Technische Universität München

Lehrstuhl für Produktion und Supply Chain Management

Rolling-horizon production planning for seasonal and uncertain demand

Alexandre Forel

Vollständiger Abdruck der von der Fakultät für Wirtschaftswissenschaften der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Rainer Kolisch
Prüfer der Dissertation:	1.	Prof. Dr. Martin Grunow
	2.	Prof. Dr. Stefan Minner

Die Dissertation wurde am 09.04.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Wirtschaftswissenschaften am 15.06.2021 angenommen.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Dr. Martin Grunow for his invaluable support throughout this fulfilling experience. I am grateful to have been given this great opportunity to learn and I deeply appreciate his continuous guidance that has been indispensable to bring our research to fruition. I would also like to thank Prof. Dr. Stefan Minner for taking the role of second examiner and Prof. Dr. Rainer Kolisch for being the chairman of the examination committee.

My most sincere thanks go to Dennis Prak for his precious advice and for stepping up as mentor of my PhD studies. I wish to thank Markus Meiler, Thorsten Ehret, Zhicheng Huang and Florian Arnold from Bayer AG for their inspiring expertise throughout our rewarding collaboration. I would also like to thank the Deutsche Forschungsgemeinschaft (DFG) for funding my research as part of the GRK 2201 ("Advanced Optimization in a Networked Economy").

My sincere thanks go to all AdONE members who have given me such a large overview of what is possible within the world of Operations Research. I am indebted to all the wonderful colleagues, past and present, at the Chair of Production and Supply Chain Management. Alex, Andy, Bryndís, Florian, Frank, Jishna, Lena, Mirko, Verena, thank you for the great time working and learning together. I also would like to thank Monika, Yoshimi and Isabel for their help and patience.

I would like to express my warmest thanks to all my friends who accompanied me over the years. I am deeply grateful to my family and to my parents, Anna and René, for their unconditional love and for instilling in me the values of perseverance and curiosity that proved so relevant in my studies. I am particularly grateful to Sophie for her unwavering support and trust. I could not have done it without you.

Merci!

Abstract

Rolling-horizon production planning is based on the periodic update of forecasts and decisions. Today, companies implement a combination of deterministic models and exogenously calculated safety stocks in rolling horizon. Yet, deterministic models only react passively to forecast updates. On the contrary, stochastic models anticipate the uncertainty of forecasts over the horizon and provide cost-optimal safety stocks. Further, stochastic programming reflects the flexibility of rolling-horizon planning through recourse decisions that adapt to uncertainty in each stage. This thesis studies the application of stochastic methods to production planning when demand is seasonal and uncertain. Different problems and solution approaches are analysed in three chapters.

First, a multi-ordering newsvendor problem is considered in which forecasts are periodically updated according to the martingale model of forecast evolution (MMFE). In each planning period, capacity is limited and holding costs are incurred for carrying inventory up to the selling season. We analyse the key trade-off between producing early to avoid lost-sales and producing late to minimise inventory costs. The optimal production policy is determined analytically for the single-product case and adapted into a heuristic to plan several correlated products. The value of forecast evolution models is evaluated in rolling-horizon simulations.

Second, we apply stochastic programming to master production scheduling. We identify general barriers preventing the application of stochastic programming such as modelling uncertainty from limited data, reflecting the planning processes in stochastic models, and obtaining accurate evaluations. We propose a framework to overcome the barriers and develop a two-stage stochastic model with production recourse that improves planning flexibility, communicability and stability. We demonstrate our approach on a real-world case study in the agrochemical industry.

Third, we focus on dynamic stochastic lot-sizing problems. We show how both additive and multiplicative MMFE can be readily integrated in lot-sizing problems and solved using piecewise-linear approximations of the expected inventory and backlogs. We introduce scenario-based recourse to allow flexible decisions. The value of forecast

evolution models and the value of recourse are quantified in rolling-horizon simulations using both artificial and real-world data.

Stochastic rolling-horizon production planning is thoroughly studied throughout this thesis. We highlight strengths and limitations of stochastic models and evaluate their performance in a wide range of problems and production environments.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1. Motivation

Demand forecasts are a key input of production planning systems. Despite constant advances in forecasting techniques, they remain subject to forecast error. To manage the resulting demand uncertainty, a common approach is to conduct planning in a rolling-horizon fashion. In each planning cycle, a production plan is determined over the horizon but only its first periods are implemented. Decisions for later periods are revised in following planning cycles using newly available forecasts. By frequently updating forecasts and decisions, improved production plans can be derived.

Rolling-horizon planning is applied extensively to a wide variety of industry settings (Sahin et al., 2013). Traditionally, a combination of deterministic models and rule-of-thumbs for safety stock calculations are used to determine production plans (Sridharan and Berry, 1990; Yano and Carlson, 1987). However, these models react only passively to updates and errors in forecasts and do not yield cost-effective decisions (Tang and Grubbström, 2002; Vargas and Metters, 2011). Conversely, stochastic models explicitly integrate uncertainty through probability distributions to calculate safety stocks that satisfy demand at minimum cost. Stochastic models have been applied with great success in production problems such as material requirement planning (Thevenin et al., 2021), lot sizing (Sereshti et al., 2020), and lot sizing and scheduling (Hu and Hu, 2018).

Using updated demand forecasts in each review period is a major strength of rolling-horizon planning. Planning performance could be further improved by characterising the forecast revision process and integrating it in stochastic planning models. The martingale model of forecast evolution (MMFE) formalised by Heath and Jackson (1994) proposes two methods to describe forecast evolution as a stochastic process. The additive version measures the absolute difference between forecasts. The multiplicative version

measures forecast updates relative to the forecasts themselves, which has been shown to better reflect forecasters' behaviour (Hausman, 1969). The MMFE is a powerful framework that has been applied to varied problems such as defining supply contracts (Donohue, 2000), capacity planning (Boyacı and Özer, 2010), inventory management (Iida and Zipkin, 2006; Özer and Wei, 2004; Wang and Tomlin, 2009), and multi-ordering newsvendor problems (Biçer and Seifert, 2017; Wang et al., 2012). The above literature shows that determining optimal policy with forecast evolution models can provide significant cost reductions.

Yet, applications of MMFE models in rolling-horizon production planning problems are limited. In particular, managing the production of multiple products with limited capacity and correlated forecast evolution is a challenging problem that has only been solved heuristically so far (Albey et al., 2016; Norouzi, 2012; Ziarnetzky et al., 2018). Further, despite their dependence on data, MMFE models have seen only limited applications to real-world data. To the best of our knowledge, the only applications of MMFE to real-world problems have been proposed by Albey et al. (2015), who apply additive MMFE in the semi-conductor industry, and Pinçe et al. (2021), who apply the multiplicative model to an agricultural supply chain. A detailed analysis of both MMFE models on real-world data is missing.

Forecast evolution models provide a probabilistic description of the forecast evolution process, which can be integrated in stochastic models. However, determining the resulting optimal production policies analytically is often too complex for realistic problems. Stochastic programming can be used instead to solve multi-stage problems by representing uncertainty as a scenario tree over the horizon (Dupačová et al., 2000). Multi-stage formulations describe the progressive resolution of uncertainty in sequential stages in which decisions can be adapted (King and Wallace, 2012). These recourse opportunities lead to less conservative first-stage decisions that reduce overall costs. Multi-stage models are closely linked to rolling-horizon planning since they explicitly model the flexible adaptation of decisions over time. Yet, the application of multi-stage stochastic models in rolling horizon has received only limited attention for production planning. A majority of existing literature considers only static evaluation of stochastic models and ignore rolling-horizon implementations. In particular, the value of recourse offered by multi-stage formulations has only been partially quantified, since static comparisons with two-stage models proposed by (Hu and Hu, 2018; Kazemi Zanjani et al., 2010) may overestimate the value of recourse (Stephan et al., 2010). Another approach to study the value of flexibility in rolling-horizon planning has been proposed by Tavaghof-Gigloo

and Minner (2020) who introduce a heuristic to reduce safety stocks when capacity is large. They do not discuss the value of recourse when limited capacity is shared by several products. Hence, a complete evaluation of the value of recourse in rolling-horizon planning is still missing.

Designing multi-stage models that fit existing rolling-horizon practice also remains an open question. An essential but often overlooked aspect of rolling-horizon planning is the central role of communicating reference plans to coordinate the different planning processes. Communicability of reference plans is essential in complex supply chains since planning is decomposed in several consecutive steps. The rolling-horizon cycle is set up so that plans are propagated through the supply chain and act as input for dependent planning steps. For instance, raw-material requirements are propagated upward the supply chain to organise purchasing, production and transportation activities. Similarly, production plans are communicated to downstream parts to schedule workforce and finished-goods deliveries. Yet, in their common form, multi-stage production planning models do not provide communicable reference plans. Instead, they determine a tree of production decisions over the planning horizon (Escudero et al., 1993; Körpeoğlu et al., 2011).

Reference plans should further remain stable in rolling horizon to efficiently coordinate upstream and downstream parts of the supply chain. Frequent plan changes create nervousness, which can reduce confidence in planning and increase overall costs (Atadeniz and Sridharan, 2020). To mitigate planning nervousness, strategies have been proposed that penalise or prohibit plan changes (Koca et al., 2018; Sridharan and Berry, 1990). While effective to increase planning stability, these methods reduce flexibility and may lead to important cost increase. Hence, the trade-off between planning communicability, stability and flexibility needs to be carefully investigated.

This research is motivated by a collaboration with a company in the agrochemical industry for a set of products with high seasonality. This dynamic setting is especially challenging since early forecasts have poor accuracy even in the close future. The flexibility of rolling-horizon planning is then essential to adapt production decisions and ensure that demand can be met. Analysing the uncertainty of forecasts and integrating it in stochastic models in rolling-horizon planning is a promising direction to improve production planning and supply chain management.

1.2. Research Questions

This thesis studies the integration of stochastic optimisation in rolling-horizon production planning when demand is uncertain and seasonal. We analyse the interactions between stochastic models, rolling-horizon planning and forecast uncertainty. Research questions that cover the main contributions proposed in the three chapters of this thesis are stated as follows.

Stochastic models are still rarely used in practice despite their complementarity with rolling-horizon planning. This suggests that there remain open questions regarding how to apply stochastic models in real-world problems. In the academic literature, a very common assumption is that demand distributions are known. However, distributions are not given in practice and have to be estimated from past forecast and demand data.

(RQ 1) *How can stochastic models be applied from the available history of forecast and demand data?*

This research question aims to foster the application of stochastic optimisation by studying the link between planning data and model development. In Chapter 3, we identify important barriers that prevent the application of stochastic programming to master productions scheduling and relate them to existing literature. We develop a framework to overcome the barriers based on data and apply it to our real-world case study in the agrochemical industry. A second approach is proposed in Chapter 4, in which we use MMFE models to characterise forecast uncertainty from data. This method is especially suitable in rolling-horizon planning since it relies on the history of forecast and demand data that is readily available to practitioners.

Forecast evolution models can be seen as a bridge between theory and practice for stochastic rolling-horizon planning. However, the integration of MMFE in complex planning environments remains challenging. In particular, limited production capacity, inventory costs and product correlations lead to difficult problems that have only be solved approximately so far. Further, despite close link to practice and its reliance on data, MMFE has seen only limited application to real-world problems. In-depth analyses of the additive and multiplicative models in practical settings are missing.

(RQ 2.1) *How can MMFE models be integrated into complex production planning environments?*

(RQ 2.2) *What are strengths and limitations of the additive and multiplicative MMFE when applied from real-world data?*

We propose several approaches to integrate forecast evolution models in production

planning. In Chapter 2, we consider a short seasonal demand and solve the dynamic programming model analytically for a single product. Properties of the optimal policy are analysed to develop a heuristic to plan several correlated products. In Chapter 4, we integrate the MMFE in a general lot-sizing problem and solve the resulting non-linear formulation using existing piecewise-linearisation techniques. Both additive and multiplicative models are applied with real-world data and their performance are compared. We identify their advantages and limitations as well as their sensitivity to problem parameters. We also provide general recommendations to practitioners.

Forecast evolution models and forecast uncertainty models can be used to formulate multi-stage optimisation models that yield less conservative decisions. However, the value of recourse offered by multi-stage formulations has not been measured precisely in rolling-horizon production planning.

(RQ 3) *What is the value of recourse in rolling-horizon planning and what parameters influence it?*

This research question is studied thoroughly in the three chapters of this thesis by performing repeated simulations in rolling horizon. The value of recourse is measured accurately by defining stochastic models without recourse as benchmarks. We perform extensive sensitivity analyses to highlight parameters that influence the value of recourse such as available capacity and product correlation. When possible, we set up out-of-sample simulations to evaluate the value of recourse when the uncertain process is unknown and has to be estimated from data.

Recourse can reduce production costs thanks to less conservative decisions. However, traditional multi-stage models do not respect the constraints of rolling-horizon planning such as providing stable reference plans in each planning cycle.

(RQ 4) *How can stochastic models that satisfy the trade-off between planning flexibility, stability and communicability be developed?*

We use several methods to determine flexible recourse decisions and stable reference plans with stochastic models. In Chapter 2, a linear policy approximation is used to calculate expected production plans. In Chapter 3, products are aggregated in optimal families so that first-stage decisions are placed on the family level and recourse decisions on the product level. In Chapter 4, we introduce partial recourse over the planning horizon so that decisions are only flexible in the later part of the horizon. Since reference plans are available in each review period, the nervousness resulting from stochastic models with recourse can be quantified. Strategies for improving planning stability such as product aggregation and freezing decisions are also analysed.

1.3. Outline

The chapters of this thesis are based on three distinct working papers. The remainder of this thesis unfolds as follows.

In Chapter 2, we study the production of seasonal goods with short selling season. Production capacity is limited so that planning starts ahead of the selling season. Demand forecasts are uncertain and updated periodically according to the MMFE. Since early production leads to inventory holding costs, we model the key trade-off between producing early to avoid lost sales and producing late to minimise costs. We formulate the problem as a dynamic programming model and solve it optimally for the single-product case. A heuristic is developed for managing the production of multiple products with correlated forecast evolution. The value of forecast evolution models and the recourse they provide is evaluated in repeated rolling-horizon simulations. Chapter 2 is based on Forel and Grunow (2020b).

In Chapter 3, we identify barriers that limit the application of stochastic models and develop strategies to overcome them. In particular, we discuss how to define and model demand uncertainty from limited available data and how to communicate stable reference plans while allowing flexible decisions. A two-stage stochastic model with recourse is derived that increases both planning flexibility and stability based on the aggregation of decisions over optimal product families. The approach is demonstrated on a real-world case study from the agrochemical industry. Chapter 3 is based on Forel and Grunow (2020c).

In Chapter 4, we combine and extend the insights from the previous two chapters as we introduce forecast evolution models for a dynamic demand realising over several periods. We integrate additive and multiplicative MMFE into general lot-sizing problems and solve the resulting non-linear models using existing piecewise-linearisation techniques. To increase flexibility, we introduce production recourse in later periods of the horizon via a multi-stage scenario tree. The value of forecast evolution models and the value of recourse are highlighted in rolling-horizon simulations using both artificial and real-world data. We identify advantages and limitations of additive and multiplicative MMFE when probability distributions are estimated from past data. Chapter 4 is based on Forel and Grunow (2020a).

In Chapter 5, we summarise our findings and answer the research questions. We discuss the limitations of our work and propose directions for future research.

Chapter 2

Production planning for a short seasonal demand with forecast evolution

Abstract

This paper studies the production of several products in a rolling-horizon setting as forecasts are periodically updated. Demand is observed over a short selling season and, since capacity is limited and carrying inventory is expensive, it is essential to make efficient production decisions ahead of the selling season. We aim to bridge the gap between current industry practice, that implements deterministic rolling-horizon planning based on forecasts, and academia, where stochastic models are often developed while ignoring their rolling-horizon implementation. We integrate forecast evolution models in a production planning environment with a fill-rate service-level constraints and inventory costs. We model the evolution of demand forecasts with the martingale model of forecast evolution, which captures the property that forecast accuracy increases as the selling season gets closer. The optimal production policy is determined through a dynamic programming model for the single-product case and is adapted into an iterative heuristic for the multi-product case. Through repeated rolling-horizon simulations, we show that stochastic models that do not account for forecast evolution often fail to reach the service-level targets. Explicit integration of forecast evolution leads to high demand satisfaction and up to 13% cost reductions. Further, the production policy with forecast evolution yields a closer link between forecast nervousness and production nervousness thus improving planning visibility. We show the importance of explicitly integrating forecast evolution in production planning. We identify the influence of correlation and timing of uncertainty on the planning policy and performance.

2.1. Introduction

Rolling-horizon planning is an effective and flexible framework for dealing with demand uncertainty that is widely used in practice. It is based on periodically reviewing demand forecasts and adapting production decisions so that a production plan is derived over the planning horizon in each review period. Often, practitioners implement simple deterministic rolling-horizon planning without explicit stochastic models for demand uncertainty. Conversely, academia is constantly pushing the boundaries of decision-making under uncertainty in terms of scalability, computational efficiency, and relying on fewer limiting assumptions. However, the interplay between stochastic optimisation and rolling-horizon planning remains understudied.

In the academic literature, the martingale model of forecast evolution (MMFE) has been introduced to model the uncertainty of the forecast revision process. It describes the amplitude and timing of forecast changes over the planning horizon with probability distributions. The MMFE captures the property that forecast uncertainty reduces over time and includes correlations between the forecast updates of different products. By explicitly integrating forecast evolution into planning and determining the corresponding production policy, we are able to fully utilise the flexibility of rolling-horizon planning to react to new knowledge. In this light, forecast evolution models implemented in rolling-horizon planning can be seen as the bridge between academic research and practitioners.

In each review period, production decisions should be determined and communicated over a prediction horizon. Hence, the optimal production policy should be translated into a production plan. The plan is not only used to implement the production decisions for the first periods but also provides a reference to upstream and downstream members of the supply chain. The production plan can support many strategic, tactical and operational decisions such as whether to outsource part of production, managing the purchasing and delivery of raw materials as well as determining the workforce schedule. Because the plan is communicated through the supply chain, it is important to ensure that there are no unnecessary changes and that the plan remains stable. In this paper, we determine the optimal policy that can be implemented flexibly, the expected production plan communicated in each review period, and evaluate its nervousness in rolling-horizon implementation.

The value of forecast evolution can be measured by comparing the MMFE model to a stochastic model that ignores forecast evolution. Comparing a priori expected performance of the two models is not sufficient since this model also benefits from updated

forecasts in rolling-horizon planning. The value of forecast evolution needs to be assessed by comparing the results of repeated rolling-horizon simulations in terms of achieved service level, operational costs and planning nervousness.

Matching supply and demand is especially difficult for seasonal goods since typical challenges include stock-out, obsolescence, and inventory left-over. These challenges are present in many industries. Volatility and unpredictability of demand have been described as one of the main causes of shortages of essential medicines and vaccines for instance (Aditi et al., 2018; Leung et al., 2016). Due to limited capacity, seasonal demand forces production to start well in advance of the selling periods (Fisher and Raman, 1996). In the fashion industry, firms may commit up to nine months before the selling season (Wang et al., 2012), and similarly, in the agricultural goods industry, production of hybrid seeds starts a year before the selling season (Bansal and Nagarajan, 2017; Jones et al., 2001). Producing ahead of the selling season leads to inventory on hand, which can be costly for the company. Inventory holding costs include warehouse facility costs, costs of handling and storing the inventory, and the opportunity costs of invested capital. Determining the optimal inventory level through the production season is essential for companies managing seasonal goods.

On the other hand, manufacturer often need to satisfy service-level agreements and are penalised for not reaching the agreed-upon target. Hence, as demand forecasts are periodically reviewed, there is a fundamental trade-off between early and late production. Late production fulfils demand in a just-in-time fashion with low inventory but has only limited flexibility to react to demand forecast increases since capacity is utilised in later periods. Early production builds early inventory that allows one to react to potential forecast increases but implies higher inventory costs. This trade-off becomes even more challenging when several correlated products share the capacity since the optimal production quantity depends on the products inventory, forecast updates and correlation. In this paper, we formalise the trade-off between early and late production using the MMFE and study the resulting production policy.

2.1.1. Outline

The remainder of this paper is organised as follows. Related literature is reviewed in Section 2.2 with a focus on forecast evolution. The problem setting and forecast evolution model are presented in Section 2.3. In Section 2.4, a fill-rate service level constraint is presented and solved for additive and multiplicative forecast evolution. In Section 2.5, a dynamic programming model is developed to determine the optimal pro-

duction policy with forecast evolution following the MMFE. The optimal production policy is determined analytically for the single-product case and is adapted into a decomposition/coordination heuristic for the multi-product case. In Section 2.6, the value of forecast evolution is assessed by performing repeated rolling-horizon simulations. We show that the traditional rolling-horizon model without forecast evolution fails to ensure high demand satisfaction in several simulation settings. The MMFE model achieves high demand satisfaction consistently thanks to efficient planning decisions throughout the production season and can reduce costs by up to 13%. For the multi-product case, the heuristic outperforms the benchmark in all instances as the MMFE model achieves higher demand satisfaction and can reduce inventory and production costs. In Section 2.7, we conclude by summarising our findings and provide an outlook for future research.

2.2. Related literature on forecast evolution

An early analysis of forecast evolution was proposed by Hausman (1969) who shows that a log-normal distribution can well model the forecast revision process of seasonal goods such as agricultural products and clothing articles. In a subsequent paper, Hausman and Peterson (1972) use this forecast evolution model to manage the production of several products with limited capacity and propose three heuristics to solve the problem. Forecast evolution was then formalised by Graves et al. (1986) and Heath and Jackson (1994) into the MMFE. Despite its apparent simplicity, the MMFE is a powerful framework suitable for a wide variety of problems including capacity planning (Boyacı and Özer, 2010) and defining supply contracts (Donohue, 2000).

Inventory management with capacity restrictions has been studied, for instance, by Özer and Wei (2004) who consider a manufacturer with limited capacity and advance demand observation. They determine the optimal policy for cases with and without fixed ordering costs. Toktay and Wein (2003) study a single-product capacitated production setting with stationary demand. They model the problem as a single queue to derive the optimal modified base-stock policy. Norouzi and Uzsoy (2014) consider correlated forecast evolution over several demand periods in a single-product environment with a capacity constraint. Recently, Ban et al. (2019) have proposed a generalisation of the additive MMFE to include covariate information and apply it to the procurement of new products with a short life cycle.

While in the above works the demand is observed over multiple periods, the MMFE is

particularly adapted to model a seasonal demand contained in a short selling season. Wang and Tomlin (2009) apply a multiplicative version of the MMFE to a newsvendor problem facing both lead-time uncertainty and demand uncertainty. They exhibit the optimal ordering policy and investigate the trade-off between supply and demand risk. Although the demand season is short, it may be possible to react to forecast evolution using a second order as presented by Milner and Kouvelis (2005) who study a single demand period with two ordering opportunities. They consider both limited capacity and inventory costs and compare three demand models including the additive MMFE. Li et al. (2009) consider a seasonal demand observed over a season of arbitrary length. They propose analytical results for the optimal ordering policy for a seasonal demand incurring inventory and backorder costs in an uncapacitated setting. The two ordering opportunity structure is especially adapted for managing the production of hybrid seed where one considers sequentially the production in northern and southern hemispheres: Jones et al. (2001) provide analytical results on a problem including harvesting costs and inventory costs and Bansal and Nagarajan (2017) extend the model to account for limited capacity.

Most closely related to our paper, recent research has looked into applying the MMFE to multi-ordering newsvendor problems in which the newsvendor has several opportunities to order before satisfying the demand of the selling period. Wang et al. (2012) study an uncapacitated setting where ordering costs increase as the selling season gets closer. They prove that a base-stock policy is optimal and that the base-stock level in each period depends linearly on the updated forecast. Biçer and Seifert (2017) extend the model by introducing a capacity limit for the quantity ordered in each period. They propose a heuristic to plan several correlated products sharing capacity. However, existing literature on the multi-ordering newsvendor does not consider that early production leads to on-hand inventory and that carrying inventory up to the selling season is costly. When facing seasonal demand, this may result in carrying expensive inventory for several months. We extend the multi-ordering newsvendor setting to specifically investigate the trade-off between early production, which anticipates demand but builds-up expensive inventory, and late production, which aims for low inventory costs but takes the risk of not being able to react to an increase in demand forecast due to limited capacity. Further, the fact that the planning would be implemented in a rolling-horizon fashion is not considered, which leads to a biased evaluation of the stochastic benchmark without forecast evolution.

To summarise, our contribution is threefold. First, we extend the multi-ordering newsven-

dor to a production environment with fill-rate service level constraints and inventory holding costs and we determine the optimal production policy. Second, we adapt the policy to an iterative heuristic for multiple correlated products sharing capacity. Third, we assess the value of integrating forecast evolution in production planning through repeated rolling-horizon simulations.

2.3. Problem setting

This section introduces the problem setting for the single-product case and describes the forecast evolution model. Consider a planning horizon of T production periods with limited capacity K in each period. The demand is uncertain and observed in period $T+1$. Missed demand is considered a lost sale. In each period, the planner reviews the initial inventory on hand x_t and decides on an immediate production quantity Q_t as well as a production plan over the planning horizon $\mathbf{Q}_t = \{Q_{t,s}, s \in t+1, \ldots, T\}$. Let x_t and y_t be the inventory on hand at the beginning and end of period t respectively. The inventory evolution is then described by $y_t = x_t + Q_t = x_{t+1}$. Per-unit production cost p and inventory cost h are incurred so that the overall cost in each period is given by $c_t = p + h \cdot (T - t)$. With positive inventory holding costs, the overall costs c_t are strictly decreasing as the season progresses. Despite demand being uncertain, the planner needs to guarantee a minimum fill-rate service level. The service-level targets β is set exogenously, for instance by upper management, and is seen as a hard constraint by the planner.

2.3.1. Additive MMFE

Let the actual demand observed in the selling period be denoted as $D_{T+1|T+1} = D_{T+1}$. At the beginning of each period t, an updated forecast $D_{T+1|t}$ is available. The evolution of the demand forecast follows the relation $D_{T+1|t} = D_{T+1|t-1} + \varepsilon_t$ where ε_t is the forecast update in period t. The main assumption of the additive MMFE is that the forecast updates are independent and identically distributed with a normal distribution $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ where σ_t is the standard deviation of the forecast update in period t. In each period, knowing the latest cumulated update forecast $A_t = \sum_{s=1}^{t} \varepsilon_s$, the demand in the selling season is a random variable with distribution $D_{T+1} \sim \mathcal{N}(\mu + A_t, \tilde{\sigma_t}^2)$ where $\mu = D_1$ is the initial demand forecast and $\tilde{\sigma_t}^2 = \sum_{s=t+1}^{T+1} \sigma_s^2$ is the residual uncertainty at time t. Thus, uncertainty decreases as the selling season gets closer. The variance

of the forecast updates over the planning horizon describes the timing of uncertainty resolution .

2.3.2. Multiplicative MMFE

Following multiplicative MMFE, the forecast updating process at the beginning of planning period t is described by $D_{T+1|t} = D_{T+1|t-1} \cdot \exp(\varepsilon_t)$. The forecast update ε_t is independent and identically distributed with a normal distribution $\varepsilon_t \sim \mathcal{N}(-\frac{\sigma_t^2}{2}, \sigma_t^2)$. An initial demand forecast D_1 is available in the first period. The expectation of the logarithm of the demand in period 1 is denoted by $\mu = \ln(D_1) - \sum_{t=1}^{T+1} \sigma_t^2/2$. In period t, the forecast follows the relation $D_{T+1|t} = D_1 \cdot \exp(\varepsilon_2 + \cdots + \varepsilon_t) = \exp(\mu + A_t)$ where $A_t = \sum_{s=2}^{t}(\varepsilon_s + \sigma_s^2/2)$ is the cumulative forecast update at the beginning of period t. In each period, knowing the cumulative forecast update A_t, the demand follows a log-normal distribution $\ln(D_{T+1}) \sim \mathcal{N}(\mu + A_t, \tilde{\sigma_t}^2)$ where $\tilde{\sigma_t}^2 = \sum_{s=t+1}^{T+1} \sigma_s^2$ is the residual uncertainty at time t.

2.3.3. Timeline of events

In each period $t \in \{1, 2, \dots T\}$, the sequence of events is as follows: (1) a forecast update ε_t is observed, (2) the updated demand forecast $\mu + A_t$ is determined, (3) on-hand inventory x_t is reviewed, (4) the planner decides on the immediate production quantity Q_t and the production plan over the rest of the prediction horizon $\mathbf{Q}_t$, (5) the inventory position is increased to y_t and (6) a production cost $c_t \cdot Q_t$ is incurred. In the selling period, demand is observed and satisfied with on-hand inventory and the service level is finally measured.

2.4. Fill-rate service level

The fill-rate service level is defined as the proportion of demand that is satisfied directly from on-hand inventory. At the end of the demand season, the achieved fill-rate service level is measured as $\beta = 1 - \max(x_{T+1} - D_{T+1}, 0)/D_{T+1}$. At time T, the fill-rate target β is expected to be reached in period $T+1$ if the following inequality holds (Silver et al., 2016; Thomas, 2005):

$$\mathbb{E}_{D_{T+1}|A_T}[\max(x_{T+1} - D_{T+1}, 0)] \leq (1 - \beta)\, \mathbb{E}_{D_{T+1}|T}[D_{T+1}]. \tag{2.1}$$

Since there is no opportunity for recourse after the last production period, the problem reduces to a newsvendor problem with a capacity limit and fill-rate constraint.

2.4.1. Additive case

In period T, demand follows a normal distribution $D_{T+1} \sim \mathcal{N}(\mu + A_T, \sigma^2_{T+1})$.

Proposition 2.1. *The optimal inventory at the beginning of the selling season is given by* $S_T(A_T) = \mu + A_T + b_T(A_T)$ *where* b_T *is a safety-stock factor depending on* A_T *and given by*

$$b_T(A_T) = \sigma_{T+1} \cdot \mathcal{L}^{-1}\left(\frac{(1-\beta)(\mu + A_T)}{\sigma_{T+1}}\right)$$

where $\mathcal{L}^{-1}(\cdot)$ *is the inverse of the first-order loss function of a standard normal distribution.*

Proposition 2.1 states the inventory target to be reached at the end of the production season. The details are provided in Appendix A.1. Since the first-order loss function of any normal distribution can be reformulated to depend on the first-order loss function of a standard normal distribution, the computation of the safety stock factor is computationally inexpensive. The inverse function $\mathcal{L}^{-1}$ can be pre-computed so that the safety stock factor can be evaluated quickly for any problem settings.

Lemma 2.1. *The optimal inventory function* S_T *is defined over* $]-\mu; +\infty[$*, is convex with a minimum value obtained in* $\underline{A}_T = \frac{\sigma_{T+1}}{1-\beta} \cdot \mathcal{L}\left(\Phi^{-1}(\beta)\right) - \mu$*, and is strictly increasing over* $\left[\underline{A}_T; +\infty\right[$.

The details of the proof are available in Appendix A.2. This proposition specifies the edge behaviour of the fill-rate constraint for a normally distributed demand when the distribution mean becomes much smaller than its standard deviation. This is a common, yet often overlooked, issue of assuming that demand follows a normal distribution.

2.4.2. Multiplicative case

In period T, demand follows a log-normal distribution $\ln(D_{T+1} \mid A_T) \sim \mathcal{N}(\mu + A_T, \sigma^2_{T+1})$.

Proposition 2.2. *The optimal inventory at the beginning of the selling season is given by*

$$S_T(A_T) = \mathcal{L}^{-1}\left((1-\beta)\exp\left(\mu + A_T + \frac{\sigma^2_{T+1}}{2}\right), D_{T+1} \mid A_T\right)$$

where $\mathcal{L}^{-1}(\cdot, D_{T+1} \mid A_T)$ *is the inverse of the first-order loss function* $\mathcal{L}(\cdot, D_{T+1} \mid A_T)$ *where* $D_{T+1} \mid A_T$ *is log-normally distributed. The optimal inventory function* S_T *is strictly increasing over* $\mathbb{R}$.

The details of the proof are available in Appendix A.3. The safety stocks reserved by the multiplicative MMFE model can be deduced as $b_T = S_T(A_T) - \mathbb{E}[D_T \mid A_T]$ even though no closed form is available.

2.4.3. Comparison of inventory targets

The optimal inventory target functions are illustrated in Figure 2.1 for initial forecast $D_1 = 100$ and service-level target $\beta = 0.95$. The part of the inventory function that is subject to the edge behaviour described in Section 2.4.1 is shown in dashed line. Figure 2.1 suggests that the inventory target is approximately linear in the forecast update for additive MMFE when the coefficient of variation is low, whereas the inventory target is exponential in the forecast update for multiplicative MMFE. The safety stock

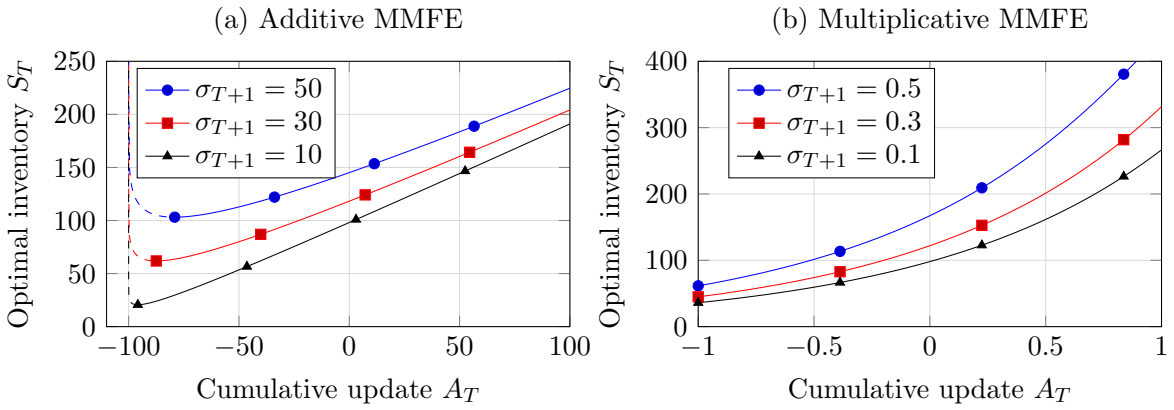

Figure 2.1.: Optimal last-period inventory as a function of the cumulative update.

is also represented as a function of the residual uncertainty for several fill-rate targets in Figure 2.2, which shows that the safety stock increases exponentially with both the standard deviation of the residual uncertainty and the service level target.

2.5. Inventory policy and production plan

The previous section describes the inventory target to reach by the end of the planning season. The MMFE model should find the optimal inventory trajectory from period 1 to T to reach the last-period inventory target with minimum cost.

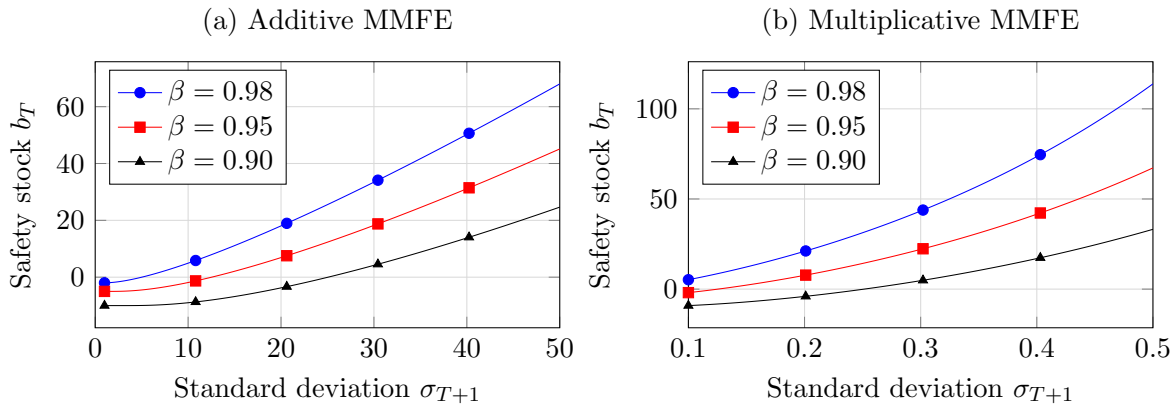

Figure 2.2.: Sensitivity of safety stock as a function of the standard deviation for $A_T = 0$.

2.5.1. No forecast evolution: traditional rolling-horizon planning

A naive approach to solve the planning problem is to ignore forecast evolution and aggregate forecast and residual uncertainties together. This traditional rolling-horizon planning model can be modelled as a linear optimisation problem

$$\min_{\mathbf{y}} \quad \sum_{\tau=t}^{T} c_\tau \cdot (y_\tau - x_\tau) \tag{2.2a}$$

$$\text{s.t.} \quad y_\tau = x_\tau + Q_\tau, \quad \forall \tau \geq t, \tag{2.2b}$$

$$Q_\tau \leq K, \quad \forall \tau \geq t, \tag{2.2c}$$

$$y_T \geq \tilde{S}_t(A_t) \tag{2.2d}$$

where constraint (2.2b) specifies the inventory balance in each period, constraint (2.2c) enforces the capacity limitation, and constraint (2.2d) ensures that the final inventory is large enough to satisfy the fill-rate constraint. The target inventory $\tilde{S}_t(A_t)$ is deduced by adapting Proposition 2.1 and Proposition 2.2 for the additive and multiplicative model respectively. Note that the model considers the total residual uncertainty $\tilde{\sigma}_t$ at time t and not only the final demand uncertainty. Since this model ignores forecast evolution and aims to minimise inventory costs, production is pushed to later periods. Hence, there is no capacity reserved in later periods and the model cannot react to a forecast increase.

