HEC MONTRÉAL
École affiliée à l'Université de Montréal

Stochastic Lot Sizing Problems with Service Level Constraints

par
Narges Sereshti

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(spécialisation gestion des opérations et de la logistique)

Juin 2022

HEC MONTRÉAL

École affiliée à l'Université de Montréal

Cette thèse intitulée :

Stochastic Lot Sizing Problems with Service Level Constraints

Présentée par :

Narges Sereshti

a été évaluée par un jury composé des personnes suivantes :

Jean-François Cordeau
HEC Montréal
Président-rapporteur

Raf Jans
HEC Montréal
Directeur de recherche

Yossiri Adulyasak
HEC Montréal
Codirecteur de recherche

Walter Rei
Université du Québec à Montréal
Membre du jury

Céline Gicquel
Université Paris Saclay
Examinatrice externe

Marie-Ève Rancourt
HEC Montréal
Représentante du directeur de HEC Montréal

Résumé

La prise de décision sous incertitude est une tâche difficile mais inévitable dans de nombreuses applications de la chaîne d'approvisionnement. Lors de la planification de production, les décisions concernant le moment optimal et les quantités à produire sont prises lors de la résolution de ce qui est communément appelé le problème de lotissement. Le problème de lotissement est affecté par de nombreux facteurs incertains, notamment la demande des clients. Dans cette thèse, nous étudions différentes variantes du problème de lotissement stochastique sous incertitude de la demande, dans lesquelles différents types de niveaux de service sont utilisés pour faire face à cette stochasticité. Considérant le problème de lotissement stochastique avec les niveaux de service comme problème principal de ces travaux, nous en étudions différentes extensions. Chacune de ces extensions étudie différents types de flexibilité dans le système et mesure la valeur de l'ajout de ces flexibilités en termes d'économies monétaires. Plus précisément, nous étudions la valeur de la flexibilité obtenue en imposant un niveau de service agrégé sur différents produits, en appliquant une stratégie plus adaptative en réponse à la demande stochastique dans un cadre multi-niveaux, et en ajoutant la possibilité de substitution de produits. Nous appliquons différentes méthodologies de résolution et techniques d'approximation pour résoudre nos problèmes, y compris des formulations de programmation mixte en nombres entiers, un algorithme de branchement et de coupes, des approximations linéaires par morceaux et une approximation par la moyenne d'échantillon. Enfin, nous montrons que la prise en compte de ces flexibilités dans le système de production se traduit par des économies monétaires notables. Les extensions mentionnées sont organisées en trois parties distinctes comme suit.

Le premier projet généralise le problème de lotissement stochastique avec des contraintes de niveau de service dans un contexte multi-produits en considérant les niveaux de service agrégés en plus des niveaux de service individuels. Le niveau de service agrégé a une pertinence pratique dans

les situations où il existe une grande variété de produits, par exemple, pour un produit disponible en différentes couleurs ou tailles. Un niveau de service agrégé peut alors être imposé sur l'ensemble des produits, alors que des niveaux de service spécifiques sont imposés au niveau des articles individuels. Dans ce projet, différents types de niveaux de service sont considérés dans des versions individuelles et agrégées, et une stratégie de planification statique est utilisée. La stratégie statique est une stratégie dans laquelle les configurations et les décisions de production sont définies au début de l'horizon de planification, décisions qui restent fixes lorsque les demandes réelles sont observées. Les problèmes sont modélisés comme des modèles de programmation stochastique à deux étapes et ils sont approximés à l'aide de fonctions linéaires par morceaux, en raison de leur non-linéarité. Grâce à des expériences numériques, nous montrons que les niveaux de service agrégés fournissent une flexibilité qui se traduit par une réduction des coûts, comparativement à une situation où des niveaux de service traditionnels sont imposés indépendamment sur chaque produit individuel. Cette réduction des coûts varie en fonction du type de niveau de service utilisé et des différents paramètres du problème. En plus de la valeur des niveaux de service agrégés, nous montrons que lorsqu'une certaine flexibilité de planification est autorisée dans le système, nous pouvons tout de même utiliser des modèles statiques dans un horizon de temps glissant / reculant, pour surmonter la limitation inhérente de ces modèles statiques et augmenter la réactivité des modèles à la réalisation de la demande. Cela entraîne une diminution des niveaux de stocks et des coûts dans le système.

Le deuxième projet est consacré au problème stochastique de lotissement à multi-niveaux dans lequel nous avons une nomenclature (BOM) et des contraintes de capacité. Le niveau de service utilisé dans ce problème est un niveau de service orienté sur le temps et la quantité. Nous avons considéré un cadre général dans lequel, en plus des produits finaux, la demande indépendante peut également se présenter au niveau des composants. Dans ce projet, nous étudions l'intérêt d'avoir une stratégie adaptative comparativement à une stratégie statique dans le problème de lotissement stochastique multi-niveaux. Dans la stratégie statique, les décisions de configuration et de production pour tous les articles de la nomenclature restent fixes, tandis que dans la stratégie adaptative, certains ou tous les produits suivent une stratégie statique-dynamique, dans laquelle les décisions de production sont mises à jour lorsque la demande réelle est observée. Nous modélisons le problème sous la forme d'un modèle stochastique à deux étapes et le résolvons à l'aide de

modèles d'approximation par la moyenne d'échantillons dans lesquels l'incertitude est reflétée via des ensembles de scénarios discrets. Trois structures de nomenclature différentes, à savoir, série, assemblage, et générale, sont considérées. Nous montrons numériquement que l'ajout de flexibilité au système entraîne des économies monétaires et que l'ampleur de ces économies dépend de l'endroit où nous ajoutons la flexibilité dans la nomenclature. Des expériences numériques et des simulations approfondies sont réalisées pour étudier l'impact de différents paramètres, notamment le niveau de service, la structure des coûts de stockage dans la nomenclature, et le temps entre les commandes, sur les économies de coûts lorsque nous appliquons une stratégie plus adaptative.

Le troisième projet est une extension du problème de lotissement stochastique avec une contrainte de niveau de service considérant la possibilité de substitution de produits. Plus précisément, ce projet présente le problème de lotissement stochastique multi-étapes avec substitution et un niveau de service α jumelé. Nous considérons un horizon temporel infini dans lequel différentes décisions, à savoir la configuration de la production, et la quantité de production et de substitution, sont dynamiquement mises à jour lorsque la demande est révélée. Nous proposons différentes politiques d'horizon glissant pour résoudre ce problème et déterminer différentes décisions de production à chaque étape de planification. Ces politiques reposent sur des modèles de programmation mixte en nombres entiers et sont fondées sur des approximations de la version finie du problème. Le modèle de politique avec des contraintes de niveau de service est résolu à l'aide d'une méthode de séparation et coupes proposée pour les modèles avec des contraintes de chance conjointes. Nous testons différentes politiques dans une procédure à horizon glissant à l'aide d'une simulation, et les comparons en ce qui concerne leur temps d'exécution et la qualité de leur solution. A l'aide d'expériences numériques, nous montrons que la possibilité de substitution réduit considérablement le coût total, à niveau de service identique ou supérieur.

Mots-clés

Planification de la production, lotissement, demande stochastique, prise de décision dans l'incertitude, niveau de service, multi-niveaux, substitution de produits, approximation de la moyenne de l'échantillon, branche et coupe

Méthodes de recherche

Recherche opérationnelle, programmation mathématique, programmation stochastique

Abstract

Decision-making under uncertainty is a challenging but inevitable task in many supply chain applications. In production planning, decisions about the optimal timing and production quantities are known as lot sizing problems. The lot sizing problem is affected by many uncertain factors including customer demand. In this thesis, we study different variants of the stochastic lot sizing problem under demand uncertainty, in which different types of service levels are used to deal with this stochasticity. Considering the stochastic lot sizing problem with service levels as the main focus in this research, we study different extensions of it. Each of these extensions investigates different types of flexibility in the system and measures the value of adding these flexibilities in terms of cost savings. More specifically, we study the value of flexibility obtained by imposing an aggregate service level over different products, applying a more adaptive strategy in response to stochastic demand in a multi-level setting and adding the possibility of product substitution. We apply different solution methodologies and approximation techniques to solve our problems, including mixed integer programming formulations, a branch-and-cut algorithm, piece-wise linear approximations, and sample average approximation. Finally, we show that considering these flexibilities in production systems results in noticeable cost savings. The mentioned extensions are organized in three separate studies as follows.

The first study generalizes the stochastic lot sizing problem with service level constraints in a multi-product context by investigating aggregate service levels in addition to individual ones. The aggregate service level has practical relevance in situations where there is a lot of product variety, e.g., for a specific type of product that comes in different colors or sizes. An aggregated service level can then be imposed at the general product level, while specific service levels are imposed at the individual item levels. In this research, different types of service levels are studied in both individual and aggregate versions. The static strategy is considered, in which both setups

and production decisions are defined at the beginning of the planning horizon, and they remain fixed when the actual demands are observed. The problems are modeled as two-stage stochastic programming models and they are approximated using piece-wise linear functions, due to their nonlinearity. Through extensive numerical experiments, we show that the aggregate service levels will provide flexibility which results in cost reduction, as opposed to traditional service levels imposed independently on each individual item. This cost reduction varies based on the type of service level and the different parameters of the problem. In addition to the value of aggregate service levels, we show that when some planning flexibility is allowed in the system, we can still use static models in a rolling/receding horizon environment to overcome its inherent limitation, and increase the responsiveness of the models to the demand realization which leads to a decrease in inventory levels and costs in the system.

The second study is dedicated to the stochastic multi-level capacitated lot sizing problem in which we have a bill of material (BOM). The service level used in this problem is a time and quantity-oriented service level. We addressed a general setting in which, in addition to the end items, the independent demand may be present at the component levels as well. In this research, we study the value of having an adaptive strategy compared to the static strategy in stochastic multi-level lot sizing. In the static strategy, the setup and production decisions for all the items in the BOM remain fixed, while in the adaptive strategy, some or all items follow a static-dynamic strategy, in which the production decisions are updated when the demand is observed. We model the problem as a two-stage stochastic model and solve it using sample average approximation models in which the uncertainty is reflected in discrete scenario sets. Three different BOM structures, i.e., serial, assembly, and general, are considered and we numerically show that adding flexibility to the system results in cost savings and the magnitude of these savings depends on where we add the flexibility in the BOM. Extensive numerical experiments and simulations are performed to investigate the impact of different parameters including the service level, the holding cost structure in the BOM, and the time between orders on the cost savings when we apply a more adaptive strategy.

The third study is an extension of the stochastic lot sizing problem with a service level constraint considering the possibility of product substitution. More specifically, this study presents the multi-stage stochastic lot sizing problem with substitution and a joint α service level. We consider

an infinite time horizon in which different decisions, including production setup, and the amount of production and substitution are dynamically updated when the demand is realized. We propose different rolling-horizon policies to solve this problem and determine different production decisions at each planning stage. These MIP-based policies are based on approximations of the finite version of the problem. The policy model with service level constraints is solved using a branch-and-cut method proposed for models with joint chance constraints. We test different policies in a rolling horizon procedure using a simulation and compare them with respect to their execution time and their solution quality. Using numerical experiments, we show that considering substitution will reduce the total cost significantly, at the same or better service level.

Keywords

Production planning, lot sizing, stochastic demand, decision making under uncertainty, service level, multi-level, product substitution, sample average approximation, branch-and-cut

Research methods

Operations research, mathematical programming, stochastic programming

Contents

List of Tables

List of Figures

List of acronyms

B&C Branch-and-cut

DP Dynamic programming

MILP Mixed integer linrear programming

MIP Mixed integer programming

BOM Bill of material

MRP Material requirement planning

SAA Sample average approximation

To my Parents,

For their Love, Patience, Trust,

and all they have Taught me

Acknowledgements

My sincerest gratitude goes to my advisors Professors Raf Jans and Yossiri Adulyasak for their support and guidance throughout this work. I was very fortunate to have their trust and enough freedom to develop my expertise as a researcher, and their encouragement to present my work at many scientific conferences around the world. Many thanks to Raf not only for his support and dedication but more importantly, for his mentorship through this Journey. Without his encouragement, it was very difficult to accomplish this. I should also thank Yossiri for accepting me as his first Ph.D. student at HEC and pushing me toward learning new technical skills. I am also very grateful to Professor Rei as my external committee member and for his scientific advice during different milestones of my Ph.D. I would like to thank Professors Céline Gicquel and Jean-François Cordeau for accepting to be a part of my jury and for their comments and positive feedback on the content of this thesis. It has been an honor to finish my Ph.D. in their presence.

Starting with an internship supported by FRQNT, in the last 2 years of my Ph.D., I had the opportunity to work with Professor Merve Bodur from the University of Toronto. Working with Merve was the best thing that happened to me during the pandemic. By having regular meetings and attending her classes and group meetings virtually, I got this opportunity to expand my knowledge of new topics in stochastic programming. She was a great mentor with wise and supportive advice for my future. During this period, I had also the honor of collaborating with Professor James Luedtke from Wisconsin University. Every single session with him was a class for me, where he walked me patiently through very difficult topics. I truly appreciate their dedication to the third research of my thesis.

I am very blessed for being affiliated with several academic institutions. Many thanks to faculty and staff members at HEC Montréal especially Professors Gilbert Laporte, and Claudia Rebolledo for what I learned from them and Line, Nathalie, and Julie for their professional support. I should

also thank the department of Logistics and Operations Management for giving me invaluable teaching opportunities and Claire Poitras for her guidance through this process. During my Ph.D. I was also very fortunate to be a member of two prestigious research centers, CIRRELT and GERAD. Thanks to them, I found many friends and a huge professional networking opportunity around the world, in different universities and many industries. A big thanks to Matthieu, Teodora, Masoud, Rahim, Khalid, Gislaine, Borzoo, Okan, Mehdi, Sajad, Mahdis, and Payman for the insightful discussions and judicious advice. I should also thank Serge, Khalid, and Guillaum for their technical assistance. It is impossible to name all staff and faculty members and friends, but I sincerely thank every single one for their support and all the pleasant moments I had with them.

I should also acknowledge the different sources of funding for my research; the Chair in Supply Chain Operations Planning at HEC Montréal, the Canada Research Chair in Supply Chain Analytics, the National Science and Engineering Research Council, the Fonds de recherche du Quebec, HEC Montréal Ph.D. fellowship and awards, CIRRELT scholarship for excellence in research, and Gerad fundings.

Last but not least, I would like to thank my family for their unconditional love and encouragement. To my father, thank you for believing in me even when I no longer did and instilling in me a love for learning. To my mother, for giving me an immense amount of love and independence and for being a role model for kindness and wisdom. To Nafiseh and Nikoo, thanks for your constant presence and your sisterly warmth which was always my fuel to continue. How fortunate I am to have you.

General Introduction

For companies manufacturing products, having an efficient Material Requirement Planning (MRP) system is important to minimize different costs in the production system. A central component of an MRP system is the production lot sizing problem. In dynamic lot sizing problems, proper inventory control and production decisions are crucial to achieve a balance between customer demand satisfaction and cost management. While insufficient inventory will lead to shortages, unnecessary stocks will increase the holding cost, which is charged for the quantity being stored at the end of each period. Furthermore, in each period in which production occurs, a setup has to be performed which incurs a fixed setup cost. The basic lot sizing problem is a multi-period production planning problem which considers the trade off between the setup costs and inventory holding cost and its goal is to define the optimal timing and quantity of production to minimize the total cost over a finite and discrete time horizon (Pochet and Wolsey, 2006). The lot sizing problem has been extensively studied and applied in real world situations and extended to include several practical cases such as multiple products, capacitated machines, or backlog costs (Jans and Degraeve, 2008).

The standard assumption in the basic lot sizing problem is that the demand is known. However, the plans created by these techniques are often very sensitive to even a small change in the demand. Decisions that do not incorporate uncertainty are known to be inferior and sometimes costly and meaningless compared to decisions resulting from models in which the uncertainty is explicitly taken into account. In this thesis, three different extensions of the lot sizing problem with demand uncertainty are investigated. In this section, we will explain briefly two common concepts which are important in all three articles in the thesis, and then introduce each of these papers, separately. The first concept is the service level and the second one relates to the possible strategies which can be used in stochastic lot sizing problems.

One common approach that the planners use to deal with uncertainty is to impose some demand

fulfillment criteria known as service levels to satisfy the uncertain demand. There are different types of service levels which can be classified into *event-oriented*, *quantity-oriented*, and *time and quantity-oriented* service levels. For example, the α service level is an *event-oriented* service level which puts limits on the probability of stock outs. The β service level put limits on the amount of unsatisfied demand in each period, and is hence a *quantity-oriented* service level. The γ service level put limits on the backlog which is the cumulative unsatisfied backorder. There are different forms of service levels. Some are defined for each period separately, and some over the whole planning horizon. Different types of service levels are investigated in this research. The details of the service levels and their formulations used in our research will be explained separately in the corresponding chapters.

The other concept which is important throughout this thesis is the strategy which is used in case of uncertainty. There are three main strategies in stochastic lot sizing problems which differ based on the timing of the setup and production decisions. These strategies are *static*, *dynamic*, and *static-dynamic* strategy (Bookbinder and Tan, 1988). In the *static* strategy, both the setup and production decisions are determined at the beginning of the planning horizon and they remain fixed when the demand is realized. In the *dynamic* strategy, both the setup and production decisions can be dynamically changed to react to the demand realizations. The *static-dynamic* strategy is a combination of these two strategies in which the setups are fixed at the beginning of the planning horizon and the production decisions can be adjusted when the demands are realized.

Each of these strategies has specific characteristics and is suitable for different situations. The *dynamic* strategy is the most responsive strategy which results in the lowest cost among the three. However, in practice, it may lead to high variations in setups and production quantities. This may be undesirable in some situations. A *static* strategy results in plans which do not exhibit nervousness (Tunc et al., 2013), since the setups and the production quantity, will remain the same regardless of the demand realization. The *static-dynamic* strategy is a compromise between the two extremes and compared to the *static* strategy results in less cost as it is more responsive, and more cost and less nervousness compared to the *dynamic* strategy.

Each of the chapters in this thesis considers a different extensions related to the lot sizing problem with stochastic demand and service levels.

In Chapter 1, we investigate the value of aggregate service levels in a two-stage stochastic lot

sizing problem. The service levels which are mentioned before are usually defined for each of the products separately, in the literature. In this work, we study the aggregate service levels which are defined jointly over all different products. Through extensive numerical experiments, we show that these joint service levels will provide flexibility which results in cost reduction, as opposed to traditional service levels imposed independently on each item. This cost reduction varies based on the type of service level and the parameters of the problem. All three categories of service levels, i.e., *event-oriented*, *quantity-oriented*, and *time and quantity-oriented* service levels are studied in this research. The strategy which is used in this research is the static strategy. The problems are modeled as two-stage stochastic models. Due to the fact that some service levels result in nonlinear functions for some of the service levels, such models are approximated and solved using piece-wise linear functions.

One of the disadvantages of the static strategy is that this strategy is not responsive to demand realizations, which potentially leads to large inventory levels and costs in the system. In this research, we will show that we can still use *static* models in a rolling/receding horizon environment to overcome some of their inherent limitations, when we allow production recourse decisions. In such an implementation, only the decisions related to the first planning periods under the static strategy are implemented. The planning horizon is next moved forward and the input parameters are updated. This process is next repeated until the end of planning horizon. This implementation leads to a plan in which production quantities in further periods can still be changed, and hence results in a reduction in the inventory levels and total cost.

The second chapter is dedicated to another extension of the stochastic lot sizing with service levels. In this research, we study the stochastic multi-level capacitated lot sizing problem in which we have a bill of material (BOM). The service level used in this problem is a *time and quantity-oriented* service level. We address a more general setting in which, in addition to the end items, their components can also have independent demand. We investigate and compare two different strategies, which are the *static* strategy and a more adaptive one in which we apply the *static-dynamic* strategy for some or all of the items. We model the problem as a two-stage stochastic model and solve it using sample average approximation models in which the uncertainty is reflected using a set of discrete scenarios. We numerically show that adding flexibility to the system can result in cost savings depending on the flexibility level in the BOM.

The third chapter considers another important extension to the stochastic lot sizing problem by considering substitution. More specifically, this chapter presents the multi-stage stochastic lot sizing problem with substitution and joint α service level. In this research, we consider the *dynamic* strategy in which both the setup and production decisions can be updated after the demands realization. We propose a dynamic programming formulation for this problem and apply different policies to determine different production decisions including setups, production, and substitution at each planning stage. We test the problem in a rolling horizon procedure using a simulation. In the following chapters, we will explain each of these research projects separately.

References

Bookbinder, J. H. and Tan, J.-Y. (1988). Strategies for the probabilistic lot-sizing problem with service-level constraints. *Management Science*, 34(9):1096–1108.

Jans, R. and Degraeve, Z. (2008). Modeling industrial lot sizing problems: a review. *International Journal of Production Research*, 46(6):619–1643.

Pochet, Y. and Wolsey, L. A. (2006). *Production Planning by Mixed Integer Programming*. Springer, Science & Business Media, New York.

Tunc, H., Kilic, O. A., Tarim, S. A., and Eksioglu, B. A. (2013). simple approach for assessing the cost of system nervousness. *International Journal of Production Economics*, 141(2):619–625.

Chapter 1

The value of aggregated service levels in stochastic lot sizing problems

Chapter information: This chapter has been published as a research article in Omega: Sereshti, Narges, Adulyasak, Yossiri and Jans, Raf, The value of aggregated service levels in stochastic lot sizing problems, Omega, 102335, (2021) *Awarded the Esdras-Minville best student paper award in 2022, HEC Montréal

Abstract

Dealing with demand uncertainty in multi-item lot sizing problems poses huge challenges due to the inherent complexity. The resulting stochastic formulations typically determine production plans which minimize the expected total operating cost while ensuring that a predefined service level constraint for each product is satisfied. We extend these stochastic formulations to a more general setting where, in addition to the individual service level constraints, an aggregate service level constraint is also imposed. Such a situation is relevant in practical applications where the service level aggregated from variety of products must be collectively satisfied. These extended formulations allow the decision maker to flexibly assign different individual service levels to different products while ensuring that the overall aggregate service level is satisfied and these aggregated service level measures can be used in conjunction with the commonly adopted individual service levels. Different mathematical formulations are proposed for this problem with different types of

service levels. These formulations are a piece-wise linear approximation for the β, γ, and δ service levels and a quantile-based formulation for the α_c service level. We also present a receding horizon implementation of the proposed formulations which can be effectively used in a dynamic environment. Computational experiments are conducted to analyze the impact of aggregate service levels and demonstrate the value of the proposed formulations as opposed to standard service levels imposed on individual items.

1.1 Introduction

In dynamic lot sizing problems, proper inventory control and production decisions are crucial to achieve a balance between customer demand satisfaction and cost management. While insufficient inventory will lead to shortages, unnecessary stocks will increase the holding cost. An inventory holding cost is charged for the quantity being stored at the end of each period. Furthermore, in each period in which production occurs, a setup has to be performed which incurs a fixed setup cost. The basic lot sizing problem hence considers the trade-off between setup costs and inventory holding costs. The goal of the standard lot sizing problem is to determine the optimal timing and production quantities in order to satisfy a known demand over a finite and discrete time horizon (Pochet and Wolsey, 2006). The lot sizing problem has been extended to include several practical cases such as multiple products, capacitated machines, or backlog costs (Jans and Degraeve, 2008).

While the standard assumption in lot sizing problems is that all the parameters are deterministic, it is inevitable that some parameters are actually uncertain in practice. From a practical point of view, even a small level of uncertainty may heavily affect the nominal solution determined by a deterministic model and make it infeasible or more costly than anticipated (Ben-Tal et al., 2009). To deal with the uncertainty in demand, safety stock levels are usually predetermined for each item under strict assumptions such as stationary demand, normality, as well as the independence of demand. The decisions resulting from models that do not incorporate uncertainty are known to be sub-optimal compared to the solution of the models in which the uncertainty has explicitly been taken into account (Mula et al., 2006). Consequently, there is a need to have methods to mitigate the risk of uncertainty and simultaneously determine the time-dependent lot size and buffer stock decisions in the dynamic lot sizing problem.

The stochastic lot sizing problem is an extension of the deterministic case in which the problem is to determine the production schedules and quantities to satisfy stochastic demand over a finite planning horizon. In the context where the planner must ensure that a service level is satisfied, the objective is to minimize the total expected cost whereas the decisions are subject to certain demand fulfillment criteria (Tempelmeier, 2007). These criteria are usually modeled as chance constraints in which the probability of reaching a service level must be greater than or equal to a predefined value (Brahimi et al., 2017). In this paper, several service level measures have been investigated. The α service level considers the probability of no stock out during the production or procurement cycle, the β service level or the fill rate is the proportion of the demand directly filled from stock, the γ service level limits the proportion of expected backlog to expected demand, and the δ service level limits the proportion of total expected backlog to the maximum expected backlog. These service levels are typically defined for each product separately.

In this research, we study an aggregate service level which is defined aggregately for multiple products in addition to individual service levels when uncertainty in the demand is present (Akçay et al., 2016). Such a situation is relevant in practical applications where there is a lot of product variety. For a specific type of clothing that comes in different colors or sizes, an aggregated service level can be imposed at the product level (i.e., for a specific piece of clothing), while specific service levels are imposed at the individual levels (i.e., for the different sizes). Consider a situation where a firm is concerned with its aggregate service level across multiple products. While it is clear that an aggregate service level of, for example, 95% can be achieved by imposing an individual service level of 95% for each item, this solution does not take advantage of the possible flexibility to have different individual service levels. The firm can impose a specific aggregate service level (e.g. 95%) while also imposing individual service levels which are less strict (e.g. 90%). This provides the flexibility to have a solution in which the resulting individual service levels for some products are less strict than the imposed aggregate level, while others are stricter. This flexibility can result in an overall cost reduction.

Different mathematical models are proposed to approximate this problem when considering different types of service level. These formulations are a piece-wise linear approximation and a quantile-based formulation. The contributions of this paper are as follows. The first contribution of this paper is to propose the idea of an aggregate service level for the stochastic lot sizing prob-

lem which generalizes the formulations presented in the literature. Next, we propose mathematical formulations to model an aggregate service level for different types of service levels considered in the literature (i.e., the α, β, γ, and δ service levels). In this problem, the imposed aggregate service levels allow the flexibility to choose a separate service level for each item and can be used in conjunction with minimum individual service level constraints used in the literature. While the aggregate service levels are parameters in the model, the individual service levels are a result of the optimization process. The proposed piece-wise linear approximation formulations for the β, γ, and δ service levels in this research are extensions of existing formulations, while the quantile-based formulation is newly proposed for these problems. Finally, we provide computational experiments and investigate the value of the aggregate service level in different situations. Another strength of the paper is the use of a unified simulation procedure for the evaluation of the approximation formulations in order to have a fair comparison of the different models and service levels. In the situations in which some level of nervousness is acceptable, the formulations under the static uncertainty can be used in a rolling/receding horizon environment to overcome some of their inherent limitations. The analysis of the application of the proposed formulations in a receding horizon fashion is also another contribution of this paper.

This paper is structured into nine different sections. In the second section, we review the existing literature. In Section 3, different aggregate service levels are introduced. Sections 4 to 6 are dedicated to the formulations for the various aggregate service levels. In each of these sections, the mathematical models and different approximations are presented. Section 7 discusses the implementation of the model in a receding horizon environment. Section 8 discusses the experimental results. Section 9 concludes the paper and discusses possible future research.

