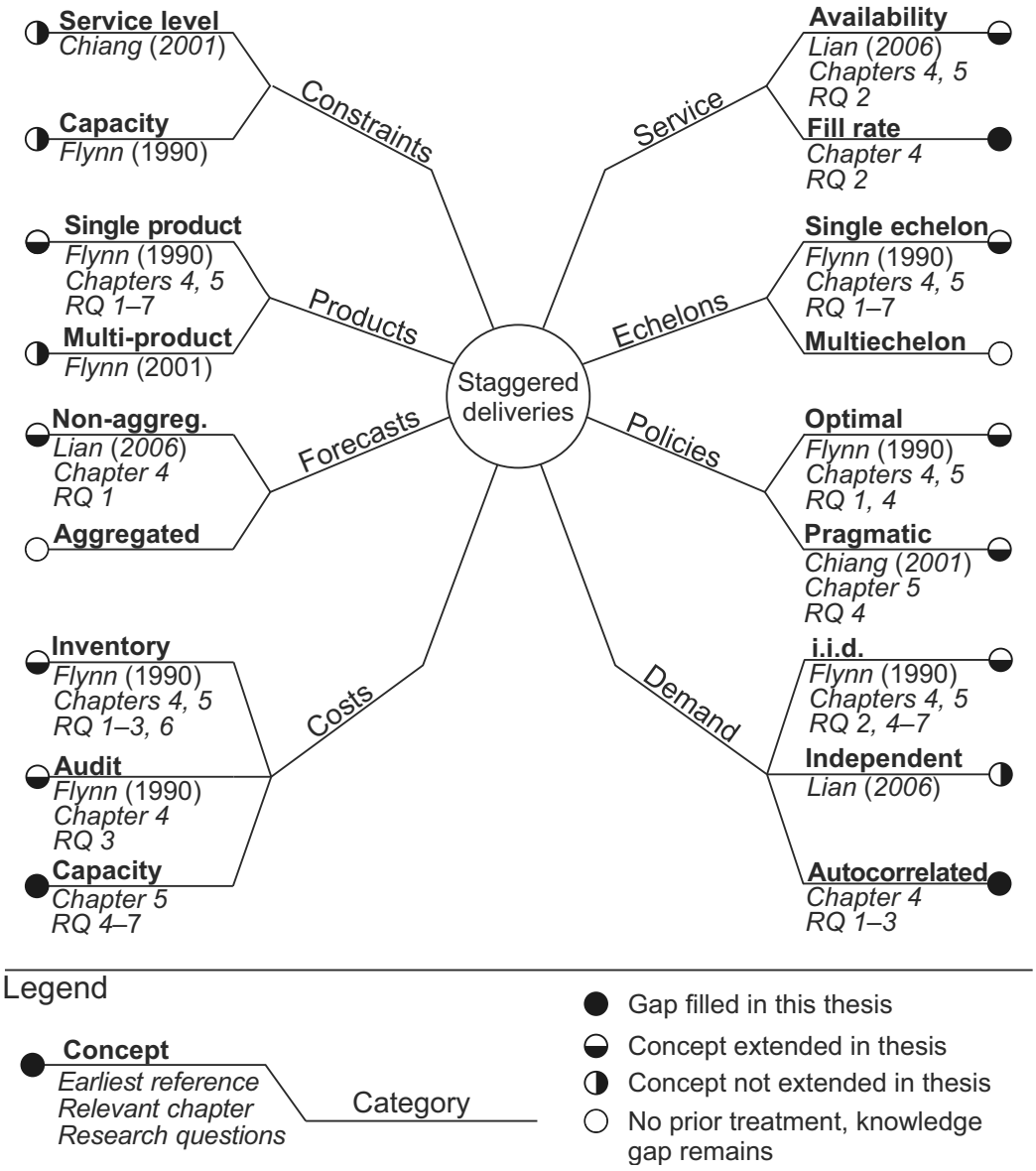

Figure 2.3: Resolved and remaining research gaps for staggered deliveries.

et al. (2006) have investigated the effects of forecasts in relation to staggered deliveries, but they did not consider temporally aggregated forecasts despite the attractiveness of this solution in relation to long order cycles. We extend the theory of non-aggregated forecasts as part of the optimal policy sought by RQ 1. System constraints were first implemented as nonnegative orders in Flynn and Garstka (1990), while an upper capacity limit (in some periods) was used first by Chiang (2009). Another constraint was used by Chiang (2001), where a service level constraint was used in lieu of backorder costs. As for service levels, only Lian et al. (2006) provided the *average* availability over the cycle, but not the availability per period. Fill rates have not been considered at all. Contributions to the theory of availability and fill rates result from addressing RQ 2.

From these observations, we deduce that the theory on staggered systems is incomplete with respect to autocorrelated demand, capacity costs, fill rates, availability, (temporally) aggregated forecasts, and multiechelon models. It is not difficult to think of other research gaps, but the ones presented are essential to our understanding of inventory and production systems. In light of these research gaps, we shall investigate staggered deliveries under autocorrelated demand, inventory costs and audit costs, and for this derive the optimal policy, and the resulting availability and fill rate. This covers two of the gaps in Figure 2.3. We shall also investigate capacity costs (for i.i.d. demand), using a policy that minimizes quadratic cost, and using a family of pragmatic policies that are staggered variants of APIOBPCS. Aggregated forecasts and multiechelon models remain open problems.

Chapter 3

Methodology

This chapter outlines the methodology used to address the research questions. It begins with some general views on knowledge and research, and then delves into the actual research design. A section on validation explains how we can ensure that this research is trustworthy.

3.1 Epistemology and ontology

Research is influenced by ones perception of reality. The set of assumptions one holds about the nature of reality is termed *ontology*, and the set of assumptions about knowledge and research is termed *epistemology* (Blaikie, 1993, p. 6). These terms come as a pair, often in relation to a *research strategy*. Sometimes, this is called a research paradigm. The importance of choosing the right research strategy can be understood by comparing two widely contrasting approaches: the *positivist* and the *phenomenological*. The difference between them can be seen in Table 3.1. The positivist approach encompasses the attitudes and methods of hard science, while the phenomenological approach is closer to the social sciences.

The choice of research strategy is not arbitrary, as it depends on the problem being researched. In this dissertation, the research questions ask how a design feature (periodic reordering) influences specific variables (service levels and cost); therefore we would do well to opt for a positivist approach.

The positivist approach described by Easterby-Smith et al. (1991) comes in various flavours, encompassing traditional positivism, critical rationalism, and realism. *Traditional positivism* assumes in its ontology a material world, where only observable or measurable events are considered to be real (Blaikie, 1993, p. 94). Traditional positivism leaves no room for studying unobservable events.

Table 3.1: Comparison of the positivist and phenomenological research paradigms (Easterby-Smith et al., 1991).

Paradigm:	Positivist	Phenomenological
Basic beliefs:	The world is external and objective;	The world is socially constructed and subjective;
	Observer is independent;	Observer is part of what is observed;
	Science is value–free.	Science is driven by human interest.
Researcher should:	Focus on facts.	Focus on meanings.
Preferred methods:	Look for causality and fundamental laws;	Try to understand what is happening;
	Reduce phenomena to simplest elements;	Look at the totality of the situation;
	Formulate hypotheses and then test them;	Develop ideas through induction from data;
	Operationalizing concepts so that they can be measured;	Using multiple methods to establish different views of phenomena;
	Taking large samples.	Small samples investigated in depth over time.

Its purpose is to predict the observable.

Critical rationalism is akin to positivism, but comes with a different epistemology. It assumes that before any observation is conducted, we must have a reason to make that observation (Blaikie, 1993, p. 95). Therefore, theory precedes observation. The practical difference between traditional positivism and critical rationalism is that the former uses observations as a foundation for theory, while the latter builds theory to be tested against observations (Blaikie, 1993, p. 96).

Realism, sometimes called *empirical realism*, assumes that reality exists independently from scientific activity. In contrast to positivism, which is concerned only with the empirical (observed events), realism also considers the entire set of events, observed or or, which is termed the *actual*, and also the mechanism which brings about these events, termed the *real*. In this way, objects can exist and events can take place without being observed, or even observable. The result is that realism, in contrast to positivism and critical rationalism, can accept something as true (real), even if it is unobservable. Because the real is assumed to exist, realist research seeks not only to predict events, but also to explain why they occur. Realism is not associated with induction or deduction, but with retroduction (Blaikie, 1993, p. 98-99), which works as follows:

> When a non-random pattern is identified, the first step is to undertake a series of experiments to determine the range of conditions under which it appears. Then the processes which generate the pattern are to be looked for in the natures of the things and materials involved. It is the fact that these are usually not known that brings into action the model building process. The creative task is to create a plausible analogue of the mechanism which is really producing the phenomenon. (Harré, 1976, p. 21)

An important challenge that realists must face is the risk of developing models that, contrary to the researcher's perception, fail to represent reality. In cases like these, the research may fail to identify the real mechanism behind the phenomenon being investigated (Bryman, 2012, p. 29). This phenomenon is the basis of *critical realism*, which seeks to identify the real mechanisms in order to effect a change to the (often social) system. To attain this, critical realists seek to understand the entirety of a situation by providing information about context, as this is believed to bring the model or description closer to reality.

3.1.1 Considerations for operational research

Operational research (OR) and operations management do not rely on a single theoretical foundation, but encompasses a multitude of perspectives and methodologies (Boer et al., 2015). With disparate methodologies such as case studies, surveys, simulation models, and mathematical models, there will sometimes be differences in the view on research. For example, Checkland and Scholes (1990, p. 22) argue that it may be useful to think in terms of systems, but questions their existence. In contrast, system dynamics asserts that systems are real, and that systems cause observable behaviour (Forrester, 1994). On a similar note, Arrow et al. (1960, p. 18) suggests that models are approximations that highlight the most important tendencies of real systems.

To connect this with research paradigms, we note that research based on mathematical modelling or simulation presupposes the existence of a causal mechanism, i.e. a system, or what realists call the *real*. In contrast, pure positivism dismisses evidence from models, as it requires empirical observations for theory-building.

Minas (1956) argues that there is a beneficial interaction between model-building and empirical work. As a practical example, consider the mathematical models of Simon (1952) and Tustin (1953) predicting the bullwhip effect, which is now supported by rich empirical evidence (Isaksson and Seifert, 2016). Models have also identified ways to mitigate problems with bullwhip, and these remedies have been demonstrated to work (Disney et al., 2013; Potter and Disney, 2010). With realist eyes, we would see this development as evidence of a useful model that in its general features reflects the real world. With positivist eyes, only the empirical evidence counts, while the models at best serve as tools for generating experimental hypotheses.

This thesis seeks to develop an understanding of staggered deliveries from first principles. The assumption is that such systems are real, and that physical implementations of staggered delivery systems will share some features with simple theoretical models of staggered deliveries. Because we assume that a system mechanism causes events and potential empirical observations, realism is the appropriate research paradigm.

The realist approach provides a good match with the research questions: A design feature (periodic reordering) is a real-world entity; we suspect it influences costs and service levels; can we build a model that captures the essential consequences of periodic reordering, even if it only approximates reality? In light of the research questions, it appears that realism is the

approach best suited for this dissertation.

3.2 Research method

This section details the methods used, why they were chosen, and how they were executed. At first, the general research area (staggered deliveries) was chosen, as it appeared to be largely ignored, even though this practice has been observed in supply chains. To find out the current state of research for this problem, a systematic literature review followed.

3.2.1 Literature review

The literature review started with a broad review of production and inventory control research. Then, a search for keywords that were thought appropriate for the problem of staggered deliveries was conducted. Initially, MetaLib, Scopus, Web of Knowledge, and Google Scholar were used to find appropriate papers, and more relevant keywords. Then, the reference lists of the identified papers were used to find further papers on this topic. The reference lists were searched for titles that appeared relevant, and for authors who had written other papers on the staggered deliveries.

Based on new keywords found from the initial searches, a structured search was conducted on Web of Knowledge and Scopus. After identifying relevant papers, they were classified based on their differences, and a research gap was identified. This formed the basis of the research questions.

3.2.2 Modelling approaches

Having opted for a realist approach, we are required to construct a model to explain the underlying mechanism of the research problem. By its very nature, the staggered deliveries problem reflects performance over time: it is dynamic. Pidd (2003) argues that one can understand management problems by building mathematical models, simulation models, and soft systems models. All of these try to explain how a set of inputs affects a set of outputs. Mathematical modelling and simulation quantify this, while soft systems modelling does not. A common feature of these approaches is the concept of a system state, and controls.

Mathematical modelling

In mathematical models we tend to consider a state vector $\mathbf{x}_t$ that describes the state of the system with respect to t. Depending on the model, the state can change in discrete time increments, $t \in \mathbb{Z}$, or in continuous time, $t \in \mathbb{R}$. An example of a system evolution in discrete time is

$$\mathbf{x}_{t+1} = f\left(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t, t\right), \tag{3.1}$$

where $\mathbf{u}_t$ is the control and $\mathbf{w}_t$ is some exogenous input; and the corresponding continuous time realization is

$$\dot{\mathbf{x}}_t = f\left(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t, t\right), \tag{3.2}$$

where $\dot{\mathbf{x}}_t$ is the derivative of $\mathbf{x}$ with respect to t. The discrete-time case is treated with difference equations, and the continuous–time case with differential equations. A thorough treatment of the topic is provided in Luenberger (1979). In some cases, the problem of determining a control $\mathbf{u}$, can be attacked heads-on; this is done in Box and Jenkins (1976). But not all situations are amenable to this direct approach. Then we can consider classic control theory, optimal control, or dynamic programming.

Control theory refers to techniques for making the output of a system conform to an input signal. It is helpful to differentiate between classical control theory, and optimal control. The classical approach aims to produce systems that perform acceptably when provided with certain input signals (Kirk, 1997, p. 3). Commonly tested signals are a single impulse, a step increase, a ramp, a parabola, or a sine wave (Nise, 2011, p. 19). The criteria for an acceptable design can include aspects of the transient response, (such as the peak overshoot and the settling time), the steady-state error, or the *stability* of the system. We note that classical control theory does not seek an optimal solution, but it can be used to find solutions that are robust to widely different input signals. Before considering modern control theory, we shall make some brief observations about stability.

The term *asymptotic stability* refers to the tendency of a system, regardless of initial conditions, to approach an equilibrium point, μ, as time progresses. If the system instead diverges to positive or negative infinity, the system is said to be *unstable*, and if the system neither converges, nor diverges, but falls into a limit cycle, it is termed *marginally stable* (Luenberger, 1979, pp. 154–159). Like Box and Jenkins (1976, p. 9), we shall drop the adjective,

and refer to asymptotically stable systems as stable. These systems have the attractive property of producing *stationary* time series as output. This simply means that the distribution of the output is time-invariant, i.e. $\mathbf{x}_t \sim \mathbf{x}_{t+n}$, a helpful property when investigating long-run performance. Stability ensues when the impulse response of a system comprises a finite number of periods, or if the impulse response covers infinite periods, but converges to zero (Box and Jenkins, 1976, p. 9). For further details about stability criteria, the reader may consult Luenberger (1979, pp. 154–159).

As an alternative to classic control theory, we may seek an optimal control policy that minimizes a penalty (cost) function. To find the optimal policy for continuous systems, we can use the Pontryagin minimum principle, or the Hamilton-Jacobi-Bellman equation, whereas for discrete-time systems, dynamic programming is the method of choice (Bertsekas, 2005).

A particular kind of policy is optimal when the system structures are linear, and the cost function is quadratic. By quadratic costs, we simply incur a penalty equal to the square difference between the actual outcome and the intended outcome. Under these conditions, the optimal policy is a linear function of the system states (Simon, 1956). The optimal policy is said to be a *linear quadratic regulator (LQR)*, and its identification depends on the solution of an algebraic Ricatti equation. This can be done directly, when we have perfect information about the system states. When there are Gaussian measurement errors, the optimal policy is an LQR combined with a Kalman filter. This summarizes the theory on linear systems. Next we will consider optimal policies for nonlinear systems.

Dynamic programming is a collection of techniques for finding an optimal sequence of controls, or an optimal policy, for a system whose state changes over time. The common denominator between these techniques is the reliance on the *principle of optimality*:

> "An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." (Bellman, 2003, p. 83).

This principle exploits that sequential decision-making problems can be divided into subproblems. When a number of decisions are sequentially contingent, we optimize the last decision, i.e. the decision that does not affect any other, and substitute this optimum in the original dynamic program. Thus, we have

eliminated one decision variable, and can continue to optimize one decision variable at a time until the optimization is complete.

The principle of optimality applies to both deterministic and stochastic problems, with both finite and infinite time horizons. For finite-time problems, the cost to be minimized is the sum of periodic costs, plus some terminal cost that depends on the final system state. The procedure can be numeric, or algebraic, depending on the problem. The difference between the two is that the analytical approach tries to identify the *structure* of an optimal policy, whereas the computational approach takes an initial condition, and identifies the optimal control as a set of numbers, or as a matrix. In the former case, with a known optimal structure, it is still necessary to identify the optimal parameter settings for the policy. In the latter case, the optimal numerical setting provides no explanation for its optimality.

Dynamic programs can also span an infinite time-horizon. Then it is common to use a discount factor, or an average-cost approach, to prevent infinite costs (Bertsekas, 2005, p. 403). In average-cost infinite-horizon problems, the initial state (and its transient effects), and the terminal cost tend not to matter, as they are averaged over an infinite time period.

Simulation modelling

Simulation models use equations or computer code to reproduce the behaviour of a system. In this way, we may see a simulation model as a digital *copy* of a real or imagined system. In contrast to analytical models, we must specify a set of numeric inputs that in turn provide a set of numeric outputs. Simulation may not produce as general results as a mathematical model would, making the connection between input and output numeric rather than analytic, but it allows complicated models, that might otherwise be intractable, to be built with relative ease. In addition, simulation models can be used to validate the results of analytical models, for both deterministic and stochastic systems.

Soft systems modelling

In contrast to mathematical modelling and simulation, soft systems modelling does not generate a quantitative model. Instead, it exploits systems thinking to help managers understand human organizations (Pidd, 2003, p. 115). Practice is emphasized over theory, and the means of modelling are diagrammatic, rather than quantitative (Mingers, 2011). Soft systems modelling deals with complicated problems (what Ackoff (1999) terms a *mess*), often involving

organizational and political aspects that may not be included in a quantitative model. As soft-systems modelling deals with specific real-world problems, it is not well suited to the abstract order cycle problem.

Choice of modelling approach

As modellers, we seek at the same time to present a rich picture of some phenomenon, general enough to cover a range of scenarios, yet specific enough to give useful results. There is a trade-off between these, and in many cases simplification may be warranted. As an example, the dynamics of physical systems are often linearized to become mathematically tractable (Nise, 2011, pp. 88–97). In the same way, we must select system structures and cost functions that are both realistic, and tractable. To obtain a balance between model realism and richness of the analytical results, the following assumptions are made:

1. *System structure.* We opt for a linear system structure. This makes it simple to calculate the statistical moments of output variables, and the covariance between variables. When the input signal is normally distributed, the output variables also have this distribution. By using a linear model, we cannot assume lost sales, nor can we assume that capacity is finite. Still, a linear model can approximate a physical system with sufficient safety stock and capacity. We must also assume that returns to suppliers are allowed, and that anything ordered from suppliers (or from production) is delivered in full.

2. *Costs.* When investigating the inventory dynamics only, we assume convex, piecewise-linear costs. For the case of capacity and inventory costs, we design an optimal policy for quadratic costs, which have been used to approximate the piecewise-linear inventory and capacity costs observed in industry (Holt et al., 1960). We also consider piecewise-linear inventory and capacity costs for four variations of a staggered policy that is used industrially.

3. *Demand.* Assumed to be normally distributed. When we investigate the inventory dynamics only, demand can have an arbitrary autocorrelation function. When we add capacity costs, demand is assumed to be independent and identically distributed.

These assumptions are reasonable in situations with high-volume products with low variability, and sufficient capacity. Other production settings may

require a different model. The particular benefits with this set of assumptions is that we can obtain fairly rich analytical results, in the form of costs and service level expressions.

3.3 Validation

It is not enough just to have a model. We must also know that it serves its purpose. To our aid, we have a range of validation techniques.

When a (physical) reference system exists, *black-box validation* can be used to test that the output of the model conforms with the reference system. Often, we can only compare a present-state model against a real system, but for models of a tentative future-state there is no real system against which to compare it. In these cases, we must demonstrate that the model operates as a physical system would operate, under the same circumstances, and for the same reasons. This reduces to demonstrating that the model's components and their interactions reflect those of the physical system. We call this validation of model internals *open-box validation* (Pidd, 2003).

Of the many validation techniques, those designed for System Dynamics models are of particular interest when we consider discrete-time systems. An extensive treatment can be found in Forrester and Senge (1978), where the tests cover three categories: model structure, model behaviour, and policy improvement. Some of these tests are described further by Sterman (2000, pp. 859–861). Based on Forrester and Senge (1978) and Sterman (2000), we shall consider the following tests, which have been selected because they are meaningful for analytical models:

- The model should be causal, exhibiting the same behaviour as the real system, for the same reasons.
- The model structure must be consistent with the descriptive knowledge of the system.
- The model must not break physical laws.
- The decision rules in the model must reflect the behaviour of the decision-makers.
- The units of the stocks and flows in the model must be consistent.
- The model should respond reasonably when parameters are set to extreme values.
- The model variables must be unambiguous and quantifiable.

- The model must be able to generate the same behaviour as related models.

This list of validation practices reveals that validation is just as much a qualitative problem, as it is a quantitative one.

Validation in this thesis

Mathematical models are functions, relating input and output variables without intermediate steps. They are black boxes. Following the derivation of a mathematical model from the start allows a certain qualitative understanding of the model to be developed. To achieve open-box validity, we build a discrete-time simulation model designed to address the same problem as the mathematical model. Not only is the simulation model easier to understand, but its results should also be comparable with those of the mathematical model, allowing one to validate progress each step of the way.

3.4 Ethics in operational research

OR frequently develops models describing how real systems should be configured, and how decisions should be made. As a result, OR models can have significant effects on the world. Apart from this are the ethical standards of people working with OR, including researchers and practitioners. This section explores both perspectives, starting with the first, and concludes with a statement of the ethical considerations taken when performing the research for this thesis.

3.4.1 Ethical considerations within models

An early influential textbook states the objective of OR:

> O.R. tries to find the best decisions relative to as large a portion of a total organization as possible (Churchman et al., 1957, p. 6).

Although primarily focused on economic performance on an organization-wide scale, Churchman et al. (1957, pp. 59–64) describes how worker attitudes, working conditions, and motivation can be important factors to consider in OR. On a more abstract level, Ackoff (1949) talks about ideals as the end goals of human pursuits. Based on this, Ackoff argues that one must define value criteria (objective functions) that indicate progress towards these ideals.

According to Brans (2002a), ethical decision-making in OR rests on three foundations: The first is *rationality*, where the entire decision problem is

reduced to a single-objective optimization problem. The second is *subjectivity*, where multiple fitness criteria are identified, and the decision-maker selects one of several Pareto-optimal solutions. The third foundation is *ethics*, where decision-makers appraise how people and the environment are affected by the potential solutions, and use this when determining what solution to implement. The foregoing approach is primarily used when the expected outcome of a decision is important, but there is always a risk that the actual outcome differs from the expected. To evaluate these situations Cairns et al. (2016) suggest the use of scenario planning in combination with an evaluation of stakeholder objectives.

3.4.2 Ethics in OR work

Apart from including ethical considerations in OR models, there is also the dimension of ethical OR practice. Brans (2002b) suggests a kind of Hippocratic oath for operational researchers in different roles. it dictates that:

- As a consultant, one should try to convince decision-makers to be ethical.
- As a decision-maker, one should consider the ecological, economic, and social consequences of ones actions.
- As a teacher, one should be honest, respect colleagues, and discuss the consequences of OR work.
- As a researcher, one should seek freedom of association, use suitable tools and methods, and realize that any work may be put to practical use.

A different perspective is provided by Diekmann (2013), who formulates four principles: *Transparency* — to provide full disclosure of models and assumptions; *integrity* — to follow professional and scientific standards, and not to distort work for personal gain; *comprehensiveness* — to consider all stakeholders and the moral implications of the work; and finally *efficacy* — to provide a detailed account of possible ethical consequences. Some of this is reflected by Gallo (2004), who states that all stakeholders must be taken into account, and that research outcomes, such as models and algorithms, should be made accessible to practitioners and the academic community.

3.4.3 Ethical considerations in this thesis

This thesis develops models, and later describes an industrial production system where staggered delvieries occur. In accord with the preceding discussion, we

shall consider how ethics relates to the models themselves, and to the research effort as a whole.

In the models, we assume the role of a planner being responsible for the MPS. Chapter 4 considers, a pure inventory model with the aim to minimize the sum of linear holding- and backorder costs, and the once-per-cycle audit cost. The model can be applied to a range of scenarios, including ones with products that are critical to human well-being, such as pharmaceuticals. When products are important in this sense, users of the model must set sufficiently high backorder costs, and they must also verify that the service levels are sufficiently high. Chapter 5 adds the complication of a capacity cost, and identifies an optimal capacity level. Improving a system following the recommendations made in the chapter may cause two changes for the workforce: First, a reduced amount of idling and overtime work, which leads to a more predictable schedule, but fewer man-hours of overtime worked per product; second, a changed capacity level, potentially affecting the staffing of the plant. This may lead to recruitment or redundancies, with consequences for individual workers.

Chapter 6 includes an industrial example outlining a production process where staggered deliveries occur. The investigation entailed to an interview with production planners, and followed the ESRC (Economic and Social Research Council) Framework for Research Ethics (ESRC, 2015, p. 4)

- Participation was voluntary, with no conflicts of interest between the researcher and the participants.
- The participants were informed of the purpose of the study, and their role therein.
- The study did not expose the participants to any safety risks.
- Anonymity was preserved for the organization, and for participating individuals.
- Transparency was achieved by providing a statement of the interview procedure (in Chapter 6).
- There were no conflicts of interest between the researcher and the participants, or the organizations involved.
- No personal information was sought or gained.

With this, we have shown how ethics has been treated with respect to the model and the industrial investigation.

3.5 Summary

We have concluded that a realist approach befits the research questions. The research strategy of choice is not induction or deduction, but retroduction. This entails to devising and testing a conceptual model that is believed to resemble a real-world phenomenon.

The conceptual model is tested by being implemented both as a mathematical model and as a simulation model. They are designed to be in exact correspondence to each other, so black-box validity can be tested by comparing the analytical and simulation results. Open-box validity is provided by the simulation model. This entails to checking that the difference equations match the rules provided by the conceptual model.

Chapter 4

Staggered order-up-to policy for autocorrelated demand

This chapter begins by presenting a model structure that realizes staggered deliveries. Then the optimal policy under autocorrelated demand is identified as a variant of the OUT policy. There are three important reasons for considering autocorrelated demand: First, it can describe stochastic processes with vastly different characteristics (Box and Jenkins, 1976). Second, demand for many products is significantly autocorrelated (Lee et al., 2000). Third, it introduces depth to the analysis, as non-trivial forecasts must be used. Expressions for costs, availability, and fill rate are provided, as well as a procedure for determining the optimal order cycle length. The variances of the system, most importantly the variance of the inventory level, are identified for generally autocorrelated demand, and for the specific cases of AR(1) and i.i.d. demand. Notable results include the optimality of time-varying safety stocks that set the availability to some fixed value, but where the fill rate fluctuates over the cycle. Constant safety stocks are suboptimal, and cause the availability to fluctuate.

4.1 Model description in a natural setting

To ease understanding, let us initially define the model in a weekly setting of seven days, where we plan once per week, but produce every day. Later, we generalize this to arbitrary planning periods of length P, but for now we consider the planning cycle to be seven days long.

Every morning the inventory level is tallied. If it is Monday, a production plan is made immediately after the inventory inspection. This production plan contains seven orders, to cover an entire week of production. This reflects

the planning cycle length $P = 7$. In more general terms, staggering deliveries means that we must determine the production rate for P days once every P periods. Between two such occasions, no new production plans are calculated, as we are committed to the established plan; Figure 1.1 illustrates this.