2.5.2. Single product: optimal policy

In this section, we consider that the last-period inventory target function S_T is strictly increasing and invertible, as is the case for multiplicative MMFE. The changes needed for the additive MMFE model to fit these properties are discussed in Section 2.6.1.

Dynamic programming formulation

Since the fill-rate constraint in Equation 2.1 is a hard-constraint, it is not possible to write the optimisation problem as a dynamic programming problem. To find a feasible inventory trajectory, we introduce a shortfall penalty cost $\gamma \geq c_T$, which penalises the amount by which the inventory target S_T cannot be attained. The minimum costs incurred in period T are now given by

$$V_T(x_T, A_T) = \min_{x_T \leq y_T \leq x_T + K} c_T(y_T - x_T) + \gamma \cdot \max(S_T(A_T) - y_T, 0). \tag{2.3}$$

The planning problem from period 1 to T can thus be formulated as a dynamic programming problem in which the cost-to-go in each period $t \in \{1, \ldots, T-1\}$ is given by

$$V_t(x_t, A_t) = \min_{x_t \leq y_t \leq x_t + K} c_t(y_t - x_t) + \mathbb{E}_{A_{t+1}|A_t}[V_{t+1}(y_t, A_{t+1})].$$

The key trade-off between early and late production is now captured by the inventory costs in all periods and the shortfall cost in the last period.

Optimal production policy

The optimal production policy can be determined by solving the dynamic programming problem by recursion using Bellman's principle of optimality.

Lemma 2.2. *The minimum cost incurred in period T is given by*

$$V_T(x_T, A_T) = \begin{cases} c_T \cdot K + \gamma \cdot (S_T(A_T) - (x_T + K)), & \text{if } x_T < S_T(A_T) - K \\ c_T \cdot (S_T(A_T) - x_T), & \text{if } S_T(A_T) - K < x_T \leq S_T(A_T) \\ 0, & \text{if } S_T(A_T) \leq x_T \end{cases}$$

and the optimal production volume is given by $Q_T^* = \max\{\min\{S_T(A_T) - x_T, K\}, 0\}$.

The proof is straightforward since the last-period cost given in Equation (2.3) is convex and piecewise-linear in y_T.

Proposition 2.3. *The optimal inventory at the end of each period $t \in \{1, \dots, T-1\}$ follow a base-stock policy. The inventory targets $S_t(A_t)$ are the solution of $g_t(y_t, A_t) = 0$ where the marginal cost function $g_t(y_t, A_t)$ is given by*

$$\begin{aligned} g_{T-1}(y_{T-1}, A_{T-1}) = c_{T-1} - c_T \left[F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1} + K)) - F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1}))\right] \\ -\gamma \left[1 - F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1} + K))\right] \end{aligned}$$

for $t = T-1$, and

$$\begin{aligned} g_t(y_t, A_t) = c_t - c_{t+1} + \int_{-\infty}^{S_{t+1}^{-1}(y_t)} g_{t+1}(y_t, a) f_{A_{t+1}|A_t}(a) da \\ + \int_{S_{t+1}^{-1}(y_t+K)}^{+\infty} g_{t+1}(y_t + K, a) f_{A_{t+1}|A_t}(a) da \end{aligned}$$

for all period $t < T-1$, where $f_{A_{t+1}|A_t}$ and $F_{A_{t+1}|A_t}$ are resp. the density and cumulative distribution functions of A_{t+1} having observed A_t.

The proof is adapted from Biçer and Seifert (2017) and is given in Appendix A.4. Proposition 2.3 provides a method to determine the optimal inventory targets in each period recursively. The optimal inventory targets depend on the cost and capacity parameters as well as the volatility of the forecast updates in each period.

Expected production plan

In each period, the planner is not only interested in the immediate production quantity Q_t but also in the production plan over the remaining horizon. This plan serves as a reference and is communicated both internally and externally. However, since the optimal production quantities in future periods depend on the forecast update, future production is uncertain. The expected production plan in period t is deduced from the expected inventory target as

$$\mathbb{E}[Q_\tau] = \mathbb{E}[\min\left\{\max\left\{S_\tau(A_\tau) - \mathbb{E}[x_\tau], 0\right\}\right], K\}, \quad \forall \tau \geq t.$$

The production plan is calculated from period t until period T such that the expected inventory $\mathbb{E}[x_\tau]$ is given by $\mathbb{E}[x_\tau] = x_t + \sum_{s=t}^{\tau-1} \mathbb{E}[Q_s]$. The expectation of the inventory targets in each period can be determined numerically, for instance by sampling future forecast updates and averaging the end-inventory targets from Proposition 2.3.

2.5.3. Multiple products: approximate policy

The manufacturer now manages a portfolio of N products sharing the same production resources with limited capacity K. Let x_t^j and y_t^j be the inventory of product j available respectively at the beginning and end of period t. The per-unit production costs p^j and inventory holding costs h^j of all products are different and reflect their value: the higher the product value, the higher its inventory cost. Without loss of generality, the products are sorted in decreasing order of value. At time t, the demand forecast is denoted by $\mathbf{D_{T+1|t}} = \begin{bmatrix} D^1_{T+1|t} & \dots & D^N_{T+1|t} \end{bmatrix}$ and the demand of product j observed in the selling period is $D^j_{T+1|T+1} = D^j_{T+1}$. An initial demand forecast D^j_1 is available for each product.

Overview of decomposition/coordination algorithm

As observed by Hausman and Peterson (1972) and Biçer and Seifert (2017), determining the optimal production policy for multiple correlated products with limited capacity is a very complex problem and one has to resort to heuristic approaches. In this section, we present a heuristic production policy for N products with correlated demand. The heuristic is based on results obtained for the single-product setting. For the sake of simplicity, we focus on the additive MMFE. The method can be easily extended to the multiplicative setting.

At the beginning of each planning period, the additive MMFE describes the forecast evolution as $\mathbf{D_{T+1|t}} = \mathbf{D_{T+1|t-1}} + \varepsilon_t$ where $\varepsilon_t = \begin{bmatrix} \varepsilon_t^1 & \dots & \varepsilon_t^N \end{bmatrix}$ is the vector of forecast updates over all products. The forecast update vector is independent, identically distributed and follows a multivariate normal distribution $\varepsilon_t \sim \mathcal{MN}(0, \mathbf{\Sigma}_t)$ where $\mathbf{\Sigma}_t$ is the covariance matrix of the forecast updating process at time t. The covariance matrix is an $N \times N$ matrix that can be expressed as $\mathbf{\Sigma}_t = (\rho_{(i,j),t}\sigma_{i,t}\sigma_{j,t})_{ij}$ where $\rho_{(i,j),t}$ is the correlation between the forecast updates of product i and product j at time t. The covariance matrix describes both the uncertainty of the forecast updates of the different products and the linear relation between the different product updates. In each period, knowing the latest cumulative update forecast $\mathbf{A_t} = \sum_{s=1}^{t} \varepsilon_s$, the demand follows a multivariate normal distribution $\mathbf{D_{T+1}} \sim \mathcal{MN}(\mu + \mathbf{A_t}, \tilde{\mathbf{\Sigma}}_\mathbf{t})$ where $\tilde{\mathbf{\Sigma}}_t = (\tilde{\sigma_{(ij),t}}^2)_{ij}$ describes the residual uncertainty at time t with $\tilde{\sigma}^2_{(ij),t} = \sum_{s=t+1}^{T} \rho_{(i,j),s}\sigma_{i,s}\sigma_{j,s}$.

The solution procedure contains two main steps. In the first step, the optimal inventory target to be reached at the end of the production season is determined for each product independently. It is based on the fill-rate service level constraint and accounts for the residual demand uncertainty. In the second step, an iterative procedure is developed

to determine an aggregated inventory target over all products and allocate it to the individual products.

Optimal inventory in the last period

In line with industry practice and recent research, the target service level β_j is set independently for each product (Meistering and Stadtler, 2017). The optimal inventory to have on hand at the beginning of the demand period for each product is then independent of the inventory target of the other products and can be deduced directly from Proposition 2.1.

Aggregated inventory target

The aggregated target is obtained by solving an aggregated problem, equivalent to a single-product problem. Consider the aggregated forecast update $A_t^{(agg)} = \sum_{j=1}^{N} A_t^j$. Under additive MMFE, the aggregated forecast update distribution follows a normal distribution $A_t^{(agg)} \sim \mathcal{N}(0, \sigma_t^{(agg)})$ where $\sigma_t^{(agg)} = [1 \dots 1]\mathbf{\Sigma}_t[1 \dots 1]^T$. We specify the aggregated initial forecast $D_1^{(agg)} = \sum_{j=1}^{N} D_1^j$ and costs $c_t^{(agg)} = \sum_{j=1}^{N} \frac{c_t^j}{D_1^j}$. The aggregated problem can be solved to optimality using the results on the single-product problem from Section 2.5.2. Thus, inventory targets $S_t^{(agg)}(A_t^{(agg)})$ are determined in each period.

Inventory target and production for each product

The aggregated target is allocated to individual products through an iterative procedure consisting of two steps: (1) an estimation of the marginal cost as a function of the inventory for each product, and (2) a maximisation of the minimum marginal cost across all products.

Let $g_t^j(\mathbf{y_t}, \mathbf{A_t})$ be the marginal cost of product j in period t as a function of the end-period inventory and the cumulative forecast update. Since capacity is shared and products are correlated, the marginal cost of a product depends on the inventory and forecast of other products. As such, determining the exact marginal cost functions analytically is a hard problem. To approximate the marginal cost functions, we decompose the problem into independent sub-problems with dedicated capacity K_t^j for each product j. The capacity allocated to a product is set as the difference between the overall capacity and the capacity expected to be used by products with greater value. This is expressed as $K_t^j = K - \sum_{i=1}^{j-1} \mathbb{E}[Q_t^i]$. This capacity allocation strategy is based on the observation that, in the case of capacity shortage, planners prioritise products with higher values

and accept shortages for lower value products. The marginal cost of product j in period t can then be deduced from Proposition 2.3 as

$$g^j_{T-1}(y^j_{T-1}, A^j_{T-1}) = c^j_{T-1} - c^j_T \left[F_{A^j_T|A^j_{T-1}}((S^j_T)^{-1}(y^j_{T-1} + K^j_T)) - F_{A^j_T|A^j_{T-1}}((S^j_T)^{-1}(y^j_{T-1}))\right] - \gamma^j \left[1 - F_{A^j_T|A^j_{T-1}}((S^j_T)^{-1}(y^j_{T-1} + K^j_T))\right]$$

in the period $T-1$ where γ^j is the shortfall cost of product j, and

$$g^j_t(y^j_t, A^j_t) = c^j_t - c^j_{t+1} + \int_{-\infty}^{(S^j_{t+1})^{-1}(y^j_t)} g^j_{t+1}(y^j_t, a) f_{A^j_{t+1}|A^j_t}(a)da + \int_{(S^j_{t+1})^{-1}(y^j_t + K^j_t)}^{+\infty} g^j_{t+1}(y^j_t + K^j_t, a) f_{A^j_{t+1}|A^j_t}(a)da$$

for previous periods. The inventory target functions S^j_t can be deduced recursively from period T to 1 as the solution of $g^j_{t+1}(y\cdot, A^j_t) = 0$ as in the single-product case.
The second part of the iterative procedure calculates a production plan for all products over the prediction horizon. Similarly to Biçer and Seifert (2017), the inventory allocation is determined through a non-linear max-min optimisation problem:

$$\begin{aligned} \max_{\mathbf{y_t}} \quad & \min\left\{g^1_t(y^1_t, A^1_t), \ldots, g^N_t(y^N_t, A^N_t)\right\} \\ \text{s.t.} \quad & \sum_{j=1}^{N} y^j_t = \mathbb{E}[S^{(agg)}_t(A^{(agg)}_t)], \\ & \sum_{j=1}^{N}(y^j_t - x^j_t) \leq K, \\ & y^j_t \geq x^j_t, \qquad \forall j, \\ & y^j_t \leq \mathbb{E}_{\mathbf{A_T}|\mathbf{A_t}}[S^j_T(A^j_T)], \quad \forall j. \end{aligned} \tag{2.4}$$

Problem 2.4 allocates the aggregated target $\mathbb{E}[S^{(agg)}_t(A^{(agg)}_t)]$ to the individual products by prioritising products with smaller marginal costs. Further, it ensures feasible production volumes through a capacity constraint. Although this problem is non-linear, it is convex and can be solved using a derivative-free algorithm, see e.g. COBYLA (Powell, 1994).
As in the single-product case, the inventory targets are uncertain over the planning horizon since they depend on yet unobserved forecast updates. We approximate the expected inventory targets in future periods as $\mathbb{E}[S^{(agg)}_t(A^{(agg)}_t)] \approx S^{(agg)}_t(\mathbb{E}[A^{(agg)}_t])$ and

$\mathbb{E}[S_t^j(A_t^j)] \approx S_t^j(\mathbb{E}[A_t^j])$. Problem 2.4 can thus be solved to obtain a production plan over the horizon. The algorithm proceeds to the next iteration using the newly calculated production plan to derive the capacities available for all products. The iterative procedure stops when the changes between two successive production plans is below a convergence criteria.

Summary

An iterative solution procedure has been presented to approximate the optimal inventory policy and deduce a production plan over the prediction horizon. An advantage of the method is the integration of product correlation when determining the aggregated inventory targets. However, correlation is not considered when allocating inventory to individual products which can lead to sub-optimal allocation especially for settings with a large number of products and a dense correlation matrix. Still, in the next section, numerical simulations show that significant improvements can be obtained compared to the traditional rolling-horizon benchmark that ignores the evolution of forecasts.

2.6. Numerical study

The numerical study is implemented in the Julia programming language (Bezanson et al., 2017), a fast scientific language with a wide environment of modules. In particular, the COBYLA algorithm for solving the non-linear optimisation Problem (2.4) is called from the NLopt module (Johnson, 2014). The simulations are run on an Intel(R) Core(TM) i7-4810MQ processor at 2.80Ghz using 16GB of RAM. The code used to produce all results and figures in this paper is openly available online.
In this section, we discuss some computational aspects of our solution approach, we describe the rolling-horizon simulation setting and assess the value of the forecast evolution models by comparing them to the traditional rolling-horizon benchmark (t-RH) introduced in Section 2.5.1.

2.6.1. Single product

Simulation setting

Simulations are performed in a rolling-horizon fashion. In each planning period, a production plan is calculated and the first period is implemented. The inventory position is updated and the simulation is rolled forward. A new forecast update is then sampled

from the true forecast evolution distribution. In the selling period, the demand and the amount of lost sales are observed. This simulation framework provides a fair evaluation of the benchmark without forecast evolution since it can still benefit from the updated forecasts received in each review period.

The planning horizon is set to $T = 4$ periods with a capacity of $K = 50$ in each period and the fill-rate target is set to $\beta = 0.95$. The production and inventory cost parameters p and h are both set to 1. The shortfall penalty costs are derived proportional to the overall costs in each period as $\gamma = g \cdot c_T$ where we define g as the shortfall penalty factor. The value of the shortfall penalty factor are found numerically, see Appendix A.5.

The performance of the two models is compared using three measures: the actual costs, the achieved fill-rate service level, and the nervousness. The actual cost is measured as the sum of production and inventory costs over the simulation period. The service level is measured at the end of the selling period as the proportion of demand satisfied directly from on-hand inventory. Nervousness is measured in each planning period as the average of the absolute changes over the prediction horizon, as $nv_t = \frac{1}{N(T-t+1)} \sum_{j=1}^{N} \sum_{s=t}^{T} \mid Q_{t,s}^{j} - Q_{t-1,s}^{j} \mid$ and the nervousness of a single simulation run is determined as the nervousness of across all plans as $anv = \frac{1}{T-1} \sum_{t=2}^{T} nv_t$.

Similarly as in Norouzi and Uzsoy (2014), we investigate the impact of the timing of uncertainty resolution on the production policy and model performance. The uncertainty resolution timing describes the periods in which the forecast updates have highest variability, which can be interpreted as periods in which most information is obtained. Three uncertainty resolution settings are considered: early, constant and late. With early uncertainty resolution, the variance of forecast updates is high in the first planning periods and low in the periods close to the selling season. It is the opposite with late uncertainty resolution. With constant uncertainty resolution, the cumulative variance decreases linearly over time, which corresponds to the setting used by Wang et al. (2012). Note that the overall forecast and demand uncertainty over the initial planning horizon is identical in all settings. The standard deviation of the forecast evolution is given in Table 2.1 for additive and multiplicative MMFE. The initial demand forecast is set to $D_1 = 100$.

Table 2.1.: Standard deviation of forecast evolution over horizon of $T = 4$ periods.

Uncertainty	Additive MMFE	Multiplicative MMFE
Early	$\sigma = \begin{bmatrix} 30 & 20 & 10 & 5 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.30 & 0.20 & 0.10 & 0.05 \end{bmatrix}$
Constant	$\sigma = \begin{bmatrix} 18.87 & 18.87 & 18.87 & 18.87 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.1887 & 0.1887 & 0.1887 & 0.1887 \end{bmatrix}$
Late	$\sigma = \begin{bmatrix} 5 & 10 & 20 & 30 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.05 & 0.10 & 0.20 & 0.30 \end{bmatrix}$

Inventory target functions under additive and multiplicative MMFE

In Section 2.5, we have identified the optimal production policy under forecast evolution assuming that the inventory target function in the last period $S_T(\cdot)$ is strictly increasing and invertible. We have also shown that these assumptions hold for multiplicative MMFE, but not for additive MMFE since the target function is convex. To allow the calculation of the production policy in the additive case, we modify the inventory target function in the last period as

$$\hat{S_T}(A_T) = \begin{cases} S_T(A_T), & \text{if } A_T \geq \underline{A_T} \\ S_T(\underline{A_T}) - m(\underline{A_T} - A_T), & \text{otherwise} \end{cases} \tag{2.5}$$

where the slope m is calculated as $m = \frac{S_T(1.5 \cdot D_1) - S_T(0.5 \cdot D_1)}{D_1}$. This allows to calculate the production policy under additive MMFE as described in Section 2.5.2.
The optimal inventory targets are presented in Figure 2.3 as a function of the cumulative update for additive and multiplicative MMFE under constant uncertainty resolution. The extended part of $S_4(A_4)$ obtained from Equation (2.5) is shown as a dotted line. The figure suggests that the last-period inventory targets increase linearly and exponentially with the cumulative forecast updates under additive and multiplicative forecast evolution respectively.

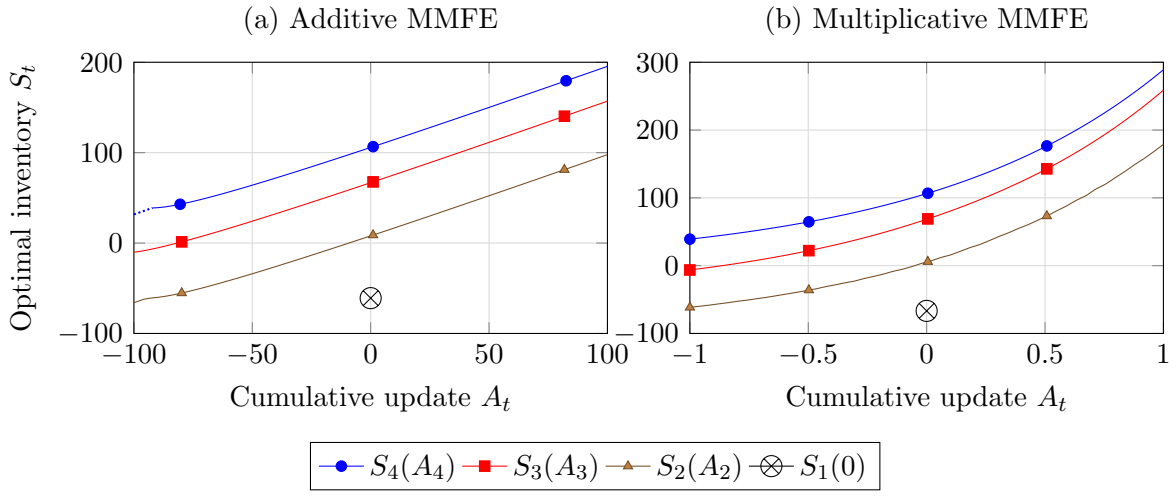

Figure 2.3.: Inventory target at the end of each planning period as a function of the cumulative update for $\gamma = 5$.

Comparison to benchmark

For each uncertainty resolution setting, a total of 1000 rolling-horizon simulations are run. Each set of 1000 runs can be computed in around 60 seconds. The average model performance are given in Table 2.2. If the average achieved service level is lower than the target, the statistical significance of the service-level shortfall is assessed through Student's t-test. The statistical significance of the difference between the cost and nervousness of the t-RH and MMFE models is assessed using Student's t-test. Statistical significance at a p-value smaller than 0.05 is indicated with a star symbol (*) in the table.

Table 2.2.: Average of performance results for the single-product case.

		Early		Constant		Late	
		t-RH	MMFE	t-RH	MMFE	t-RH	MMFE
Additive MMFE	Service level	0.9200*	0.9483	0.9388*	0.9486	0.9542	0.9447
	Objective	154.88	206.91	185.636	200.309	215.796	198.216
	(relative in %)		(+33.59*)		(+7.90*)		(−8.15*)
	Nervousness	6.777	7.413	6.445	7.197	4.153	3.309
	(relative in %)		(+13.32*)		(+9.38*)		(−20.31*)
Multiplic. MMFE	Service level	0.9143*	0.9480	0.9362*	0.9440	0.9560	0.9443
	Objective	151.22	248.35	185.63	197.81	230.48	198.81
	(relative in %)		(+64.22*)		(+6.56*)		(−13.74*)
	Nervousness	7.639	8.361	8.157	8.267	6.393	4.395
	(relative in %)		(+9.46*)		(+1.35*)		(−31.25*)

The simulations results show that, despite accounting for aggregated forecast and demand uncertainty, the traditional rolling-horizon benchmark fails to satisfy the service-level targets in the early and constant uncertainty resolution settings. As uncertainty resolution is shifted to later periods, the assumptions of the t-RH model are more accurate, which explains the increase in achieved service level. The production decisions of the two planning models also become more similar.

The forecast evolution model reaches the target service level in all instances. For early and constant uncertainty resolution, the MMFE model builds the minimum safety stock amount in early periods to manage forecast uncertainty, thus increasing operational costs but ensuring that the target service level is reached. Table 2.2 also shows that nervousness increases as uncertainty resolution is shifted to early periods, especially for the MMFE model, suggesting that early resolution implies more planning efforts. When

uncertainty is resolved late, the MMFE provides cost reductions while reaching the service level target. This is because the t-RH model overestimates the safety stock needed in the last period. The results are similar for additive and multiplicative MMFE. The value of forecast evolution is overall higher for multiplicative MMFE.

Planning strategy

To identify the planning strategy of the MMFE model, the production plans determined in the first period by both models are presented in Figure 2.4 for the three uncertainty resolution settings under additive MMFE. Since the traditional rolling-horizon benchmark does not explicitly consider forecast evolution, its production plan is independent of the uncertainty resolution setting. The figure shows the impact of the uncertainty resolution timing on planning and the importance of early production and capacity reservation when uncertainty is resolved early.

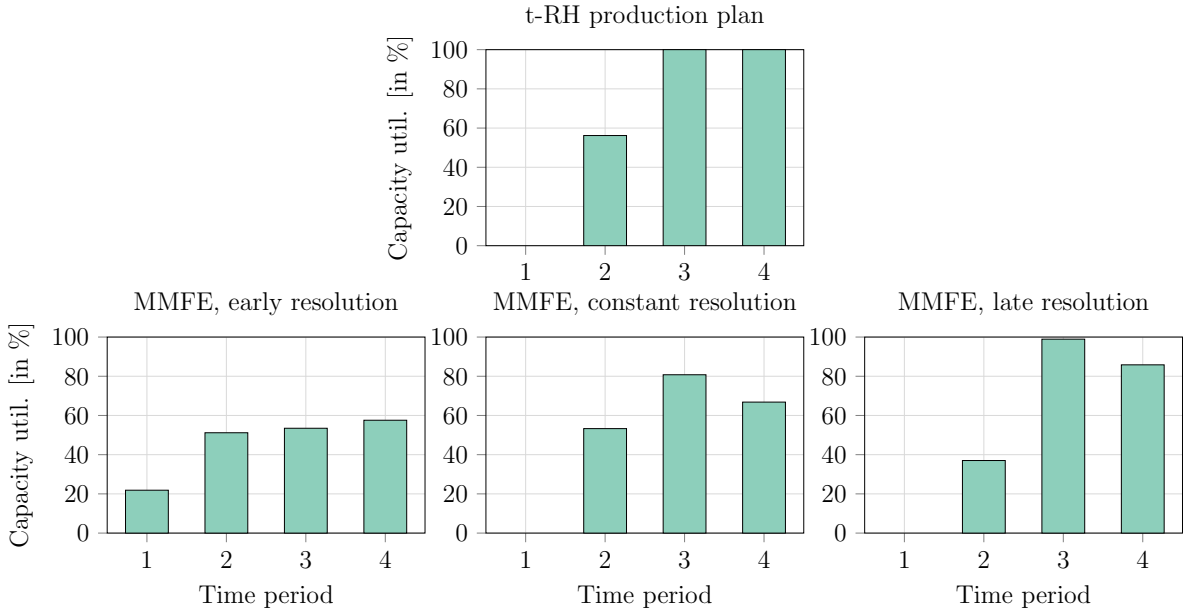

Figure 2.4.: Production plan of t-RH and MMFE models determined in the first planning period.

Another interesting analysis is to relate the nervousness of the production plan to the nervousness of the forecast. In Figure 2.5, we present the nervousness of the plan as a function of the nervousness of the forecast under additive MMFE. Each point is the average nervousness *anv* over a single simulation run. The figure illustrates that integrating forecast evolution provides a stronger link between forecast nervousness and

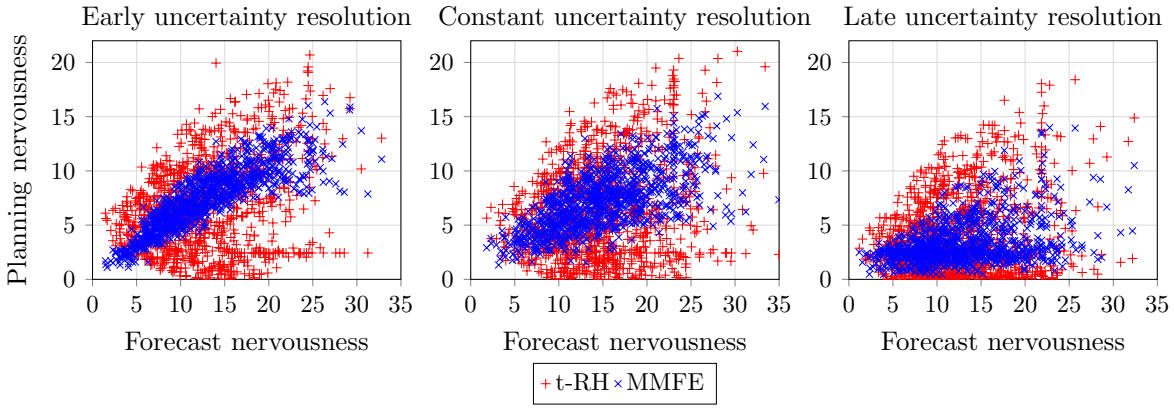

Figure 2.5.: Planning nervousness as a function of forecast nervousness for each simulation run.

production nervousness for all uncertainty resolution scenarios. This implies that the plan provided by the MMFE model would change only when the forecast changes, contrary to the t-RH model for which planning nervousness is erratic. Integrating forecast evolution in decision-making thus improves planning predictability as the planner can estimate planning nervousness when observing the demand forecast.

Summary

In single product environments, integrating forecast evolution in production planning appears beneficial in all settings. The MMFE model satisfies the expected service level constraint in all instances and can achieve significant cost savings when uncertainty is resolved late. The numerical study highlights the value of integrating forecast evolution as well as the importance of explicitly taking into account the timing of uncertainty resolution. One important recommendation to planners dealing with seasonal demand is to analyse the timing of forecast updates and to prepare accordingly: early uncertainty suggests building pre-emptive inventory and reserve capacity buffers while late uncertainty resolution suggests a more wait-and-see strategy with a high utilisation in later periods. Further, we have showed that integrating forecast evolution strengthens the relation between forecast nervousness and planning nervousness, so that the planner can anticipate planning nervousness when observing the updated forecast.

2.6.2. Multiple products

Simulation setting

We extend the previous simulation setting to include $N = 2$ products whose demand and forecast updates may be correlated. The capacity level is set to $K = 80$ in each period. The cost parameter p and h are both set to 2 for the first product and 1 for the second product. Considering that inventory costs are calculated proportional to the product value, product 1 can be seen as twice as valuable as product 2. The fill-rate service-level target is set to $\beta^j = 0.95$ for each product. The value of the shortfall costs are given in Appendix A.5. The standard deviation of the forecast evolution process is set equal for both products. We use the same uncertainty resolution settings as for the single-product case. To analyse the impact of forecast evolution correlation between the products, we consider three cases: $\rho = 0$, $\rho = 0.6$ and $\rho = -0.6$. The correlation coefficient is kept constant throughout the periods. The initial forecast of both products is $D_1 = 100$.

Convergence of heuristic

In our simulations, we observe that the iterative procedure of the multi-product converges in a few iterations when using the t-RH production plan as an initial guess to determine the product-specific capacities. In fact, the procedure converges in a single iteration for two uncorrelated products under constant uncertainty resolution. This is shown in Figure 2.6, where the relative change between two successive plans is calculated as $rc_i = \frac{\sum_t \sum_j |(Q_t^j)^i - (Q_t^j)^{i-1}|}{\sum_t \sum_j (Q_t^j)^i}$ where $(Q_t^j)^i$ is the production planned in iteration i. In the following, we stop the iterative procedure if the relative change is such that $rc_i < 1e^{-4}$.

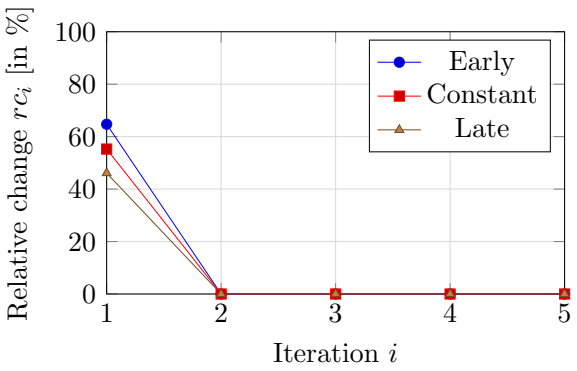

Figure 2.6.: Convergence of iterative procedure for two products with non-correlated forecast evolution.

Planning strategy

The impact of product correlation and uncertainty resolution on planning is analysed in Figure 2.7 which shows the production plan determined in the first planning period for both models. The production plan of the t-RH model is identical for all combinations since it ignores both forecast evolution and product correlation. The MMFE production plans are then presented where each column corresponds to an uncertainty resolution type and each row to a product correlation. The MMFE model exhibits common patterns over the different settings. For instance, product 2, the lowest priority product, is always scheduled first since it has the lowest inventory costs, and the last period is almost always dedicated to product 1, the most valuable product. The timing and volume of capacity reservation is strongly impacted by the product correlation and uncertainty resolution. Capacity buffers are reserved in later periods when there is higher uncertainty resolution in early periods. As product correlation increases, the model increases early production and reserves more capacity. The same pattern is observed when uncertainty resolution is shifted to earlier periods. Thus, forecast evolution have highest impact on planning when both conditions are observed: uncertainty resolution is non-constant and products are correlated, either positively or negatively.

Simulation results

For each combination of product correlation and uncertainty resolution, 1000 simulations are run in a rolling-horizon fashion. The computation time is around 600 seconds for each set of 1000 runs. The majority of the computation time is spent solving the non-linear optimisation Problem (2.4). The average of the performance indicators is provided in Table 2.3. A statistical significance analysis is performed similarly as in the numerical study of the single-product case.

The simulations results show that the forecast evolution model outperforms the traditional rolling-horizon benchmark on all simulation instances. As in the single-product case, the MMFE model is able to reach the service-level target in all instances. The benchmark fails to reach the service level target in all but one instance, when uncertainty is resolved late and forecast updates are negatively correlated. The instance with positive product correlation and early uncertainty resolution sees the highest shortfall violation of the service level targets by the model without forecast evolution. Correspondingly, it is the setting in which the forecast evolution model yields the highest cost increase. The forecast evolution model can decrease costs in three out of nine instances.

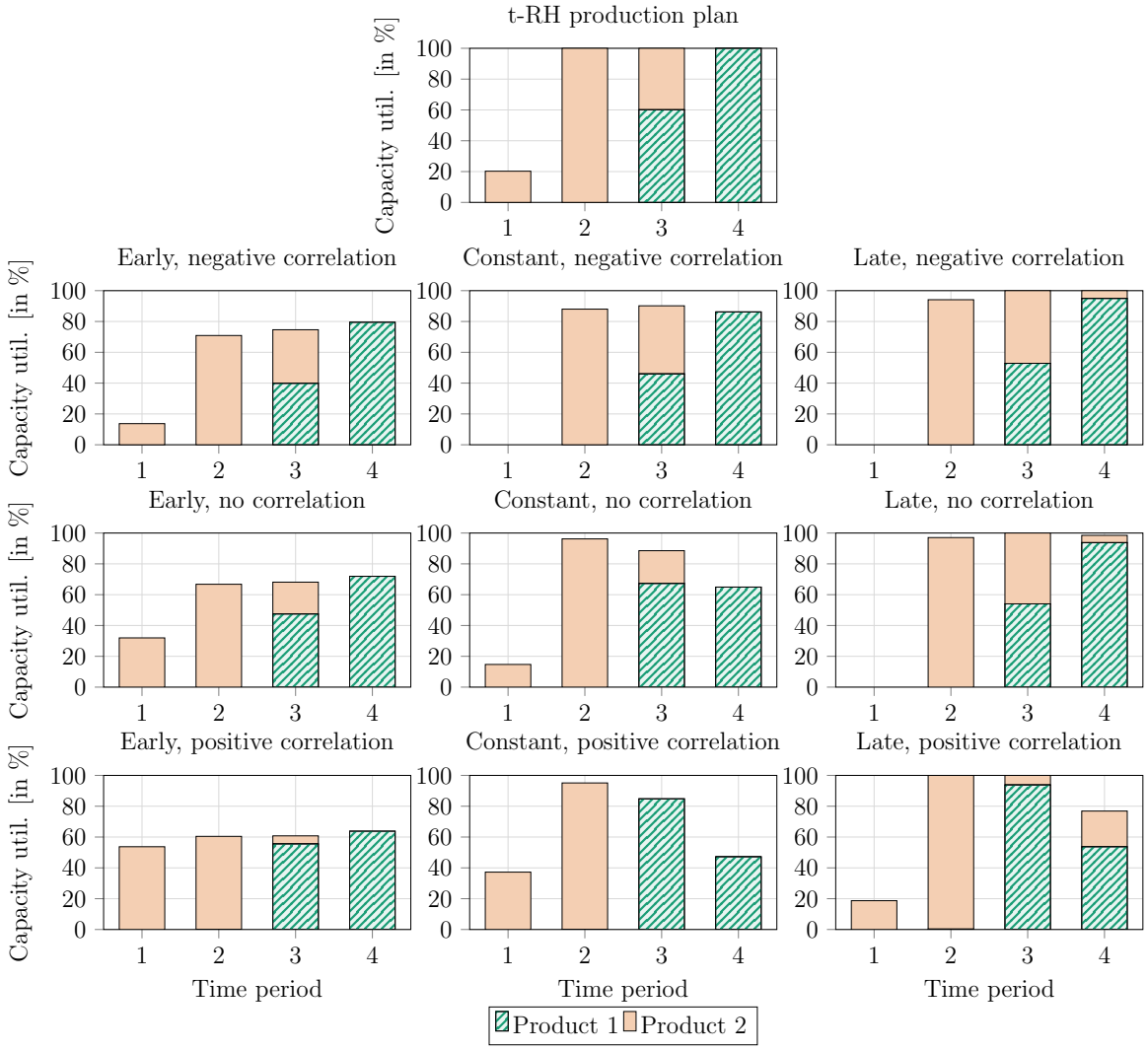

Figure 2.7.: Production plan in first period varying for different uncertainty resolution timing and product correlation.

The highest cost reduction is obtained for late uncertainty resolution and negative correlation. It is also interesting to observe that the achieved costs of the forecast evolution model vary between the simulation setting. The highest achieved costs are obtained for late uncertainty resolution and positive correlation, suggesting that this is the most challenging instance.

2.6.3. Managerial recommendations

Considering aggregated demand uncertainty is not sufficient to reach a service-level target in almost all instances. Hence, it is essential to explicitly integrate forecast evolution

Table 2.3.: Average values of simulation results for different product correlation and uncertainty resolution timing.