1.2 Literature review

Although imposing individual demand fulfillment criteria is the most common approach to deal with multiple products in inventory management, the idea of defining an integrated service level has been investigated in the inventory management literature. Kelle (1989) stated that having different items with different demand, cost, and delivery characteristics, requires different service levels, and defining fixed service levels for different groups of items to insure an aggregate service

level is a challenge. A common approach to deal with the huge numbers of products or stock-keeping units (SKUs) in the inventory system in many real cases, is using ABC classification to group the items. Companies usually impose a fixed service level for all the products in the same group. Teunter et al. (2010) showed that this approach based on the classical classifications results in solutions which are far from the optimal solutions. They introduce a more efficient approach for ABC classification in which they define an overall fill rate over all SKUs, but they did not consider a setup cost in their calculations (Teunter et al., 2010). Akçay et al. (2016) introduce a multi-product, joint service-level model for an inventory control problem without any lot sizing decisions. They used *order fill rate, line item fill rate* and *dollar fill rate*. Each of these service levels is joint across random customer orders with different products and correlated demands. Escalona et al. (2019) investigated the effect of not having similar service levels for fast-moving items under different inventory policies. In their study, they consider different types of service levels (α and β) and propose different models for the combination of them for two categories of items belonging to a different customers' class. Shivsharan (2012) considered an inventory control system for a large number of spare parts with highly random and in some cases sparse demand. He mentioned that in such a case achieving the desired service level imposes a huge inventory cost. To deal with this problem, he proposed a model to minimize the safety stock cost while achieving an aggregate service level. Gruson et al. (2018) and Stadtler and Meistering (2019) investigate different types of service levels including various aggregate service levels in the deterministic lot sizing problem.

It is worthwhile to mention that there is a difference between joint and aggregate service levels. Although both of these ideas consider multiple products simultaneously, in the literature, the joint service levels refer to the case where the service level requirements are imposed on all of the products simultaneously as chance-constraints based on the joint distributions of product demands whereas the aggregate service levels we consider here refer to the case where the constraints are imposed on aggregated values of the service levels associated with individual products. The use of the aggregate service levels allows us to extend the models with individual service levels in a scalable manner to deal with a practical case where companies must ensure that the aggregate service level of a group of products is collectively satisfied. Assuming the same value for joint, aggregate, and individual service levels, the joint service levels results in more strict constraints compared to individual service level while the aggregate service level results in more relaxed and

flexible constraints. In addition, the models with joint service level can be much more difficult to solve and not tractable (Jiang et al., 2017).

In the literature, several service level measures have been proposed when dealing with demand uncertainty (Helber et al., 2013). The α service level ensures that the probability of no stock out during the production or procurement cycle is more than α. The β service level or the fill rate is the proportion of the demand directly filled from stock and it is equal to one minus the expected backorders to the expected demand. The γ service level is one minus the proportion of expected backlog to expected demand. Note that while the γ service level considers backlog, the β service level deals with backorders. The backorder level in period t is the quantity of unmet demand in period t whereas the backlog in period t represents the cumulative backorders from period 1 to period t that have not been filled by the end of period t (Gade and Küçükyavuz, 2013). The δ service level ensures that the proportion of total expected backlog to the maximum expected backlog is less than or equal to $1-\delta$. It is stated that this service level transparently considers the amount of backlog and the waiting time together (Helber et al., 2013). These service levels are defined for each of the products individually. These service levels and their mathematical representations will be further discussed in Section 3.

Many papers studied the lot sizing problem with service level constraints using different strategies and different types of service levels (Tempelmeier, 2007; Helber et al., 2013; Tempelmeier, 2011; Tempelmeier and Herpers, 2011; Tempelmeier and Hilger, 2015; Tunc et al., 2014). Table 1.1 summarizes the most relevant papers. As can be seen in this table, none of the reviewed papers consider the aggregate service level in a stochastic context. This research addresses this gap in the literature.

Bookbinder and Tan (1988) investigated three different strategies to deal with a probabilistic single-stage lot sizing problem with service level constraints. The first strategy is the *static* uncertainty in which the decisions for all periods are made at the beginning of the planning horizon and cannot be changed. These decisions are the setup and production level decisions. In the second strategy, which is called *dynamic* uncertainty, the setups and production levels are decided dynamically as the information is revealed during the planning horizon. The third strategy is the combination of the two previous strategies in which the setup periods are determined at the beginning of the planning horizon and remain fixed, whereas the production quantities are determined

dynamically depending on the realized demand. This strategy is called the *static-dynamic* strategy. Each of these strategies has specific characteristics and are suitable for different situations. In this research, we select the static strategy. A *static* strategy results in plans which do not exhibit any nervousness (Tunc et al. (2013), Koca et al. (2018)), since the production plan, both with respect to the setups and the production quantity, will remain the same regardless of the demand realization. In addition, among the three strategies only the *static* strategy is able to deterministically consider the capacity requirements (Tempelmeier, 2013). In other words, in this strategy once a feasible production plan has been found, it will be fixed and the production quantities remain the same, and hence the capacity limitation will not be violated (Tempelmeier, 2011). As can be seen in Table 1.1, all of the papers with capacity constraints are analysed under the *static* strategy. For sure, the *static-dynamic* uncertainty strategy results in less cost as it is more responsive, however, in practice, it may lead to high variance in production quantities which may be undesirable in some situations. For example, in MRP systems, changes in the production of the parent item lead to changes in the replenishment of its components, and there may be some negative consequences for the whole supply chain due to the bullwhip effect. Second, the random changes in the timing and quantity of production results in random resource requirements, which is referred to in the literature as planning nervousness. In some cases it may make the problem infeasible, the planned due dates may be missed, and the changes may be also unfavorable or prohibited by labour agreements (Tempelmeier, 2013).

In the situations in which production quantity fluctuations are acceptable, we can still use *static* models in a rolling/receding horizon environment to overcome some of their inherent limitations. In such an implementation, only the decisions related to the first (or first few) planning period(s) are implemented. The planning horizon is next moved forward and the information is updated. This process is next repeated. The difference between the rolling and receding horizon approach is that in the rolling horizon approach the end of the planning horizon is also moved, while in the receding horizon the end of the planning horizon remains fixed. This implementation will automatically lead to a plan in which production quantities in further periods can still be changed. While it is true that for a given static horizon, the *static-dynamic* model will result in lower costs compared to the *static* models, Bookbinder and Tan (1988) conclude (for the α service level) that "this advantage will be lost in the rolling schedule situation", and hence it is sensible to use a *static*

model in a rolling horizon fashion. Dural-Selcuk et al. (2019) showed for the single-item non-stationary stochastic lot sizing problem with backorders that the performance of both the *static-dynamic* and *static* policies is improved in a receding horizon. The authors also mentioned that the improvement is substantial for the *static* policy and the performance difference between these policies becomes very small in a receding horizon. Bookbinder and Tan (1988) discussed how to implement the static uncertainty strategy in a rolling-schedule environment for the α service level. Meistering and Stadtler (2017) proposed a stabilized-cycle strategy which combines the idea of *static* strategy and rolling schedule and come up with the concept of *stabilized-cycle* to use the production stability of the former and the ability of responding to uncertain data of the latter. In this research we also explain how to apply the proposed model in a receding horizon plan.

Table 1.1: Overview of literature on lot sizing with service level constraints

Authors	Strategy			Service level type				Individual	Aggregate	Capacity
	static	dynamic	static-dynamic	α	β	γ	δ			
Bookbinder and Tan (1988)	+	+	+	+				+		
Tarim and Kingsman (2004)			+	+				+		
Tempelmeier (2007)			+	+	+			+		
Tempelmeier and Herpers (2010)	+				+			+		+
Tempelmeier (2011)	+				+			+		+
Tempelmeier and Herpers (2011)	+				+			+		
Gade and Küçükyavuz (2013)						+		+		
Helber et al. (2013)	+						+	+		+
Tunc et al. (2014)			+	+				+		
Rossi et al. (2015)			+	+	+			+		
Tempelmeier and Hilger (2015)	+				+			+		+
Tunc et al. (2018)			+	+	+			+		
Gruson et al. (2018)		Deterministic		+	+		+	+	+	+
Stadtler and Meistering (2019)		Deterministic		+	+	+		+		+
Our Work	+			+	+	+	+	+	+	+

1.3 Individual and aggregate service levels

In all the models proposed in this paper, a static strategy in which all the decisions are made at the beginning of the planning horizon is considered and the production quantity decisions cannot be changed when demands are realized. In addition to the deterministic multi-item lot sizing problem assumptions, we assume that the demand for different products in different periods is not known, but the distributions are known and they are independent for each product. In the case of a stock out, the unmet demand is backlogged and fulfilled as soon as possible.

In this research, we investigate four different types of aggregate service levels. These service levels are based on the α_c, β, γ, and δ service levels (Tempelmeier, 2013). Table 1.2 defines these different types of service levels in separate and aggregate format. Some of these service levels are defined globally over the whole planning horizon (Tempelmeier, 2013) and some others are imposed for each planning period. The α_c is the minimum service level provided in each planning period considering the total planning horizon. It is also possible to consider the average service level across the planning horizon which is called α_p (Tempelmeier, 2013). Let K be the set of products and T the set of time periods. The first type of service level is the β service level which is a quantity oriented service level. Considering $\overline{BO}_{kt}$ as the backorder and $\overline{D}_{kt}$ the demand of product k in period t, $E[\overline{BO}_{kt}]$ and $E[\overline{D}_{kt}]$ are the expected backorder and expected demand for product k in period t, respectively. The aggregate service level in the global case and based on the β service level is equal to 1 minus the total expected backorders for all products in all planning periods divided by the total average demand over all products and periods. Another format of this service level is β_p which is imposed in each period. This service level considers the expected backorder divided by the expected demand in each planning period.

The second type of aggregate service level is based on the γ service level which is time and quantity oriented. This service level in the global case is equal to one minus the total expected backlog divided by total expected demand (Helber et al., 2013). In this service level $\overline{B}_{kt}$ is the backlog and $E[\overline{B}_{kt}]$ is the expected backlog for product k in period t. We can define γ_p as the gamma service level per period.

The third type of aggregate service level is based on the δ service level which is equal to 1 minus the total expected backlog divided by the total maximum expected backlog (Helber et al.,

Table 1.2: Different types of service level and their separate and aggregate forms

SL	Separate	Aggregate
	Quantity oriented service level	
β	$\frac{\sum_{t\in T} E[\overline{BO}_{kt}]}{\sum_{t\in T} E[\overline{D}_{kt}]} \leq 1-\beta \quad \forall k \in K$	$\frac{\sum_{t\in T}\sum_{k\in K} E[\overline{BO}_{kt}]}{\sum_{t\in T}\sum_{k\in K} E[\overline{D}_{kt}]} \leq 1-\beta$
β_p	$\frac{E[\overline{BO}_{kt}]}{E[\overline{D}_{kt}]} \leq 1-\beta_p \quad \forall k \in K, \forall t \in T$	$\frac{\sum_{k\in K} E[\overline{BO}_{kt}]}{\sum_{k\in K} E[\overline{D}_{kt}]} \leq 1-\beta_p \quad \forall t \in T$
	Time and quantity oriented service level	
γ	$\frac{\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T} E[\overline{D}_{kt}]} \leq 1-\gamma \quad \forall k \in K$	$\frac{\sum_{t\in T}\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K} E[\overline{D}_{kt}]} \leq 1-\gamma$
γ_p	$\frac{E[\overline{B}_{kt}]}{E[\overline{D}_{kt}]} \leq 1-\gamma_p \quad \forall k \in K, \forall t \in T$	$\frac{\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{k\in K} E[\overline{D}_{kt}]} \leq 1-\gamma_p \quad \forall t \in T$
δ	$\frac{\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T}(T-t+1)E[\overline{D}_{kt}]} \leq 1-\delta \quad \forall k \in K$	$\frac{\sum_{t\in T}\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K}(T-t+1)E[\overline{D}_{kt}]} \leq 1-\delta$
δ_p	$\frac{E[\overline{B}_{kt}]}{\sum_{j=1}^{t} E[\overline{D}_{kj}]} \leq 1-\delta_p \quad \forall k \in K, \forall t \in T$	$\frac{\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{j=1}^{t}\sum_{k\in K} E[\overline{D}_{kj}]} \leq 1-\delta_p \quad \forall t \in T$
	Event oriented service level	
α_c	$\min_{t\in T}(pr(I_{k0}+\sum_{j=1}^{t}(x_{kj}-\overline{D}_{kj})\geq 0)) \geq \alpha_c \quad \forall k \in K$	$\sum_{k\in K} w_k \min_{t\in T}(pr(I_{k0}+\sum_{j=1}^{t}(x_{kj}-\overline{D}_{kj})\geq 0)) \geq \alpha_c^{agg}$

2013) in the global case. This service level is also a time and quantity oriented service level. The δ_p service level is defined as the δ service level per period. This service level is 1 minus the expected backlog in each period divided by the maximum expected possible backlog until period t. The maximum expected backlog in period t is equal to the cumulative expected demand until period t.

The fourth type of service level is the α service level which ensures that the probability of having a stock-out for each product is less than or equal to $1-\alpha$. This service level is an event oriented one which is typically measured over each replenishment cycle (α_c). In order to model this in a multi-period problem, the required service level must be imposed in each specific period and the resulting service level is the minimum service level over all planning periods (Tempelmeier, 2013). The initial inventory of product k is indicated by I_{k0}. Considering w_k as the non-negative

weight of each product such that $\sum_{k \in K} w_k = 1$ and x_{kt} the production for product k in period t, the separate and aggregate version of this service level are shown in Table 1.2. The aggregate service level guarantees that the weighted average of the resulting individual service levels is at least α_c^{agg} ($\alpha_c^{agg} \in [0,1]$).

For the global aggregate β, γ, and δ service levels and for specific weights (w_k), the weighted sum of the separate service levels is equal to the aggregate service level as defined in Table 2. More explanations and the proofs are given in Appendix A. For the per period service levels, we can define such weights for each planning period, as there are T constraints for each service level (See Appendix A). Note that the aggregate version of the α_c service level is already defined as a weighted sum of separate service levels.

It is also possible to define different aggregate service levels for different product families. In this case, the service levels are aggregately defined over all products within a family of products (See Appendix B).

Before moving forward to the mathematical models of each service level, we will provide an example based on the β service level. The main purpose of the example is to illustrate the new concepts of aggregate service levels and compare it to the traditional individual service levels. Table 1.3 provides this example with 5 products and 5 periods for 3 different situations. The first column shows the result for the 95% service level imposed for each individual product. The second and third columns show the result for the less tight individual service levels of 90% and 85%, respectively, in addition to an aggregate service level of 95% which is defined over all SKUs. Adding the flexibility of the aggregate service level to the model results in a cost reduction and different individual service levels. In the case of aggregate service level, the assigned individual service levels are not arbitrary and are the result of the optimization process. In this example, the SKUs are sorted based on their holding costs. The first SKU has the lowest holding cost and the last one has the highest. The model satisfies the aggregate service level by stocking less of the products with a high inventory cost, leading to a lower individual service level for these products. At the same time, the model compensates this by stocking more of the products with a low inventory cost, leading to a higher individual service level for these products. Hence, adding the flexibility of an aggregate service level results in a 7% cost reduction and an increase in the service level for 3 products at the expense of a service level reduction for two other products.

Table 1.3: Small example with β separate and aggregate service levels

		Separate 95%	Aggregate 95% Separate 90%	Aggregate 95% Separate 85%
Individual service level	SKU 1	95%	99%	99%
	SKU 2	95%	98%	99%
	SKU 3	95%	98%	97%
	SKU 4	95%	90%	94%
	SKU 5	95%	90%	85%
Aggregate Service Level		95%	95%	95%
Total Cost		23,563	23,165	22,023
Cost Decrease		0%	2%	7%

1.4 Models with aggregate β service levels

In this section, we investigate the aggregate β service level, imposing that the total expected amount of backorder divided by the total expected demand should be less than a predefined percentage. The expected inventory and backorder in each planning period is a non-linear function of the cumulative production in each planning period. To solve this problem we use a piece-wise linear approximation.

1.4.1 Problem definition and mathematical model

The parameters and decision variables are presented in Table 1.4. The mathematical model for the stochastic capacitated lot sizing problem with aggregate β service level is as follows:

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}E[\bar{I}_{kt}]) \tag{1.1}$$

subject to:

$$\bar{I}_{k,t-1} + x_{kt} + \bar{B}_{kt} = \bar{I}_{kt} + \bar{D}_{kt} + \bar{B}_{k,t-1} \quad \forall t \in T, \forall k \in K \tag{1.2}$$

$$x_{kt} \leq My_{kt} \quad \forall t \in T, \forall k \in K \tag{1.3}$$

$$\sum_{k\in K}(st_{kt}y_{kt} + pt_{kt}x_{kt}) \leq Cap_t \qquad \forall t \in T \tag{1.4}$$

Table 1.4: Parameters and decision variables of the models with β service level

Sets	
K	Set of products
T	Set of planning periods
Parameters	
β	Target fill rate as an aggregate service level
cap_t	Production capacity in period t
hc_{kt}	Inventory holding cost for product k in period t
I_{k0}	The initial inventory for product k
M	A sufficiently large number
pt_{kt}	Unit production time for product k in period t
sc_{kt}	Setup cost for product k in period t
st_{kt}	Setup time for product k in period t
Random variables	
$\overline{B}_{kt}$	Amount of backlog for product k at the end of period t
$\overline{BO}_{kt}$	Amount of backorder for product k at the end of period t
$\overline{D}_{kt}$	Demand for product k in period t (model input)
$\overline{I}_{kt}$	Amount of physical inventory for product k at the end of period t
Decision variables	
x_{kt}	Amount of production for product k in period t
y_{kt}	Binary variable which is equal to 1 if there is a setup for product k in period t, 0 otherwise

$$E[\overline{BO}_{kt}] = E[\max\{0, \sum_{j=1}^{t}(\overline{D}_{kj} - x_{kj}) - I_{k0}\}] - E[\max\{0, \sum_{j=1}^{t-1}\overline{D}_{kj} - \sum_{j=1}^{t}x_{kj} - I_{k0}\}] \quad \forall t \in T, \forall k \in K \tag{1.5}$$

$$\frac{\sum_{t\in T}\sum_{k\in K}E[\overline{BO}_{kt}]}{\sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]} \leq 1-\beta \tag{1.6}$$

$$y_{kt} \in \{0,1\} \quad \forall t \in T, \forall k \in K \tag{1.7}$$

$$x_{kt} \geq 0 \quad \forall t \in T, \forall k \in K \tag{1.8}$$

$$\overline{I}_{kt} \geq 0 \quad \forall t \in T, \forall k \in K \tag{1.9}$$

$$\overline{B}_{kt} \geq 0 \quad \forall t \in T, \forall k \in K \tag{1.10}$$

The objective function of the model (1.1) minimizes the setup and expected inventory holding costs. Constraints (1.2) are the flow conservation constraint. Constraints (1.3) guarantee the setup forcing in case there is production. Constraints (1.4) enforce the capacity limitation. Constraints (1.5) calculate the expected backorder level for product k in period t (Van Pelt and Fransoo, 2018). The first part calculates the backlog in period t, and the second part calculates the amount of

cumulative demand until the period $(t-1)$ which is not satisfied by the cumulative production up to period t . This calculation is based on the FIFO assumption. Constraint (1.6) ensures the aggregate β service level. Constraints (1.7) to (1.10) show the domain of the different variables in the model. In this model the expected value of inventory level ($E[\bar{I}_{kt}]$) can also be calculated by equation (1.11) instead of having constraint (1.2).

$$E[\bar{I}_{kt}] = E[\max\{0, I_{k0} + \sum_{j=1}^{t}(x_{kj} - \overline{D}_{kj})\} \quad \forall t \in T, \forall k \in K \tag{1.11}$$

As can be seen, in this mathematical formulation, $E[\overline{BO}_{kt}]$ and $E[\bar{I}_{kt}]$ are non-linear functions of the cumulative production which are shown in (1.5) and (1.11), respectively. In the next section, a mathematical model is presented to approximate this non-linear model.

1.4.2 Piece-wise linear approximation

Rossi et al. (2014) performed a comprehensive study on the first order loss function and its linear upper and lower bounds. They also presented efficient piece-wise linear upper and lower bounds for normally distributed random variables which have been used in piece-wise linear approximation models for stochastic lot sizing problem with demand uncertainty (Rossi et al., 2015; Tempelmeier and Hilger, 2015; Tunc et al., 2018; Van Pelt and Fransoo, 2018). Considering the static strategy, the formulation presented here is an extension of the model proposed by Van Pelt and Fransoo (2018), in which, the expected inventory, backlog, and backorder are calculated by (1.12), (1.13), and (1.14) respectively. Q_{kt} is the cumulative production of product k up to period t, which is equal to $\sum_{j=1}^{t} x_{kj}$. $\overline{CD}_{kt}$ is the cumulative demand for product k until period t, and $\Gamma^1_{\overline{CD}_{kt}}(Q_{kt})$ is the first order loss function of $\overline{CD}_{kt}$ based on Q_{kt}. Equations (1.5), (1.11), and (1.24) are equivalent to (1.14),(1.12), and (1.13), respectively, but take into account the initial inventory. These non-linear functions can be approximated using piece-wise linear functions based on the cumulative production quantities. Assuming the normal distribution for demand, Van Pelt and Fransoo (2018) show that the expected backorder (1.14) is a non-convex function and to insure that the pieces are selected sequentially, additional binary variable need to be added to the model. These additional variables are not needed in case of convexity of the functions. The parameters and decision variables are presented in Table 1.5.

$$E[\bar{I}_{kt}] = Q_{kt} - E[\overline{CD}_{kt}] + \Gamma^1_{\overline{CD}_{kt}}(Q_{kt}) \tag{1.12}$$

$$E[\overline{B}_{kt}] = \Gamma^1_{\overline{CD}_{kt}}(Q_{kt}) = E[max\{0, \overline{CD}_{kt} - Q_{kt}\}] \tag{1.13}$$

$$E[\overline{BO}_{kt}] = \Gamma^1_{\overline{CD}_{kt}}(Q_{kt}) - \Gamma^1_{\overline{CD}_{k,t-1}}(Q_{kt}) \tag{1.14}$$

Table 1.5: Parameters and decision variables of piece-wise linear model

Sets	
L	Set of linear segments
Parameters	
u_{0kt}	Lower limit of segment 1 for product k in period t
u_{lkt}	Upper limit of segment l for product k in period t
$\Delta_{I_{0kt}}$	Expected physical inventory associated with 0 cumulative production for product k in period t
$\Delta_{BO_{0kt}}$	Expected backorder associated with 0 cumulative production for product k in period t
$\Delta_{I_{lkt}}$	Slope of inventory function associated with segment l for product k in period t
$\Delta_{BO_{lkt}}$	Slope of backorder function associated with segment l for product k in period t
Decision variables	
w_{lkt}	Cumulated production quantity associated with segment l for product k in period t
λ_{lkt}	The binary variable which is equal to 1 if w_{lkt} takes a positive value.

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}(\Delta_{I_{0kt}} + \sum_{l\in L}\Delta_{I_{lkt}}w_{lkt})) \tag{1.15}$$

subject to constraints (1.3), (1.4), (1.7), (1.8), and:

$$x_{kt} = \sum_{l\in L} w_{lkt} - \sum_{l\in L} w_{lk,t-1} \qquad \forall t \in T, \forall k \in K \tag{1.16}$$

$$w_{l-1,kt} \geq (u_{l-1,kt} - u_{l-2,kt})\lambda_{lkt} \qquad \forall t \in T, \forall k \in K, \forall l \in L, l \geq 2 \tag{1.17}$$

$$w_{lkt} \leq (u_{lkt} - u_{l-1,kt})\lambda_{lkt} \qquad \forall t \in T, \forall k \in K, \forall l \in L \tag{1.18}$$

$$\sum_{l\in L} w_{lk,t-1} \leq \sum_{l\in L} w_{lkt} \qquad \forall t \in T, \forall k \in K \tag{1.19}$$

$$\frac{\sum_{t\in T}\sum_{k\in K}(\Delta_{BO_{0kt}} + \sum_{l\in L}\Delta_{BO_{lkt}}w_{lkt})}{\sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]} \leq 1-\beta \tag{1.20}$$

$$w_{lkt} \geq 0 \qquad \forall t \in T, \forall k \in K, \forall l \in L \tag{1.21}$$

$$\lambda_{lkt} \in \{0,1\} \qquad \forall t \in T, \forall k \in K, \forall l \in L \tag{1.22}$$

The objective function (1.15) is to minimize the setup cost plus the approximated expected value of the holding costs. Constraints (1.16) calculate the production amount based on the selected segments. Constraints (1.17) to constraints (1.19) guarantee that the segments are selected in

sequential order as proposed in Van Pelt and Fransoo (2018). Constraint (1.20) is the aggregate service level constraint in which the total average backorders divided by the total average demand is less than or equal to $1-\beta$. Constraints (1.21) and (1.22) show the domain of the different variables in the model.

1.5 Models with aggregate γ and δ service levels

In this section, we investigate an aggregate version of time and quantity oriented service levels (i.e., γ and δ). First, we explain the model for the γ service level which imposes that the total expected backlog divided by the total expected demand should be less than a predefined percentage. We then modify the model to consider an aggregate δ service level.

1.5.1 Problem definition and mathematical model

The mathematical model for this problem is presented as follows:

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt}+hc_{kt}E[\bar{I}_{kt}]) \tag{1.23}$$

subject to constraints (1.2), (1.3), (1.4), (1.7), (1.8), (1.9), (1.10), and:

$$E[\overline{B}_{kt}]=E[\max\{0,\sum_{j=1}^{t}\overline{D}_{kj}-\sum_{j=1}^{t}x_{kj}-I_{k0}\}] \quad \forall t\in T,\forall k\in K \tag{1.24}$$

$$\frac{\sum_{t\in T}\sum_{k\in K}E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]}\leq 1-\gamma \tag{1.25}$$

The objective function of the model (1.23) minimizes the setup and expected inventory holding costs. The expected value of backlog ($E[\overline{B}_{kt}]$) is calculated by constraint (1.24). Constraint (1.25) guarantees the γ aggregate service level.

It is also possible to define the aggregate γ service level in each planning period (γ_p). In this case, constraint (1.25) is replaced by constraints (1.26).

$$\frac{\sum_{k\in K}E[\overline{B}_{kt}]}{\sum_{k\in K}E[\overline{D}_{kt}]}\leq 1-\gamma_p \quad \forall t\in T \tag{1.26}$$

It is also possible to use the δ service level instead of the γ service level. In this case, constraint (1.25) is replaced by constraint (1.27). Both the γ and δ service levels work with the expected backlog for each product in each planning period.

$$\frac{\sum_{t\in T}\sum_{k\in K}E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K}(T-t+1)E[\overline{D}_{kt}]} \leq 1-\delta \tag{1.27}$$

In this mathematical formulation $E[\overline{I}_{kt}]$ and $E[\overline{B}_{kt}]$ are non-linear functions of the cumulative production which are shown in (1.11) and (1.24), respectively. In the next section, a mathematical model is presented to approximate this non-linear model.