Let t number the individual days (periods), and let Mondays occur when when t/P is an integer. Suppose it is the start of Monday morning, and that we must plan the orders for the next cycle $\{o_{t,1}, o_{t,2}, \ldots, o_{t,7}\}$, numbered in the same sequence as they will be produced. Every order o corresponds to a future inventory receipt r, according to

$$o_{t,k} = \begin{cases} r_{t+\tau} = r_{t+k+L} & \text{when } t/P \in \mathbb{Z} \\ \varnothing & \text{otherwise;} \end{cases} \tag{4.1}$$

where $k \in \{1, 2, \ldots, P\}$ is the order release offset due to staggering, L is a non-negative integer lead time, and $\tau = k + L$ is the effective lead time. In this case, $L = 4$, meaning that these orders will register as received in the periods $\{t + 5, t + 6, \ldots, t + 11\}$, as Figure 4.1 illustrates. The information available when determining $o_{t,k}$ is i_t, all past demand observations up to and including d_t, as well as all previously planned receipts $\{\ldots, r_{t+k+L-2}, r_{t+k+L-1}\}$.

Immediately when a cycle's order quantities have been determined, they are sent to production scheduling, where the required receipt rates r_{t+k+L} are disaggregated into a detailed schedule of individual jobs. Each job then leads to the production of one or more lots. Our model does not place any specific restriction on jobs, except that they must be released so that the production completions between $t+\tau-1$ and $t+\tau$ equal $r_{t+\tau}$. From a production planning / inventory modelling perspective, we need not know the exact timing of job releases, as long as the planned quantity arrives in the right period. The lead time L reflects the time required to effect a new production plan, including the time to schedule production, to allocate resources, and to produce the goods. Note that if $L = 0$, the first order would be released to production immediately, and be completed in less than one day. It would therefore contribute to the inventory level measured at time $t + 1$.

The receipts resulting from the staggered deliveries are placed in inventory. We assume that there is no shrinkage, and that the inventory level increases with receipts (r_t) and decreases with demand (d_t),

$$i_t = i_{t-1} + r_t - d_t. \tag{4.2}$$

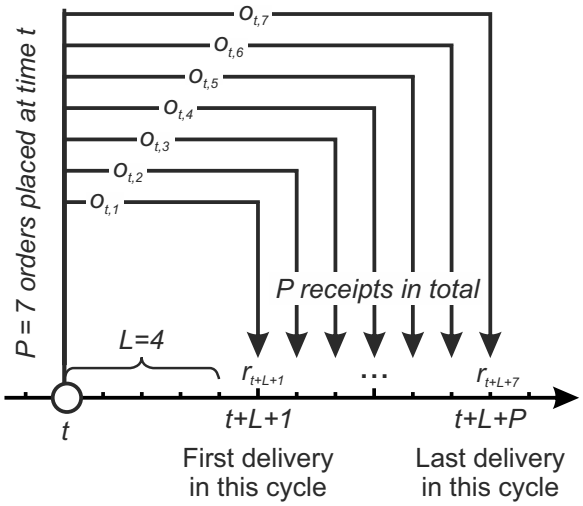

Figure 4.1: Staggering means that even though we plan intermittently, we receive orders every period.

If there are no goods on-hand, the excess demand is backlogged and subtracted from the inventory level; then the inventory level falls below zero. Backlogged demand is satisfied immediately when new goods are received. A sequence of events consistent with this description can be described as follows:

1. At the very beginning of period t, observe the inventory level i_t.
2. Immediately thereafter, place all orders $o_{t,1}$ through $o_{t,P}$ (in sequence) if t marks the start of an order cycle, otherwise do not order.
3. After any orders are placed, receive r_{t+1}.
4. Thereafter, the final event in period t is to satisfy demand d_{t+1}.
5. After demand has been subtracted from inventory, time advances to period $t + 1$, starting with an inspection of the inventory level i_{t+1}.

We assume that the system is linear, therefore negative orders are permitted, reflecting costless returns (in distribution scenarios), or that goods are sold off at a price equalling the variable cost of production (in manufacturing scenarios). Negative demand is also permitted, reflecting returns from customers to our inventory. Both the storage and the production facilities have unlimited capacity.

4.2 The optimal ordering rule

The inventory costs are incurred once per period, and consist of a holding cost h per unit of inventory, and a backorder cost b per unit of unsatisfied demand, which we model as negative inventory,

$$j(i_t) = h\,(i_t)^+ + b\,(-i_t)^+ = h\,i_t + (b+h)\,(-i_t)^+ . \tag{4.3}$$

where $(x)^+ = \max(x, 0)$. Let the lead-time demand be denoted as $f_{t,\tau} = \sum_{n=1}^{\tau} d_{t+n}$, and the WIP as $w_{t,\tau} = \sum_{n=1}^{\tau-1} r_{t+n}$; then the inventory level can be expressed as $i_{t+\tau} = i_t + w_{t,\tau} + o_{t,k} - f_{t,\tau}$, when $P/t \in \mathbb{Z}$. The single-period cost $j(i_{t+\tau})$ is thus convex in $o_{t,k}$ due to (4.2). Taking the expectation gives

$$J(i_{t+\tau}) = \mathbb{E}\left[j(i_{t+\tau})|d_t, d_{t-1}, \cdots\right], \tag{4.4}$$

where $J(i_{t+\tau})$ is the expected inventory cost in period $t+\tau$. Note that the expectation is a linear operator, and as such, it preserves convexity. Therefore, $J(i_{t+\tau})$ is convex in $o_{t,k}$. As $o_{t,k}$ can be set freely, the influence of $i_t + w_{t,\tau}$ can be removed from $i_{t+\tau}$, meaning that $J(i_{t+\tau})$ depends only on the decision variable $o_{t,k}$, and on the lead-time demand $f_{t,\tau}$. Therefore, the optimal order policy is myopic, meaning that the optimal solution for an n-period problem can be found by solving n independent single-period problems (Heyman and Sobel, 1984, p. 63-71). In practice, we need only to consider the immediate consequences of each decision, as it has no bearing on the cost incurred in other (future) periods.

The expected inventory cost (4.4) is convex, and a minimum exists because both b and h are positive. Therefore, there exists an optimal expected inventory level that minimizes $J(i_{t+\tau})$. This optimum is referred to as the *safety stock*, $i^*_{t+\tau}$. It is identified in the following lemma.

Lemma 4.1 (The optimal safety stock level)**.** *The expected inventory cost, $J(i_{t+\tau})$, is minimized when*

$$i^*_{t+\tau} = \Phi^{-1}_{i_{t+\tau}}\left(\frac{b}{b+h}\right), \tag{4.5}$$

where $\Phi^{-1}_{i_{t+\tau}}$ is the inverse of the inventory level's cumulative distribution function at time $t+\tau$.

Proof. The expected inventory cost J is structurally identical to the single-period newsvendor problem, with b and h representing the underage and overage

costs. See Theorem A.1 in Appendix A for a proof. ■

With $i^*_{t+\tau}$ known, all that remains is to specify the policy that sets $i^*_{t+\tau} = \mathbb{E}\left[i_{t+\tau}|d_t,\, d_{t-1}\,\cdots\right]$. This is done in the following theorem.

Theorem 4.2. *When there are P staggered orders per cycle, the expected inventory cost is minimized by the policy*

(a) for the first order in a cycle, when $k = 1$,

$$o_{t,1} = \hat{f}_{t,L+1} + i^*_{t+L+1} - i_t - w_{t,L+1}, \tag{4.6}$$

where $\hat{f}_{t,\tau} = \mathbb{E}\left[f_{t,\tau}|d_t,\, d_{t-1}, \cdots\right]$ is the forecast of lead-time demand;

(b) for all subsequent orders in the cycle, when $k > 1$,

$$o_{t,k} = i^*_{t+\tau} - i^*_{t+\tau-1} + \hat{d}_{t,\tau}, \tag{4.7}$$

where $\hat{d}_{t,\tau} = \mathbb{E}\left[d_{t+\tau}|\{d_t, d_{t-1},\, ...\}\right]$ is the single-period forecast, made at time t, for $d_{t+\tau}$.

Proof. Given in Appendix B.1. ■

The policy can be interpreted as an OUT policy with increasing lead times over the cycle and a simplified ordering rule for all periods but the first. Note that our simplified rule, (4.7), is different from the simplified rule in Chiang (2009), who assumes that $o_{t,2} = o_{t,3} = \cdots$. Instead, we order the single-period forecast of demand plus any desired change in safety stock, $i^*_{t+\tau} - i^*_{t+\tau-1}$. When $P = 1$ the policy simplifies to the regular OUT policy.

4.2.1 Demand specification

To gain further insights about the optimal policy and its dynamic performance, we shall assume that demand is autocorrelated,

$$d_t = \mu + \sum_{n=0}^{\infty} \varepsilon_{t-n}\theta_n, \tag{4.8}$$

where θ_n is the autocorrelation function, $\mu = \mathbb{E}[d_t]$, and ε_t is an independent and identically distributed (i.i.d.) random variable drawn from the normal distribution. We call ε_t the *error term*. It has a mean of zero and a variance of σ^2_ε. The mechanism of such demand processes is well documented in Box and

Jenkins (1976), where they are described as moving-average processes. One important property of this type of demand signal is that its variance can be obtained as $\text{var}(d_t) = \sigma_\varepsilon^2 \sum_{m=0}^{\infty} \theta_m^2$.

The nature of i_t^* can be determined from three facts: The system is assumed to be linear, the policy given in (4.6) and (4.7) is linear, and demand is assumed to be normally distributed. Taken together, they imply that the inventory levels are normally distributed, which we will later verify in (B.10). Because the inventory level follows a normal distribution, the mean and variance are sufficient to specify the inventory distribution. The mean inventory can be set to i_t^* with r_t, but the inventory variance is a function of τ and the demand process. Before identifying the inventory variance required to calculate i_t^*, let us define the service levels.

4.2.2 Service levels

Not only does the optimal safety stock minimize the total cost of the system, it also sets the system's availability (Silver et al., 1998) to the critical ratio $b/(b+h)$. The availability (α), or type 1 service level, refers to the probability of not encountering a stock-out in any given period,

$$S_1 = \mathbb{P}(i_t \geq 0). \tag{4.9}$$

The fill rate (S_2), or type 2 service level, is sometimes considered a more appropriate measure in customer-facing settings, as it measures the fraction of demand fulfilled immediately from stock (Johnson et al., 1995). As the exact fill rate in (Johnson et al., 1995) is restricted to normal i.i.d. demand, it was necessary to develop a fill rate expression that copes with autocorrelated demand. Appendix C makes explicit the limitation of the conventional fill rate definition. The exact fill rate when demand is autocorrelated and possibly negative is of the following form:

$$S_2^- = \frac{\mathbb{E}\left\{[\min(d_t, i_t + d_t)]^+\right\}}{\mathbb{E}\left[(d_t)^+\right]}. \tag{4.10}$$

This exact fill rate takes the expectation of immediately satisfied demand, and divides it by the expected positive demand. This works well when demand is stationary, the fill rate is undefined for nonstationary demand. Although the fill rate definition does not permit nonstationary demand, it is possible to obtain an analogous measure if one measures the fill rate over a finite time period, and

if the initial conditions of the system are known (Strijbosch et al., 2011). Here, we do not pursue this path, but limit the scope to the conventional fill rate, which is obtained as follows. If the variables in (4.10) are normally distributed, we can obtain the fill rate via the following lemma.

Lemma 4.3. *The exact fill rate for normally distributed demand, where periods with negative demand do not contribute to the fill rate, is*

$$S_2^- = \frac{\int_{x=0}^{\infty} \varphi^-(x) x \, dx}{\sigma(d_t)\, g\left[-\mu/\sigma(d_t)\right]}. \tag{4.11}$$

Here $\sigma(d_t) = \sqrt{\text{var}(d_t)}$ is the standard deviation of d_t, $\varphi^-(x)$ is the probability distribution function (pdf) of the minimum of the normally distributed bivariate random variables d_t and $(d_t + i_t)$. $g\,(x) = \varphi\,(x) - x\,[1 - \Phi\,(x)]$ is the standard normal loss function, where $\varphi\,(x)$ is the standard normal pdf, and $\Phi\,(x)$ is the standard normal cumulative density function (Axsäter, 2006, p. 91).

Proof. This follows directly from (4.10) and the assumption of a linear system. ■

Remark. The pdf of the minimum of bivariate random normal variables, $\varphi^-(x)$, is given in Cain (1994), as $\varphi^-(x) = \varphi_1^-(x) + \varphi_2^-(x)$, where

$$\varphi_1^-(x) = \frac{\varphi\left(\frac{x-\mathbb{E}[i_t+d_t]}{\sigma(i_t+d_t)}\right)}{\sigma(i_t+d_t)}\ \Phi\left[\frac{\rho\left(\frac{x-\mathbb{E}[i_t+d_t]}{\sigma(i_t+d_t)}\right) - \frac{x-\mu}{\sigma(d_t)}}{\sqrt{1-\rho^2}}\right], \tag{4.12}$$

$$\varphi_2^-(x) = \frac{\varphi\left(\frac{x-\mu}{\sigma(d_t)}\right)}{\sigma(d_t)}\ \Phi\left[\frac{\rho\left(\frac{x-\mu}{\sigma(d_t)}\right) - \frac{x-\mathbb{E}[i_t+d_t]}{\sigma(i_t+d_t)}}{\sqrt{1-\rho^2}}\right], \tag{4.13}$$

where the correlation coefficient is

$$\rho = \frac{\text{cov}\,(i_t + d_t, d_t)}{\sqrt{\text{var}(i_t + d_t)\text{var}(d_t)}}. \tag{4.14}$$

It is often necessary to evaluate (4.11) numerically. This is usually done with software like Mathematica or Matlab, but it can also be achieved with Microsoft Excel using the macro provided in Disney et al. (2015).

4.2.3 Identifying the variances of the system

To calculate the exact fill rate we must know the variances $\text{var}(i_t)$ and $\text{var}(i_t+d_t)$, and also the correlation coefficient ρ. For autocorrelated demand, we identify these according to Theorem 4.4.

Theorem 4.4. *If planning took place at time $t-\tau$, the characteristics of the inventory level are as follows:*

(a) The inventory variance is

$$\operatorname{var}(i_t) = \sigma_\varepsilon^2 \sum_{n=0}^{\tau-1} \left(\sum_{m=0}^{n} \theta_m \right)^2 . \tag{4.15}$$

(b) The variance of $i_t + d_t$ is

$$\operatorname{var}(i_t + d_t) = \sigma_\varepsilon^2 \left\{ \left[\sum_{m=1}^{\tau-1} \left(\sum_{n=0}^{m-1} \theta_n \right)^2 \right] + \sum_{x=\tau}^{\infty} \theta_x^2 \right\} . \tag{4.16}$$

(c) The covariance between demand, d_t, and $i_t + d_t$, is

$$\operatorname{cov}(d_t, i_t + d_t) = \sigma_\varepsilon^2 \left\{ \left[\sum_{n=1}^{\tau-1} \left(- \sum_{m=0}^{n-1} \theta_m \right) \theta_n \right] + \sum_{x=\tau}^{\infty} \theta_x^2 \right\} . \tag{4.17}$$

Proof. Presented in Appendix B.2. ■

The inventory variance (4.15) increases in τ, regardless of θ_t, and is finite for all demands, stationary or nonstationary. The variance of the state variable $i_t + d_t$ is also increasing in τ, but is only finite for stationary demand. The covariance (4.17) between demand and initial inventory exists only for stationary demand. The main insight from (4.15) is that the inventory variance increases over the cycle. As inventory costs are minimized when $\mathbb{P}(i_t \geq 0) = b/(b+h)$, we find a *time-varying* safety stock to be optimal. This safety stock is increasing in τ. It is also clear from (4.15) that autocorrelation can amplify or attenuate inventory heteroskedasticity. A heteroskedastic time series is one where the variance changes over time. It is worth noting that this property appears when sampling the inventory level over a range of consecutive periods $\{i_t, i_{t+1}, ..., i_{t+n}\}$, but not when sampling only for a specific τ, i.e. $\{i_t, i_{t+P}, ..., i_{t+nP}\}$. Then the heteroskedasticity disappears.

4.2.4 Total cost and the optimal order cycle length

Under normally distributed demand and linear transformations, the inventory level is also normally distributed. The expected inventory cost is

$$\begin{aligned} J(i_{t+\tau}) &= h\mathbb{E}[i_{t+\tau}] - \frac{b+h}{\sigma_{i,k}} \int_{-\infty}^{0} \varphi\left(\frac{x - \mathbb{E}[i_{t+\tau}]}{\sigma_{i,k}}\right) x\, dx \\ &= h\mathbb{E}[i_{t+\tau}] + (b+h)\sigma_{i,k}\, g\left(\frac{\mathbb{E}[i_{t+\tau}]}{\sigma_{i,k}}\right), \end{aligned} \tag{4.18}$$

where $\mathbb{E}[i_{t+\tau}]$ denotes the safety stock, and $\sigma_{i,k} = \sqrt{\text{var}(i_{t+k+L})}$. As the error terms are i.i.d., $J(i_t) = J(i_{t+P})$. Therefore, the average cost is obtained by averaging over P successive periods. When the optimal safety stocks $i^*_{t+\tau}$ are used, the average cost from (4.18) simplifies to

$$J^*_P = \frac{1}{P}\sum_{k=1}^{P} J(t+k) = \bar{\sigma}_{i,P}(b+h)\varphi\left[\Phi^{-1}\left(\frac{b}{b+h}\right)\right], \tag{4.19}$$

where $\bar{\sigma}_{i,P} = P^{-1}\sum_{k=1}^{P}\sigma_{i,k}$ is the average standard deviation of the inventory level. This variable is essential for characterizing P^*.

Consider a fixed audit cost per cycle, v, leading to an average audit plus inventory cost per period of $C_P = J^*_P + v/P$. Let $\lambda = v/\psi$ where

$$\psi = v + (b+h)\varphi\left[\Phi^{-1}\left(\frac{b}{b+h}\right)\right]. \tag{4.20}$$

The total cost can then be expressed as a linear function of $\lambda \in [0,1]$,

$$C_P(\lambda) = \psi\left[\bar{\sigma}_{i,P} + \lambda\left(P^{-1} - \bar{\sigma}_{i,P}\right)\right]. \tag{4.21}$$

With this formulation, it is possible to find a cost combination λ_P for which P minimizes the total cost.

Theorem 4.5. *When $\sigma_{i,P}$ is increasing in P, the order cycle length P minimizes $C_P(\lambda)$ for $\lambda \in [\lambda_{P-1}, \lambda_P]$ where $\lambda_0 = 0$ and*

$$\lambda_P = 1 - \frac{1}{1 + P\left(\sigma_{i,P+1} - \bar{\sigma}_{i,P}\right)}. \tag{4.22}$$

Proof. Let λ_P denote the intersection $C_P(\lambda_P) = C_{P+1}(\lambda_P)$. Solving for λ_P provides (4.22). Suppose λ_P is increasing in P. Then, as $P = 1$ minimizes the cost for $\lambda \in [0, \lambda_1]$, the reorder period P minimizes $C_P(\lambda)$ for $\lambda \in [\lambda_{P-1}, \lambda_P]$.

To see that λ_P is increasing in P, recall that $\sigma_{i,P} \leq \sigma_{i,P+1}$. This provides

$$P\sigma_{i,P+1} - \sum_{k=1}^{P} \sigma_{i,k} \leq (P+1)\,\sigma_{i,P+2} - \sum_{n=1}^{P+1} \sigma_{i,n}, \tag{4.23}$$

which leads to $P\,(\sigma_{i,P+1} - \bar{\sigma}_{i,P}) \leq (P+1)\,(\sigma_{i,P+2} - \bar{\sigma}_{i,P+1})$. This is equivalent to $\lambda_P \leq \lambda_{P+1}$, completing the proof. ■

This procedure lets us specify a P and provides the range of cost configurations $[\lambda_{P-1}, \lambda_P]$ for which this P is optimal. Through this indirect approach, several properties of P^* are revealed: it is increasing in the audit cost v, and it is decreasing in the factors that drive inventory cost, namely b, h, L, and σ_ε. This follows from the influence of (4.15) and (4.19) on λ. Furthermore, given a cost balance λ and an arbitrary P, it is immediately clear if $P < P^*$, $P = P^*$, or $P > P^*$, as λ_P is increasing in P. These observations hold for generally autocorrelated demand as (4.15) is increasing in τ.

To find P^*, it is sufficient to identify two values P_1 and P_2, such that $\lambda_{P_1} \leq \lambda \leq \lambda_{P_2}$; a binary search between these values then provides the optimum. As an alternative, we may plot the first few λ_P, and then seek P^* graphically. This simpler approach does not guarantee that P^* will be in the range plotted, but it is nonetheless reasonable when the audit cost is moderate in relation to the inventory cost.

4.3 The optimal policy for first-order autoregressive demand

To better understand the model, it is helpful to consider a simple case. Here we choose the AR(1) demand process. Its autocorrelation function is determined by a single parameter, ϕ; AR(1) time series are stationary and invertible for $|\phi| < 1$ (Box and Jenkins, 1976). The corollary below provides necessary expressions for calculating inventory costs, availability, the fill rate, and λ_P.

Corollary 4.6 (AR(1) demand)**.** *Using $\theta_m = \phi^m$ as the autocorrelation function of demand in Theorem 4.4, we obtain the following variance expressions:*

(a) The inventory variance,

$$\operatorname{var}(i_t) = \sigma_\varepsilon^2 \left[\frac{\tau}{(\phi-1)^2} + \frac{\phi(\phi^\tau - 1)(\phi^{\tau+1} - \phi - 2)}{(\phi+1)\,(\phi-1)^3} \right]. \tag{4.24}$$

(b) The variance of $i_t + d_t$,

$$\text{var}\,(i_t + d_t) = \sigma_\varepsilon^2 \left[\frac{\tau}{(\phi-1)^2} - \frac{2\phi^\tau}{(\phi-1)^3} + \frac{1+\phi\,(2-(\phi-2)\phi^{2\tau})}{(\phi-1)^3\,(\phi+1)} \right]. \tag{4.25}$$

(c) The covariance between $i_t + d_t$ *and* d_t,

$$\text{cov}(d_t, i_t + d_t) = \sigma_\varepsilon^2 \left[\frac{(\phi+1)\phi^\tau - \phi - \phi^{1+2\tau}}{(\phi-1)^2\,(\phi+1)} \right]. \tag{4.26}$$

Knowing the inventory variance for AR(1) demand (4.24), we have sufficient information to compute the optimal order quantities for each period in the planning cycle.

4.3.1 Determining the production quantities

To understand how this policy works in practice, consider the following numerical example using the AR(1) demand process.

Example 4.7. Consider the staggered system in Figure 4.1, where $L = 4$ and $P = 7$. The current period is $t = 0$, and we are ordering for the receipts in periods 5 through 11. Demand is known to be first-order autocorrelated (i.e. $\theta_n = \phi^n$) with $\phi = 0.70$, with a mean $\mu = 10$, and error terms that are normally distributed $\varepsilon \sim \mathcal{N}(0, 1)$. We have also observed that the inventory level is $i_0 = 5.20$, and that the work-in-progress inventory is $w_{0,5} = 41.30$. Adding these we obtain the current inventory position $i_0 + w_{0,5} = 46.50$. The optimal order quantities are calculated as follows:

1. Make the *lead-time demand forecast* for the first period. As demand is autocorrelated, our last demand observation, $d_0 = 8.71$, is sufficient to make a forecast of lead time demand $\mu(L+1) + (d_0 - \mu)\sum_{n=1}^{L+1}\phi^n = 10 + (8.71 - 10) \times 1.94117 = 47.5$.

2. Make the *single period forecasts* for the remaining periods, $k = 2$ to $k = 7$, or equivalently $\tau = 6$ to $\tau = 11$. We obtain this as $f_{t,\tau} = \mu + (d_t - \mu)\phi^\tau$, which for the second order of the cycle gives $f_{t,6} = 10 + (8.71 - 10) \times (0.7)^6 = 9.85$. The remaining periods are obtained in the same way, after incrementing τ.

3. Calculate the *time-varying safety stocks*. These are of the form $i^*_{t+\tau} = \sigma_{i,k}\Phi^{-1}\,[b/\,(b+h)]$, where $\sigma_{i,k}$ is the square root of the inventory variance

found in (4.24). Thus, the first safety stock for $k = 1$ is $i_5^* = \sqrt{22.7923} \times \Phi^{-1}(0.9) = 6.12$. For the following periods, increment k and perform the calculation again. For example $i_6^* = \sqrt{31.4428} \times \Phi^{-1}(0.9) = 7.19$.

4. Determine the *safety stock increase* between periods. Starting with $k = 2$, this is done by the subtraction $i_{t+\tau}^* - i_{t+\tau-1}^*$. The first change in safety stock, occurring at $\tau = 6$ is $i_6^* - i_5^* = 7.19 - 6.12 = 1.07$. The remaining safety stock increases are obtained by incrementing τ.

5. Calculate the *first receipt* according to the standard OUT policy. We order the lead-time forecast of demand, plus the safety stock, minus the inventory position, according to $\hat{d}_{0,5} + i_5^* - (i_0 + w_{0,5}) = 47.5 + 6.12 - 46.5 = 7.12$.

6. Calculate the *remaining receipts* using a simpler formula. The second receipt of the cycle, with $\tau = 6$, takes the single-period forecast, plus the increase in safety stock, $r_6 = \hat{d}_{0,6} + i_6^* - i_5^* = 9.85 + 1.07 = 10.92$. The remaining receipts of the cycle are calculated in the same way, with τ incremented.

Table 4.1 presents the optimal order quantities for the entire cycle, as well as the intermediate results. Contrary to the worked example, the table has been calculated with machine precision, so the last decimal of the calculations may vary.

4.3.2 Cost and service implications

Now consider the case where a plan made at time t will generate its first receipt in time for it to affect i_{t+1} (that is, $L = 0$). The previous setting $L = 4$ has been substituted for $L = 0$ to highlight the effects of inventory heteroskedasticity.

Figure 4.2 illustrates the inventory standard deviation for a range of AR(1) demands. The configurations where $|\phi| < 1$ reflect stationary demand, while other configurations reflect nonstationary demand. In either case, the inventory level is stationary. As we can see from (4.15) the inventory standard deviation is increasing in τ. The consequences of this are clear: staggering increases the total inventory cost, particularly when there is significant autocorrelation. Staggering is least harmful when demand is negatively autocorrelated and stationary $(-1 < \phi < 0)$.

Corollary 4.8. *Some special cases of the first-order autoregressive inventory variance can be identified.*

Table 4.1: Calculating the planned receipts for the numerical example.

Period (t)	0	1	2	3	4	5	6	7	8	9	10	11
k	3	4	5	6	7	1	2	3	4	5	6	7
τ	7	8	9	10	11	5	6	7	8	9	10	11
Inventory level (i)	5.2	-	-	-	-	-	-	-	-	-	-	-
Work-in-progress (w)	41.3	-	-	-	-	-	-	-	-	-	-	-
Demand (d)	8.71	-	-	-	-	-	-	-	-	-	-	-
Lead-time forecast[a] ($\hat{f}$)	-	-	-	-	-	47.50	-	-	-	-	-	-
Single-period forecast ($\hat{d}$)	-	-	-	-	-	-	9.85	9.89	9.93	9.95	9.96	9.97
Safety stock required (i^*)	-	-	-	-	-	6.12	7.19	8.19	9.12	10.00	10.83	11.61
Change in safety stock	-	-	-	-	-	-	1.07	1.00	0.94	0.88	0.83	0.78
Planned receipts (r)	-	-	-	-	-	7.12	10.92	10.89	10.86	10.83	10.79	10.76

Dashes (-) refer to values not needed for calculating the order quantities in this cycle.
[a] Forecasted demand over the lead time.

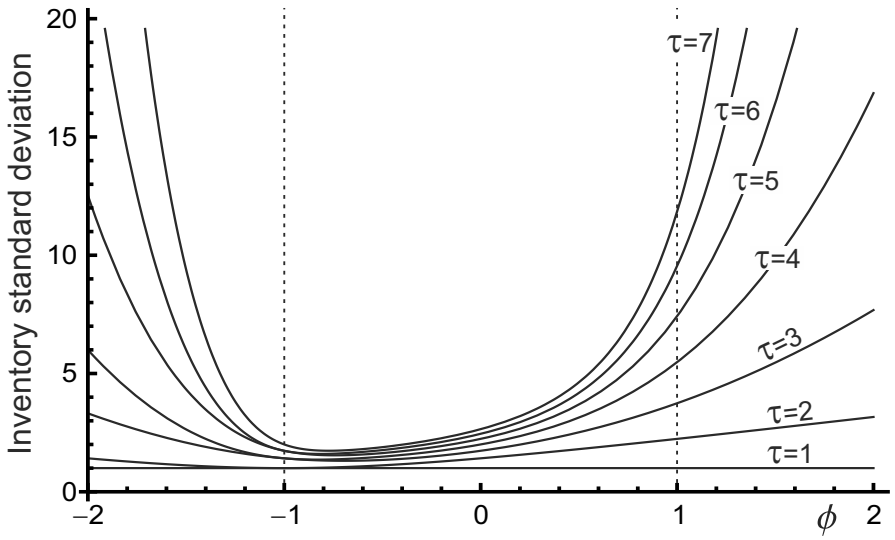

Figure 4.2: The inventory standard deviation for AR(1) demand is increasing in τ.