Product correlation		Early		Constant		Late	
		t-RH	MMFE	t-RH	MMFE	t-RH	MMFE
$\rho = -0.6$	SL prod 1	0.9295*	0.9487	0.9423*	0.9477	0.9607	0.9555
	SL prod 2	0.9700	0.9617	0.9779	0.9576	0.9721	0.9506
	Objective	507.25	526.18	600.95	554.53	684.91	612.09
	(relative in %)		+3.73*		−7.72*		−10.63*
	Nervousness	19.23	17.25	13.89	15.74	7.84	13.53
	(relative in %)		−10.30*		+13.30*		+72.49*
$\rho = 0$	SL prod 1	0.9008*	0.9479	0.9231*	0.9591	0.9395*	0.9488
	SL prod 2	0.9700	0.9551	0.9776	0.9510	0.9785	0.9519
	Objective	496.46	560.75	586.89	609.50	671.47	614.48
	(relative in %)		+12.94*		+3.85*		−8.48*
	Nervousness	19.30	17.14	14.08	18.57	8.08	11.89
	(relative in %)		−11.22*		+31.89*		+47.03*
$\rho = 0.6$	SL prod 1	0.8750*	0.9474	0.8989*	0.9620	0.9327*	0.9672
	SL prod 2	0.9701	0.9548	0.9772	0.9504	0.9787	0.9505
	Objective	477.35	601.90	573.18	673.41	663.12	696.85
	(relative in %)		+26.09*		+17.48*		+5.08*
	Nervousness	19.32	15.73	14.47	17.56	8.33	32.32
	(relative in %)		−18.57*		+21.41*		+287.81*

in planning. When forecasts are periodically updated, the MMFE model suggests that high demand satisfaction can be achieved only through early inventory. Late uncertainty resolution and negative correlation is a case with highest potential since considering forecast evolution can reduce costs. The simulation results highlight the importance for companies to analyse the timing of uncertainty resolution and understand its impact on planning. Since actual costs are highest when uncertainty resolves late, it is clearly beneficial to strive for more information gain in the early periods, for instance, by investing in forecasting or by increasing communication through the supply chain. Further, negative correlation between product demands and forecast updates suggest that companies should group products with negative demand correlation and substitution effects together on the same production lines to benefit most from forecast updates.

2.7. Conclusion

In this article, we have investigated the situation of a planner facing an uncertain seasonal demand with forecast evolution when production is limited by capacity and inventory is costly. Both single- and multiple-product cases have been considered. For the single-product case, we have identified the optimal production policy to satisfy an exogenous fill-rate service-level target. For the multi-product setting, we have proposed an iterative heuristic derived from our understanding of the single-product problem. The performance of the heuristic has been demonstrated in an extensive rolling-horizon simulation study. In all instances, the MMFE model was able to provide either higher demand satisfaction or lower operational costs. Our study extends the existing literature on stochastic planning by highlighting that managing demand uncertainty alone does not immunise against shortages and that it is necessary to implement a more detailed forecast evolution model. Further, we analysed the planning strategy of the forecast evolution model to provide guidelines to planners facing the typical trade-off between building early inventory or waiting for a more precise forecast with limited capacity.
A potential research direction would be to consider that not only the volume but also the timing of the season is uncertain. This is often the case for agricultural goods, for instance, since the selling season depends on unpredictable parameters such as weather conditions. In this context, forecasts on both the selling season volume and timing would be periodically updated.

Chapter 3

Stochastic programming in master production scheduling: overcoming barriers to industry application

Abstract

This paper stems from the observation that companies still rely on deterministic rolling-horizon planning despite a substantial body of literature on stochastic planning models. To foster practical applications, we identify barriers that limit the widespread application of stochastic programming in master production scheduling and develop a framework to overcome them. Our solutions include modelling uncertainty from available data, reflecting planning processes in the optimisation model and evaluating its performance accurately. A two-stage stochastic model with production recourse is introduced to improve planning flexibility, stability and communicability. It is applied on a real-world case study with large product portfolio, complex production processes and uncertain seasonal demand. Out-of-sample rolling-horizon simulations show that well-defined stochastic models can provide high demand satisfaction and low inventory costs while improving planning stability. In particular, planning nervousness can be reduced by 40% and raw-material nervousness by 80% compared to our industry partner's current production scheduling solution.

3.1. Introduction

Even when demand is highly uncertain, companies still rely on deterministic rolling-horizon planning and rule-of-thumbs for safety-stock calculations in master production

scheduling (Meistering and Stadtler, 2017). Yet, recent applications of stochastic programming have shown impressive results in controlled simulation environments (Gruson et al., 2021; Thevenin et al., 2021). Stochastic programming can accurately determine the volume and timing of safety stocks based on probabilistic uncertainty models. Further, they reflect the flexibility of rolling-horizon planning through recourse decisions that adapt to uncertainty as it unfolds. When demand is dynamic and forecasts have poor accuracy, planning flexibility is critical to ensure that demand can be met. Nonetheless, stochastic programming is far from widely applied in practice despite promising complementarity with rolling-horizon planning . This observation is especially surprising considering the breadth of existing research on stochastic programming. It suggests that existing models still contain important shortcomings that prevent their application. In particular, there appears to be a lack of discussion on how to translate models from academic settings, that rely on simplifying assumptions, to real-world problems and their complexity.

In this paper, we study how to overcome barriers facing practitioners when applying stochastic programming to master production scheduling. First, we identify barriers that still prevent the application of stochastic programming. We distinguish barriers relating to the identification and representation of uncertainty, relating to the development of stochastic planning models that fit existing planning structures, and relating to the computational challenges of evaluating model performance. Second, we propose a decision framework to set up stochastic models in master production scheduling. We present novel strategies to overcome the barriers such as aggregating products into optimal families, which increases planning stability and allows flexible production recourse. A two-stage stochastic model is developed to integrate the above strategies and determine a master production schedule that provides high demand satisfaction for low inventory costs and ensures planning stability on both the production and raw-material levels. We demonstrate our approach on a real-world case study in the agrochemical industry and evaluate its performance through out-of-sample rolling-horizon simulations.

The remainder of this paper is organised as follows. In Section 3.2, barriers limiting the application of stochastic planning models are presented and related to existing literature. In Section 3.3, the real-world case study is introduced. We discuss the specific form of the barriers identified in the previous section and we provide an overview of our strategies to overcome them. In Section 3.4, uncertainty models are derived from available historical data and used to construct scenario trees. In Section 3.5, stochastic planning models are developed to improve planning flexibility, stability and communicability. In

Section 3.6, sensitivity analyses are conducted to tune the model and a comparison with industry benchmarks is presented. In Section 3.7, we summarise our work and propose directions for future research.

3.2. The barriers of applying stochastic programming in master production scheduling

This section details barriers that prevent the widespread application of stochastic planning models. We categorise the barriers in three groups relating to modelling uncertainty, the planning environment and the numerical challenges.

3.2.1. Modelling uncertainty from data

In practice, probability distributions are not available to describe demand uncertainty. Instead, uncertainty models have to be constructed from available data.

Barrier 1 (Data scarcity). *Data is essential when setting up stochastic models, yet it is especially scarce in master production scheduling.*

Master production scheduling derives tactical decisions typically aggregated on a monthly granularity with planning horizons between 6 months and 2 years. Data sets cover only several years of historical data at most, hence only a limited number of observations are available. Further, the relevance of older data is limited by product life cycles and changes in market conditions (Chopra and Meindl, 2013). If demand is dynamic, for instance if there is a yearly seasonality, only few observations of the entire demand process are available in the data set. Yet, data is essential to measure the uncertainty of the planning environment and to evaluate model performance in simulations. Data scarcity is an ever-present problem for planners who require quantitative methods to support their decisions. This limits the application of sophisticated planning techniques. In particular, recent developments in data-driven operations research (Mišić and Perakis, 2020) may not be applicable.

Barrier 2 (Uncertainty definition). *Identifying the nature, number and stationarity or lack thereof of uncertain processes influencing demand is critical.*

There are two main methods to characterise demand uncertainty. Most common is the assumption that demand itself can be modelled as an uncertain process. This method relies on past demand observations to predict future demand. It may be especially

relevant when there is (a) a stationary demand process, (b) seasonality, or (c) if demand follows a type of auto-regressive process (Klabjan et al., 2013). For instance, Ban (2020) considers seasonal goods whose demand realises over a long season. Demand periods within the season are assumed correlated but observations of the full season are assumed independent identically distributed. Li and Disney (2017) model demand as a simple first-order auto-regressive process. In effect, the first approach assumes that demand is a seasonal but stationary process, whereas the second approach assumes that demand evolves over time but is locally stationary. The second method to characterise demand uncertainty focuses on the error caused by inaccurate forecasts. The key stochastic process is then the forecast error, which can be modelled as a probability distribution (Prak et al., 2017; Trapero et al., 2019).

Both approaches ultimately provide uncertainty models for the demand over the planning horizon. However, the resulting uncertainty models vary drastically when measuring either demand or forecast uncertainty from data. Using wrong assumptions may lead to severely inaccurate models with long-lasting consequence. This first step is fundamental when applying stochastic models from data but is often overlooked in the literature. Practical guidelines describing methods to identify and measure uncertainty from limited data are still missing.

Barrier 3 (Uncertainty model). *It is not clear when to use past data directly and when to estimate distributions, which is challenging with scarce data.*

Once the uncertain processes are defined and samples have been measured from historical data, the question arises of whether to use these samples directly in a data-driven fashion (Kleywegt et al., 2002) or to assume that they are observations of an underlying probability distribution. Creating scenario trees directly from data allows to capture correlation between products and period while avoiding distribution assumptions. However, it may fail to generalise from the data set and lead to overfitting.

In the literature, demand is commonly assumed to follow a known distribution, often normal, which is fitted to the data (Silver et al., 2016). These distributions can be sampled to create scenario trees over the horizon (Heitsch and Römisch, 2009; Homem-de-Mello and Bayraksan, 2014). Still, there is no guarantee that demand follows a probability distribution. Further, distribution parameters cannot be estimated precisely from scarce data and are subject to estimation error (Prak and Teunter, 2019). In a multi-dimensional setting, estimation error plays an even larger role since the number of observations may be much smaller than the number of parameters to estimate. An alternative to estimating probability distribution is to use distribution-free methods such

as robust optimisation (Bertsimas et al., 2018a) or distributionally robust optimisation (Ben-Tal et al., 2013; Wiesemann et al., 2014). Yet, these methods have been applied to problems for which hundreds of observations are available and may not be suitable to problems with scarce data.
Barrier 2 and 3 are closely related but focus on different problems. Barrier 2 describes the challenges in identifying the source of uncertainty and obtaining relevant samples from past data whereas Barrier 3 discusses the different methods to process the samples.

Contributions. The strategies to overcome the barriers are typically problem specific. We propose two approaches to measure uncertainty from limited data based on seasonal demand uncertainty and forecast error respectively. For both uncertainty definitions, we compare the use of empirical and estimated probability distributions. The models are evaluated in simulations using real-world data. We show that accurately defining uncertainty is critical to ensure high demand satisfaction. In fact, deterministic models with simple rule-of-thumbs for safety stock but accurate uncertainty definition outperform stochastic models with wrong uncertainty definition. Thus, Barrier 2 is found more critical for performance than Barrier 3, which is remarkable considering that existing literature mostly focuses on the latter barrier at the expense of the former.

3.2.2. Reflecting the planning process

Stochastic models can reduce costs through recourse decisions that adapt to uncertainty as it unfolds. However, they also need to respect the constraints of the planning processes. The interaction of recourse models with planning flexibility, communicability and stability remains understudied.

Barrier 4 (Flexibility representation). *Stochastic programming models must be designed to properly represent the planning flexibility resulting from rolling-horizon planning processes.*

Since scenario-based stochastic programming can introduce recourse variables that adapt to the uncertain process as it unfolds (King and Wallace, 2012), it can capture the flexibility of rolling-horizon planning. Flexibility in production planning has been studied in early works by Escudero et al. (1993) and Brandimarte (2006) who compare different recourse structures in lot-sizing problems. Recently, Tavaghof-Gigloo and Minner (2020) propose a heuristic to integrate re-planning opportunities in a single-stage stochastic model by reducing safety stock levels when capacity is unlimited. Yet, recourse deci-

sions should also improve costs when capacity is tight thanks lower safety stocks and better prioritisation of products over the horizon. Hence, how to best match the flexibility of stochastic programming (i.e. the definition of stages and recourse decisions) to the flexibility of the production environment is also an open question.
Quantifying the value of recourse in rolling-horizon planning has only be done partially. Existing works conduct static comparison of two-stage and multi-stage formulations in problems such as production planning with demand and yield uncertainty (Kazemi Zanjani et al., 2010), and lot-sizing and scheduling (Hu and Hu, 2018). Static evaluations ignore the rolling-horizon implementation of planning models, which provides flexible re-planning opportunities to stochastic models without recourse even if they are not explicitly modelled. A notable exception has been proposed by Stephan et al. (2010), who accurately measure the value of multi-stage models in capacity planning problems by using a rolling two-stage benchmark. Hence, practitioners cannot estimate the value of applying recourse models in rolling horizon.

Barrier 5 (Communicability). *Scenario-independent reference plans need to be communicated to upstream and downstream members of the supply chain.*

Recourse models typically ignore the communicability requirement of rolling-horizon planning, which is essential throughout the supply chain. Contrary to deterministic or stochastic models without recourse, there is no unique plan obtained when solving a model with recourse. Instead, a tree of decisions is derived over the planning horizon that merely represents what-if statements. However, unconditional production plans need to be communicated to downstream parts of the supply chain to coordinate production schedules as well as the distribution and sale of finished goods.
In the same vein, raw-material orders are communicated to upstream parts of the supply chain to coordinate production and purchasing activities. Considerations of raw-material ordering and availability in production planning problems are rare and seem restricted to settings in which raw materials exhibit specific properties. For instance, Cunha et al. (2018) determine raw-material purchases with quantity-based discounts. Bollapragada et al. (2015) investigate the stochastic optimisation of procurement and production decisions in a make-to-order environment with supply uncertainty. More generally, Kanyalkar and Adil (2010) develop a two-stage stochastic model for the procurement, production and distribution including raw materials but consider a simple product structure with a single raw material. New formulations are thus needed to ensure communicability of a reference plan while allowing the flexibility of stochastic models with recourse.

Further, scenario-based multi-stage solutions are typically not implementable in practice unless the true uncertainty distribution is discrete and completely captured in the scenario tree. Thevenin et al. (2021) investigate this issue by proposing several methods to determine a production policy from scenario-based multi-stage solutions. Yet, it is not discussed how to translate the obtained policy into a reference plan that provides long-term visibility.

Barrier 6 (Plan stability). *Reference plans should be stable in rolling horizon with only limited changes between successive review periods, which may restrict the flexibility of recourse decisions.*

Significant plan changes create nervousness, which hinders supply chain performance, leads to loss of confidence, confusion through the supply chain and ultimately higher costs (Atadeniz and Sridharan, 2020). Seminal works analyse the nervousness resulting from lot-sizing heuristics in single-level (Carlson et al., 1979; Sridharan et al., 1988) or multi-level environments (Blackburn et al., 1986; Ho, 1989; Zhao et al., 2001). They develop strategies to mitigate nervousness such as freezing periods or penalising plan changes. Recent research studies the nervousness resulting from optimal planning models. Lin and Uzsoy (2016) compare chance-constraint formulations to capture demand uncertainty and their impact on planning stability. Herrera et al. (2016) integrate different nervousness penalty costs in the objective function to identify a balance between stability and operational costs. Meistering and Stadtler (2017) propose a stabilised-cycle strategy that allows changes in production decisions only when necessary to reach the target service level

Existing nervousness mitigation strategies are based on restricting planning flexibility, which may reduce planning performance when short-term uncertainty is high. While stochastic models should derive optimal production volumes despite the limited flexibility, it is not clear how they would perform when distributions are not known but modelled from data. Further, since freezing periods inherently prohibit recourse opportunities, the trade-off between traditional nervousness reduction methods and stochastic programming with recourse remains open.

Contributions. We note that existing stochastic models with recourse do not evaluate the resulting nervousness, since reference plans are not determined in existing stochastic programming models. By providing reference plans when solving stochastic models with recourse, we can bridge the gap between research on planning stability and stochastic programming.

We contribute to existing literature in several ways. First, we develop a two-stage model that provides recourse and reference plans based on aggregating products into optimal families. Second, we measure the value of recourse in rolling-horizon planning with real-world data. In particular, we show that recourse is especially beneficial when capacity is limited. Finally, we compare the use of traditional nervousness mitigation strategies based on frozen decisions and our novel approach based on product aggregation. We show that freezing decisions on the raw-material level does not limit planning flexibility while providing significant stability improvements. On the other hand, the aggregation-based strategy can improve planning flexibility, communicability and stability, thus outperforming the traditional strategy of freezing production decisions.

3.2.3. Computational challenges

The evaluation of stochastic models is challenging due to several factors including long computation times, the need for complex simulation settings, and the strong dependence of results on the assumptions used in simulations.

Barrier 7 (Tractability). *Stochastic models often exhibit a trade-off between accuracy and long computation times.*

Stochastic programming approaches, and especially multi-stage formulations, lead to notoriously long computation times. Significant attention has been given to designing scenario trees with optimal size. In particular several methods have been developed to reduce the size of scenario trees while retaining their accuracy (Dupačová et al., 2003; Heitsch and Römisch, 2003). Other approaches to improve computation times include decomposition techniques such as progressive hedging (Watson and Woodruff, 2011) and stochastic dual dynamic programming (Shapiro, 2011). Yet, solving times depend not only on the scenario tree but also on the recourse structure. The trade-off between computation times and flexibility offered by recourse also needs to be analysed.

Barrier 8 (Evaluation). *The performance of stochastic models should be evaluated accurately despite limited available data.*

A reliable assessment of expected performance is essential to foster the adoption of new models. This reliability can be achieved by simulating the model in a setting close to its practical use. Simulations can be implemented in a rolling-horizon fashion to respect the planning structure and performed in an out-of-sample fashion to accurately evaluate the uncertainty model. To the best of the authors' knowledge, out-of-sample evaluations have not been applied in production planning to evaluate stochastic models

based on real-world data. Out-of-sample evaluations have been more commonly applied to inventory management and in particular to newsvendor problems (Bertsimas et al., 2018b; Beutel and Minner, 2012; Huber et al., 2019; Oroojlooyjadid et al., 2020). When data is scarce and is used for both model calibration and evaluation, carefully designing the simulation experiments is crucial.

Contributions. We study the trade-off between model accuracy and tractability by varying the scenario size as well as the recourse structure. In both cases, we show that efficient trade-offs can be found. To tune and evaluate the models, we propose the first out-of-sample rolling-horizon evaluation of stochastic production planning models with real-world data. We highlight the importance of out-of-sample evaluation by measuring the bias of in-sample evaluations.

3.3. Real-world case study

In this section, we introduce the industry problem and show the relevance of the barriers identified above. While barriers may be common to many production planning problems, we believe that solution approaches are inherently problem specific. We discuss the form of the barriers in the case study and provide an overview of the strategies to overcome them.

3.3.1. Problem setting

Our industry partner is a world-leading agrochemical company managing a global supply chain with a large product portfolio, long production lead times and complex planning problems. We focus on the production of a restricted product portfolio of pesticides that embodies the planning challenges of the firm. Since the use of pesticides follows the crops' growth cycle, demand patterns exhibit strong seasonality, and accurately forecasting demand is limited by unpredictable parameters such as weather conditions.
The production of synthetic pesticides contains two main steps: the active ingredient synthesis, in which the molecules forming the base of the finished products are synthesised, and the formulation step, in which one or several active ingredients are combined and diluted. The active ingredient synthesis is the most complex process with important capital investment, long lead times and low flexibility. At this level, production is conducted in long campaigns that realise over several months to a year. Short-term changes

to campaigns are limited by cleaning operations that can last up to several weeks. As the most value-adding process, the active ingredient synthesis highlights the inherent challenge of agrochemical supply-chain management: production has low flexibility and long lead times whereas demand is dynamic and hard to predict even in the short future. Production and supply planners are thus looking for advanced strategies to manage demand uncertainty and to ensure efficient operations throughout the supply chain.
Because of the complexity of the global network, the active ingredient synthesis and formulation are planned sequentially. Formulation planners derive the intended production over the planning horizon and deduce the active ingredient requirements that are communicated to upstream planners. The aim of our industry collaboration is to improve the formulation planning step to derive plans that satisfy the uncertain demand while ensuring that stable raw-material orders are provided to upstream planners. In effect, this improved formulation planning would act as a dampening step, reducing the uncertainty of the demand forecast as it propagates through the supply chain.

3.3.2. Overcoming the barriers

From the identification of the uncertain processes to the model development and evaluation, this industry problem encompasses the barriers of stochastic programming described in Section 3.2. We discuss the specific forms taken by the barriers in this industry case and present an overview of our strategies to overcome them.

Uncertainty. Historical forecasts and past demands are available for the last four years. Because of the seasonality of demand, this data set corresponds to only few observation of the entire demand process. Defining the uncertain processes from this limited data set is challenging since demand is dynamic and forecasts are inaccurate. To overcome Barrier 2 (Uncertainty definition), we derive seasonal models of uncertainty. Both demand-driven and forecast-driven are analysed based on the uncertainty of demand and forecasts respectively. The two approaches provide different samples for the empirical demand distributions, which can be either used directly or to estimate probability distributions. To overcome Barrier 3 (Uncertainty models), we implement both approaches, estimating normal and uniform distributions from the empirical samples. Scenarios trees are created and integrated in two-stage stochastic models.

Planning processes. The supply chain and production processes of the industry case are complex. In particular, the active ingredients have long production lead times and

are especially sensitive to planning nervousness. Yet, flexibility is essential to ensure that demand can be met despite poor forecasts accuracy. We overcome the barriers linked to planning flexibility, stability and communicability in several way. To overcome Barrier 5 (Communicability), we ensure that a reference plan is always available on both the production plan and raw-material levels. Raw-material orders and inventory are explicitly modelled. Long-term visibility is essential for raw-material planning. However, a detailed production plan is only required by downstream planners to determine the schedule of formulation campaigns. Hence, we can aggregate communications on the production plan level by defining product families. The definition of the families is a key part of our approach. To ensure that aggregated plans provide the information necessary to derive production schedules, families are defined through a multi-objective optimisation models with custom rewards and constraints that reflect production processes. Product families allow to overcome Barrier 4 (Flexibility representation) by introducing production recourse. First-stage capacity reserves are placed on the family level, which can be used flexibly by products within in the family through recourse decisions. We observe that plan changes within product families tend to compensate in rolling horizon so that aggregating decisions on the family level also improves planning stability. Thus, we compare the nervousness mitigation techniques of freezing and aggregating decisions to overcome Barrier 6 (Plan stability). The different planning strategies are integrated in a mixed-integer linear problems that optimally determines the share of first-stage and recourse production decisions.

Numerical study. The models are evaluated through rolling-horizon simulations and extensive sensitivity analyses are performed. We overcome Barrier 7 (Tractability) by studying the effect of the size of the scenario trees and the number of product families that both increase the number of recourse variables. In both cases, efficient trade-offs can be found between solution quality and model complexity. The simulation setting is crucial to overcome Barrier 8 (Evaluation). The out-of-samples rolling-horizon simulation framework proposed is especially powerful to make generalisable conclusions from the limited amount of data and avoid in-sample bias. To finalise the model evaluation, meaningful benchmarks are defined from historical company data, providing an accurate assessment of expected improvements compared to current practice. Barrier 1 (Data scarcity) is the most fundamental and challenging barrier to overcome. It underlies all strategies applied in this paper, and is only overcome at the end of the numerical study, once we finally identify the best model configuration and prove its benefits experimen-

tally.

3.4. Modelling the uncertain process

In this section, we discuss the definition and representation of the uncertain process. We present two uncertainty models based on demand uncertainty and forecast error respectively. Scenarios are obtained for both models and used to construct two-stage scenario trees.

3.4.1. Seasonal demand uncertainty

A data set is available covering Y seasons of S periods each. Historical demand of the portfolio of K products has been observed where $d_k^{s,y}$ is the demand for product k observed in period s of season y. To reflect seasonality, the first uncertainty model assumes that demand follows a stationary distribution in each period of the season and that demand periods within the season may be correlated. The planner can either use the empirical distribution $\mathcal{D}_s = \{d_k^{s,y}, y \in \{1, \dots, Y\}\}$ derived from past observations of demand in period s of the season, or estimate a probability distribution to derive additional scenarios. This uncertainty model is based solely on past demand data and ignores forecasts available in each review period. It is a static approach that does not benefit from forecast updates obtained in rolling horizon.

3.4.2. Seasonal forecast error

In rolling horizon, an updated forecast is obtained in each review cycle covering a planning horizon of T periods. Let $f_{k,t}^{s,y}$ be the forecast for product k in period t of the planning horizon as seen in review period s of season y. We introduce an additional time index to distinguish the different versions of forecast relating to the same demand period.

To model uncertainty in a forecast-driven fashion and reflect the seasonality of both the demand and forecast processes, we introduce the concept of seasonal forecast error. The forecast error associated to planning period s of season y is defined by by $\mathbf{e^{s,y}} = (e_{k,t}^{s,y}) \in \mathbb{R}^{K \times T}$ where

$$e_{k,t}^{s,y} = d_k^{s+t-1,y} - f_{k,t}^{s,y}, \quad \forall k \in \mathcal{K}, t \in \mathcal{T}.$$

The novelty of this model is to assume that each planning period s has its own forecast error distribution, which is stationary across seasons. For each review period s in the season, the set of forecast error $\mathcal{E}_s = \{\mathbf{e}^{\mathbf{s,y}}, y \in \{1, \ldots, Y\}\}$ is the empirical distribution of the (unknown) multivariate random forecast error distribution. The forecast error can be measured a posteriori for all periods for which the actual demand has been observed. The empirical distribution can be used to estimate the parameters of an assumed distribution and sampled to create additional forecast error scenarios. Since it is not straightforward to decide a priori which uncertainty model provides the best results, we compare their performance numerically through out-of-sample simulations in Section 3.6.2.

3.4.3. Two-stage scenario tree

Let y_o be the current season for which we want to derive a production plan using Y past seasons. Demand-driven samples obtained in Section 3.4.1 can be used directly to form a scenario tree.

Forecast error samples can also be used to generate a scenario tree by correcting the currently available forecast with forecast error samples. Let $N-1$ be the number of forecast error samples equal to Y if one uses the empirical distribution. A two-stage scenario tree can be constructed as a fan containing N equiprobable sample paths. The first path is set to the deterministic demand forecast $f_{k,t}^{1,s,y_o} = f_{k,t}^{s,y_o}$ for all products over the planning horizon. The remaining scenarios can be determined as

$$f_{k,t,n+1}^{s,y_o} = f_{k,t}^{s,y_o} + e_{k,t,n}^{s}, \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, n \in \{1, \ldots, N-1\}.$$

to correct the deterministic forecast with the seasonal forecast error samples of the same review period. Scenarios with negative demand are corrected to take the value zero.

3.4.4. Summary

We have shown strategies to overcome Barrier 2 (Uncertainty definition) and Barrier 3 (Uncertainty model) with limited available data. Seasonal uncertainty models based on forecast and demand data have been presented and integrated in two-stage scenarios trees. The use of scenarios based on the empirical distribution and estimated distribution have been considered.

3.5. Stochastic planning model with flexibility, stability and communicability

The uncertainty model and scenario tree presented in the previous section are now integrated in a stochastic planning model. We present different recourse structures and strategies to ensure communicability and flexibility of reference plans for both production and raw-material decisions. The notation for this section is summarised in Table B.1 in the appendix.

3.5.1. Stochastic model without recourse

The planner manages a portfolio of K products made from A raw materials and needs to determine a production plan over a horizon of T periods. Consider a general product structure in which raw material can be used for several products and each product can require multiple raw materials. The bill of material is given by $\mathbf{U} = (u_{k,a}) \in \mathbb{R}^{K \times A}$ where $u_{k,a}$ is the amount of raw material a required to produce one unit of product k. The planner is responsible for several production sites that serve a regional market. Each site contains parallel lines with different capacity κ_l and product portfolio. The set of production lines at site w is denoted by $\mathcal{L}_w$. The set of products that can be formulated on line l is given by $\mathcal{K}_l = \{k \in \mathcal{K} \mid \rho_{k,l} = 1\}$. Raw-material inventory is kept in a single warehouse and shared over the production sites whereas finished goods are held at the production sites. At the end of each review period, the company incurs a per-unit holding costs ν_a for raw-material a and $\mu_{k,w}$ for product k in site w. The problem setting is illustrated in Figure 3.1 for two production sites and five lines.

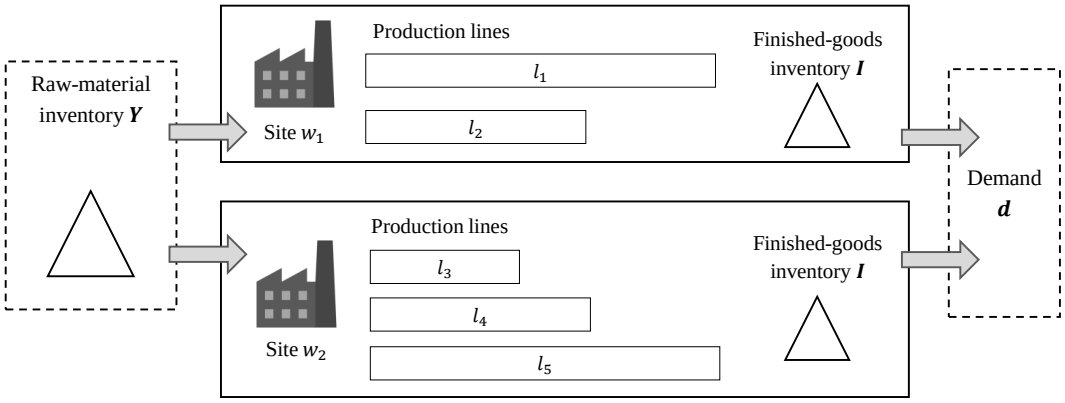

Figure 3.1.: Supply, production and inventory system for $W = 2$ sites and $L = 5$ lines.

In each review period, the planner uses a scenario tree $f \in \mathbb{R}^{(K\times T\times N)}$ to determine a production plan over the horizon and to communicate raw-material orders to the upstream level. The planner's goal is to satisfy the uncertain demand while minimising the inventory costs of raw materials and finished goods. Unmet demand is considered a lost sale and penalised with per-unit cost γ_k. The planning model is formulated as a two-stage stochastic model. Production decisions and raw-material orders are set as first-stage variable in order to provide a reference plan over the horizon. The inventory, sales and lost-sales decisions are set as recourse variables. The model is presented in Problem (3.1) where the season and review period indices are dropped for clarity.

$$\min \quad \sum_{t=1}^{T}\left(\frac{1}{N}\sum_{k=1}^{K}\sum_{w=1}^{W}\mu_{k,w}\sum_{n=1}^{N} i_{k,w,t,n} + \sum_{a=1}^{A}\nu_a \cdot y_{a,t} + \frac{1}{N}\sum_{k=1}^{K}\gamma_k\sum_{n=1}^{N}\cdot b_{k,t,n}\right) \tag{3.1a}$$

$$\text{s.t.} \quad i_{k,w,t,n} = i_{k,w,t-1,n} + \sum_{l\in\mathcal{L}_w} q_{k,l,t} - s_{k,w,t,n}, \qquad \forall k,w,t,n \tag{3.1b}$$

$$f_{k,t,n} = b_{k,t,n} + \sum_{w=1}^{W} s_{k,w,t,n}, \qquad \forall k,t,n \tag{3.1c}$$

$$\sum_{k=1}^{K} q_{k,l,t} \le \kappa_l, \qquad \forall l,t \tag{3.1d}$$

$$y_{a,t} = y_{a,t-1} + z_{a,t} - \sum_{k=1}^{K}\sum_{l=1}^{L}\beta_{k,a}\cdot q_{k,l,t}, \qquad \forall a,t \tag{3.1e}$$

$$q_{k,l,t}, y_{a,t}, z_{a,t} \ge 0, \qquad \forall k,w,l,a,t \tag{3.1f}$$

$$i_{k,w,t,n}, b_{k,t,n}, s_{k,t,n} \ge 0, \qquad \forall k,l,w,t,n \tag{3.1g}$$

The objective function in (3.1a) minimises the expected costs of inventory and lost sales over the different scenarios where the lost-sales penalty cost γ_k adjusts the conservativeness of the solution. Constraint (3.1b) describes the inventory balance at the production sites. Constraint (3.1c) ensures that demand is satisfied from sales or accounted as a lost sale in each scenario path. Constraint (3.1d) limits the production on each line to its

capacity in each period. Constraint (3.1e) describes the raw-material inventory balance. Constraints (3.1f) and (3.1g) specify the domain of the first-stage and recourse decisions variables respectively.

To improve planning stability, production and raw-material decisions can be frozen over the short-term horizon, prohibiting changes from decisions made in the previous review period. The frozen horizon can be implemented through the additional constraints

$$z_{a,t} = z^0_{a,t}, \qquad \forall a, t \leq \tau_a \tag{3.2a}$$

$$q_{k,l,t} = q^0_{k,l,t}, \quad \forall k, l, t \leq \tau_k \tag{3.2b}$$

where $z^0_{a,t}$ and $q^0_{k,l,t}$ are raw-material orders and production values determined in the previous review period. The length of the frozen horizon for production decisions τ_k and raw-material orders τ_a is chosen by the planner.

The stochastic model presented in (3.1) overcomes Barrier 5 (Communicability) by providing a reference plan on both the production and raw-material levels. Barrier 6 (Plan stability) can be overcome by freezing decisions on either the raw-material, the production levels, or both. However, the resulting model provides low flexibility since there is no recourse production and short-term decisions are frozen.

3.5.2. Improving flexibility through production recourse

By allowing recourse, decisions can be adapted to each scenario leading to less conservative here-and-now decisions. However, recourse variables limit planning communicability since the planner does not determine a unique reference plan but a tree of decisions. We introduce a stochastic model with recourse that provides high flexibility and communicability. The model is based on product families built through a data-driven optimisation model with custom rewards and constraints that reflect the product structure. The families are integrated in the planning model that reserves capacity on the family level through first-stage decisions. Thus, a reference plan is obtained on the family level. Recourse production decisions that consume the reserved capacity are implemented for products within families. In the numerical study, we show that aggregating production decisions over products improves planning stability since production changes tend to compensate within product families.

Product families: a multi-objective problem

The product-to-family assignment problem is a multi-criteria decision problem. Together with our industrial partner, we identify properties that the final assignment should exhibit: (1) the family assignment should cover many products, (2) a large share of the demand should be covered in product families, and (3) products with high uncertainty should be prioritised in the allocation. Each property is formulated as a normalised reward function, so that the rewards can be weighted easily to reflect planners' preferences. The product families should also respect the operational constraints and provide high visibility to site planners and schedulers. The product families are built in a data-driven fashion by using the historical data set of Y seasons.

Custom reward functions. Let $x_{k,f}$ be the binary variable equal to 1 if product k is assigned to family f. The first reward function is given by

$$\psi_1(X) = \frac{1}{K} \sum_{f=1}^{F} \sum_{k=1}^{K} x_{k,f}$$

and simply counts the number of products assigned to families. The second reward quantifies the share of demand covered by assigned products. It is given by

$$\psi_2(X) = \frac{1}{\sum_{k=1}^{K} td_k} \sum_{f=1}^{F} \sum_{k=1}^{K} td_k \cdot x_{k,f}$$

where $td_k = \sum_{y=1}^{Y} \sum_{s=1}^{S} d_k^{s,y}$ is the total demand of product k over the data set. The third reward prioritises products with high uncertainty. It is expressed by

$$\psi_3(X) = \frac{1}{\sum_{i=1}^{K} fe_i} \sum_{f=1}^{F} \sum_{k=1}^{K} fe_k \cdot x_{k,f}$$

where fe_k represents the difficulty to forecast product k. In this paper, we measure the uncertainty of a product using the weighted mean absolute percent error (wMAPE). This measure is normalised and allows to compare the forecast error of products with different demand share. The wMAPE forecast error in review period s of season y is given by

$$fe_{k,t}^{s,y} = \frac{\sum_{t=1}^{T} \omega_t \mid d_k^{s+t-1,y} - f_{k,t}^{s,y} \mid}{\sum_{t=1}^{T} \omega_t \cdot d_k^{s+t-1,y}} \quad \forall k, t.$$

where the weighting factor ω_t emphasises forecast error over the short-term horizon. The average product forecast error is then calculated as $fe_k = \frac{1}{T}\sum_{t=1}^{T} fe_{k,t}$.