1.5.2 Piece-wise linear approximation

The expected inventory and backlog in each planning period are non-linear functions of the cumulative production in each planning period. In this formulation, these non-linear functions are approximated based on the linearization of the first order loss function of the normal distribution which is convex in cumulative production (Rossi et al., 2014). As the non-linear functions for the expected inventory and expected backlog are convex, different segments on the piece-wise linear functions will be selected in sequential order and there is no need to add extra binary decision variables to ensure this, which is different from the model with the β service level. Table 1.6 indicates the new parameters and decision variables of this model.

Table 1.6: Parameters and decision variables of piece-wise linear model

Parameters	
$\Delta_{B_{lkt}}$	Slope of backlog function associated with segment l product k in period t
γ	Target aggregate γ service level
δ	Target aggregate δ service level

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}(\Delta_{I_{0kt}} + \sum_{l\in L}\Delta_{I_{lkt}}w_{lkt})) \tag{1.28}$$

subject to constraints (1.3), (1.4), (1.7), (1.8), (1.16), (1.21), and:

$$w_{lkt} \leq u_{lkt} - u_{l-1,kt} \quad \forall t \in T, \forall k \in K, \forall l \in L \tag{1.29}$$

$$\frac{\sum_{t\in T}\sum_{k\in K}(\Delta_{B_{0kt}}+\sum_{l\in L}\Delta_{B_{lkt}}w_{lkt})}{\sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]}\leq 1-\gamma \tag{1.30}$$

The objective function (1.28) is to minimize the setup cost plus the expected value of the holding costs. Constraints (1.29) define the maximum amount that the production quantity associated with segment l can take in period t. Constraint (1.30) is the aggregate service level constraint in which the total average backlog divided by total average demand is less than or equal to $1-\gamma$.

To change the model to consider the δ service level, constraint (1.30) should be replaced by constraint (1.31) in which the total average backlog divided by the total maximum expected backlog is less than or equal to $1-\delta$.

$$\frac{\sum_{t\in T}\sum_{k\in K}(\Delta_{B_{0it}}+\sum_{l\in L}\Delta_{B_{lkt}}w_{lkt})}{\sum_{t\in T}\sum_{k\in K}(T-t+1)E[\overline{D}_{kt}]}\leq 1-\delta \tag{1.31}$$

1.6 Models with α_c aggregate service level

In this section, we present the model for the α_c aggregate service level with a capacity constraint. First we define the mathematical model for this case. Next we present a quantile-based mathematical model to approximate the actual model.

1.6.1 Problem definition and mathematical model

This model is an extension of the model presented by Tempelmeier (2007) in which the α_c service levels are defined for each product separately. We investigate the combination of aggregate and individual service levels. The value of the minimum individual and aggregate α_c level are decided by the managers based on the cost of shortfalls and it is possible to ignore this cost in the model (Bookbinder and Tan, 1988). The minimum aggregate service level is a parameter in the model. Each item also has an individual minimum service level, which is less tight than the minimum aggregate service level. The actual individual service level is an output of the model since it depends on the decisions and can be better than the minimum imposed one. The α_c service level is considered in both the aggregate and individual constraints. The parameters and decision variables are presented in Table 1.7.

Table 1.7: Parameters of the model with α service level

Parameters	
w_k	The weight of product k such that $\sum_k w_k = 1$
α_c^{agg}	Minimum aggregate service level
α_c^{min}	Minimum required service level for each product

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}E[\bar{I}_{kt}]) \tag{1.32}$$

subject to constraints (1.2), (1.3), (1.4), (1.7), (1.8), (1.9), (1.10) and:

$$pr(I_{k0} + \sum_{j=1}^{t}(x_{kj} - \overline{D}_{kj}) \geq 0) \geq \alpha_c^{min} \quad \forall k \in K, \forall t \in T \tag{1.33}$$

$$\sum_{k\in K} w_k \min_{t\in T}(pr(I_{k0} + \sum_{j=1}^{t}(x_{kj} - \overline{D}_{kj}) \geq 0)) \geq \alpha_c^{agg} \tag{1.34}$$

The objective function (1.32) minimizes the setup and expected holding cost. The minimum service level for each product is imposed through the chance constraints (1.33). These constraints guarantee that the probability of a stock out is not larger than $(1-\alpha_c^{min})$. Constraint (1.34) imposes the aggregate service level. This constraint guarantees that the weighted sum of the resulting individual service levels is greater than or equal to the imposed aggregate service level.

1.6.2 Quantile-based approximation

In this section, we approximate the model with aggregate α_c service level using a quantile approach. The α service level is an event oriented service level which is expressed as a chance constraint to model the individual service level. In the literature, the common approach to model this problem under the static strategy with an individual service level is to use a predetermined service level as an input parameter (Bookbinder and Tan, 1988; Tempelmeier, 2013). However, in the model with the aggregate service level, the choice of the service level α_c becomes the decision variable for each product which will be used in the aggregate service level constraint, as they will be defined in the model and the minimum aggregate service level is the model parameter. In other words, in addition to setups and production amounts, the individual service levels are also decision variables. This will require a model that is different from the piece-wise linear approximations

which were used for the β, γ and δ service levels. In this model, we assume that the average net inventory is positive. This is a reasonable assumption for high service levels as the amount of negative inventory is negligible (Tempelmeier, 2007). In this formulation, the choices of possible service levels for each item are discretized using the set N, which leads to an approximation for the real problem. For each of the products one of the service levels in set N will be selected in the model such that the actual service level can take any value equal to or above the selected service level. The minimum service level (α_c^{min}) for each of the products is imposed by the minimum value in the set N. The new notation is presented in Table 1.8. The mathematical model is as follows:

Table 1.8: Parameters and decision variables of the quantile-based model

Set	
N	Set of service levels
Parameters	
$\overline{CD}_{kt}$	Cumulative demand for product k until period t
$F^{-1}_{\overline{CD}_{kt}}(\alpha_c)$	The minimum value of cd of cumulated t-period demand for which $P\{\overline{CD}_{kt} \leq cd\} \geq \alpha_c$ (Tempelmeier, 2013)
α_c^{kn}	Minimum probability of no stock out for item k based on service level n in each planning period
Decision variables	
I_{kt}	The amount of inventory for product k at the end of period t
s_{kn}	The binary variable which is equal to 1 if service level α_c^{kn} is selected for product k

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}I_{kt}) \tag{1.35}$$

subject to constraints (1.3), (1.4), (1.7), (1.8), and:

$$I_{kt} = I_{k0} + \sum_{j=1}^{t}(x_{kj} - E[\overline{D}_{kj}]) \qquad \forall t \in T, \forall k \in K \tag{1.36}$$

$$I_{k0} + \sum_{t=1}^{j} x_{kt} \geq F^{-1}_{\overline{CD}_{kj}}(\alpha_c^{kn})s_{kn} \quad \forall j \in T, \forall k \in K, \forall n \in N \tag{1.37}$$

$$\sum_{n\in N} s_{kn} = 1 \qquad \forall k \in K \tag{1.38}$$

$$\sum_{k\in K}\sum_{n\in N} w_k \alpha_c^{kn} s_{kn} \geq \alpha_c^{agg} \tag{1.39}$$

$$s_{kn} \in \{0,1\} \qquad \forall k \in K, \forall n \in N \tag{1.40}$$

$$I_{kt} \geq 0 \qquad \forall k \in K, \forall t \in T \tag{1.41}$$

The objective function (1.35) minimizes the sum of setup and holding costs. Constraints (1.36) are the inventory balance constraints in which $E[\overline{D}_{kt}]$ is the expected value of demand of product k in period t. It should be noted that I_{kt} is not a random decision variable in this model since we consider only the average demand. Constraints (1.37) are the individual service level constraints in which the discrete choice service level is defined for each item using a binary variable. These constraints ensure that the sum of the initial inventory and production quantities up to period t is at least equal to the cumulative demand required for the selected minimum service level. These constraints are equal to the chance constraints (1.42) and for cases that the demand follows a normal distribution the value of $F^{-1}_{\overline{CD}_{kj}}(\alpha_c^{kn})$ is easy to calculate. These chance constraints ensure that for each period the probability of a stock out is less than or equal to $(1-\alpha_c^{kn})$ for the chosen level n of the service level. For the case where there is only one choice of service level, this constraint will be the same as constraint (1.43) proposed for the single item problem (Tempelmeier, 2013). Constraints (1.38) guarantee that exactly one service level for each item is selected. Constraint (1.39) imposes the aggregate service level in which the weighted average of the selected individual service levels is larger than or equal to the imposed aggregate service level.

$$pr(I_{k0} + \sum_{j=1}^{t} x_{kj} - \sum_{j=1}^{t} \overline{D}_{kj} \geq 0) \geq \alpha_c^{kn} s_{kn} \quad \forall t \in T, \forall k \in K, \forall n \in N \tag{1.42}$$

$$I_{k0} + \sum_{t=1}^{j} x_{kt} \geq F^{-1}_{\overline{CD}_{kj}}(\alpha_c^k) \qquad \forall j \in T, \forall k \in K \tag{1.43}$$

1.7 Receding horizon model

The models which are presented in this paper, assume the static strategy in which the setup and production quantity decisions remain unchanged during the planning horizon. Although this characteristic is important in some applications to avoid nervousness, it will reduce the responsiveness of the plan and potentially incur additional costs to the system. In order to deal with the case when the setup and quantity decisions can be continuously adjusted once the demand information is revealed, we can apply the receding approach using the static models proposed in this paper. In the first step, the complete model (with static strategy assumption) is solved for the horizon 1 to T, and the setup and production decisions of the first period are fixed. Then, based on the realized demand in the first period, the amounts of backlog and inventory are calculated at the end of the first period which will be the initial inventory and backlog at the beginning of the second period. In the following steps the model will be run for the rest of the periods in the planning horizon with updated initial inventory and backlog. This procedure continues until the end of the planning horizon.

As the service level constraint may be violated with the demand realization, the violation of this constraint will be incorporated in the objective function with a high penalty (P). Table 1.9 illustrates the new parameters for this model in the i^{th} iteration which are calculated based on the fixed production in the periods before period i. The mathematical model for the i^{th} iteration is as follows:

Table 1.9: Parameters and decision variables of the i^{th} iteration of the receding horizon model

Parameters	
$\hat{BO}_{kt}$	The backorder for product k in period t, $t < i$, which is calculated by (1.44)
$\hat{B}_{kt}$	The backlog for product k in period t, $t < i$, which is calculated by (1.45)
d_{kt}	The realized demand for product k in period t, $t < i$
$\hat{I}_{kt}$	The inventory for product k in period t, $t < i$, which is calculated by (1.46)
P	The penalty cost for service level violation
$\hat{x}_{kt}$	The fixed amount of production for product k in period t, $t < i$
Decision variable	
ε	The service level violation

$$\hat{BO}_{kt} = \max\{0, \sum_{j=1}^{t}(d_{kj} - \hat{x}_{kj}) - I_{k0}\} - \max\{0, \sum_{j=1}^{t-1} d_{kj} - \sum_{j=1}^{t} \hat{x}_{kj} - I_{k0}\} \quad \forall k \in K, \forall t < i \tag{1.44}$$

$$\hat{B}_{kt} = \max\{0, \sum_{j=1}^{t} d_{kj} - \sum_{j=1}^{t} \hat{x}_{kj} - I_{k0}\} \quad \forall k \in K, \forall t < i \tag{1.45}$$

$$\hat{I}_{kt} = \max\{0, I_{k0} + \sum_{j=1}^{t} \hat{x}_{kj} - \sum_{j=1}^{t} d_{kj}\} \quad \forall k \in K, \forall t < i \tag{1.46}$$

$$\text{Min} \sum_{t=i}^{T} \sum_{k \in K} (sc_{kt} y_{kt} + hc_{kt} E[\bar{I}_{kt}]) + P\varepsilon \tag{1.47}$$

subject to:

$$\hat{I}_{k,i-1} + x_{ki} + \overline{B}_{ki} = \bar{I}_{ki} + \overline{D}_{kt} + \hat{B}_{k,i-1} \quad \forall k \in K \tag{1.48}$$

$$\bar{I}_{k,t-1} + x_{kt} + \overline{B}_{kt} = \bar{I}_{kt} + \overline{D}_{kt} + \overline{B}_{k,t-1} \quad \forall t \in T, t > i, \forall k \in K \tag{1.49}$$

$$x_{kt} \leq M y_{kt} \quad \forall t \in T, t \geq i, \forall k \in K \tag{1.50}$$

$$\sum_{k \in K} (st_{kt} y_{kt} + pt_{kt} x_{kt}) \leq Cap_t \quad \forall t \in T, t \geq i \tag{1.51}$$

$$y_{kt} \in \{0,1\} \quad \forall t \in T, t \geq i, \forall k \in K \tag{1.52}$$

$$x_{kt} \geq 0 \quad \forall t \in T, t \geq i, \forall k \in K \tag{1.53}$$

$$\bar{I}_{kt} \geq 0 \quad \forall t \in T, t \geq i, \forall k \in K \tag{1.54}$$

$$\overline{B}_{kt} \geq 0 \quad \forall t \in T, t \geq i, \forall k \in K \tag{1.55}$$

The objective function (1.47) minimizes the total setup cost, expected inventory holding cost, and the violation of the service level for period i until the end of the horizon. Constraints (1.48) are the inventory balance constraints for period i in which we have the initial inventory and backlog from period $i-1$. Constraints (1.49) are the inventory balance constraints for the periods after period i until the end of planning horizon. Constraints (1.50) and (1.51) are the setup production constraints and the capacity constraints, respectively. Constraints (1.52) to (1.55) define different variables of the model.

In addition to these constraints, based on the type of service level, the following constraints should

be added to the model. For the β service level, constraints (1.56) and (1.57), for the γ service level constraint (1.58), for the δ service level constraint (1.59), and for the α_c service level constraint (1.60) will be added to the model.

$$E[\overline{BO}_{kt}] = E[\max\{0, \sum_{j=i}^{t}(\overline{D}_{kj} - x_{kj}) - \hat{I}_{k,i-1} + \hat{B}_{k,i-1}\}] \\ -E[\max\{0, \sum_{j=i}^{t-1}\overline{D}_{kj} - \sum_{j=i}^{t} x_{kj} - \hat{I}_{k,i-1} + \hat{B}_{k,i-1}\}] \qquad \forall t \in T, t \geq i, \forall k \in K \quad (1.56)$$

$$\frac{\sum_{t=1}^{i-1}\sum_{k\in K}\hat{BO}_{kt} + \sum_{t=i}^{T}\sum_{k\in K}E[\overline{BO}_{kt}]}{\sum_{t=1}^{i-1}\sum_{k\in K}d_{kt} + \sum_{t=i}^{T}\sum_{k\in K}E[\overline{D}_{kt}]} \leq 1 - \beta + \varepsilon \quad (1.57)$$

$$\frac{\sum_{t=1}^{i-1}\sum_{k\in K}\hat{B}_{kt} + \sum_{t=i}^{T}\sum_{k\in K}E[\overline{B}_{kt}]}{\sum_{t=1}^{i-1}\sum_{k\in K}d_{kt} + \sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]} \leq 1 - \gamma + \varepsilon \quad (1.58)$$

$$\frac{\sum_{t=1}^{i-1}\sum_{k\in K}\hat{B}_{kt} + \sum_{t=i}^{T}\sum_{k\in K}E[\overline{B}_{kt}]}{\sum_{t=1}^{i-1}\sum_{k\in K}(T-t+1)d_{kt} + \sum_{t\in T}\sum_{k\in K}(T-t+1)E[\overline{D}_{kt}]} \leq 1 - \delta + \varepsilon \quad (1.59)$$

$$\sum_{k\in K} w_k \min_{t\in T, t\geq i}(pr(I_{k0} + \sum_{j=1}^{i-1}\hat{x}_{kj} + \sum_{j=i}^{t}x_{kj} - \sum_{j=1}^{i-1}d_{kj} - \sum_{j=i}^{t}\overline{D}_{kj} \geq 0)) \geq \alpha_c^{agg} + \varepsilon \quad (1.60)$$

It is important to note that the efficiency of the rolling/receding horizon scheduling depends on the defined parameters such as the frozen horizon (which is equal to one in our case), and the length of the planning interval (Meistering and Stadtler, 2017). This opens a new direction for research which is not within the scope of this paper.

1.8 Computational experiments

To investigate the effect of an aggregate service level and gain computational insights into the benefits of the solutions based on different types of service levels, we conduct different computational experiments. First, the data generation procedure is explained. Next, the parameters of the different models such as the number of service level options in the quantile-based approximation and the number of segments in the piece-wise linear models are analysed. In the third section, we evaluate the results of different service levels based on an initial data set. The fourth section is dedicated to extensive sensitivity analysis of the value of the aggregated service levels based on different parameters and service levels. In the last sections, the effect of individual service levels on the value of aggregate service level is presented.

1.8.1 Instance generation

In this section, we explain the data which are used to test the models. We have two different sets, one set for the initial and more general tests (set A) and one for more specific tests and the sensitivity analysis (set B). For both sets, we follow the same procedure to generate data as in Helber et al. (2013) with some modification.

The set A is used to investigate the difference between the aggregate and individual service level for all types of service levels. As the original data set which was proposed by Helber et al. (2013) is generated based on the δ service level, for other service levels some instances may be infeasible due to the capacity constraint. The first modification is to reduce the utilization factor to increase the capacity. The other modification is to assign different holding costs to different products. If all the products have the same holding cost, the aggregate service level does not show a big advantage over the separate service levels (as indicated later in the sensitivity analysis). Table 1.10 shows the average long-term demand for each of the products ($E[\overline{D}_k]$) and Table 1.11 shows the parameters used for the generation of these test instances. VC^{ip} is the inter-period coefficient of variation which is used to generate dynamic time series based on $E[\overline{D}_k]$ and defines the average demand for each product in each planning period ($E[\overline{D}_{kt}]$). More specifically, the average demand for product k in period t ($E[\overline{D}_{kt}]$) is taken from a normal distribution with an average equal to the average long-term demand ($E[\overline{D}_k]$) and a coefficient of variation of VC^{ip}. A lower VC^{ip} results in more moderate variability between demands in different periods and a higher one results in larger differences. The VC^d refers to the coefficient of variation of the demand in a specific period. The standard deviation of the demand is equal to the average demand multiplied by the VC^d. Note that in practice, if the forecasted demand is used, one can calculate VC^d of forecast errors and use it here. TBO is the time between orders, which shows the number of periods between two consecutive orders. TBO is used to define the value of the setup cost based on the average demand and holding cost. A detailed explanation of the data generation procedure can be found in (Helber et al., 2013). In set A there are 432 instances with all the combinations of parameters presented in Table 1.11. The sizes of these samples which are defined based on the number of products $|K|$ and number of periods $|T|$, are $\{|K| = 5, |T| = 5\}$, $\{|K| = 10, |T| = 5\}$, and $\{|K| = 5, |T| = 10\}$.

The second data set, set B, is used for the sensitivity analyses. To this end, 10 base instances

Table 1.10: Series of expected demand $E[\overline{D}_k]$ (Helber et al., 2013)

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$E[D_k]$	67	135	105	80	79	72	85	136	56	100	150	108	150	126	63	114	66	75	117	93

Table 1.11: Parameters of the test instances

Parameters	
Inter-period coefficient of variation of expected demand	VC^{ip} = 0.2 , 0.3
Coefficient of variation of demands	VC^{d} = 0.1 , 0.3
Time between orders	TBO = 1, 2, 4
Utilization of resource due to processing	Util = 0.4 , 0.5
Setup time as fraction of period processing time	ST = 0.0 , 0.25
Service-level target	Service Level = 0.8 , 0.9 , 0.95
Holding cost	hc = 1, 2, 3 , ..., $\|K\|$
Product weight	$w_k = 1/\|K\|$

with the size of $\{|K| = 10, |T| = 10\}$ are generated with the same parameters and different demand values which are randomly generated based on the normal distribution. The parameters used to generate the instances are listed in Table 1.12. In the sensitivity analysis section, we will explain how different scenarios for each of the parameters are generated.

Table 1.12: Parameters of the base case instances for the sensitivity analysis

$\|K\|$ = *10*	$\|T\|$ = *10*	VC^{ip} = *0.3*	VC^{d} = *0.3*
TBO = 3	*Util = 0.65*	*ST = 0.0*	*Service Level = 0.95*
$hc \in \{1,2,3,...,\|K\|\}$			

To evaluate the solutions formed by the approximate formulations, we use simulation. The results of the models including the setup decisions and production levels for each product and in each period are the input of this process. 10,000 demand scenarios are generated based on a normal distribution with the same average and variance as the input to the model. The objective function and service levels are then evaluated using the simulation. Using exactly the same 10,000 scenario set for all the service levels, will help us to have a fair comparison between different service levels.

1.8.2 Determining the number of linear segments and service levels

To define the number of linear segments for the piece-wise linear models, we performed tests using the model with separate γ service levels and solved it with different numbers of segments for each of the 432 small instances. Assuming that the models have the same characteristics, among

different models, the model with separate γ service level is selected. Each instance is solved with 5, 10, 15, 20, 30, 40, and 50 linear segments and evaluated using the same set of scenarios. Figure 1.1 shows the average solution time and average accuracy of the solution over all solved instances. The accuracy measures are the cost accuracy (1.61) and service level accuracy (1.62). The cost accuracy is the percentage of absolute difference between the model objective function and the evaluated objective function. The service level accuracy is the absolute difference between the evaluated service level and the target service level which is calculated in percentage point (%p). Each of these accuracy measures are calculated for each instance (in), and their average over all instances are used for the analysis. Considering the trade-off between solution time and accuracy measures, the model with 20 segments is selected. The result of this study is generally in line with the study of Helber et al. (2013) who used 18 segments for their piece-wise linear model for the δ service level and Tempelmeier and Hilger (2015) for the β service level. Without loosing generality the pieces are in equidistant intervals and we will use the same number of segments for all the piece-wise linear models. These intervals will be the same for all the service levels with piece-wise linear approximation (i.e., β, γ, δ).

$$Cost\ Accuracy_{in}(\%) = \frac{|Evaluated\ total\ cost_{in} - Model\ objective\ function_{in}|}{Model\ objective\ function_{in}} \tag{1.61}$$

$$Service\ Level\ Accuracy_{in}(\%p) = |Evaluated\ service\ level_{in} - Target\ service\ level_{in}| \tag{1.62}$$

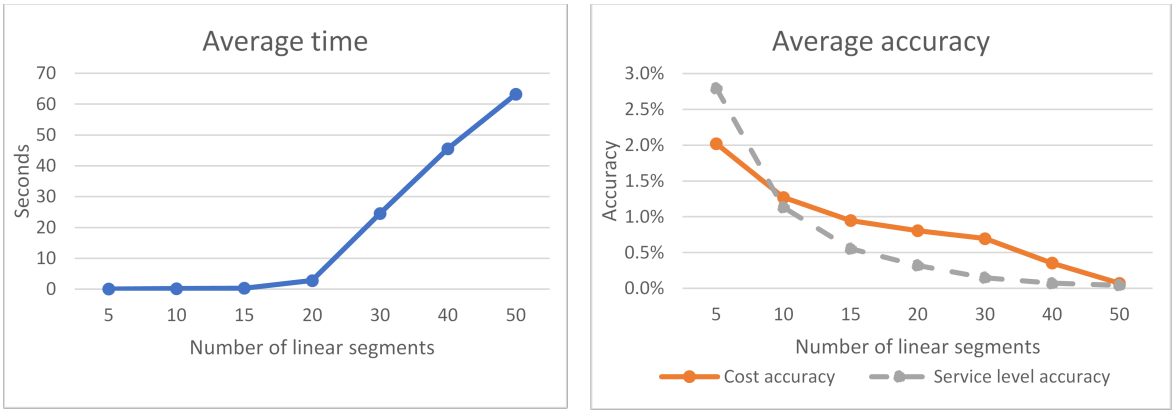


Figure 1.1: Execution time and accuracy of the piece-wise linear model for the γ service level based on the number of linear segments

To test the impact of the number of service level options for each of the products in the quantile approach, we solve the aggregate model for the α_c service level with different numbers of service level options. Each instance is solved with 2, 5, 11, 15, 21, and 30 service level options and evaluated using the same set of scenarios. The service levels are equally distributed between the minimum service level, which is 80%, and 99.99%. For example, the 21 service levels are 80%, 81%, ...,99%, 99.99%. Note that the option 99.99% is used instead of 100% since the 100% service level results in an infinite amount of inventory. Figure 1.2 shows the average solution time and accuracy of the solutions. Considering the trade-off between time and accuracy measures, and the fact that the accuracy measures do not decrease notably when increasing the number of service level options to more than 11, 11 service level choices are used in the subsequent experiments.