(a) When $\phi \to 0$ the inventory variance is a linear function of τ,

$$\mathrm{var}(i_t) = \sigma_\varepsilon^2 \tau. \tag{4.27}$$

(b) When $\phi \to 1$ demand is a random walk in discrete time (Box and Jenkins, 1976, p. 123), and the inventory variance is a cubic function, increasing in τ,

$$\mathrm{var}(i_t) = \sigma_\varepsilon^2 \frac{\tau}{6}(1+\tau)(1+2\tau). \tag{4.28}$$

This expression is the variance of the error term, multiplied with a square pyramidal number.

(c) When $\phi \to (-1)$ the inventory variance is an increasing affine function for odd or even values of τ,

$$\mathrm{var}(i_t) = \sigma_\varepsilon^2 \left[\frac{1-(-1)^\tau}{4} + \frac{\tau}{2} \right]. \tag{4.29}$$

From (4.29), we see that the inventory variance only increases strictly for odd values of τ when $\phi = (-1)$. Consequently, when the lead time, L, is even, $\mathrm{var}(i_{t+L+1}) = \mathrm{var}(i_{t+L+2})$, $\mathrm{var}(i_{t+L+3}) = \mathrm{var}(i_{t+L+4})$ and so forth. For odd L the pattern starts with $\mathrm{var}(i_{t+L+2}) = \mathrm{var}(i_{t+L+3})$.

Equation (4.27) reveals that when demand lacks autocorrelation, the variance increases linearly, meaning that the inventory standard deviation is pro-

portional to the square root of τ. This is a fundamental result, demonstrating that the staggered policy behaves like an OUT policy that is iterated, with increasing lead times for each order in the cycle. The inventory variance increases step-wise in (4.29). This is noticeable in the inventory standard deviation of Figure 4.2, where for $\phi = (-1)$ every curve coincides with another curve (except for $\tau = 7$, which would coincide with $\tau = 8$ if $P \geq 8$).

Consider the costs $b = 1$ and $h = 9$. They imply that the inventory costs are minimized when the availability is 90%, for every τ. As we know, the inventory variance changes with τ, and hence time-varying safety stocks are optimal. If we insist on using constant safety stocks, the availability will change over the cycle. For example, a constant safety stock can be based on the worst-case inventory variance, obtained at the end of the cycle, providing $i^*_{t+k+L} = \sigma_{i,P}\Phi^{-1}(0.9)$. This is not cost-optimal, but it simplifies the order quantity calculations. The results of this alternative strategy can be seen in Figure 4.3, where the availability, α, is given by

$$\alpha = \Phi\left(\frac{i^*_t}{\sigma_{i,k}}\right). \tag{4.30}$$

For any constant safety stock setting, availability degrades as τ increases. This is due to $\sigma_{i,k}$ being increasing in k. For the safety stock setting under discussion, $\sigma_{i,P} \geq \sigma_{i,k}$, making the availability lower-bounded at 90%.

A more sophisticated constant safety stock setting could be based on the average inventory variance,

$$i^*_t = \Phi^{-1}\left(\frac{b}{b+h}\right)\sqrt{P^{-1}\textstyle\sum_{n=1}^{P}\sigma^2_{i,n}}. \tag{4.31}$$

This safety stock setting is obtained if one ignores the cyclical heteroskedasticity and takes the variance of the inventory process as a whole. The resulting availability is shown in Figure 4.4, illustrating that the target availability of 90% is no longer a lower bound. On average, however, the availability is above the target of 90%. This always results when $h < b$ and when $i^*_t \geq \Phi^{-1}\{\bar{\sigma}_{i,P}\Phi[b/(b+h)]\}$ (which for this strategy is true, due to Jensen's inequality) as a consequence of Corollary 4.9.

Corollary 4.9. *For fixed safety stocks $i^*_t \geq 0$,*

$$\Phi\left(\frac{i^*_t}{\bar{\sigma}_{i,P}}\right) \leq \frac{1}{P}\sum_{k=1}^{P}\Phi\left(\frac{i^*_t}{\sigma_{i,k}}\right); \tag{4.32}$$

for $i_t^* \leq 0$ the inequality is reversed.

Proof. Observe that $\Phi(x)$ is concave for $x \geq 0$. Then (4.32) is an immediate result of Jensen's inequality. On the domain $x \leq 0$, $\Phi(x)$ is convex, and the inequality in (4.32) is reversed. This completes the proof. ■

The takeaway from this corollary is that the availability estimate $\hat{\alpha} = \Phi\left(i_t^*/\bar{\sigma}_{i,P}\right)$ is less than the realized availability, i.e. $\hat{\alpha} \leq \alpha$, when $\hat{\alpha} \geq 0.5$, or equivalently when $h \leq b$. Conversely, when $\hat{\alpha} < 0.5$, $\hat{\alpha}$ overestimates α.

The cost differential between these three strategies is worth considering. Figure 4.5 verifies that the optimal time-varying safety stock outperforms all constant settings. The worst economic performance results from the constant safety stock setting based on the end-of-cycle inventory variance. This is clear for nonstationary demand, but when there is little autocorrelation, the two fixed safety stock strategies are nearly cost-equal.

To see if these observations hold under different cost settings, we may consult Figure 4.6, where b and h assume different values, but in all cases $b + h = 10$. These settings imply an optimal availability of 60%, 90%, 95%, or 99%. Regardless of the cost configuration, we notice that the demand autocorrelation drives the cost differential between the constant and the time-varying safety safety stock settings. This effect appears for all of the cost settings, particularly when demand is nonstationary. In the $b = 9.9, h = 0.1$ setting, corresponding to an optimal availability of 99%, the superior performance of

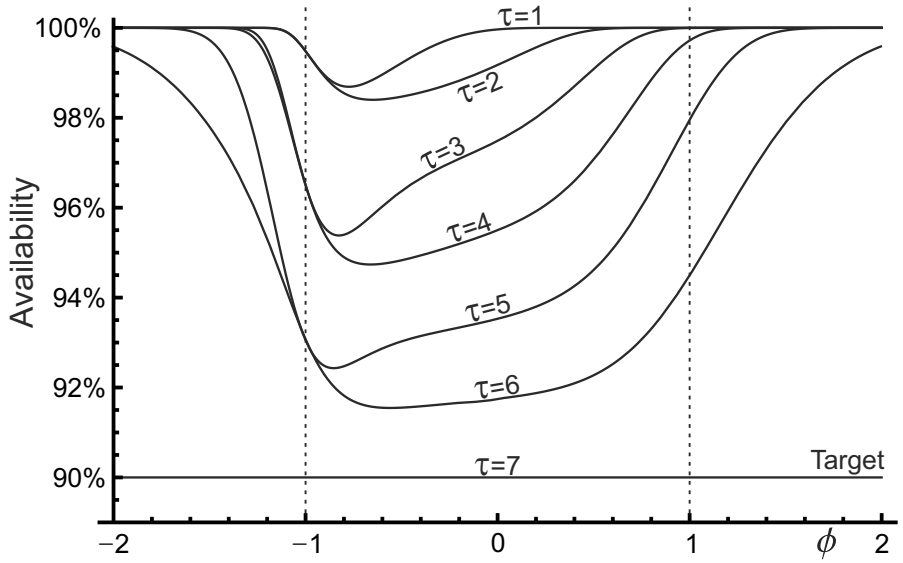

Figure 4.3: Availability for a fixed safety stock based on the inventory variance at the end of the order cycle.

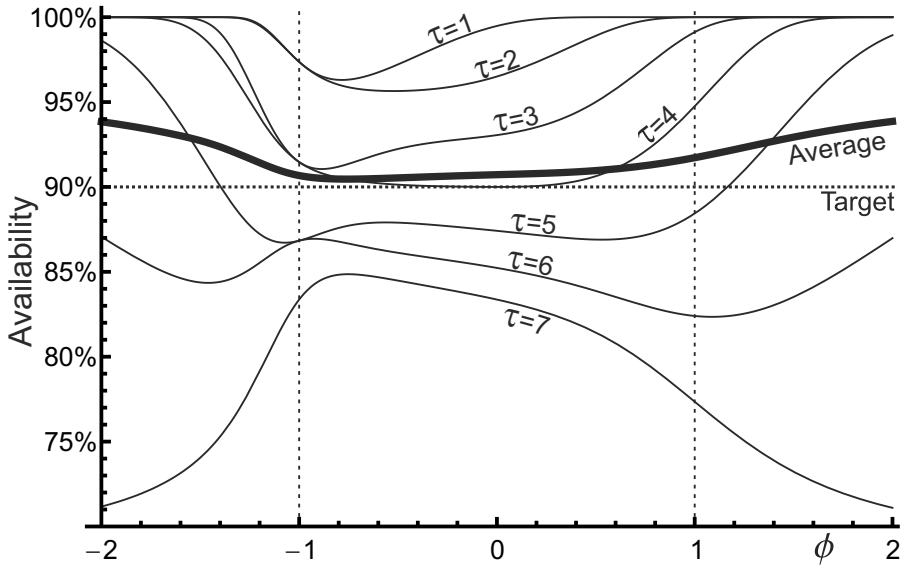

Figure 4.4: Availability for a fixed safety stock based on the average inventory variance (4.31).

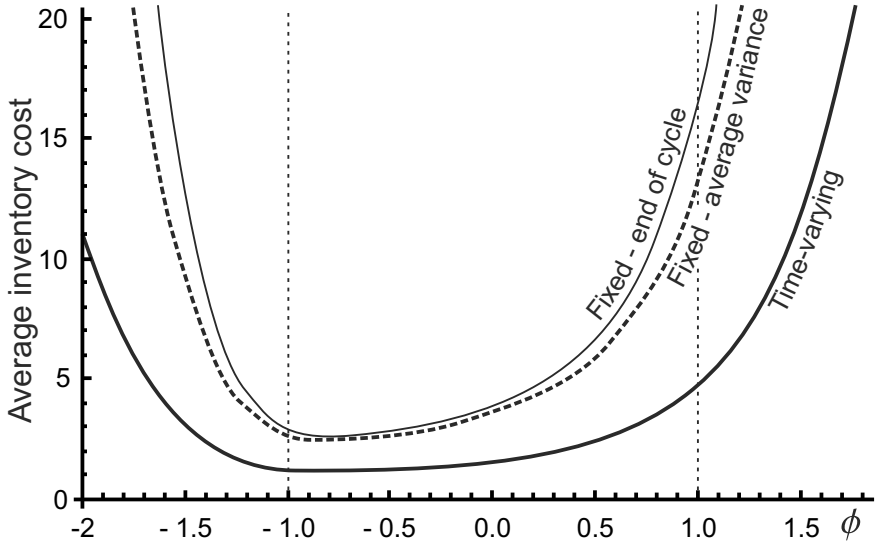

Figure 4.5: Average inventory cost over a seven-day cycle for three different safety stock settings with $h = 1, b = 9$.

the time-varying safety stock becomes clear, leading to a fundamental insight: time-varying safety stocks are most important when demand exhibits strong autocorrelation, and when high service levels are required.

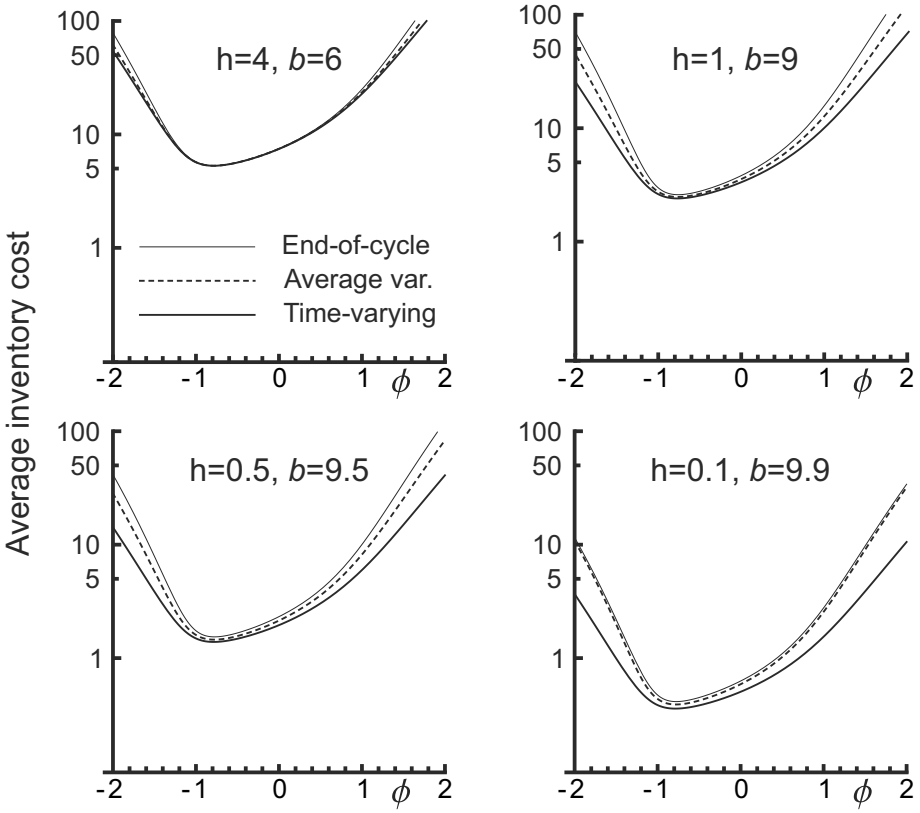

Figure 4.6: Average inventory cost over a seven-day cycle for various settings of h and b.

As the variable safety stock is the strategy of choice – providing the required availability at the lowest cost – we may wish to understand how the fill rate develops over a cycle, when time-varying safety stocks are in place. Though the availability remains constant, we see in Figure 4.7 (restricted to $|\phi| < 1$ as the fill rate can only be defined for stationary demand) that the fill rate fluctuates over τ, and that it depends on ϕ. Furthermore, Figure 4.7 indicates that the fill rate degrades as τ increases, particularly when demand is positively autocorrelated. Although the fill rate is undefined for nonstationary demand, we observe that the fill rate approaches 100% as $|\phi| \to 1$. This can be verified by taking the limit of the correlation coefficient ρ (based on the covariances in Corollary 4.6): $\lim_{|\phi|\to 1} \rho = 1$, and considering Corollary 4.10.

Corollary 4.10. *The fill rate is 100% when* $\rho = 1$ *and* $\mathbb{E}\left[(d_t)^+\right]$ *exists.*

Proof. When $\rho = 1$, the bivariate distribution degenerates to a univariate distribution. Therefore, $\min(d_t, i_t + d_t) = d_t$. Inserting this in (4.10) gives $S_2^- = \mathbb{E}\left[(d_t)^+\right] / \mathbb{E}\left[(d_t)^+\right] = 1$, when $\mathbb{E}\left[(d_t)^+\right]$ exists. ■

From the definition of the fill rate (4.10) and the knowledge that the inventory variance is increasing in τ, we can make some observations about the fill rate under constant safety stock settings. For the end-of-cycle constant safety stock setting, the fill rate at the end of the cycle, with $\tau = 7$, is identical to the fill rate of the optimal time-varying safety stock. For $\tau < 7$, the fill rate is higher. The other constant safety stock setting, based on the average inventory variance, does not have the fill rate of the optimal time-varying safety stock at $\tau = 7$ as a lower bound.

4.3.3 Determining the optimal planning cycle length

Figure 4.8 shows P^* for AR(1) demand under six different lead times, using (4.22) and (4.24). Each area between the contour lines indicates that a particular P^* is optimal; P^* is increasing in λ (every time we cross a contour in Figure 4.8 from below, P^* increases by one). For the setting $\phi = (-1)$, $L + P^*$ is always even, as a consequence of (4.22) in conjuction with the odd-even effect in (4.29). Therefore, with an even lead time, P^* is also even, and vice versa.

The area a is Figure 4.8 denotes the case when $P^* \geq 20$ but we have not drawn the contours as they become indistinguishable from each other. The following numerical example describes the optimization procedure.

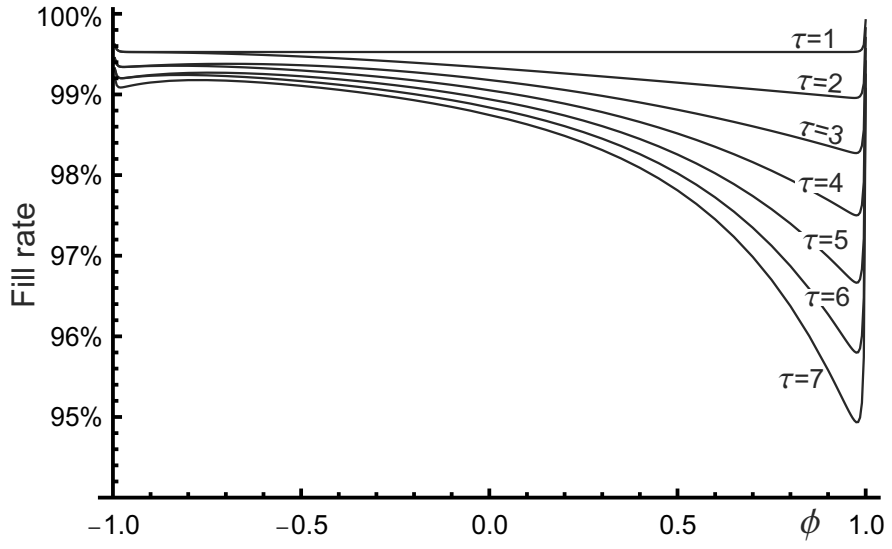

Figure 4.7: Fill rates are affected by staggering and by autocorrelation.

Example 4.11. To determine the optimal planning cycle length, start by identifying the auditing, inventory holding, and backlog costs. Then use (4.20) to determine ψ for this set of costs, and more importantly $\lambda = v/\psi$. Finally, exploit Theorem 4.5 to find P^*, either by inspecting Figure 4.8, or by finding two reorder cycle lengths P_1 and P_2, such that $\lambda_{P_1} \leq \lambda \leq \lambda_{P_2}$, and then performing a binary search for λ between P_1 and P_2, until a P is found such that $\lambda_{P-1} \leq \lambda \leq \lambda_P$. Then $P^* = P$ denotes the optimum.

Suppose $b = 9, h = 1, v = 10$. This leads to $\lambda = 0.695$. With zero lead time and i.i.d demand ($\phi = 0$), the open circle in Figure 4.8 shows that $P^* = 4$. Were demand instead positively correlated with $\phi = 0.9$, then $P^* = 2$, illustrated by the closed circle in Figure 4.8. Were $L = 4$, then $P^* = 5$ with $\phi = 0$, and $P^* = 2$ with $\phi = 0.9$. This illustrates that positive autocorrelation favours short planning cycles, and also that the physical production lead time influences P^*.

4.4 Conclusion

4.4.1 Theoretical contribution

We have identified the inventory-optimal policy under staggered deliveries and autocorrelated demand. The strategy is to correct all inventory errors for the first order of the cycle, and then to order only the forecasted demand for the period in question and the required change in the safety stock, according to (4.6) and (4.7). This makes real the optimal policy identified by the Flynn and Chiang papers by applying the OUT policy to autocorrelated demand.

The policy in this chapter is not identical to Flynn and Garstka (1990), as $o_{t,k}$ may be negative. Conservative settings of μ, and σ_ε can make the effect of negative orders negligible, meaning that our policy (under i.i.d. demand) becomes computationally consistent with Flynn and Garstka (1990).

There is a limited but important overlap between our policy and Chiang's (2009) simplified policy (CSP), which orders a variable amount in the first period of a cycle, and a constant quantity in the remaining periods. For i.i.d. demand, our model differs from CSP only in the management of safety stocks. Chiang's constant order quantity allows only for a linearly increasing safety stock over the cycle, but the optimal policy requires a nonlinear increase. However, when $P \leq 2$, CSP is identical to the optimal policy.

Our model allows for constant safety stocks if desired, but we find that not only are time-varying safety stocks more economical, they also ensure that the

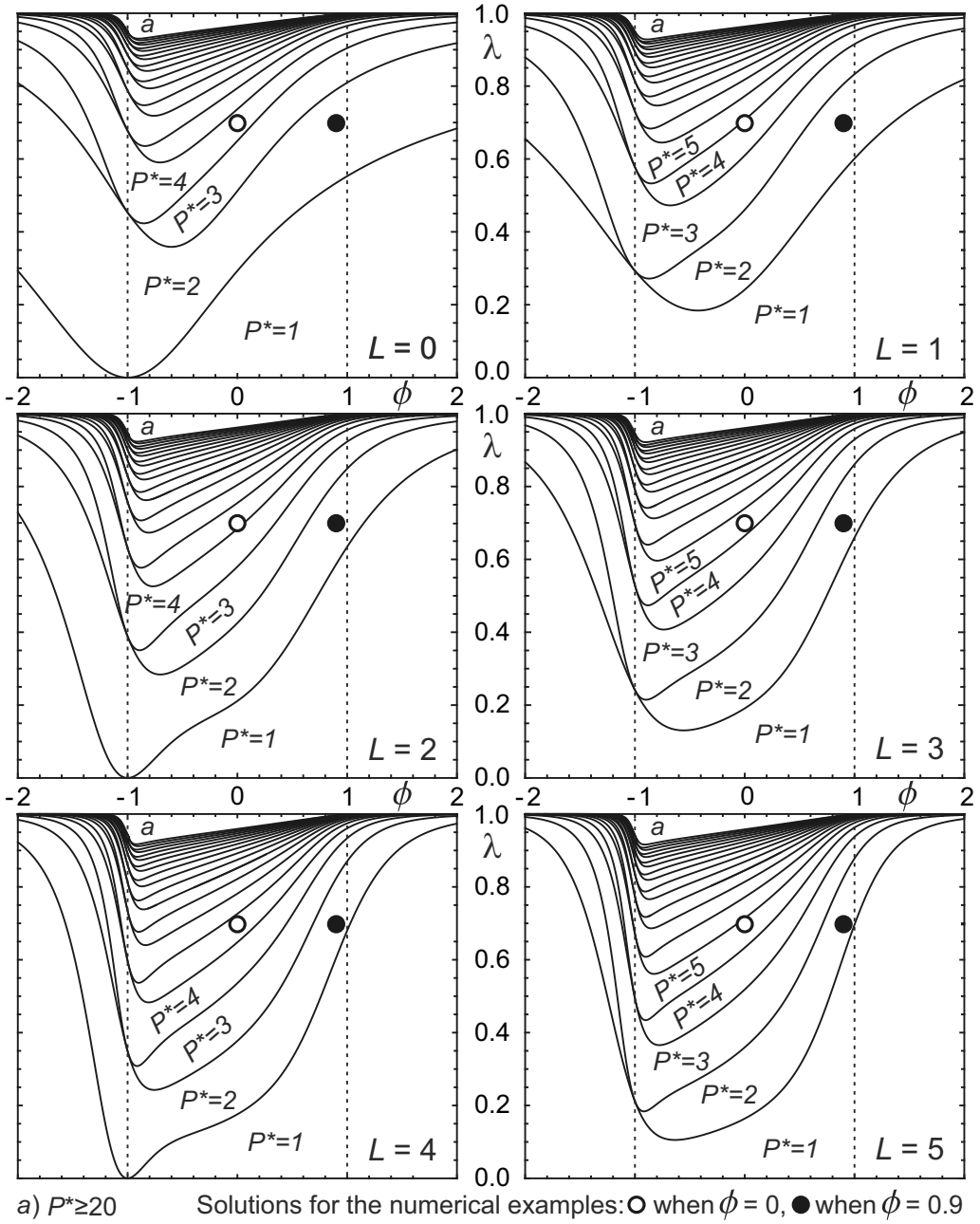

Figure 4.8: Optimal order cycle lengths P^* for some values of ϕ, λ, and L when $P^* < 20$.

target availability is achieved consistently. The overall safety stock is affected by the autocorrelation of demand and increases with the order cycle length. In the special case of an AR(1) demand process with $\phi = (-1)$, the safety stock only needs to be changed for periods when τ is odd. Our model and analysis also captures the nonstaggered case when $P = 1$, which results in a regular OUT policy.

The inventory variance is increasing over the order cycle for any demand autocorrelation function, and the heteroskedasticity affects fill rates, even when the availability is kept constant. This causes the fill rate to fluctuate cyclically. We have also provided an exact approach for determining P^*, the optimal length of the planning cycle, when auditing, holding, and backlog costs are present. The optimization procedure reveals that P^* is an increasing function of the audit cost v, and a decreasing function of b, h, L, and σ_ε. When τ is specified, the inventory level is an MA(τ) process, as can be seen from (B.10). As a result, the inventory level is always stationary, for any demand autocorrelation (Box and Jenkins, 1976, p. 79).

4.4.2 Managerial insights

If a production system requires consistent availability, it is necessary to take into account the time-varying inventory variance. Ignoring the heteroskedasticity of inventory will result in either excessive service levels and unnecessary costs, or poor service on predictable days of the planning cycle. However, if time-varying safety stocks are deemed impractical or too complicated, we recommend a safety stock setting based on the average inventory variance over the cycle. Then the availability will fluctuate over the cycle, but on average it will exceed the critical fractile $b/(b+h)$ when $b \geq h$.

Even with optimal time-varying safety stocks, fill rates may degrade over the cycle, particularly when demand is positively autocorrelated. Reducing the length of the planning cycle provides an opportunity to reduce inventory costs, and is especially attractive when the initial planning cycle is long, and demand exhibits strong autocorrelation. However, short planning cycles require frequent audits, incurring a cost. The balance between inventory and audit costs must be regulated carefully via the order cycle length.

4.4.3 Summary

This chapter has identified the optimal staggered policy under autocorrelated demand (Theorem 4.2, answering RQ 1), when piecewise-linear inventory costs, and an audit cost are present. Conditions for the optimal order cycle length were also given, as well as a search procedure for its identification (RQ 3). Apart from cost expressions, we derived fill rate and availability expressions, resolving RQ 2. A result of particular interest is that the optimal policy maintains a constant availability by changing the safety stock levels dynamically within each cycle. While the availability is constant, the fill rate deteriorates over the cycle.

This chapter has revealed that the OUT policy is still optimal when demand is autocorrelated, which extends the findings of Flynn and Garstka (1990), who considered i.i.d. demand. We also provided a method for identifying P^*, which can be compared to the procedure for i.i.d. demand in Flynn and Garstka (1997). In addition, we have extended the findings of Lian et al. (2006) by considering forecasts for autocorrelated demand, and by providing availability and fill rate service levels per period. One unresolved question is that of alternative forecasting methods, and how they affect inventory systems with staggered deliveries. The policy presented in this chapter will be referred to as ARSTOUT in the following chapters.

Chapter 5

Bullwhip and capacity costs

The preceding chapter presented an inventory-optimal policy for autocorrelated demand. Here we shall assume that the audit cost is zero, and investigate how staggering affects bullwhip and capacity costs under i.i.d. demand. The reason for restricting the analysis is to keep the analysis tractable, and to highlight the trade-off between order variance and inventory variance in a simple setting. The assumption of i.i.d. demand is not unrealistic, as it has been observed industrially (Disney et al., 2016). The audit cost was also ignored, to illustrate clearly how the order cycle length relates to a trade-off between inventory and capacity costs. First, we identify the policy that minimizes the weighted sum of production variance (bullwhip) and inventory variance. Then we investigate how this policy performs for a different cost function, and compare this with the performance of two similar policies with different constraints on overtime work, including having all overtime work collected to a short period, such as a shift, or having the overtime work distributed evenly over the cycle.

5.1 Linear quadratic control

We seek policy that minimizes the weighted total variance of inventory and receipts. Again, we consider the staggered system described by (4.2) and (4.1), but we now assume that demand is i.i.d. Recalling that we have perfect state information, and that the inventory balance equation is linear, we expect a Linear Quadratic Regulator (LQR) to be optimal (Kirk, 1997).

In principle, we could model the system with a state vector $\mathbf{x_t}$ of dimension $(P + L) \times 1$, to keep track of all WIP and all inventory levels in a cycle, and with a control vector $\mathbf{u_t}$ of dimension $P \times 1$ to keep track of the orders made per cycle. For simple cases with *a priori* given lead times and cycle lengths,

this naïve implementation works well. But when lead times and cycle lengths are arbitrary, the Ricatti equation for this formulation becomes complicated.