Model formulation. The optimisation model is formulated in Problem (3.3).

$$\max_{x} \quad \sum_{i=1}^{3} w_i \cdot \psi_i(x) \tag{3.3a}$$

$$\text{s.t.} \quad \sum_{f=1}^{F} x_{k,f} \leq 1, \quad \forall k \tag{3.3b}$$

$$x_{k_1,f} \cdot x_{k_2,f} \leq \rho_{k_1,l} \cdot \rho_{k_2,l}, \quad \forall k_1, k_2, l, f \tag{3.3c}$$

$$x_{k_1,f} \cdot x_{k_2,f} \leq 1 - m_{k_1,a_1} \cdot m_{k_2,a_2} \cdot (1 - m_{k_1,a_2}) \cdot (1 - m_{k_2,a_1}) \tag{3.3d}$$

$$- m_{k_1,a_2} \cdot m_{k_2,a_1} \cdot (1 - m_{k_1,a_1}) \cdot (1 - m_{k_2,a_2}), \quad \forall k_1, k_2, a_1, a_2, l, f,$$

$$x_{k,f} \in \{0; 1\}, \quad \forall k, f. \tag{3.3e}$$

The objective function in (3.3a) maximises the weighted sum of rewards. Constraint (3.3b) ensures that a product is assigned to at most one family. Constraint (3.3c) specifies that all products within a family must be produced on the same set of production lines. Constraint (3.3d) states that there is always a sequence of products feasible without cleaning operation within a product family. Although cleaning operations are outside the scope of tactical planning, we ensure that the reference plan on the family level provides high visibility for the site schedulers. In the agrochemical industry, cleaning operations are conducted each time a raw material is removed when switching equipment from one product to the next. Let $\mathbf{m} = (m_{k,a})$ be the raw-material usage matrix where $m_{k,a}$ is equal to 1 if product k requires raw material a and 0 otherwise. Constraint (3.3d) holds for any number of products and raw materials. Although the above formulation is non-linear, the product of binary variables in Constraints (3.3c) and (3.3d) can be linearised by adding auxiliary variables $z_{k1,k2,f}$ and the following constraints:

$$z_{k_1,k_2,f} \leq x_{k_1,f}, \quad \forall k_1, k_2, f \tag{3.4}$$

$$z_{k_1,k_2,f} \leq x_{k_2,f}, \quad \forall k_1, k_2, f \tag{3.5}$$

$$z_{k_1,k_2,f} \geq x_{k_1,f} + x_{k_2,f} - 1, \quad \forall k_1, k_2, f. \tag{3.6}$$

Two-stage stochastic model with production recourse

The product families are integrated in a stochastic model that allows recourse production decisions. The extended stochastic model with family reserves and production recourse is formulated in Problem (3.7).

$$\min \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{w=1}^{W}\mu_{k,w}\sum_{n=1}^{N}\frac{1}{N}i_{k,w,t,n}+\sum_{a=1}^{A}\nu_a\cdot\sum_{n=1}^{N}\frac{1}{N}y_{a,t,n}\ +\sum_{k=1}^{K}\gamma_k\sum_{n=1}^{N}\frac{1}{N}\cdot b_{k,t,n} \quad (3.7a)$$

$$\text{s.t.} \quad i_{k,w,t,n}=i_{k,w,t-1,n}+\sum_{l\in\mathcal{L}_w}\Big(q_{k,l,t}+r_{k,l,t,n}\Big)-s_{k,w,t,n}, \qquad \forall k,w,t,n \quad (3.7b)$$

$$f_{k,t,n}=b_{k,t,n}+\sum_{w=1}^{W}s_{k,w,t,n}, \qquad \forall k,t,n \quad (3.7c)$$

$$\sum_{k=1}^{K}q_{k,l,t}+\sum_{f=1}^{F}h_{f,l,t}\le\kappa_l, \qquad \forall l,t \quad (3.7d)$$

$$\sum_{k\in\mathcal{K}_f}r_{k,l,t,n}\le h_{f,l,t}, \qquad \forall f,l,t,n \quad (3.7e)$$

$$r_{k,l,t,n}\le\sum_{f=1}^{F}x_{k,f}\cdot h_{f,l,t}, \qquad \forall k,l,t,n \quad (3.7f)$$

$$q_{k,l,t}+r_{k,l,t,n}\le\kappa_l\cdot\rho_{k,l}, \qquad \forall k,l,t,n \quad (3.7g)$$

$$y_{a,t,n}=y_{a,t-1,n}+z_{a,t}-\sum_{k=1}^{K}\sum_{l=1}^{L}\beta_{k,a}\cdot(q_{k,l,t}+r_{k,l,t,n}), \qquad \forall a,n,t \quad (3.7h)$$

$$u_{k,t,n}\cdot\sum_{l=1}^{L}\kappa_l\cdot\rho_{k,l}\ge\sum_{l=1}^{L}r_{k,l,t,n}, \qquad \forall k,t,n \quad (3.7i)$$

$$\sum_{n=1}^{N}u_{k,t,n}=N-1, \qquad \forall k,t \quad (3.7j)$$

$$\sum_{k\in\mathcal{K}_f}\sum_{l=1}^{L}r_{k,l,t,n}\ge\sum_{l=1}^{L}h_{f,l,t}-v_{f,t,n}\cdot\sum_{l=1}^{L}\kappa_l, \qquad \forall f,t,n \quad (3.7k)$$

$$\sum_{n=1}^{N}v_{f,t,n}=N-1, \qquad \forall f,t \quad (3.7l)$$

$$r_{k,l,t,n},h_{f,l,1}=0, \qquad \forall k,l,f,n \quad (3.7m)$$

$$z_{a,t}=z^0_{a,t}, \qquad \forall a,t\le\tau_a \quad (3.7n)$$

$$q_{k,l,t}=q^0_{k,l,t}, \qquad \forall k,l,t\le\tau_k \quad (3.7o)$$

$$h_{f,l,t} = h^0_{f,l,t}, \qquad \forall f, l, t \leq \tau_k \quad (3.7p)$$

$$q_{k,l,t}, h_{f,l,t}, z_{a,t}, \geq 0, \qquad \forall k, l, a, f, t \quad (3.7q)$$

$$i_{k,w,t,n}, b_{k,t,n}, s_{k,t,n}, y_{a,t,n}, r_{k,l,t,n} \geq 0, \qquad \forall k, l, w, a, t, n \quad (3.7r)$$

$$u_{k,t,n}, v_{f,t,n} \in \{0; 1\}, \qquad \forall f, k, t, n. \quad (3.7s)$$

The objective function in (3.7a) minimises the expected costs of finished-goods inventory, raw-material inventory and lost sales. Constraint (3.7b) describes the inventory balance of finished goods at each production site. Constraint (3.7c) tracks the demand satisfaction from the sites. Constraint (3.7d) ensures that production and capacity reserves on each line do not exceed available capacity. Constraint (3.7e) states that recourse production within a family is restricted by its capacity reserve in each scenario. Constraint (3.7f) ensures that there is no recourse production for unassigned products. Constraint (3.7g) specifies that production on a line is restricted to its feasible portfolio. Constraint (3.7h) describes the raw material balance. Constraints (3.7i) and (3.7j) ensure that the minimum recourse production over all scenarios is zero for each product and time period. Constraints (3.7k) and (3.7l) force the maximum recourse production over all scenario to be equal to the capacity buffer reserved for each product in each period. These two sets of constraints ensure that the capacity buffer reserved for each family accounts exactly for the volatile part of demand. Constraint (3.7m) states that there is no recourse variable in the first period. Constraints(3.7n) implements a frozen horizon on the raw-material orders. Constraints (3.7o) and (3.7p) implement a frozen horizon on first-stage production decisions and capacity reserves respectively. Constraints (3.7q) and (3.7r) express the domain of the continuous first-stage and recourse variables respectively. Constraint (3.7s) defines the auxiliary binary variables to identify the minimum and maximum production recourse over the scenarios.

3.5.3. Summary

The stochastic model with recourse determines the optimal first-stage raw-material orders, production and capacity reserves that allows flexible second-stage production decisions, overcoming Barrier 4 (Flexibility representation). Recourse production can only be used for products within families if enough resources have been reserved. Although it uses scenarios, the model overcomes Barrier 5 (Communicability) by providing an aggregated reference plan defined so that a detailed production schedule can still be derived by downstream planners. Barrier 6 (Stability) is overcome by aggregating first-stage

decisions on the family level. The model also explicitly distinguishes between parts of the plan likely to be conducted (first-stage production) and parts of the plan potentially subject to changes (recourse production).

3.6. Numerical study

The numerical study applies the strategies developed in the previous sections to the real-world case study and details the final steps to overcome barriers of stochastic programming in practice. Due to the large number of parameters and performance indicators, it is difficult to investigate their interactions in a full factorial experiment. Instead, we analyse sequentially the strategies related to the uncertainty process from Section 3.4 and the planning structure from Section 3.5. First, we present the simulation setting, the performance metrics and the problem parameters. Second, we compare the performance of demand-driven and forecast-driven uncertainty models as well as the use of empirical or estimated distributions. Third, we evaluate the stochastic models with varying nervousness mitigation strategies. Finally, we compare our model to the current practice of our industrial partner.
Simulations are implemented in the Julia programming language (Bezanson et al., 2017) and are run on an Intel(R) Core(TM) i7-4810MQ processor at 2.80Ghz using 16GB of RAM. The optimisation problems are formulated using JuMP (Dunning et al., 2017) and solved with Gurobi 9.0. The relative MIP gap is set to 0.1% for all instances of the stochastic model with production recourse.

3.6.1. Simulation setting

All simulations from model parameterisation to final evaluation are conducted in an out-of-sample rolling-horizon fashion. In each review period, the following steps are taken: (1) a production plan is calculated over the planning horizon using the available forecast or scenario tree, (2) the production quantity of the first period is added to the on-hand inventory, (3) the actual demand is observed, (4) sales are subtracted from the inventory and lost sales are observed if demand is higher than the on-hand inventory, and (5) the new inventory position is determined.
We gather a data set containing the planning history of our industrial partner over $Y = 4$ seasons of $S = 12$ months each. Rolling-horizon simulations are run for each season independently using the other $(Y - 1)$ seasons to construct the uncertainty model. In

each review period, the forecast and demand are taken from the historical data set of our industrial partner. This simulation setting allows to carry Y independent out-of-sample simulations.
The initial inventory in the first period of the season is set to the historical inventory of the company. Each simulation is started $\tau = \max(\tau_a, \tau_k)$ periods earlier than the first period of the season and the demand and forecast are set to zero during this warm-up phase. The corresponding review periods are ignored for the model evaluation. To neglect the interactions between consecutive seasons, we replace demand and forecast values by zero for all periods later than the last period of the current simulation season.

Key performance indicators

The models are evaluated using four key performance indicators: the service level, the inventory costs, the planning nervousness and the nervousness of the raw-material orders.

Key trade-off: service level and inventory. The service level is measured as the proportion of satisfied demand over the season given by

$$sl = \frac{\sum_{s=1}^{S}\sum_{k=1}^{K}(d_{k,s} - b_{k,s}^{(r)})}{\sum_{s=1}^{S}\sum_{k=1}^{K} d_{k,s}}$$

where $b^{(r)}$ are the realised lost sales of product k in simulation period s respectively. The inventory costs are measured as the sum of finished-goods and raw-material inventory costs over the season as

$$ic = \sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{w=1}^{W} \mu_{k,w} \cdot i_{k,w,s}^{(r)} + \sum_{s=1}^{S}\sum_{a=1}^{A} \nu_a \cdot y_{a,s}^{(r)}$$

where $i_{k,w,s}^{(r)}$ and $y_{a,s}^{(r)}$ are the realised finished-goods inventory and raw-material inventory observed at the end of period s. All inventory costs reported are normalised by dividing them by the company's average historical inventory costs.

Planning stability. There is a large body of literature discussing how to measure nervousness. Quantity-oriented and setup-oriented nervousness measures have been distinguished, which are particularly relevant in lot-sizing contexts (Tunc et al., 2013). Early measures focus on setup-oriented nervousness such as Carlson et al. (1979) who account only for the nervousness induced by adding a new setup in the plan. Sridharan et al.

(1988) have proposed a quantity-oriented measure that assigns weights to periods in the horizon in order to emphasise short-term stability. While the majority of existing measures are absolute, relative nervousness measures are more interpretable. Jensen (1993) proposes a normalised nervousness measures that relate nervousness to the maximum nervousness possible, which can be determined from the available capacity. However, this measure has several practical shortcomings: it cannot be applied if the maximum nervousness is unbounded, and it might give a false sense of stability if capacity is large. We propose a novel quantity-oriented nervousness measure that is relative to the plan itself. This measure provides high interpretability and allows to compare several planning steps. For instance, we compare production planning nervousness, raw-material nervousness and forecast nervousness in Section 3.6.4.

Planning stability is measured independently on both the production and raw material levels. Planning nervousness is measured as the average sum of absolute changes between production volumes aggregated on the product family level. It is based on the observation that nervousness within a family is negligible compared to nervousness between families. Planning nervousness is measured as

$$
\begin{aligned}
nsf = \frac{1}{S-1}\sum_{s=2}^{S} & \frac{1}{\max\left(\sum_{t=1}^{T-1}\sum_{l=1}^{L}\sum_{k=1}^{K} Q_{k,l,t}^{(s)} + \sum_{f=1}^{F} HK_{f,l,t}^{(s)}, \sum_{t=1}^{T-1}\sum_{l=1}^{L}\sum_{k=1}^{K} Q_{k,l,t+1}^{(s-1)} + \sum_{f=1}^{F} HK_{f,l,t+1}^{(s-1)}\right)} \\
& \times \Big(\sum_{t=1}^{T-1}\sum_{f=1}^{F} \mid \sum_{l=1}^{L} (HK_{f,l,t}^{(s)} + \sum_{k\in\mathcal{K}_f} Q_{k,l,t}^{(s)} - HK_{f,l,t+1}^{(s-1)} - \sum_{k\in\mathcal{K}_f} Q_{k,l,t+1}^{(s-1)}) \mid \\
& + \sum_{k\in\mathcal{K}\setminus\mathcal{KF}} \mid \sum_{l=1}^{L} Q_{k,l,t}^{(s)} - Q_{k,l,t+1}^{(s-1)} \mid \Big)
\end{aligned}
\tag{3.8}
$$

where $\mathcal{K} \setminus \mathcal{KF}$ is the set of products not assigned to any family. Raw-material orders nervousness is given by

$$
nsa = \frac{1}{S-1}\sum_{s=2}^{S} \frac{\sum_{t=1}^{T-1}\sum_{a=1}^{A} \mid Z_{a,t}^{(s)} - Z_{a,t+1}^{(s-1)} \mid}{\max\left(\sum_{t=1}^{T-1}\sum_{a=1}^{A} Z_{a,t}^{(s)}, \sum_{t=1}^{T-1}\sum_{a=1}^{A} Z_{a,t+1}^{(s-1)}\right)}. \tag{3.9}
$$

The nervousness measures are quantity oriented. They account for plan changes due to both volume and timing. Nervousness is calculated relative to the reference plan. This normalisation does not guarantee that nervousness is always between 0 and 1 but

increases interpretability and allows the comparison of different models.

Problem parameters

The product portfolio contains $K = 55$ products formulated from $A = 13$ raw materials. There are $W = 2$ production sites with 2 production lines at the first site and 3 lines at the second site. The demand is seasonal with periodicity $S = 12$ periods and the planning horizon is set to $T = 12$ periods. The raw-material frozen horizon is set to $\tau_a = 2$ periods. There is no frozen horizon for production decisions. The line capacities, bill of materials, each line's product portfolio and the inventory costs have been collected together with our industrial partner. The lost-sales penalty cost is set proportional to the product inventory cost as $\gamma_k = \lambda \cdot \max_{w \in \mathcal{W}}(\mu_{k,w})$.

3.6.2. Evaluation of uncertainty models

Pareto fronts

The stochastic models are implemented in out-of-sample rolling-horizon simulations to measure the value of different uncertainty models and decide on the optimal configuration. We compare the performance of the empirical distribution and estimated distributions as well as the use of forecast-driven and demand-driven models. The value of augmenting the scenario tree with scenarios sampled from assumed distribution is also measured.

Normal and uniform distributions are fitted to the empirical samples. A normal distribution is estimated from the empirical mean and variance of the forecast error independently for each product and time period. The bounds of the uniform distribution are taken as 80% and 120% of the minimum and maximum empirical forecast error for each product and time period. We refrain from estimating covariance parameters since the number of samples $(Y - 1)$ is significantly smaller than the number of parameters to estimate $(K^2 \times T^2)$. Demand scenarios are then sampled using Descriptive Sampling (Saliby, 1990). To identify the value of stochastic programming, we also show the Pareto front of deterministic models with exogenous safety stock calculations. The deterministic optimisation model presented in B.1 is implemented with additional exogenous safety stocks determined by $ss_{k,t} = z \cdot \sigma_{k,t}$ where σ is the standard deviation of demand (DD) or forecast error (DF) and z is a conservativeness parameter set by the planner.

Since the planning problem has four objectives, several trade-offs exist between the performance indicators presented in Section 3.6.1. We focus on the most important

trade-off between realised service level and inventory costs. While the planner is interested in achieving high demand satisfaction, it comes at the price of more conservative decisions yielding higher inventory costs. We determine the Pareto front of stochastic forecast-driven (SF) and demand-driven (SD) models using empirical, normal and uniform distributions. We study the effect of varying the number of scenarios by setting $N \in \{4, 8, 16, 32\}$ for the normal and uniform distributions while the number of empirical scenarios remains equal to $N = 4$. For each scenario tree, a sensitivity analysis of the lost-sales penalty cost factor is performed with $\lambda \in \{1, 2, 5, 10, 15, 20, 25, 30, 50, 200\}$. Similarly, the conservativeness of the deterministic models is adjusted through parameter $z \in \{0, 0.2, 0.4, \ldots, 1.8, 2\}$.

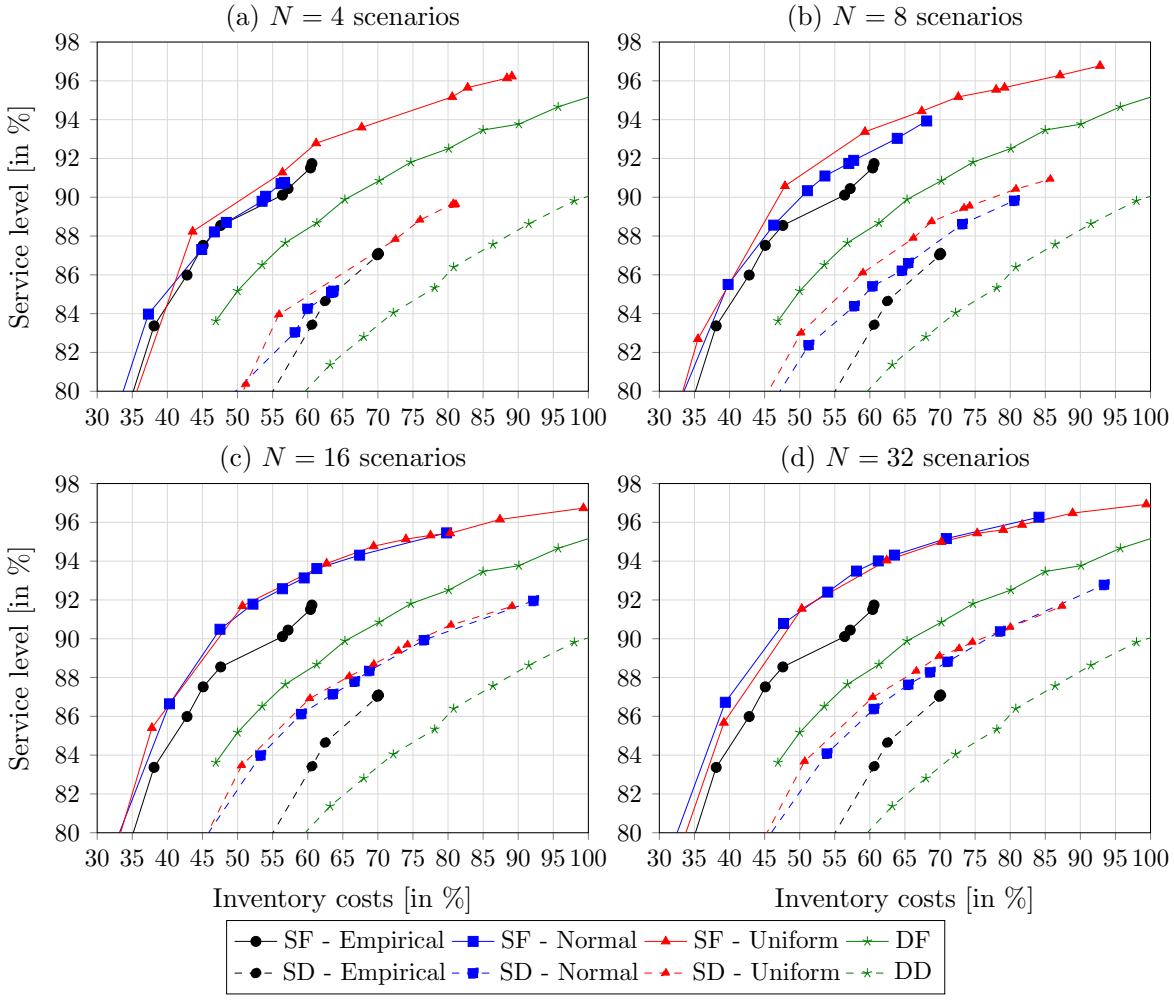

Figure 3.2.: Pareto front between service level and inventory costs for different model configurations.

The Pareto fronts are shown in Figure 3.2 where each mark corresponds to the average

performance over the Y seasons for a given lost-sale penalty cost or deterministic safety factor. Several conclusions can be drawn from the simulation results: (a) forecast-driven models outperform demand-driven models, (b) stochastic models dominate deterministic models with exogenous safety stock calculations, (c) the value of sampling additional scenarios decreases quickly so that only few scenarios are necessary to achieve good performance, and (d) the uniform distribution dominates other distributions for small scenario trees but appears equivalent to the normal distribution for larger tree sizes.
This analysis highlights the importance of the uncertainty modelling step identified in Barrier 2 (Uncertainty definition) when applying stochastic programming from data. Notably, the deterministic forecast-driven model outperforms all stochastic demand-driven models, confirming our intuition that defining uncertainty correctly may be more important that applying advanced stochastic techniques. Still, using a stochastic model instead of a deterministic model provides significant benefits. For small sample sizes, the uniform distribution provides the best results, which may be explained by the fact that it contains more extreme scenarios that allows it to reach high service levels. For large scenario trees, which provides a more accurate evaluation of the distribution quality, the normal and uniform probability distributions yield similar performance. The results suggest that overcoming Barrier 2 (Uncertainty definition) is even more important than Barrier 3 (Uncertainty model), even though the latter has received much more attention in the literature.

Out-of-sample regret

To highlight the importance of performing out-of-sample simulations, we compare the results of in-sample and out-of-samples simulations. We investigate the relative out-of-sample regret, which is defined as the difference between the average performance obtained with in-sample and out-of-sample simulations divided by the in-sample performance.
The relative regret of service level and inventory costs is shown on Figure 3.3 as a function of the lost-sales penalty factor. The figure shows that all service level regrets are negative. In-sample simulations have an optimistic bias, which is consistent over all uncertainty models. Similarly, the inventory regret shows that out-of-samples inventory costs are overall higher than their in-sample estimates. Interestingly, the empirical distribution shows the highest regret on both the service level and inventory costs. Increasing the size of the scenario trees does not reduce the out-of-sample regret of estimated distributions. On the contrary, it leads to overall higher service level regret.

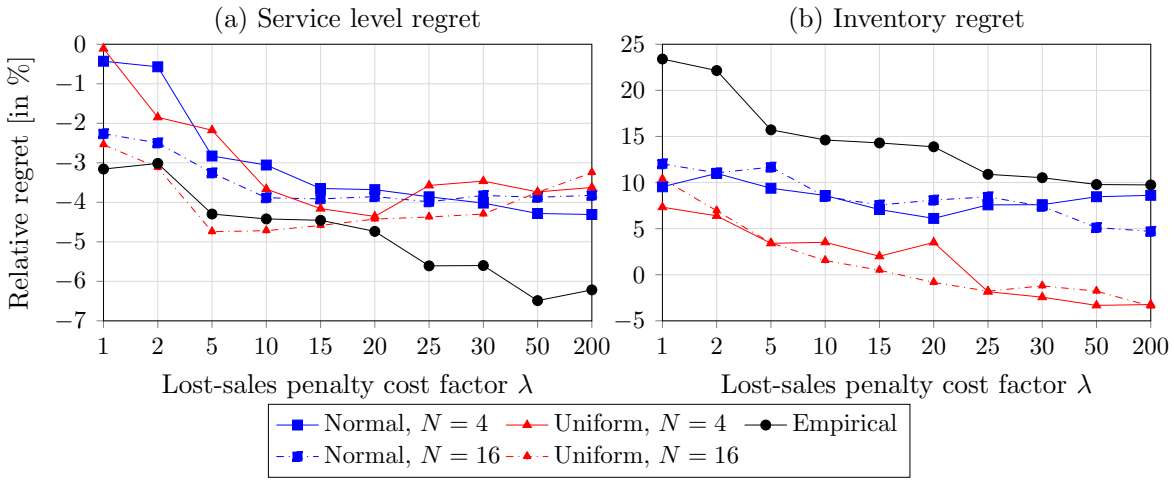

Figure 3.3.: Out-of-sample regret of realised (a) service level and (b) inventory cost.

Summary

Determining the Pareto fronts of the models with different uncertainty process definition and representation allows to overcome Barrier 2 (Uncertainty definition) and Barrier 3 (Uncertainty model). The out-of-sample evaluations are a key component for overcoming Barrier 1 (Data scarcity) and Barrier 8 (Evaluation). They provide an accurate and unbiased estimate of model performance. They also highlight the ability to generalise from past observations by estimating probability distributions and sampling from them. We overcome Barrier 7 (Tractability) by observing that a small scenario tree is enough to provide good out-of-sample performance. In the remainder of the numerical study, we use the forecast-driven stochastic model with $N = 8$ scenarios sampled from the uniform distribution. The lost-sales penalty cost factor is set to $\lambda = 15$, which ensures a satisfying trade-off between between service level and inventory costs.

3.6.3. Stochastic programming, recourse and planning stability

In this part, we investigate the trade-off between planning flexibility, stability and communicability. First, we illustrate the reference plan obtained with product families. Then, we evaluate the value of recourse and compare the effect of freezing and aggregating production decisions to mitigate nervousness.

Planning communicability

The stochastic model with production recourse presented in Section 3.5.2 determines a reference plan as a combination of capacity reserves and first-stage production decisions. An example is shown in Figure 3.4 for $F = 4$ product families, where production is shown relative to available capacity in each period. The figure shows the first-stage decisions aggregated over all products as well as the capacity reserves for the four families. It illustrates the variation in volume and timing between the different families over the planning horizon. The capacity reserves can be understood as the volatile part of the plan since they are used differently in each recourse scenario by the products in the family. Hence, a reference plan can be communicated while allowing flexible product-specific decisions.

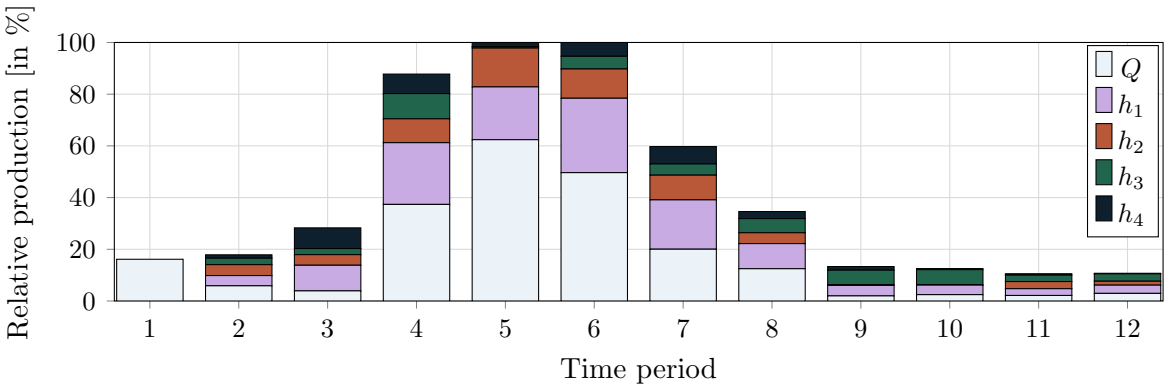

Figure 3.4.: Capacity reserves and first-stage production decisions relative to available capacity.

Raw-material stability

We analyse the impact of freezing raw-material ordering decisions by varying the frozen horizon length τ_a within the set $\{0, 1, 2, 3, 4, 5, 6\}$. The results are shown in Figure 3.5, which shows that freezing raw-material ordering decisions is an effective strategy to improve raw-material stability although it leads to increased inventory costs.
The stochastic model maintains high service level by increasing safety inventory, suggesting that the scenario tree accurately captures the raw-material uncertainty over the prediction horizon. Interestingly, there is no distinguishable effect on planning nervousness. Freezing raw-material orders does not reduce planning flexibility if enough safety inventory is available on the raw-material level. Hence, we fix the raw-material lead time to $\tau_a = 2$ to decrease raw-material nervousness with acceptable inventory costs increase.

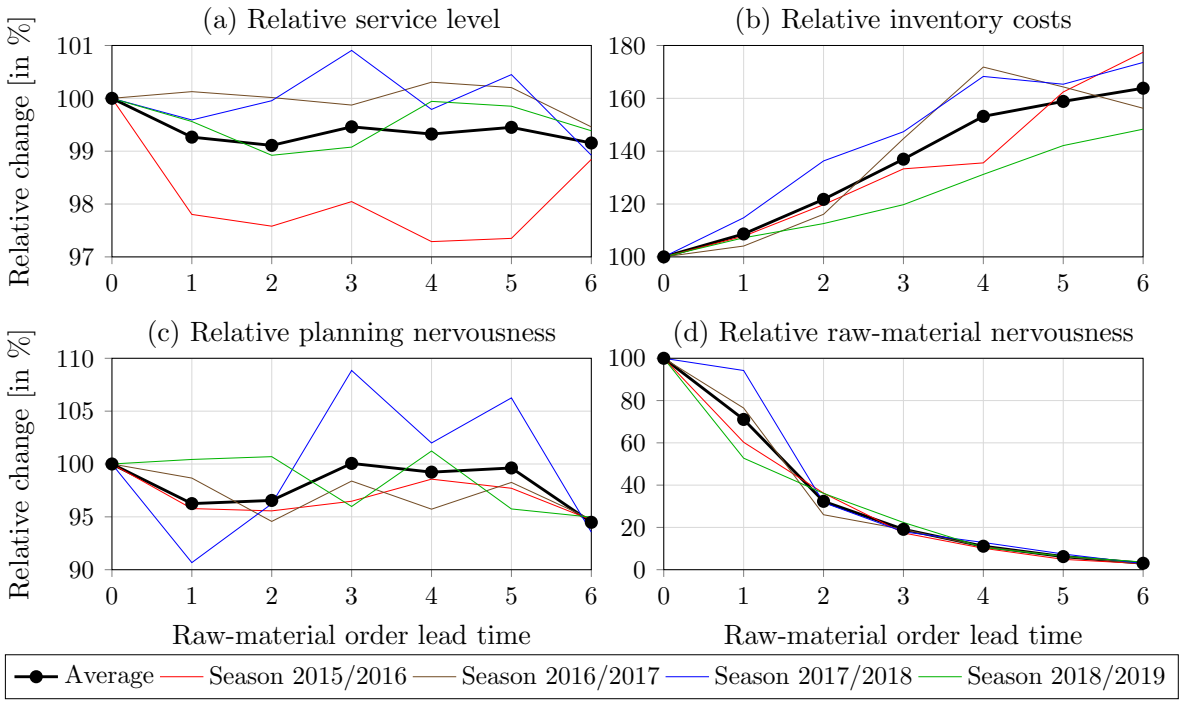

Figure 3.5.: Sensitivity analysis of raw-material ordering lead time.

Plan stability

In Section 3.5, two methods are presented to mitigate planning nervousness. The first strategy freezes production decisions over the short-term horizon while the second aggregates decisions over optimally defined families. We evaluate and compare their performance in a sensitivity analysis. In Figure 3.6, we show a side-by-side comparison of the effect of increasing the length of the frozen horizon and increasing the number of product families.

As for raw materials, implementing a frozen horizon on the production level gives significant reduction in planning nervousness. However, it leads to small decrease in average service level and comes at the cost of increased inventory costs. Freezing the production horizon also has a stabilising effect on raw-material orders since production flexibility is strongly reduced. On the other hand, the stochastic model with recourse provides high stability, high demand satisfaction and low costs. Since the product-to-family assignment model prioritises the assignment of products with high demand and large forecast errors, few product families are sufficient to observe large improvements in planning stability. As the number of families increases, planning nervousness decreases with diminishing marginal returns. For $F = 4$ families, the model can reduce inventory costs

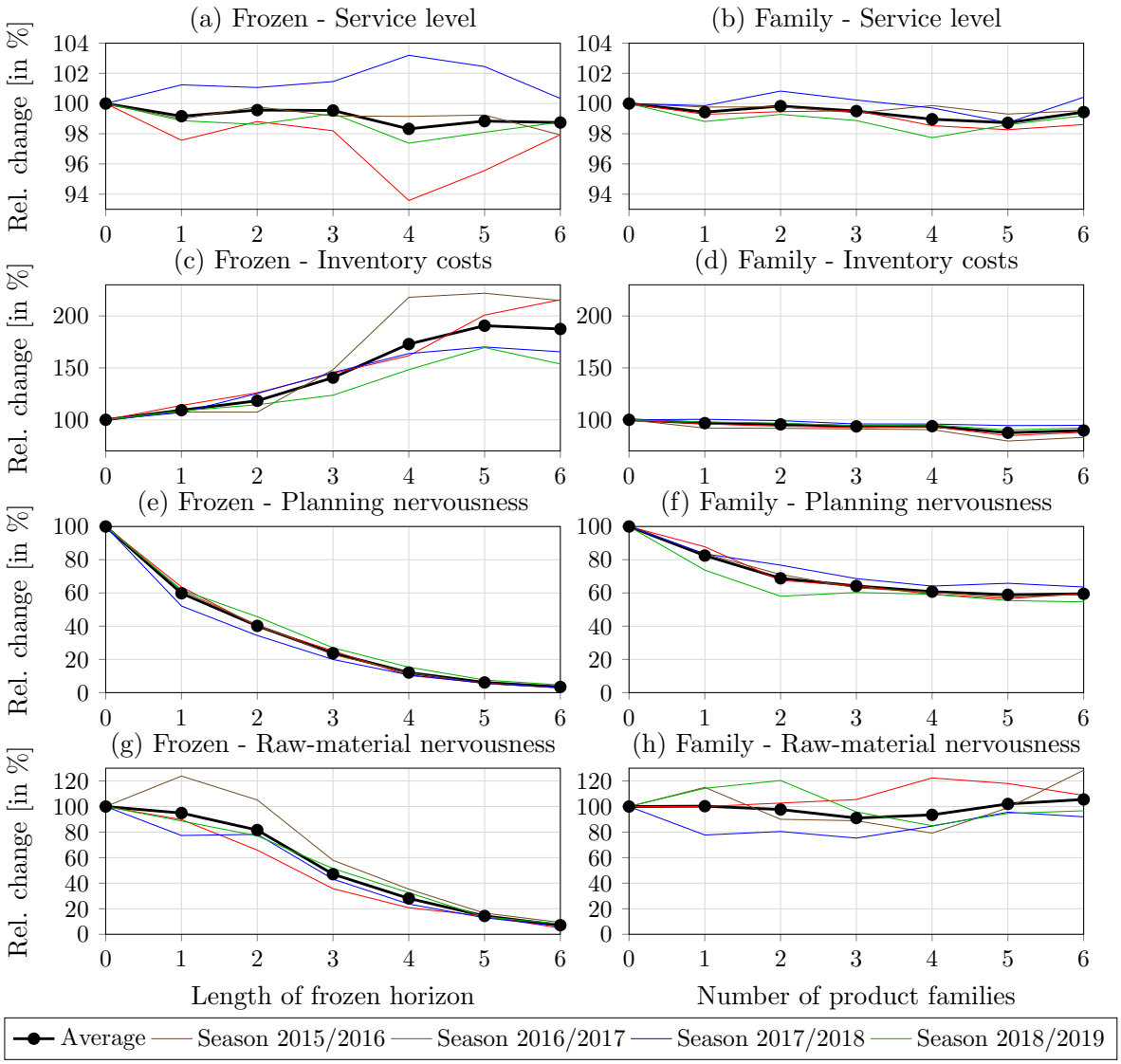

Figure 3.6.: Comparison of nervousness mitigation strategies on planning level.

by 6% while the average service level is decreased by only 1% and planning nervousness is reduced by 40%. On the contrary, freezing production decisions overly restricts the flexibility of the model, which may lead to unacceptable cost increase.

Value of recourse under varying capacity utilisation

In the agrochemical industry, capacity is expensive and capacity planning is an important long-term problem. To demonstrate the robustness of our approach in diverse settings, we analyse performance under varying capacity. Available capacity is reduced in 5% increments from 100% to 40%. The average performance are shown in Figure 3.7 for the

stochastic model without recourse corresponding to $F = 0$ family, the stochastic model with recourse and $F = 4$ families, as well as the stochastic model without family and frozen horizon $\tau_k = 1$.

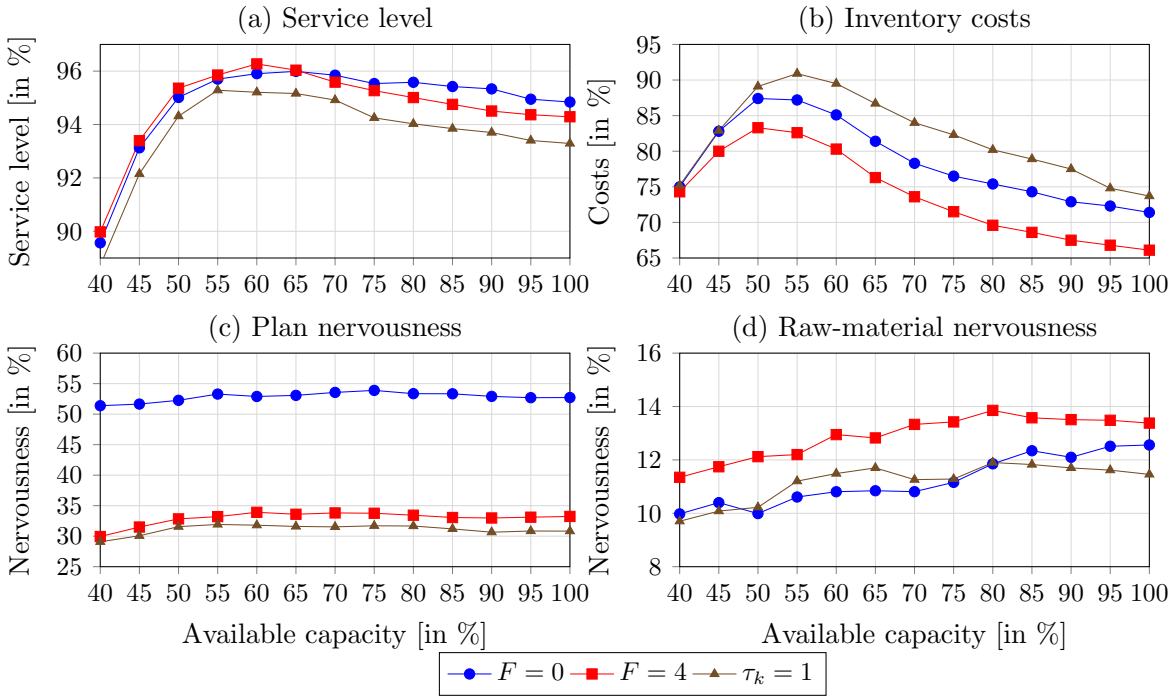

Figure 3.7.: Performance of stochastic models under varying capacity.