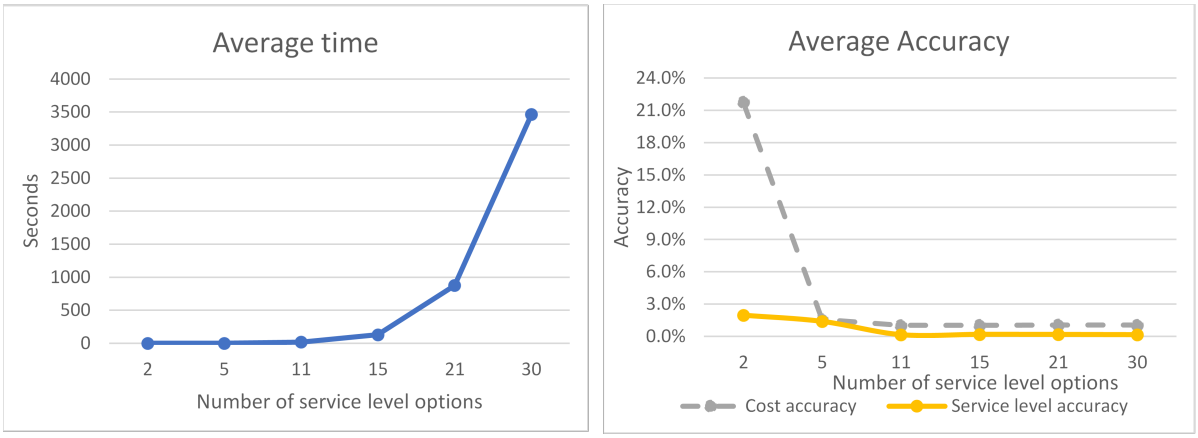


Figure 1.2: Execution time and accuracy of the quantile model for the α_c service level based on the number of service level options

1.8.3 Performance evaluation based on different service levels

This section shows the results of the experiments for different types of service levels using the 432 instances in set A. The aim of these experiments is to provide insights with respect to the models with different service levels and to give a general overview of the difference between the aggregate and separate service levels. For the experiments, we used the CPLEX 12.8.1.0 and Python libraries. We performed these experiments on a 2.1 GHz Intel Broadwell processor with only one thread on the Compute Canada Graham computing grid. Table 1.13 summarizes these results. To analyze

the results, two versions of the models, i.e., with aggregate and separate service levels, are solved using the approximated models for each type of service levels. In our comparisons, in addition to average cost accuracy and average service level accuracy, new measures are considered. Again, each of these measures are calculated for each instance (in), and their averages over all instances are used for the analysis. We analyze the average cost from the evaluation (Average Cost), the average deviation of the actual service level from the defined target (Service Level Deviation (1.63)), the average percentage difference between the evaluated cost of the model objective and the model objective function (Cost Deviation (1.64)), the average solution time in seconds (Average Time), and the average difference between models with aggregate and separate service levels ($\Delta Cost$). $\Delta Cost$ is shown in the last column of Table 1.13 and shows the advantage of the aggregate service level over the separate one based on the total cost increase percentage (1.65).

$$Service\ Level\ Deviation_{in}(\%p) = Evaluated\ service\ level_{in} - Target\ service\ level_{in} \tag{1.63}$$

$$Cost\ Deviation_{in}(\%) = \frac{Evaluated\ total\ cost_{in} - Model\ objective\ function_{in}}{Model\ objective\ function_{in}} \tag{1.64}$$

$$\Delta Cost_{in}(\%) = \frac{Cost\ of\ separate\ service\ level_{in} - Cost\ of\ aggregate\ service\ level_{in}}{Cost\ of\ aggregate\ service\ level_{in}} \tag{1.65}$$

The first service level is the β service level. To analyze the results, the two versions of the model,

Table 1.13: Results of the approximation models for different types of service level

Service Level	Version	Average Cost	Service Level Deviation (%p)	Service level Accuracy (%p)	Cost Deviation (%)	Cost Accuracy (%)	Average Time(s)	$\Delta Cost(\%)$
β	Aggregate	27441.9	0.1	0.1	-0.1	0.1	394.1	6.2
	Separate	29452.7	0.2	0.2	-0.2	0.2	247.7	
β_p	Aggregate	29836.5	0.0	0.3	-0.3	0.3	309.4	15.1
	Separate	34607.3	0.1	0.2	-1.0	1.0	22.3	
γ	Aggregate	28090.6	0.1	0.1	-0.1	0.1	0.5	5.2
	Separate	29877.7	0.3	0.3	-0.2	0.2	3.1	
γ_p	Aggregate	30753.0	0.0	0.1	-0.2	0.3	14.2	12.5
	Separate	34732.6	0.1	0.2	-1.0	1.0	30.1	
δ	Aggregate	17070.3	0.0	0.0	-0.1	0.1	0.3	9.7
	Separate	18622.7	0.1	0.1	-0.5	0.5	0.6	
δ_p	Aggregate	22593.2	0.0	0.0	-0.4	0.4	8.5	26.6
	Separate	27818.0	0.0	0.1	-1.1	1.1	16.7	
α_c	Aggregate	38453.8	-0.2	0.2	1.5	1.5	2478.8	1.5
	Separate	38970.9	-0.2	0.2	1.1	1.1	0.3	

aggregate and separate, are solved using the piece-wise linear approximation. As can be seen, the

average cost which is the result of the evaluation has the lower value when there is an aggregate service level constraint compared to the separated one. It is worthwhile to mention that the piece-wise linear model generally overestimates the inventory and backlog. The average β deviation, cost deviation and both accuracy measures are close to 0, which shows that the piece-wise linear model provides a very good estimation. The positive percentage of service level deviation shows that the imposed service levels are satisfied. The last column shows the advantage of the aggregate β service level over the separate one which is about 6%. In terms of execution time the separate model is faster on average. When the β service level is defined per period (β_p) the $\Delta Cost$ is increased to 15.2% which is more than twice the value for the global case (β). Furthermore, the execution time for the aggregate model is slightly increased but remains in the same order of magnitude, whereas the time for the separate service level has been reduced by an order of magnitude. The execution time is reduced from aggregate to separate and from global service level to per period service level.

The next service level is the γ service level. The deviations and accuracy measures are close to 0 for both aggregate and separate models. This means that the piece-wise linear model provides a good approximation. The $\Delta Cost$ is about 5% which shows the average cost reduction for the aggregate case compared to the separate case. In terms of execution time the model is very fast compared to the β service level. One of the main reasons is the presence of extra binary variables in the models for the β service level. It is also possible to investigate the difference between the separate and aggregate service level per period (γ_p). The advantage of the aggregate service level over the separate one is about 12%. These differences are more distinctive compared to the global case (γ) where the service level is defined over the whole planning horizon. The execution time is higher compared to the global version.

The next service levels are the δ and δ_p service levels. As can be seen, the difference between the model and evaluated cost and service level is very small and in most cases close to 0. For both service levels, the cost of the aggregate models are less than the cost of the separate models and the average differences are 9.7% and 26.6% for δ and δ_p, respectively. The execution times of the models with δ and δ_p are slightly lower than for the γ and γ_p service levels, respectively, but follow the same pattern and they are lower in the global cases compared to per period ones.

The last service level is the α_c service level. Based on the preliminary test, 11 service levels for

the quantile-based model are selected. The minimum service level for each product is set to 80%. The models are solved with the aggregate and separate α_c service levels using the quantile-based model. Same as previous service levels, we have three levels for the α_c target: 80%, 90%, and 95%. For example, for the case with α_c equal to 90%, we solve the separate model imposing individual α_c levels of 90% for each product, and we solve the aggregate model imposing individual α_c levels of 80% and an aggregate α_c level of 90%. The α_c deviation and accuracy are close to 0% and therefore the model has a very good performance in terms of service level. Based on the cost accuracy, although the quantile-based approximation has a good performance, the performance of the piece-wise linear approximation for other service levels outperforms quantile-based approximation for the defined number of segments and service level options. In terms of execution time the separate model is much faster than the aggregate model. The difference between the cost of the models with aggregate and separate α_c service level is about 1.5%. Note that the case with the aggregate service level equal to 80% results in 0% $\Delta Cost$ since we assume a minimum individual service level of 80%. Excluding this case the $\Delta Cost$ is equal to 2.53%.

In general, the $\Delta Cost$ for the global service levels (β, γ, δ) are less than the $\Delta Cost$ for the service levels imposed in each period $(\beta_p, \gamma_p, \delta_p)$. Based on the total cost, we can conclude that with the same value for the service levels, the α_c service level is the most strict and the δ service level is the least strict service level among the four. This can conclude that the advantage of an aggregate service level is more noticeable for the less strict service levels because of the higher flexibility in these service levels. In more strict service levels there is less possibility for the aggregate model to maneuver.

1.8.4 Sensitivity analysis

To have a better understanding of the effect of different parameters on the cost difference between aggregate and separate service levels, sensitivity analyses are conducted. In these experiments, the effect of holding cost, demand variation, capacity, *TBO*, service level, number of products, and periods are investigated. Different values which are used for the sensitivity analysis are provided in Table 2.8. A low level for *Util* shows the loose capacity and a high level for *Util* pertains to the tight one. Table 1.15 shows different cases for the holding cost and their variance.

Figure 1.3 presents the results for the sensitivity analysis of the γ service level. The capacity

Table 1.14: Parameter values for the sensitivity analysis

Parameter	Values
$\|K\|$	*5, 10, 15, 20*
$\|P\|$	*5, 10, 15, 20*
Util	*0.45, 0.55, 0.65, 0.75, 0.85*
TBO	*1, 2, 3, 4, 5*
VC^d	*10%, 20%, 30%, 40%, 50%*
Service Level	*91%, 93%, 95%, 97%, 99%*

Table 1.15: Different options of the holding cost for the sensitivity analysis

Case	Holding cost for 10 products	Standard Deviation
1	*[1, 1, 1, 1, 1, 10, 10, 10, 10, 10]*	4.74
2	*[1, 1, 1, 5.5, 5.5, 5.5, 5.5, 10, 10, 10]*	3.67
3	*[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]*	3.03
4	*[3, 3, 3, 5.5, 5.5, 5.5, 5.5, 8, 8, 8]*	2.04
5	*[4, 4, 4, 5.5, 5.5, 5.5, 5.5, 7, 7, 7]*	1.22
6	*[5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5]*	0.00

constraints affect more the models with separate service level compared to aggregate ones as in the separate case it is not possible to compensate the production of some products with others. Because of the aggregate model flexibility, the capacity does not affect it at lower utilization. At higher utilization when the capacity is tighter, the capacity will affect the aggregate model as well and this causes a very small reduction in $\Delta Cost$.

For the *TBO*, there is a reduction in $\Delta Cost$ when *TBO* increases from 1 to 3 and there is an increase in $\Delta Cost$ when *TBO* increase from 3 to 5. When *TBO* is 1 the advantage of the aggregate model is reflected in the total inventory cost. The aggregate model satisfies the service level constraint by storing less from the products which have the higher holding costs. This flexibility does not exist in the model with the separate service level. When *TBO* is equal to 5 the advantage of the aggregate model is more reflected in the total setup cost. There is the possibility of not producing a product in the aggregate model as it is possible to compensate it with other products. This flexibility does not exist in the separate model.

The $\Delta Cost$ generally decreases when the variance increases. The higher variance will increase the total expected cost in general for both the aggregate and separate model. Due to the flexibility of the aggregate model, it is less affected at the lower variance, and this causes the ΔCost to decrease as the variance increases.

The $\Delta Cost$ generally decreases when the target aggregate service level increases. This is logical because, when there is a lower service level, the aggregate model has a higher flexibility, which is not the case for higher service levels.

Based on the plots, the $\Delta Cost$ for the γ service level in its global version exibits no obvious trend for the number of products and periods. The lowest value for the $\Delta Cost$ is when the number of products is equal to 5 and highest value is when it is equal to 10.

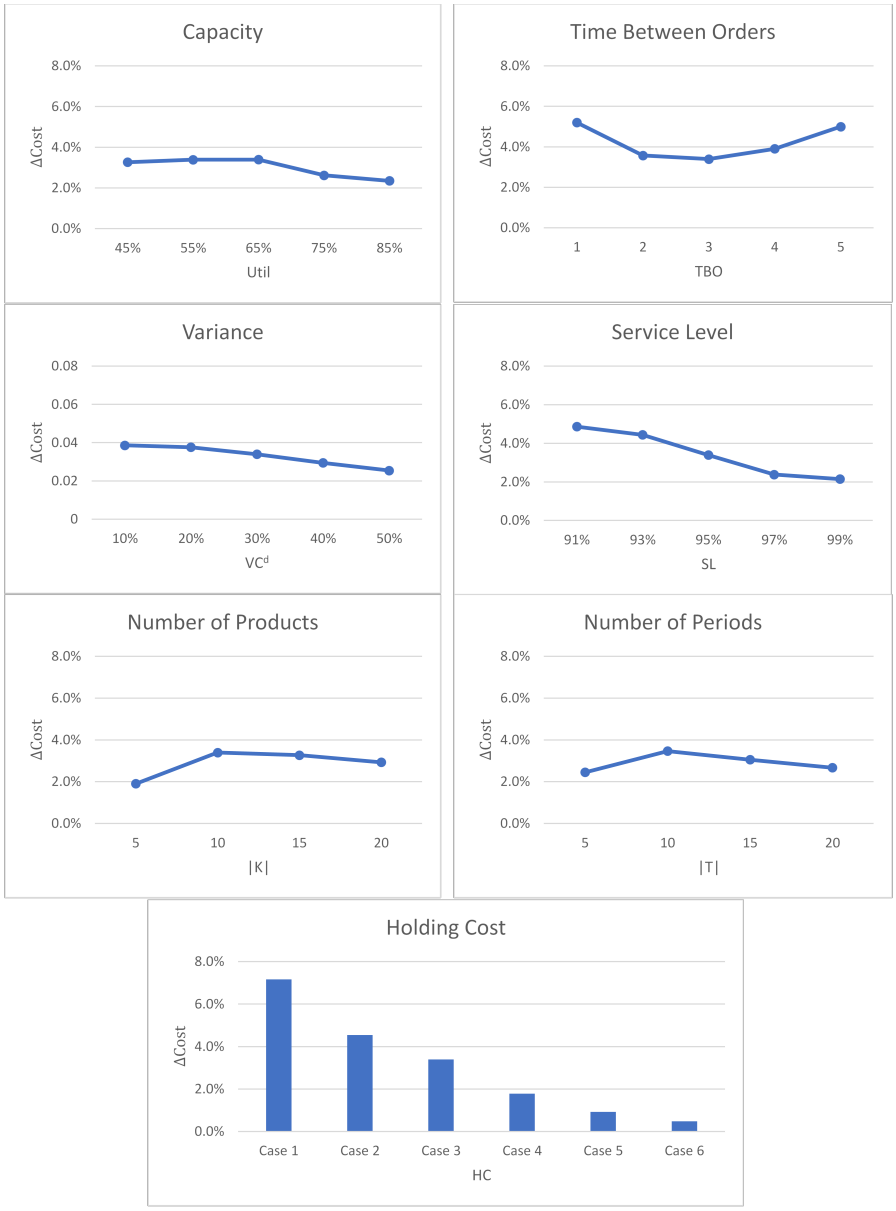


Figure 1.3: Sensitivity analysis plots for γ service level

The last plot in Figure 1.3 shows the sensitivity analysis based on the holding cost. The $\Delta Cost$ increases as the variance in the holding costs increases as the variation of holding costs increase from case 6 to case 1. The advantage of the aggregate service level is more noticeable when the products have higher differences in their holding costs. This shows that, when there is a limited capacity and high variation in holding costs an aggregate service level will allow us to obtain significantly lower costs. Figure 1.4 illustrates different plots for the sensitivity analysis of the γ_p service level. The advantage of the aggregate service level in this case is much higher than for the γ service level as indicated by the general higher level of $\Delta Cost$. Unlike for the γ service level, which was defined globally, at higher capacity utilization $\Delta Cost$ will increase for the γ_p service level. In addition to the cost saving with the aggregate service level, there is also a higher probability of infeasibility in the model with a separate service level. It is worthwhile to mention that there is also a higher probability of infeasibility in γ_p compared to the γ service level. For example, when the utilization factor is equal to 0.85, the models with γ_p service levels are infeasible which is not the case for the γ service level.

Similar to the case of the γ service level, in the case of the γ_p service level, the $\Delta Cost$ decreases when the target aggregate service level increases. This is because, when there is a lower service level, the aggregate model has a higher flexibility, which is not the case for higher service levels. When the service level is equal to 99% all the 10 models with separate service levels were infeasible, while 6 of them were infeasible in the aggregate case. Unlike the global version, γ, in which the $\Delta Cost$ is not sensitive to the number of periods, for the γ_p service level the $\Delta Cost$ increases as the number of periods increases.

The last plot in Figure 1.4 shows the sensitivity analysis based on the holding cost. This plot follows a similar trend as in the case of the γ service level but the values of $\Delta Cost$ are much higher in all cases. The $\Delta Cost$ increases as the variance in the holding costs increases. The advantage of the aggregate service level is more distinctive when the products have higher differences in terms of holding cost.

Appendix C shows the sensitivity analysis diagrams for β, β_p, δ, and δ_p service levels. In these diagrams the global service levels, β and δ have similar trends as observed for γ. The β_p and δ_p service levels show similar patterns as the γ_p service level.

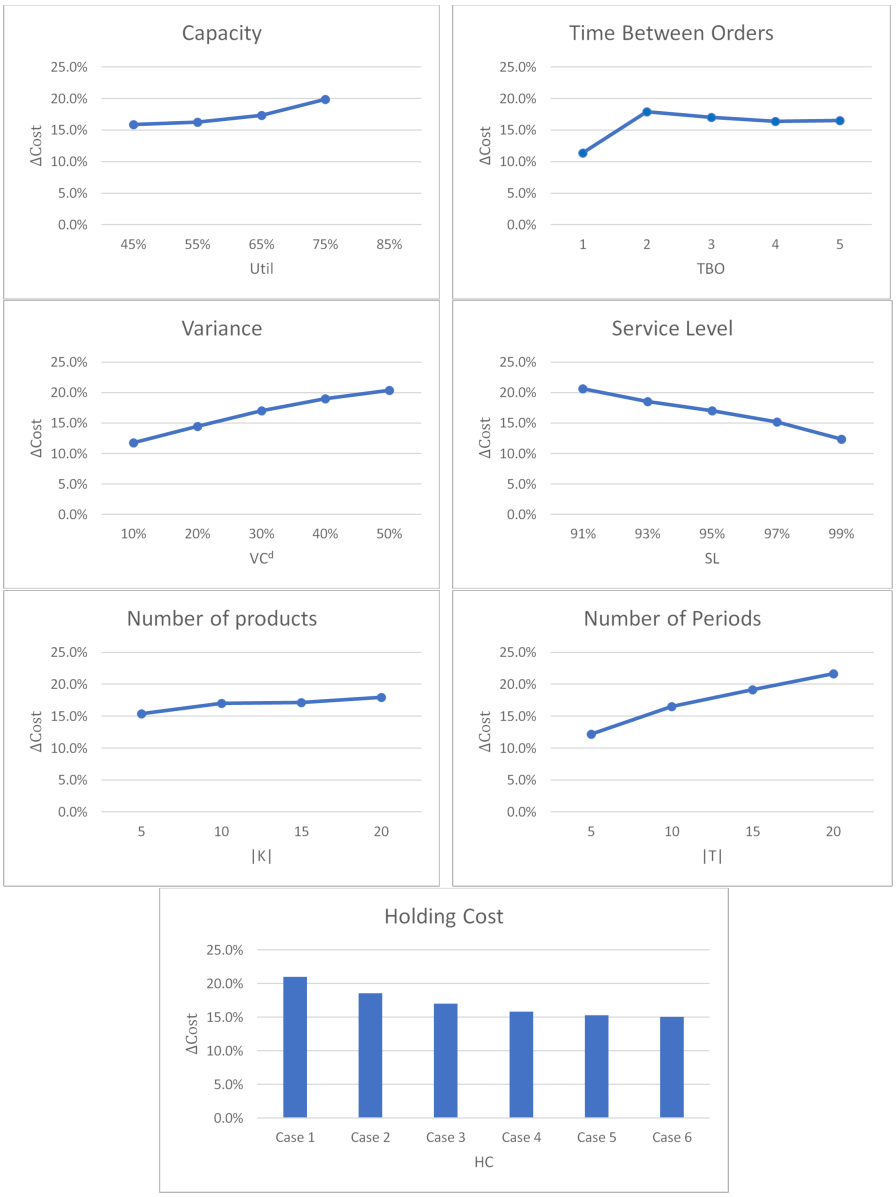


Figure 1.4: Sensitivity analysis plots for γ_p service level

1.8.5 The effect of minimum individual service level

In the previous section, there was no minimum individual service level imposed in the aggregate models except for the α_c service level since this model requires the explicit modeling of a discrete number of service level options. In this section, we investigate the case when both the individual and aggregate service levels are imposed collectively for selected service levels. To avoid infeasibilities the *Util* is changed from 0.65 to 0.5. Figure 1.5 shows the plots for the γ service level.

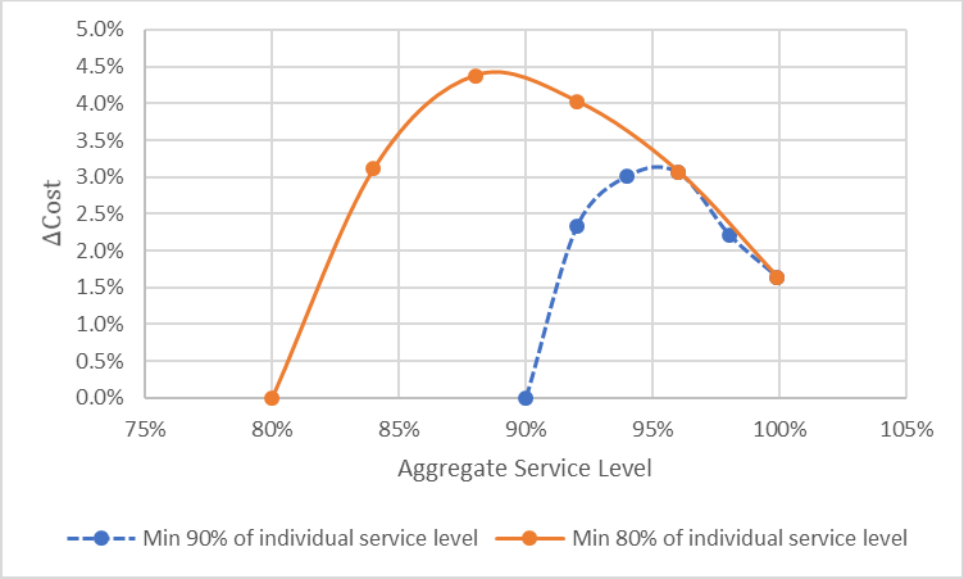


Figure 1.5: Effect of individual service levels (γ)

There are two series in this plot. One of them shows the cost difference between the aggregate and separate service levels when there is an individual γ service level of 80% imposed together with an aggregate γ constraint. The other series which is shown by the dashed line is the cost difference when the individual service level of 90% is imposed together with an aggregate constraint. Both of these series follow a similar pattern. When the separate service level is equal to the minimum individual service levels in the aggregate model, $\Delta Cost$ is equal to 0. When the difference between the minimum individual and aggregate service level increases, the $\Delta Cost$ will also increase to a certain point. After that there is a decrease in the $\Delta Cost$. This shows that at high service levels the difference between aggregate and separate service levels will decrease. This is logical as in the lower service levels there is more flexibility for the aggregate model, while at higher service levels the amount of allowable backlog for both separate and aggregate service level is very low. Figure 1.6 shows the similar diagrams for $\beta, \beta_p, \gamma_p, \delta$, and δ_p. Note that for convenience, we include the diagram for the γ service level as well. The trends in these diagrams for the global service levels are similar to the γ service level which is explained before. The trends for the per period service levels ($\beta_p, \gamma_p, \delta_p$) are also similar to each other. In these latter diagrams there is no point for the 99.9% as in all of these cases the models with separate service levels were infeasible and it was not possible to calculate the $\Delta Cost$. Despite the similar trends in the diagrams, there are

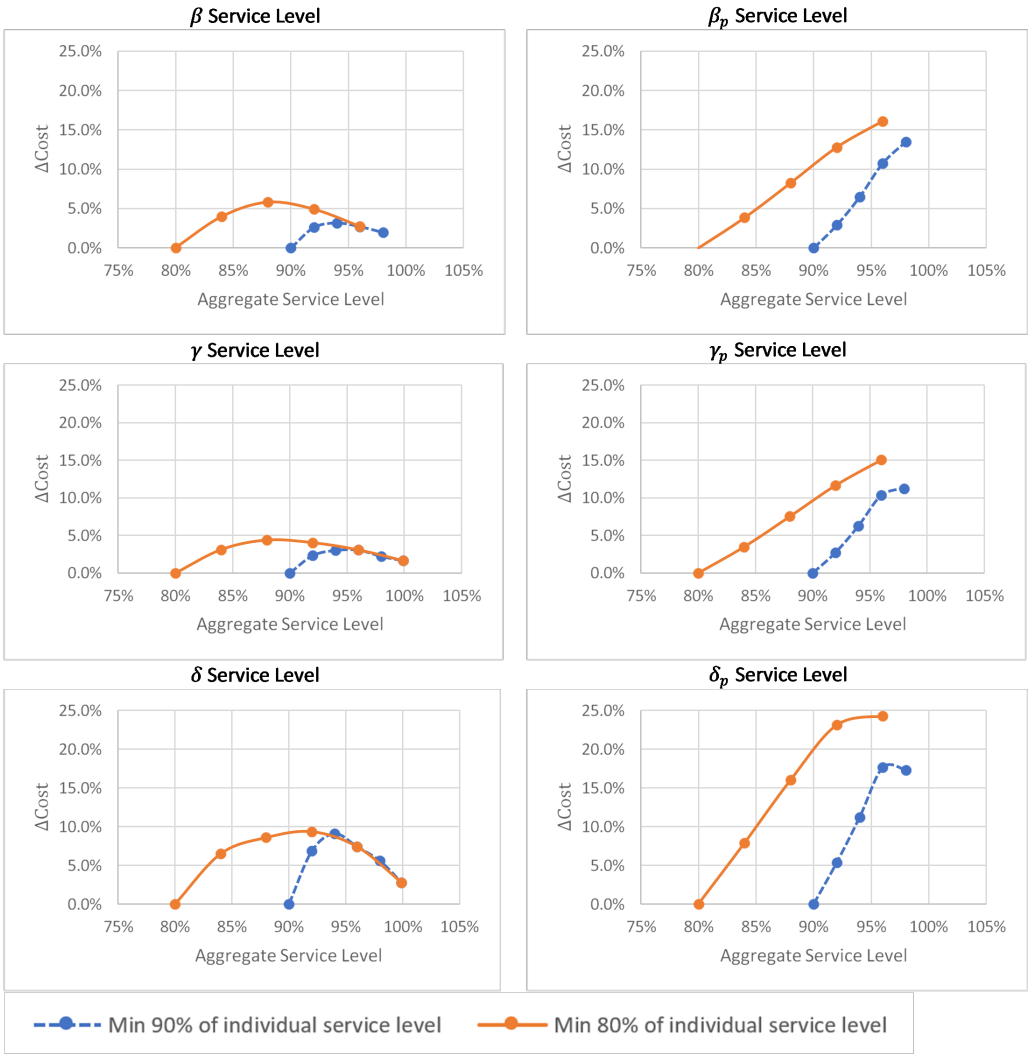


Figure 1.6: Effect of individual service levels

differences in the value of $\Delta Cost$. Based on these diagram we can conclude that the value of the aggregate service level in the per period cases is more than for the global case at the same value of service level. The β and γ service level are very close to each other in terms of the value of $\Delta Cost$, and δ service has the highest $\Delta Cost$ at the same value of service level.

1.8.6 Receding horizon implementation

In this section, we discuss experiments for the receding horizon implementation for the γ service level. To this end, we use the 10 instances of the base case in set B and for each instance we

generate 1,000 scenarios (which are different in the realized demand).

We first evaluate the static model when applied in a static way. In order to do this, we proceed as in the previous experiments. For each of the 10 instances, we solve the static model one time (taking into account the information about the average demand and the demand variability) over the full horizon. Next, this static solution is evaluated over 1,000 scenarios.

In order to evaluate the application of the static model in a receding horizon fashion, we need a slightly different approach. This approach is also evaluated over 1,000 scenarios. However, for the different scenarios, the resulting solution can be different. Hence, for each scenario we need to determine first the decisions that we would have taken using the static model in a receding horizon way, if that scenario happened. We hence need to resolve the model multiple times for each scenario. In our case, following the procedure explained in Section 8, the model is resolved at every period once the demand realization of the period preceding it has been revealed. Thus, for the problem with 10 periods and 1,000 scenarios, we solve the model $10 \times 1,000 = 10,000$ times for each instance. Second, once these decisions for a specific scenario are determined, we calculate the resulting service level and total cost for that specific scenario. This procedure is repeated for all scenarios and the results are aggregated to evaluate the quality of the solutions determined by the receding horizon implementation over the entire scenario set.

The results of these experiments are presented in Table 1.16. As can be seen, the aggregate service level results in a cost reduction in the receding horizon model as well. In both the separate and the aggregate service level, the receding horizon approach results in a cost reduction compared to the static approach. This is because the amount of inventory will be modified after each demand realization and hence the receding horizon approach provides a higher level of flexibility. This finding is in line with the experimental results in Dural-Selcuk et al. (2019) who find that for the single item non-stationary stochastic lot sizing problem with backorders, the application of the static strategy in a receding horizon framework provides very good results.