To find a workable solution, we must recast the problem. Let us call the inventory position x_t. The inventory position and the inventory level are related by

$$i_{t+\tau} = x_t - \sum_{n=1}^{\tau} d_{t+\tau}, \tag{5.1}$$

where τ is the effective lead time, including information delays. When demand is i.i.d., the inventory position is not correlated with future demand noise, that is $\mathrm{cov}(x_t,\, d_{t+n} - \mathbb{E}\left[d_{t+n}\right]) = \mathrm{cov}(x_t,\, \varepsilon_{t+n}) = 0$ for non-negative n, therefore

$$\mathrm{var}(i_{t+\tau}) = \mathrm{var}(x_t) + \mathrm{var}\left[\sum_{n=1}^{\tau} d_{t+n}\right]. \tag{5.2}$$

The equation above reveals that the variance of the inventory level has two components: the variance of the inventory position, and the variance of lead-time demand. Equation (5.2) also shows that for variance control, there is no benefit to keeping track of WIP and inventory separately. Instead, it suffices to monitor only the inventory position. With this established, let us model the case when $P = 1$, which is the same as the time-invariant case without staggered deliveries.

The standard form for a linear-time invariant system is $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t$. We have the state vector $\mathbf{x}_t = x_t$, the control $\mathbf{u}_t = r_{t+L}$, and we find that $\mathbf{A} = 1$, $\mathbf{B} = 1$, and $\mathbf{w_t} = -d_t$ satisfy the evolution of the inventory position[1]. The standard form for the cost associated with this problem is (Bertsekas, 2005, p. 148):

$$\lim_{N\to\infty} E\left\{\mathbf{x}_N'\mathbf{Q}\mathbf{x}_N + \sum_{k=0}^{N-1}\left(\mathbf{x}_k'\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k'\mathbf{R}\mathbf{u}_k\right)\right\}, \tag{5.3}$$

and as we have only one state variable and one control, we choose the cost vectors $\mathbf{Q} = \lambda$ and $\mathbf{R} = 1 - \lambda$. This means that the real variable $\lambda \in [0,\, 1]$ can express all possible preferences between inventory variance and production rate variance. Note that we interpret λ differently for this quadratic cost model, here inventory costs dominate for $\lambda = 1$, and capacity costs for $\lambda = 0$. This switch leads to more compact expressions. As we have a linear time-invariant system with quadratic costs, the optimal control law is of the form $\mathbf{L}\mathbf{x}_t$, where $\mathbf{L} = -\left(\mathbf{B}'\,\mathbf{K}\,\mathbf{B} + \mathbf{R}\right)^{-1}\mathbf{B}'\,\mathbf{K}\,\mathbf{A}$ (Bertsekas, 2005, p. 151). The variable $\mathbf{K}$ is a

[1] To keep the notation simple, we present one-dimensional vectors as scalars.

solution to the algebraic Ricatti equation

$$\mathbf{K} = \mathbf{A}'(\mathbf{K} - \mathbf{K}\,\mathbf{B}(\mathbf{B}'\,\mathbf{K}\,\mathbf{B} + \mathbf{R})^{-1}\,\mathbf{B}'\,\mathbf{K})\mathbf{A} + \mathbf{Q}. \tag{5.4}$$

For the problem at hand, this becomes

$$\mathbf{K} = \mathbf{K} - \frac{\mathbf{K}^2}{\mathbf{K} + 1 - \lambda} + \lambda, \tag{5.5}$$

thus

$$\mathbf{K} = \frac{\lambda}{2} \pm \sqrt{\lambda(4 - 3\lambda)}; \tag{5.6}$$

only one of these solutions gives the optimal control law. By inserting both in $\mathbf{L}$, we obtain

$$\mathbf{L} = \frac{\lambda \pm \sqrt{\lambda(4 - 3\lambda)}}{2 - 2\lambda} \tag{5.7}$$

One of these solutions has positive feedback, rendering it unstable. The other solution is stable,

$$\mathbf{L} = \frac{\lambda - \sqrt{\lambda(4 - 3\lambda)}}{2 - 2\lambda}. \tag{5.8}$$

The optimal policy is then obtained as $\mathbf{L}\mathbf{x}_t$. For every period in a non-staggered system, we observe the inventory position, multiply it by $\mathbf{L}$, and produce the resulting amount. For future periods, we repeat the calculation.

Under staggering, we must make several production decisions in each period. Let us call the initial observed inventory position x_t, and the inventory positions resulting from our ordering decisions in the same cycle as $x_{t,k}$. This notation helps us to cope with the cycle / period duality by the relation

$$x_t = x_{t-P} + \sum_{k=1}^{P} (o_{t-P,k} - d_{t-P+k-1}) = x_{t-P,P} + o_{t-P,P} - d_{t-1}. \tag{5.9}$$

With this notation, the first ordering decision is $o_{t,1} = \mu + \mathbf{L}x_t$. For the next period, the inventory position has been raised by the amount ordered, $x_{t,1} = x_t + o_{t,1}$. As no demand has been observed, it has not been subtracted from the inventory position. For the next ordering decision $o_{t,2}$, we notice that there is an updated inventory position $x_{t,1}$, and that the effective lead time τ has increased by one, as this order will arrive one period later. However, for $k = 2$ the optimal policy is independent of the lead time, and has the same structure and parameters as for $k = 1$. We simply need to iterate the control for the updated inventory position.

$$x_{t,k} = \begin{cases} (1 + \mathbf{L})\, x_t + \mu & \text{when } k = 1, \\ x_{t,k-1} + \mathbf{L} x_{t,k-1} + \mu & \text{otherwise.} \end{cases} \tag{5.10}$$

Since we are still in period t, no demand has been subtracted from the inventory position. Note that this staggered inventory position can be expressed in the same standard form as the non-staggered inventory position. Hence, it has the same optimal control policy. As the initial inventory position is known, the above recursion can be expressed as

$$o^*_{t,k} = x_{t,1}\,(\xi - 1)\,\xi^{k-1}, \tag{5.11}$$

where $\xi = \mathbf{L} + 1$.

Example 5.1. Suppose that we are doubtful of the control policy in (5.11) and want to solve this problem numerically. Assume a lead time $L = 0$, and reorder period $P = 3$. We have a preference toward reducing inventory variance, $\lambda = 0.6$.

We begin by expressing this system on the standard form, with $\mathbf{x}$ representing the three inventory levels over the cycle, and $\mathbf{u}$ representing the three orders over the cycle. Each inventory level in the cycle consists of the initial inventory at the start of the cycle, plus the cumulative orders up to that point, minus demand. Demand is a vector of cumulative demand over the cycle, with an element for each period. While the elements within a single cycle are correlated, they are uncorrelated between cycles. Therefore, we do not deviate from the standard assumptions for linear quadratic regulators. As we have known and constant system matrices, certainty equivalence holds, and the demand noise can be ignored (Bertsekas, 2005, p. 160). It is easy to see that

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix},$$

and the cost matrices take the form $\mathbf{Q} = 0.6\,\mathbf{I}(3)$ and $\mathbf{R} = 0.4\,\mathbf{I}(3)$. As an alternative to solving the Ricatti equation by hand, we can employ the Mathematica command `LQRegulatorGains` to obtain the gain matrix

$$-\mathbf{L} = \begin{bmatrix} 0. & 0. & 0.686141 \\ 0. & 0. & 0.215352 \\ 0. & 0. & 0.0675901 \end{bmatrix}. \tag{5.12}$$

Inserting the same cost preference, $\lambda = 0.6$, into the optimal order rate equation (5.11) gives $\{o^*_{t,1}, o^*_{t,2}, o^*_{t,3}\} = x_t\{-0.686141, -0.215352, -0.0675901\}$. Although we used a different formulation, the result matches (5.11). As predicted, it does not matter if we minimize inventory level variance directly, or by using the inventory position as a proxy.

5.1.1 Variances of the bullwhip-optimal policy

Theorem 5.2.

(a) If orders are placed in period t, the inventory variance is

$$\operatorname{var}(i_{t+k+L}) = \operatorname{var}(\varepsilon)\left[\frac{P\xi^{2k}}{1-\xi^{2P}} + k + L\right]; \tag{5.13}$$

(b) the order rate variance is

$$\operatorname{var}\left(o^*_{t,k}\right) = \operatorname{var}(\varepsilon)\frac{P(\xi-1)^2\xi^{2k-2}}{1-\xi^{2P}}. \tag{5.14}$$

Proof.

(a) To find the inventory variance, we shall express the inventory position as a weighted sum of error terms ε, calculate the variance of the inventory position, and then use the relation between the inventory position and the inventory variance (5.2). Recalling (5.11) as $o^*_{t,k} = x_t(\xi-1)\xi^{k-1}$, we have

$$\begin{aligned} x_{t,k} &= \xi^k\left[x_{t-P} + \sum_{m=1}^{P} o^*_{t-P,m} - \varepsilon_{t-P+m}\right] \\ &= \xi^k\left[\xi^P x_{t-P} - \sum_{m=1}^{P}\varepsilon_{t-P+m}\right] \\ &= \xi^{2P+k}x_{t-2P} - \sum_{m=1}^{P}\left(\xi^k\varepsilon_{t-P+m} + \xi^{P+k}\varepsilon_{t-2P+m}\right). \end{aligned} \tag{5.15}$$

Iterating this N cycles back yields

$$x_{t,k} = \xi^{NP+k}x_{t-NP} - \sum_{j=0}^{N}\sum_{k=1}^{P}\xi^{jP+k}\varepsilon_{t-jP+m}. \tag{5.16}$$

As $\lim_{N\to\infty}\xi^N = 0$, we have

$$x_{t,k} = -\sum_{j=0}^{\infty}\sum_{k=1}^{P}\xi^{jP+k}\varepsilon_{t-jP+m}. \tag{5.17}$$

Taking the variance gives

$$\begin{aligned} \operatorname{var}(x_{t,k}) &= \operatorname{var}(\varepsilon_{t,k})\, P \sum_{j=0}^{\infty} \left(\xi^{jP+k}\right)^2 \\ &= \operatorname{var}(\varepsilon_{t,k})\, \frac{P\xi^{2k}}{1-\xi^{2P}}. \end{aligned} \tag{5.18}$$

The inventory variance is obtained as

$$\begin{aligned} \operatorname{var}(i_{t+k}) &= \operatorname{var}(x_{t,k}) + \operatorname{var}\left[\textstyle\sum_{n=1}^{k+L} \varepsilon_{t+n}\right] \\ &= \operatorname{var}(\varepsilon) \left[k + L + \frac{P\xi^{2k}}{1-\xi^{2P}}\right], \end{aligned} \tag{5.19}$$

and the first part of the proof is complete.

(b) The relation $o^*_{t,k} = x_t(\xi - 1)\xi^{k-1}$ holds. Hence,

$$\operatorname{var}\left(o^*_{t+k}\right) = \operatorname{var}\left(x_t(\xi-1)\xi^{k-1}\right) = \operatorname{var}(\varepsilon_t)\, \frac{(\xi-1)^2 P\xi^{2k-2}}{1-\xi^{2P}}, \tag{5.20}$$

and the proof is complete.

■

5.1.2 Properties of the bullwhip-optimal policy

The bullwhip-optimal policy differs from the inventory-optimal policy as an OUT policy is no longer optimal. Instead, it is optimal to recover a fraction of the expected deviation from the target stock level. As expected, if inventory costs are high in relation to the order rate variance, we recover inventory more aggressively. This is illustrated in Figure 5.1, which presents the optimal order quantity as a fraction of the inventory position's deviation from equilibrium. Even with $\lambda = 1$, no bullwhip is generated by this policy under i.i.d. demand, but we may speculate that bullwhip can appear if demand is autocorrelated, and if the forecasted demand is added to the order quantity without smoothing. The quadratic cost increases with both P and with L. Accordingly, it is preferable not to stagger deliveries ($P^* = 1$), if we limit our cost model to capacity and inventory. As a final step, we shall verify that the system is stationary. Note that the inventory position can be expressed as a first-order system of the form $x_{t,k} = \xi x_{t,k-1}$, which is asymptotically stable (and therefore generates stationary time series) if $|\xi| < 1$ (Luenberger, 1979, p. 154–157). From the

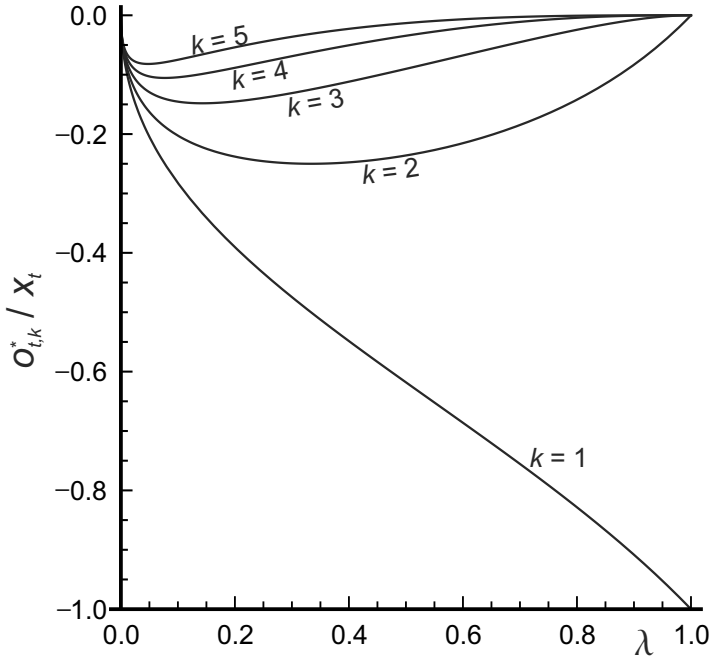

Figure 5.1: Optimal order quantities as a fraction of the inventory position's deviation from its target.

definition of ξ, we find that the inventory position is stationary for $0 < \lambda \leq 1$. However, due to (5.11) the orders are stationary when $0 \leq \lambda \leq 1$.

5.2 Capacity costs and overtime

Consider a system, with the same inventory mechanism as we have considered thus far, using (4.1) and (4.2). The inventory costs are not quadratic, but piecewise-linear (4.3), as in the preceding chapter. Demand is assumed to be i.i.d. and normally distributed. We shall explore four policies that differ over two dimensions: the allocation of overtime within the cycle, and production smoothing between cycles. The myopic policy of the preceding chapter, we call the staggered order-up-to policy (STOUT); when this is adapted by working an equal amount of overtime in every period it is referred to as STOUT-E. The smoothing policy that allocates all of the overtime to the first period of each cycle is called the staggered proportional order-up-to policy (SPOUT), and when the overtime is distributed equally over the cycle, it is called SPOUT-E. Figure 5.2 illustrates the overtime allocation of the four policies.

As a first step, we shall formulate the inventory-optimal SPOUT policy from the last chapter in terms of optimal OUT levels x^*_k. Recalling our previous definition of the inventory position, we can use (4.2) and (4.1) to obtain a

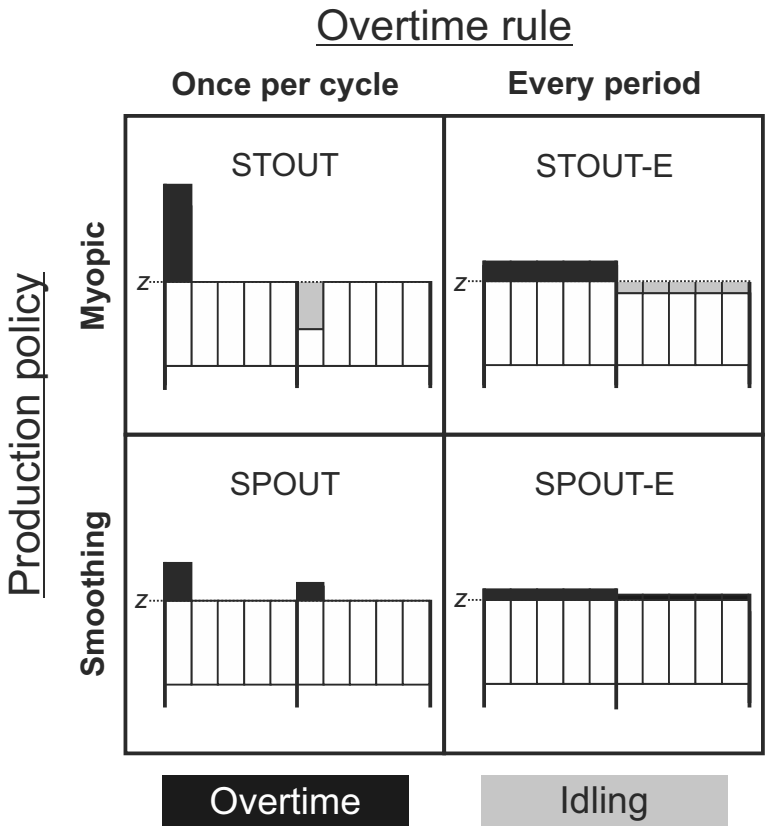

Figure 5.2: Illustration of the four pragmatic policies.

future inventory level as

$$i_{t+k+L} = i_t + \sum_{n=1}^{k+L} (r_{t+n} - d_{t+n}) = x_{t,k} - \sum_{n=1}^{k+L} d_{t+n} \quad \text{when } t/P \in \mathbb{Z}, \tag{5.21}$$

We can influence the inventory position $x_{t,k}$ as it contains $r_{t+k+L} = o_{t,k}$. The optimal safety stock (4.5) can therefore be expressed as

$$\mathbb{E}[i_{t+k+L}] = x_k^* - \mu\,(k+L) = \Phi_{i,k}^{-1}\left(\frac{b}{b+h}\right), \tag{5.22}$$

where $\Phi_{i,k}^{-1}$ is the inverse of the inventory level's cumulative distribution function. The policy that attains this optimal safety stock is

$$o_{t,k} = x_k^* - x_{t,k-1}, \tag{5.23}$$

where $x_{t,0} = x_{t,1} - o_{t,1}$ is the observed inventory position before any orders have been placed. This is simply another representation of (4.6) and (4.7). Note that only the first order $o_{t,1}$ is a random variable, and that the other orders,

with $k \geq 2$, are deterministic. Therefore, this policy will cause overtime to be worked in the first period of each cycle, corresponding to case 2 in Figure 1.3.

5.2.1 Capacity costs

The inventory-optimal policy has another attractive feature: the production volume is stochastic only for the first order of every cycle. Despite not being designed for capacity costs, we may still be interested to see how this policy performs when capacity costs are present. The costs follow a model with guaranteed hours and overtime (Hosoda and Disney, 2012). Workers are guaranteed compensation for a daily output of z_k products at the normal rate u dollars per product; in cases when the production quantity is greater than z_k, the excess is paid for with the overtime rate v per product,

$$a_{t,k} = uz_k + v(o_{t,k} - z_k)^+. \tag{5.24}$$

As $o_{t,k}$ does not vary with z_k, and as it is a linear function of the system state, the orders will be normally distributed. Disney and Grubbström (2004) show that this problem has a newsvendor-type solution. The optimal capacity level is $z^* = \sigma_{o,k}\Phi^{-1}\left[(v-u)/v\right] + x_k^* - x_{k-1}^*$, where $\sigma_{o,k}$ is the order rate variance of order k. The same reference provides the expected capacity cost when z^* is used,

$$\mathbb{E}\left[a_{t,k}\right] = v\sigma_{o,k}\varphi\left[\Phi^{-1}\left(\frac{v-u}{v}\right)\right] + u\left(x_k^* - x_{k-1}^*\right). \tag{5.25}$$

The average order cost per period is

$$A_P = \frac{1}{P}\sum_{k=1}^{P}\mathbb{E}\left[a_{t,k}\right] = v\bar{\sigma}_{o,P}\varphi\left[\Phi^{-1}\left(\frac{v-u}{v}\right)\right] + u\mu, \tag{5.26}$$

where $\bar{\sigma}_{o,P} = P^{-1}\sum_{k=1}^{P}\sigma_{o,k}$ is the average standard deviation of the orders. The total average cost per period is then

$$C_P = J_P + A_P, \tag{5.27}$$

including both inventory costs and capacity costs. Before the average costs in (4.19) and (5.26) can be calculated, we must know the standard deviations of both inventory and orders. For the STOUT policy, they are trivial:

Lemma 5.3. *For the STOUT policy,*

(a) the inventory variance is

$$\sigma_{i,k}^2 = \sigma_d^2 (k + L); \tag{5.28}$$

(b) the order rate variance is

$$\sigma_{o,k}^2 = \begin{cases} \sigma_d^2 P & \textit{when } k = 1, \\ 0 & \textit{otherwise.} \end{cases} \tag{5.29}$$

Proof.

(a) Taking the variance of (5.21) provides

$$\sigma_{i,k}^2 = \operatorname{var}(x_{t,k}) + \operatorname{var}\left(\sum_{n=1}^{k+L} d_{t+n}\right) = \sigma_d^2 (k + L), \tag{5.30}$$

as future demand is uncorrelated with the inventory position. For this policy $x_{t,k} = x_k^*$ is constant across time (for each k), and therefore $\operatorname{var}(x_{t,k}) = 0$.

(b) From (4.1), (5.21), and (5.23) we obtain

$$o_{t+P,1} = x_1^* - i_{t+P} = x_1^* - \left(x_P^* - \sum_{n=1}^{P} d_{t+n}\right), \tag{5.31}$$

from which it is easy to see that $\sigma_{o,1}^2 = \sigma_d^2 P$. For the remaining periods with $k > 1$, $o_{t,k} = \mathbb{E}(o_{t,k}) = x_k^* - x_{k-1}^*$, and therefore $\sigma_{o,k}^2 = 0$. This completes the proof. ■

Having specified STOUT and its variances, we find that the average standard deviation of the inventory increases with P, $\bar{\sigma}_{i,P} = \sigma_d P^{-1} \sum_{k=1}^{P} \sqrt{k + L}$ and that the average standard deviation of the orders is decreasing in P, $\bar{\sigma}_{o,P} = \sigma_d \sqrt{P^{-1}}$. This shows that the order cycle contains a crude mechanism for production smoothing, and implies that the optimal reorder period may be distinct from one. The optimization problem is non-convex, and can be solved with an inverse-function approach. First, we write the total cost as

$$C_P(\lambda) = \psi \left[\bar{\sigma}_{i,P} + \lambda (\bar{\sigma}_{o,P} - \bar{\sigma}_{i,P})\right] + \mu u, \tag{5.32}$$

where ψ is a scaling factor,

$$\psi = v\varphi\left[\Phi^{-1}\left(\frac{v - u}{v}\right)\right] + (b + h)\,\varphi\left[\Phi^{-1}\left(\frac{b}{b + h}\right)\right], \tag{5.33}$$

and

$$\lambda = \frac{v\varphi\left\{\Phi^{-1}\left[(v-u)/v\right]\right\}}{\psi}, \tag{5.34}$$

is a continuous variable from zero to one, providing the balance between inventory costs and overtime costs. The setting $\lambda = 1$ represents capacity costs only ($b = 0$, or $h = 0$, or $b + h = 0$) and $\lambda = 0$ indicates inventory costs only ($v = 0$). The important result from (5.32) is that the total cost is a linear function of λ for any fixed P.

Theorem 5.4. *The order cycle length P minimizes the total cost $C_P(\lambda^*)$ for $\lambda^* \in [\lambda_{P-1}, \lambda_P]$, where $\lambda_0 = 0$, and*

$$\lambda_P = \frac{\bar{\sigma}_{i,P+1} - \bar{\sigma}_{i,P}}{\bar{\sigma}_{i,P+1} - \bar{\sigma}_{i,P} + \bar{\sigma}_{o,P} - \bar{\sigma}_{o,P+1}}. \tag{5.35}$$

Proof. Let λ_P be the point at which we are indifferent between the choice of P or $P+1$, occurring when $C_P(\lambda_P) = C_{P+1}(\lambda_P)$. Solving for λ_P gives (5.35), which is equivalent to

$$\lambda_P = 1 - \left[1 + \frac{P\left(\sqrt{P} + \sqrt{P+1}\right)}{\sqrt{P(P+1)}} \cdot \frac{\sigma_{i,P+1} - \bar{\sigma}_{i,P}}{\sigma_\varepsilon}\right]^{-1}. \tag{5.36}$$

From this expression, it is clear that λ_P is increasing in P. Therefore, every P is optimal when $\lambda \in [\lambda_{P-1}, \lambda_P]$. This completes the proof. ■

5.2.2 Staggered order-up-to policy with equal overtime

To this point, overtime work has been allocated to the first period of each order cycle. We may be interested to see the effects of distributing the overtime work evenly over every period in the cycle. Starting with the STOUT policy, we divide the overtime work into P equal parts. This provides the STOUT-E policy:

$$o_{t,k} = x_k^* - x_{k-1}^* + P^{-1}\left(x_0^* - x_{t,0}\right), \tag{5.37}$$

where $x_0^* = \mathbb{E}\left[x_{t,0}\right] = x_P^* - \mu P$. New values must be computed for x_k^*. The variances of inventory and orders also change.

Lemma 5.5. *For the STOUT-E policy,*

(a) the inventory variance is

$$var\left(i_{t+L+k}\right) = \sigma_d^2\left[k + L + \frac{(P-k)^2}{P}\right]. \tag{5.38}$$

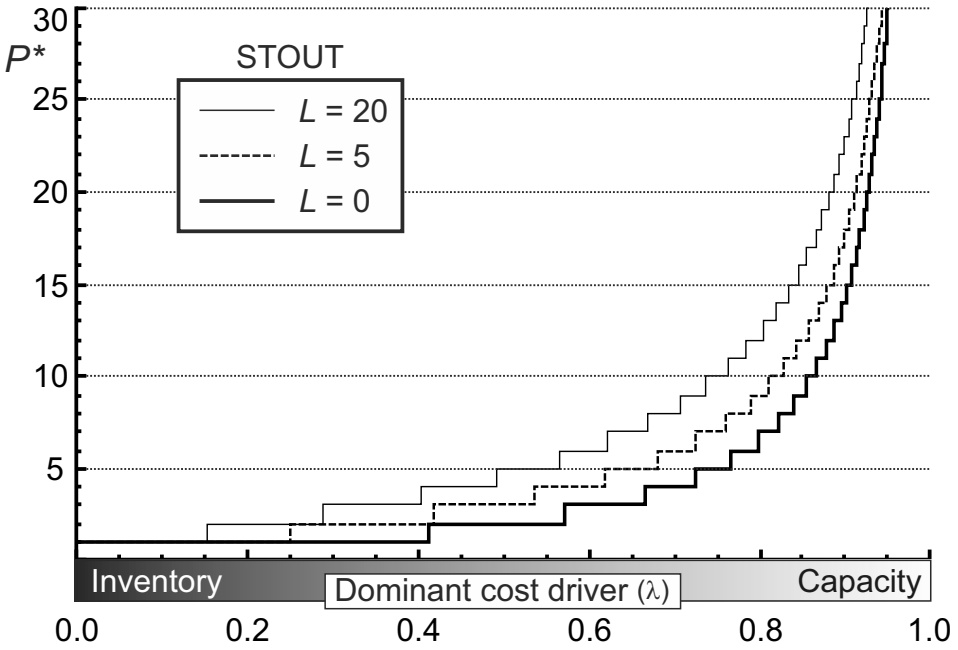

Figure 5.3: Optimal order cycle lengths, P^*, under the STOUT policy.

(b) the order rate variance is

$$\sigma_{o,k}^2 = \sigma_d^2/P \text{ for all } k. \tag{5.39}$$

Proof.

(a) From (5.21) and (5.37) we see that $x_{t,P} = x_P^*$. Consequently, $x_{t,0} = x_P^* - \sum_{n=1}^{P} d_{t-P+n}$. The inventory position can then be expressed as

$$x_{t,k} = x_k^* - \frac{P-k}{P} \sum_{n=1}^{P} (d_{t-n} - \mu) \tag{5.40}$$

which provides the variance of the inventory position as,

$$\text{var}\,(x_{t,k}) = \sigma_d^2 \frac{(P-k)^2}{P} \tag{5.41}$$

Recalling that future demand is uncorrelated with the inventory position

$$\sigma_{i,k}^2 = \text{var}(x_{t,k}) + \text{var}\left(\sum_{n=1}^{k+L} d_{t+n}\right) = \sigma_d^2 \left[k + L + \frac{(P-k)^2}{P}\right], \tag{5.42}$$

and the first part of the proof is complete.