Figure 3.7 shows that inventory increases as capacity is further reduced, leading to higher costs but also higher service level. When capacity is severely limited, a steep decline in service level is observed. The simulations highlight that the value of recourse is robust over a wide range of capacity settings. Remarkably, the stochastic model with recourse provides highest service level and lower costs when capacity is highly utilised, which corresponds to capacity reduction of 65% and lower. On the contrary, the stochastic model with frozen production decisions yields the highest inventory costs and lowest service level over all instances. Hence, production flexibility is essential to manage short-term uncertainty when capacity is limited.

It is interesting to observe that realised service level is overall higher when capacity is limited. With little available capacity, production starts earlier and uses less accurate forecasts. Hence, additional safety stock are placed, leading to both higher inventory and service level. Yet, we note that the obtained solutions is a dominated solution on the Pareto front shown in Figure 3.2. The Pareto analysis performed in Section 3.6.2 should be performed with the new capacity to decide on the optimal trade-off between service

level and inventory costs and identify the lost-sales penalty cost factor that achieves the target service level.

Summary

The barriers linked to the planning processes have been overcome thanks to several strategies. Barrier 5 (Communicability) is solved by explicitly integrating raw-material orders and determining a reference plan on the family level. Barrier 4 (Flexibility representation) is overcome through recourse decisions, which proves especially relevant when capacity is highly utilised. Barrier 6 (Plan stability) is resolved by implementing a frozen horizon on raw-material orders and aggregating decisions over families. We show that there is not necessarily a trade-off between planning stability and flexibility. The proposed approach based on product aggregation is especially successful since it overcomes the above barriers jointly.

3.6.4. Comparison with industry benchmarks

To conclude the numerical study, we compare our approach to the current practices of our industrial partner. The stochastic model with $N = 8$ scenarios sampled from a uniform distribution and $F = 4$ families is compared to two benchmarks based on the historical data of our industry partner.

Benchmark definition

The *forecast* benchmark assesses the quality of the demand forecast. The service level of the benchmark is measured through a rolling simulation in which the on-hand inventory is set equal to the demand forecast, thus evaluating the forecast accuracy of the first period in the horizon. Planning and raw-material nervousness are determined by applying Equation (3.8) and Equation (3.9) respectively using the demand forecasts and forecasts translated into raw materials using the bill-of-material. The forecast benchmark does not lead to inventory costs, which are not reported.
The *company* benchmark represents the practice of our industrial partner. Currently, a combination of deterministic automated MRP software and expert knowledge is used to derive a production plan in each review period. The benchmark is based on the history of production plans and inventory levels. The service level of the company benchmark is measured by comparing the sum of historical on-hand inventory and the production plan implemented in rolling horizon to the demand. The inventory costs are determined

from the historical inventory of raw materials and finished goods. Planning nervousness is measured using Equation (3.8). Raw-material orders are obtained by converting the finished-goods production plan on the raw-material level and by accounting for on-hand raw-material inventory. Raw-material nervousness is then deduced using Equation (3.9).

Simulation results

The out-of-sample rolling-horizon simulation results are presented in Figure 3.8 for all seasons. The average results are given in Table 3.1. The forecast accuracy is poor since the forecast benchmark yields lowest service level in all seasons. Interestingly, the deterministic model provides higher service level although it uses the same demand forecasts. This can be explained by the fact that the deterministic model carries inventory from one period to the next if it produces more than the actual demand. This highlights a bias in the forecasting process: demand tends to be forecast earlier than it actually realises, which leads to inventory build up that is used in later periods. Overall the company benchmark achieves a high service level and outperforms the forecast and deterministic models, highlighting the value of planner expertise.
The stochastic model with $F = 4$ families achieves high service level consistently over the four seasons. It reduces inventory costs by more than 33% compared to the company benchmark, which suggests an efficient placement of safety stocks. It also yields substantial improvements in stability as planning nervousness is reduced by 40% thanks to the aggregation of planning decisions on the family level. Raw-material nervousness is reduced by almost 80% on average, which results in lower nervousness than the forecast benchmark. Thus, the planning step acts as a dampening step. Short-term demand variability is effectively mitigated, which provides a robust ordering signal to upstream raw-material planners.
The simulation setting and benchmark definition allows us to overcome Barrier 8 (Evaluation). The results show that the stochastic model with production recourse improves all performance indicators compared to the company historical practice: customer satisfaction is increased, inventory costs are reduced and planning is more stable on both the finished-goods and raw-material orders levels.

3.7. Conclusion

This paper aims to foster the use of stochastic programming in master production scheduling. First, we identify barriers that challenge the application of stochastic pro-

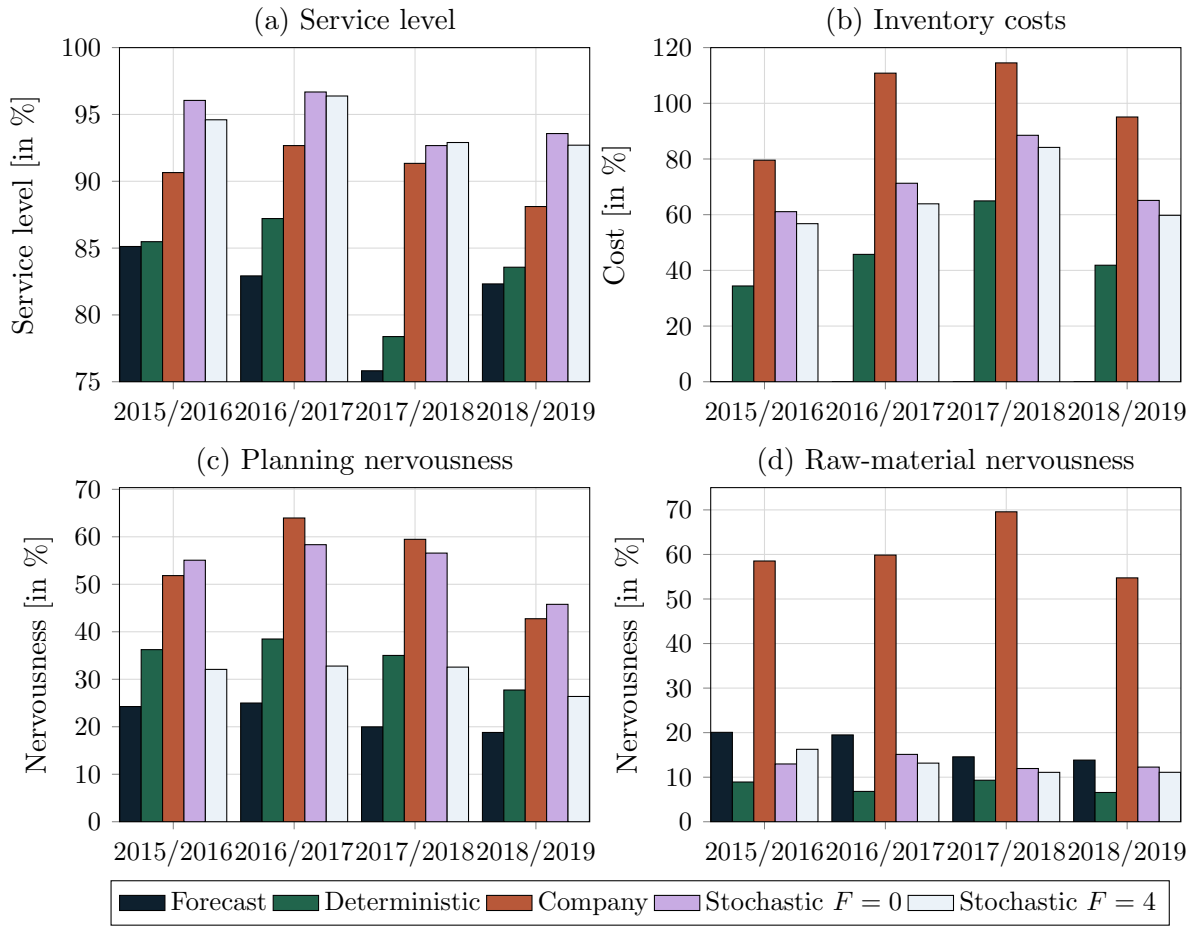

Figure 3.8.: Simulation results over four seasons: (a) service level, (b) inventory, (c) planning nervousness, and (d) raw-material nervousness.

gramming and relate them to a real-world case study in the agrochemical industry. Then, we discuss how to model the uncertainty from limited available data and construct scenario trees to represent future demand. The scenario trees are integrated into stochastic planning model that reflect the planning processes. The trade-off between planning communicability, stability and flexibility are integrated in a two-stage stochastic model that determine the optimal recourse production volumes. Finally, we demonstrate our framework on the case study, determine the best model configuration through sensitivity analyses and compare its result to industry practice.

The results of this paper extend beyond the scope of the case study considered. The barriers identified are common to a wide array of manufacturing environment. We hope to stimulate the discussion on their relevance and encourage the development of solutions suitable to varied production settings. The simulation study allows us to emphasise the importance of the definition and representation of the uncertain processes. Whereas

Table 3.1.: Average value of KPIs over all seasons and value relative to company.

KPI	Company	Forecast	Determ.	Stoch. $F = 0$	Stoch. $F = 4$
Service level [in %]	90.69	81.54	83.66	94.74	94.14
relative [in %]	100	89.9	92.2	104.5	103.8
Rel. inv. costs [in %]	100	-	46.7	71.5	66.2
Rel. fin.-goods inv. costs [in %]	100	-	33.8	80.3	76.2
Rel. raw-mat. inv. costs [in %]	100	-	62.2	61	54.2
Planning nervousness [in %]	54.5	22	34.4	53.9	31
relative [in %]	100	40.4	63.1	99	56.8
Raw-mat. nervousness [in %]	60.7	17	7.9	13.1	12.9
relative [in %]	100	28	13	21.6	21.3

existing literature overwhelmingly assumes that uncertainty models are available, we show that the carefully modelling uncertainty is critical. Indeed, a simple deterministic model with the right uncertainty model outperforms advanced stochastic models with inaccurate uncertainty definition.

We discern several directions for future research. The analysis of nervousness mitigation by aggregating production decisions could be extended to a multi-level supply chain. The application of advanced models to characterise the forecast revision process such as the Martingale Model of Forecast Evolution of Heath and Jackson (1994) could be investigated to further exploit available data and derive robust uncertainty models.

Chapter 4

Dynamic stochastic lot sizing with forecast evolution in rolling-horizon planning

Abstract

Stochastic lot-sizing problems arise in many production settings in which inventory and setup costs are incurred and demand is uncertain. While existing approaches assume that demand follows known probability distributions, many industries struggle to determine stochastic distributions from available data. Instead, they implement a rolling-horizon planning framework based on frequent forecast updates and deterministic models. Using the Martingale model of forecast evolution, we integrate stochastic forecast evolution in lot-sizing problems and solve the models efficiently using piecewise-linear approximation of the expected inventory and backlogs. The formulation is extended with production recourse through discrete demand scenarios to reflect the flexibility of rolling-horizon planning. Extensive rolling-horizon simulations on both synthetic and real-world data show the value of forecast evolution models. Forecast evolution models reduce realised costs by 14% on average compared to traditional deterministic planning. The advantage of the extended model with production recourse depends on several factors including capacity, correlation and uncertainty. Sensitivity analyses show that recourse can reduce cost by up to 10%. We conclude by identifying advantages and limitations of forecast evolution models and provide general recommendations to practitioners.

4.1. Introduction

Rolling horizon is a planning framework based on the periodic updating of demand forecasts and production decisions. This paradigm underlies the planning cycle of many companies. In academia, demand uncertainty has been studied extensively in the stochastic programming community using probability distributions to model the uncertain demand. However, only limited attention has been given to the performance of these stochastic models in rolling horizon, and the value of these methods compared to traditional deterministic planning is not always clear to practitioners. Further, there is an overall lack of pragmatic guidelines describing how to determine demand distributions from the data available to planners. In this light, we believe that forecast evolution models can allow the successful application of stochastic models in practical settings. The martingale model of forecast evolution (MMFE) developed by Graves et al. (1986) and Heath and Jackson (1994) is based on measuring forecast revisions in successive planning periods and modelling future forecast changes as a stochastic process. Not only do forecast evolution models rely on an existing rolling-horizon planning framework, they benefit directly from the history of past demand and forecasts routinely collected by practitioners.

Stochastic models, such as chance-constrained models, can be formulated to account for forecast uncertainty. Solving the models in each review period provides a production plan that can be implemented in a rolling-horizon fashion. However, this approach does not capture the progressive resolution of uncertainty as described by forecast evolution models. Planners can react to the updated forecast in each period and adapt their production plan accordingly.

Traditional stochastic approaches that formulate all decisions as first-stage variables over the planning horizon typically ignore this re-planning opportunity. Resulting production decisions are overly conservative. They create unnecessary safety stock and increase inventory costs. On the contrary, scenario-based stochastic models can explicitly integrate re-planning opportunities through recourse decisions. In particular, multi-stage scenario trees include a recourse opportunity in all periods of the planning horizon. By capturing the inherent flexibility of rolling-horizon planning, recourse models provide less conservative decisions and reduce operational costs. However, it is well known that the computational times of multi-stage models may become prohibitively long for large scenario trees.

This paper is motivated by a collaboration with a large company in the chemical in-

dustry that manages expensive multi-purpose equipment in the face of uncertain and seasonal demand. Since capacity is limited, production often starts ahead of the peak selling season, which can lead to expensive on-hand inventory. Extensive cleaning operations have to be conducted each time the equipment is setup for a different product family. The company's planning problem exhibits the key trade-off between demand satisfaction, inventory costs, and setup costs that is captured by the lot-sizing problem. Because early forecasts often have poor accuracy, planning is implemented in a rolling-horizon fashion to benefit from frequent forecast updates.

Even without recourse, stochastic lot-sizing problems are notoriously hard to solve. In this paper, we use the piecewise-linear approximation (PLA) introduced by Helber et al. (2013) to determine the expected inventory and backlogs over the planning horizon. We show that forecast evolution can be readily integrated into lot-sizing problems and solved efficiently using PLA. Further, we develop an extended stochastic model that combines the strengths of PLA and scenario methods to allow production recourse while maintaining tractable computations.

There are two methods to model the forecast evolution process according to the MMFE: additive and multiplicative. The additive model measures the difference between successive forecasts and assumes that the forecast evolution follows a multivariate normal distribution. The multiplicative model measures the ratio between successive forecasts and assumes that demand follows a log-normal distribution. While it has been argued that the multiplicative model is more relevant when demand fluctuates over time, extensive comparisons of the two MMFE models are still missing. In particular, the cost of modelling error, that is to use the additive or multiplicative model when the true process is unknown, has not been evaluated so far in the MMFE literature.

Despite the central role of data in forecast evolution models, application of MMFE to real-world case studies are rare and many questions remain open regarding the choice and tuning of the right forecast evolution model. To the best of the authors' knowledge, the application of MMFE to real-world data has only been presented by Albey et al. (2015) for the additive model and Pinçe et al. (2021) for the multiplicative model. Through a simulation study using both synthetic and real-world data, we aim to provide insights on the value of forecast evolution models and derive guidelines to choose, tune and apply MMFE models.

Our contributions revolve around the application of forecast evolution models to stochastic lot-sizing problems in rolling horizon. We extend the state of the art by (i) presenting a general method to integrate forecast evolution in lot sizing and solve the problems ef-

ficiently, (ii) identifying the value of forecast evolution models for lot-sizing problems in rolling-horizon planning, (iii) assessing the strengths and weaknesses of the additive and multiplicative models when the true forecast evolution process is unknown, (iv) developing an extended model that allows production recourse while preserving computational tractability, and (v) assessing the value of recourse in stochastic lot sizing in rolling-horizon planning.

In the following section, a brief review of related literature is presented. In Section 4.3, we introduce the additive and multiplicative MMFE and recall how to obtain the distributions of demand and cumulative demand from the forecast evolution process. In Section 4.4, the stochastic lot-sizing problem is formulated and solved using piecewise-linear approximations. We then introduce a multi-stage model that provides production recourse thanks to a scenario-based representation of demand uncertainty. In Section 4.5, we assess the value of forecast evolution models and the value of recourse through extensive rolling-horizon simulations using synthetic and real-world data. Our findings are summarised in Section 4.6, where we also provide suggestions for future research.

4.2. Literature review

In this section, we review literature on stochastic lot-sizing problems and forecast evolution models. We locate our work at the intersection of the two research streams and highlight gaps in the existing literature.

4.2.1. Stochastic lot sizing

While deterministic lot-sizing problems have been extensively studied in the academic literature (Buschkühl et al., 2010), less attention has been given to the stochastic version of the problem. Seminal works represent demand uncertainty using discrete scenario trees. Escudero et al. (1993) use a multi-stage scenario tree to model demand uncertainty and present several model formulations to allow increasing level of recourse. Brandimarte (2006) investigate the value of scenario-based stochastic lot-sizing in rolling horizon through repeated simulations. They show that scenario models are flexible and allow recourse decisions but require long computation times. Thevenin et al. (2021) use a combination of heuristics, advanced sampling techniques and rolling-horizon implementation to efficiently solve stochastic lot-sizing problems with multi-stage scenario trees in a material requirement planning context.

To improve computational performance, Helber et al. (2013) develop approximations of the expected inventory and backlog functions through piecewise-linear function and show that they outperform scenario-based formulations without recourse. Tempelmeier and Hilger (2015) and Pelt and Fransoo (2018) adapt the formulation to fill-rate service level constraints. Rossi et al. (2015) use PLA to determine the parameters of near-optimal production policies. PLA-based formulations are efficient and flexible. Sereshti et al. (2020) show that PLA can be used to formulate several types of service-level constraints in stochastic lot-sizing. De Smet et al. (2020) extend the model to include sequence-dependent changeovers in a lot-sizing and scheduling problem.

A downside of PLA methods is that they define all decision variables as first-stage decisions and do not allow production recourse. This can lead to overly conservative decisions and goes against the idea of rolling-horizon planning, which is based on a periodic update of forecasts and decisions. Tavaghof-Gigloo and Minner (2020) developed a heuristic to incorporate the replanning opportunity in lot-sizing problems by reducing the safety-stock with a replanning opportunity coefficient.

We contribute to this research stream in two ways. First, all above cited works assume that the demand distributions are known. However, in practice, these distributions are seldom available. We show that forecast evolution models can provide meaningful demand distributions from available data and be readily integrated and solved in lot-sizing with PLA. Second, we extend existing PLA formulations to allow production recourse at discrete scenarios. We combine the strengths of PLA and scenario methods to allow flexible decisions and ensure fast computation times.

4.2.2. Forecast evolution models

Since early analyses of the forecast revision process conducted by Hausman (1969) and Hausman and Peterson (1972), the MMFE has been applied to a wide variety of problems including defining supply contracts (Donohue, 2000), capacity planning (Boyacı and Özer, 2010), and inventory management (Biçer and Seifert, 2017; Iida and Zipkin, 2006; Özer and Wei, 2004; Wang et al., 2012; Wang and Tomlin, 2009). While the aforementioned works focus on determining optimal policies analytically, they do not integrate forecast evolution in rolling-horizon planning.

A second research stream studies the rolling-horizon implementation of forecast evolution models. Norouzi and Uzsoy (2014) determine key properties of the uncertain demand under additive and multiplicative MMFE and derive the optimal base-stock policy for a single-product, uncapacitated planning problem with a chance-constraint.

Albey et al. (2015) extend their work with a heuristic to solve a multi-product problem with exogenous capacity allocation. They evaluate the rolling-horizon performance of the MMFE model on a real-world case study in the semiconductor industry. Ziarnetzky et al. (2018) adapted the method to a multiplicative MMFE and evaluate it in rolling horizon with synthetic data. Albey et al. (2016) combine the model with a genetic algorithm to allocate capacity to products. They show the benefits of the improved allocation in a simulation study under additive MMFE.
We extend the research stream on MMFE by further relaxing the limiting assumptions of the model. We consider a general lot-sizing setting with multiple products, limited capacity, inventory holding costs and fixed costs for setup operations. The model does not rely on an a priori allocation of capacity and is solved to arbitrary optimality using PLA. An extended model is introduced that combines PLA and scenario-based production recourse. Further, we provide insights on the strengths and weaknesses of the additive and multiplicative MMFE, analyse the risk of forecast evolution model mismatch, and evaluate performance through rolling-horizon simulations.

4.3. Martingale model of forecast evolution

The methodology to solve the stochastic lot-sizing problem with forecast evolution contains two main parts: determining the cumulative demand probability distributions over the horizon and solving the resulting stochastic lot-sizing problem with PLA. In this section, we introduce the additive and multiplicative MMFE as formalised by Heath and Jackson (1994). For each model, we recall the probability distributions underlying the demand and cumulative demand over the planning horizon. The results on demand covariance and cumulative demand distributions are adapted from Norouzi and Uzsoy (2014). We also compare the effect of forecast update correlation on the cumulative demand covariance for the additive and multiplicative MMFE. The cumulative demand distributions obtained in this section are used to solve the stochastic lot-sizing problem in Section 4.4.

4.3.1. Problem setting

Consider the rolling-horizon planning of K products with a horizon of T periods. In each review period, updated forecasts are observed and used to calculate a production plan. Let $\mathbf{F^s} \in \mathbb{R}^{(K \times T)}$ be the forecast vector obtained at the beginning of period s given

by $\mathbf{F^s} = \begin{bmatrix} F^s_{1,1} & \dots & F^s_{1,T} & \dots & F^s_{k,t} & \dots & F^s_{K,1} & \dots & F^s_{K,T} \end{bmatrix}^\intercal$ where $F^s_{k,t}$ is the forecast of product k in period t obtained in review period s. An initial forecast vector is available in the first review period denoted by $\mathbf{F^1}$. In each review period, a new demand forecast is also obtained for the last period in the planning horizon. The MMFE models the evolution of the forecasts to the demand realisation as a stochastic process. The demand observed at the end of period s is denoted by D^s_k. After the demand has been observed, the forecast is not further updated.

4.3.2. Additive MMFE

The additive MMFE describes the evolution of the forecast vector in each review period by the relation

$$\mathbf{F^s} = \mathbf{F^{s-1}} + \varepsilon^{\mathbf{s}} \tag{4.1}$$

where the forecast update vector $\varepsilon^{\mathbf{s}}$ is observed at the beginning of review period s. The forecast update vector follows a multivariate normal distribution $\varepsilon^{\mathbf{s}} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Sigma})$. The covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{(K\times T, K\times T)}$ can be expressed as

$$\mathbf{\Sigma} = \begin{bmatrix} (\sigma^1_1)^2 & \dots & \rho^{1,T}_{1,K}\sigma^1_1\sigma^T_K \\ \dots & \rho^{t_1,t_2}_{k_1,k_2}\sigma^{t_1}_{k_1}\sigma^{t_2}_{k_2} & \dots \\ \rho^{T,1}_{K,1}\sigma^T_K\sigma^1_1 & \dots & (\sigma^T_K)^2 \end{bmatrix}$$

where σ^t_k is the standard deviation of the t-th period of the forecast updating process for product k, and $\rho^{t_1,t_2}_{k_1,k_2}$ is the correlation between the forecast update of product k_1 at time t_1 and product k_2 at time t_2. The covariance matrix describes the uncertainty of the forecast updating process over the horizon as well as the correlation between the forecast updates of different products and time periods.

Demand distribution

The demand realisation follows the same updating process as the forecast and is given by $D_{k,s} = F^s_{k,1} + \varepsilon^{s+1}_{k,1}$. In any review period s, the demand for the t-th period in the planning horizon is subject to t forecast updates. As such, the demand realisation in period $s+t-1$ as seen from period s follows the relation

$$D^{s+t-1}_k = F^s_{k,t} + \sum_{\tau=1}^{t} \varepsilon^{s+\tau}_{k,t-\tau+1}.$$

Since the forecast update vectors ε are independent and normally distributed, the demand in period $s+t-1$ follows a normal distribution $D_k^{s+t-1} \sim \mathcal{N}(F_{k,t}^s, \tilde{\sigma_{k,t}}^2)$ where $\tilde{\sigma_{k,t}}^2 = \sum_{\tau=1}^{t} (\sigma_k^\tau)^2$ is the residual uncertainty of the t-th period in the planning horizon. The residual uncertainty depends only on how far the demand period is in the planning horizon and is a direct measure of the forecast accuracy over the horizon. The demand and forecast revision process are illustrated in Figure 4.1 for three review periods.

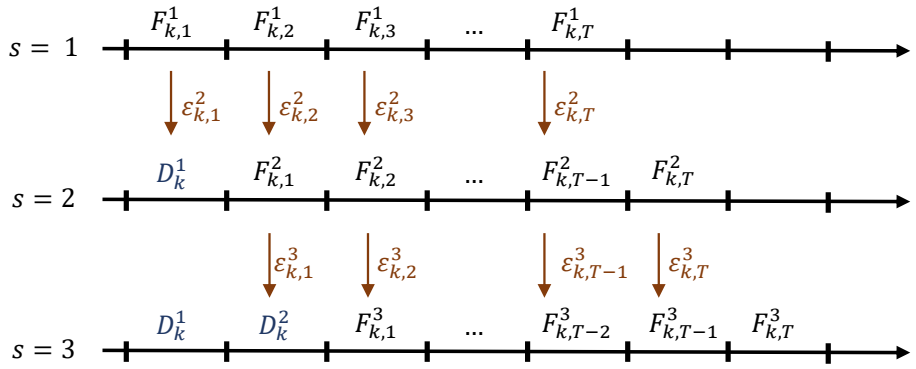

Figure 4.1.: Demand and forecast observed at three successive review periods.

Demand covariance

Although the forecast update vectors are observed independently in each review period, the forecast updating process in Equation (4.1) can exhibit correlations between the updates of different products and time periods. It follows that demand distributions of a product k in different periods of the planning horizon may be correlated. In review period s, the covariance between the demands of product k in period t_1 and t_2 of the planning horizon is given by

$$\gamma_k^{t_1,t_2} = \mathrm{Cov}\left(D_k^{s+t_1-1}, D_k^{s+t_2-1}\right) = \sum_{\tau=1}^{\min(t_1,t_2)} \rho_{k,k}^{t_1-\tau+1,t_2-\tau+1} \sigma_k^{t_1-\tau+1} \sigma_k^{t_2-\tau+1}.$$

The demand correlation depends only on how many forecast update vectors are observed in which the two periods are both in the planning horizon. The covariance between demands in different periods is necessary to determine the distribution underlying the cumulative demand.

Cumulative demand distribution

The cumulative demand of product k in period t of the planning horizon at review period s, $CD^s_{k,t} = \sum_{\tau=1}^{t} D_k^{s+\tau-1}$, is uncertain since demand is uncertain over the planning horizon. As a sum of correlated, normally distributed random variables, the cumulative demand $CD^s_{k,t}$ follows a normal distribution with mean $\sum_{\tau=1}^{t} F^s_{k,\tau}$ and variance $\sum_{t_1=1}^{t}\sum_{t_2=1}^{t} \gamma_k^{t_1,t_2}$.
The variance of the cumulative demand of product k depends only on the variance of the demands and covariance between the demands of different periods for the same product k. The variance of the cumulative demand depends linearly on the time correlation coefficient. The variance increases (resp. decrease) linearly with the forecast update correlation between time periods. The cumulative demand distribution describes the demand uncertainty over the planning horizon and allows to solve the stochastic lot-sizing problem with cumulative demand formulation introduced in Section 4.4.

4.3.3. Multiplicative MMFE

In the multiplicative model introduced by Heath and Jackson (1994), the forecast evolution process follows the relation

$$F^s_{k,t} = F^{s-1}_{k,t} \cdot \exp(\varepsilon^s_{k,t}) \tag{4.2}$$

where the forecast update vector $\varepsilon^{\mathbf{s}}$ follows a multivariate normal distribution $\varepsilon^{\mathbf{s}} \sim \mathcal{MN}(\mu, \mathbf{\Sigma})$ and each marginal distribution is given by $\varepsilon_{k,t} \sim \mathcal{N}\left(-\frac{\sigma^2_{k,t}}{2}, \sigma^2_{k,t}\right)$.
As for the additive model, the forecast updating process is unbiased. However, there is a key difference between the two models: in the additive model, forecast uncertainty depends only on the variance of the forecast update distribution, whereas in the multiplicative MMFE the uncertainty associated to a forecast update is relative to the forecast value. The variance of the forecast updating process depends both on the forecast update covariance matrix $\mathbf{\Sigma}$ and on the forecast vector $\mathbf{F^s}$. Because of these properties, the multiplicative MMFE has been described as more relevant in practice since forecasts tend to be reviewed in a relative manner. The multiplicative model can also be suitable when demand has important fluctuations over time since relative forecast updates remain of similar magnitude. Yet, there remain many open questions on how to apply the multiplicative model from available forecast and demand data. In the numerical study in Section 4.5, we detail the estimation process of the multiplicative model from data

and assess its expected and actual performance in rolling-horizon planning.

Demand distribution

The demand of product k in each review period s follows the same relation as the forecast update so that $D_k^s = F_{k,1}^s \cdot \exp(\varepsilon_{k,1}^{s+1})$. From this relation and Equation (4.2), the demand in period $s+t-1$ as seen from review period s is given by

$$D_k^{s+t-1} = F_{k,t}^s \cdot \exp\left(\sum_{\tau=1}^{t} \varepsilon_{k,t-\tau+1}^{s+\tau}\right).$$

The demand in period $s+t-1$ follows a log-normal distribution

$$\log\left(D_k^{s+t-1}\right) \sim \mathcal{N}\left(\log\left(F_{k,t}^s\right) - \frac{\tilde{\sigma_{k,t}}^2}{2}, \tilde{\sigma_{k,t}}^2\right)$$

where $\tilde{\sigma_{k,t}}^2 = \sum_{\tau=1}^{t} (\sigma_k^\tau)^2$ is the residual uncertainty of the t-ahead period. As for the additive model, the residual uncertainty in the log domain is independent of the review period. However, demand variance depends on both the forecast update variance and the value of the forecast.

Demand covariance

The demands of product k in period t_1 and t_2 of the planning horizon in review period s are correlated with covariance

$$\gamma_k^{t_1,t_2} = \mathrm{Cov}\left(\log\left(D_k^{s+t_1-1}\right), \log\left(D_k^{s+t_2-1}\right)\right) = \sum_{\tau=1}^{\min(t_1,t_2)} \rho_{k,k}^{t_1-\tau+1,t_2-\tau+1} \sigma_k^{t_1-\tau+1} \sigma_k^{t_2-\tau+1}.$$

The demand covariance can be deduced similarly as for the additive case by analysing the covariance of the forecast evolution process in the log domain. The covariance of the demand periods is used to estimate the parameters of the distribution underlying the cumulative demand.

Cumulative demand distribution

Contrary to the additive case, there is no closed-form expression for the cumulative demand since it is the sum of correlated log-normal distributions. However, it has been observed that the sum of log-normal distributions can be well approximated by a

log-normal distribution. To estimate the cumulative demand distributions with multiplicative MMFE, we follow the approach of Norouzi and Uzsoy (2014) and apply the Fenton-Wilkinson approximation (FWA). The method is attractive because of its computational simplicity and overall high approximation quality over a wide range of parameters. The approximation is based on matching the first two moments of the approximating log-normal distribution with the moments of the sum of the correlated log-normal distributions (Abu-Dayya and Beaulieu, 1994).
Following the moment-matching approximation, the cumulative demand $CD_{k,t}$ approximately follows a log-normal distribution, $\log(CD_{k,t}) \sim \mathcal{N}(m_{k,t}, v_{k,t})$, with parameters $m_{k,t} = 2\log(u_1) - \frac{1}{2}\log(u_2)$ and $v_{k,t} = \log(u_2) - 2\log(u_1)$ where $u_1 = \sum_{\tau=1}^{t} F_{k,\tau}$ and

$$u_2 = \sum_{\tau=1}^{t}(F_{k,\tau})^2 \exp\left(\tilde{\sigma_{k,\tau}}^2\right) + 2\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} F_{k,i}F_{k,j} \exp\left(\sum_{\tau=1}^{\min(i,j)} \rho_{k,k}^{i-\tau+1,j-\tau+1}\sigma_k^{i-\tau+1}\sigma_k^{j-\tau+1}\right).$$

This approximate cumulative demand distribution is used in Section 4.4 to solve the stochastic lot-sizing problem.

4.3.4. Influence of forecast update correlation on variance of cumulative demand

The variance of the cumulative demand has been shown to depend linearly on the forecast update correlation for the additive model. In the multiplicative model, although the relation between the forecast update correlation and the cumulative demand variance appears exponential, it is approximately linear over the relevant domain.

Proposition 4.1. *Under multiplicative MMFE, the variance of the cumulative demand of product k in period t, $Var(CD_{k,t})$, is approximately linear in the forecast update correlation $\rho_{k,k}^{t_1,t_2}$ for $t_1, t_2 \leq t$ with slope given by*

$$\frac{\partial\, Var(CD_{k,t})}{\partial \rho_{k,k}^{t_1,t_2}} \approx 2\sigma_k^{t_1}\sigma_k^{t_2} \sum_{i=1}^{t-t_2+1} F_{k,t_1+i-1}F_{k,t_2+i-1}.$$

Proof. The variance of the cumulative demand is given by

$$\mathrm{Var}(CD_{k,t}) = u_2 - (u_1)^2 = \sum_{\tau=1}^{t}(F_{k,\tau})^2 \exp\left(\tilde{\sigma_{k,\tau}}^2\right)$$

$$+ 2\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} F_{k,i}F_{k,j}\exp\left(\sum_{\tau=1}^{\min(i,j)} \rho_{k,k}^{i-\tau+1,j-\tau+1}\sigma_k^{i-\tau+1}\sigma_k^{j-\tau+1}\right) - \left(\sum_{\tau=1}^{t} F_{k,\tau}\right)^2.$$

It can be expressed as

$$\mathrm{Var}(CD_{k,t}) = \alpha + 2\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} F_{k,i}F_{k,j}\exp\left(\sum_{\tau=1}^{\min(i,j)} \rho_{k,k}^{i-\tau+1,j-\tau+1}\sigma_k^{i-\tau+1}\sigma_k^{j-\tau+1}\right)$$

, where α is independent of $\rho_{k,k}^{t_1,t_2}$. Clearly, if $t_1 > t$ or $t_2 > t$, the variance is independent of the correlation coefficient $\rho_{k,k}^{t_1,t_2}$. Without loss of generality, we set $t_1 < t_2 \leq t$ and deduce

$$\mathrm{Var}(CD_{k,t}) = \alpha + 2\sum_{i=1}^{t-t_2+1} F_{k,t_1+i-1}F_{k,t_2+i-1}\exp\left(\sum_{\tau=1}^{t_1+i-1} \rho_{k,k}^{t_1+i-\tau,t_2+i-\tau}\sigma_k^{t_1+i-\tau}\sigma_k^{t_2+i-\tau}\right),$$

which can be further simplified as

$$\mathrm{Var}(CD_{k,t}) = \alpha + 2\sum_{i=1}^{t-t_2+1} F_{k,t_1+i-1}F_{k,t_2+i-1}\exp\left(\beta_i + \rho_{k,k}^{t_1,t_2}\sigma_k^{t_1}\sigma_k^{t_2}\right)$$

where β_i is independent of $\rho_{k,k}^{t_1,t_2}$. Since the covariance parameters of the multiplicative MMFE are small, the variance of the cumulative demand is well approximated by its Taylor expansion as $\mathrm{Var}(CD_{k,t}) \approx \alpha + 2\sum_{i=1}^{t-t_2+1} F_{k,t_1+i-1}F_{k,t_2+i-1}\left(\beta_i + \rho_{k,k}^{t_1,t_2}\sigma_k^{t_1}\sigma_k^{t_2}\right)$. □

Proposition 4.1 states that the variance of the cumulative demand depends linearly on the forecast update correlation between two time periods of the same product. This implies that ignoring the correlation between demand periods can lead to under- (resp. over-) estimation of the cumulative demand variance if the correlation is positive (resp. negative). The effect of the correlation coefficient is proportional not only to the variance but also to the forecast values. Thus, ignoring correlation has a greater impact for large forecast values.

We illustrate the evolution of the variance of the cumulative demand distribution with the forecast update correlation and compare additive and multiplicative MMFE. We consider a single product planned over a horizon of $T = 2$ periods and investigate the effect of forecast update correlation on the cumulative demand CD^2. The forecast updating process is defined with standard deviation $\sigma^1 = \sigma^2 = 20$ for the additive model and $\sigma^1 = \sigma^2 = 0.2$ for the multiplicative model. The initial forecast in periods 1 and 2

are set equal $F^1 = F^2$ and chosen within the set $\{50, 100, 150\}$.