Figure 1.7 illustrates the total inventory of all products in each period for the static and receding horizon approach. The left diagram presents the case of the individual γ service level and the right diagram presents the case of the aggregate γ service levels. In the first period, the amount of inventory is the same for the static and receding horizon approach as the initial inventory and backlog quantities are the same for both cases. As we move forward through the planning hori-

Table 1.16: Static model VS receding horizon model (Service level = 95%)

	Static γ-separate		Static γ-aggregate		Receding γ-separate		Receding γ-aggregate	
#	SL	Total Cost	SL	Total Cost	SL	Total Cost	SL	Total Cost
1	96.2%	143517.7	95.5%	138657.3	96.7%	137078.1	95.2%	125529.5
2	96.3%	157445.4	95.6%	153294.4	96.3%	148328.2	95.2%	136340.4
3	96.1%	151985.9	95.4%	148021.8	96.3%	144516.0	95.2%	132131.7
4	96.1%	145469.0	95.6%	141612.7	96.7%	139169.4	95.3%	127617.8
5	96.1%	154312.7	95.4%	150089.6	96.5%	146423.8	95.3%	135092.9
6	96.0%	145791.0	95.2%	141559.2	96.6%	140404.8	95.2%	129132.2
7	96.1%	150583.3	95.4%	146112.8	96.4%	142223.8	95.3%	131618.1
8	96.2%	144425.4	95.5%	140364.5	96.5%	137671.5	95.2%	126662.5
9	96.2%	153857.0	95.5%	149216.9	96.4%	144683.2	95.2%	132660.1
10	96.0%	141645.6	95.4%	137595.8	96.6%	134989.5	95.2%	124515.5
Average	96.1%	148903.3	95.4%	144652.5	96.5%	141548.8	95.2%	130130.1

zon, the receding approach determines the setups and production amounts considering the realized demand whereas the inventory and backlog quantities are updated. The difference in the amounts of inventory between these two approaches increases gradually towards the end of the planning horizon. This shows that applying the static model in the receding horizon approach can reduce the stock level which alleviates some of the inherent limitations of the static model, in which the inability to react to changes during the horizon leads to large levels of inventory towards the end of the planning horizon. In practice, the use of the rolling/receding horizon implementation depends largely on the flexibility in the production to allow the production plan to be changed reactively in order to realize the benefits. This receding horizon implementation of the static model can be adapted to specific production configurations in practice by changing the update mechanisms to be aligned with the frozen periods used in the configuration.

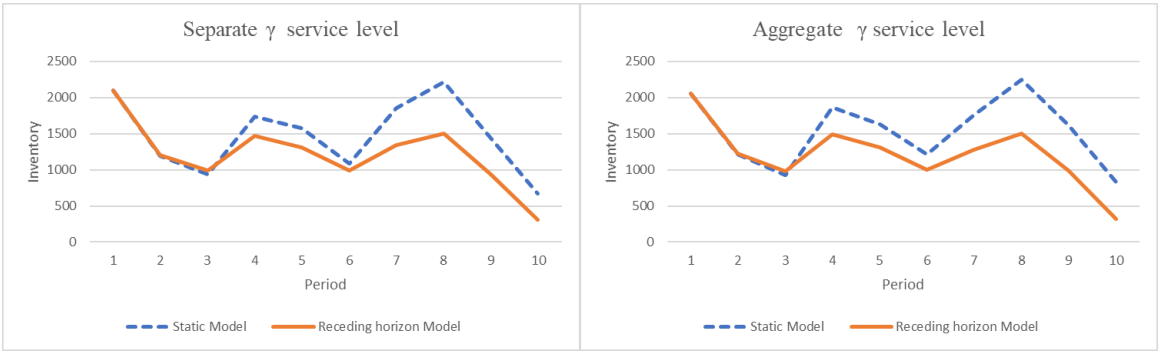


Figure 1.7: Average of total inventory per period (Static model VS receding horizon approach)

1.9 Conclusion

In this research, different aggregate service levels have been investigated in the context of multi-item capacitated lot-sizing problems. Such aggregate service levels allow the planner to flexibly assign different service levels to individual products so that they collectively satisfy the aggregate service level measures. These aggregate service levels can be used in conjunction with the commonly adopted service levels imposed on individual products. These service levels are the extensions of the well-known α_c, β, γ, and δ service levels. Having all these service levels investigated simultaneously, this research helps the companies to understand the difference and computational implications of the different service levels.

Since the mathematical models are non-linear, different approximations schemes are developed which are piece-wise linear and quantile-based approximations. Extensive numerical experiments are conducted to analyze the flexibility and cost savings of the aggregate service level. Using the aggregate service level provides flexibility to the problem which result in overall cost reductions. This cost reduction varies depending on different service levels and parameters. The numerical experiments show that the cost reduction is higher in the case where the aggregate service level is imposed in each period compared to the global case, and in quantity and time oriented service levels (β, γ, δ) compared to the event oriented one (α_c). The value of the aggregate service level is more obvious when there is a higher variability in the holding cost of the different products. This is also the case when service levels are more loose. With looser service levels there is more flexibility and the plant can save a lot considering an aggregate service level. At more strict service levels, when there is a limited capacity, there is a higher probability for the models with individual service level to be infeasible compared to the aggregate one. In general, the aggregate service level will provide some flexibility to the model which allows the production system to use its limited capacity more efficiently. It is also possible to consider the individual service level simultaneously with the aggregate service levels.

Investigating the static-dynamic strategy in the non-capacitated version of the problem and comparing it with the static strategy and the rolling/receding horizon approach is an interesting future research direction. Extending the problem to more general cases such as the distribution free case and no i.d.d assumption in demand also constitutes an interesting research direction.

To this end, the next step of this research is to approximate the problem using scenario-based formulations which can be used in dealing with any demand distributions. Such modeling scheme, albeit general, may not be scalable and this hence also justifies further research on the development of efficient solution frameworks.

References

Akçay, A., Biller, B., and Tayur, S. R. (2016). *Beta-Guaranteed aggregate Service Levels*. Available at SSRN.

Ben-Tal, A., Laurent, E. G., and Arkadi, N. (2009). *Robust Optimization*. Prinston University Press.

Bookbinder, J. H. and Tan, J.-Y. (1988). Strategies for the probabilistic lot-sizing problem with service-level constraints. *Management Science*, 34(9):1096–1108.

Brahimi, N., Absi, N., Dauzère-Pérès, S., and Nordli, A. (2017). Single-item dynamic lot-sizing problems: An updated survey. *European Journal of Operational Research*, 263(3):838–863.

Dural-Selcuk, G., Rossi, R., Kilic, O. A., and Tarim, S. A. (2019). The benefit of receding horizon control: Near-optimal policies for stochastic inventory control. *Omega*, 97(10209):1.

Escalona, P., Angulo, A., Weston, J., Stegmaier, R., and Kauak, I. (2019). On the effect of two popular service-level measures on the design of a critical level policy for fast-moving items. *Computers & Operations Research*, 107:107–126.

Gade, D. and Küçükyavuz, S. (2013). Formulations for dynamic lot sizing with service levels. *Naval Research Logistics*, 60(2):87–101.

Gruson, M., Cordeau, J.-F., and Jans, R. (2018). The impact of service level constraints in deterministic lot sizing with backlogging. *Omega*, 79:91–103.

Helber, S., Sahling, F., and Schimmelpfeng, K. (2013). Dynamic capacitated lot sizing with random demand and dynamic safety stocks. *OR Spectrum*, 35(1):75–105.

Jans, R. and Degraeve, Z. (2008). Modeling industrial lot sizing problems: a review. *International Journal of Production Research*, 46(6):619–1643.

Jiang, Y., Xu, J., Shen, S., and Shi, C. (2017). Production planning problems with joint service-level guarantee: a computational study. *International Journal of Production Research*, 55(1):38–58.

Kelle, P. (1989). ,optimal service levels in multi-item inventory systems. *Engineering Costs and Production Economics*, 15:375–379.

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

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

Mula, J., Poler, R., Garcia-Sabater, J. P., and Lario, F. C. (2006). Models for production planning under uncertainty: A review. *International Journal of Production Economics*, 103(1):271–285.

Pochet, Y. and Wolsey, L. A. (2006). *Production Planning by Mixed Integer Programming*. Springer, Science & Business Media, New York.

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

Rossi, R., Tarim, S. A., Prestwich, S., and Hnich, B. (2014). Piecewise linear lower and upper bounds for the standard normal first order loss function. *Applied Mathematics and Computation*, 231:489–502.

Shivsharan, C. T. (2012). *Optimizing the Safety Stock Inventory Cost Under Target Service Level Constraints*. Master of science), University of Massachusetts Amherst.

Stadtler, H. and Meistering, M. (2019). Model formulations for the capacitated lot-sizing problem with service-level constraints. *OR Spectrum*, 41(4):1025–1056.

Tarim, S. A. and Kingsman, B. G. (2004). The stochastic dynamic production/inventory lot-sizing problem with service-level constraints. *International Journal of Production Economics*, 88(1):105–119.

Tempelmeier, H. (2007). On the stochastic uncapacitated dynamic single-item lotsizing problem with service level constraints. *European Journal of Operational Research*, 181(1):184–194.

Tempelmeier, H. (2011). A column generation heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. *Omega*, 39(6):627–633.

Tempelmeier, H. (2013). Stochastic lot sizing problems. *Handbook of Stochastic Models and Analysis of Manufacturing System Operations*, pages 313–344.

Tempelmeier, H. and Herpers, Sascha, A. B. C. (2010). β–a heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. *International Journal of Production Research*, 48(17):5181–5193.

Tempelmeier, H. and Herpers, S. (2011). Dynamic uncapacitated lot sizing with random demand under a fillrate constraint. *European Journal of Operational Research*, 212(3):497–507.

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

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

Tunc, H., Kilic, O. A., Tarim, S. A., and Eksioglu, Burak, A. (2014). reformulation for the stochastic lot sizing problem with service-level constraints. *Operations Research Letters*, 42(2):161–165.

Tunc, H., Kilic, O. A., Tarim, S. A., and Eksioglu, B. A. (2013). simple approach for assessing the cost of system nervousness. *International Journal of Production Economics*, 141(2):619–625.

Tunc, H., Kilic, O. A., Tarim, S. A., and Rossi, R. (2018). An extended mixed-integer programming formulation and dynamic cut generation approach for the stochastic lot-sizing problem. *INFORMS Journal on Computing*, 30(3):492–506.

Van Pelt, T. D. and Fransoo, Jan C., A. (2018). note on “linear programming models for a stochastic dynamic capacitated lot sizing problem". *Computers & Operations Research*, 89:13–16.

Chapter 2

Flexibility in the Stochastic Multi-level Lot Sizing Problem with Service Level Constraints

Abstract

We investigate the stochastic multi-level lot sizing problem with a service level and in a general setting in which it is possible to have independent demand for the components as well. In this work, we present a systematic approach to evaluate the value of adding flexibility in such context. To this end, the problem with uncertain demand is modeled as a two-stage stochastic program considering different demand scenarios. We first consider at all levels a static strategy in which both the setup decisions and the production quantities are determined in the first stage before the demand is realized. We also model a more adaptive strategy to be more responsive to the realized demand when production quantity decisions of some items can be treated as recourse decisions. We investigate the value of applying such an adaptive strategy and adding more flexibility in the system under different settings. Three different bill of material (BOM) structures (serial, assembly, and general) are considered. We numerically show that adding flexibility to the system results in a cost savings depending on where we add the flexibility in the BOM. While controlling the variation in the plans is very important in the multi-level system, this research show that even having a small

degree of flexibility may result in a reasonable amount of cost savings.

2.1 Introduction

Being cost efficient is a crucial imperative in a competitive business environment. For manufacturing companies, having an efficient production plan in the context of a material requirements planning (MRP) system is important to minimize different costs of production and inventory control. In MRP, time-phased production and inventory plans are crucial decisions to make a balance between customers' demand satisfaction and cost management. While insufficient inventory will lead to shortages, unnecessary stocks will increase the inventory holding cost.

The standard lot sizing problem aims to determine the optimal timing and production quantities in order to satisfy known demand over a finite and discrete time horizon (Pochet and Wolsey, 2006). One of the extensions to the standard lot sizing problem is to consider the multi-level product structure which is common in MRP systems. While only independent demand exists for each of the products in the single level lot sizing problem, there is also dependent demand due to the bill of material (BOM) structure in a multi-level lot sizing problem. There are different product structures in the literature including serial, assembly, and general structures (Pochet and Wolsey, 2006).

Within the optimization models for production planning, where all the levels of the BOM are optimized simultaneously, the decision variable related to the backlog only exists for the items that have independent demand. This is due to the fact that, in order to produce the items at the lower levels, their components need to be available at the required time, and it is hence not possible to have backlog for the dependent demand (Hung and Chien, 2000). Indeed, if we would allow backlog at the component level, then a solution can exist in which there is some backlog at the component level, while there is no backlog at the end item level. In this research, we address a more general setting in which, in addition to the end items, each of the components in the BOM may also have an independent demand. Therefore, it is possible to have backlog for them due to this independent portion. This problem with demand at multiple levels in the BOM structure has practical relevance in industries with production and aftermarket services, which require spare parts (Wagner and Lindemann, 2008). A good example is the aerospace industry where, in addition

to the demand for end items, the components also have independent demand, which has to be taken into account in the planning process.

Even though demand is typically stochastic in nature, the calculations used in the MRP systems are based on the deterministic demand assumption while safety stocks of items with independent demand are separately determined to hedge against demand uncertainty. The use of safety stocks in such context can potentially lead to sub-optimal solutions since the calculations are performed in isolation under different assumptions than the deterministic model (Tempelmeier and Herpers, 2011). Unlike this approach which can potentially result in sub-optimal decisions, in this research, we will use stochastic optimization models to deal with demand uncertainty in a single framework. In these models, the lot sizing and safety stock level decisions are jointly determined as the demand's probability distributions are considered in the model (Helber et al., 2013; Thevenin et al., 2021).

Different forms of service levels are widely used in the calculations of safety stock to deal with demand uncertainty in stochastic lot-sizing problems. However, most of the research has been focused on the single level problem. In the multi-level problem, as it is not possible to have backlog for the dependent demand in the BOM, the service levels are defined only for the independent demand of the end products and the components. The service level which we consider in this research is closely related to the δ service level proposed by Helber et al. (2013). Here, instead of limiting average backlog, we limit the maximum proportion, taken over all demand realizations, of total backlog to the total possible backlog over the whole planning horizon.

In the single level lot sizing problem, there are three main strategies to deal with multi-period lot sizing problems with stochastic demand and these have a different approach for the setup and production decisions, namely the static, dynamic, and static-dynamic strategy (Bookbinder and Tan, 1988). In the static strategy, the setup and production decisions will be defined at the beginning of the planning horizon and they remain unchanged with the demand realization. In the dynamic strategy, both setups and production decisions may be modified after the demand realization. The static-dynamic strategy is the combination of these two strategies in which the setups are fixed at the beginning of the planning horizon and the production decisions are made after the demand realization. These strategies can also be applied in multi-level lot sizing problems. In the system in which we have independent demand for the components as well, we can apply different

strategies at different levels in the BOM to increase the responsiveness in the system, while keeping the nervousness under control. By allowing some production level decisions to be made in the second stage, we gain more flexibility and hence lower costs. We provide an illustrative example to demonstrate the benefits of such flexibility, later in this section.

In this research, we model the stochastic multi-level lot sizing problem as a two-stage stochastic programming model which is solved using the sample average approximation (SAA) formulation (Kleywegt et al., 2002; Shapiro et al., 2014). The contributions of this research can be stated as follows. First, we investigate the stochastic multi-level lot sizing problem with service level constraints. We specifically consider the case where independent demand exists not only for the end items but at the component level as well. Second, we model different variants of the problem which allow a different level of flexibility (i.e., static strategy versus static-dynamic strategy) at different levels in the BOM structure. Third, we apply the SAA method to empirically evaluate the solution quality with different number of scenarios. Fourth, we propose a systematic way to calculate the cost savings. Based on that, we perform extensive computational experiments to empirically validate the value of flexibility and derive managerial insights for this problem under different settings.

2.1.1 An illustrative example

The aim of this short section is to offer some intuition on how we may benefit from the production recourse in a multi-level lot sizing problem. We will use a small example which only considers one period. Assume we have two items in the BOM, one end item and one component. The average external demand for each of the items is 100 units. The stochastic demand for each item is represented by 3 independent demand scenarios with equal probability and the demand values of 50, 100, 150. Therefore, we have 9 scenarios in total based on different combinations of external demand for the two items as illustrated in Table 2.1. Based on these demand scenarios, we consider and compare three different cases. One without any flexibility and two with flexibility in which we have recourse for the end item production. In the first case, there is no flexibility, and the production decisions for both the component and end product are taken before the demand realization. For illustrative purposes, we assume that we will produce equal to the average demand. This results in a production of 100 for the end item and 200 for the component. The 200 units for the component

are based on the average external demand of the component itself and the internal demand coming from the end item. Since the production of both the end item and the component is fixed, there is no flexibility. More specifically, the 200 available units of the component will always be allocated in the same way: 100 units to satisfy the internal demand generated by the fixed production of 100 units of the end item, and the other 100 units to satisfy the external demand of the component. These allocation decisions are indicated in the table for each scenario. Because there is no flexibility, a situation might arise such as in scenario 3, where we have enough components (i.e., 200) to satisfy the external demand of the component (i.e., 50) and end item (i.e., 150), but because of the fixed production at of 100 units the end item level we end up with an inventory of 50 units at the component level while having a backlog of 50 at the end item level. The average backlog and inventory levels are calculated over the 9 different scenarios.

In the second case, we consider some level of flexibility in which we have a production recourse for the end item. As in the first case, we will produce 200 units for the component, but how much to produce for the second item is a recourse decision and will be defined based on the observed demand. The flexibility with respect to the production quantity of the end item results in flexibility in the allocation of the 200 units of the component, to satisfy the external demand of the component or to produce end item. This flexible allocation will define what portion of the produced component should be used for its own external demand and how much should be used for the end item production. In the second case, the production quantity for the end item is determined so that it satisfies as much as possible the external demand for this end item, while avoiding any inventory for the end item. In Table 2.1, the allocation of the 200 available units of the component to the component and to the end product are given. These decisions are now different in each scenario because the production decision for the end item is now a recourse decision. We observe here that for scenario 3, the flexibility in the production quantity for the end product now allows a flexible allocation of the 200 components: 50 to satisfy the external demand for the component and 150 to be allocated to the end product. The result is that demand for both the end product and the component is exactly satisfied without creating any backlog or inventory. In this case, the recourse decisions taken lead to an average of 22.2 units backlog and the same amount of inventory for the component, while the average backlog and inventory for the end item is equal to zero.

Case 3 is similar to case 2, but with slightly different recourse decisions taken, resulting in

Table 2.1: An illustrative example

Demand scenarios			Solution 1 with no flexibility						Solution 2 with flexibility at the end item						Solution 3 with flexibility at the end item					
			Comp	End	Comp	End	Comp	End	Comp	End	Comp	End	Comp	End	Comp	End	Comp	End	Comp	End
Production			200	100					200	Recourse					200	Recourse				
#	Comp	End	Alloc	Alloc	Inv	Inv	Back	Back	Alloc	Alloc	Inv	Inv	Back	Back	Alloc	Alloc	Inv	Inv	Back	Back
1	50	50	100	100	50	50	0	0	150	50	100	0	0	0	100	100	50	50	0	0
2	50	100	100	100	50	0	0	0	100	100	50	0	0	0	100	100	50	0	0	0
3	50	150	100	100	50	0	0	50	50	150	0	0	0	0	50	150	0	0	0	0
4	100	50	100	100	0	50	0	0	150	50	50	0	0	0	150	50	50	0	0	0
5	100	100	100	100	0	0	0	0	100	100	0	0	0	0	100	100	0	0	0	0
6	100	150	100	100	0	0	0	50	50	150	0	0	50	0	50	150	0	0	50	0
7	150	50	100	100	0	50	50	0	150	50	0	0	0	0	150	50	0	0	0	0
8	150	100	100	100	0	0	50	0	100	100	0	0	50	0	100	100	0	0	50	0
9	150	150	100	100	0	0	50	50	50	150	0	0	100	0	100	100	0	0	50	50
Avg	100	100	100	100	16.7	16.7	16.7	16.7	100	100	22.2	0	22.2	0	100	100	16.7	5.6	16.7	5.6

an average inventory and backlog level of 16.7 units for the component and 5.6 units for the end item. This result dominates the result of the first case. As we can see, in general the flexibility can reduce the average inventory and backlog in the system, but we may have several solutions to use this flexibility, which can be defined optimally based on the structure and different costs in the system. This small illustration makes clear that the flexibility with respect to the production quantity of the end product results in a flexible decision on the allocation of the fixed production quantity of the component.

2.2 Literature review

We organized the literature review into two sections. The first part discusses the lot sizing problem with service level constraints, and the second one discusses the multi-level lot sizing problem.

2.2.1 Stochastic lot sizing problem with service level constraints

The stochastic lot sizing problem with service level constraints has been studied extensively. Several service level measures have been proposed which can be classified as event-oriented service levels, quantity-oriented service levels, and time and quantity-oriented service levels (Sereshti et al., 2020). The α service level is an event-oriented service level which imposes a limit on the probability of stock out. The β service level or the fill rate is the proportion of the demand directly filled from stock and it is calculated based on the expected backorders to the expected demand. This service level is a quantity oriented service level. The γ service level limits the pro-

portion of expected backlog to expected demand. The δ service level is based on the proportion of total expected backlog to the maximum expected backlog. Both γ and δ service levels are time and quantity oriented service levels.

Many papers studied the lot sizing problem with service level constraints using different strategies and different types of service levels (Helber et al., 2013; Tavaghof-Gigloo and Minner, 2021; Tempelmeier, 2011; Tempelmeier and Herpers, 2011; Tempelmeier and Hilger, 2015; Tunc et al., 2014). These service levels are commonly considered in systems with one item, or multiple end items, but they are not considered in multi-level systems where we have BOM structures. The service level which we consider in this research is closely related to the δ service level proposed by Helber et al. (2013). Here, instead of limiting average backlog we limit the maximum proportion, taken over all demand realisations, of total backlog, to the total possible backlog over the whole planning horizon, for each product with external demand. This service level which we denote by δ' is more strict compared to the standard δ service level. While the δ service level is defined based on the averages over all scenarios, δ' is imposed for each of the scenarios separately. This per-scenario service level is also adopted in other similar problems. For example, Alvarez et al. (2020) investigate the inventory routing problem with stochastic supply and demand with a service level in which they limit the proportion of total lost sale to the total demand for each scenario.

2.2.2 Multi-level lot sizing problem

Several formulations have been considered for the multi-level lot sizing problem. Some formulations use the concept of echelon stock Afentakis et al. (1984); Afentakis and Gavish (1986), Pochet and Wolsey (2006) Akartunalı and Miller (2009)). This reformulation results in a separation of the different levels and allows the use of strong cuts or reformulations based on the single-level lot sizing problem, resulting in better bounds. Wu et al. (2011) investigated the capacitated multi-level lot sizing problem with backlogging, and proposed different mathematical models. These models are the common multi-level lot sizing model, the model based on echelon variables, the model based on the facility location formulation, and the model based on the shortest path formulation. As mentioned before, it is not possible to allow backlog decisions for the dependent demand in the BOM structure, and these backlogs are defined only for the independent demands in the system. Hung and Chien (2000) proposed a mathematical model for a multi-class multi-level capacitated

Table 2.2: Multi-level lot sizing research

Authors	Capacity	Backlogging	Independent demand at the component level	Stochasticity	Service level
Tempelmeier and Derstroff (1996)	+	-	-	-	-
Hung and Chien (2000)	+	+	+	-	-
Stadtler (2003)	+	-	-	-	-
Sahling et al. (2009)	+	-	-	-	-
Akartunalı and Miller (2009)	+	+	-	-	-
Almeder (2010)	+	-	-	-	-
Wu et al. (2011)	+	+	-	-	-
Seeanner et al. (2013)	+	-	-	-	-
Xiao et al. (2014)	-	-	-	-	-
Toledo and da Silva Arantes (2015)	+	+	-	-	-
You et al. (2019)	-	-	-	-	-
Thevenin et al. (2021)	-	-	-	+	-
Quezada et al. (2020)	-	-	+	+	-
Gruson et al. (2021)	-	-	-	+	-
Our work	+	+	+	+	+

lot sizing problem, where two classes of orders, i.e., confirmed and predicted, are considered. In this model, there are three different constraints for the inventory balance constraints; one for end products, one for components with external demand, and one for components without external demand. In addition to these constraints, there is also a set of constraints to ensure that there is no backlog for the dependent demand of the components. To solve this problem, the authors used simulated annealing, tabu search, and genetic algorithm heuristics. Table 2.2 shows some of the related works on multi-level lot sizing problem, and their similarities and differences with our work.

2.3 Mathematical formulation

In this section, we propose the mathematical models for the deterministic and stochastic multi-level lot sizing problems with a service level constraint. As mentioned, in this model, there is a possibility of having external (independent) demand for the components as well.

2.3.1 Deterministic model

The deterministic model is an extension of the model proposed by Hung and Chien (2000) in which, for each type of product in the system, there is a different set of inventory balance constraints. Table 2.3 provides the list of sets, parameters and decision variables in the model. In this

model, the structure of the BOM is considered by the successors of each item. We have multiple machines and the capacity is also defined for each machine separately. Furthermore, we consider the possibility of overtime.