(b) Note that $x_{t,0} = x_0^* - \sum_{n=1}^{P} (d_{t-P+n} - \mu)$, which when inserted in (5.37)

provides

$$o_{t,k} = x_k^* - x_{k-1}^* + \sum_{n=1}^{P} \frac{d_{t-P+n} - \mu}{P}. \tag{5.43}$$

Taking the variance of the above gives $\sigma_{o,k}^2 = \sigma_d^2/P$, completing the proof. ■

The inventory variance of the STOUT-E policy (5.38) is greater than that of STOUT. Note that $\bar{\sigma}_{o,P}$ is identical for STOUT and STOUT-E. As a result, the realized capacity cost will be the same, despite the difference in overtime strategy. Thus for any relative weighting of inventory and capacity costs, the STOUT policy dominates the STOUT-E policy. For this reason, we shall not optimize it, as one should switch to the STOUT policy first. Notably, the STOUT-E inventory variance is not always increasing in k. Instead, the STOUT-E inventory variance is minimized when $k = P/2$ for even P, or when $k = (P \pm 1)/2$ for odd P. See Figure 5.4 for an example with $P = 5$. If one uses a constant safety stock, the changing inventory variance will cause the availability to fluctuate, as was shown in Chapter 4. Figure 5.5 illustrates the availability fluctuations when the safety stock is a constant based on the end-of-cycle inventory variance, $\mathbb{E}[i_{t+k+L}] = x_0^*$. This is clearly not optimal but it is probably a common way to manage safety stocks. When safety stocks are set to minimize the expected inventory costs in each period, the availability is constant over the cycle.

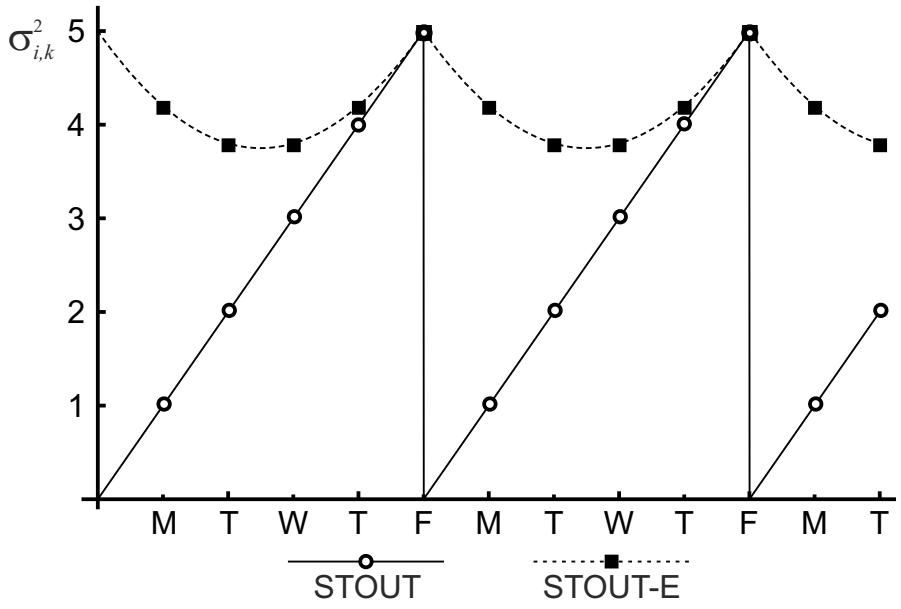

Figure 5.4: The inventory variance is cyclically heteroskedastic.

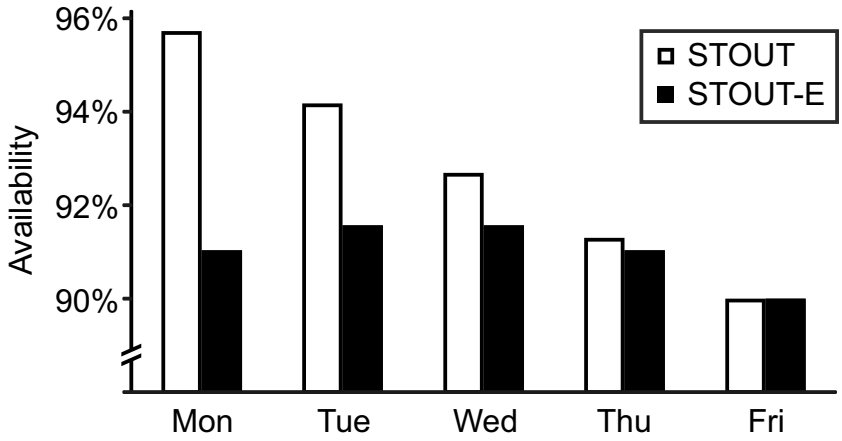

Figure 5.5: Availability fluctuations depend on the overtime strategy used.

5.3 Staggered proportional order-up-to policy

It is clear that the OUT policy can exploit the order cycle P to strike a favourable balance between the different cost drivers, but the balance between inventory and capacity costs can also be managed by the addition of a proportional feedback controller, which corrects a multiple α of the inventory position's error each time reordering takes place (Simon and Holt, 1954). For convenience, define $x_0^* = x_P^* - \mu P$, then

$$o_{t,k} = \begin{cases} x_1^* - x_0^* + \alpha\left(x_0^* - x_{t,0}\right) & \text{when } k = 1, \\ x_k^* - x_{k-1}^* & \text{otherwise.} \end{cases} \tag{5.44}$$

Theorem 5.6.

(a) The inventory variance under the SPOUT policy is

$$\sigma_{i,k}^2 = \sigma_d^2 \left[k + L + \frac{P\left(1-\alpha\right)^2}{\alpha\left(2-\alpha\right)} \right], \quad \textit{when } 0 < \alpha < 2. \tag{5.45}$$

(b) The variance of the orders is

$$\sigma_{o,k}^2 = \begin{cases} \sigma_d^2 \alpha P \left(2-\alpha\right)^{-1}, & \text{when } \{k = 1, 0 \leq \alpha < 2\} \\ 0, & \text{otherwise.} \end{cases} \tag{5.46}$$

Proof. The proof is presented in Appendix B.3. ■

To see when the SPOUT policy is stationary, observe that for all k, $x_{t,k} - \mathbb{E}\left[x_{t,k}\right] = \left(1-\alpha\right)\left(x_{t,0} - \mathbb{E}\left[x_{t,0}\right]\right)$. As this is a first-order system, we have a

stationary inventory position when $|1-\alpha| < 1$, or equivalently when $0 < \alpha < 2$. The orders are stationary when $0 \leq \alpha < 2$, following $|\alpha(1-\alpha)| < 1$.

5.3.1 Finding the optimal smoothing setting α^*

To obtain the optimal α, it is sufficient to set $P = 1$, and then to differentiate C_P with regard to α, and then to solve for zero. Except for the trivial case $L = 0 \rightarrow \alpha^* = 1 - \lambda$, the resulting expression is very large. Due to its size, it is effectively unreadable, and has therefore been omitted. Figure 5.6 shows α^* for some values of L, where it is clear that $\alpha^* \leq 1 - \lambda$, and that the required damping increases with the lead time. For the special case $\{P = 1, L = 0\}$ the optimal total cost is $C_P^* = \sqrt{1-\lambda^2}$. Whenever $\lambda > 0$, this cost is lower than the minimum cost obtainable via STOUT. This results from the following lemma.

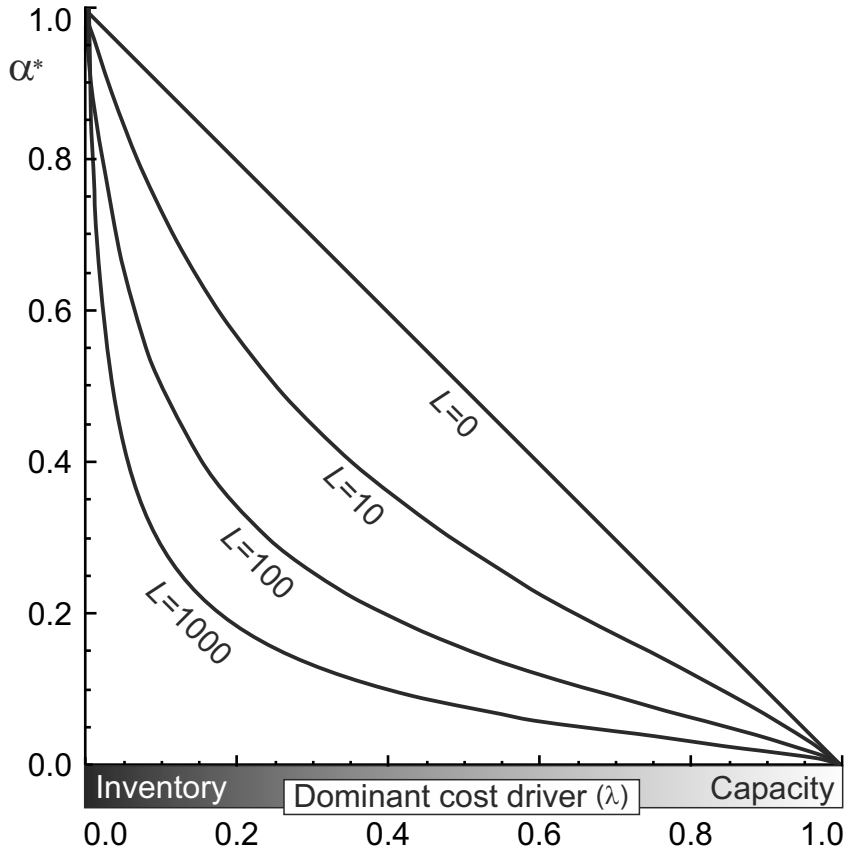

Figure 5.6: Optimal values for the feedback controller α when $P = 1$.

Lemma 5.7. *When $L = 0$, the total cost of STOUT (with arbitrary P) is no less than the optimal non-staggered SPOUT cost ($P = 1$ and $\alpha = \alpha^*$). We*

express this as

$$(C_1^*|_{\text{SPOUT}}) \leq (C_P|_{\text{STOUT}}) \tag{5.47}$$

Proof. The proof is presented in Appendix B.4. ■

As a result, it is more economical to embed production smoothing in the order policy, than to produce it via manipulating the order cycle length of the STOUT policy.

5.3.2 Staggered proportional policy with equal overtime

Just as the STOUT policy has an equal-overtime variant, so can one be identified for SPOUT. We define the SPOUT-E policy as

$$o_{t,k} = x_k^* - x_{k-1}^* + \alpha P^{-1}\left(x_0^* - x_{t,0}\right). \tag{5.48}$$

The variances required to calculate costs and availability are given below.

Theorem 5.8. *For the SPOUT-E policy,*

(a) the inventory variance is

$$\sigma_{i,k}^2 = \sigma_d^2\left[k + L + \frac{(P-\alpha k)^2}{\alpha P\,(2-\alpha)}\right], \quad \textit{when } 0 < \alpha < 2. \tag{5.49}$$

(b) The order rate variance is

$$\sigma_{o,k}^2 = \frac{\sigma_d^2 \alpha}{P\,(2-\alpha)}, \quad \textit{when } 0 \leq \alpha < 2. \tag{5.50}$$

Proof.

(a) In theorem 5.6(a), replacing the SPOUT orders (5.44) with the SPOUT-E orders (5.48), produces the same expression for $x_{t,0}$, which is (B.16). As $x_{t,k} = x_{t,0} - \sum_{n=1}^{k} o_{t,n}$ we obtain

$$\begin{aligned} x_{t,k} &= x_k^* - x_0^* + \frac{\alpha k}{P}\left(x_0^* - x_{t,0}\right) \\ &= x_k^* - \sum_{n=1}^{P}\sum_{m=1}^{q} \frac{P-\alpha k}{P}\left(1-\alpha\right)^{m-1} \varepsilon_{t-mP+n}. \end{aligned} \tag{5.51}$$

Taking the variance of $x_{t,k}$ and adding the variance of lead-time demand, $\sigma_d^2\,(k+L)$ gives (5.49) completing this part of the proof.

(b) Inserting (B.16) in (5.48) provides

$$o_{t,k} = x_k^* - x_{k-1}^* + \sum_{n=1}^{P} \sum_{m=1}^{q} \frac{\alpha}{P} (1-\alpha)^{m-1} \varepsilon_{t-mP+n}. \tag{5.52}$$

Taking the variance gives (5.50), completing the proof. ■

The order variances, and hence the capacity costs, of SPOUT-E and SPOUT are identical, but the inventory cost of SPOUT-E is higher, rendering the equal-overtime policy inferior to the once-per-cycle setting. Similarly, under piecewise-linear costs, the bullwhip-optimal policy performs worse than SPOUT but better than SPOUT-E, as a result of the overtime allocation. The SPOUT-E policy is stationary under the same conditions as SPOUT, as the total order quantity over an entire cycle is identical between the two policies. To confirm this, observe that under both SPOUT and SPOUT-E,

$$x_{t,0} - \mathbb{E}[x_{t,0}] = (1-\alpha)(x_{t-P,0} - \mathbb{E}[x_{t-P,0}]) - \sum_{n=0}^{P-1} \varepsilon_{t-P+n}, \tag{5.53}$$

where it is evident that $x_{t,0}$ is a first-order system, stable when $|(1-\alpha)| < 1$. It follows that the inventory positions under SPOUT-E are stationary, as they are linearly dependent on $x_{t,0}$ following $x_{t,k} - \mathbb{E}[x_{t,k}] = (1 - \alpha k P^{-1})(x_{t,0} - \mathbb{E}[x_{t,0}])$.

5.4 Numerical study

To further explore the consequences of the four strategies, consider the set-up $\{\mu = 10,\ \sigma_d = 1,\ h = 1,\ b = 9,\ u = 40,\ v = 60\}$. With these settings, the optimal STOUT configuration is $P^* = 17$, and the optimal SPOUT setting is $\{P^* = 1,\ \alpha^* = 0.074\}$ for the SPOUT policy. The total cost of each strategy is illustrated in Figure 5.7, where safety stocks are optimal, and α^* has been optimized numerically for STOUT[-E] configurations where $P > 1$.

The SPOUT policy gives the lowest total cost, regardless of P. The cost advantage that can be gained from improving the ordering policy in these two areas depends on P, as the adoption of production smoothing has a significant economic impact if P is small. When P is large the greatest savings come from changing the overtime strategy so that overtime production is collected at the start of the order cycle.

When the order cycle is short, the production strategies without smoothing (STOUT[-E]) suffer from high costs, while both of the smoothing policies

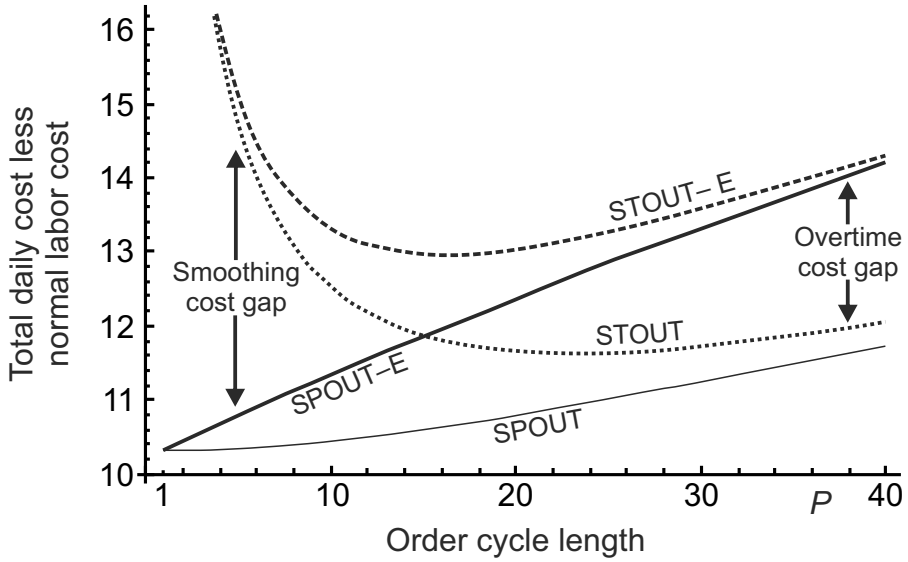

Figure 5.7: Avoid the capacity cost trap by improving the order policy before reducing the order cycle length.

perform better. In this case, the economic potential of smoothing production is greater than that of changing overtime strategy. The opposite holds true when the order cycle is long, as smoothing then has a small impact on the total cost. In these cases, smart overtime management is more important.

Irrespective of demand variability and cost factors, the production control policy should collect the inventory corrections to a short period, such as a weekend, always at the beginning of the production cycle. It is then desirable to react in a moderate but timely fashion to keep the overtime costs in check, hence the need for production smoothing.

The STOUT policy does not smooth production, and suffers from impaired efficiency. When the policy is used, the order cycle plays an important role in the balancing of capacity and inventory costs. If the order cycle is too long, inventory costs dominate; if it is too short, capacity costs inflate. We shall now detail how the four policies are applied in practice.

Example 5.9. Suppose we are to place orders for a system with the same parameter settings as the preceding numerical example, and that the initial observed inventory position is $x_{0,0} = 47$. Machine precision should be used; accordingly the numerical calculations below are truncated instead of rounded. To calculate the orders in a cycle:

1. Determine the target inventory positions, x_k^*, in the cycle. From (5.22) we obtain $x_k^* = \mu(k+L) + \sigma_{i,k}\Phi^{-1}\left[b/(b+h)\right] = 10(k+5) + \sqrt{k+5}\Phi^{-1}(0.9)$. For the first period, inserting $k = 1$ gives $x_1^* = 63.13$. Increment k in the

same calculation to obtain the remaining values of x_k^*. For policies other than STOUT, $\sigma_{i,k}$ is calculated differently: for STOUT-E, use (5.38) for SPOUT, use (5.45); and for SPOUT-E, use (5.49).

2. Obtain the deterministic production requirement by calculating the difference between the target inventory positions of consecutive periods. Take $x_k^* - x_{k-1}^*$ for every value of k. For $k = 1$ this is $x_1^* - x_0^* = 63.13 - 54.05 = 9.08$, continuing with $k = 2$, $x_2^* - x_1^* = 10.25$, and so forth. For each period k (and for every policy), order the deterministic production requirement.

3. Calculate the deficit between the target and the actual inventory position, $x_{t,0}^* - x_{t,0} = 63.13 - 47 = 16.13$.

   STOUT: Add the entire deficit to the first order ($o_{t,1}$).

   STOUT-E: Divide the deficit by P; add to every order ($o_{t,1}$ to $o_{t,P}$).

   SPOUT: Multiply the deficit by α; add to the first order.

   SPOUT-E: Multiply the deficit by α; divide by P, add the resulting amount to every order.

Complete calculations are performed in Table 5.1.

5.5 Managerial insights

In a practical situation we are likely to start with an existing system where the order cycle is already defined. There will be a selection among possible improvements, where the ones with a high return-on-investment will be chosen. To find suitable candidates we note that:

- *Production smoothing* is likely to be the most important improvement to a planning system when the order cycle is short (e.g. daily or weekly). For infrequent planning (e.g. monthly), the benefit is less pronounced.
- Assigning *overtime* only to the first order in each cycle reflects a good opportunity to reduce the average replenishment time, and hence the inventory costs.
- *Safety stock optimization* is a well-known practice for reducing inventory costs, but we have shown that time-invariant safety stocks lead to unnecessary and avoidable costs. The optimal solution is a dynamic safety stock setting.

Table 5.1: Calculating the orders to be received in periods 6–10.

	Period t	5	6	7	8	9	10	11
	Index k	5	1	2	3	4	5	1
STOUT	x_k^*	-	63.13	73.39	83.62	93.84	104.05	-
	$x_k^* - x_{k-1}^*$	-	9.08	10.25	10.23	10.21	10.20	-
	$x_0^* - x_{0,0}$	-	7.05	↓	↓	↓	↓	-
	$o_{0,k}$	-	16.13	10.25	10.23	10.21	10.20	-
STOUT-E	x_k^*	-	63.88	73.80	83.80	93.88	104.05	-
	$x_k^* - x_{k-1}^*$	-	9.83	9.91	10	10.08	10.16	-
	$(x_0^* - x_{0,0})/P$	-	1.41	1.41	1.41	1.41	1.41	-
	$o_{0,k}$	-	11.24	11.32	11.41	11.49	11.57	-
SPOUT[a]	x_k^*	-	64.77	74.94	85.10	95.26	105.41	-
	$x_k^* - x_{k-1}^*$	-	9.35	10.16	10.16	10.15	10.15	-
	$\alpha(x_0^* - x_{0,0})$	-	8.41	↓	↓	↓	↓	-
	$o_{0,k}$	-	17.77	10.16	10.16	10.15	10.15	-
SPOUT-E[b]	x_k^*	-	65.45	75.44	85.44	95.45	105.47	-
	$x_k^* - x_{k-1}^*$	-	9.98	9.99	10	10.00	10.01	-
	$\alpha(x_0^* - x_{0,0})/P$	-	1.69	1.69	1.69	1.69	1.69	-
	$o_{0,k}$	-	11.67	11.68	11.69	11.70	11.70	-

Dashes (-) refer to values unrelated to the present ordering decision.
[a] $\alpha^* = 0.217944$.
[b] $\alpha^* = 0.211445$; numerically optimized.

- *Lead time reduction* helps to decrease the inventory variance, and is viable for both long and short order cycles. Unlike the previous improvements, this requires changes outside of the planning system, which may be costly to effect.
- Shortening the *order cycle* allows supply chains that already use smooth production to reduce their costs further. However, if smoothing is not in place, capacity costs may soar as the order cycle is shortened.

To identify which of these improvement strategies are best for a particular supply chain, we may build a System Dynamics simulation model based on the system equations (4.2) and (4.1) in combination with one of the replenishment policies (5.23), (5.37), (5.44), or (5.48). Company-specific factors can also be included, as well as nonlinearities and specific cost assumptions, to give deeper insight into the supply chain at hand.

5.6 Conclusion

The previous chapter showed that the inventory variance becomes heteroskedastic when inventory inspections take place more frequently than production planning. This effect is aggravated by distributing overtime evenly over the order cycle, and can be counteracted by shortening the order cycle, or by collecting overtime to the beginning of the cycle. However, short order cycles may themselves be harmful. Unless an appropriate production smoothing policy is in place, capacity costs may increase as the production system alternates between states of high and low utilization. Before the order cycle is shortened, supply chains must have the capability for production smoothing. These insights allow us to formulate a recipe for the design of highly efficient planning systems. Its constituents are: fast reordering, smooth production, and smart overtime planning.

In terms of the research questions, this chapter has introduced three linear policies capable of production smoothing: the bullwhip-optimal policy, SPOUT, and SPOUT-E, resolving RQ 4. Showing that SPOUT and STOUT were preferable to SPOUT-E and STOUT-E revealed that overtime should be placed at the start of each cycle, answering RQ 5. If one uses the STOUT policy, we have provided a procedure for finding the optimal order cycle length, and shown that it may be greater than unity; this answers RQ 6. Finally, we have shown that it is preferable to embed production smoothing in the order policy (SPOUT), instead of extending the order cycle length of a policy without smoothing (STOUT), which addresses RQ 7. We note that no previous investigation in the literature has considered capacity costs. The closest would be the pragmatic policy proposed by Chiang (2009), which operates like STOUT with a *linear* time-varying safety stock. The optimization of the order cycle length based on the capacity cost also takes the idea from Flynn and Garstka (1997) in another direction, as we have shown that the optimal order cycle length can be driven by other trade-offs than the one between inventory and audit costs. The bullwhip-optimal policy is also new for staggered systems, but reflects on the non-staggered literature, in particular Disney et al. (2004), which identifies the same optimal configurations for non-staggered systems as this chapter identifies for the first period of each cycle ($k = 1$ in Figure 5.1).

Chapter 6

Verification and validation

The analytical results have exposed many properties of PIC systems with staggered deliveries. The proofs provide that these results are consistent with the underlying assumptions, but we have yet to test the validity of the models. This chapter takes a three-pronged approach to validation: First, an industrial example reveals that staggered deliveries occur as specified in this thesis, with an identifiable order cycle length P and lead time L. Second, the models are shown to be consistent with familiar results from the literature. Third, the numerical output of the mathematical models is compared with the equivalent output from simulation models. Fourth, tests against validation criteria reveal that the analytical model responds as expected under a range of conditions, that it is consistent with established theory, and that staggered deliveries are used industrially.

6.1 Industrial example

To see if staggered deliveries appear in an industrial context, the order fulfilment process of a factory was mapped. Process maps must be made at a level of aggregation appropriate for the problem at hand Rummler and Brache (1995, p. 33). As this thesis models staggered deliveries on the MPS level, we opted to map the process as a master planner would see it. To obtain this view, two planners were interviewed at the same occasion, allowing them to discuss and clarify their views on the process before describing it. The planners were then asked to describe how an order passes through the planning system and production, from the time it is entered, to the time it is fulfilled with a physical delivery (Rummler and Brache, 1995, pp. 50–52). In addition, the planners were asked about how performance was measured, materials supply, and personnel

and material resources necessary for the process to operate. After collecting the data, the process, as it had been interpreted, was described to the planners, so that they could correct any misinterpretations.

6.1.1 The order fulfilment process

A company in Western Europe (referred to as Alpha for anonymity) produces coinage and commemorative medals. We shall focus on coins for circulation, for which there are seven parallel production lines, operating largely in the same way. For simplicity, we shall focus on the production lines that use steel as a raw material. In essence, the production ofha aims for full capacity utilization, which they achieve by managing an order book.

Customers issue a request for tenders. Alpha submits a tender, if there is sufficient capacity to deliver the full quantity on time — timely deliveries are critical, as late fulfilment leads to heavy penalties. Awarded contracts enter an order book, usually several months before they are due. From the order book, a tentative production plan is generated, where contracts are prioritized by their due dates. The sequence of order releases is fixed every Tuesday, for seven days of production (i.e. $P = 7$), but the orders are released to the shop floor on Friday, three days later. Production planning does not place a limit on the maximum production quantity, as the intention is to produce as much as possible. If one week's planned production is completed ahead of time, the orders for the next week enter production ahead of schedule. The time spent in physical production varies between six and twelve days, providing an order-to-receipt lead time $L \in [8, 14]$, where the nominal sequence-of-events delay has been subtracted.

In detail, the production process consists of two major subprocesses, with a strategic buffer in between. The first subprocess is blanking, and the second is a straight flow (first in first out) though the operations of plating, annealing, finishing, striking, and telling, all joined by intermediate buffers that absorb variations in throughput (see Figure 6.1). This design allows the process to be managed via the theory of constraints (Goldratt, 2004) — the bottleneck operation (plating) operates around the clock, including weekends, and appears at the start of the second subprocess. To ensure that the bottleneck is not starved of materials, the preceding buffer contains materials for 32–72 hours of production. This buffer level is maintained by controlling, in continuous time, the throughput of the first subprocess (blanking). After the second subprocess ends, the coins change ownership to the customer, and are either stored or

shipped, based on customer preference. Large orders are split into multiple shipments, sent with a frequency that usually ranges from a week to a month; this is reminiscent of the lot splitting model in Chiang (2001).

The raw materials needed for the first subprocess is steel, which is ordered once per week, but delivered once or twice per day. Here there are up to fourteen staggered deliveries per order cycle, but should we assume that the inventory is tallied only at the end of each day, then $P = 7$. When a new order is placed, the first shipment arrives sometime the following day, i.e. $L = 0$ for plans made at the end of the day. The steel orders are planned weekly so that the raw materials inventory should contain sufficient materials for 36 hours of production. The order quantities were based on judgment, but it is likely that the order decisions resembled the STOUT policy, as it immediately brings the inventory position to its target value.

6.1.2 Comparison with the model in this thesis

This industrial example includes one true example of staggered deliveries, which is the weekly ordering of steel, with deliveries occurring every day. This supply loop includes a random demand component, which is the material consumption by blanking. As the steel supply is not concerned with capacity costs, it matches closely the model in Chapter 4, with $P = 7$ or $P = 14$ and $L = 0$. Hence, we have established the existence of a system that follows the concept of staggered deliveries. Although the case verified that staggering occurs, this occurred in a raw materials buffer, and the demand (i.e. consumption) reflects the consumption of the plating subprocess, plus eventual yield losses in blanking. It would be desirable to numerically test for the heteroskedasticity induced by staggered deliveries, but this industrial case was unsuitable primarily because there was no record of demand (plating consumption), which is required to predict the variances of inventory and orders. In addition, we do not know how well the ordering of raw materials corresponds with STOUT, as planners may decide to round the order quantity to full truckloads, or make other adjustments. Taking these uncertainties into account, it may be difficult to provide definite evidence of the effects of staggering.