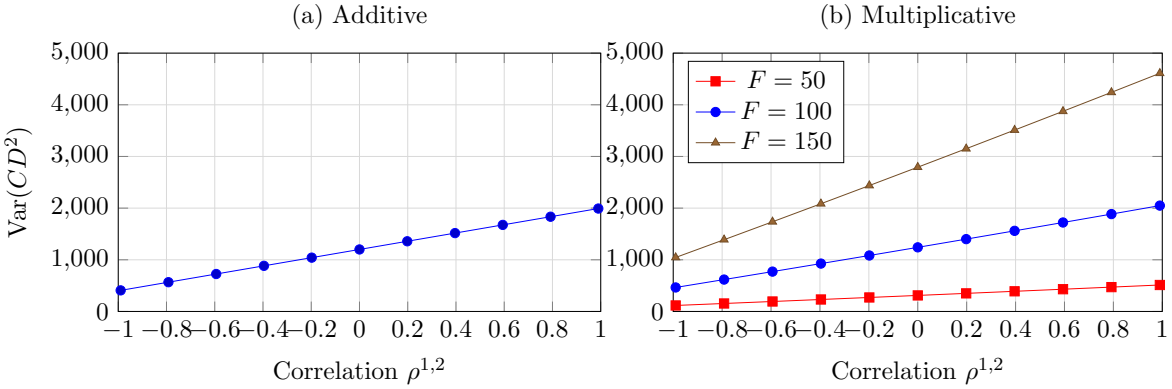

Figure 4.2.: Evolution of variance with correlation coefficient for (a) additive and (b) multiplicative MMFE.

Figure 4.2 shows the evolution of the variance of the cumulative demand in period 2 with varying correlation for the additive and multiplicative models. The figure highlights the linear relationship between the forecast update correlation and the variance of the cumulative demand for both the additive and multiplicative cases. It further illustrates the impact of the forecast value on the variance of the cumulative demand for the multiplicative model.

4.3.5. Summary

In this section, the multivariate forecast evolution process has been introduced for additive and multiplicative MMFE. The parameters of the resulting demand and cumulative demand distributions have been obtained. The cumulative demand distributions can be determined exactly for the additive model and approximately for the multiplicative model. Finally, we have analysed the dependency of the cumulative demand variance on the forecast update correlation coefficient. In the next section, we derive efficient formulations for the stochastic lot-sizing problem based on the cumulative demand distributions estimated from the MMFE models.

4.4. Stochastic lot sizing

We integrate the additive and multiplicative MMFE in lot-sizing problems through the cumulative demand distributions described in the previous section. We introduce the

PLA to solve the problem efficiently and extend the model with scenario-based production recourse. The extended model combines the strengths of PLA and scenario methods, providing fast computations and flexible decisions.

4.4.1. Non-linear stochastic lot-sizing formulation based on cumulative demand

Consider the planning of K products over a horizon of T periods. The planner aims to satisfy the uncertain demand while minimising costs. The operational costs include inventory costs incurred at the end of each period with unit cost hc_k for product k and setup costs incurred each time a new product is set up with per-unit cost sc_k. Unsatisfied demand is backordered and penalised with per-unit cost bc_k. The products share the same equipment with limited capacity *cap* in each period.
In each review period, the planner determines the production quantity $Q_{k,t}$ for all products over the horizon. Since demand is uncertain, the inventory $I_{k,t}$ and backlog $B_{k,t}$ at the end of each period are random variables. The initial inventory is denoted by in_k^0 and can be positive or negative depending on whether there is on-hand inventory or backlogs. The stochastic lot-sizing problem that minimises expected costs is a mixed-integer non-linear problem given by

$$\min \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\left(hc_k \cdot \mathbb{E}[I_{k,t}] + bc_k \cdot \mathbb{E}[B_{k,t}] + sc_k \cdot X_{k,t}\right) \tag{4.3a}$$

$$\text{s.t.} \quad \mathbb{E}[I_{k,t}] = \mathbb{E}\left[\max\left(in_k^0 + \sum_{\tau=1}^{t} Q_{k,\tau} - \sum_{\tau=1}^{t} D_k^\tau, 0\right)\right], \forall k, t, \tag{4.3b}$$

$$\mathbb{E}[B_{k,t}] = \mathbb{E}\left[\max\left(\sum_{\tau=1}^{t} D_k^\tau - in_k^0 - \sum_{\tau=1}^{t} Q_{k,\tau}, 0\right)\right], \forall k, t, \tag{4.3c}$$

$$\sum_{k=1}^{K} Q_{k,t} \leq cap, \qquad \forall t, \tag{4.3d}$$

$$Q_{k,t} \leq cap \cdot X_{k,t}, \qquad \forall k, t, \tag{4.3e}$$

$$Q_{k,t} \geq 0, \qquad \forall k, t, \tag{4.3f}$$

$$X_{k,t} \in \{0; 1\}, \qquad \forall k, t. \tag{4.3g}$$

The objective function in (4.3a) minimises the expected costs of inventory, backlogs and setup for all products over the horizon. Constraints (4.3b) and (4.3c) determine the

expected inventory and backlogs at the end of each period as a function of the uncertain cumulative demand. Constraint (4.3d) ensures that the production over all products is limited by the available capacity in each period. Constraint (4.3e) states that production of a product can occur only if a setup operation is conducted. Constraints (4.3f) and (4.3g) describe the domain of positive and binary variables respectively.
Since the expected inventory and backlog in constraints (4.3b) and (4.3c) depend on the production quantity, Problem (4.3) is non-linear and cannot be solved directly. However, it has been shown that the expected inventory and backlog function could be well approximated by piecewise-linear functions.

4.4.2. Stochastic lot-sizing model with PLA

The PLA method is based on evaluating the first-order loss function at a selected number of breakpoints and determining the slope of the expected inventory and backlog between the breakpoints. The first-order loss function of a real variable x and random variable ω with p.d.f. ϕ and c.d.f. Φ is defined as

$$\mathcal{L}(x,\omega) = \mathbb{E}\left[\max(\omega - x, 0)\right] = \int_x^{+\infty} \max(t-x,0)\cdot\phi(t)dt = \int_x^{+\infty} (1-\Phi(t))\,dt \quad (4.4)$$

Let $\mathbf{u} = (u_{k,t,l})$ be the set of $L+1$ breakpoints determined independently for each product and time period. The first breakpoint is set to $u_{k,t,0} = in_k^0$, which can be either positive or negative, and the last breakpoint is set to the highest inventory position attainable at the end of period t with full capacity utilisation as $u_{k,t,L} = in_k^0 + cap \cdot t$. The remaining breakpoints are set uniformly between these two bounds. For each segment, the slope of the expected inventory and backlog can be determined as

$$\Delta^l_{B_{k,t}} = \frac{\mathcal{L}\left(u_{k,t,l+1}, CD_{k,t}\right) - \mathcal{L}\left(u_{k,t,l}, CD_{k,t}\right)}{u_{k,t,l+1} - u_{k,t,l}} \quad (4.5)$$

$$\Delta^l_{I_{k,t}} = \frac{\mathcal{L}\left(u_{k,t,l+1}, CD_{k,t}\right) + u_{k,t,l+1} - \mathcal{L}\left(u_{k,t,l}, CD_{k,t}\right) - u_{k,t,l}}{u_{k,t,l+1} - u_{k,t,l}} \quad (4.6)$$

where $CD_{k,t}$ is the cumulative demand distribution of product k in period t as determined in Section 4.3. Calculating the slopes of the L segments of the expected inventory and backlog requires evaluating the first-order loss function $K \cdot T \cdot (L+1)$ times at each review period. This evaluation is computationally very cheap for a normal probability distribution since the first-order loss function of a normal variable can be expressed as a function of the first-order loss function of a standard normal (Rossi et al., 2014), and can

thus be calculated offline. The calculation is more intensive for a log-normal variable since it requires evaluating many integrals as in Equation (4.4). Note also that the domain of the c.d.f. of a log-normal variable needs to be extended for negative values since the initial inventory in each period may be negative. The PLA of two demand distributions following a normal and log-normal with equal mean and similar variance is shown on Figure 4.3. The figure shows that the expected inventory and backlog functions can be well approximated with only $L = 6$ segments for both distributions.

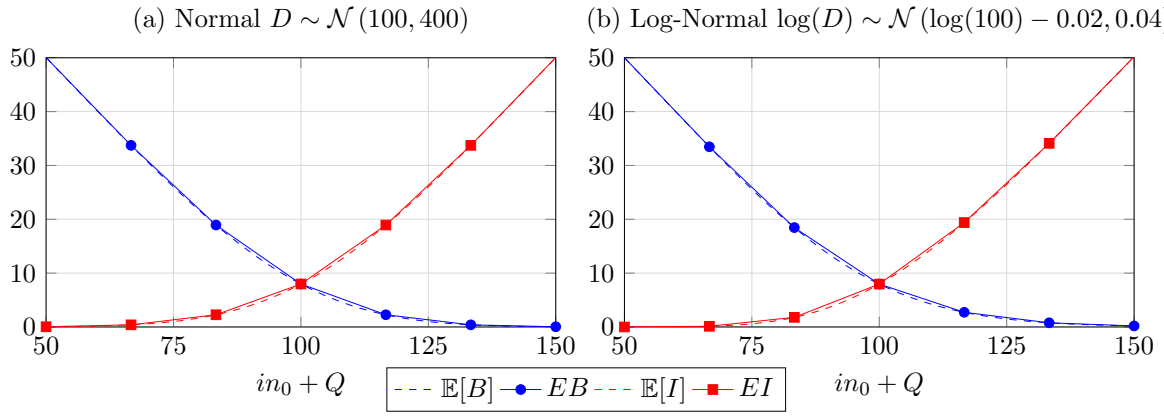

Figure 4.3.: Piecewise-linear approximation of expected inventory and backlog for demand following (a) normal distribution and (b) log-normal distribution.

The stochastic lot-sizing problem solved using PLA approximates the expected inventory and backlog with variables $EI_{k,t}$ and $EB_{k,t}$ respectively. The formulation requires the introduction of auxiliary variables $w_{k,t,l}$ to measure the cumulative production from period 1 to t associated to segment l and binary auxiliary variables $\lambda_{k,t,l}$ to ensure that the L segments are used consecutively. The formulation is adapted from Pelt and Fransoo (2018) to penalty cost for backlogs, and is given by

$$\min \quad \sum_{t=1}^{T} \sum_{k=1}^{K} (hc_k \cdot EI_{k,t} + bc_k \cdot EB_{k,t} + sc_k \cdot X_{k,t}) \tag{4.7a}$$

$$\text{s.t.} \quad EI_{k,t} = \Delta^0_{I_{k,t}} + \sum_{l=1}^{L} \left(\Delta^l_{I_{k,t}} \cdot w_{k,t,l} \right), \ \forall k, t, \tag{4.7b}$$

$$EB_{k,t} = \Delta^0_{B_{k,t}} + \sum_{l=1}^{L} \left(\Delta^l_{B_{k,t}} \cdot w_{k,t,l} \right), \forall k, t, \tag{4.7c}$$

$$\sum_{l=1}^{L} (w_{k,t,l} - w_{k,t-1,l}) = Q_{k,t}, \qquad \forall k, t, \tag{4.7d}$$

$$w_{k,t,l-1} \geq (u_{k,t,l-1} - u_{k,t,l-2})\,\lambda_{k,t,l}, \qquad \forall k,t,l \geq 2, \tag{4.7e}$$

$$w_{k,t,l} \leq (u_{k,t,l} - u_{k,t,l-1})\,\lambda_{k,t,l}, \qquad \forall k,t,l, \tag{4.7f}$$

$$X_{k,t}, \lambda_{k,t,l} \in \{0;1\}, \qquad \forall k,t,l, \tag{4.7g}$$

Constraints (B.1d) – (B.1f)

Constraints (4.7b) and (4.7c) approximate the expected inventory and backlog using the slopes of the first-order loss function previously determined. Constraint (4.7d) determines the production volume from the cumulative production over the linearisation segments. Constraints (4.7e) and (4.7f) ensure that the linearisation segments are used in increasing order thank to the auxiliary variable $\lambda_{k,t,l}$.

4.4.3. Extended lot-sizing formulation with PLA and production recourse

The stochastic lot-sizing formulation in (4.7) provides significant computational improvements compared to traditional scenario-based stochastic formulations. However, it ignores that the planner has the opportunity to react to forecast updates in each review period. More precisely, Problem (4.7) defines all production decisions as first-stage decision variables, which can lead to overly conservative decisions.
Scenario trees can model multi-stage stochastic processes allowing recourse decisions. However, they require notoriously long computation times that grow exponentially with the problem size and scenario tree. The main idea of our extension is to combine PLA and scenario trees in a single model to allow fast computations and flexible recourse decisions. In this section, we describe the integration of the two methods, build multi-stage scenario trees from the MMFE models, and formulate the extended model.

Combining PLA and scenario-based recourse

The extended model combines PLA and scenarios over the planning horizon. The first periods are modelled with PLA so that all production decisions are set as first-stage variables. This provides accurate approximation of the expected inventory and backlogs over the short-term horizon. In parallel, a multi-stage scenario tree is created to describe the demand and forecast uncertainty over the planning horizon. Applying the first-stage decisions from PLA, several inventory and forecast positions are reached when following the scenario tree. The multi-stage scenario tree allows recourse in each of the inventory/forecast positions to react to the different situations created by the first-stage

decisions. Because of the recourse opportunities in later periods, the model can take less conservative first-stage decisions in the short-term horizon. Formally, we define $t_b \in \{1, \dots T\}$ such that PLA is applied from period 1 to t_b, and scenario recourse is applied from period $t_b + 1$ to T. Clearly, $t_b = 1$ and $t_b = T$ reduce to a multi-stage scenario lot-sizing formulation and the PLA model in (4.7) respectively.

There are two main motivations for applying first PLA and then scenario trees with recourse. First, since the scenario tree grows exponentially over the horizon, and the precision of scenario tree increases with the number of scenarios, this decomposition allows PLA to provide high accuracy on periods with few scenarios. Second, since recourse production decisions reduce the visibility of the planner as there is no single reference plan, the proposed method ensures the availability of a reference plan over the short-term horizon, which is often indispensable. The trade-off between flexibility and visibility is adjusted through the parameter t_b.

The scenario-based extension of the PLA model can be seen as an approximation of the optimal production policy that would be obtained if the corresponding dynamic programming model could be solved. The scenario part of the model acts as a look-ahead approximation of the optimal policy (Powell, 2016). In this sense, the scenario part approximates the true problem in which decisions are revised in each period, whereas PLA provides the solution of a more conservative version of the problem in which there is no production recourse. Our approach combines the two methods efficiently to allow fast computations and good approximation of the optimal policy.

The combination of PLA and scenario-based recourse is illustrated in Figure 4.4. The multi-stage scenario tree is generated with a branching factor of 2 over a planning horizon of $T = 6$ periods with $t_b = 3$. Thus, there are $[2, 4, 8, 16, 32, 64]$ demand scenarios, $[1, 2, 4, 8, 16, 32, 64]$ inventory positions, and $[1, 1, 1, 8, 16, 32]$ production decisions over the horizon.

Generating scenario trees from forecast evolution

The demand scenario tree is generated from the MMFE from period 1 to T by updating the initial forecast $F_{k,t}$ with forecast updates vectors sampled in each node. The forecast update vectors are drawn from the multivariate forecast evolution distribution. The forecast of product k in period t of the horizon in scenario node n can be expressed as $F^n_{k,t} = F_{k,t} + \sum_{\tau=0}^{t-1} \varepsilon^{a_\tau(n)}_{k,t-\tau}$ for the additive MMFE and $F^n_{k,t} = F_{k,t} \cdot \exp\left(\sum_{\tau=0}^{t-1} \varepsilon^{a_\tau(n)}_{k,t-\tau}\right)$ for the multiplicative MMFE where $\varepsilon^{a_\tau(n)}$ is the forecast update vector obtained at node $a_\tau(n)$, the τ-th ancestor node of node n with $a_0(n) = n$.

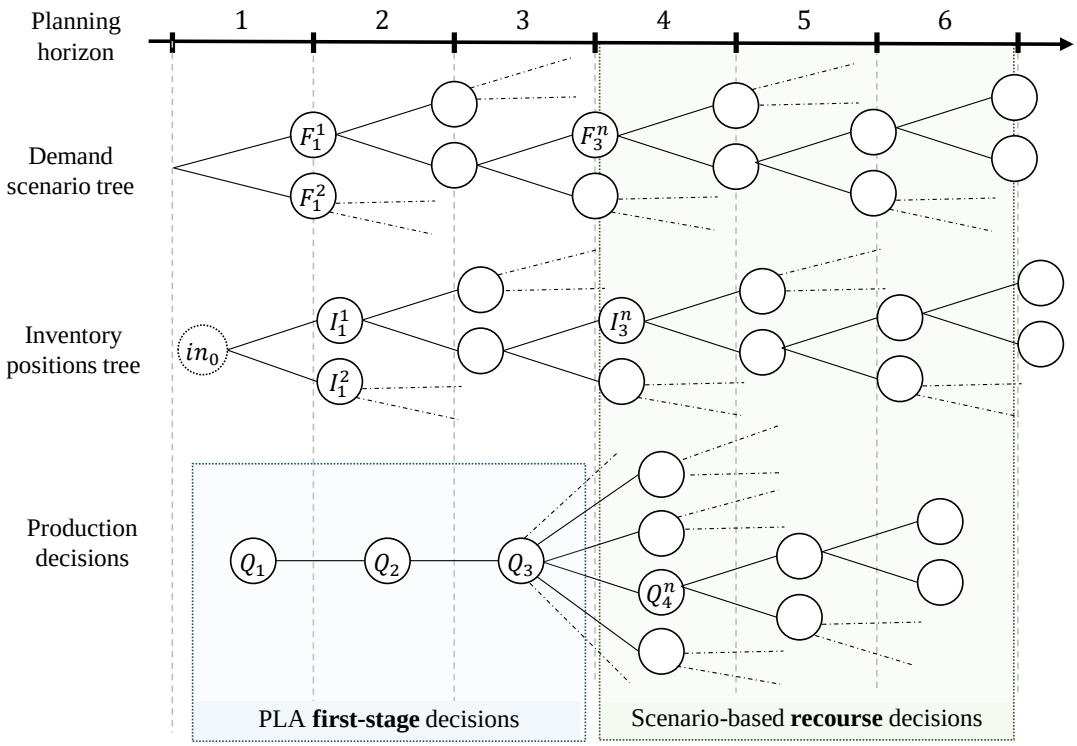

Figure 4.4.: Demand realisations, production decisions and inventory trajectories over $T = 6$ periods with $t_b = 3$.

Advanced sampling techniques have been extensively studied for sampling univariate distributions but are less commonly applied to multivariate distributions. To sample the high-dimensional, correlated forecast update vectors in each node, we apply the Latin Hypercube with multivariate uniformity (LHMU) method developed by Deutsch and Deutsch (2012). The method is designed to reduce sampling variability and increase uniformity over all dimensions by applying Latin Hypercube on deterministic strata.

Extended lot-sizing formulation

The extended stochastic lot-sizing formulation with PLA and scenario-based recourse is given by

$$\min \quad \sum_{k=1}^{K}\sum_{t=1}^{t_b}\left(hc_k \cdot EI_{k,t} + bc_k \cdot EB_{k,t}\right) + \sum_{t=t_b+1}^{T}\left(\frac{hc_k}{N}\cdot\sum_{n=1}^{N} I_{k,t,n} + \frac{bc_k}{N}\cdot\sum_{n=1}^{N} B_{k,t,n}\right) + \sum_{t=1}^{T} sc_k \cdot X_{k,t} \tag{4.8a}$$

s.t.
$$EI_{k,t} = \Delta^0_{H_{k,t}} + \sum_{l=1}^{L} (\Delta^l_{H_{k,t}} w_{k,t,l}), \qquad \forall k, t \leq t_b, \tag{4.8b}$$
$$EB_{k,t} = \Delta^0_{B_{k,t}} + \sum_{l=1}^{L} (\Delta^l_{B_{k,t}} w_{k,t,l}), \qquad \forall k, t \leq t_b, \tag{4.8c}$$
$$\sum_{l=1}^{L} (w_{k,t,l} - w_{k,t-1,l}) = Q_{k,t}, \qquad \forall k, t \leq t_b, \tag{4.8d}$$
$$w_{k,t,l-1} \geq (u_{k,t,l-1} - u_{k,t,l-2}) \lambda_{k,t,l}, \qquad \forall k, t \leq t_b, l \geq 2, \tag{4.8e}$$
$$w_{k,t,l} \leq (u_{k,t,l} - u_{k,t,l-1}) \lambda_{k,t,l}, \qquad \forall k, t \leq t_b, l, \tag{4.8f}$$
$$\sum_{k=1}^{K} Q_{k,t} \leq cap, \qquad \forall t \leq t_b, \tag{4.8g}$$
$$Q_{k,t} \leq cap \cdot X_{k,t}, \qquad \forall k, t \leq t_b, \tag{4.8h}$$
$$I_{k,t,n} - B_{k,t,n} = I_{k,t-1,n} - B_{k,t-1,n} + Q_{k,t} - F^n_{k,t}, \quad \forall k, t \leq t_b, \tag{4.8i}$$
$$I_{k,t,n} - B_{k,t,n} = I_{k,t-1,n} - B_{k,t-1,n} + Q_{k,t,n} - F^n_{k,t}, \forall k, t > t_b, \tag{4.8j}$$
$$\sum_{k=1}^{K} Q_{k,t,n} \leq cap, \qquad \forall n, t > t_b, \tag{4.8k}$$
$$Q_{k,t,n} \leq cap \cdot X_{k,t}, \qquad \forall k, n, t > t_b, \tag{4.8l}$$
$$Q_{k,t,n} \text{ non-anticipative}, \tag{4.8m}$$
$$Q_{k,t} \geq 0, \qquad \forall k, t, \tag{4.8n}$$
$$Q_{k,t,n}, I_{k,t,n}, B_{k,t,n} \geq 0, \qquad \forall k, t, n, \tag{4.8o}$$
$$X_{k,t}, \lambda_{k,t,l} \in \{0; 1\}, \qquad \forall k, t, l. \tag{4.8p}$$

The objective function in (4.8a) minimises the PLA expected inventory and backlog costs over the first t_b periods and the sample average inventory and backlog costs over the remaining $T - t_b + 1$ periods. Constraints (4.8b) to (4.8h) are adapted from the PLA model to the first t_b periods. Constraints (4.8i) and (4.8j) describe the discrete inventory positions through the planning horizon with first-stage and recourse production decisions respectively. Note that the discrete inventory and backlog determined in Constraint (4.8i) are not used in the objective function. They determine sample inventory positions resulting from the first-stage decisions, at which the scenario model starts to be applied. Constraint (4.8m) describe the so-called *non-anticipative* structure of the recourse decisions, which ensures that production decisions in a certain time period cannot use information obtained in later periods, as is also illustrated on Figure 4.4.

4.4.4. Summary

In this section, we have presented a general stochastic lot-sizing formulation and a high-quality approximation based on piecewise-linear functions. The model was extended to production recourse through a discrete scenario tree. The forecast evolution models presented in Section 4.3 are integrated in the lot-sizing problem through the cumulative demand and forecast evolution distributions. We have shown that both additive and multiplicative MMFE models can be readily included in lot-sizing problems through PLA and that additional flexibility can be provided through recourse decisions.

4.5. Numerical study

The numerical study investigates several questions relating to the use of forecast evolution models in practice from model estimation to its application. We aim to provide a fair assessment of forecast evolution models by investigating

- how can MMFE model parameters be estimated from real data? How sensitive are they to data?
- What are the advantages and weaknesses of the additive and multiplicative MMFE? What are the risks of using a misspecified MMFE model?
- What is the value of forecast evolution models in practice?
- What is the value of recourse provided by multi-stage formulations in rolling-horizon planning and what factors influence it?

The numerical study is composed of two parts based on synthetic and real-world data respectively. First, we construct forecast evolution distributions for the additive and multiplicative models and specify their parameters. As the forecast evolution process is fully known, this ideal situation allows us to assess the cost of modelling error. We evaluate the risk of using the additive model when the actual forecast evolution follows a multiplicative model and conversely. Further, we quantify the value of recourse for the MMFE model with known forecast evolution process. Sensitivity analyses are set up to identify parameters that drive the performance of forecast evolution models including capacity, forecast update covariance, and demand pattern. In a second part, we solve the real-world case study of a global company in the process industries. A large data set of forecast and demand history is used to estimate the MMFE models and assess their performance. Simulation are run in an *out-of-sample* setting in which the forecast evolution process is unknown and can only be estimated with past historical data. We

conclude with general recommendations on the use of MMFE models.
The numerical study is implemented in Julia, a fast scientific language with a large environment of modules (Bezanson et al., 2017). The optimisation problems are modelled in JuMP (Dunning et al., 2017) and solved with Gurobi 9.0. The simulations are run on an Intel(R) Core(TM) i7-4810MQ processor at 2.80Ghz using 16GB of RAM. The code used to produce all results and figures using synthetic data in this paper is made publicly available on the online repository.

4.5.1. Synthetic data

We consider the rolling-horizon planning of $K = 2$ products over a prediction horizon of $T = 6$ periods. Each simulation contains $S = 12$ review periods. The inventory holding cost is sampled randomly for the two products as $h_c \sim \mathcal{U}[1, 1.5]$. The backlog cost is set to $b_c = 10 \cdot h_c$ and the setup cost is set to $s_c = 150$. The initial inventory is set to $in^0 = 50$ for each product.

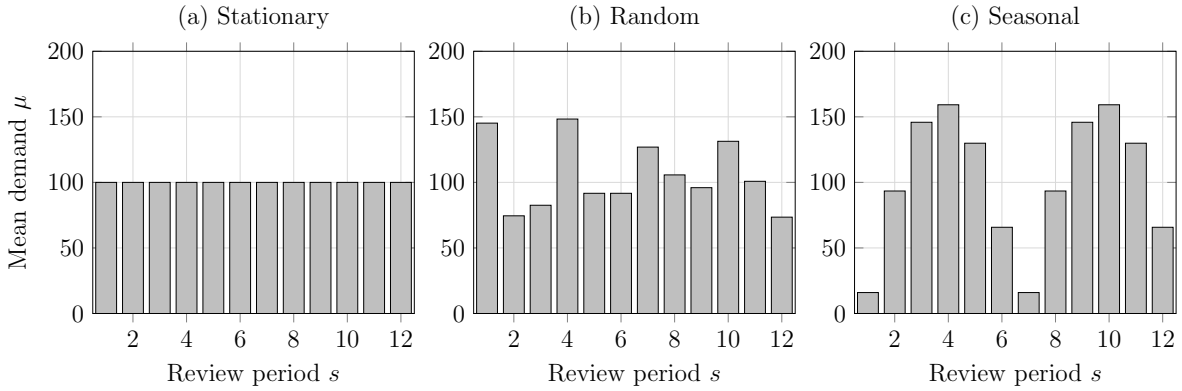

Figure 4.5.: Mean demand for (a) stationary, (b) random, and (c) seasonal patterns over simulation of 8 periods.

First, we perform repeated rolling-horizon simulations under varying capacity, uncertainty and demand pattern. The capacity in each period is set to $cap \in \{300, 500\}$ to consider limited and ample capacity settings. The initial forecast values are generated from three demand patterns with different dynamics. Stationary, random and seasonal patterns are used as illustrated in Figure 4.5. The stationary pattern sets the initial forecasts to a constant value $F = 100$ over the simulation length. The random pattern samples each forecast from a uniform distribution $\mathcal{U}[50, 150]$. The seasonal pattern is generated with periodicity $S/2$ periods from a sinusoid function with values $[16, 93, 145, 159, 129, 65]$ over the season length. The three demand patterns have similar

average demand over the simulation length but different dynamics, which may impact the additive and multiplicative MMFE models differently.

MMFE models: known and mismatched

For each demand pattern, we run two distinct sets of simulations in which the true forecast evolution process follows the additive and multiplicative MMFE respectively. The forecast evolution models are set unbiased and uncorrelated with equal forecast update variance for all products and time periods. Low, medium and high forecast uncertainty settings are defined with variance $\sigma^2 \in \{100, 400, 700\}$ for the additive model and $\sigma^2 \in \{0.01, 0.04, 0.07\}$ for the multiplicative model.

In practice, it is unknown whether the forecast evolution follows an additive or multiplicative model and practitioners have to decide a priori what is the most suitable MMFE model. To estimate the risks associated with using a mismatched forecast evolution model, we simulate the forecast evolution process over 1 million review periods for each MMFE model and demand pattern. The sampled forecast updates are then measured according to the mismatched MMFE model and used to estimate its distribution parameters.

While the procedure is straightforward for the additive model, it is less consistent for the mismatched multiplicative model. When the forecast evolution process follows an additive MMFE, demand is normally distributed and there is a positive probability that demand is zero or negative. In simulations with additive MMFE, we truncate the distributions so that demand cannot be negative. It is also possible that a forecast with zero value is updated to a positive forecast. Although these two cases realistically occur in practical settings, they are incompatible with the relative forecast measure of the multiplicative MMFE. Thus, when estimating the parameters of the multiplicative MMFE we remove all sampled forecast updates in which the initial or updated forecasts for at least one product in a time period are zero. For the setting with low uncertainty, this amounts to removing 0%, 1% and 41% of samples for the stationary, random and seasonal patterns respectively, 13%, 26% and 64% for the medium uncertainty setting, and 37%, 50% and 74% for the high uncertainty setting. Clearly, more sample updates are removed from the data set as the demand pattern is more dynamic and as uncertainty increases. The high number of unusable samples is an important shortcoming of multiplicative MMFE. Collecting data is often an expensive process and it is highly undesirable to discard such large portions of the data set.

Results

A full-factorial sensitivity analysis of the forecast evolution models to the demand patterns, uncertainty levels, and capacity settings is performed. The PLA model is implemented with $L = 40$ segments. The extended model uses a scenario tree with $[3, 6, 12, 24, 48, 48]$ nodes over the planning horizon sampled with LHMU. We set $t_b = 3$ so that the first half of the planning horizon is modelled with PLA and the second half with scenario recourse. Model performance are measured as the sum of realised inventory, backlog and setup costs. Each of the 36 simulation setting is repeated 1000 times.

Table 4.1.: Average simulation results over all configurations for additive MMFE.

Demand	Uncert.	*cap*	Det.	Additive PLA	Multiplicative PLA	Ext. Add. PLA
Stationary	Low	300	4701.1	4158.9 (88.8% (*))	4127.3 (88.2% (*))	4122.1 (88.1% (*))
		500	4664.4	4169.6 (89.8% (*))	4145.7 (89.3% (*))	4137.4 (89.1% (*))
	Medium	300	5247.7	4641.0 (89.4% (*))	5129.3 (99.0% (*))	4571.2 (88.1% (*))
		500	5068.1	4589.7 (91.3% (*))	4992.3 (99.4%)	4530.9 (90.1% (*))
	High	300	5896.5	4930.7 (85.7% (*))	6712.2 (116.7% (*))	4848.2 (84.1% (*))
		500	5456.2	4821.6 (89.5% (*))	5794.9 (107.7% (*))	4740.0 (88.0% (*))
Random	Low	300	4585.3	4017.8 (88.1% (*))	4110.1 (90.1% (*))	3982.1 (87.3% (*))
		500	4511.9	3981.2 (88.7% (*))	4050.4 (90.2% (*))	3945.4 (87.9% (*))
	Medium	300	5301.2	4571.8 (87.5% (*))	5622.0 (107.6% (*))	4484.9 (85.8% (*))
		500	5052.8	4468.8 (89.3% (*))	5174.6 (103.5% (*))	4406.5 (88.0% (*))
	High	300	5990.8	4921.4 (84.2% (*))	7299.0 (125.2% (*))	4823.0 (82.5% (*))
		500	5532.4	4764.1 (87.2% (*))	6038.6 (110.7% (*))	4675.9 (85.6% (*))
Seasonal	Low	300	4611.3	3973.7 (87.0% (*))	5789.2 (127.0% (*))	3892.7 (85.2% (*))
		500	4194.7	3716.6 (89.1% (*))	4567.0 (109.7% (*))	3671.8 (88.1% (*))
	Medium	300	5714.1	4569.0 (82.4% (*))	8238.2 (149.4% (*))	4485.1 (80.8% (*))
		500	4951.6	4238.9 (86.5% (*))	5986.6 (122.4% (*))	4137.2 (84.4% (*))
	High	300	6554.3	5095.4 (82.0% (*))	9704.7 (158.6% (*))	4973.5 (79.6% (*))
		500	5431.7	4526.7 (84.7% (*))	8029.5 (150.6% (*))	4447.0 (83.2% (*))
Average			5192.6	4453.2 (85.8% (*))	5861.8 (112.9% (*))	4381.9 (84.4% (*))

The simulation results under additive and multiplicative MMFE are presented in Table 4.1 and Table 4.2. The statistical significance of all costs relative to the deterministic models are assessed using Student's t-test. Statistical significance is indicated with a $(*)$ symbol for all relatives values for which the associated $p-$value is strictly smaller than 5%.

Our first observation is that the average costs of the deterministic model have large variations among the simulation settings while the true MMFE models yield more stable costs over the instances. The deterministic costs are especially high when capacity is low, uncertainty is high and when demand is dynamic such as for the random and sea-

Table 4.2.: Average simulation results over all configurations for multiplicative MMFE.

Demand	Uncert.	*cap*	Det.	Additive PLA	Multiplicative PLA	Ext. Mult. PLA
Stationary	Low	300	4694.8	4186.5 (89.6% (*))	4140.6 (88.6% (*))	4121.7 (88.2% (*))
		500	4649.2	4184.3 (90.5% (*))	4147.9 (89.7% (*))	4114.9 (88.9% (*))
	Medium	300	5455.8	4871.1 (91.3% (*))	4831.9 (90.9% (*))	4717.2 (88.4% (*))
		500	5098.7	4731.8 (93.8% (*))	4680.3 (93% (*))	4634.1 (92.0% (*))
	High	300	6309.1	5419.3 (90.0% (*))	5440.6 (91.4% (*))	5188.6 (86.1% (*))
		500	5516.5	5080.5 (93.8% (*))	5007.0 (93.1% (*))	4897.7 (90.8% (*))
Random	Low	300	4598.3	4059.4 (88.8% (*))	4006.2 (87.7% (*))	3978.9 (87.0% (*))
		500	4503.9	4015.2 (89.6% (*))	3955.6 (88.3% (*))	3925.1 (87.6% (*))
	Medium	300	5594.2	4884.6 (89.8% (*))	4791.5 (88.8% (*))	4675.1 (86.0% (*))
		500	5080.3	4655.9 (92.9% (*))	4549.8 (91.1% (*))	4472.2 (89.3% (*))
	High	300	6732.9	5710.8 (89.7% (*))	5674.0 (91.0% (*))	5481.1 (85.3% (*))
		500	5600.8	5130.1 (93.9% (*))	5053.5 (93.4% (*))	4870.0 (89.4% (*))
Seasonal	Low	300	4854.6	4172.8 (87.2% (*))	4085.5 (85.7% (*))	4026.9 (84.3% (*))
		500	4270.2	3858.1 (90.9% (*))	3743.0 (88.3% (*))	3699.1 (87.2% (*))
	Medium	300	6440.4	5363.9 (88.3% (*))	5309.0 (89.1% (*))	5066.6 (82.6% (*))
		500	5164.1	4628.1 (91.4% (*))	4425.9 (88.2% (*))	4313.8 (85.7% (*))
	High	300	8261.4	6373.7 (85.4% (*))	6485.8 (91.5% (*))	6188.9 (81.6% (*))
		500	5997.5	5199.3 (89.8% (*))	5079.8 (89.5% (*))	4838.3 (84.3% (*))
Average			5490.2	4807.0 (87.6% (*))	4744.9 (86.4% (*))	4622.8 (84.2% (*))

sonal patterns. It is also in these settings that the MMFE models with known forecast evolution process provide the highest cost reduction. On average, the additive and multiplicative MMFE models can reduce realised costs by 14% compared to the deterministic model when the forecast evolution process is known. The cost of modelling error is low for the additive model and high for the multiplicative model. Table 4.1 shows that a mismatched multiplicative model can significantly increase costs compared to traditional deterministic planning. Over all instances, the costs of the multiplicative are larger by 13% on average and more than 50% when demand is seasonal and uncertainty is high.

Value of recourse

The value of recourse is quantified by comparing the costs of the stochastic model without recourse solved with PLA and the extended stochastic model combining PLA and scenario-based recourse. Table 4.3 and Table 4.4 provide a statistical overview of the value of recourse on the previous simulation instances under additive and multiplicative MMFE respectively. The average, median and quartiles of the relative cost of extended model compared to the PLA model are provided. The statistical significance of the average relative cost is assessed through Student's t-test. It is indicated with a star symbol $(*)$ if the p-value is smaller than 5%.

Table 4.3.: Value of production recourse under additive MMFE.