Table 2.3: Notation for the multi-level deterministic lot sizing problem

Sets	**Definition**
$\mathscr{K}$	Set of products, indexed by $1,...,K$
$\mathscr{S}_k$	Set of immediate successors of product k
$\mathscr{T}$	Set of planning periods, indexed by $1,...,T$
$\mathscr{EI}$	Set of end items
$\mathscr{CW}$	Set of components without external demand
$\mathscr{CE}$	Set of components with external demand
$\mathscr{MC}$	Set of machines
$\mathscr{K}_m$	Set of products that are produced on machine m
Parameters	**Definition**
cap_{mt}	Production capacity for machine m in period t
d_{kt}	External demand of item k in period t
bc_{kt}	Backlog cost for product k in period t
hc_{kt}	Inventory holding cost for product k in period t
$\bar{M}_{kt}, M'_{kt}$	A sufficiently large number
pt_{kt}	Unit production time for product k in period t
r_{ki}	Number of units of item k required to produce one unit of the immediate successor item i
oc_{mt}	Overtime cost for machine m in period t
sc_{kt}	Setup cost for product k in period t
st_{kt}	Setup time for product k in period t
δ, δ'	Target service level
Decision variables	**Definition**
y_{kt}	Binary variable which is equal to 1 if there is a setup for product k in period t, 0 otherwise
x_{kt}	Amount of production for product k in period t
o_{mt}	Overtime for machine m in period t
B_{kt}	Amount of backlog for product k at the end of period t
I_{kt}	Amount of physical inventory for product k at the end of period t

$$\text{Min} \sum_{t\in\mathscr{T}}\sum_{k\in\mathscr{K}}(sc_{kt}y_{kt}+hc_{kt}I_{kt})+\sum_{t\in\mathscr{T}}\sum_{m\in\mathscr{MC}}oc_{mt}o_{mt}+\sum_{k\in\mathscr{K}}bc_{kT}B_{kT} \tag{2.1a}$$

Subject to:

$$I_{k,t-1}+x_{kt}+B_{kt}=I_{kt}+B_{k,t-1}+d_{kt} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{EI} \tag{2.1b}$$

$$I_{k,t-1}+x_{kt}=I_{kt}+\sum_{i\in\mathscr{S}_k}r_{ki}x_{it} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{CW} \tag{2.1c}$$

$$I_{k,t-1}+x_{kt}+B_{kt}=I_{kt}+B_{k,t-1}+d_{kt}+\sum_{i\in\mathscr{S}_k}r_{ki}x_{it} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{CE} \tag{2.1d}$$

$$B_{kt} - B_{k,t-1} \leq d_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CE} \quad (2.1e)$$

$$x_{kt} \leq \bar{M}_{kt} y_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \quad (2.1f)$$

$$\sum_{k \in \mathscr{K}_m} (st_{kt} y_{kt} + pt_{kt} x_{kt}) \leq cap_{mt} + o_{mt} \qquad \forall m \in \mathscr{MC}, \forall t \in \mathscr{T} \quad (2.1g)$$

$$\frac{\sum_{t \in T} B_{kt}}{\sum_{t \in T} (T - t + 1) d_{kt}} \leq 1 - \delta \qquad \forall k \in \mathscr{EI} \cup \mathscr{CE} \quad (2.1h)$$

$$y_{kt} \in \{0, 1\} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \quad (2.1i)$$

$$B \in \mathbb{R}_+^{KT}, I \in \mathbb{R}_+^{KT}, x \in \mathbb{R}_+^{KT}, o \in \mathbb{R}_+^{MCT} \quad (2.1j)$$

The objective function (2.1a) minimizes the setup cost, holding cost, overtime cost and the cost of unsatisfied demand at the end of the planning period. Constraints (2.1b-2.1d) are the inventory balance constraints. Constraints (2.1b) are the inventory balance constraints for each of the end items in each planning period, and constraints (2.1c) and (2.1d) are for the components without and with external demand, respectively. Constraints (2.1e) ensure that the amount of backorder in each planning period cannot be more than the period external demand (Hung and Chien, 2000). These constraints are only imposed for the components with external demand and they ensure that any backlog is only directly related to the independent demand of the component and there is no backlog for dependent demands. Constraints (2.1f) are the production setup constraints. $\bar{M}_{kt}$ which is the maximum possible production is calculated using equations (2.2) and (2.3) in a similar fashion as in Toledo and da Silva Arantes (2015). These calculations need to be done recursively, starting from the end items. Constraints (2.1g) are the capacity constraints. Constraints (2.1h) are the δ service level constraints in the deterministic setting (Gruson et al., 2018).

$$D_{k(t..T)} = \sum_{u=t}^{T} d_{ku} + \sum_{i \in \mathscr{S}_k} r_{ki} D_{i(t..T)} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \quad (2.2)$$

$$\bar{M}_{kt} = D_{k(1..T)} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \quad (2.3)$$

2.3.2 Static stochastic model with service level

In this section, we present the model for the stochastic capacitated multi-level lot sizing problem. In this first variant, as in (Bookbinder and Tan, 1988) we assume that the strategy is static which

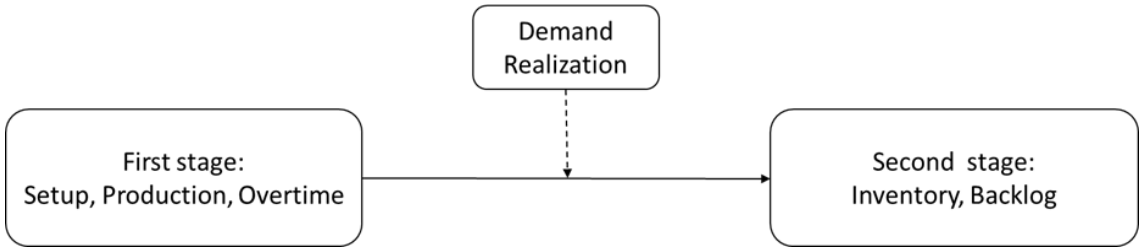


Figure 2.1: Sequence of events for the case with no flexibility

implies that the setup and production quantity decisions are determined at the beginning of the planning horizon and cannot be changed. Figure 2.1 illustrates the sequence of decisions in this problem. The setup and production, and overtime decisions are the first stage variables which are defined before the demand realization. It is worthwhile to mention that in general the overtime is a part of the recourse decisions in our problem definition. However, in the static case, there is no benefit in defining overtime as a second stage variable since all the production amounts are determined in the first stage and hence the overtime does not depend on the demand realization. Thus, for the static case, we include this variable in the first-stage model for simplicity. After the demand realization (for the entire planning horizon), the resulting inventory and backlog levels are determined for each scenario in the second stage (Helber et al., 2013). In this problem, the model guarantees that for each product with external demand, the proportion of backlog divided by the maximum possible backlog considering any realization is less than $(1-\delta')$. As mentioned earlier, this service level is more strict than the standard δ service level which is defined based on the expected value of the backlog (Helber et al., 2013). We model this problem as a two-stage stochastic mixed integer program. To account for the stochasticity, a random vector $d = (\tilde{d}_{11},...,\tilde{d}_{KT})$ is considered, where $\tilde{d}_{kt}$ represents the random demand for product k, in period t. This model is represented in (2.4a)-(2.4e).

$$v^* := \text{Min } F(y,o) + \mathbb{E}_d[Q(x,d)] \tag{2.4a}$$

Subject to:

$$x_{kt} \leq M'_{kt} y_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \tag{2.4b}$$

$$\sum_{k \in \mathscr{K}_m} (st_{kt} y_{kt} + pt_{kt} x_{kt}) \leq cap_{mt} + o_{mt} \qquad \forall m \in \mathscr{MC}, \forall t \in \mathscr{T} \tag{2.4c}$$

$$y \in \{0,1\}^{KT} \tag{2.4d}$$

$$x \in \mathbb{R}_+^{KT}, o \in \mathbb{R}_+^{MC,T} \tag{2.4e}$$

In this model, the first-stage cost function is defined as (2.5a) which minimizes the setup and overtime costs, and the second stage cost function is represented in (2.5b-2.5h).

$$F(y,o) = \sum_{t\in\mathscr{T}} \Big(\sum_{k\in\mathscr{K}} sc_{kt} y_{kt} + \sum_{m\in\mathscr{MC}} oc_{mt} o_{mt} \Big) \tag{2.5a}$$

$$Q(x,d) = min \sum_{t\in\mathscr{T}} \sum_{k\in\mathscr{K}} hc_{kt} I_{kt} + \sum_{k\in\mathscr{K}} bc_{kT} B_{kT} \tag{2.5b}$$

Subject to:

$$I_{k,t-1} + x_{kt} + B_{kt} = I_{kt} + B_{k,t-1} + \tilde{d}_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{EI} \tag{2.5c}$$

$$I_{k,t-1} + x_{kt} = I_{kt} + \sum_{i\in\mathscr{S}_k} r_{ki} x_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CW} \tag{2.5d}$$

$$I_{k,t-1} + x_{kt} + B_{kt} = I_{kt} + B_{k,t-1} + \tilde{d}_{kt} + \sum_{i\in\mathscr{S}_k} r_{ki} x_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CE} \tag{2.5e}$$

$$B_{kt} - B_{k,t-1} \leq \tilde{d}_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CE} \tag{2.5f}$$

$$\frac{\sum_{t\in T} B_{kt}}{\sum_{t\in T}(T-t+1)\tilde{d}_{kt}} \leq 1-\delta' \qquad \forall k \in \mathscr{EI} \cup \mathscr{CE} \tag{2.5g}$$

$$I \in \mathbb{R}_+^{KT}, B \in \mathbb{R}_+^{KT} \tag{2.5h}$$

The objective function (2.5b) minimizes the holding cost, and the cost of unsatisfied demand at the end of the planning period. Constraints (2.5c-2.5e) are the inventory balance constraints. Constraints (2.5f) limit the maximum amount of backorder in each planning period. Constraints (2.5g) are the service level constraints.

2.3.3 Stochastic model with service level and production recourse

In the previous section, we presented the model based on the static strategy which does not allow any flexibility in the production decisions. In this section, we consider the case where we have production recourse and flexibility can be allowed for some products by assuming that they can follow a static-dynamic strategy, while other products keep following the static strategy.

In this model, the production amounts of the products with no flexibility are part of the first-stage decisions, and the production for the rest of them are part of the second stage decisions which

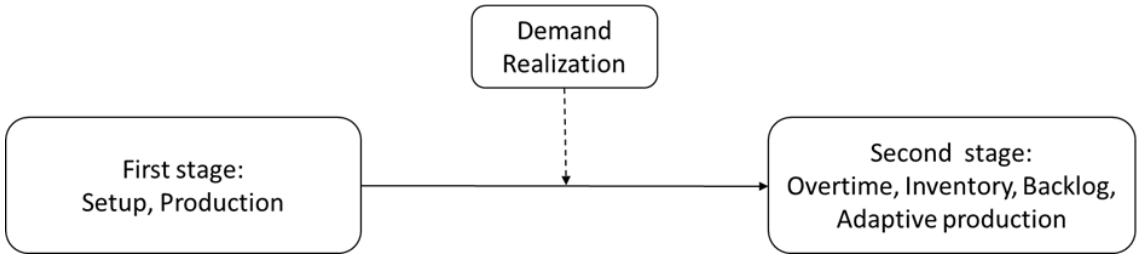


Figure 2.2: Sequence of events for the case with flexibility

are determined after the demand realization (Figure 2.2). Table (2.4) presents the additional set and variables required for this model.

Table 2.4: Additional notation for the stochastic model with production flexibility

Set	**Definition**
$Flex$	Set of products with flexible production
Random variables	**Definition**
X_{kt}	Amount of production for product $k \in Flex$ in period t
Decision variables	**Definition**
x_{kt}	Amount of production for product $k \in \mathscr{K} \setminus Flex$ in period t

The stochastic multi-level lot sizing problem with flexibility is presented in (2.6a)-(2.6d).

$$v^* := \text{Min } F'(y) + \mathbb{E}_d[Q'(y,x,d)] \tag{2.6a}$$

Subject to:

$$x_{kt} \leq M'_{kt} y_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K}, k \setminus Flex \tag{2.6b}$$

$$y \in \{0,1\}^{KT} \tag{2.6c}$$

$$x \in \mathbb{R}_+^{KT} \tag{2.6d}$$

In model 2.6, the first-stage cost function is defined as (2.7a) and the second stage recourse model is represented in (2.7b-2.7m).

$$F'(y) = \sum_{t \in \mathscr{T}} \sum_{k \in \mathscr{K}} sc_{kt} y_{kt} \tag{2.7a}$$

$$Q'(y,x,d) = \sum_{t \in \mathscr{T}} \Big(\sum_{k \in \mathscr{K}} hc_{kt} I_{kt} + \sum_{m \in \mathscr{M}\mathscr{C}} oc_{mt} O_{mt} \Big) + \sum_{k \in \mathscr{K}} bc_{kT} B_{kT} \tag{2.7b}$$

Subject to:

$$X_{kt} \leq M_{kt} y_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K}, k \in Flex \quad (2.7c)$$

$$\sum_{k \in \mathscr{K}_m} st_{kt} y_{kt} + \sum_{k \in \mathscr{K}_m \setminus Flex} pt_{kt} x_{kt} + \sum_{k \in \mathscr{K}_m \cap Flex} pt_{kt} X_{kt}$$

$$\leq cap_{mt} + O_{mt} \qquad \forall m \in \mathscr{MC}, \forall t \in \mathscr{T} \quad (2.7d)$$

$$I_{k,t-1} + x_{kt} + B_{kt} = I_{kt} + B_{k,t-1} + \tilde{d}_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{EI}, k \setminus Flex \quad (2.7e)$$

$$I_{k,t-1} + X_{kt} + B_{kt} = I_{kt} + B_{k,t-1} + \tilde{d}_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{EI}, k \in Flex \quad (2.7f)$$

$$I_{k,t-1} + x_{kt} = I_{kt} + \sum_{i \in \mathscr{S}_k \setminus Flex} r_{ki} x_{it} + \sum_{i \in \mathscr{S}_k \cap Flex} r_{ki} X_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CW}, k \setminus Flex \quad (2.7g)$$

$$I_{k,t-1} + X_{kt} = I_{kt} + \sum_{i \in \mathscr{S}_k \setminus Flex} r_{ki} x_{it} + \sum_{i \in \mathscr{S}_k \cap Flex} r_{ki} X_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CW}, k \in Flex \quad (2.7h)$$

$$I_{k,t-1} + x_{kt} + B_{kt} = I_{kt} + B_{k,t-1} + \tilde{d}_{kt} + \sum_{i \in \mathscr{S}_k \setminus Flex} r_{ki} x_{it}$$

$$+ \sum_{i \in \mathscr{S}_k \cap Flex} r_{ki} X_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CE}, k \setminus Flex \quad (2.7i)$$

$$I_{k,t-1} + X_{kt} + B_{kt} = I_{kt} + B_{k,t-1} + \tilde{d}_{kt} + \sum_{i \in \mathscr{S}_k \setminus Flex} r_{ki} x_{it}$$

$$+ \sum_{i \in \mathscr{S}_k \cap Flex} r_{ki} X_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CE}, k \in Flex \quad (2.7j)$$

$$B_{kt} - B_{k,t-1} \leq \tilde{d}_{kt} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{CE} \quad (2.7k)$$

$$\frac{\sum_{t \in T} B_{kt}}{\sum_{t \in T} (T - t + 1) d_{kt}} \leq 1 - \delta' \qquad \forall k \in \mathscr{EI} \cup \mathscr{CE} \quad (2.7l)$$

$$I \in \mathbb{R}_+^{KT}, B \in \mathbb{R}_+^{KT}, X \in \mathbb{R}_+^{KT}, O \in \mathbb{R}_+^{MCT} \quad (2.7m)$$

Notable differences between model 2.5 and model 2.7 are in the presence of the recourse variable X_{it} which shows the adaptive production, and of the recourse variable O_{it} which shows the overtime and cannot be moved to the first stage decisions anymore. Each of the inventory balance constraints in model 2.5 should be separately considered for the items with and without adaptive production in model 2.7.

2.4 Sample average approximation

We cannot directly solve the two-stage programming models (2.4) and (2.6) due to the presence of the random demand vector d and the expectation terms in their objective functions (2.4a) and

(2.6a). Having different random variables, we need to consider all possible realizations of the random parameters to have exact evaluation of this expectation term. To tackle this computational difficulty, we apply the sample average approximation (SAA) method. SAA is a Monte Carlo simulation-based method to solve the stochastic optimization problems in which the random distribution is replaced by a finite number of scenarios and the true expected value of the objective function is approximated by the average cost over the scenarios (Shapiro et al., 2014). To generate a scenario sample for the random vector $d = \{d_s\}_{s \in S}$, Monte Carlo sampling is used and an equal probability is assigned to each scenario.

The quality of the approximation in the stochastic approach mainly depends on the number of scenarios used (Mousavi et al., 2021). The challenge here is to define a proper number of scenarios which can provide a near-optimal approximation of the original problem. This number is defined based on the statistical lower bound and upper bound of the optimal solution. The SAA procedure, the definition of these bounds, and the gap between them are explained next.

In the SAA procedure, we have a set of scenario $\mathscr{S} = \{1, 2, ..., S\}$. First we choose the initial sample sizes S, and the number of SAA replications M. Then, for each replication $m = 1$ to M, an instance with S scenarios is generated and the corresponding SAA model is solved based on the chosen set of scenarios. Let the $\hat{v}_m^S$ and $\hat{\sigma}_m^S$, be the optimal objective function value and the solution for replication m, respectively. Equation (2.8) defines the expected value of the lower bound of v^*, the optimal objective value for the original problem, based on M replications of size S, denoted by $L_{M,S}^{mean}$.

$$L_{M,S}^{mean} = E(\hat{v}_M) = \frac{1}{M} \sum_{m=1}^{M} \hat{v}_m^S \tag{2.8}$$

To estimate the upper bound, we generate a large enough sample set $\mathscr{S}^{eval} = \{1, 2, ..., S^{eval}\}$, where $S^{eval} \gg S$. Having a solution $\hat{\sigma}$ as the first stage decision, $U_{M,S}$ (2.9) is an estimation of the upper bound, in which $\hat{g}(\mathscr{S}^{eval}, \hat{\sigma})$ is the objective function of the SAA formulation when the first stage solution, $\hat{\sigma}$, is fixed and when scenario set S^{eval} is used.

$$U_{M,S} = \hat{g}(\mathscr{S}^{eval}, \hat{\sigma}) \tag{2.9}$$

It should be noted that in this procedure, we have M different feasible solutions to calculate the upper bound. Among those we will choose the one with the smallest estimated objective value

which is calculated based on a scenario set $\mathscr{S}^{eval}$, denoted by σ^* (Verweij et al., 2003) (see eq. 2.10).

$$\sigma^* = arg\ min\{\hat{g}(\mathscr{S}^{eval}, \hat{\sigma}) : \hat{\sigma} \in \{\hat{\sigma}_1^S, ..., \hat{\sigma}_M^S\}\} \tag{2.10}$$

The empirical gap between the $U_{M,S}$ and $L_{M,S}^{mean}$ is used to define the proper number of scenarios to run the experiments to investigate the flexibility. In the following section we will present the SAA formulation of the two-stage models (2.4) and (2.6), as models (2.11) and (2.14), respectively. Applications of this method can also be found in Contreras et al. (2011), Taş et al. (2019), and Mousavi et al. (2021).

2.4.1 Reformulating the SAA problem

In this section, we present the extensive forms of SAA formulation (Mousavi et al., 2021), for both the static and adaptive models. Additional parameters and decision variables are presented in Table 2.5 and the stochastic models are presented afterward.

Table 2.5: Additional parameters and variables used in the SAA formulations

Parameters	**Definition**
d_{kts}	Demand for product k in period t in scenario s
Decision variables	**Definition**
X_{kts}	The production amount for product $k, k \in Flex$, in period t for scenario s
O_{mts}	The backorder for machine m in period t for scenario s
B_{kts}	Backlog for product k in period t for scenario s
I_{kts}	Amount of inventory for product k at the end of period t in scenario s

The static SAA model

Following the *s*tatic strategy, the setup and production, and overtime variables are the first stage variables and they do not have any index for the different scenarios in this model. The resulting inventory and backlog for each scenario are determined in the second stage.

$$\text{Min} \sum_{t\in\mathscr{T}}\sum_{k\in\mathscr{K}} sc_{kt}y_{kt} + \sum_{t\in\mathscr{T}}\sum_{m\in\mathscr{MC}} oc_{lt}o_{mt} + \sum_{t\in\mathscr{T}}\sum_{k\in\mathscr{K}} hc_{kt}\frac{\sum_{s\in\mathscr{S}} I_{kts}}{S} + \sum_{k\in\mathscr{K}} bc_{kT}\frac{\sum_{s\in\mathscr{S}} B_{kTs}}{S} \tag{2.11a}$$

Subject to constraints (2.4b) - (2.4d), and:

$$I_{k,t-1,s} + x_{kt} + B_{kts} = I_{kts} + B_{k,t-1,s} + d_{kts} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{E}\mathscr{I}, \forall s \in \mathscr{S} \qquad (2.11b)$$

$$I_{k,t-1,s} + x_{kt} = I_{kts} + \sum_{i \in \mathscr{S}_k} r_{ki} x_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{C}\mathscr{W}, \forall s \in \mathscr{S} \qquad (2.11c)$$

$$I_{k,t-1,s} + x_{kt} + B_{kts} = I_{kts} + B_{k,t-1,s} + d_{kts} + \sum_{i \in \mathscr{S}_k} r_{ki} x_{it} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{C}\mathscr{E}, \forall s \in \mathscr{S} \qquad (2.11d)$$

$$B_{kts} - B_{k,t-1,s} \leq d_{kts} \qquad \forall t \in \mathscr{T}, \forall k \in CE, \forall s \in \mathscr{S} \qquad (2.11e)$$

$$\frac{\sum_{t \in T} B_{kts}}{\sum_{t \in T} (T - t + 1) d_{kts}} \leq 1 - \delta' \qquad \forall k \in \mathscr{E}\mathscr{I}, CE, \forall s \in \mathscr{S} \qquad (2.11f)$$

$$B \in \mathbb{R}_+^{KTS}, I \in \mathbb{R}_+^{KTS} \qquad (2.11g)$$

The objective function (2.11a) is to minimize the setup cost, overtime cost, and the expected value of the inventory holding costs and the backlog in the last period. Constraints (2.11b-2.11d) are the inventory balance constraints which are defined for all products in all periods, and for all the scenarios. Constraints (2.11e) define the limit for the maximum amount of backorder for each component with external demand in each period for each scenario. Constraints (2.11f) show the service level for each product and each scenario, in which the proportion of total backlog to the maximum possible backlog for each scenario should not exceed the threshold percentage set by service level δ'. The parameter M'_{kt} in the production setup constraint is calculated recursively using equations (2.12) and (2.13).

$$D_{k(t..T)s} = \sum_{u=t}^{T} d_{kus} + \sum_{i \in \mathscr{S}_k} r_{ki} D_{i(t..T)s} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K}, \forall s \in \mathscr{S} \qquad (2.12)$$

$$M'_{kt} = \max_{s \in \mathscr{S}} D_{k(1..T)s} \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \qquad (2.13)$$

The SAA model with flexibility

Similar to the model without flexibility, in the SAA formulation for case with flexibility, the model (2.6) is reformulated as model (2.14).

$$\text{Min} \sum_{t \in \mathscr{T}} \sum_{k \in \mathscr{K}} (sc_{kt} y_{kt} + h_{kt} \frac{\sum_{s \in \mathscr{S}} I_{kts}}{S}) + \sum_{t \in \mathscr{T}} \sum_{m \in \mathscr{M}\mathscr{C}} oc_{mt} \frac{\sum_{s \in \mathscr{S}} O_{mts}}{S} + \sum_{k \in \mathscr{K}} bc_k \frac{\sum_{s \in \mathscr{S}} B_{kTs}}{S} \qquad (2.14a)$$

Subject to constraints (2.6b), (2.6c), (2.11e), (2.11f), and:

$$\sum_{k\in\mathscr{K}_m} st_{kt}y_{kt} + \sum_{k\in\mathscr{K}_m\setminus Flex} pt_{kt}x_{kt} + \sum_{k\in\mathscr{K}_m\cap Flex} pt_{kt}X_{kts}$$

$$\leq cap_{mt} + O_{mts} \qquad \forall s\in S, \forall m\in\mathscr{MC}, \forall t\in\mathscr{T} \quad (2.14b)$$

$$I_{k,t-1,s} + x_{kt} + B_{kts} = I_{kts} + B_{k,t-1,s} + d_{kts} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in\mathscr{EI}, k\setminus Flex \quad (2.14c)$$

$$I_{k,t-1,s} + X_{kts} + B_{kts} = I_{kts} + B_{k,t-1,s} + d_{kts} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in\mathscr{EI}, k\in Flex \quad (2.14d)$$

$$I_{k,t-1,s} + x_{kt} = I_{kts} + \sum_{i\in\mathscr{S}_k\setminus Flex} r_{ki}x_{it}$$

$$+ \sum_{i\in\mathscr{S}_k\cap Flex} r_{ki}X_{its} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in\mathscr{CW}, k\setminus Flex \quad (2.14e)$$

$$I_{k,t-1,s} + X_{kts} = I_{kts} + \sum_{i\in\mathscr{S}_k\setminus Flex} r_{ki}x_{it}$$

$$+ \sum_{i\in\mathscr{S}_k\cap Flex} r_{ki}X_{its} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in\mathscr{CW}, k\in Flex \quad (2.14f)$$

$$I_{k,t-1,s} + x_{kt} + B_{kts} = I_{kts} + B_{k,t-1,s} + d_{kts}$$

$$+ \sum_{i\in\mathscr{S}_k\setminus Flex} r_{ki}x_{it} + \sum_{i\in\mathscr{S}_k\cap Flex} r_{ki}X_{its} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in\mathscr{CE}, k\setminus Flex \quad (2.14g)$$

$$I_{k,t-1,s} + X_{kts} + B_{kts} = I_{kts} + B_{k,t-1,s} + d_{kts}$$

$$+ \sum_{i\in\mathscr{S}_k\setminus Flex} r_{ki}x_{it} + \sum_{i\in\mathscr{S}_k\cap Flex} r_{ki}X_{its} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in\mathscr{CE}, k\in Flex \quad (2.14h)$$

$$X_{kts} \leq M'_{kt}y_{kt} \qquad \forall s\in S, \forall t\in\mathscr{T}, \forall k\in Flex \quad (2.14i)$$

$$B\in\mathbb{R}_+^{KTS}, I\in\mathbb{R}_+^{KTS}, x\in\mathbb{R}_+^{KT}, X\in\mathbb{R}_+^{KTS}, O\in\mathbb{R}_+^{MC,T} \quad (2.14j)$$

The objective function (2.14a) minimizes the setup cost, plus the expected value of inventory holding cost, overtime cost, and unsatisfied demand at the end of planning period. Constraints (2.14b) are the capacity constraints for each production level, each scenario in each period. In this constraint the overtime is different for each demand scenario. Constraints (2.14c - 2.14h) are the inventory balance constraints. The difference between these constraints and inventory balance constraints in the model without any flexibility is that we have separate constraints for the items that have flexibility and the items that do not have it.

SAA implementation

In the SAA procedure, we first solve the SAA model to define the first stage decisions which we refer as the planning phase. In this phase, we solve the models using a set of sampled demand scenarios. In the second phase or evaluation phase, after fixing the first stage solution, we solve the model using a new and larger set of scenarios. Since each of the evaluation scenarios differ from the scenarios considered initially in the planning phase, this may cause infeasibility due to the service level constraint. Note that the overtime decision, even in the case of the static-dynamic strategy, does not rule out this infeasibility if the flexibility level is insufficient. More specifically, we may have some cases where we are not allowed to produce more of an item since the item's component production is fixed in the planning phase or the first stage and it is not possible to increase it. In case of full flexibility, i.e., all the production quantities are determined after the demand realization, the model will always be feasible with the larger set of scenarios.

The SAA method requires that the model must be feasible in order to calculate the SAA bounds. Thus, to alleviate this issue, we make use of a set of auxiliary variables which are associated with a penalty cost. To this end, we need to change the service level constraint from a hard one to a soft constraint in the the evaluation model. There is an additional penalty variable, ε_{ks}, for the violation in the service level constraint and a penalty cost, P, for this violation in the objective function. To have a consistent model in both the planning and the evaluation, we use constraints (2.16) instead of the service level constraints (2.11f) and (2.15) instead of the objective function (2.14a) in the extensive form model (2.14) which will be used both in planning and evaluation.

$$Min \sum_{t\in\mathscr{T}}\sum_{k\in\mathscr{K}}(sc_{kt}y_{kt}+h_{kt}\frac{\sum_{s\in\mathscr{S}}I_{kts}}{S})+\sum_{t\in\mathscr{T}}\sum_{m\in\mathscr{MC}}oc_{mt}\frac{\sum_{s\in\mathscr{S}}O_{mts}}{S}$$
$$+\sum_{k\in\mathscr{K}}(bc_k\frac{\sum_{s\in\mathscr{S}}B_{kTs}}{S}+\frac{\sum_{s\in\mathscr{S}}P\varepsilon_{ks}}{S}) \tag{2.15}$$

$$\frac{\sum_{t\in T}B_{kts}}{\sum_{t\in T}(T-t+1)d_{kts}}\leq 1-\delta'+\varepsilon_{ks} \qquad \forall k\in\mathscr{K},\forall s\in\mathscr{S} \tag{2.16}$$

In our experiments, the value of P is set high enough so that the value of ε in the planning phase becomes zero, which guarantees that the plan is feasible with respect to all the demand scenarios used in the planning phase, and the service level constraint (2.11f) is not violated in this phase.