The production mechanism itself is not staggered by our definition. An order release sequence is fixed once per week, but the production pipeline is not committed to any order quantity per period. Instead, orders are released to the shop floor continuously, based on material consumption in the strategic buffer after blanking. However, the shipping to customers is staggered, with

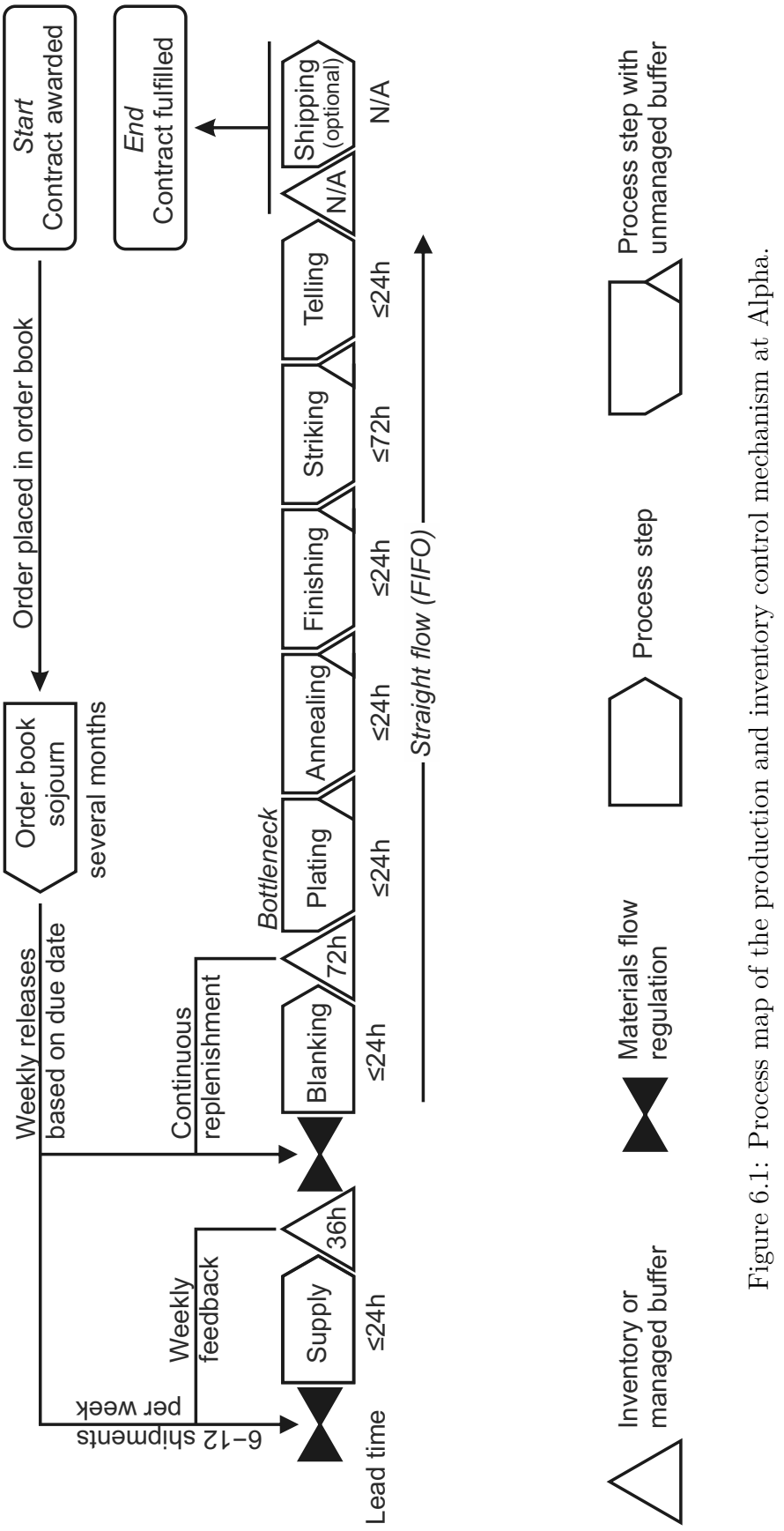

Figure 6.1: Process map of the production and inventory control mechanism at Alpha.

the particular property that each contract involves a single order cycle, making it akin to a newsvendor model with staggered deliveries.

6.2 Testing the analytical models

To see if the analytical models are reasonable, we may compare them with established models. For the optimal staggered policy under autocorrelated demand, the optimal policy is of the OUT type. This implies that the OUT policy is optimal for i.i.d. demand, which was proven by Flynn and Garstka (1990). Compared to non-staggered settings, we find that the inventory variance (4.15) is identical to the variance of the forecast error over τ periods, which is given as the equation (5.1.16) in Box and Jenkins (1976). This is even clearer for i.i.d. demand, where the standard deviation of the inventory level is $\sigma_\varepsilon\sqrt{\tau}$ — an expression commonly used when calculating optimal safety stocks. Setting $P = 1$ provides the regular OUT policy under a minimum-mean-square-error forecast. The inventory variance for such a system is provided in equation (3.4) of Lee et al. (2000) as

$$\sigma_\varepsilon^2 \sum_{n=1}^{L+1} \left(\sum_{m=0}^{n-1} \phi^m \right)^2, \tag{6.1}$$

expressed in the notation of this thesis. This expression is identical to (4.15) in this thesis, with $\theta_t = \phi^t$. Simplifying (6.1) also produces our variance for AR(1) demand (4.24), when $k = 1$. We have thus verified that the ARSTOUT policy is a generalization of the inventory-optimal non-staggered policy for AR(1) demand. It is easy to see that this is the also the case for any autocorrelated demand process.

To test the bullwhip-optimal policy and the SPOUT[-E] policies, we set $P = 1$ to produce a non-staggered system. The three policies then coincide, as they differ only in the distribution of overtime work over the order cycle. As the policies are proportional, they should be identical to the DE-APIOBPCS model in Disney and Towill (2003) with $\mathrm{Ta} \to \infty$. Taking the impulse response of orders in (Disney and Towill, 2003, equation [5]) and substituting $\alpha = \mathrm{Ti}^{-1}$ for notational consistency, we obtain the impulse response of orders:

$$\alpha(1-\alpha)^t. \tag{6.2}$$

This can be compared with (B.18), where the same form appears. This verifies that the staggered proportional policies are generalizations of DE-APIOBPCS

when Ta $\to \infty$.

6.3 Simulation model and results

To test the analytical model, a simulation model was built in Microsoft Excel, chosen due to the prevalence of spreadsheet models in industry and research (Ciancimino et al., 2012). This provided new insights on how the staggered delivery concept may be implemented and analysed in an industrial setting. Compared to two early prototypes built in C++ and Matlab, the Excel model provided some distinct advantages and shortcomings. First, Excel provides immediate numerical feedback for every formula entered; if one knows what results to expect, this serves as a mechanism for testing the validity of each component (not just the output) in the simulation model. There were two main disadvantages to model development in Excel: First, the program flow (sequence of execution) is not obvious from a spreadsheet, and one must be cautious to ensure that the flow and dependencies are as intended. Second, large Excel formulas may be difficult to write and to read.

6.3.1 Model design

The basic structure of the simulation model was implemented using (4.1) and (4.2), along with w as as intermediary variable. AR(1) demand was modelled as

$$d_t = \phi(d_{t-1} - \mu) + \mu + \varepsilon_t, \tag{6.3}$$

where $\varepsilon_t = \sigma_\varepsilon \Phi^{-1}(y_t)$, in which y_t is a uniformly distributed random variable on $[0, 1]$, generated in Excel. Negative demand (returns) is permitted, causing an increase of the inventory level, in agreement with (4.2). For the STOUT policy under autocorrelated demand, (4.6) and (4.7) were used, while for the STOUT[-E] / SPOUT[-E] policies, (5.23) (5.37), (5.44), and (5.48) were used. Both orders $o_{t,k}$ and receipts r_{t+k+L} were modelled to make it clear when orders were determined and when they were received as inventory.

6.3.2 Output variables

The analytical work presented several output variables to be tested in the model. The inventory costs were calculated as the average inventory cost per period using (4.3), while the availability (S_1) was obtained by calculating the number of periods with $i_t > 0$ and dividing it the total number of periods in the time

range. The exact fill rate was calculated by taking the sum (over all periods in the time range) of the ability to fulfil demand $\sum_t [\min(d_t, i_t + d_t)]^+$, and by dividing this by the sum of all demand that it was possible to fulfil $\sum_t (d_t)^+$. Also measured is the inventory variance for each k, the overall inventory variance $\hat{\sigma}^2(i_t)$ (measured for all k at once), and the variance of the orders (for each k).

Table 6.1: Description of simulation output variables.

Symbol	Description
$\hat{J}$	Average inventory cost
$\hat{A}$	Average capacity cost
$\hat{S}_1$	Availability
$\hat{S}_2^-$	Exact fill rate
$\hat{\sigma}^2(i_t)$	Variance of the inventory level

6.3.3 Experiment design

The analytical results depend strongly on the autocorrelation of demand, ϕ, and it is therefore desirable to test the system performance under different settings. For this reason, seven configurations of AR(1) demand are tested. For the other parameters, the setting $\{b = 9, h = 1, \mu = 10, \sigma_\varepsilon = 1, L = 4, P = 5\}$ was chosen. With this setting of b and h, costs are minimized when $S_1 = 90\%$.

For the SPOUT/STOUT[-E] policies it was assumed that $\{b = 19, h = 1, \mu = 10, \sigma_\varepsilon = 1, P = 5\}$, implying that costs are minimized when $S_1 = 95\%$. A short and a long lead time, $L = 0$ and $L = 8$, were considered. The smoothing parameter was optimized numerically in Mathematica, as the minimizing value of A, defined in (5.27). Table 6.2 presents the optimized values of α.

Table 6.2: Optimized settings for the smoothing parameter α.

L		α^*
0	SPOUT	0.354821
	SPOUT-E	0.328498
8	SPOUT	0.274583
	SPOUT-E	0.267431

When simulating the variance-optimal policy, $L = 2$ and $\sigma_d = 1$. As we are interested in variances only, $\mu = 0$, without loss of generality. For each parameter set, the simulation result represents the average output of 200

simulations, each containing 50k simulated periods. As we have assumed $P = 5$, each simulation contains 20k data points for each period of the cycle.

6.3.4 The impulse response

The analytical expressions for order and inventory variances rely on algebraic manipulation of the systems equations (4.8), (4.1), and (4.2), along with one of the policy equations (5.23) (5.37), (5.44), and (5.48). To obtain the variance of a variable in a linear system analytically, we must first express the variable of interest as a weighted sum of (time-shifted) i.i.d. random variables. This leads to a straightforward calculation of the variance (Box and Jenkins, 1976; Tsypkin, 1964). As the deterministic components of the process are known, the variance calculation is simplified by setting $\mu = 0$ and all safety stocks $i_k^* = 0$ and OUT levels $x_k^* = 0$.

The coefficients in this weighted sum of random variables can also be obtained from a simulation model where the i.i.d. random noise is replaced with an impulse, i.e. $d_t = \delta(t - n), n \in \{0, 1, \ldots, P - 1\}$, where $\delta(\cdot)$ is the Kronecker delta function,

$$\delta(t) = \begin{cases} 1 & \text{if } t = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

Note the occurrence of n in the impulse response. This is a consequence of the system having multiple inputs, which also requires one impulse to be generated for each n. The variance of the inventory level is then obtained as

$$\sigma^2(i_{t+k+L}) = \sum_{n=0}^{P-1} \sum_{m=0}^{\infty} i_{t+k+L+mP}^2. \tag{6.5}$$

where n affects i_t via d_t. The variance of the orders is calculated in the same way. This has no bearing on the variances produced. The impulse response lets us obtain the variance from a deterministic simulation, complementing the analytical results, and the stochastic simulation.

6.3.5 Results

The results for the optimal policy under autocorrelated demand are presented next to the analytical results in Table 6.3 and Table 6.4. For all output variables, the simulation results are close to the analytical results. The results are less

precise for high values of ϕ, as this increases the inventory variance. If more precision is required, the simulation horizon may be extend. Notably, in tables 6.4, 6.6 and 6.7, the simulated impulse response variances are numerically identical to the analytically obtained variances. This is expected.

The results for the STOUT/SPOUT[-E] policies are presented in Table 6.5, and the variances appear in Table 6.6. While $\phi = 0$ here, the setting $L = 8$ increases the inventory variance, making these measurements less precise than when $L = 0$. This is also an expected result.

6.4 Tests against validation criteria

In Chapter 3, we identified some criteria for model validation. This section tests the staggered models of this thesis against the validation criteria.

1. *The model structure must be consistent with the descriptive knowledge of the system.* There should be no doubt that the relation between cyclical orders (4.1) and the inventory mechanism (4.2) represent the staggered deliveries concept as defined in Chapter 1. The industrial example provided in this chapter illustrates that the variables P and L used in the model can be identified in an industrial setting.

2. *The model should be causal, exhibiting the same behaviour as the real system, for the same reasons.* There was no empirically observed or expected output behaviour for costs or service levels, against which to compare the model. However, looking at the order variances in Table 6.6, it is clear that SPOUT allocates the overtime to one period, whereas SPOUT-E distributes it over the cycle. This result is consistent with our expectations of the model's behaviour.

3. *The model must not break physical laws.* The model follows the conservation-of-materials principle, which means that the inventory level equals the difference between the inflow (receipts, returns from customers) and outflow (demand, returns to supplier) of materials. No explicit capacity limits are implemented, which means that the order quantity can be large. This does not break any physical laws, but it limits the applicability of the model to systems with sufficient capacity. In addition, the model

Table 6.3: Simulation and analytical results comparison.

ϕ	J	$\hat{J}$	$\hat{S}_1$	S_2^-	$\hat{S}_2^-$	$\sigma^2(i_t)$	$\hat{\sigma}^2(i_t)$
−0.95	3.2095	3.2095	89.98%	99.13%	99.13%	3.41	3.41
−0.7	3.0514	3.0514	90.00%	99.18%	99.18%	3.08	3.07
−0.5	3.2968	3.2977	90.00%	99.11%	99.11%	3.60	3.60
0	4.6190	4.6154	90.00%	98.75%	98.75%	7.11	7.12
0.5	8.0529	8.0411	90.05%	97.84%	97.83%	22.00	22.05
0.7	11.1233	11.1175	90.00%	97.02%	97.02%	43.11	43.20
0.95	18.6677	18.4635	90.23%	95.41%	95.16%	132.19	132.66

Table 6.4: Inventory variances for each period in the cycle.

		ϕ						
	k	−0.95	−0.7	−0.5	0	0.5	0.7	0.95
Simulated	1	2.75	2.39	2.67	5.00	13.55	22.79	47.02
	2	2.77	2.66	3.11	6.00	17.43	31.39	74.81
	3	3.52	3.06	3.56	6.99	21.36	40.71	110.91
	4	3.56	3.37	4.00	7.99	25.32	50.54	155.72
	5	4.25	3.74	4.45	8.98	29.29	60.75	210.67
Impulse	1	2.75	2.39	2.68	5.00	13.58	22.79	47.17
	2	2.76	2.66	3.11	6.00	17.46	31.44	75.24
	3	3.52	3.06	3.56	7.00	21.40	40.80	111.64
	4	3.55	3.37	4.00	8.00	25.36	50.67	156.96
	5	4.25	3.74	4.45	9.00	29.35	60.90	211.64
Analytical	1	2.75	2.39	2.68	5	13.58	22.79	47.17
	2	2.76	2.66	3.11	6	17.46	31.44	75.24
	3	3.52	3.06	3.56	7	21.40	40.80	111.64
	4	3.55	3.37	4.00	8	25.36	50.67	156.96
	5	4.25	3.74	4.45	9	29.35	60.90	211.64

Table 6.5: Performance metrics of the STOUT/SPOUT[-E] policies.

L		J	$\hat{J}$	A	$\hat{A}$	$\hat{S}_1$	$\sigma^2(i_t)$	$\hat{\sigma}^2(i_t)$
0	STOUT	3.46	3.46	409.8	410.1	94.99%	3.51	3.51
	SPOUT	5.25	5.25	404.5	404.7	94.99%	6.78	6.76
	STOUT-E	4.22	4.22	409.8	409.8	95.00%	4.23	4.23
	SPOUT-E	6.17	6.17	404.3	404.4	94.99%	8.95	8.95
8	STOUT	6.83	6.83	409.8	410.1	95.02%	11.12	11.11
	SPOUT	8.38	8.39	403.9	404.0	94.97%	16.64	16.68
	STOUT-E	7.20	7.19	409.8	409.8	94.99%	12.21	12.21
	SPOUT-E	8.91	8.89	403.8	403.9	95.04%	18.67	18.60

is causal in the sense that it does not assume prescience, but uses only information available at the time of planning. This is evident from the replenishment policies, i.e. Theorem 4.2 for ARSTOUT, and (5.11), (5.23), (5.37), (5.44), or (5.48) for the policies considering overtime work.

4. *The decision rules in the model must reflect the behaviour of the decision-makers.* As the industrial case has shown, staggered deliveries occur similarly as in the model. The OUT policy is a standard policy. The SPOUT[-E] policies are based on the proportional OUT policy, which has seen industrial use for many years, as evidenced by Simon and Holt (1954) and Disney et al. (2013).

5. *The units of the stocks and flows in the model must be consistent.* With a discrete-time model operating on a per-period basis, both inventory levels and receipts are measured in units; hence all terms in the inventory equation (4.2) are expressed in units. The periodic holding and backorder costs in (4.3) are expressed in pounds sterling (£) per unit, as are the capacity and overtime costs in (5.24). The audit cost is expressed as pounds sterling per order cycle, which in (4.20) is divided by the number of periods per order cycle, producing the average audit cost in one period. For the total cost, regardless of model, we obtain a cost in pounds, implicitly per period. Service levels are dimensionless.

6. *The model should respond reasonably when parameters are set to extreme values.* An increase of any cost factor, $\{b, h, u, v\}$, causes the corresponding cost, (4.3), (5.24), or (4.20), to increase. The total cost also increases with the base variability of the system $\{\sigma_\varepsilon^2, \sigma_d^2\}$, and the lead time L. The increased cost is a reasonable response, as we are either increasing the

Table 6.6: Variances of the STOUT/SPOUT[-E] policies.

	L		$\sigma^2_{o,1}$	$\sigma^2_{o,2}$	$\sigma^2_{o,3}$	$\sigma^2_{o,4}$	$\sigma^2_{o,5}$	$\sigma^2_{i,1}$	$\sigma^2_{i,2}$	$\sigma^2_{i,3}$	$\sigma^2_{i,4}$	$\sigma^2_{i,5}$
Simulated	0	STOUT	5.006	0.000	0.000	0.000	0.000	0.999	1.999	3.000	4.002	5.006
		STOUT-E	0.200	0.200	0.200	0.200	0.200	4.198	3.794	3.800	4.196	4.996
		SPOUT	1.076	0.000	0.000	0.000	0.000	4.553	5.551	6.549	7.547	8.545
		SPOUT-E	0.039	0.039	0.039	0.039	0.039	8.953	8.879	8.875	8.948	9.107
	8	STOUT	4.992	0.000	0.000	0.000	0.000	9.003	9.989	11.024	11.976	12.935
		STOUT-E	0.200	0.200	0.200	0.200	0.200	12.198	11.798	11.803	12.200	12.997
		SPOUT	0.798	0.000	0.000	0.000	0.000	14.603	15.618	16.620	17.627	18.634
		SPOUT-E	0.031	0.031	0.031	0.031	0.031	18.605	18.564	18.562	18.628	18.739
Impulse	0	STOUT	5.000	0.000	0.000	0.000	0.000	1.000	2.000	3.000	4.000	5.000
		STOUT-E	0.200	0.200	0.200	0.200	0.200	4.200	3.800	3.800	4.200	5.000
		SPOUT	1.078	0.000	0.000	0.000	0.000	4.565	5.565	6.565	7.565	8.565
		SPOUT-E	0.039	0.039	0.039	0.039	0.039	8.949	8.870	8.870	8.949	9.106
	8	STOUT	5.000	0.000	0.000	0.000	0.000	9.000	10.000	11.000	12.000	13.000
		STOUT-E	0.200	0.200	0.200	0.200	0.200	12.200	11.800	11.800	12.200	13.000
		SPOUT	0.796	0.000	0.000	0.000	0.000	14.554	15.554	16.554	17.554	18.554
		SPOUT-E	0.031	0.031	0.031	0.031	0.031	18.668	18.606	18.606	18.668	18.791
Analytical	0	STOUT	5	0	0	0	0	1	2	3	4	5
		STOUT-E	0.2	0.2	0.2	0.2	0.2	4.2	3.8	3.8	4.2	5
		SPOUT	1.078	0	0	0	0	4.565	5.565	6.565	7.565	8.565
		SPOUT-E	0.039	0.039	0.039	0.039	0.039	8.949	8.870	8.870	8.949	9.106
	8	STOUT	5	0	0	0	0	9	10	11	12	13
		STOUT-E	0.2	0.2	0.2	0.2	0.2	12.2	11.8	11.8	12.2	13
		SPOUT	0.796	0	0	0	0	14.554	15.554	16.554	17.554	18.554
		SPOUT-E	0.031	0.031	0.031	0.031	0.031	18.668	18.606	18.606	18.668	18.791

Table 6.7: Variances of the bullwhip-optimal policy.

	λ	$\sigma^2_{o,1}$	$\sigma^2_{o,2}$	$\sigma^2_{o,3}$	$\sigma^2_{o,4}$	$\sigma^2_{o,5}$	$\sigma^2_{i,1}$	$\sigma^2_{i,2}$	$\sigma^2_{i,3}$	$\sigma^2_{i,4}$	$\sigma^2_{i,5}$
Simulation	0.0	0.000	0.000	0.000	0.000	0.000	-	-	-	-	-
	0.1	0.414	0.213	0.110	0.056	0.029	5.633	5.344	5.705	6.350	7.168
	0.3	1.126	0.311	0.086	0.024	0.007	4.371	4.374	5.099	6.015	6.976
	0.5	1.909	0.279	0.041	0.006	0.001	3.726	4.093	5.007	5.989	6.982
	0.7	2.849	0.170	0.010	0.001	0.000	3.282	3.995	4.960	5.957	6.969
	0.9	4.133	0.035	0.000	0.000	0.000	3.050	4.018	5.014	5.991	6.999
	1.0	5.001	0.000	0.000	0.000	0.000	2.994	3.994	4.999	6.010	6.991
Impulse	0.0	0.000	0.000	0.000	0.000	0.000	-	-	-	-	-
	0.1	0.414	0.213	0.110	0.056	0.029	5.672	5.376	5.709	6.365	7.188
	0.3	1.128	0.311	0.086	0.024	0.007	4.383	4.382	5.105	6.029	7.008
	0.5	1.910	0.279	0.041	0.006	0.001	3.730	4.106	5.016	6.002	7.000
	0.7	2.853	0.171	0.010	0.001	0.000	3.299	4.018	3.001	4.000	5.000
	0.9	4.125	0.035	0.000	0.000	0.000	3.042	4.000	3.000	4.000	5.000
	1.0	5.000	0.000	0.000	0.000	0.000	3.000	4.000	5.000	6.000	7.000
Analytical	0.0	0	0	0	0	0	-	-	-	-	-
	0.1	0.414	0.213	0.110	0.056	0.029	5.672	5.376	5.709	6.365	7.188
	0.3	1.128	0.311	0.086	0.024	0.007	4.383	4.382	5.105	6.029	7.008
	0.5	1.910	0.279	0.041	0.006	0.001	3.730	4.106	5.016	6.002	7.000
	0.7	2.853	0.171	0.010	0.001	0.000	3.299	4.018	5.001	6.000	7.000
	0.9	4.125	0.035	0.000	0.000	0.000	3.042	4.000	5.000	6.000	7.000
	1.0	5	0	0	0	0	3	4	5	6	7

cost of variability, or increasing the variability of the system. The order cycle length is a special case. We have established that the inventory cost is minimized when $P = 1$. Note that as $P \to \infty$, the inventory variance and the inventory cost also tend toward infinity, while the audit or overtime costs per period tend toward zero.

7. *The model variables must be unambiguous and quantifiable.* Both inventory levels and orders are quantified as units, and all costs are applied on a per-period basis, except the audit cost, which is applied once per order cycle. All of the variables are measurable in principle, over an infinite time horizon.

8. *The model must be able to generate the same behaviour as related models.* We have established that the ARSTOUT model is a generalization of the OUT policy under AR(1) demand via Lee et al. (2000). The bullwhip-optimal policy and SPOUT[-E] are all generalizations of the DE-APIOBPCS policy with Ta $\to \infty$, as evidenced by a comparison with Disney and Towill (2003).

6.5 Summary

This chapter has shown that when the staggering is removed by setting $P = 1$, the policies in this thesis correspond to established models in the literature. For the inventory-optimal, the correspondence is to the forecast error over the lead time, while for the smoothing policies, the correspondence is to DE-APIOBPCS. For empirical justification, an industrial example from Alpha revealed that staggering occurs in the orders to the steel supplier, and that L and P are readily identifiable. The simulation model provided that the analytical variances of inventory and orders match those obtained when simulating the impulse response. Furthermore, inventory costs, capacity costs, service levels, and variances are consistent between the analytical models and stochastic simulation models. This correspondence is not exact, as the simulation is sampled over a limited time, but it is reasonably close for all output variables. This discrepancy is the greatest when the order or inventory variances are large, either due to L or to ϕ, as expected. In conclusion, the analytical models operate as intended.

Chapter 7

Conclusion

We set out to understand staggered deliveries and their implications. Answers have been found, at least for the set of assumptions we posited. This final chapter ties together the work, by returning to the research questions, and stating the resolution to each of them. Moving on, we discuss the theoretical results as a whole, and illuminate noteworthy features of the investigation. This is followed by some comments on the limitations of this study, as well as further work.

7.1 Review of research questions

1. *What is the inventory-optimal policy under staggered deliveries and autocorrelated demand?* The inventory optimal policy is a staggered variant of the OUT policy, combined with a minimum-mean-squared-error forecast. This is provided in Theorem 4.2. In addition to this, Chapter 4 reveals that each receipt in the cycle needs its own forecast and safety stock level. The optimal safety stock levels tend to change over the cycle.

2. *How do costs and service levels develop under staggered deliveries and autocorrelated demand?* Chapter 4 shows that costs and fill rate vary over the periods in a planning cycle, but availability remains constant (Lemma 4.1). Observing individual periods, we see that the inventory variance is increasing, as we progress from the first to the last receipt of a cycle (Theorem 4.4). The reason for the constant availability is that the cost-optimal policy is myopic, sharing the same structure as the newsvendor problem. Therefore, we set the safety stock (or base stock level) so that the probability of encountering a stockout is constant.

3. *Under inventory costs and audit costs, can an optimal reorder cycle length be identified when demand is autocorrelated?* Yes, Theorem 4.5 proves this, and a procedure for identifying the optimal order cycle length appears in Chapter 4. Longer order cycle lengths lead to increased inventory costs, but a lower audit cost per period. The optimal order cycle length varies with the autocorrelation of demand.

4. *How can a linear production smoothing policy be applied under staggered deliveries and i.i.d. demand?* Chapter 5 presents various ways to do this: It can be applied differently to every period in the cycle (to minimize the sum of the order rate variance and the inventory variance), to the first period of the cycle (minimizing the inventory variance). Distributing the overtime evenly over the cycle is uneconomic. The policies considered for this research question are the bullwhip-optimal policy (5.11), STOUT (5.23), STOUT-E (5.37), SPOUT (5.44) and SPOUT-E (5.48). These policies are investigated in detail in Chapter 5.

5. *How do overtime work rules affect the performance of systems with staggered deliveries?* We have considered two cost models. For the quadratic one, costs are minimized when different amounts of overtime are worked within each period, as provided by (5.11) and the analysis that precedes it. For the piecewise linear costs, we should opt to do all overtime work in the first period of the cycle (Considering the contribution of the inventory and order variances in Lemma 5.3 and 5.5 to the total cost (5.27), reveals that STOUT and STOUT-E have identical *average* capacity costs, but that STOUT-E has a higher inventory cost). We may choose between fixed or variable safety stocks. The fixed setting increases inventory costs, and causes availability to fluctuate. The variable setting minimizes costs, and keeps availability constant over the cycle.

6. *Can an optimum order cycle length be identified when capacity and inventory costs are present?* Yes, for the STOUT policy we may find an optimal cycle length distinct from unity, as long order cycles level production (Theorem 5.4). The bullwhip-optimal policy minimizes the quadratic costs under a unit order cycle length, as can be seen from Theorem 5.2. The optimal order cycle length is not known for SPOUT.