Demand	Uncert.	*cap*	PLA Cost	Extended PLA with recourse Cost	Relative avg.	First quart.	Median	Third quart.
Stationary	Low	300	4158.9	4122.1	99.1% (*)	97.9%	99.1%	100.3%
		500	4169.6	4137.4	99.3% (*)	97.1%	99.2%	101.4%
	Medium	300	4641.0	4571.2	98.6% (*)	96.1%	98.3%	101.1%
		500	4589.7	4530.9	98.9% (*)	95.4%	98.8%	102.0%
	High	300	4930.7	4848.2	98.5% (*)	94.6%	97.9%	101.9%
		500	4821.6	4740.0	98.5% (*)	94.3%	98.4%	102.5%
Random	Low	300	4017.8	3982.1	99.2% (*)	97.5%	99.1%	100.6%
		500	3981.2	3945.4	99.1% (*)	97.7%	99.1%	100.5%
	Medium	300	4571.8	4484.9	98.2% (*)	95.0%	98.0%	100.8%
		500	4468.8	4406.5	98.7% (*)	96.4%	98.3%	100.9%
	High	300	4921.4	4823.0	98.2% (*)	94.8%	97.8%	101.3%
		500	4764.1	4675.9	98.3% (*)	94.9%	98.0%	101.6%
Seasonal	Low	300	3973.7	3892.7	98.0% (*)	96.0%	98.1%	100.0%
		500	3716.6	3671.8	98.8% (*)	97.4%	98.5%	99.9%
	Medium	300	4569.0	4485.1	98.3% (*)	94.7%	98.0%	101.6%
		500	4238.9	4137.2	97.7% (*)	95.3%	97.3%	99.8%
	High	300	5095.4	4973.5	97.6% (*)	92.9%	96.9%	101.1%
		500	4526.7	4447.0	98.4% (*)	95.0%	98.0%	101.3%
All settings			4453.2	4381.9	98.5% (*)	95.9%	98.4%	101.0%

The value of recourse varies over the simulation settings similarly for both MMFE models and is overall higher under multiplicative MMFE. It is higher for more complex planning settings: when demand is dynamic, uncertainty is high, and capacity is limited. On average, the value of recourse is around 1.5% and 2.5% over all simulation settings and can reach 2.3% and 6.2% for the additive and multiplicative models respectively. It is interesting to note that the median relative cost is always smaller than the average value. This suggests that the relative cost distribution is skewed: in the majority of cases, observed costs are smaller than the average value.

Sensitivity analysis of capacity

To further investigate the value of recourse in stochastic models and to identify settings in which it is most beneficial, we perform several sensitivity analyses. First, we analyse the influence of the available capacity in each period by varying $cap \in \{250, 275, 300, 325, 350, 375, 400, 450, 500, 600\}$. We fix the demand pattern to seasonal and set medium uncertainty as this setting is close to the real world-case study. For each capacity setting, rolling-horizon simulations are repeated 1000 times. The results are shown in Figure 4.6 and Figure 4.7 for the additive and multiplicative models respec-

Table 4.4.: Value of production recourse under multiplicative MMFE.

Demand	Uncert.	*cap*	PLA	Extended PLA with recourse				
			Cost	Cost	Relative avg.	First quart.	Median	Third quart.
Stationary	Low	300	4140.6	4121.7	99.6% (*)	96.9%	99.3%	102.0%
		500	4147.9	4114.9	99.3% (*)	96.8%	99.0%	101.4%
	Medium	300	4831.9	4717.2	97.7% (*)	93.0%	97.4%	101.7%
		500	4680.3	4634.1	99.2% (*)	94.7%	98.7%	103.1%
	High	300	5440.6	5188.6	95.2% (*)	89.0%	94.1%	99.9%
		500	5007.0	4897.7	98.0% (*)	92.4%	97.0%	102.5%
Random	Low	300	4006.2	3978.9	99.4% (*)	97.3%	99.1%	101.2%
		500	3955.6	3925.1	99.3% (*)	97.2%	98.9%	101.1%
	Medium	300	4791.5	4675.1	97.5% (*)	92.6%	96.5%	101.5%
		500	4549.8	4472.2	98.4% (*)	94.7%	97.7%	101.7%
	High	300	5674.0	5481.1	95.4% (*)	87.5%	93.2%	99.8%
		500	5053.5	4870.0	96.4% (*)	91.1%	96.0%	101.3%
Seasonal	Low	300	4085.5	4026.9	98.6% (*)	95.5%	98.1%	100.9%
		500	3743.0	3699.1	98.9% (*)	97.0%	98.5%	100.5%
	Medium	300	5309.0	5066.6	94.6% (*)	86.3%	91.7%	98.6%
		500	4425.9	4313.8	97.6% (*)	93.6%	97.2%	101.1%
	High	300	6485.8	6188.9	93.8% (*)	81.2%	87.7%	98.5%
		500	5079.8	4838.3	95.3% (*)	89.0%	94.0%	100.3%
All settings			4744.9	4622.8	97.4% (*)	92.9%	97.5%	101.2%

tively. Each figure shows the average absolute cost for the PLA and extended model as well as the relative improvement of the extended model with recourse. The 95% confidence intervals of the average values are represented with error bars.

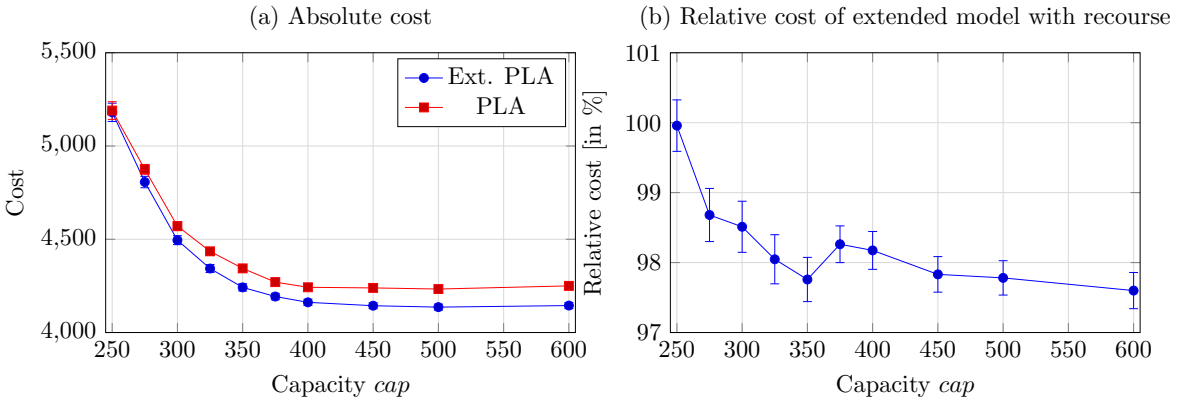

Figure 4.6.: (a) Absolute and (b) relative cost of PLA and extended model for additive MMFE with varying capacity.

For both forecast models, the average costs decrease exponentially as the available capacity increases. Under additive MMFE, the value of recourse increases almost monotonously with capacity. Under multiplicative MMFE, the value of recourse is higher when capac-

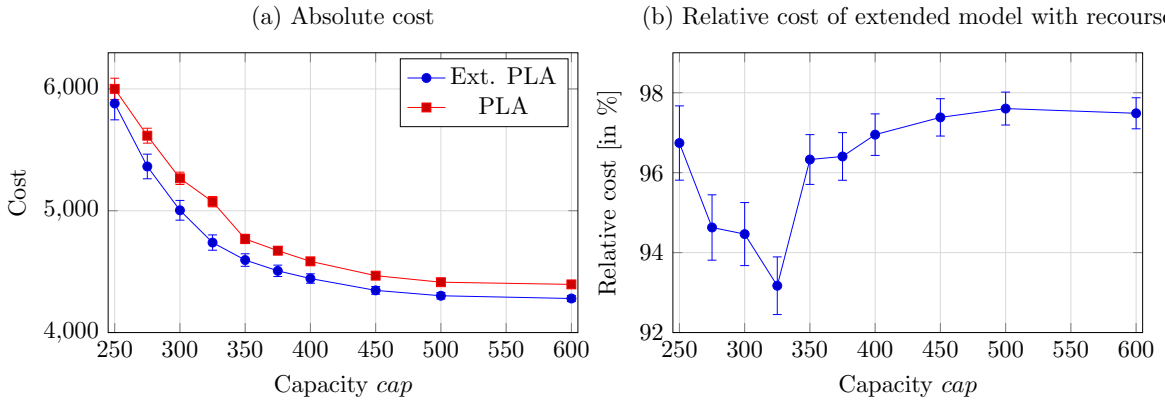

Figure 4.7.: (a) Absolute and (b) relative cost of PLA and extended model for multiplicative MMFE with varying capacity.

ity is constrained. Cost reductions of 6.8% are provided by the recourse model when are $cap = 325$, which corresponds to a setting with limited capacity but not overly constrained. Overall, the multiplicative model benefits most from recourse, achieving substantial cost reductions over all capacity settings.

Sensitivity analysis of scenario extension

The extended model with production recourse can provide significant cost savings thanks to more flexible decisions. However, it implies longer computation times due to the exponential scenario structure. The trade-off between cost reduction and increased computations can be managed in two ways: by adapting the scenario structure and by deciding the split between PLA and scenario models through t_b. In this sensitivity analysis, we compare three scenario tree structures: a small tree with nodes $[2, 4, 8, 8, 16, 16]$, an intermediate tree with nodes $[3, 6, 12, 24, 48, 48]$ and a large tree with nodes $[3, 9, 18, 36, 36, 72]$ in each stage. We perform $N = 1000$ rolling-horizon repetitions with varying parameter $t_b \in \{2, 3, 4, 5\}$ and compare the results of the extended model to the PLA model without recourse, equivalent to $t_b = 6$. We focus on the simulation setting with seasonal demand, medium uncertainty and large capacity $cap = 300$. The average costs relative to the PLA model and the average solver times are shown in Figure 4.8 with a 95% confidence interval.

The value of recourse as a function of t_b has different patterns for the additive and multiplicative models. Under additive MMFE, the costs of the extended model decrease as more recourse decisions are included for both the intermediate and large trees. The small scenario tree does not show further cost reduction below $t_b = 5$, which suggests

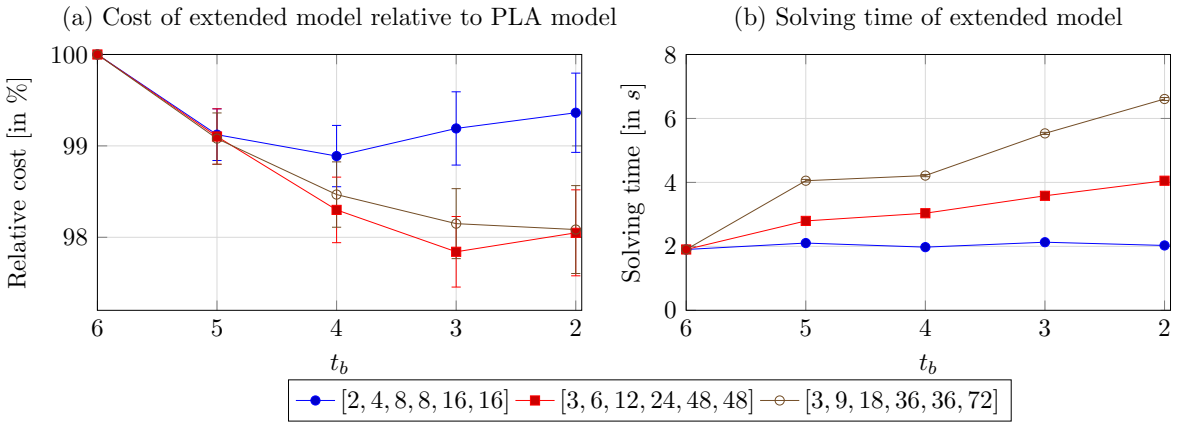

Figure 4.8.: (a) Relative cost and (b) solving time of extended model for varying t_b and scenario tree under additive MMFE.

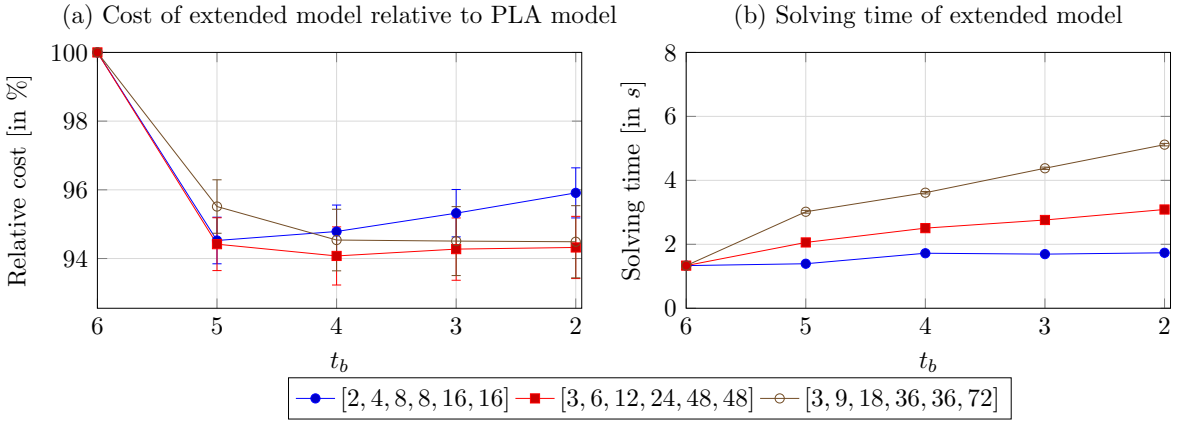

Figure 4.9.: (a) Relative cost and (b) solving time of extended model for varying t_b and scenario tree under multiplicative MMFE.

that the tree size is too small to accurately capture the forecast evolution and recourse opportunities. Under multiplicative MMFE, the value of recourse is higher but does not seem to be impacted by t_b. For both MMFE models, the intermediate tree size provides a good trade-off between observed costs and computations times. The sensitivity analysis shows that it is possible to obtain cost reductions through recourse by keeping solving times low. Indeed even the small tree can reduce costs on all instances with almost no computation times increase compared to the PLA model.

Sensitivity analysis of correlation structure

In Section 4.3 we have shown that positive (resp. negative) forecast time correlation was equivalent to higher (resp. lower) cumulative demand variance for both MMFE

models. For the extended model with recourse, correlation has an even larger impact since the recourse model can react to correlated forecast updates. We analyse the impact of the correlation structure on the value of recourse for both forecast evolution models on the simulation setting with seasonal demand and medium uncertainty. The capacity is fixed to $cap = 300$. The intermediate tree is used for the extended model with $[3, 6, 12, 24, 48, 48]$ nodes over the horizon.
The influence of both product and time correlation are investigated. Product correlation is set constant over the horizon as $\rho_{1,2}^{t,t} = \rho_k$ with $\rho_k \in \{-0.6, 0, 0.6\}$. Time correlation is set between the first and second periods of the horizon for both products $\rho_{k,k}^{1,2} = \rho_t$ with $\rho_t \in \{-0.6, 0, 0.6\}$. If both product and time correlation parameters are non-zero, then the first and second periods of the two products are also correlated and $\rho_{1,2}^{1,2} \neq 0$.

Table 4.5.: Value of recourse for different correlation structure.

	Additive MMFE			Multiplicative MMFE		
	$\rho_k = -0.6$	$\rho_k = 0$	$\rho_k = 0.6$	$\rho_k = -0.6$	$\rho_k = 0$	$\rho_k = 0.6$
$\rho_t = -0.6$	97.4(*)	96.9(*)	96.8(*)	92.1(*)	91.2(*)	89.2(*)
$\rho_t = 0$	98.3(*)	98.5(*)	99.0(*)	93.9(*)	94.6(*)	93.3(*)
$\rho_t = 0.6$	98.7(*)	100.1	100.9(*)	97.4(*)	97.1(*)	95.8(*)

The relative costs of the extended model with recourse compared to the PLA model without recourse are presented in Table 4.5. Statistical significance of the relative cost is assessed with Student's t-test and is shown with a star symbol $(*)$ if the p-value is below $p < 0.05$. The correlation structure has a strong influence on the value of recourse but impacts the two forecast evolution models differently. Under additive MMFE, the value of recourse appears to increase monotonously as correlation coefficients decrease. The value of recourse is again higher for the multiplicative model. Time correlation has a strong effect on the value of recourse, which increases as time correlation decreases. The value of recourse is highest for negative time correlation and positive product correlation as costs can be reduced by more than 10% compared to the stochastic model without recourse. This analysis shows the importance of including correlation in stochastic planning especially when using recourse models.

4.5.2. Real-world case study

We now apply our approach to the real-world case study of a large company manufacturing chemical products used in agriculture. The demand follows the growth cycle of crops and therefore exhibit strong seasonality and high uncertainty. The products are

grouped into $K = 6$ families for which a cleaning operations is required each a time a new family is set up. Together with our industrial partner, we gather the history of forecasts and demand realisations on a monthly granularity over 4 years. The historical demand is shown in Figure 4.10, which clearly shows the yearly seasonal pattern. The planning horizon is set to $T = 8$ to capture the majority of the season while keeping computation times low. The inventory costs are determined together with our industry partner and range between 0.04 and 0.1 over the product families. The backlog costs are set to $b_c = 15 \cdot h_c$ and the setup costs to $s_c = 15$. The initial inventory is set to zero. The monthly production capacity, $cap = 4934$, is large relative to the demand so that the case study strongly resembles the setting previously investigated with synthetic data. The PLA model is implemented with $L = 60$ segments. The extended model uses

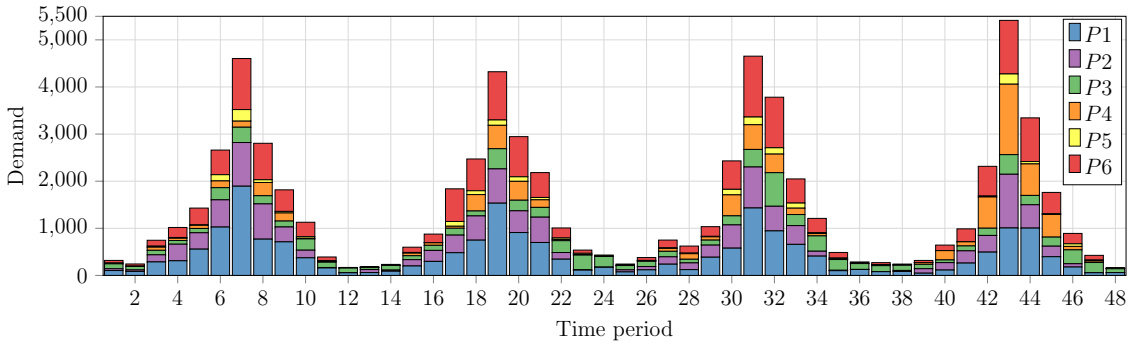

Figure 4.10.: Demand evolution over the four years of historical data for 6 product families.

a scenario tree with $[3, 6, 6, 12, 12, 24, 24, 48]$ nodes over the planning horizon sampled with LHMU. We set the extended model with $t_b = 4$ so that the first half of the planning horizon is modelled with PLA and the second half with scenario-based recourse.

Estimation of MMFE models

The first step for using the MMFE is to measure the forecast updates from the forecast and demand realisation history following the additive and multiplicative processes described in Section 4.3. As in the analysis with synthetic data, the occurrence of zero values for the forecast and demand realisation complicates the estimation of the multiplicative model parameters. All forecast and demand vectors in which at least one value is zero are removed from the data set, which amounts to around 50% of the data set.
The mean and covariance matrix of the MMFE model are estimated using the sample mean and covariance respectively. To conform to the unbiased assumption of the MMFE

models, we correct the sample bias. In Section 4.3, special attention has been given to the correlation of forecast updates in different time periods. The correlation matrix of the two MMFE models is represented in Figure 4.11. Interestingly, the additive MMFE exhibits strong positive correlation for the first three products over the horizon, while the multiplicative MMFE has high positive time correlations for the two last products.

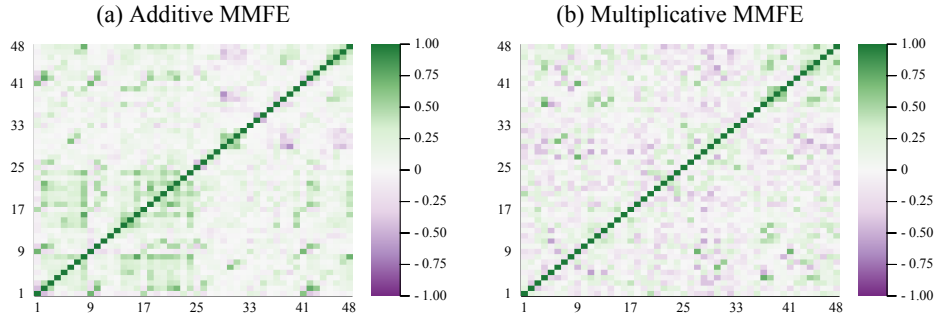

Figure 4.11.: Correlation matrix of (a) additive and (b) multiplicative MMFE models.

After estimating the distribution parameters, a practitioner might be interested in evaluating the goodness-of-fit of the update samples to the assumed distributions of the additive and multiplicative models. Intuitively, one would think that the goodness-of-fit provides a first measure of the expected performance of the MMFE models. The goodness-of-fit is assessed through a Shapiro-Wilk test performed on each marginal normal distribution of the additive and multiplicative models. The p-values of the tests are provided in Table 4.6 and Table 4.7 for the additive and multiplicative samples respectively. The statistical tests reject the assumptions that the forecast updates are normally distributed for all products and all time periods for the additive model. The results are more nuanced for the multiplicative model as some p-values are strictly greater than 5%. This first analysis suggests that the multiplicative model, having a better fit to the data, is likely to provide good results whereas the additive model should perform poorly.

Table 4.6.: p-value of Shapiro-Wilk normality test for additive samples.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8
k=1	2.50E-04	4.52E-05	0.077	1.18E-03	7.07E-03	5.50E-08	6.83E-09	5.70E-10
k=2	5.83E-03	8.63E-06	7.61E-09	7.29E-07	1.19E-08	1.58E-10	1.16E-09	1.65E-12
k=3	5.68E-05	1.48E-06	1.78E-10	6.56E-07	9.98E-05	1.58E-02	3.24E-03	2.60E-09
k=4	7.82E-07	1.12E-06	1.80E-11	4.21E-13	8.15E-12	2.28E-10	2.41E-13	9.97E-11
k=5	7.16E-05	1.75E-07	2.68E-07	1.35E-08	1.24E-08	6.79E-11	7.59E-09	1.64E-10
k=6	1.84E-03	9.99E-05	7.45E-07	6.85E-11	3.52E-10	3.82E-08	1.42E-09	3.45E-09

Table 4.7.: p-value of Shapiro-Wilk normality test for multiplicative samples.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8
k=1	0.072	1.93E-02	0.779	1.79E-05	2.54E-06	0.114	4.37E-03	1.20E-05
k=2	6.51E-04	4.60E-02	7.34E-04	2.08E-05	4.81E-07	0.147	2.91E-07	6.35E-05
k=3	2.42E-02	0.912	0.115	0.068	0.681	0.149	7.06E-04	0.054
k=4	2.41E-03	0.459	5.43E-04	2.14E-03	1.39E-06	1.61E-03	5.80E-04	1.38E-03
k=5	4.01E-04	6.45E-06	2.43E-05	7.87E-04	2.11E-06	5.77E-06	2.05E-05	1.30E-04
k=6	4.43E-02	9.31E-05	7.05E-05	3.43E-02	0.05	8.07E-05	6.27E-05	3.76E-03

Out-of-sample simulation results

To accurately assess the value of the MMFE model, the simulation is run in an out-of-sample fashion. Only past observations of the forecast evolution process are used to estimate the MMFE parameters in each review period. The simulation start at period 25, so that half the data set is available to estimate the MMFE models at the first simulation period, and half the set is used for rolling-horizon simulations. In each period, a new forecast update is observed and the model parameters are re-estimated in an online fashion. Forecast updates are taken from the data set. Thus, the simulation assesses the value of the models in a practical setting and, in particular, the ability to generalise the forecast evolution process based on past observations. The realised costs over the 24 periods out-of-sample simulation are presented in Table 4.8.

The additive MMFE model with PLA reduces realised costs by 11% thanks to relevant safety stock that increase inventory costs but provide significant reduction of backlog and setup costs. The extended additive model with production recourse further reduces costs by 3% through less conservative inventory decisions. The multiplicative MMFE model performs poorly over the simulation as it builds large inventory reserves. The results are contradicting with the goodness-of-fit analysis suggesting that goodness-of-fit is not a reliable a priori criterion to decide on the best MMFE model to use.

Table 4.8.: Results of out-of-sample case study.

Model	Total cost (rel.)	Inventory cost (rel.)	Backlog cost (rel.)	Setup cost (rel.)
Deterministic	3513 (100%)	796 (100%)	1442 (100%)	1275 (100%)
Additive PLA	3116 (89%)	1293 (162%)	803 (56%)	1020 (80%)
Extended Add. PLA	3015 (86%)	1112 (140%)	868 (60%)	1035 (81%)
Multipicative PLA	3612 (103%)	2424 (304%)	288 (20%)	900 (71%)
Extended Mult. PLA	3560 (101%)	2431 (305%)	259 (18%)	870 (68%)

4.5.3. Summary and recommendations

The numerical study shows that integrating forecast evolution models in stochastic lot-sizing problem can significantly improve planning quality. The additive MMFE model is robust and performs well over all simulation instances investigated: (1) when the forecast evolution process is known, (2) when it is unknown and estimated from mismatched updates, and (3) when it is learned from real-world past data. On the contrary, the multiplicative model is undermined by several limitations: it requires costly evaluations of the first-order loss function, it cannot use demand or forecast having zero value, and it can lead to important cost increases as shown in the mismatched analysis and real-world case study. Thus, we recommend practitioners to prioritise the implementation of the additive MMFE. Further, we highlight that the choice of relevant MMFE models should not be based on an a priori goodness-of-fit analysis but instead on evaluating model performance through out-of-sample rolling-horizon simulations using historical data.
The extended model with production recourse can provide consistent cost reductions on both synthetic and real-world data. Over all simulation settings, the value of recourse is higher for the multiplicative model. Still, recourse can consistently provide lower costs for the additive model as well. Several parameters that impact the value of recourse have been analysed such as the demand pattern, uncertainty, capacity, and correlation. In particular we have identified that the value of recourse is especially high when demand is dynamic, uncertainty is high and when forecast updates exhibit negative time correlation.
The extended model with recourse requires managerial decisions as it impacts planning in several ways including longer computation times and lower planning visibility due to the presence of recourse decisions. In our analysis, we have provided some first guidelines to tune the model and find a good compromise between its advantages and limitations.

4.6. Conclusion

In this paper, forecast evolution models are integrated in dynamic, stochastic, capacitated lot-sizing problems. After presenting the forecast evolution process, we have shown that determining the cumulative demand distributions is the only step needed to solve stochastic lot-sizing problems efficiently with PLA. The model was extended with a scenario-tree representation of uncertainty to allow production recourse over the planning horizon and provide flexible decisions. The value of forecast evolution models

has been quantified in a large-scale numerical study using both synthetic and real-world data. Forecast evolution models have been shown to provide significant cost reductions compared to traditional deterministic planning. However, when the forecast evolution process is unknown, the multiplicative MMFE tend to generalise poorly from available data and to increase realised costs due to important inventory build-ups. On the contrary, the additive model performs robustly over all simulation instances. We advise practitioners to focus on the additive MMFE because of its ability to generalise the stochastic forecast evolution process from available data.

Production recourse has been shown to consistently reduce costs on average for both additive and multiplicative MMFE thanks to less conservative inventory decisions. Important parameters that impact the value of recourse have been identified through sensitivity analyses.

This work proposes the first numerical comparison of the performance of additive and multiplicative MMFE in rolling-horizon planning. As several shortcomings have been identified for the multiplicative model, further work could investigate more robust techniques to estimate the model parameters. In particular, estimation techniques that better handle cases with zero forecast or demand are needed to use available data more efficiently. More generally, MMFE models cannot differentiate between volume and timing change of forecasts. In practice, it is often the case that an order is shifted in time with the same volume. An interesting research direction would be to adapt the MMFE framework to this setting.

Chapter 5

Conclusions

5.1. Summary

This thesis studies the integration of stochastic optimisation in rolling-horizon production planning. Several planning problems with uncertain and seasonal demand have been presented and solved with different approaches. In Chapter 2, a multi-ordering newsvendor problem with inventory carrying costs and forecast evolution is solved as a dynamic programming model. In Chapter 3, two-stage stochastic programming with optimally defined product families is applied to a real-world case study. In Chapter 4, additive and multiplicative martingale model of forecast evolution (MMFE) are integrated in a general lot-sizing problem and combined with a multi-stage scenario tree to allow production recourse. We summarise our findings by answering the research questions stated in Section 1.2.

(RQ 1) *How can stochastic models be applied from the available history of forecast and demand data?*

We have proposed several approaches to develop stochastic models from available forecast and demand data. In Chapter 3, demand-driven and forecast-driven uncertainty models are defined and used to build scenario trees. We compare the use of empirical distributions and estimated distributions that are sampled to obtain additional scenarios. We show that forecast-driven uncertainty models with estimated distributions achieve high demand satisfaction at low cost on out-of-samples simulations with real-world data and outperform demand-driven uncertainty models. In Chapter 4, a second approach to apply stochastic models from data focuses on modelling the forecast revision process. Additive and multiplicative MMFE models are estimated from available forecast and demand data and integrated in lot-sizing problems. Out-of-sample simulations with

real-world data show that forecast evolution models consistently improve demand satisfaction compared to deterministic models.

(RQ 2.1) *How can MMFE models be integrated into complex production planning environments?*

(RQ 2.2) *What are strengths and limitations of the additive and multiplicative MMFE when applied from real-world data?*

We develop two methods to solve stochastic models with forecast evolution. In Chapter 2, the optimal production policy is determined analytically for the single-product case and extended to a heuristic for multiple correlated products. The heuristic is based on a decomposition/coordination procedure that iteratively allocates capacity between products and determines the products' inventory targets independently. The numerical study shows that explicitly modelling forecast evolution is essential to reach target service levels when products have correlated forecast evolution and when uncertainty resolution is not constant. In Chapter 4, we show that MMFE models can be integrated in lot-sizing problems by determining the cumulative demand distributions over the planning horizon. The resulting non-linear problems can be solved efficiently using existing piecewise-linearisation techniques. The additive MMFE can be solved to arbitrary optimality whereas the multiplicative model relies on an approximation to determine probability distributions of the cumulative demand. We compare the additive and multiplicative MMFE models in extensive rolling-horizon simulations using both synthetic and real-world data. In particular, we quantify the risk of modelling error due to using an inappropriate MMFE model. The additive model is shown to be the more robust MMFE model as it consistently reduces production costs over a wide range of problem settings. On the contrary, the multiplicative model appears sensitive to modelling error and is shown to increase costs in several instances when compared to a simple deterministic benchmark. Overall, we show that the value of forecast evolution is closely linked to the recourse opportunities given by multi-stage stochastic models.

(RQ 3) *What is the value of recourse in rolling-horizon planning and what parameters influence it?*

The three chapters of this thesis measure the value of recourse through repeated rolling-horizon simulations. Parameters that affect the value of recourse are identified over the different problems. In particular, we show that the value of recourse is high when there is a complex correlation structure between products and time periods. Further, recourse is especially beneficial when capacity is limited, suggesting that planning models with recourse determine a better prioritisation of production. In Chapter 2 and 4, we show

that stochastic models that ignore forecast evolution and recourse fail to meet service level targets and provide overall higher inventory costs. Hence, forecast evolution and recourse are intimately linked and are both necessary to provide optimal decisions in rolling-horizon.

(RQ 4) *How can stochastic models that satisfy the trade-off between planning flexibility, stability and communicability be developed?*

Multi-stage stochastic models with recourse can derive cost-efficient decisions but are challenging to set up and solve while respecting the constraints of rolling-horizon planning. In Chapter 2, we use a linear policy approximation to obtain reference plans from the optimal and approximate production policies. We compare the nervousness resulting from planning models with and without explicit forecast evolution. We show that using a flexible policy does not increase nervousness on average and lead to a linear relationship between forecast nervousness and production nervousness. In Chapter 3 we solve the trade-off between planning flexibility, stability and communicability by determining optimal product families that reflect production processes. The families allow to reserve capacity through first-stage decisions that can be used by individual products in recourse decisions. This approach ensures that a reference plan is available while providing high flexibility. It is shown to also improve planning stability as plan changes compensate within product families. This strategy is notable in the way that it improves flexibility and stability simultaneously, whereas existing techniques such as freezing decisions or penalising changes inherently reduce planning flexibility. In Chapter 4, multi-stage stochastic programming and piecewise-linearisation techniques are combined so that recourse is available for later periods in the horizon. Flexible decisions are allowed but a unique reference plan is determined on the short-term horizon. This approach provides substantial cost reductions and remains tractable as small scenario trees are sufficient to represent the multi-stage process accurately.

This thesis has highlighted the importance of integrating recourse in stochastic production planning models. Several methods have been developed to derive forecast uncertainty models from data and to formulate optimisation problems with successive decision stages. Stochastic planning models with recourse have been solved with specific techniques pertinent to each problem. To allow the implementation of models with recourse in existing rolling-horizon planning practice, we have put emphasis on determining stable reference plans in each review period. We have set up extensive rolling-horizon simulations that show that stochastic models with recourse can increase demand satisfaction, reduce inventory costs and even provide more stable plans.

5.2. Outlook

Several research directions are identified to extend the main research topics in this thesis.

In the three chapters of this thesis, we use recourse to adapt decisions to uncertainty in rolling-horizon planning. We have shown that a linear policy approximation provides good results on multi-ordering newsvendor problems. We have also applied a more traditional multi-stage scenario-based model to lot-sizing problems. Using linear policy approximations to introduce recourse in lot-sizing problems could be a first extension, which may resolve two of the main difficulties of scenario-based models: their long computation times and the discretisation errors due to sampling. The linear policy approximation could be included directly in the lot-sizing model presented in Chapter 4 and solved by adapting the piecewise-linear approximation. This method would be similar to Adjustable Robust Optimisation (Yanıkoğlu et al., 2019) but applied in a stochastic context. Scenario-based recourse and linear policy approximations could also be combined so that different variables use different recourse formulations in the same optimisation model.

In Chapter 3, we develop a strategy to obtain flexible and stable decisions based on the aggregation of products in families. The definition of the product families is done in a optimisation problem that reflects the constraints of the production processes of the real-world case study. The generalisation of this method to different industries and problem settings could be investigated. The value of aggregating decisions could also be evaluated for varied product structures.

Forecast evolution models act as a bridge between academia and practitioners as they allow the application of stochastic models with meaningful uncertainty distributions estimated from readily available data. The additive and multiplicative versions of the MMFE have been applied in Chapter 4 and their strengths and limitations have been identified. In particular, we have shown that multiplicative MMFE is sensitive to modelling error when demand is dynamic even though it has been designed precisely for situations with fluctuating demand. We argue that two aspects of forecast evolution models are currently conflated: (1) how to measure forecast update samples from data, and (2) identifying the underlying probability distribution. While the two existing methods measure forecast updates as absolute difference or log-ratios, there may be other ways to measure forecast evolution such as ratios or percentages. The additive and multiplicative models both use the assumption that forecast updates follow a normal distribution. While this distribution provides nice mathematical properties, it

may not be suitable to most problem settings. Other probability distributions could be investigated to model additive or relative forecast evolution. Another noteworthy direction is to apply distribution-free approaches such as robust or distributionally robust optimisation to model forecast evolution from data.
The intermittency of demand is another practical aspect that is not captured by existing MMFE models. In Chapter 4, we show that the occurrence of zero values for demand and forecasts is frequent in real-world data but can lead to estimation problems, especially for multiplicative MMFE. Extending forecast evolution models to capture demand intermittency could allow the application of MMFE to more realistic problem settings.

Finally, the interaction of planning nervousness and forecast evolution has been only tangentially studied. A challenging but promising research direction is to develop event-driven planning frameworks that calculate updated production plans only when necessary. In Chapter 2, we have shown that using the optimal production policy with forecast evolution creates a linear relationship between forecast nervousness and production nervousness. This result could be used to trigger replanning actions in an event-driven fashion. This new planning paradigm could be linked to existing research in control theory and production scheduling to provide an automated planning framework with high flexibility and stability.

Bibliography

Abu-Dayya, Adnan A and Norman C Beaulieu (1994). Outage probabilities in the presence of correlated lognormal interferers. *IEEE Transactions on Vehicular Technology* 43 (1), pp. 164–173.

Aditi, Karnad, Anelia Boshnakova, and Annie Pannelay (2018). Medicine and vaccine shortages: What is the role of global regulatory complexity for post approval changes? Accessed September 25, 2020, `http://graphics.eiu.com/upload/topic-pages/medicine-shortages/Medicine-and-vaccine-shortages-EIU.pdf`. The Economist Intelligence Unit, EIU Healthcare.

Albey, E., R. Uzsoy, and K. G. Kempf (2016). A chance constraint based multi-item production planning model using simulation optimization. *2016 Winter Simulation Conference (WSC)*, pp. 2719–2730.

Albey, Erinc, Amirhosein Norouzi, Karl G Kempf, and Reha Uzsoy (2015). Demand modeling with forecast evolution: an application to production planning. *IEEE Transactions on Semiconductor Manufacturing* 28 (3), pp. 374–384.

Atadeniz, Sukran N and Sri V Sridharan (2020). Effectiveness of nervousness reduction policies when capacity is constrained. *International Journal of Production Research* 58 (13), pp. 4121–4137.

Ban, Gah-Yi (2020). Confidence intervals for data-driven inventory policies with demand censoring. *Operations Research* 68 (2), pp. 309–326.

Ban, Gah-Yi, Jérémie Gallien, and Adam J. Mersereau (2019). Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing & Service Operations Management* 21 (4), pp. 798–815.

Bansal, Saurabh and Mahesh Nagarajan (2017). Product portfolio management with production flexibility in agribusiness. *Operations Research* 65 (4), pp. 914–930.

Ben-Tal, Aharon, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59 (2), pp. 341–357.

Bertsimas, Dimitris, Vishal Gupta, and Nathan Kallus (2018a). Data-driven robust optimization. *Mathematical Programming* 167 (2), pp. 235–292.

— (2018b). Robust sample average approximation. *Mathematical Programming* 171 (1–2), 217–282.