For the evaluation of the decisions, we generate a large number of scenarios, and we solve the extensive form of the problem. The average of the objective functions over all of these scenarios will be the result of the evaluation which is also the the upper bound of the optimal value which was explained in section 5 in the SAA procedure. In the numerical experiments, the percentage of violated scenarios and how much the service level constraint is violated are also calculated in the evaluation phase.

2.5 Numerical experiments

2.5.1 Instance generation

This research combines a multi-level lot sizing problem with external demand for the components, and the stochastic lot sizing with service level constraints. To the best of our knowledge, instances for this problem are not available in the literature. To this end, we make use of two data sets in the literature and adapt them to the problem considered in this paper. More specifically, we modified the method used by Helber et al. (2013) for the stochastic lot sizing problem with service level, and adapted the instance generation method presented by Tempelmeier and Derstroff (1996) for the multi-level lot sizing and different BOM structures, as follows:

- We consider three different structures, serial, assembly and general (Figure 2.3). For the last two structures, we follow the Tempelmeier and Derstroff data (Tempelmeier and Derstroff, 1996).

- The holding cost of an item is equal to the sum of the holding cost of its components, multiplied by $(1+HValue)$. A High *HValue* means higher value adding operations at each level. We will also consider a case in which all the values for *HValue* are equal to 0 which means that the holding cost of an item is equal to the sum of the holding costs of its components, without any added value (Tempelmeier and Derstroff, 1996).

- Following Tempelmeier and Derstroff (1996), five different Time Between Orders (TBO) profiles are considered (Table 2.6). For example, for the serial BOM, in the first profile, the

TBO for all the items is equal to 1, and in the fourth profile, the TBO for the first 2 items is equal to 1, for the second item it is equal to 2 and for the last two items it is equal to 4.

- The capacity is defined for each machine. In our instances, exactly one machine is assigned to each level of BOM, and all the unit processing times are equal to 1. The capacity of each machine is equal to the sum of the average demand of the items assigned to a machine, divided by a parameter $Util$. The average demand of an item is equal to the sum of dependent and independent demand. Note that $Util$ is a parameter for data generation and it does not show the actual utilized capacity. We also considered the case without any capacity.

- The setup cost is determined based on the TBO, average demand and the holding cost, using equation (2.17) (Helber et al., 2013).

$$sc_{kt} = \frac{E[\overline{D}_{kt}] \times TBO^2 \times hc_{kt}}{2} \tag{2.17}$$

- We will have different levels of flexibility, which is defined based on the items for which production quantities are not fixed at the beginning of the planning horizon and they may be modified when the demand is observed. 0 means no flexibility, and i means flexibility for all the items until the $(i+1)^{th}$ level of BOM structure.

- Average demand profile for different structures. There are three different patterns for the external demand average. The first one is constant average independent demand for all the items at different levels. The second one is in increasing order and the third one is in decreasing orders from the end item level to the following levels (lowest to the highest level). For a given item, the average demand remains the sames over the horizon. The patterns and their detailed demand generations are summarized as follows:

    1. Constant external demand in which the average external demand ($dl = 100$) is multiplied by 1 at subsequent levels. For example if there is an external demand for any of the items at different levels it is equal to 100.

    2. The increasing order of demand in which the average external demand ($dl = 100$) is multiplied by the (level of the item +1). For example the external demand for the items at level 0 is equal to 100, and for the items at level 1 is equal to 200, is there is any.

Table 2.6: Different TBO profiles

	Serial product structure			Assembly product structure			General product structure		
TBO profile	TBO = 1	TBO = 2	TBO = 4	TBO = 1	TBO = 2	TBO = 4	TBO = 1	TBO = 2	TBO = 4
1	1 … 5	-	-	1 … 10	-	-	1 … 10	-	-
2	-	1 … 5	-	-	1 … 10	-	-	1 … 10	-
3	-	-	1 … 5	-	-	1 … 10	-	-	1 … 10
4	1,2	3	4,5	1	2 … 4	5 … 10	1 … 4	5 … 7	8 … 10
5	4,5	3	1,2	5 … 10	2 … 4	1	8 … 10	5 … 7	1 … 4

3. The decreasing order of demand in which the average external demand ($dl = 100$) is multiplied by (Max level - level of the item +1). For example, the external demand for the item at level 0 of the serial structure is equal to 500, and for the assembly and general structures it is equal to 300.

- The various combinations of average demand profiles and external demand profiles are provided in Table 2.7. In the assembly and general structure, the external demand is added level by level to some of the components. This is in line with reality in which some of the components may have external demand, and some may not. Based on these classes and numbers, the demand for each item in each period is randomly generated based on the normal distribution, the average demand profiles, and a 30% coefficient of variance. For each of these settings, 5 different random replications are generated.

- We consider 5 planning periods and 1 machine at each level. Without loss of generality, we assume that the lead time is equal to 0. The processing time (pt_{kt}) and the setup times (st_{kt}) are equal to 1 and 0, respectively. The BOM coefficient (r_{ki}) is equal to 1. Table 2.8 illustrates the value of the parameters, in the base case (set A) and for the sensitivity analysis.

In the following sections, we provide the numerical results for different product structures and investigate the value of flexibility in the multi-level lot sizing problem. For the experiments, we used the CPLEX 12.8.1.0 and Python libraries. We performed these experiments on a 2.4 GHz Intel Gold processor with only one thread on the Compute Canada computing grid.

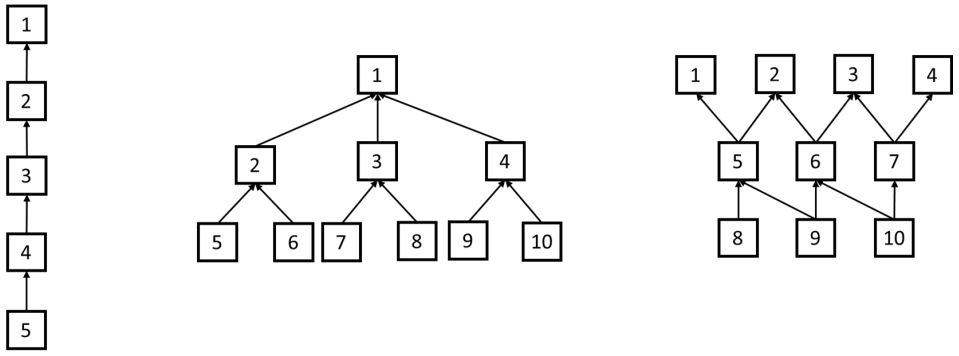


Figure 2.3: Different BOM structure (adapted from (Tempelmeier and Derstroff, 1996))

Table 2.7: Demand Profiles for different structures

		Average demand category		
		1	2	3
		Serial structure		
External demand	1	(100,0,0,0,0)	(100,0,0,0,0)	(500,0,0,0,0)
	2	(100,100,0,0,0)	(100,200,0,0,0)	(500,400,0,0,0)
	3	(100,100,100,0,0)	(100,200,300,0,0)	(500,400,300,0,0)
	4	(100,100,100,100,0)	(100,200,300,400,0)	(500,400,300,200,0)
	5	(100,100,100,100,100)	(100,200,300,400,500)	(500,400,300,200,100)
		Assembly structure		
External demand	1	(100,0,0,0,0,0,0,0,0,0)	(100,0,0,0,0,0,0,0,0,0)	(300,0,0,0,0,0,0,0,0,0)
	2	(100,100,0,0,0,0,0,0,0,0)	(100,200,0,0,0,0,0,0,0,0)	(300,200,0,0,0,0,0,0,0,0)
	3	(100,100,100,0,0,0,0,0,0,0)	(100,200,200,0,0,0,0,0,0,0)	(300,200,200,0,0,0,0,0,0,0)
	4	(100,100,0,0,100,0,0,0,0,0)	(100,200,0,0,300,0,0,0,0,0)	(300,200,0,0,100,0,0,0,0,0)
	5	(100,100,100,0,100,0,100,0,0,0)	(100,200,200,0,300,0,300,0,0,0)	(300,200,200,0,100,0,100,0,0,0)
		General structure		
External demand	1	(100,100,100,100,0,0,0,0,0,0)	(100,100,100,100,0,0,0,0,0,0)	(300,300,300,300,0,0,0,0,0,0)
	2	(100,100,100,100,100,0,0,0,0,0)	(100,100,100,100,200,0,0,0,0,0)	(300,300,300,300,200,0,0,0,0,0)
	3	(100,100,100,100,100,100,0,0,0,0)	(100,100,100,100,200,200,0,0,0,0)	(300,300,300,300,200,200,0,0,0,0)
	4	(100,100,100,100,100,0,0,100,0,0)	(100,100,100,100,200,0,0,300,0,0)	(300,300,300,300,200,0,0,100,0,0)
	5	(100,100,100,100,100,100,0,100,0,100)	(100,100,100,100,200,200,0,300,0,300)	(300,300,300,300,200,200,0,100,0,100)

2.5.2 SAA analysis

In this section, we perform the SAA analysis on the basic set A which has been defined in Table 2.8, with different numbers of scenarios. We determine a reasonable number for the rest of the experiments based on the SAA Gap, solution time and memory limits. Table 2.9 illustrates these experiments with $M = 10$, and $S^{eval} = 10000$ for different numbers of scenarios S in each replication. To calculate the gap and its standard deviation for each instance, and to determine a proper

Table 2.8: Parameter values for the base case and the sensitivity analysis

Parameter	Base case	Sensitivity analysis
Util	0.5	0.1 , 0.5 , 0.9
HValue	1	0, 1, 10
TBO profile	2	1, 2, 3, 4, 5
Service Level	95%	80%, 90%, 95% , 99%

number of scenarios in the SAA, we used the version with full flexibility. The time per replication in seconds (labeled as Time) is also reported for the version with full flexibility, which has the highest execution time compared to all versions with lower levels of flexibility. For the case with S= 1000 and full level of flexibility, for about 54% of the instances with an assembly structure and for all of the instances with a general structure, an optimal solution could NOT be found within the time limit of 7200 seconds or due to memory limitations. Considering the execution time, we will use 500 scenarios for the rest of the experiments to solve the model, and 10000 scenarios for evaluation. Among the three structures, the general structure has the highest execution time and the serial structure has the lowest one.

Table 2.9: SAA analysis

	Serial			Assembly			General		
# Scenario	Avg Gap (%)	Std Gap (%)	Time	Avg Gap (%)	Std Gap (%)	Time	Avg Gap (%)	Std Gap (%)	Time
100	0.17	0.010	8.2	0.32	0.008	54.5	0.21	0.005	129.6
250	0.14	0.006	39.9	0.25	0.005	390.8	0.17	0.003	800.5
500	0.14	0.004	219.8	0.14	0.005	1314.7	0.16	0.003	3802.0
1000	0.14	0.006	524.7	0.13	0.004	5544.7			

Table 2.10: SAA analysis, infeasibility percentage

	Serial		Assembly		General	
# Scenario	Infeasibility percentage	ε (%)	Infeasibility percentage	ε (%)	Infeasibility percentage	ε (%)
100	3.35	0.07	3.79	0.08	7.14	0.13
250	1.82	0.04	1.76	0.03	3.49	0.06
500	1.11	0.02	1.10	0.02	2.19	0.04
1000	0.37	0.01	0.59	0.01	1.14	0.02

As discussed in Section 2.4.1, it is possible to have violated service levels for some scenarios in the evaluation phase of the SAA method, except for the case with full flexibility. We analyse

the extent of these infeasibilities. In Table 2.10, the “Infeasibility percentage" shows the average percentage of scenarios for which a violation occurs out of the 10000 scenarios in the evaluation phase. The average value of service level violation is reported as ε. These two measures are calculated based on all levels of flexibility for each instance, except the full flexibility. We can see that these two measures are also acceptable for 500 scenarios.

2.5.3 The value of stochastic solution

In this section, we calculate the value of stochastic solution for the problem which is the cost difference between the cost of the optimal solution of the stochastic model and the deterministic model. To this end, the deterministic model (2.1) is solved in which the demand is equal to the expected demand. The solution of this model is then fixed as the first stage solution, and the cost of this model is calculated by optimally solving the second stage problem using the 10000 demand scenarios.

However, the current problem requires an additional consideration because of the service level. Although we can also consider the service level in the deterministic case, it is calculated based on the expected demand, and not several scenarios. This will result in a significant difference in the level of production between between the deterministic and stochastic case. This difference is more pronounced, when there is no flexibility in the model and all production decisions are defined in the first stage based on the expected value. In the deterministic case, based on the level of flexibility, the total evaluated cost (excluding the service level violation penalty) may be lower than the cost of the stochastic solution due to this lower level of production, but it also results in a high service level violation. Therefore, we do two separate analyses to show the value of the stochastic solution. First, for the case with full flexibility, we calculate the traditional VSS, calculated as the relative cost difference between the value of the deterministic solution and the stochastic solution. This VSS is equal to 10.6% for the serial structure, 11.9% for the assembly structure and 1.8% for the general structure. The second analysis focuses on the cases with a lower level of flexibility. For these cases, we focus on the service level violations in order to show the superiority of the stochastic model. Table 2.11 shows the service level violation of the mean value solution at different levels of flexibility for different structures. We conclude hence that the mean value deterministic approach cannot provide solutions that satisfy the service level.

Table 2.11: Service level violation, $\varepsilon(\%)$

LoF	Serial	Assembly	General
0	9.7	8.2	13.9
1	7.6	6.5	9
2	5.9	5.4	6.4
3	4.8	n/a	n/a
4	4.2	n/a	n/a

2.5.4 Serial structure

In this section, we investigate the effect of adding flexibility in the serial structure. This effect is measured using the ratio of the total cost decrease ($\Delta Cost$) when we have some level of flexibility (LoF) compared to the case when there is no flexibility and every decision is fixed at the beginning of the planning horizon. To solve the models we use 500 scenarios for the first stage, and the evaluation phase is performed using 10000 randomly generated scenarios.

Figure 2.4 shows the effect of adding flexibility for the base case, set A, as defined in Table 2.8. When referring to the BOM, the lowest level (starting from 0) refers to the level of the end item and the highest level refers to the incoming components. The horizontal axis is classified in two categories, the upper one for the level of flexibility which is from 0 to 5, and the lower one for the external demand profile from 1 to 5. In demand profile 1, only the end item has the external demand. In demand profile 2, we also have the external demand for the component at level 1 of the BOM. This is explained in detail in the previous section and Table 2.7. Figure 2.4 illustrates that increasing the flexibility will result in cost reduction in all the demand profile. Even if there is external demand only for the end items (External demand profile 1), it is still beneficial (with decreasing marginal benefits) to increase the flexibility at the component levels. If external demand exists also at the component level, the benefits (in terms of relative cost reduction) will increase compared to the case with only external demand for the end item. So both an increased level of flexibility and the presence of external demand at the component level lead to larger relative cost reductions.

In the next set of experiments, we perform sensitivity analyses, in which all the parameters of the base case A, except the parameter of interest, remain fixed. Figure 2.5 illustrates the sensitivity analysis on the effect of adding flexibility, considering changes in the TBO profile, service level,

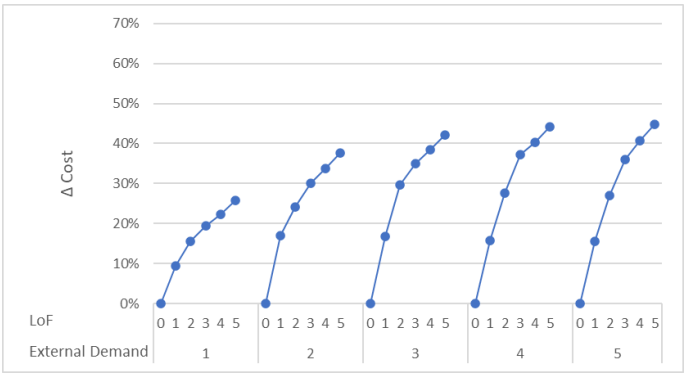


Figure 2.4: Effect of adding flexibility for serial structure

HValue, and *Util* parameters.

The smaller the *TBO*, the higher is the cost decrease by adding the flexibility. Low *TBO* means that you have many setup periods, and hence many opportunities to adjust the production (if flexible). As can be seen $\Delta Cost$ is the highest in *TBO* profile 1 compared to the other profiles. Comparing the *TBO* profiles 4 and 5, we can see that a high *TBO* at the lower level in the BOM (i.e., end items and lower level components) as in *TBO* profile 5 leads to lower benefits compared to a high *TBO* at the higher levels. This is logical as the production of the products at the lower level, determines the internal demand for the higher level components in the BOM and directly influences their production. *TBO* profile 4 has a higher rate of cost decrease as the *TBO* for the items at the lower level is less than the *TBO* of the items at the higher level.

HValue is related to the amount of added value to the components at different levels of the BOM. When it is equal to 0, it means there is no difference in the holding cost of different items at different levels. Adding some flexibility only provides a limited cost decrease. Indeed, because the inventory holding cost is the same for holding an end item or for holding all its components, the total holding cost cannot be improved by a better redistribution of the inventory at different levels in the BOM. In case of full flexibility, we observe a sudden jump in the cost reduction as the production of all items is now reactive to the demand realization and the total amount of inventory in the system is reduced. This will be further explained when we provide a more detailed analysis of this at the end of this section. On the other hand, when *HValue* is high, for example equal to 10, the inventory holding cost of the end item compared to other components is very high, and adding flexibility at higher levels where the inventory cost is much lower, will not result in a very high cost reduction. However, adding flexibility only for the end item reduces the costs by almost 30% since

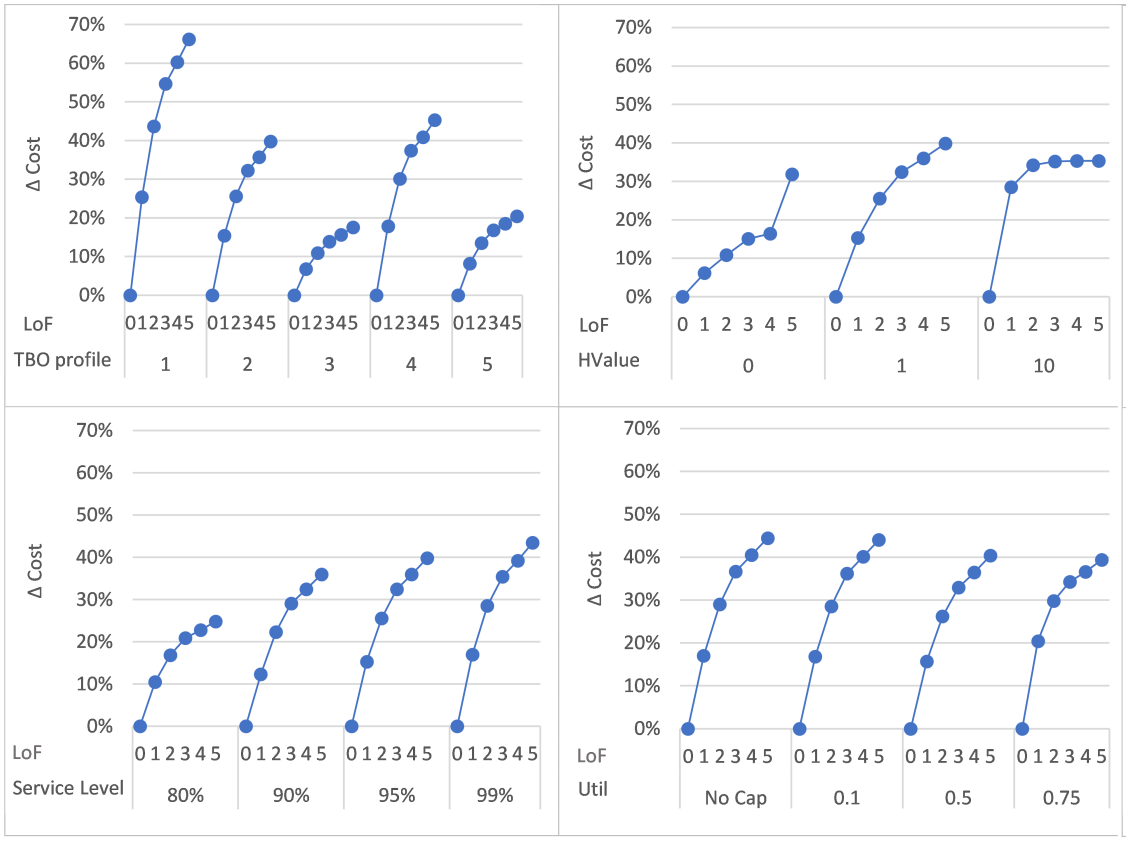


Figure 2.5: Sensitivity analysis for serial structure

this flexibility with respect to the production of the end item already allows some redistribution of the inventory between levels 0 and 1. If demand for the end items is low, we do not have to produce excess end items (with a high holding cost) and we can keep inventory at the component level (where holding costs are cheaper).

At the higher service level, the value of flexibility is bigger compared to the lower ones. Having a higher service level result in more production and inventory to mitigate the uncertainty. In this case, having more reactive inventory system will cause a higher cost reduction. It is also interesting that adding only one level of flexibility will result in higher cost reduction at higher service level compared to lower ones.

The last diagram presents the effect of capacity on $\Delta Cost$. When there is no capacity limitation or when the capacity is loose, we have a slightly higher cost reduction. This is logical as the capacity limitation and overtime cost will limit the production even if it is flexible. To show the scale of the total costs in different setting, Table 2.12 illustrates the sensitivity analysis for the total

cost. As can be seen, in the average column, by increasing the $HValue$, the total cost increases as this parameter has a direct effect on the holding cost. Regarding the TBO profiles, profile 1 which has a TBO of 1 for all the items, has the lowest cost, while TBO profile 3 which has the TBO equal to 4 for all the items has the highest cost as this value increases the setup cost. Comparing the total cost of profile 4 and 5 shows that, when the item at the lower level of BOM has a higher TBO, the total cost is higher compared to the case when the higher TBO is at the higher level. Increasing the service level results in a total cost increase as the total production increases in the system. Increase in $Util$ means tighter capacity which imposes more cost to the production system. All the mentioned trends are valid not only for the average cost over all levels of flexibility, but also at each level of flexibility individually.

As can be seen in Figure 2.5, the trend of $\Delta Cost$ for different options of TBO profiles, Service levels, and $Util$ parameter are relatively similar. However, for the parameter $HValue$, the trends are also different for different options. Figure 2.6 presents a more detailed analysis for this parameter. In addition to the level of flexibility and $HValue$, the horizontal axis is also categorized based on the external demand profile which is shown in the second level from 1 to 5. When $HValue$ is equal to 0, and the external demand profile is equal to 1, i.e., we have only external demand for the end item at level 0, adding flexibility does not decrease the cost unless we have the full flexibility at all levels. The reason behind it is that there is no advantage in keeping inventory at a higher level of BOM, as the unit holding cost is not different. However, when we have full flexibility, the total amount of inventory decreases in the system and the total cost will decrease. When $HValue$ is equal to zero and when we have independent demand for the components as well (demand profile 2 to 5), depending on the level of these components, we have a cost decrease. The reason is that when we have proper level of flexibility, the amount of production is reactive to the demand realization, and it result in more efficient production and a cost decrease. At the levels where there is no independent demand, we see that the cost remains unchanged, until having the full flexibility. Having full flexibility, the total production is reactive to the demand realization and the the total amount of inventory is reduced in the system. When $HValue$ is equal to 10, we have a very high added value BOM. As can be seen, the marginal cost decrease is most significant when adding the first level of flexibility, compared to adding the further levels of flexibility. This is due to the fact that the cost of inventory for the item which is stored at the lowest level of the BOM is much lower

Table 2.12: Sensitivity analyses of the total cost for the serial structure

	Level of flexibility						
Parameter	0	1	2	3	4	5	Average
External demand							
1	89633	79940	73356	69428	66876	63617	73808
2	81441	67467	61108	56013	53048	49969	61508
3	91223	74407	62840	58016	55006	51873	65561
4	94821	79508	68198	59693	56895	53371	68748
5	94753	78750	67842	59697	55732	52213	68165
TBO profile							
1	48842	35491	26368	21035	18400	15657	27632
2	89267	74842	65458	59444	56495	53021	66421
3	227301	211502	201694	195318	191472	187463	202458
4	73859	60001	50601	45025	42305	39010	51800
5	192183	176345	165968	159916	157022	153779	167536
HValue							
0	12835	12045	11382	10767	10550	8484	11010
1	90427	75734	66191	59937	56842	53537	67111
10	35885935	26206487	24374210	24079449	24037381	24020933	26434066
Service Level							
80%	61672	54832	50850	48413	47256	46067	51515
90%	76737	66782	58787	53618	51149	48542	59269
95%	90277	75801	66172	59907	56812	53495	67078
99%	109196	89797	77158	69896	66069	61492	78935
Util							
No cap	81026	66436	56428	50199	47140	44005	57539
0.1	80797	66558	56626	50373	47299	44168	57637
0.5	90943	76176	66217	60010	56880	53469	67282
0.75	122711	95964	84240	78668	75814	72549	88324

compared to the same amount of inventory stored for an item at a higher level. When the $HValue$ is equal to 1, which is between the two extreme cases of $HValue = 0$ and $HValue = 10$, we can see a smooth cost reduction by adding multiple levels of flexibility.

2.5.5 Assembly structure

In this section, we investigate the value of flexibility for the assembly structure. Similar to the serial structure, we will compare the possible cost decrease ($\Delta Cost$) when we consider a more adaptive strategy. For this structure, flexibility will be added to the BOM level by level. Having 3 levels in the BOM for the assembly structure, we define 4 options for flexibility, from no flexibility to full

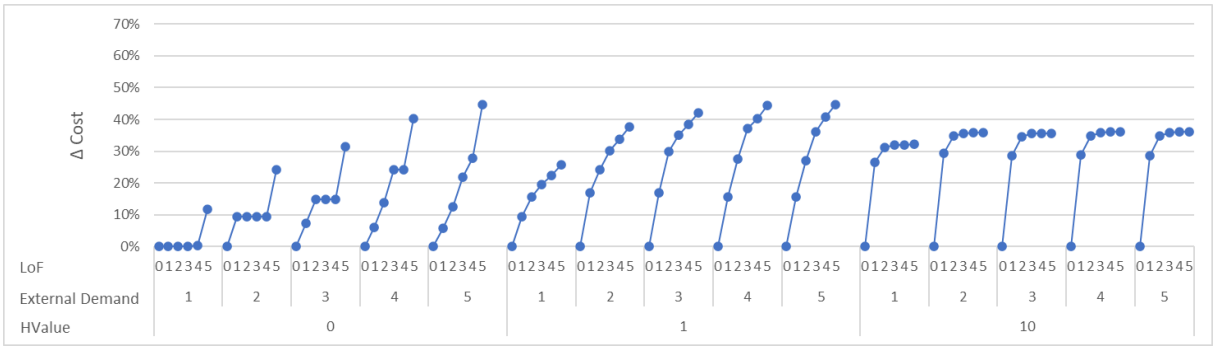
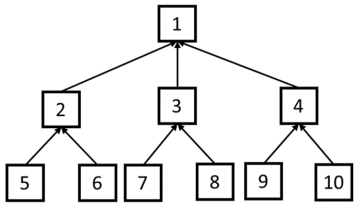


Figure 2.6: Effect of adding flexibility for serial structure

Table 2.13: Levels of flexibility for assembly structure

# Flexibility	# product									
	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1
Level	0	1	1	1	2	2	2	2	2	2



flexibility. Assuming 0 as no flexibility for an item, and 1 for flexibility, Table 2.13 illustrates these levels of flexibility considered for this structure. The first option (0) has no flexibility, the second option (1) has flexibility for the end item only. The last option (3) has flexibility for all the items.