7. *How does the order cycle length interact with production smoothing?* It depends on the control policy. Given a policy capable of production

smoothing, there are no benefits of extending the order cycle (Lemma 5.7). Then we should opt for a non-staggered policy. But if we have a simple policy, such as the order-up-to policy, then we can see smoothing benefits from staggered deliveries. On the flip side, shortening the reorder cycle can lead to increased capacity costs, if a primitive ordering policy is used (illustrated by Figure 5.7).

7.2 Review of results

Staggering fixes orders for a limited time in the future, giving us a sense of constancy in the face of fluctuating demand. But this act of decoupling comes at a cost. In effect, we turn a blind eye to the state of the system for a while. When we finally decide to check, the state may have drifted from its last position. This happens for autocorrelated demand. As we have seen in (4.15), the inventory variance is a non-decreasing function of k, denoting the k'th day since the start of the order cycle.

Deriving the optimal policy for autocorrelated demand and linear holding and backlog costs, we found that it is a sequence of myopic order-up-to decisions. The inventory variance is the same as the forecast error from the planning occasion to the corresponding inventory receipt. As the receipts in a single cycle are separated in time, we effectively have different lead times, and the inventory variance increases through the individual days of a cycle. The inventory costs are proportional with the standard deviation of the inventory level, so we know they are non-decreasing with the cycle length P.

Service levels are affected by the non-decreasing inventory variance. For the cost-optimal policy, the safety stock varies to accommodate these changes in variance, leading to a fixed availability corresponding to the critical fractile $b/(b+h)$ obtained from the holding and shortage costs. This is not the case for the fill rate, which still varies over the cycle. In a numerical example, we found that this difference is most significant for positively autocorrelated demand.

First, we considered inventory costs only. The introduction of capacity costs adds further depth to our analysis. To keep the model tractable, yet insightful, we assumed that demand is i.i.d. and not generally autocorrelated. This reduces the number of variables, and helps to avoid the curse of dimensionality (Bellman, 1956). First considering quadratic (variance-based) costs, we find that the optimal way to plan overtime includes a different amount of overtime or idling in every period of the cycle, as (5.11) shows. The optimal policy is linear,

and takes the inventory position as its sole state variable. The policy is also a special case of APIOBPCS, and particularly, the Deziel-Eilon configuration, where the same feedback parameter is applied to WIP deficits and to inventory deficits.

The optimal order quantity can be thought of as an ordinary proportional policy where we, for every order, correct a fraction of the remaining deficit in the inventory position, based on the inventory position at the start of the cycle. The principle of certainty equivalence permits us to do this, and as it states, produce the quantity that would be optimal, if no further random deviations would occur. Only expected values need to be accounted for. Under the quadratic-optimal policy, the optimal reorder cycle length is unity.

Circumstances may dictate that overtime should be done once per cycle, or that the overtime work should be equally distributed over all of the days in the cycle. For a low quadratic cost, the former is preferable to the latter. These two policies have the same order rate variance for the same amount of smoothing, despite having different implementations. The overtime-once policy obtains its minimum inventory variance in the first period of the cycle, whereas the equal-overtime policy achieves is minimum inventory variance in the middle of the cycle.

While designs for quadratic costs are excellent for keeping system states close to a target value, they do not distinguish between deviations that are are above or below the average. One way to achieve asymmetric costs has been via piecewise-linear convex functions, where the costs of normal and overtime production differ. Under such a cost model, the optimal policy is a non-staggered threshold policy (Sobel, 1970). Of the linear policies we presented, the overtime-once policy is preferable to the optimal policy under quadratic cost, which in turn is preferable to the equal-overtime policy.

In practice, we may find that simple policies are used, such as the OUT policy. The overtime work rules considered are as before: overtime-once and equal-overtime. In this case, *the optimum reorder period need not be unity* for piecewise-linear costs. This results from an order rate smoothing effect. With optimal capacity levels, the expected capacity cost is proportional to the standard deviation of the orders in each period. Therefore, we have a pooling effect when the order-up-to policy is staggered. By no means does this imply that companies should increase the reorder cycle length while maintaining a primitive policy. Rather, they should seek to implement a more suitable control policy. A greater danger may lie in well-intentioned attempts to reduce

lead times by shortening the order cycle. Unless an appropriate policy is in place, capacity costs may increase to such an extent that inventory savings are nullified.

7.3 Managerial implications

Taking a manager's point of view, we are likely to have an existing production and inventory system, designed for functionality, but not necessarily optimized for performance. Therefore, we would do well to map out the path from a simple functioning system, to a highly efficient one. Figure 7.1 illustrates such a path.

Suppose we start with a primitive policy, OUT for example, with a long reorder cycle. This situation is far from ideal, but possible. Given the choice of speeding up the reorder cycle, or to improve the production control policy, we should opt for the latter. Doing otherwise — shortening the reorder cycle while keeping a primitive policy — may cause both excessive overtime and idling. We could call this an overtime inefficiency trap, where a well-intended improvement gives rise to unexpected costs. Maintaining the order cycle length while improving the policy is the more conservative change.

Changing the order policy means that the order quantities are determined in a different way, usually to smooth the production rate. This might lead to increased safety stock levels, and greater inventory fluctuations, but otherwise there should be no negative effects on the production system.

When a proportional policy is in place and the order cycle is long, the overtime work should be done as close to the first period of the order cycle as possible; this can lead to a dramatic reduction of inventory fluctuations. At the same time, time-varying safety stocks can be introduced. For this overtime strategy, the safety stock will be depleted in the first period of every cycle, and then gradually accumulated over the remaining periods. Therefore, we require less capacity in the first period than in the rest of the order cycle.

7.4 Limitations and research opportunities

The industrial prevalence of cyclical planning suggests that staggered deliveries may be common. Despite this, the literature on staggered deliveries is minimal. This thesis expands our understanding of staggered systems, but has its own limitations, and there are still many research gaps to fill.

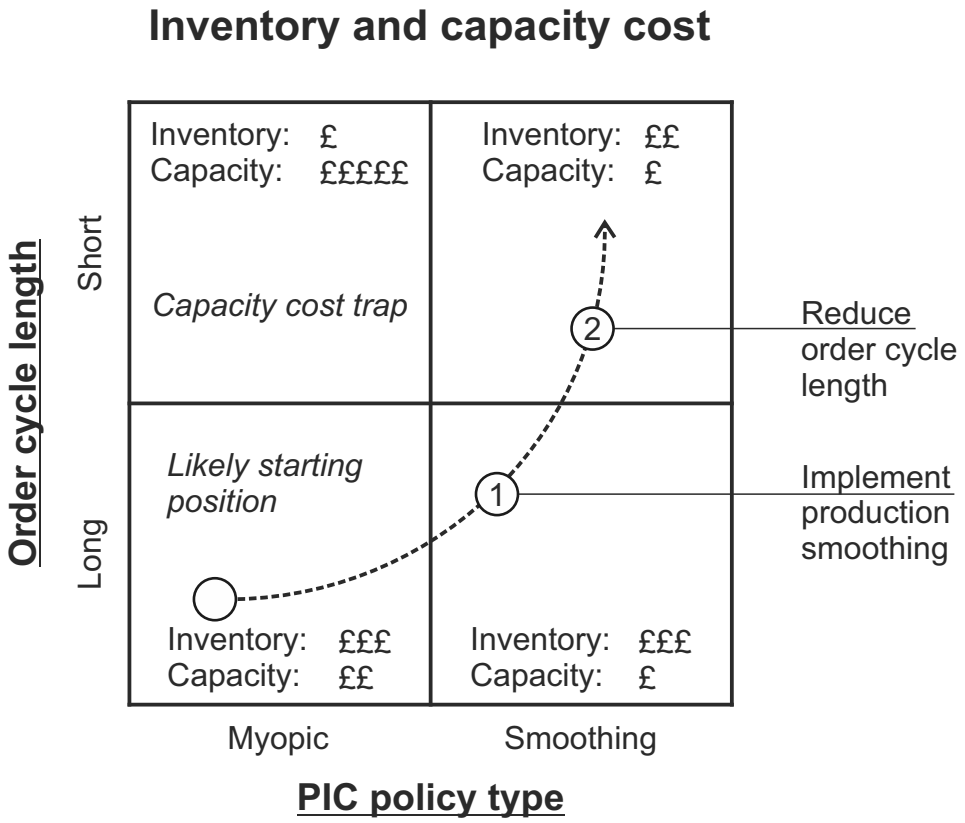

Figure 7.1: Avoid the capacity cost trap by improving the order policy before reducing the order cycle length.

The choice of developing a mathematical model has allowed us to reason, from first principles, about the expected consequences of staggered deliveries. This revealed the cyclical effects on the inventory variance. The analytical results are consistent with simulation, but the cyclical effect remains to be observed empirically. For general autocorrelated demand, this investigation is limited to inventory performance and the optimal order cycle length. Our numerical investigation of the fill rate reveals a rich behaviour that calls for further study.

Furthermore, we have assumed perfect knowledge of the autocorrelation function (ACF) of demand, that the ACF does not change over time, and that we can observe past demand from the beginning of time. In a real setting we must estimate the ACF from a limited set of past observations. This may introduce specification errors, and robustness tests could be considered, perhaps along the lines set out in Hosoda and Disney (2009). Further consideration could also be given to the mis-specification of the demand distribution, as in

Akcay et al. (2011) and Lee (2014). Identifying the effect in the real world is difficult; not only are accurate historical inventory records necessary (and these are often not kept in enterprise resource planning systems), but the company must use a policy that generates significant heteroskedasticity. For example, the [AR]STOUT policy generates a stronger change in the inventory variance than the SPOUT-E policy. In some cases the effect may exist, but be masked by random fluctuations, data inaccuracies, and limited sample sizes.

To generate models that provide meaningful results, we have assumed a linear production-inventory system, which may result in negative orders. Therefore, the outcome of this research is most accurate for high-volume production with low variability. There is an opportunity for developing staggered inventory models with non-negativity, or capacity constraints. Similarly, an opportunity remains for identifying the optimal staggered policy when piecewise-linear capacity costs are present. Topics that are relevant to non-staggered production and inventory control systems can be just as interesting, if not more, for staggered systems. This includes stochastic lead times, perhaps with crossovers, estimation and misspecification of demand processes, stochastic yield, and perhaps the effect of shared capacity.

Recalling that "all models are wrong" (Box, 1976) we should understand that the models in this thesis are simpler than the models we would build for an actual supply chain, but this need not be a bad thing:

> "The solution to the model will not be applicable in all its quantitative features; but it will call attention to the chief qualitative features of the appropriate policy, the form it will take, and the directions in which it can be expected to vary with changes in the underlying parameters." (Arrow et al., 1958, p. 18)

7.5 Summary

This chapter has reviewed the research questions and presented answers based on the analysis in this thesis. Academic and managerial implications of this work have been presented, and the limitations of this research and the resulting models have been discussed.

We have confirmed Flynn's assertion that staggered deliveries can be optimal, and expanded the case to autocorrelated demand. The autocorrelation leads to different optimal safety stock levels, while also affecting inventory costs, fill rates, and the length of the optimal order cycle. If we were to apply the

inventory-cost optimal policy (STOUT) to a system with capacity costs, we may get a P^* value distinct from one. In this setting, going after a reduction of the order cycle length, perhaps as part of a well-intended lead time reduction programme, can have dire economic consequences. We can get around this problem by first implementing a policy capable of production smoothing. In addition, we now know it is better to collect the overtime to the beginning of each order cycle, than to spread the overtime work evenly over the cycles.

Even after developing these insights, much work remains to be done. Many problems in the non-staggered literature can be extended to staggered deliveries. In particular, we may want to consider the impact of multi-product scenarios, and the effect of forecasts, and misspecified demand distributions.

Bibliography

Ackoff, R. L., 1949. On a Science of Ethics. *Philosophy and Phenomenological Research*, 9(4):663–672.

Ackoff, R. L., 1999. *Ackoff's Best: His classic writings on management*. Wiley New York.

Akcay, A., Biller, B., and Tayur, S., 2011. Improved inventory targets in the presence of limited historical demand data. *Manufacturing & Service Operations Management*, 13(3):297–309.

Arrow, K. J., Harris, T., and Marschak, J., 1951. Optimal inventory policy. *Econometrica*, 19(3):250–272.

Arrow, K. J., Karlin, S., and Scarf, H., 1958. Structure of inventory problems. In *Studies in the Mathematical Theory of Inventory and Production*, pp. 16–36. Stanford University Press, Stanford.

Arrow, K. J., Karlin, S., and Suppes, P., 1960. *Mathematical methods in the social sciences, 1959*. Stanford University Press.

Axsäter, S., 2006. *Inventory Control*. Springer, New York.

Babai, M. Z. and Dallery, Y., 2009. Dynamic versus static control policies in single stage production-inventory systems. *International Journal of Production Research*, 47(2):415–433.

Beckmann, M. J., 1961. Production smoothing and inventory control. *Operations Research*, 9(4):456–467.

Bellman, R., 1956. Dynamic programming and Lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767.

Bellman, R., 2003. *Dynamic Programming*. Dover Publications, New York.

Bertsekas, D. P., 2005. *Dynamic Programming and Optimal Control*, vol. 1. Athena Scientific, Belmont, MA.

Blaikie, N., 1993. *Approaches to Social Enquiry*. Polity Press, Cambridge.

Boer, H., Holweg, M., Kilduff, M., Pagell, M., Schmenner, R., and Voss, C., 2015. Making a meaningful contribution to theory. *International Journal of Operations & Production Management*, 35(9):1231–1252.

Box, G. E. P., 1976. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Box, G. E. P. and Jenkins, G. M., 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Bradley, J. R. and Conway, R. W., 2003. Managing cyclic inventories. *Production and Operations Management*, 12(4):464–479.

Brans, J. P., 2002a. Ethics and decision. *European Journal of Operational Research*, 136(2):340–352.

Brans, J. P., 2002b. OR, Ethics and Decisions: the OATH of PROMETHEUS. *European Journal of Operational Research*, 140(2):191–196.

Bryman, A., 2012. *Social Research Methods*. Oxford University Press, Oxford.

Burbidge, J., 1983. Five golden rules to avoid bankruptcy. *Production Engineer*, 62(10):13–14.

Burbidge, J. L., 1961. The" new approach" to production. *Production Engineer*, 40(12):769–784.

Burbidge, J. L., 1989. *Production Flow Analysis for Planning Group Technology*. Clarendon Press, Oxford.

Cachon, G. and Terwiesch, C., 2009. *Matching Supply with Demand*. McGraw-Hill, Singapore.

Cairns, G., Goodwin, P., and Wright, G., 2016. A decision-analysis-based framework for analysing stakeholder behaviour in scenario planning. *European Journal of Operational Research*, 249(3):1050–1062.

Checkland, P. B. and Scholes, J., 1990. *Soft Systems Methodology in Action*. Wiley, Chichester, UK.

Chiang, C., 2001. Order splitting under periodic review inventory systems. *International Journal of Production Economics*, 70(1):67–76.

Chiang, C., 2009. A periodic review replenishment model with a refined delivery scenario. *International Journal of Production Economics*, 118(1):253–259.

Churchman, C. W., Ackoff, R. L., and Arnoff, E. L., 1957. *Introduction to Operations Research*. John Wiley & Sons, New York.

Ciancimino, E., Cannella, S., Bruccoleri, M., and Framinan, J. M., 2012. On the bullwhip avoidance phase: The synchronised supply chain. *European Journal of Operational Research*, 221(1):49–63.

Clark, A. J. and Scarf, H., 1960. Optimal policies for a multi-echelon inventory problem. *Management Science*, 6(4):475–490.

Cárdenas-Barrón, L. E., Treviño-Garza, G., and Wee, H. M., 2012. A simple and better algorithm to solve the vendor managed inventory control system of multi-product multi-constraint economic order quantity model. *Expert Systems with Applications*, 39(3):3888–3895.

Dejonckheere, J., Disney, S. M., Lambrecht, M. R., and Towill, D. R., 2003. Measuring and avoiding the bullwhip effect: A control theoretic approach. *European Journal of Operational Research*, 147(3):567–590.

Deziel, D. P. and Eilon, S., 1967. A linear production-inventory control rule. *Production Engineer*, 46(2):93.

Diekmann, S., 2013. Moral mid-level principles in modeling. *European Journal of Operational Research*, 226(1):132–138.

Diks, E. B., de Kok, A. G., and Lagodimos, A. G., 1996. Multi-echelon systems: A service measure perspective. *European Journal of Operational Research*, 95(2):241–263.

Disney, S. M., Gaalman, G. J. C., Hedenstierna, C. P. T., and Hosoda, T., 2015. Fill rate in a periodic review order-up-to policy under auto-correlated normally distributed, possibly negative, demand. *International Journal of Production Economics*, 170, Part B:501–512.

Disney, S. M. and Grubbström, R. W., 2004. Economic consequences of a production and inventory control policy. *International Journal of Production Research*, 42(17):3419–3431.

Disney, S. M., Hoshiko, L., Polley, L., and Weigel, C., 2013. Removing bullwhip from Lexmark's toner operations. In *Proceedings of the 24th Annual Conference of the Production and Operations Management Society*. Denver, CO. May 3rd – 6th, 10 pages.

Disney, S. M., Maltz, A., Wang, X., and Warburton, R. D. H., 2016. Inventory management for stochastic lead times with order crossovers. *European Journal of Operational Research*, 248(2):473–486.

Disney, S. M. and Towill, D. R., 2002. A discrete transfer function model to determine the dynamic stability of a vendor managed inventory supply chain. *International Journal of Production Research*, 40(1):179–204.

Disney, S. M. and Towill, D. R., 2003. On the bullwhip and inventory variance produced by an ordering policy. *Omega*, 31(3):157–167.

Disney, S. M., Towill, D. R., and Van de Velde, W., 2004. Variance amplification and the golden ratio in production and inventory control. *International Journal of Production Economics*, 90(3):295–309.

Easterby-Smith, M., Thorpe, R., and Lowe, A., 1991. *Management Research: An Introduction*. Sage Publications, Thousand Oaks, CA.

ESRC, 2015. ESRC Framework for research ethics. Available at: http://www.esrc.ac.uk/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/ [Accessed on 2016-01-10].

Flynn, J., 2000. Selecting T for a periodic review inventory model with staggered deliveries. *Naval Research Logistics (NRL)*, 47(4):329–352.

Flynn, J., 2001. Selecting review periods for a coordinated multi-item inventory model with staggered deliveries. *Naval Research Logistics (NRL)*, 48(5):430–449.

Flynn, J., 2008. An effective heuristic for the review period in an inventory model with staggered deliveries and normal demands. *European Journal of Operational Research*, 186(2):671–680.

Flynn, J. and Garstka, S., 1990. A dynamic inventory model with periodic auditing. *Operations Research*, 38(6):1089–1103.

Flynn, J. and Garstka, S., 1997. The optimal review period in a dynamic inventory model. *Operations Research*, 45(5):736–750.

Forrester, J. W., 1958. Industrial dynamics: a major breakthrough for decision makers. *Harvard Business Review*, 36(4):37–66.

Forrester, J. W., 1994. System dynamics, systems thinking, and soft OR. *System Dynamics Review*, 10(2-3):245–256.

Forrester, J. W. and Senge, P. M., 1978. *Tests for building confidence in system dynamics models*. Technical report, System Dynamics Group, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

Gaalman, G., 2006. Bullwhip reduction for ARMA demand: The proportional order-up-to policy versus the full-state-feedback policy. *Automatica*, 42(8):1283–1290.

Gallo, G., 2004. Operations research and ethics: Responsibility, sharing and cooperation. *European Journal of Operational Research*, 153(2):468–476.

Gardner, E. S., 1990. Evaluating forecast performance in an inventory control system. *Management Science*, 36(4):490–499.

Goldratt, E., 2004. *The Goal: A Process of Ongoing Improvement*. North River Press, Great Barrington, MA.

Guijarro, E., Cardós, M., and Babiloni, E., 2012. On the exact calculation of the fill rate in a periodic review inventory policy under discrete demand patterns. *European Journal of Operational Research*, 218(2):442–447.

Hadley, G. F. and Whitin, T. M., 1963. *Analysis of Inventory Systems: Prentice Hall International Series in Management and Quantitative Methods Series*. Prentice-Hall, Englewood Cliffs, NJ.

Harré, R., 1976. The constructive role of models. In *The use of models in the social sciences*, pp. 16–43. Tavistock, London.

Harris, F. W., 1990. How many parts to make at once. *Operations Research*, 38(6):947–950.

Hedenstierna, C. P. T. and Disney, S. M., 2012. Impact of scheduling frequency and shared capacity on production and inventory costs. In *Pre-prints of the 18th International Working Seminar on Production Economics*, vol. 2, pp. 277–288. Innsbruck, Austria.

Hedenstierna, C. P. T. and Disney, S. M., 2014. The inventory ripple effect in periodic review systems with auto-correlated demand. In R. W. Grubbström and H. H. Hinterhuber, eds., *Pre-prints of the 18th International Working Seminar on Production Ecnonomics*, vol. 3, pp. 207–218. Innsbruck, Austria.

Hedenstierna, P., 2009. *Modelling the Benefits of Supply Chain Segmentation: An Empirical Study in a Global Fast Moving Consumer Goods (FMCG) Company*. M.Sc. thesis, Cranfield University, Cranfield.

Heyman, D. P. and Sobel, M. J., 1984. *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization*. McGraw-Hill, New York.

Holmström, J., 1998. Business process innovation in the supply chain – a case study of implementing vendor managed inventory. *European Journal of Purchasing & Supply Management*, 4(2–3):127–131.

Holt, C. C., Modigliani, F., Muth, J. F., and Simon, H. A., 1960. *Planning Production, Inventories, and Work Force*. Prentice-Hall, Englewood Cliffs, NJ.

Hosoda, T. and Disney, S. M., 2009. Impact of market demand mis-specification on a two-level supply chain. *International Journal of Production Economics*, 121(2):739–751.

Hosoda, T. and Disney, S. M., 2012. On the replenishment policy when the market demand information is lagged. *International Journal of Production Economics*, 135(1):458–467.

Howard, R. A., 1963. System Analysis of Linear Models. In H. E. Scarf, D. M. Gilford, and M. W. Shelly, eds., *Multistage Inventory Models and Techniques*, pp. 143–184. Stanford University Press, Stanford.

Isaksson, O. H. D. and Seifert, R. W., 2016. Quantifying the bullwhip effect using two-echelon data: A cross-industry empirical investigation. *International Journal of Production Economics*, 171, Part 3:311–320.

John, S., Naim, M., and Towill, D. R., 1994. Dynamic analysis of a WIP compensated decision support system. *International Journal of Manufacturing System Design*, 1(4):283–297.

Johnson, M. E., Lee, H. L., Davis, T., and Hall, R., 1995. Expressions for item fill rates in periodic inventory systems. *Naval Research Logistics*, 42(1):57–80.

Jonsson, P. and Mattsson, S.-A., 2005. *Logistik: Läran om effektiva materialflöden*. Studentlitteratur, Lund. In Swedish.

Jonsson, P. and Mattsson, S.-A., 2013. *Lagerstyrning i svensk industri: 2013 års användning, användningssätt och trender*. Technical report, Chalmers University of Technology, Gothenburg. In Swedish.

Kaku, B. K. and Krajewski, L. J., 1995. Period Batch Control in group technology. *International Journal of Production Research*, 33(1):79–99.

Kelle, P. and Silver, E. A., 1990. Decreasing expected shortages through order splitting. *Engineering Costs and Production Economics*, 19(1–3):351–357.

Kirk, D. E., 1997. *Optimal Control Theory: An introduction*. Dover Publications, New York.

Lansburgh, R. H., 1928. *Industrial Management*. John Wiley & Sons, New York.

Lee, H. and Whang, S., 1999. Decentralized multi-echelon supply chains: Incentives and information. *Management Science*, 45(5):633–640.

Lee, H. L., Billington, C., and Carter, B., 1993. Hewlett-Packard gains control of inventory and service through design for localization. *Interfaces*, 23(4):1–11.

Lee, H. L., Padmanabhan, V., and Whang, S., 1997. Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43(4):546–558.

Lee, H. L., So, K. C., and Tang, C. S., 2000. The value of information sharing in a two-level supply chain. *Management Science*, 46(5):627.

Lee, Y. S., 2014. A semi-parametric approach for estimating critical fractiles under autocorrelated demand. *European Journal of Operational Research*, 234(1):163–173.

Li, Q., Disney, S. M., and Gaalman, G., 2014. Avoiding the bullwhip effect using damped trend forecasting and the order-up-to replenishment policy. *International Journal of Production Economics*, 149:3–16.

Lian, Z., Deshmukh, A., and Wang, J., 2006. The optimal frozen period in a dynamic production model. *International Journal of Production Economics*, 103(2):648–655.

Luenberger, D. G., 1979. *Introduction to Dynamic Systems: Theory, Models, and Applications*. John Wiley & Sons, New York.

Maister, D. H., 1976. Centralisation of inventories and the "Square Root Law". *International Journal of Physical Distribution*, 6(3):124–134.

Metters, R., 1997. Quantifying the bullwhip effect in supply chains. *Journal of Operations Management*, 15(2):89–100.

Minas, J. S., 1956. Formalism, Realism and Management Science. *Management Science*, 3(1):9–14.

Mingers, J., 2011. Soft OR comes of age—but not everywhere! *Omega*, 39(6):729–741.

Modigliani, F. and Hohn, F. E., 1955. Production planning over time and the nature of the expectation and planning horizon. *Econometrica, Journal of the Econometric Society*, 23(1):46–66.

Nathan, J. and Venkataraman, R., 1998. Determination of master production schedule replanning frequency for various forecast window intervals. *International Journal of Operations & Production Management*, 18(8):767–777.

Nise, N. S., 2011. *Control Systems Engineering*. John Wiley & Sons, Hoboken, NJ.

Ohno, T., 1988. *Toyota Production System: Beyond Large-Scale Production*. Productivity Press, New York.

Pagh, J. D. and Cooper, M. C., 1998. Supply chain postponement and speculation strategies: How to choose the right strategy. *Journal of Business Logistics*, 19(2):13–33.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K., 2014. 'Horses for Courses' in demand forecasting. *European Journal of Operational Research*, 237(1):152–163.

Pidd, M., 2003. *Tools for Thinking*. Wiley, Chichester, UK.

Potter, A. and Disney, S. M., 2010. Removing bullwhip from the Tesco supply chain. In *Production and Operations Management Society Annual Conference, May 7th – 10th*. Vancouver, Canada. Paper No. 015-0397, 19 pages.

Prak, D., Teunter, R., and Riezebos, J., 2015. Periodic review and continuous ordering. *European Journal of Operational Research*, 242(3):820–827.

Rao, P. P., 1990. A dynamic programming approach to determine optimal manpower recruitment policies. *The Journal of the Operational Research Society*, 41(10):983–988.

Rosenshine, M. and Obee, D., 1976. Analysis of a standing order inventory system with Emergency orders. *Operations Research*, 24(6):1143–1155.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., and Ducq, Y., 2013. Demand forecasting by temporal aggregation. *Naval Research Logistics (NRL)*, 60(6):479–498.

Roundy, R., 1986. A 98%-effective lot-sizing rule for a multi-product, multi-stage production / inventory system. *Mathematics of Operations Research*, 11(4):699–727.

Rummler, G. A. and Brache, A. P., 1995. *Improving Performance: How to Manage the White Space in the Organization Chart*. Jossey-Bass, San Francisco.

Scarf, H. E., 1959. The optimality of (S,s) policies in the dynamic inventory problem. In K. J. Arrow, S. Karlin, and P. Suppes, eds., *Mathematical Methods in the Social Sciences, 1959. Proceedings of the First Stanford Symposium*, pp. 196–202. Stanford University Press, Stanford.

Schmitt, A. J. and Singh, M., 2012. A quantitative analysis of disruption risk in a multi-echelon supply chain. *International Journal of Production Economics*, 139(1):22–32.

Sethi, S. P., Yan, H., and Zhang, H., 2003. Inventory models with fixed costs, forecast updates, and two delivery modes. *Operations Research*, 51(2):321–328.

Shingo, S., 1989. *A Study of the Toyota Production System: From an Industrial Engineering Viewpoint*. Productivity Press, New York.

Silver, E. A. and Bischak, D. P., 2011. The exact fill rate in a periodic review base stock system under normally distributed demand. *Omega*, 39(3):346–349.