Beutel, Anna-Lena and Stefan Minner (2012). Safety stock planning under causal demand forecasting. *International Journal of Production Economics* 140 (2), pp. 637–645.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B Shah (2017). Julia: A fresh approach to numerical computing. *SIAM Review* 59 (1), pp. 65–98.

Biçer, Işık and Ralf W Seifert (2017). Optimal dynamic order scheduling under capacity constraints given demand-forecast evolution. *Production and Operations Management* 26 (12), pp. 2266–2286.

Blackburn, J. D., D. H. Kropp, and R. A. Millen (1986). A comparison of strategies to dampen nervousness in MRP systems. *Management Science* 32 (4), pp. 413–429.

Bollapragada, Ramesh, Saravanan Kuppusamy, and Uday S Rao (2015). Component procurement and end product assembly in an uncertain supply and demand environment. *International Journal of Production Research* 53 (3), pp. 969–982.

Boyacı, Tamer and Özalp Özer (2010). Information acquisition for capacity planning via pricing and advance selling: When to stop and act? *Operations Research* 58 (5), pp. 1328–1349.

Brandimarte, Paolo (2006). Multi-item capacitated lot-sizing with demand uncertainty. *International Journal of Production Research* 44 (15), pp. 2997–3022.

Buschkühl, Lisbeth, Florian Sahling, Stefan Helber, and Horst Tempelmeier (2010). Dynamic capacitated lot-sizing problems: a classification and review of solution approaches. *OR Spectrum* 32 (2), pp. 231–261.

Carlson, Robert C, James V.; Jucker, and Dean H Kropp (1979). Less nervous MRP systems: A dynamic economic lot-sizing approach. *Management Science* 25 (8), pp. 754–761.

Chopra, Sunil and Peter Meindl (2013). *Supply chain management: Strategy, planning & operation*. 5th. Pearson Prentice Hall Inc.

Cunha, Artur Lovato, Maristela Oliveira Santos, Reinaldo Morabito, and Ana Barbosa-Póvoa (2018). An integrated approach for production lot sizing and raw material purchasing. *European Journal of Operational Research* 269 (3), pp. 923–938.

De Smet, Niels, Stefan Minner, El-Houssaine Aghezzaf, and Bram Desmet (2020). A linearisation approach to the stochastic dynamic capacitated lotsizing problem with

sequence-dependent changeovers. *International Journal of Production Research* 58 (16), pp. 4980–5005.

Deutsch, Jared L and Clayton V Deutsch (2012). Latin hypercube sampling with multidimensional uniformity. *Journal of Statistical Planning and Inference* 142 (3), pp. 763–772.

Donohue, Karen L (2000). Efficient supply contracts for fashion goods with forecast updating and two production modes. *Management Science* 46 (11), pp. 1397–1411.

Dunning, Iain, Joey Huchette, and Miles Lubin (2017). JuMP: A modeling language for mathematical optimization. *SIAM Review* 59 (2), pp. 295–320.

Dupačová, Jitka, Giorgio Consigli, and Stein W Wallace (2000). Scenarios for multistage stochastic programs. *Annals of operations research* 100 (1), pp. 25–53.

Dupačová, Jitka, Nicole Gröwe-Kuska, and Werner Römisch (2003). Scenario reduction in stochastic programming. *Mathematical programming* 95 (3), pp. 493–511.

Escudero, Laureano F, Pasumarti V Kamesam, Alan J King, and Roger JB Wets (1993). Production planning via scenario modelling. *Annals of Operations research* 43 (6), pp. 309–335.

Fisher, Marshall and Ananth Raman (1996). Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* 44 (1), pp. 87–99.

Forel, Alexandre and Martin Grunow (2020a). Dynamic stochastic lot sizing with forecast evolution. Working paper.

— (2020b). Production planning for a short seasonal demand with forecast evolution. Working paper.

— (2020c). Stochastic programming in master production scheduling: overcoming barriers to industry application. Working paper.

Graves, Stephen C, Harlan C Meal, Sriram Dasu, and Yuping Qui (1986). Two-stage production planning in a dynamic environment. *Multi-Stage Production Planning and Inventory Control*. Ed. by Sven Axsäter, Christoph Schneeweiss, and Edward Silver. Berlin, Heidelberg: Springer, pp. 9–43.

Gruson, Matthieu, Jean-François Cordeau, and Raf Jans (2021). Benders decomposition for a stochastic three-level lot sizing and replenishment problem with a distribution structure. *European Journal of Operational Research* 291 (1), pp. 206–217.

Hausman, Warren H (1969). Sequential decision problems: A model to exploit existing forecasters. *Management Science* 16 (2), B–93.

Hausman, Warren H and Rein Peterson (1972). Multiproduct production scheduling for style goods with limited capacity, forecast revisions and terminal delivery. *Management Science* 18 (7), pp. 370–383.

Heath, David C and Peter L Jackson (1994). Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions* 26 (3), pp. 17–30.

Heitsch, Holger and Werner Römisch (2003). Scenario reduction algorithms in stochastic programming. *Computational optimization and applications* 24 (2), pp. 187–206.

— (2009). Scenario tree modeling for multistage stochastic programs. *Mathematical Programming* 118 (2), pp. 371–406.

Helber, Stefan, Florian Sahling, and Katja Schimmelpfeng (2013). Dynamic capacitated lot sizing with random demand and dynamic safety stocks. *OR Spectrum* 35 (1), pp. 75–105.

Herrera, Carlos, Sana Belmokhtar-Berraf, André Thomas, and Víctor Parada (2016). A reactive decision-making approach to reduce instability in a master production schedule. *International Journal of Production Research* 54 (8), pp. 2394–2404.

Ho, Chrwan-Jyh (1989). Evaluating the impact of operating environments on MRP system nervousness. *International Journal of Production Research* 27 (7), pp. 1115–1135.

Homem-de-Mello, Tito and Güzin Bayraksan (2014). Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science* 19 (1), pp. 56–85.

Hu, Zhengyang and Guiping Hu (2018). A multi-stage stochastic programming for lot-sizing and scheduling under demand uncertainty. *Computers & Industrial Engineering* 119, pp. 157–166.

Huber, Jakob, Sebastian Müller, Moritz Fleischmann, and Heiner Stuckenschmidt (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research* 278 (3), pp. 904–915.

Iida, Tetsuo and Paul H Zipkin (2006). Approximate solutions of a dynamic forecast-inventory model. *Manufacturing & Service Operations Management* 8 (4), pp. 407–425.

Jensen, Thomas (1993). Measuring and improving planning stability of reorder-point lot-sizing policies. *International Journal of Production Economics* 30, pp. 167–178.

Johnson, Steven G (2014). The NLopt nonlinear-optimization package. Accessed September 25, 2020, `http://github.com/stevengj/nlopt`.

Jones, Philip C, Timothy J Lowe, Rodney D Traub, and Greg Kegler (2001). Matching supply and demand: The value of a second chance in producing hybrid seed corn. *Manufacturing & Service Operations Management* 3 (2), pp. 122–137.

Kanyalkar, Atul P and Gajendra K Adil (2010). A robust optimisation model for aggregate and detailed planning of a multi-site procurement-production-distribution system. *International Journal of Production Research* 48 (3), pp. 635–656.

Kazemi Zanjani, Masoumeh, Mustapha Nourelfath, and Daoud Ait-Kadi (2010). A multi-stage stochastic programming approach for production planning with uncertainty in the quality of raw materials and demand. *International Journal of Production Research* 48 (16), pp. 4701–4723.

King, Alan J and Stein W Wallace (2012). *Modeling with stochastic programming*. Springer Science & Business Media, New York.

Klabjan, Diego, David Simchi-Levi, and Miao Song (2013). Robust stochastic lot-sizing by means of histograms. *Production and Operations Management* 22 (3), pp. 691–710.

Kleywegt, Anton J, Alexander Shapiro, and Tito Homem-de-Mello (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12 (2), pp. 479–502.

Koca, E., H. Yaman, and M. S. Aktürk (2018). Stochastic lot sizing problem with nervousness considerations. *Computers and Operations Research* 94, pp. 23–37.

Körpeoğlu, Ersin, Hande Yaman, and M Selim Aktürk (2011). A multi-stage stochastic programming approach in master production scheduling. *European Journal of Operational Research* 213 (1), pp. 166–179.

Leung, Ngai-Hang Z, Ana Chen, Prashant Yadav, and Jérémie Gallien (2016). The impact of inventory management on stock-outs of essential drugs in Sub-Saharan Africa: Secondary analysis of a field experiment in Zambia. *PloS ONE* 11 (5), e0156026.

Li, Jian, Suresh Chand, Maqbool Dada, and Shailendra Mehta (2009). Managing inventory over a short season: models with two procurement opportunities. *Manufacturing & Service Operations Management* 11 (1), pp. 174–184.

Li, Qinyun and Stephen M. Disney (2017). Revisiting rescheduling: MRP nervousness and the bullwhip effect. *International Journal of Production Research* 55 (7), pp. 1992–2012.

Lin, Po Chen and Reha Uzsoy (2016). Chance-constrained formulations in rolling horizon production planning: an experimental study. *International Journal of Production Research* 54 (13), pp. 3927–3942.

Meistering, Malte and Hartmut Stadtler (2017). Stabilized-cycle strategy for capacitated lot sizing with multiple products: fill-rate constraints in rolling schedules. *Production and Operations Management* 26 (12), pp. 2247–2265.

Milner, Joseph M and Panos Kouvelis (2005). Order quantity and timing flexibility in supply chains: The role of demand characteristics. *Management Science* 51 (6), pp. 970–985.

Mišić, Velibor V and Georgia Perakis (2020). Data analytics in operations management: A review. *Manufacturing & Service Operations Management* 22 (1), pp. 158–169.

Norouzi, Amirhosein (2012). The effect of forecast evolution on production planning with resources subject to congestion. PhD thesis. Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh.

Norouzi, Amirhosein and Reha Uzsoy (2014). Modeling the evolution of dependency between demands, with application to inventory planning. *IIE Transactions* 46 (1), pp. 55–66.

Oroojlooyjadid, Afshin, Lawrence V Snyder, and Martin Takáč (2020). Applying deep learning to the newsvendor problem. *IISE Transactions* 52 (4), pp. 444–463.

Özer, Özalp and Wei Wei (2004). Inventory control with limited capacity and advance demand information. *Operations Research* 52 (6), pp. 988–1000.

Pelt, Thomas D. van and Jan C. Fransoo (2018). A note on "Linear programming models for a stochastic dynamic capacitated lot sizing problem". *Computers and Operations Research* 89, pp. 13–16.

Pinçe, Çerağ, Enver Yücesan, and Prithveesha Govinda Bhaskara (2021). Accurate response in agricultural supply chains. *Omega* 100, p. 102214.

Powell, M. J. D. (1994). A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation. *Advances in Optimization and Numerical Analysis*. Ed. by Susana Gomez and Jean-Pierre Hennart. Dordrecht: Springer Netherlands, pp. 51–67.

Powell, Warren B (2016). Perspectives of approximate dynamic programming. *Annals of Operations Research* 241 (1), pp. 319–356.

Prak, Dennis and Ruud Teunter (2019). A general method for addressing forecasting uncertainty in inventory models. *International Journal of Forecasting* 35 (1), pp. 224–238.

Prak, Dennis, Ruud Teunter, and Aris Syntetos (2017). On the calculation of safety stocks when demand is forecasted. *European Journal of Operational Research* 256 (2), pp. 454–461.

Rossi, Roberto, Onur A Kilic, and S Armagan Tarim (2015). Piecewise linear approximations for the static–dynamic uncertainty strategy in stochastic lot-sizing. *Omega* 50, pp. 126–140.

Rossi, Roberto, S Armagan Tarim, Steven Prestwich, and Brahim Hnich (2014). Piecewise linear lower and upper bounds for the standard normal first order loss function. *Applied Mathematics and Computation* 231, pp. 489–502.

Sahin, Funda, Arunachalam Narayanan, and E. Powell Robinson (2013). Rolling horizon planning in supply chains: Review, implications and directions for future research. *International Journal of Production Research* 51 (18), pp. 5413–5436.

Saliby, Eduardo (1990). Descriptive sampling: a better approach to Monte Carlo simulation. *Journal of the Operational Research Society* 41 (12), pp. 1133–1142.

Sereshti, Narges, Yossiri Adulyasak, and Raf Jans (2020). The value of aggregate service levels in stochastic lot sizing problems. *Omega*. In press. DOI: https://doi.org/10.1016/j.omega.2020.102335.

Shapiro, Alexander (2011). Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research* 209 (1), pp. 63–72.

Silver, Edward A, David F Pyke, and Douglas J Thomas (2016). *Inventory and Production Management in Supply Chains*. 4th. CRC Press, Boca Raton.

Snyder, Lawrence V and Zuo-Jun Max Shen (2019). *Fundamentals of supply chain theory*. 2nd. John Wiley & Sons.

Sridharan, Sri V, William L Berry, and V Udayabhanu (1988). Measuring master production schedule stability under rolling planning horizons. *Decision Sciences* 19 (1), pp. 147–166.

Sridharan, V and William L Berry (1990). Freezing the master production schedule under demand uncertainty. *Decision Sciences* 21 (1), pp. 97–120.

Stephan, Holger A, Timo Gschwind, and Stefan Minner (2010). Manufacturing capacity planning and the value of multi-stage stochastic programming under Markovian demand. *Flexible Services and Manufacturing Journal* 22 (3-4), pp. 143–162.

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

Tavaghof-Gigloo, Dariush and Stefan Minner (2020). Planning approaches for stochastic capacitated lot-sizing with service level constraints. *International Journal of Production Research*. In press. DOI: 10.1080/00207543.2020.1773003.

Tempelmeier, Horst and Timo Hilger (2015). Linear programming models for a stochastic dynamic capacitated lot sizing problem. *Computers & Operations Research* 59, pp. 119–125.

Thevenin, Simon, Yossiri Adulyasak, and Jean-François Cordeau (2021). Material requirements planning under demand uncertainty using stochastic optimization. *Production and Operations Management* 30 (2), pp. 475–493.

Thomas, Douglas J (2005). Measuring item fill-rate performance in a finite horizon. *Manufacturing & Service Operations Management* 7 (1), pp. 74–80.

Toktay, L Beril and Lawrence M Wein (2003). Analysis of a forecasting-production-inventory system with stationary demand. *Management Science* 47 (9), pp. 1268–1281.

Trapero, Juan R, Manuel Cardós, and Nikolaos Kourentzes (2019). Empirical safety stock estimation based on kernel and GARCH models. *Omega* 84, pp. 199–211.

Tunc, Huseyin, Onur A. Kilic, S. Armagan Tarim, and Burak Eksioglu (2013). A simple approach for assessing the cost of system nervousness. *International Journal of Production Economics* 141 (2), pp. 619–625.

Vargas, Vicente and Richard Metters (2011). A master production scheduling procedure for stochastic demand and rolling planning horizons. *International Journal of Production Economics* 132 (2), pp. 296–302.

Wang, Tong, Atalay Atasu, and Mümin Kurtuluş (2012). A multiordering newsvendor model with dynamic forecast evolution. *Manufacturing & Service Operations Management* 14 (3), pp. 472–484.

Wang, Yimin and Brian Tomlin (2009). To wait or not to wait: Optimal ordering under lead time uncertainty and forecast updating. *Naval Research Logistics* 56 (8), pp. 766–779.

Watson, Jean-Paul and David L Woodruff (2011). Progressive hedging innovations for a class of stochastic mixed-integer resource allocation problems. *Computational Management Science* 8 (4), pp. 355–370.

Wiesemann, Wolfram, Daniel Kuhn, and Melvyn Sim (2014). Distributionally robust convex optimization. *Operations Research* 62 (6), pp. 1358–1376.

Yanıkoğlu, İhsan, Bram L Gorissen, and Dick den Hertog (2019). A survey of adjustable robust optimization. *European Journal of Operational Research* 277 (3), pp. 799–813.

Yano, Candace Arai and Robert C Carlson (1987). Interaction between frequency of rescheduling and the role of safety stock in material requirements planning systems. *International Journal of Production Research* 25 (2), pp. 221–232.

Zhao, Xiande, Jinxing Xie, and Qiyuan Jiang (2001). Lot-sizing rule and freezing the master production schedule under capacity constraint and deterministic demand. *Production and Operations Management* 10 (1), pp. 45–67.

Ziarnetzky, Timm, Lars Mönch, and Reha Uzsoy (2018). Rolling horizon, multi-product production planning with chance constraints and forecast evolution for wafer fabs. *International Journal of Production Research* 56 (18), pp. 6112–6134.

Appendix

Appendix A

Production planning for a short seasonal demand with forecast evolution

A.1. Proof of Proposition 2.1

The lost-sales on the right-hand side of Equation (2.1) correspond to the well-studied *first-order loss function*. The first-order loss function of a random variable ω with cumulative distribution F_ω and probability distribution f_ω evaluated at a scalar x is defined as $\mathcal{L}(x,\omega) = \mathbb{E}[\max(\omega - x; 0)] = \int_{-\infty}^{\infty} \max(t-x,0) g_\omega(t) dt$.
Let ω be a normal distribution with mean μ and variance σ^2, probability density function ϕ and cumulative density function Φ. The first-order loss function of ω can be expressed as a function of the first-order loss function of a standard normal distribution $\mathcal{L}(x,\omega) = \sigma\mathcal{L}(\frac{x-\mu}{\sigma})$ (Snyder and Shen, 2019, p. 663). This property allows significant computational advantages since values of the first-order loss function of a standard normal can be pre-computed. The first-order loss function of a standard normal can further be expressed as $\mathcal{L}(x) = \phi(x) - x \cdot (1 - \Phi(x))$, which further facilitates calculations (Snyder and Shen, 2019, p. 98). It is straightforward to show that the first-order loss function of a standard normal distribution is strictly decreasing and invertible, with $\mathcal{L}^{-1}$ being defined over $]0; +\infty[$.

A.1.1. Proof of optimal inventory in the last-period

The expected lost sales as a function of the inventory position y_T is strictly decreasing. The optimal inventory in the last period is the smallest inventory y_T for which

the constraint in Equation (2.1) holds, namely $\mathbb{E}_{D_{T+1}|T}[\max(D_{T+1} - y_T^*; 0)] = (1 - \beta)\,\mathbb{E}_{D_{T+1}|T}[D_{T+1}]$. Since demand follows a known normal distribution, it follows from the properties of the first-order loss function that the optimal last-period inventory y_T^* is such that

$$\mathcal{L}_T(y_T^*, D_{T+1}) = \sigma_{T+1} \cdot \mathcal{L}\left(\frac{y_T^* - \mathbb{E}_{D_{T+1}|T}[D_{T+1}]}{\sigma_{T+1}}\right) = (1 - \beta)\,\mathbb{E}_{D_{T+1}|T}[D_{T+1}].$$

The final expression is obtained knowing that the first-order loss function is invertible.

A.2. Proof of Lemma 2.1

A.2.1. Preliminary result: derivative of inverse of first-order loss function

First note that the inverse of the first-order loss function of a standard normal distribution is differentiable as the inverse of a differentiable function. The first-order loss function can be also expressed $\mathcal{L}(x) = \int_x^\infty (1 - \Phi(t))dt$. Applying the inverse of the first-order loss function to both sides of the equation gives

$$\mathcal{L}^{-1}\left(\mathcal{L}\left(x\right)\right) = \mathcal{L}^{-1}\left(\int_x^\infty (1 - \Phi(t))dt\right) = x.$$

Now since the first-order loss function and its inverse are differentiable, the following holds

$$\frac{d}{dx}\left(\mathcal{L}^{-1}\left(\mathcal{L}\left(x\right)\right)\right) = \frac{d}{dx}\left(\mathcal{L}^{-1}\left(\int_x^\infty \left(1 - \Phi\left(t\right)\right)dt\right)\right) = 1.$$

Finally, the derivative of the inverse of the first-order loss function is solution to the following differential equation

$$\frac{d}{dy}\left(\mathcal{L}^{-1}(y)\right) = \frac{1}{(\Phi(\mathcal{L}^{-1}(y)) - 1)} \tag{A.1}$$

A.2.2. Proof of convexity

S_T is differentiable since $\mathcal{L}$ is differentiable. Using Equation (A.1), the derivative of S_T can be calculated as

$$\frac{d}{dA}(S_T(A)) = \sigma_{T+1} \cdot \frac{d}{dA}\left[\mathcal{L}^{-1}(\frac{(1-\beta)(\mu + A)}{\sigma_{T+1}})\right] + 1$$

$$= \sigma_{T+1} \cdot \left[\frac{1-\beta}{\sigma_{T+1}} \cdot \left(\frac{d}{dA} \left(\mathcal{L}^{-1} \right) \right) \left(\frac{(1-\beta)(\mu+A)}{\sigma_{T+1}} \right) \right] + 1$$
$$\overset{(A.1)}{=} \frac{1-\beta}{\Phi\left(\mathcal{L}^{-1}\left(\frac{(1-\beta)(\mu+A)}{\sigma_{T+1}} \right) \right) - 1} + 1.$$

The derivate of S_T is increasing since $\mathcal{L}^{-1}$ is decreasing, as it is the inverse of a decreasing function, Φ is increasing, and so $\Phi(\mathcal{L}^{-1}(\cdot))$ is decreasing. The root of the derivative of S_T can be found as

$$\begin{aligned} \frac{d}{dA}(S_T(A)) = 0 &\implies 1-\beta = -\left(\Phi\left(\mathcal{L}^{-1}\left(\frac{(1-\beta)(\mu+A)}{\sigma_{T+1}} \right) \right) - 1 \right) \\ &\implies \mathcal{L}^{-1}\left(\frac{(1-\beta)(\mu+A)}{\sigma_{T+1}} \right) = \Phi^{-1}(\beta) \\ &\implies \underline{A} = \frac{\sigma_{T+1}}{1-\beta} \cdot \mathcal{L}\left(\Phi^{-1}(\beta) \right) - \mu. \end{aligned}$$

The root $\underline{A}$ always exists since $\frac{\sigma_{T+1}}{1-\beta} \cdot \mathcal{L}\left(\Phi^{-1}(\beta)\right)$ is positive for all $\beta \in]0,1[$ and $\underline{A} \in]-\mu; \infty[$.

A.3. Proof of Proposition 2.2

The proof is similar to the additive setting and requires to exhibit a few properties of the first-order loss function of a log-normal distribution. Let ω be a random variable that follows a log-normal distribution with log-mean μ and log-variance σ^2 and let g_ω and G_ω be its p.d.f and c.d.f respectively. The first-order loss function is defined for any positive x as $\mathcal{L}(x,\omega) = \mathbb{E}\left[\max(\omega - x), 0\right] = \mathcal{L}(x,\omega) = \int_x^\infty (t-x) g_\omega(t) dt$.

Lemma A.1. *The first-order loss function $\mathcal{L}(\cdot,\omega)$ is strictly decreasing over its domain $[0,+\infty[$.*

Proof. Consider the function $h(\cdot,\omega)$ defined over $[0,\infty[$ as $h(x,\omega) = \int_x^b (t-x) g_\omega(t) dt$ where $b > x$. h is differentiable and using Leibniz's integral rule we have $\frac{dh}{dx}(x,\omega) = \int_x^b -g_\omega(t)dt = G_\omega(x) - G_\omega(b)$. Knowing that $\lim_{b\to+\infty} h(x,\omega) = \mathcal{L}(x,\omega)$ we deduce $\frac{d\mathcal{L}}{dx}(x,\omega) = G_\omega(x) - 1 < 0$. □

Lemma A.2. *The first-order loss function $\mathcal{L}(\cdot,\omega)$ is invertible and its inverse is defined on $]0, \mathbb{E}[\omega]]$.*

Proof. The function $\mathcal{L}(\cdot,\omega)$ is strictly decreasing, $\lim_{x\to+\infty} \mathcal{L}(x,\omega) = 0$ and $\mathcal{L}(0,\omega) =$

$\int_0^\infty t \cdot g_\omega(t)dt = \mathbb{E}[\omega]$. Hence the minimum inventory that satisfies the service level constraint is found by inverting the first-order loss function. □

A.3.1. Proof that S_T is increasing

Without loss of generality let x be a fixed positive real number. Let ω be a log-normal random variable defined as $\ln(\omega) \sim \mathcal{N}(\mu + A_T, \sigma_{T+1}^2)$ and denote its cumulative distribution function $F_\omega(x, A_T)$. With a small notation change, define the first-order loss function as $\mathcal{L}(x, A_T) = \int_x^\infty (1 - F_\omega(t, A_T))dt$. The cumulative distribution function can be expressed using the error function as

$$\mathcal{L}(x, A_T) = \int_x^\infty \left(1 - \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln(t) - \mu - A_T}{\sqrt{2}\sigma}\right)\right]\right).$$

This function is differentiable with regards to A_T and we can derive the partial derivative of $\mathcal{L}$ as

$$\frac{\partial \mathcal{L}(x, A_T)}{\partial A_T} = \frac{1}{\sigma\sqrt{2\pi}} \int_x^\infty \exp\left(-\left(\frac{\ln(t) - \mu - A_T}{\sqrt{2}\sigma}\right)^2\right)$$

which is positive for all values of A_T and x. The first-order loss function $\mathcal{L}(x, D_{T+1} \mid A_T)$ and its inverse are thus strictly increasing in A_T for any fixed x. Now, note that $\mathcal{L}(S_T(A_T), A_T) = (1 - \beta)\exp(\mu + A_T + \frac{\sigma_{T+1}^2}{2})$ is strictly increasing in A_T. The inventory target function S_T is then strictly increasing as the inverse of a strictly increasing function.

A.4. Proof of Proposition 2.3

The optimal target inventory as a function of the forecast update $S_t(A_t)$ in each period can be found by recursion. Define the auxiliary cost function $G_t(y_t, A_t) = c_t y_t + \mathbb{E}_{A_{t+1}|A_t}[V_{t+1}(y_t, A_{t+1})]$ such that $V_t(x_t, A_t) = \min_{x_t \le y_t \le x_t + K} G_t(y_t, A_t) - c_t \cdot x_t$. The minimum costs incurred in period T are known from Lemma 2.2. Hence, the auxiliary costs

in period $T-1$ are given by

$$\begin{aligned} G_{T-1}(y_{T-1}, A_{T-1}) &= c_{T-1} \cdot y_{T-1} + \mathbb{E}_{A_T|A_{T-1}}[V_T(y_{T-1}, A_T)] \\ &= c_{T-1} \cdot y_{T-1} + \int_{S_T^{-1}(y_{T-1})}^{S_T^{-1}(y_{T-1}+K)} c_T(S_T(a) - y_{T-1}) f_{A_T|A_{T-1}}(a) da \\ &\quad - c_T \cdot K \cdot F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1}+K)) \\ &\quad + \int_{S_T^{-1}(y_{T-1}+K)}^{+\infty} \gamma \cdot (S_T(a) - y_{T-1} - K) f_{A_T|A_{T-1}}(a) da. \end{aligned}$$

Since S_T is strictly increasing and differentiable, $G_{T-1}(y_{T-1}, A_{T-1})$ is differentiable with regards to y_{T-1} as a sum of differentiable functions. The partial derivative of G_{t-1} with regards to y_{T-1} gives the marginal costs and is determined as

$$\begin{aligned} g_{T-1}(y_{T-1}, A_{T-1}) = c_{T-1} &+ \frac{\partial}{\partial y_{T-1}} \left(\int_{S_T^{-1}(y_{T-1})}^{S_T^{-1}(y_{T-1}+K)} c_T \cdot (S_T(a) - y_{T-1}) f_{A_T|A_{T-1}}(a) da \right) \\ &- c_T \cdot K \cdot f_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1}+K)) \cdot \frac{dS_T^{-1}(y_{T-1}+K)}{dy_{T-1}} \\ &+ \frac{\partial}{\partial y_{T-1}} \left(\int_{S_T^{-1}(y_{T-1}+K)}^{+\infty} \gamma \cdot (S_T(a) - y_{T-1} - K) \cdot f_{A_T|A_{T-1}}(a) da \right). \end{aligned}$$

We use Leibniz's integral rule to find that

$$\begin{aligned} &\frac{\partial}{\partial y_{T-1}} \left(\int_{S_T^{-1}(y_{T-1})}^{S_T^{-1}(y_{T-1}+K)} c_T \cdot (S_T(a) - y_{T-1}) f_{A_T|A_{T-1}}(a) da \right) \\ &\qquad = c_T \cdot K \cdot f_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1}+K)) \cdot \frac{dS_T^{-1}(y_{T-1}+K)}{dy_{T-1}} \\ &\qquad\quad - c_T \cdot \left[F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1}+K)) - F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1})) \right] \end{aligned}$$

and

$$\begin{aligned} &\frac{\partial}{\partial y_{T-1}} \left(\int_{S_T^{-1}(y_{T-1}+K)}^{+\infty} \gamma \cdot (S_T(a) - y_{T-1} - K) \cdot f_{A_T|A_{T-1}}(a) da \right) = \\ &\qquad\qquad - \gamma \cdot \left[1 - F_{A_T|A_{T-1}}(S_T^{-1}(y_{T-1}+K)) \right] \end{aligned}$$

and deduce the expression of the marginal cost function $g_{T-1}(y_{T-1}, A_{T-1})$. Since $\gamma \geq c_{T-1}$, the marginal cost function is strictly increasing in y_{T-1}. Being first negative and then positive, it has a unique root. The inventory target $S_{T-1}(A_{T-1})$ is exactly the root

of the marginal cost function.
The inventory targets for previous periods $t < T-1$ can be found by recursion using the same principle as the proof as Biçer and Seifert (2017). Let $t \in \{1, \dots, T-2\}$ such that $S_{t+1}(A_{t+1})$ is known, strictly increasing and invertible. The cost-to-go in period $t+1$ depends on the action in period t and is given by

$$V_{t+1}(y_t, A_{t+1}) = \begin{cases} G_{t+1}(y_t, A_{t+1}) - c_{t+1} \cdot y_t, & \text{if } y_t > S_{t+1}(A_{t+1}) \\ G^*_{t+1}(A_{t+1}) - c_{t+1} \cdot y_t, & \text{if } y_t + K > S_{t+1}(A_{t+1}) \geq y_t \\ G_{t+1}(y_t + K, A_{t+1}) - c_{t+1} \cdot y_t, & \text{if } S_{t+1}(A_{t+1}) \leq y_t + K \end{cases}$$

where $G^*_{t+1}(A_{t+1}) = G_{t+1}(S_{t+1}(A_{t+1}), A_{t+1})$ is the minimum cost-to-go in period $t+1$. The auxiliary cost function in period t can then be expressed as

$$\begin{aligned} G_t(y_t, A_t) &= c_t \cdot y_t + \mathbb{E}_{A_{t+1}|A_t}[V_{t+1}(y_t, A_{t+1})] \\ &= y_{T-1}(c_t - c_{t+1}) + \int_{-\infty}^{S^{-1}_{t+1}(y_t)} G_{t+1}(y_t, a) f_{A_{t+1}|A_t}(a) da \\ &\quad + \int_{S^{-1}_{t+1}(y_t)}^{S^{-1}_{t+1}(y_t+K)} G^*_{t+1}(a) f_{A_{t+1}|A_t}(a) da + \int_{S^{-1}_{t+1}(y_t+K)}^{+\infty} G_{t+1}(y_t + K, a) f_{A_{t+1}|A_t}(a) da \end{aligned}$$

The marginal cost g_t are deduced as $g_t(y_t, A_t) = \frac{\partial}{\partial y_t} G_t(y_t, A_t)$.

A.5. Shortfall penalty costs

A.5.1. Single product: sensitivity analysis of shortfall costs

For each uncertainty resolution setting, a total of 1000 rolling-horizon simulations are run with varying shortfall penalty factor. For selected values of the penalty cost factor, the average of the key performance indicators over the simulations are given in Table A.1. If the average achieved service level is lower than the target, the statistical significance of the service-level shortfall is assessed through Student's t-test using a p-value of 0.05. Statistical significance is shown with a star symbol (*). The value of the shortfall cost factor is found to satisfy the service level target with lowest costs. The final values chosen for the penalty cost factor are shown in bold in Table A.1.

Table A.1.: Sensitivity analysis of penalty cost factor g.

	Uncertainty	g	Service level	Objective	Nervousness
Additive MMFE	Early	25	0.9433*	181.82	7.241
		50	0.9483	206.90	7.413
		75	0.9498	220.97	7.445
	Constant	5	0.9242*	167.42	5.720
		10	0.9486	200.33	7.195
		15	0.9534	212.28	7.541
	Late	4.5	0.9386*	189.89	2.879
		5	0.9447	198.21	3.309
		5.5	0.9488	204.55	3.644
Multiplicative MMFE	Early	50	0.9460*	228.67	8.306
		75	0.9480	248.35	8.361
		100	0.9498	265.96	8.365
	Constant	5	0.9199*	160.84	6.221
		10	0.9440	197.81	8.267
		15	0.9491	211.87	8.591
	Late	4.5	0.9368*	186.27	3.732
		5	0.9442	198.81	4.394
		5.5	0.9486	207.95	4.892

A.5.2. Multiple products: shortfall costs

The shortfall penalty costs are set as $\gamma^j = g^j \cdot c_t^j$ and $g^2 = 3 \cdot g^1$. The final shortfall cost value is found by using the values found in the single-product sensitivity analysis and progressively adjusting them so that the MMFE model reaches the expected service-level target for both products with minimum cost. The final values are given in Table A.2.

Table A.2.: Values of shortfall penalty factor g.

	Early	Constant	Late
Two products, negative correlation	50	30	$g^1 = g^2 = 60$
Two products, no correlation	25	20	10
Two products, positive correlation	25	20	15

Appendix B

Stochastic programming in master production scheduling

B.1. Deterministic model

The deterministic production planning and raw-material ordering problem can be formulated in Problem (B.1) as a deterministic optimisation problem.

$$\min \quad \sum_{t=1}^{T}\left(\sum_{k=1}^{K}\sum_{w=1}^{W}\mu_{k,w}i_{k,w,t}+\sum_{a=1}^{A}\nu_a\cdot y_{a,t}+\sum_{k=1}^{K}\gamma_k\cdot b_{k,t}\right) \tag{B.1a}$$

$$\text{s.t.} \quad i_{k,w,t}=i_{k,w,t-1}+\sum_{l\in\mathcal{L}_w}q_{k,l,t}-s_{k,w,t}, \quad \forall k,w,t \tag{B.1b}$$

$$f_{k,t}+ss_{k,t}=b_{k,t}+\sum_{w=1}^{W}s_{k,w,t}, \quad \forall k,t \tag{B.1c}$$

$$\sum_{k=1}^{K}q_{k,l,t}\leq\kappa_l, \quad \forall l,t \tag{B.1d}$$

$$y_{a,t}=y_{a,t-1}+z_{a,t}-\sum_{k=1}^{K}\sum_{l=1}^{L}\beta_{k,a}\cdot q_{k,l,t}, \quad \forall a,t \tag{B.1e}$$

$$q_{k,l,t},i_{k,w,t},b_{k,t},s_{k,t},y_{a,t},z_{a,t}\geq 0, \quad \forall k,w,l,a,t \tag{B.1f}$$

The objective function in (B.1a) minimises the inventory costs of finished goods and raw materials as well as the lost-sales costs over the planning horizon. The lost-sale costs relax the problem to allow feasible solutions when capacity or raw materials are insufficient to satisfy demand. As such, the lost-sales penalty cost is typically set to a high value ($\lambda = 1000$ in our numerical study). Constraint (B.1b) describes the inventory

balance of finished goods in each site. Constraint (B.1c) ensures that demand is either satisfied by each site's production or counted as lost sales. Constraint (B.1d) limits the production of each line to its capacity in each period. Constraint (B.1e) describes the raw-material inventory balance. Constraint (B.1f) specifies the domain of the continuous decision variables.

B.2. Notation of stochastic models.

Table B.1.: Notation of stochastic models

Sets	
$\mathcal{A}$	Set of raw materials $\{1, \ldots, A\}$
$\mathcal{L}$	Set of production lines $\{1, \ldots, L\}$
$\mathcal{K}$	Set of products $\{1, \ldots, K\}$
$\mathcal{K}_f$	Set of products within family f
$\mathcal{T}$	Set of time periods $\{1, \ldots, T\}$
$\mathcal{W}$	Set of production sites $\{1, \ldots, W\}$
$\mathcal{L}_w$	Set of lines $\{1, \ldots, L_w\}$ at site w
Parameters	
$f_{k,t}$	Demand forecast for product k in period t
κ_l	Capacity of line l
γ_k	Lost-sale penalty cost of product k
$\beta_{k,a}$	Consumption of raw material a per unit of product k
$\mu_{k,w}$	Inventory holding cost of product k at site w
ν_a	Inventory holding cost of raw material a
$\rho_{k,l}$	Equal to 1 if product k can be produced on line l, 0 otherwise
$x_{k,f}$	Product-to-family assignment, equal to 1 if product k is assigned to family f
Decision variables	
$q_{k,l,t}$	Production volume of product k on line l in period t
$z_{a,t}$	Order of raw material a for period t
$b_{k,t,n}$	Lost sales of product k at the end of period t in scenario n
$s_{k,w,t,n}$	Sales of product k assigned to site w in period t in scenario n
$i_{k,w,t,n}$	Inventory of product k on hand at site w at the end of period t in scenario n
$r_{k,l,t,n}$	Recourse production of product k on line l in period t in scenario n
$Y_{a,t,n}$	Inventory of raw material a at the end of period t in scenario n
$u_{k,t,n}$	Auxiliary variable to track minimum recourse production over all scenarios n for product k in period t
$v_{f,t,n}$	Auxiliary variable to track maximum recourse production over all scenarios n for family f in period t