Silver, E. A., Pyke, D. F., and Peterson, R., 1998. *Inventory Management and Production Planning and Scheduling*. Wiley, New York.

Simchi-Levi, D., 2002. *Designing and Managing the Supply Chain: Concepts, Strategies, and Cases*. McGraw-Hill, New York.

Simon, H. A., 1952. On the application of servomechanism theory in the study of production control. *Econometrica: Journal of the Econometric Society*, 20(2):247–268.

Simon, H. A., 1956. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24(1):74–81.

Simon, H. A. and Holt, C. C., 1954. The control of inventories and production rates-a survey. *Journal of the Operations Research Society of America*, 2(3):289–301.

Slack, N., 2015. Runners, Repeaters, and Strangers. In *Wiley Encyclopedia of Management*. John Wiley & Sons, Hoboken, NJ.

Sloan, A. P., 1963. *My Years with General Motors*. Doubleday, New York.

Sobel, M. J., 1970. Making short-run changes in production when the employment level is fixed. *Operations Research*, 18(1):35–51.

Sobel, M. J., 2004. Fill rates of single-stage and multistage supply systems. *Manufacturing & Service Operations Management*, 6(1):41–52.

Sterman, J. D., 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw-Hill Education, New York.

Strijbosch, L. W. G., Syntetos, A. A., Boylan, J. E., and Janssen, E., 2011. On the interaction between forecasting and stock control: The case of non-stationary demand. *International Journal of Production Economics*, 133(1):470–480.

Tang, O. and Grubbström, R. W., 2002. Planning and replanning the master production schedule under demand uncertainty. *International Journal of Production Economics*, 78(3):323–334.

Teunter, R. H., 2009. Note on the fill rate of single-stage general periodic review inventory systems. *Operations Research Letters*, 37(1):67–68.

Teunter, R. H., Babai, M. Z., and Syntetos, A. A., 2010. ABC classification: service levels and inventory costs. *Production and Operations Management*, 19(3):343–352.

Towill, D. R., 1982. Dynamic analysis of an inventory and order based production control system. *The International Journal of Production Research*, 20(6):671–687.

Tsypkin, I. Z., 1964. *Sampling Systems Theory and its Application*, vol. 1. Pergamon, Oxford.

Tustin, A., 1953. *The Mechanism of Economic Systems: An Approach to the Problem of Economic Stabilization from the Point of View of Control-System Engineering*. Heinemann, London.

Vassian, H. J., 1955. Application of discrete variable servo theory to inventory control. *Journal of the Operations Research Society of America*, 3(3):272–282.

Veinott, A. F., 1965. The Optimal Inventory Policy for Batch Ordering. *Operations Research*, 13(3):424–432.

Veinott, A. F., 1966. On the Optimality of (s, S) Inventory Policies: New Conditions and a New Proof. *SIAM Journal on Applied Mathematics*, 14(5):1067–1083.

Wagner, H. M. and Whitin, T. M., 1958. Dynamic version of the economic lot size model. *Management Science*, 5(1):89–96.

Wang, X. and Disney, S. M., 2016. The bullwhip effect: Progress, trends and directions. *European Journal of Operational Research*, 250(3):691–701.

Zhang, J. and Zhang, J., 2007. Fill rate of single-stage general periodic review inventory systems. *Operations Research Letters*, 35(4):503–509.

Zhao, X. and Lee, T. S., 1993. Freezing the master production schedule for material requirements planning systems under demand uncertainty. *Journal of Operations Management*, 11(2):185–205.

Zhou, L., Disney, S., and Towill, D. R., 2010. A pragmatic approach to the design of bullwhip controllers. *International Journal of Production Economics*, 128(2):556–568.

Appendices

Appendix A

Piecewise linear cost models

This appendix presents proofs for the optimal inventory and capacity levels under piecewise linear costs. Special cases of these costs appear in Disney and Grubbström (2004) and Hosoda and Disney (2012) along with optimal solutions. To make the thesis self-contained, this appendix presents a proof based on the solution to the newsvendor problem in Churchman et al. (1957, p. 207-214). The solution has been reinterpreted for inventory costs and capacity costs. A measure-theoretic approach is used, as this simplifies the proofs; assume f and o to be random variables on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where $\mathcal{A}$ is a Borel σ-algebra, and $\omega \in \Omega$.

A.1 Inventory costs

Under linear holding and backlog costs, the optimal OUT level is identical to the order quantity in the newsvendor problem. To see this, suppose there is an OUT level S, a lead-time demand f, and an inventory level $i = S - f$. with costs given by

$$j(S) = h(i)^+ + b(i)^-, \tag{A.1}$$

where $(x)^+ = x[x < 0]$ returns x if x is positive, and $(x)^- = (-x)[0 < x]$ returns $(-x)$ if x is negative.

Theorem A.1 (Safety stock optimization)**.** *The expected inventory cost,* $J(S) = \mathbb{E}[j(S)]$*, is minimized when*

$$\mathbb{P}\{i < 0\} = \frac{h}{b+h}. \tag{A.2}$$

Proof. Equation (A.1) is convex in S. Since convexity is preserved under linear

transformations, and when taking the expectation, it is sufficient to demonstrate that (A.2) is the solution obtained when differentiating $J(S)$ and solving for zero. Thus, take (A.1) and rewrite it as

$$j(S) = h(S - f) + (b + h)(S - f)^{-}. \tag{A.3}$$

Recalling that the $(\cdot)^{-}$ operator switches the sign, we obtain

$$\begin{aligned} J(S) &= h \int_{\Omega} (S - f)\, d\mathbb{P} + (-1)(b + h) \int_{i<0} (S - f)\, d\mathbb{P} \\ &= h(S + \mathbb{E}[f]) - (b + h) \int_{i<0} (S - f)\, d\mathbb{P}. \end{aligned} \tag{A.4}$$

When differentiating $J(S)$, we exploit Leibniz's rule to obtain,

$$\begin{aligned} J'(S) &= \frac{d}{dS} h(S - \mathbb{E}[f]) - (b + h) \frac{d}{dS} \int_{i<0} (S - f)\, d\mathbb{P} \\ &= h - (b + h) \int_{i<0} \frac{d}{dS} (S - f)\, d\mathbb{P} \\ &= h - (b + h) \int_{i<0} d\mathbb{P} \\ &= h - (b + h)\mathbb{P}\{i < 0\}. \end{aligned} \tag{A.5}$$

Setting $J'(S) = 0$ gives

$$\mathbb{P}\{i < 0\} = \frac{h}{b + h}, \tag{A.6}$$

and the proof is complete. ■

Remark. This result is the well-known solution to the structurally identical single-period problem, or newsboy problem as it is also called. The above derivation has the benefit of not assuming either a discrete or continuous distribution of the random variable.

A.2 Capacity costs

Definition A.2 (Reactive capacity costs)**.** A similar solution is obtained under reactive capacity. Hosoda and Disney (2012) present a proof that assumes normally distributed orders, but here we make no assumptions about the distribution of orders. Let there be a guaranteed labour capacity level z, which is always paid for with the unit cost u, regardless of production volume. When the production quantity o exceeds z, an overtime charge v, where $u < v$, is

incurred for all orders exceeding z, i.e. $o - z$. The cost is then

$$a(z) = uz + v(o - z)^+, \tag{A.7}$$

and the expected capacity cost is

$$A(z) = \mathbb{E}\left[a\left(z\right)\right] = uz + \int_{z \leq o} v\left(o - z\right)\ d\mathbb{P}. \tag{A.8}$$

Theorem A.3. *The expected capacity cost for reactive capacity* (A.8)*, is minimized when*

$$\mathbb{P}\{o < z\} = \frac{v - u}{v}. \tag{A.9}$$

Proof. We take the expected capacity cost (A.8), differentiate, and solve for zero:

$$\begin{aligned} A'\left(z\right) &= \frac{d}{dz}\left[uz + \int_{z \leq o} v\left(o - z\right)\ d\mathbb{P}\right] \\ &= \frac{d}{dz}\left[uz + v\int_{z \leq o} (o - z)\ d\mathbb{P}\right] \\ &= \frac{d}{dz}uz + v\int_{z \leq o} \frac{d}{dz}(o - z)\ d\mathbb{P} \\ &= u - v\int_{z \leq o} d\mathbb{P} \\ &= u - v\,\mathbb{P}\left\{z \leq o\right\} \\ &= u - v\left(1 - \mathbb{P}\left\{o < z\right\}\right). \end{aligned} \tag{A.10}$$

Setting $A'\left(z\right) = 0$ gives

$$\mathbb{P}\left\{o < z\right\} = \frac{v - u}{v}, \tag{A.11}$$

and the proof is complete. ■

Remark. The variables o and z need not represent production units; instead, they may be counted in man-hours, if the costs u and v also reflect this.

Appendix B

Proofs

B.1 Proof of Theorem 4.2

Proof.

(a) Assume that we are to place an order at time t, to be received in period $t+\tau$. By setting the conditional expectation $\mathbb{E}\left[i_{t+\tau}|\{d_t, d_{t-1}, \cdots\}\right] = i^*_{t+\tau}$, and solving for $r_{t+\tau}$, we obtain the policy that minimizes the inventory cost for any τ. Finally we set $\tau = L+1$, to obtain the solution for the first order in a cycle.

To begin, we use induction on the inventory balance equation (4.2) to obtain $i_{t+1} = i_{t-1} + r_t + r_{t+1} - d_t - d_{t+1}$. Extending this to $i_{t+\tau}$ yields

$$i_{t+\tau} = i_t + \sum_{n=1}^{\tau} (r_{t+n} - d_{t+n}) = i_t + w_{t,\tau} + r_{t+\tau} - f_{t,\tau}. \tag{B.1}$$

The inventory costs are convex, and we seek to minimize them for an arbitrary period by setting the expected inventory level to $i^*_{t+\tau}$. Therefore, let $i^*_{t+\tau}$ equal the expectation of (B.1), conditional on our observations of demand up to time t, when the order is determined. We obtain

$$\mathbb{E}\left[i_{t+\tau}|\{d_t, d_{t-1}, \cdots\}\right] = i^*_{t+\tau} = i_t + r_{t+\tau} + w_{t,\tau} - \hat{f}_{t,\tau}. \tag{B.2}$$

The receipt rate $r_{t+\tau}$ can be found by rearranging (B.2),

$$r_{t+\tau} = \hat{f}_{t,\tau} + i^*_{t+\tau} - i_t - w_{t,\tau}. \tag{B.3}$$

Finally, we let $k = 1$, so that $\tau = L + 1$ and obtain

$$o_{t,1} = r_{t+L+1} = \hat{f}_{t,L+1} + i^*_{t+L+1} - i_t - w_{t,L+1}, \tag{B.4}$$

This concludes the first part of the proof.

(b) Assume that $L + 1 < \tau \leq L + P$, so that $\tau + 1$ corresponds to $k > 1$. Inserting the receipts (B.3) back into the inventory equation (B.1) gives

$$i_{t+\tau} = i^*_{t+\tau} + \hat{f}_{t,\tau} - f_{t,\tau}. \tag{B.5}$$

Rearranging the inventory balance equation (4.2) yields $r_{t+\tau} = i_{t+\tau} - i_{t+\tau-1} + d_{t+\tau}$. Replacing both inventory levels with their equivalent form in (B.5), we obtain

$$r_{t+\tau} = o_{t,k} = i^*_{t+\tau} - i^*_{t+\tau-1} + \hat{d}_{t,\tau}, \tag{B.6}$$

where $\hat{d}_{t,\tau+1} = \hat{f}_{t,\tau+1} - \hat{f}_{t,\tau}$ is the single-period forecast for $d_{t+\tau+1}$ made at time t. This completes the proof. ■

B.2 Proof of Theorem 4.4

Proof. We shall express i_t and $i_t + d_t$ as a weighted sum of independent error terms, and then take the variance or covariance.

(a) Recall the inventory equation (B.5)

$$i_t = i^*_t + \hat{f}_{t-\tau,\tau} - \sum_{n=0}^{\tau-1} d_{t-n}. \tag{B.7}$$

Let us express the lead-time demand as a weighted sum of error terms,

$$\sum_{a=0}^{\tau-1} d_{t-a} = \left(\sum_{m=0}^{\tau-1} \varepsilon_{t-m} \sum_{n=0}^{m} \theta_n \right) + \left(\sum_{x=\tau}^{\infty} \varepsilon_{t-x} \sum_{y=0}^{\tau-1} \theta_{x-y} \right), \tag{B.8}$$

and express the corresponding forecast as a weighted sum of error terms,

$$\hat{f}_{t-\tau,\tau} = \sum_{x=\tau}^{\infty} \varepsilon_{t-x} \sum_{y=0}^{\tau-1} \theta_{x-y}. \tag{B.9}$$

Substituting the lead-time demand (B.8) and the forecast (B.9) into the

inventory equation, we obtain

$$i_t = i_t^* - \sum_{m=0}^{\tau-1} \varepsilon_{t-m} \sum_{n=0}^{m} \theta_n. \tag{B.10}$$

Clearly, $\mathbb{E}[i_t] = i_t^*$. Taking the variance, we obtain

$$\operatorname{var}(i_t) = \sigma_\varepsilon^2 \sum_{m=0}^{\tau-1} \left(\sum_{n=0}^{m} \theta_n \right)^2, \tag{B.11}$$

completing the first part of the proof.

(b) Without loss of generality, assume $i_t^* = 0$. We can then characterize $i_t + d_t$ as

$$\begin{aligned} i_t + d_t &= \left(\sum_{x=0}^{\infty} \varepsilon_{t-x}\, \theta_x \right) - \left(\sum_{m=0}^{\tau-1} \varepsilon_{t-m} \sum_{n=0}^{m} \theta_n \right) \\ &= \left(\sum_{x=0}^{\infty} \varepsilon_{t-x}\, \theta_x \right) - \left[\varepsilon_t\, \theta_0 + \sum_{m=1}^{\tau-1} \varepsilon_{t-m} \left(\theta_m + \sum_{n=0}^{m-1} \theta_n \right) \right] \\ &= \left(\sum_{x=\tau}^{\infty} \varepsilon_{t-x}\, \theta_x \right) - \left\{ \sum_{m=1}^{\tau-1} \varepsilon_{t-m} \left[\left(\sum_{n=0}^{m} \theta_n \right) - \theta_m \right] \right\}. \end{aligned} \tag{B.12}$$

With $i_t + d_t$ of this form, we take the variance

$$\operatorname{var}(i_t + d_t) = \sigma_\varepsilon^2 \left\{ \left[\sum_{m=1}^{\tau-1} \left(\sum_{n=0}^{m-1} \theta_n \right)^2 \right] + \sum_{x=\tau}^{\infty} \theta_x^2 \right\}, \tag{B.13}$$

completing this part of the proof.

(c) To obtain $\operatorname{cov}(d_t,\, i_t + d_t)$, we exploit that (B.10) and (B.12) are already of the required form. Taking the covariance gives (4.17), completing the proof. ■

B.3 Proof of Theorem 5.6

Proof.

(a) First, we express $x_{t,0}$ in terms of $x_{t-P,0}$

$$\begin{aligned} x_{t,0} &= x_{t-P,0} + \sum_{n=1}^{k} o_{t-P,n} - d_{t-P+n} \\ &= x_{t-P,0} + x_P^* - x_0^* + \alpha\left(x_0^* - x_{t-P,0}\right) - \sum_{n=1}^{P} d_{t-P+n} \\ &= (1-\alpha)\, x_{t-P,0} + \alpha x_0^* - \sum_{n=1}^{P} \varepsilon_{t-P+n} \end{aligned} \tag{B.14}$$

Continuing the recursion gives

$$x_{t,0} = \alpha x_0^* + \sum_{m=1}^{q} (1-\alpha)^{m-1} \left[\alpha x_0^* + x_{t-qP,0} - \sum_{n=1}^{P} \varepsilon_{t-mP+n}\right]. \tag{B.15}$$

When $q \to \infty$ we obtain

$$x_{t,0} = x_0^* - \sum_{n=1}^{P} \sum_{m=1}^{q} (1-\alpha)^{m-1} \varepsilon_{t-mP+n}, \tag{B.16}$$

which reveals the expectation $\mathbb{E}\left[x_{t,0}\right] = x_0^*$. As $x_{t,k} = x_{t,0} + \sum_{n=1}^{k} o_{t,n}$

$$x_{t,k} = x_k^* - \sum_{n=1}^{P} \sum_{m=1}^{q} (1-\alpha)^{m} \varepsilon_{t-mP+n}. \tag{B.17}$$

Taking the variance of $x_{t,k}$ and adding the variance of lead-time demand gives

$$\sigma_{i,k}^2 = \sigma_d^2 \left[k + L + \frac{P\left(1-\alpha\right)^2}{\alpha\left(2-\alpha\right)}\right],$$

completing this part of the proof.

(b) Inserting (B.16) in (5.44) provides

$$o_{t,1} = x_1^* - x_0^* + \alpha \sum_{n=1}^{P} \sum_{m=1}^{q} (1-\alpha)^{m-1} \varepsilon_{t-mP+n}. \tag{B.18}$$

Taking the variance gives

$$\sigma_{o,1}^2 = \frac{\sigma_d^2 \alpha P}{2-\alpha}.$$

For $k \neq 1$, the orders are constant, therefore $\sigma_{o,k}^2 = 0$, and the proof is complete. ■

B.4 Proof of Lemma 5.7

Proof. The cost functions are

$$(C_P^*|_{\text{STOUT}}) = \psi \left[\frac{\lambda}{\sqrt{P}} + (1-\lambda)\,\bar{\sigma}_{i,P} \right] + \mu u, \tag{B.19}$$

where $\bar{\sigma}_{i,P} = \sigma_\varepsilon P^{-1} \sum_{k=1}^{P} \sqrt{k}$ is the average inventory standard deviation under STOUT; and

$$(C_1^*|_{\text{SPOUT}}) = \psi\sqrt{1-\lambda^2} + \mu u. \tag{B.20}$$

When $P = 1$, it is trivial to see that $(C_1^*|_{\text{SPOUT}}) \leq (C_1|_{\text{STOUT}})$. The cases when STOUT has $P > 1$ can be solved as follows. Since $(C_P^*|_{\text{STOUT}})$ and $(C_1^*|_{\text{SPOUT}})$ are continuous on $\lambda \in [0,1]$ it is sufficient to see that these cost functions do not intersect. If these cost functions were to intersect, then $(C_1^*|_{\text{SPOUT}}) = (C_P|_{\text{STOUT}})$, or equivalently

$$\sqrt{1-\lambda^2} = \frac{\lambda}{\sqrt{P}} + (1-\lambda)\bar{\sigma}_{i,P}. \tag{B.21}$$

Solving for λ gives

$$\lambda = \frac{P\bar{\sigma}_{i,P}^2 - \bar{\sigma}_{i,P}\sqrt{P} \pm \sqrt{P\left(P+1-2\bar{\sigma}_{i,P}\sqrt{P}\right)}}{P\bar{\sigma}_{i,P}^2 - 2\bar{\sigma}_{i,P}\sqrt{P} + P + 1}. \tag{B.22}$$

If a real solution exists, $\text{Im}\,(\lambda) = 0$, which is equivalent to $\sqrt{P}\,(P+1) - 2\sum_{k=1}^{P}\sqrt{k} \geq 0$. For $P = 1$ we obtain $\lambda = 0$, which is consistent with the trivial case. Using induction we can show that $\sqrt{P}\,(P+1) - 2\sum_{k=1}^{P}\sqrt{k}$ is decreasing in P. This follows from

$$0 > \left[(P+2)\sqrt{P+1} - 2\sum_{m=1}^{P+1}\sqrt{m} \right] - \left[(P+1)\sqrt{P} - 2\sum_{n=1}^{P}\sqrt{n} \right], \tag{B.23}$$

$$0 > \sqrt{P}\left[\sqrt{P\,(P+1)} - (1+P) \right], \tag{B.24}$$

which holds true. Therefore $\sqrt{P}\,(P+1) - 2\sum_{k=1}^{P}\sqrt{k} \geq 0$ is false, and we have a contradiction in (B.21). In other words, when $P > 1$ the cost functions do not intersect on $\lambda \in [0,1]$. This completes the proof. ■

Appendix C

On the exact fill rate

The conventional expression for the exact fill rate implicitly assumes that demand is positive (Sobel, 2004; Teunter, 2009). When demand can be negative (returns from customers), the conventional fill rate sometimes produces results indicating a fill rate above 100% or below 0% (Disney et al., 2015). When normal demand is assumed, these problems appear, as demand can be negative. Here we investigate the conventional exact fill rate, identify where the discrepancy appears for negative demand, and propose a refined fill rate definition that works when demand is negative. This new definition is a consistent elaboration of the exact fill rate for non-negative demand.

Consider following assumptions; later we shall relax Assumption C.1(c). Demand is defined as $d = \eta - \kappa$.

Assumption C.1.

(a) $\eta \in \mathbb{R}$ denotes the inventory level (on-hand inventory less backorders) just after a delivery, and $\kappa \in \mathbb{R}$ denotes the inventory level just before the *next* delivery.

(b) The random variables $\{\eta(\omega), \kappa(\omega)\}$ are defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where $\mathcal{A}$ is a Borel σ-algebra, and $\omega \in \Omega$.

(c) Between two subsequent deliveries, there is a non-negative demand, such that $0 \leq \eta - \kappa$.

Definition C.2. The *fill rate*, S_2, is the long-run fraction of demand satisfied immediately from inventory,

$$S_2 = \frac{\mathbb{E}[\text{DFI}]}{\mathbb{E}[\eta - \kappa]}, \tag{C.1}$$

where DFI denotes the *demand filled immediately* from stock. It is calculated as follows: If $\eta < 0$, no demand can be fulfilled in an inventory cycle. If η is positive and κ is negative, $\{0 < \eta\} \cap \{\kappa \leqslant 0\}$, then η units of demand have been satisfied. Finally, if $0 < \kappa$, all demand $(\eta - \kappa)$ has been satisfied. DFI is thus defined (cf. Silver and Bischak, 2011),

$$\mathrm{DFI} = \begin{cases} 0 & \text{if } \eta < 0, \\ \eta & \text{if } \{0 < \eta\} \cap \{\kappa \leqslant 0\}, \\ \eta - \kappa & \text{if } 0 < \kappa. \end{cases} \tag{C.2}$$

Theorem C.3. *The fill rate for any inventory system that corresponds to Assumption C.1 can be expressed as*

$$S_2 = \frac{\int_{0<\eta} \eta \, d\mathbb{P} - \int_{0<\kappa} \kappa \, d\mathbb{P}}{\int_\Omega \eta - \kappa \, d\mathbb{P}}. \tag{C.3}$$

Proof. It is sufficient to demonstrate that (C.1) equals (C.3). We begin with the numerator by taking the expectation of DFI.

$$\mathbb{E}[\mathrm{DFI}] = \int_{\{0<\eta\}\cap\{\kappa\leqslant 0\}} \eta \, d\mathbb{P} + \int_{0<\kappa} (\eta - \kappa) \, d\mathbb{P} = \int_{\{0<\eta\}\cap\{\kappa\leqslant 0\}} \eta \, d\mathbb{P} + \int_{0<\kappa} \eta \, d\mathbb{P} - \int_{0<\kappa} \kappa \, d\mathbb{P}. \tag{C.4}$$

Then, from Assumption C.1(c) we have that $0 < \kappa$ implies $0 < \eta$. Therefore $\{0 < \kappa\} = \{0 < \eta\} \cap \{0 < \kappa\}$.

$$\mathbb{E}[\mathrm{DFI}] = \int_{\{0<\eta\}\cap\{\kappa\leqslant 0\}} \eta \, d\mathbb{P} \; + \int_{\{0<\eta\}\cap\{0<\kappa\}} \eta \, d\mathbb{P} - \int_{0<\kappa} \kappa \, d\mathbb{P}. \tag{C.5}$$

Note that $\{\{0 < \eta\} \cap \{\kappa \leq 0\}\}$ and $\{\{0 < \eta\} \cap \{0 < \kappa\}\}$ are disjoint. Since $\{\{0 < \eta\} \cap \{\kappa \leq 0\}\} \cup \{\{0 < \eta\} \cap \{0 < \kappa\}\} = \{0 < \eta\}$, we can write (C.5) as

$$\mathbb{E}[\mathrm{DFI}] = \int_{0<\eta} \eta \, d\mathbb{P} - \int_{0<\kappa} \kappa \, d\mathbb{P}, \tag{C.6}$$

which is equal to the numerator in (C.3). Note that the denominator of (C.3), is obtained directly when taking the expectation, $\mathbb{E}[\eta - \kappa] = \int_\Omega \eta - \kappa \, d\mathbb{P}$. Thus,

$$S_2 = \frac{\int_{0<\eta} \eta \, d\mathbb{P} - \int_{0<\kappa} \kappa \, d\mathbb{P}}{\int \eta - \kappa \, d\mathbb{P}}, \tag{C.7}$$

and the proof is complete. ■

This analysis holds for non-negative demand, but suppose we relax this assumption. Then the proof halts at (C.4), as $0 < \kappa$ no longer implies $0 < \eta$. Therefore, (C.3) is not the correct expression for the fill rate, when demand can be negative. Another theoretical issue can occur when the expectation of demand is negative (more returns than sales), also leading to a negative fill rate. To find a reasonable fill rate when demand can be negative, we must revisit the definition. Let the fill rate be *the long-run fraction of satisfied demand, to the demand that can be satisfied.*

Definition C.4 (Exact fill rate for normally distributed demand)**.** The *exact fill rate* for normally distributed demand and inventory levels relies on the following definiton of DFI^-,

(a) Let $\mathrm{DFI}^- = (\min(\eta, \eta - \kappa))^+$ be the demand filled immediately, when negative demand is present, i.e. the smallest value of the start-of-cycle inventory, or demand, if either term is greater than zero.

(b) The fill rate that excludes negative demand is then

$$S_2^- = \frac{\mathbb{E}[\mathrm{DFI}^-]}{\mathbb{E}[(d)^+]}, \tag{C.8}$$

where $\mathbb{E}[(d)^+]$, is the expected positive demand.

A realization of S_2^- for normally distributed demand is presented in Lemma 4.3.

Nomenclature

λ	The relative cost of audits or capacity to inventory
α	The smoothing parameter in the SPOUT-[E] policy
$\bar{\sigma}_{i,P}$	The average inventory standard deviation over a cycle of length P
δ	The Kronecker delta function.
$\hat{f}$	The expected lead-time demand (minimum-mean-squared-error forecast)
λ_P	The value λ which is minimized by an order cycle length of P
$\mathbb{E}$	The expectation operator
$\mathbb{P}$	The probability of some event
$\mathbb{Z}^*$	The set of nonnegative integers
$\mathbb{Z}^+$	The set of strictly positive integers
$\mathbb{Z}$	The set of integers
μ	The expectation of demand in a single period
Φ	The cumulative density function of the standard normal distribution
ϕ	The autocorrelation parameter of a first-order autoregressive process.
Φ^{-1}	The inverse of Φ
ψ	A cost scaling factor
ρ	The Pearson correlation coefficient
$\sigma(\cdot)$	The standard deviation of some variable
σ_d^2	The variance of the demand process

σ_d	The standard deviation of demand
τ	The actual lead time for a specific receipt, including staggering delays
θ	The autocorrelation coefficient of demand
ε	The uncorrelated random component of (periodic) demand
φ	The probability density function of the standard normal distribution
ξ	The optimal smoothing parameter to minimize quadratic costs
A	The expected capacity cost
a	The singe-period capacity cost
b	The backlog cost per unit per period
C	The expected total cost of inventory and audits or capacity
c	The total cost of inventory and audits or capacity
d	Demand in a single period
f	The demand over the actual lead-time
g	The standard normal loss function
h	The holding cost per unit per period
i	The inventory level; on-hand inventory minus backorders
i_t^*	The optimal safety stock in period t
J	The expected inventory cost
j	The singe-period inventory cost
k	Index for the k'th period in a cycle
L	The lead time, i.e. the number of periods until the first receipt arrives.
$o_{t,k}$	The k'th order placed in period t
P	The order cycle length (number of inspections between orders)
P^*	The optimal reorder period length

r_t	The quantity received in period t
S_1	The availability service metric
S_2	The (unit) fill rate service metric
S_2^-	The (unit) fill rate service metric when demand can be negative
t	Time period
u	The unit cost at the regular rate (no overtime compensation)
v	The audit cost (Chapter 4), or the overtime cost per unit (Chapter 5)
w	Work-in-progress; orders placed, but not yet received
x	The inventory position
z_k	The regular (non-overtime) capacity in the k'th period of each cycle
DFI	Demand filled immediately