Figure 2.7 illustrates the cost decrease percentage ($\Delta Cost$) when increasing the flexibility level by level for different external demand profile (see Table 2.7). For all cases, adding flexibility only for the end item results in the highest rate of cost decrease, compared to the case of adding flexibility to the component levels. Comparing Figure 2.7 and Figure 2.4, we can see similar trends in the assembly structure compared to what has been observed for the serial structure.

Figure 2.8 illustrates the sensitivity analysis of $\Delta Cost$ for the assembly structure based on the TBO, $HValue$, service level and $Util$ parameters considering full flexibility options at different levels of BOM.

When the TBO is equal to 1 for all the items (TBO profile 1), we have the highest percentage of cost decrease, and when TBO is at its highest value for all the items (TBO profile 3), we have the lowest rate for the cost decrease by adding flexibility. Between these two extreme cases, having lower values of TBO for the items at the lower levels of BOM (TBO profile 4), results in higher rate of cost decrease compared to the case when we have higher value of TBO at these levels (TBO

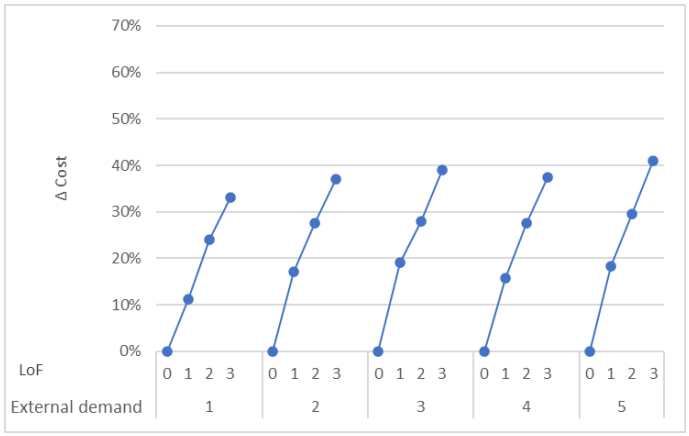


Figure 2.7: Analysis of adding flexibility per level for the base case in assembly structure

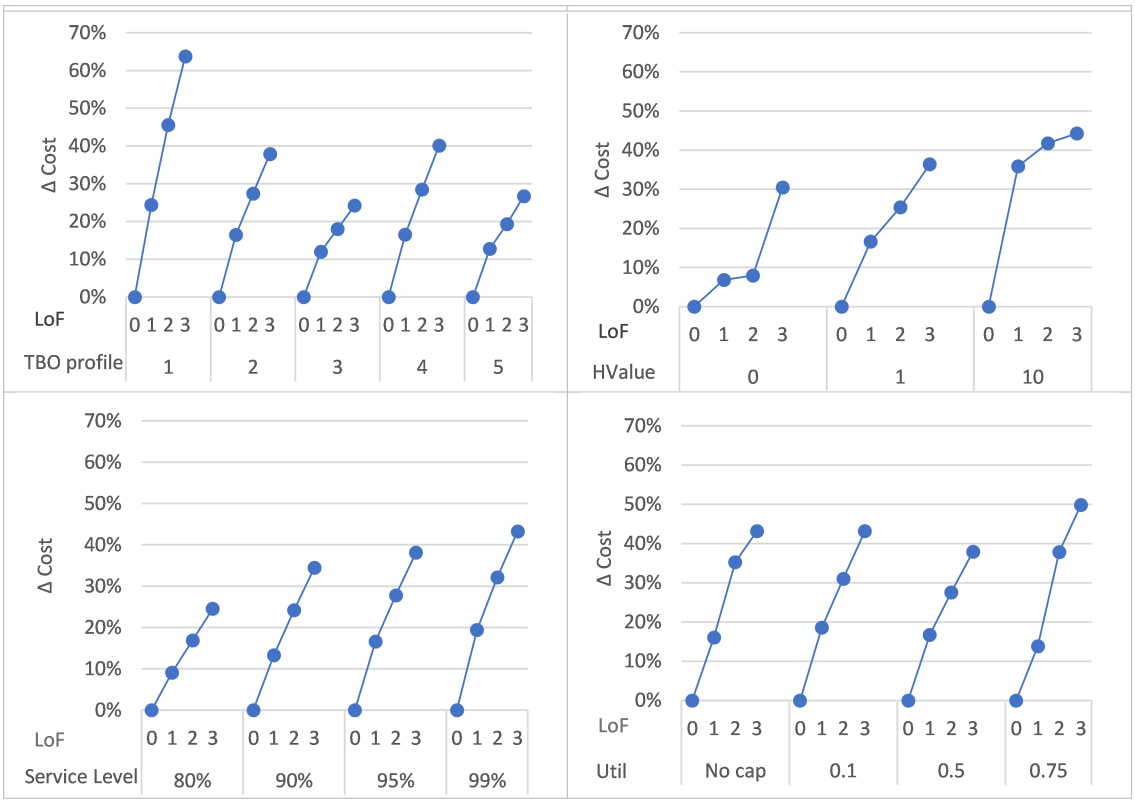


Figure 2.8: Sensitivity analysis for assembly structure

profile 5). Considering the *HValue* parameter, we can see that different patterns for the added value in the product structure, result in different pattern of cost decrease, when we add flexibility. When the *HValue* is equal to 0, there is no advantage to keep the components at the higher level of BOM to save the holding cost. In this case, when we add the flexibility at the highest level of

BOM, we see a significant increase in the rate of decrease, as the total amount of inventory in the system will decrease. We hence observe a similar effect as in the serial case.

At different service levels, we have the same pattern for $\Delta Cost$ but there is a higher cost decrease at the higher service level when we add the flexibility. While at 80% of service level we have about 25% of cost decrease at the full flexibility, the same value is about 45%, when the service level is equal to 99%. Considering different values for the $Util$ parameter, we see a similar trend for all cases. When we have a very tight capacity ($Util = 0.75$), we have a slightly higher rate of cost decrease.

Table 2.14 illustrates the sensitivity analysis for the total costs. The patterns we can see in assembly structure are similar to ones of the serial structure. In summary, higher $HValue$, higher TBO, and higher service levels result in higher cost. In addition, higher $Util$ which means tighter capacity imposes extra cost to the system.

2.5.6 General structure

In this section, we study the general structure and adding different levels of flexibility to different items in this structure. Table 2.15 illustrates different levels of flexibility for this structure. The levels of flexibility start from no flexibility to full flexibility. In this section, we only discuss the full flexibility per level.

Figure 2.9 illustrates the cost decrease percentage ($\Delta Cost$) with respect to the level of flexibility for different external demand profiles. Having more items with external demand generally results in a slightly higher cost decrease. For all cases, adding flexibility only for the end items (flexibility level 1) results in a rate of cost decrease of about 25%. In this structure, we have higher cost decrease compared to the assembly structure, and in general lower variability in different trends per external demand profiles, where there are more items with external demand, in the system.

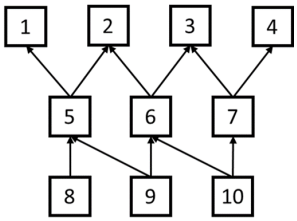
Figure 2.10 shows the sensitivity analysis for the general structure based on different parameters. The sensitivity analysis for the TBO shows similar patterns of cost decrease when adding flexibility compared to the serial and assembly structures. A lower TBO results in a higher cost decrease when adding flexibility. At higher service levels, it is more beneficial to add the flexibility. We can see that the diagrams for different service levels have similar patterns, but $\Delta Cost$ increases

Table 2.14: Sensitivity analyses of the total cost for the assembly structure

	LoF				
Parameter	0	3	5	8	Average
TBO profile					
1	41734	31217	22266	14661	27469.5
2	79300	65849	56977	48566	62673
3	201413	177611	165053	152651	174182
4	73000	60797	51697	43066	57140
5	148742	130453	120892	110171	127564.5
Hvalue					
0	28858	26828	26494	19796	25494
1	78454	64894	57787	49030	62541.25
10	1539484	1002998	913605	869498	1081396
Service Level					
80%	55965	51072	46423	42162	48905.5
90%	68036	58752	51168	44083	55509.75
95%	80408	66805	57606	49191	63502.5
99%	97996	77919	65430	54406	73937.75
Util					
No cap	64093	53203	40662	35488	48361.5
0.1	72615	58412	49203	40304	55133.5
0.5	80817	66925	57873	49378	63748.25
0.75	131791	111196	79101	62467	96138.75

Table 2.15: Levels of flexibility for general structure

# Flexibility	# product									
	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0	0
2	1	1	1	1	1	1	1	0	0	0
3	1	1	1	1	1	1	1	1	1	1
Level	0	0	0	0	1	1	1	2	2	2



as we increase the service levels. When the capacity is very tight, where the model should use a significant amount of overtime, adding flexibility results in a higher cost reduction.

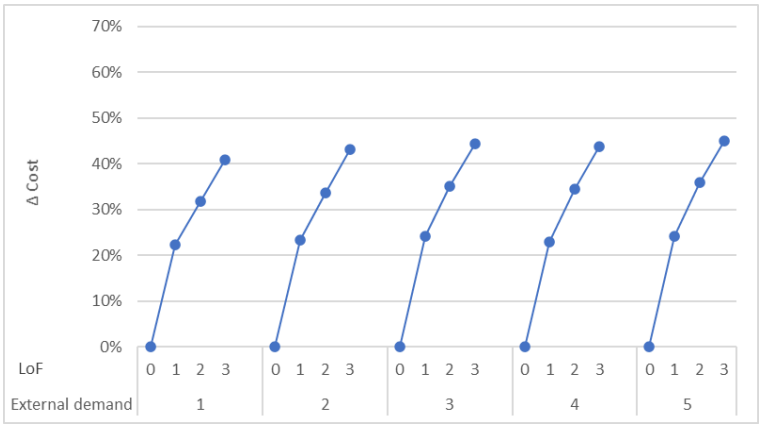


Figure 2.9: Analysis of adding flexibility per level for the base case in general structure

Table 2.16: Sensitivity analyses of the total cost for the general structure

	LoF				
Parameter	0	3	5	8	Average
TBO profile					
1	60103	37252	26928	18910	35798
2	103831	79093	67675	58054	77163
3	227132	196767	184271	172151	195080
4	82730	58683	48221	39423	57264
5	188368	160128	148268	138152	158729
Hvalue					
0	32983	28602	26563	19518	26917
1	104371	79535	68062	58196	77541
10	2494477	1497842	1386434	1364212	1685741
Service Level					
80%	64460	51680	46521	41379	51010
90%	84653	65693	57226	49164	64184
95%	104102	79415	67851	58239	77402
99%	124723	90922	77581	67258	90121
Util					
No cap	96557	72533	61467	52223	70695
10%	98110	73387	62160	52826	71621
50%	104147	79254	67829	58105	77334
75%	147411	88238	74653	64337	93660

In the previous experiments, the flexibility was added to the system level by level. Having multiple items per level in the assembly and general structures, we may add flexibility to some of

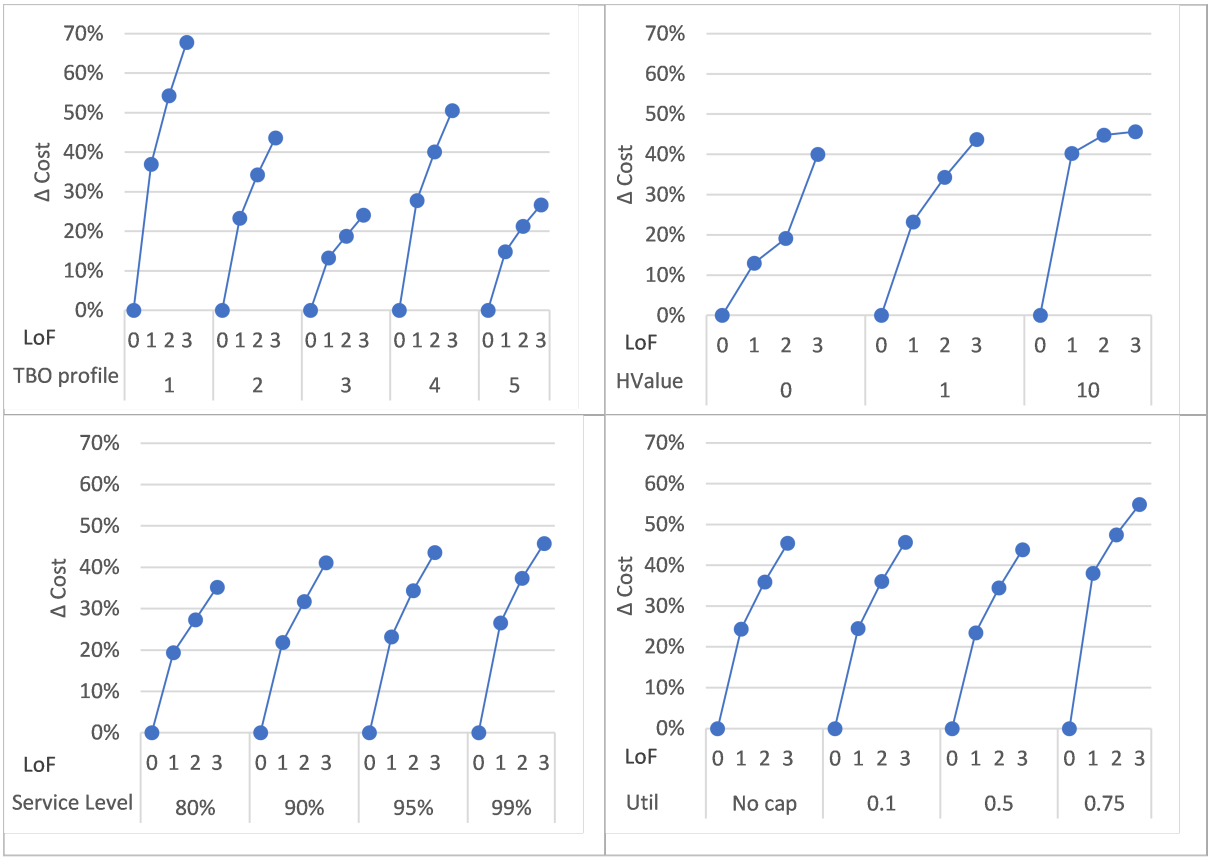


Figure 2.10: Sensitivity analysis for general structure

the items in each level. We investigate the value of partial flexibility with extra experiments for the base case which presented in Appendix D.

2.5.7 Insights

In this section we will present some insight based on our findings in the numerical experiments. First and in general, adding flexibility reduces the cost in the system. This cost reduction depends on the level where the flexibility happens and the external demand as well. In all structures, adding flexibility at the lower level, i.e., for the end items leads to more cost saving.

Second, different parameters in the problem affect the cost savings as well. The ratio between the setup cost and the inventory holding cost plays an important role in the cost savings obtained by adding flexibility. When the time between orders is low, for example one, based on the trade-off between the ordering costs and the holding costs, it is less costly that the production covers a

smaller number of periods and there are hence more frequent setups. The value of adding production recourse is higher in this case, as you can adjust the production levels in each period. As the ratio between setup cost and holding cost may be different for different items, we should note that the end items, and the items at the lower levels of BOM have a higher impact in this matter.

Third, having flexibility results in holding cost reduction and there are two reasons behind that. First, production flexibility generally reduces the amount of inventory in the system, as the production are more responsive to the demand and there is less need for safety stock. Second, having flexibility will increase the option of where we can keep our inventory. More specifically, considering the added value in the BOM, keeping stock at higher level of BOM, where we have lower holding cost, and use them when needed, will reduce the total holding cost in the system.

2.6 Conclusion

In this research, we study the stochastic multi-level lot sizing problem with service level and investigate the benefits of production flexibility in the BOM based on static and adaptive strategies. The problems are modeled as two-stage stochastic models which are approximated using a finite number of scenarios and solved by the SAA method. Extensive numerical experiments and simulations are conducted for different BOM structures and under different parameter settings. The results show that increasing the production flexibility leads to a cost reduction, even in the case where there is no external demand for any of the components, in all BOM structures. Sensitivity analyses have been performed to demonstrate the effect of changing different parameters on the cost reduction by adding flexibility and allowing more adaptive decisions. In general, the value of adding flexibility is more significant at higher service levels compared to lower ones and with lower TBO compared to higher ones. The value added in the BOM structure also has effect on the pattern of cost decrease in different structures. In the adaptive version of problem, we consider the static-dynamic strategy for some or all items and we modeled it as a two-stage stochastic model. Considering the dynamic strategy for these items is an interesting future research direction which needs a multi-stage stochastic model. This will make the problem more challenging to solve but it may result in more responsive plans.

References

Afentakis, P. and Gavish, B. (1986). Optimal lot-sizing algorithms for complex product structures. *Operations Research*, 34(2):237–249.

Afentakis, P., Gavish, B., and Karmarkar, U. (1984). Computationally efficient optimal solutions to the lot-sizing problem in multistage assembly systems. *Management Science*, 30(2):222–239.

Akartunalı, K. and Miller, Andrew J., A. (2009). heuristic approach for big bucket multi-level production planning problems. *European Journal of Operational Research*, 193(2):396–411.

Almeder, C. (2010). A hybrid optimization approach for multi-level capacitated lot-sizing problems. *European Journal of Operational Research*, 200(2):599–606.

Alvarez, A., Cordeau, J.-F., Jans, R., Munari, P., and Morabito, R. (2020). Inventory routing under stochastic supply and demand. *Omega*, pages 102–304.

Bookbinder, J. H. and Tan, J.-Y. (1988). Strategies for the probabilistic lot-sizing problem with service-level constraints. *Management Science*, 34(9):1096–1108.

Contreras, I., Cordeau, J.-F., and Laporte, G. (2011). Stochastic uncapacitated hub location. *European Journal of Operational Research*, 212(3):518–528.

Gruson, M., Cordeau, J.-F., and Jans, R. (2018). The impact of service level constraints in deterministic lot sizing with backlogging. *Omega*, 79:91–103.

Gruson, M., Cordeau, J.-F., and Jans, R. (2021). Benders decomposition for a stochastic three-level lot sizing and replenishment problem with a distribution structure. *European Journal of Operational Research*, 291(1):206–217.

Helber, S., Sahling, F., and Schimmelpfeng, K. (2013). Dynamic capacitated lot sizing with random demand and dynamic safety stocks. *OR Spectrum*, 35(1):75–105.

Hung, Y.-F. and Chien, Kuo-Liang, A. (2000). multi-class multi-level capacitated lot sizing model. *Journal of the Operational Research Society*, 51(11):1309–1318.

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

Mousavi, K., Bodur, M., and Roorda, M. (2021). Stochastic last-mile delivery with crowd-shipping and mobile depots.

Pochet, Y. and Wolsey, L. A. (2006). *Production Planning by Mixed Integer Programming*. Springer, Science & Business Media, New York.

Quezada, F., Gicquel, C., Kedad-Sidhoum, S., and Vu, Dong Quan, A. (2020). multi-stage stochastic integer programming approach for a multi-echelon lot-sizing problem with returns and lost sales. *Computers & Operations Research*, 116(10486):5.

Sahling, F., Buschkühl, L., Tempelmeier, H., and Helber, S. (2009). Solving a multi-level capacitated lot sizing problem with multi-period setup carry-over via a fix-and-optimize heuristic. *Computers & Operations Research*, 36(9):2546–2553.

Seeanner, F., Almada-Lobo, B., and Meyr, H. (2013). Combining the principles of variable neighborhood decomposition search and the fix&optimize heuristic to solve multi-level lot-sizing and scheduling problems. *Computers & Operations Research*, 40(1):303–317.

Sereshti, N., Adulyasak, Y., and Jans, R. (2020). The value of aggregated service levels in stochastic lot sizing problems. *Omega*, 102335.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

Stadtler, H. (2003). Multilevel lot sizing with setup times and multiple constrained resources: Internally rolling schedules with lot-sizing windows. *Operations Research*, 51(3):487–502.

Taş, D., Gendreau, M., Jabali, O., and Jans, Raf, A. (2019). capacitated lot sizing problem with stochastic setup times and overtime. *European Journal of Operational Research*, 273(1):146–159.

Tavaghof-Gigloo, D. and Minner, S. (2021). Planning approaches for stochastic capacitated lot-sizing with service level constraints. *International Journal of Production Research*, 59(17):5087–5107.

Tempelmeier, H. (2011). A column generation heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. *Omega*, 39(6):627–633.

Tempelmeier, H. and Derstroff, Matthias, A. (1996). Lagrangean-based heuristic for dynamic multilevel multiitem constrained lotsizing with setup times. *Management Science*, 42(5):738–757.

Tempelmeier, H. and Herpers, S. (2011). Dynamic uncapacitated lot sizing with random demand under a fillrate constraint. *European Journal of Operational Research*, 212(3):497–507.

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

Thevenin, S., Adulyasak, Y., and Cordeau, J.-F. (2021). Material requirements planning under demand uncertainty using stochastic optimization. *Production and Operations Management*, 30(2):475–493.

Toledo, C. F. M. and da Silva Arantes (2015). Márcio and hossomi, marcelo yukio bressan and França, paulo morelato and akartunalı, kerem, a relax-and-fix with fix-and-optimize heuristic applied to multi-level lot-sizing problems. *Journal of Heuristics*, 21(5):687–717.

Tunc, H., Kilic, O. A., Tarim, S. A., and Eksioglu, Burak, A. (2014). reformulation for the stochastic lot sizing problem with service-level constraints. *Operations Research Letters*, 42(2):161–165.

Verweij, B., Ahmed, S., Kleywegt, A. J., Nemhauser, G., and Shapiro, A. (2003). The sample average approximation method applied to stochastic routing problems: a computational study. *Computational optimization and applications*, 24(2):289–333.

Wagner, S. M. and Lindemann, Eckhard, A. (2008). case study-based analysis of spare parts management in the engineering industry. *Production Planning & Control*, 19(4):397–407.

Wu, T., Shi, L., Geunes, J., and Akartunalı, K. (2011). An optimization framework for solving capacitated multi-level lot-sizing problems with backlogging. *European Journal of Operational Research*, 214(2):428–441.

Xiao, Y., Zhang, R., Zhao, Q., Kaku, I., and Xu, Yuchun, A. (2014). variable neighborhood search with an effective local search for uncapacitated multilevel lot-sizing problems. *European Journal of Operational Research*, 235(1):102–114.

You, M., Xiao, Y., Zhang, S., Zhou, S., Yang, P., and Pan, X. (2019). Modeling the capacitated multi-level lot-sizing problem under time-varying environments and a fix-and-optimize solution approach. *Entropy*, 21(4):377.

Chapter 3

Stochastic dynamic lot sizing with substitution and service level constraints

This study is done as part of an internship at the University of Toronto, funded by FRQNT. The result of this work will be submitted as a research article and the coauthors are Merve Bodur and James Luedtke.

Abstract

We consider a multi-stage stochastic lot-sizing problem with service level constraints and product substitution. A firm has multiple products and it has the option to meet demand from substitutable products at a cost. Considering the uncertainty in future demands and the production leadtime, the firm wishes to make ordering decisions in every period such that the probability that all demands can be met in the next period is at least equal to a minimum service level. We propose a rolling-horizon policy in which a two-stage joint chance-constrained stochastic program is solved to make decisions in each time period. We demonstrate how to effectively solve this formulation. In addition, we propose two policies based on deterministic approximations and demonstrate that the proposed chance-constraint policy can achieve the service levels more reliably and at a lower cost. We also explore the value of product substitution in this model, demonstrating that the substitution option allows achieving service levels at significantly reduced costs, about 7% to 25% in our experiments and that the majority of the benefit can be obtained with limited levels of substitution

allowed.

3.1 Introduction

The basic lot sizing problem is a multi-period production planning problem that considers the trade-off between setup costs and inventory holding costs and defines the optimal timing and quantity of production to minimize the total cost. In situations where there exists uncertainty in demand, which is inevitable in real-world applications, the decision-maker needs to determine the production policy to minimize the expected cost. Here, it is inevitable to have stock outs and the challenge is to keep them under control. Common approaches to deal with this challenge are to consider the backorder cost which includes both tangible and intangible effects which are difficult to estimate or to impose a service level criterion. In this research, we study the stochastic lot sizing problem with an α service level which is an event-oriented service level and impose limits on the probability of stock outs. This service level which is frequently used in a variety of applications is usually defined as a chance constraint.

When an item is out of stock, sometimes the firm has the option to substitute it with another product. This type of substitution which is initiated by the supplier is called supplier-driven substitution and may result in reducing the stock outs, increasing revenue, cost savings, and customer satisfaction, especially when dealing with demand uncertainty. This problem has practical relevance in the electronics and steel industries where it is possible to substitute a lower-grade product with a higher-grade one. Semiconductors or microchips are good examples of these types of products (Lang and Domschke, 2010).

In this research, we consider the stochastic lot sizing problem with substitution and joint service level constraint over multiple products. The possibility of product substitution results in risk-pooling associated with uncertain demand (Shin et al., 2015). The substitution option and joint service level are hence in line with each other as in both cases, we consider the risk of stockouts jointly over all products. In other words, having the substitution option, it is not appropriate to impose the service levels individually for each product.

We consider an infinite-horizon problem in which we need to sequentially make decisions on setup timings, and production and substitution amounts based on the current state of the system

reflected as the amount of available inventory and backlog. We follow the “dynamic" strategy (Bookbinder and Tan, 1988) for which different decisions can be dynamically updated throughout the planning horizon when the demands are observed. As determining the optimal solution is computationally intractable, to solve this problem, we consider a finite-horizon problem and apply it in a rolling-horizon environment. The aim is to propose decision policies which map the state of the system to different decisions. These policies are based on mixed integer programming model for the problem, in which the random demand is represented as a scenario set.

The challenge of these stochastic models is that with increasing the number of scenarios the solution time will increase extensively and this makes it difficult to reach a reasonable solution in a reasonable amount of time. To deal with this challenge we choose two different approaches. First, we propose policies based on the deterministic approximation of the model, without directly imposing a service level constraint. Second, we propose a two-stage approximation for the multi-stage model and an efficient branch-and-cut (B&C) algorithm to solve the model with service level constraint.

We simulate the rolling-horizon framework and demonstrate that the proposed chance-constraint policy has the advantage of respecting the service levels more reliably and at a lower cost, over the deterministic policies. We illustrate and analyze this superiority under different settings. We also explore the value of product substitution in this model, demonstrating that the substitution option allows achieving service levels at significantly reduced costs and that the majority of the benefits can be obtained with limited levels of substitution allowed.

The contribution of this research can be summarized as follows.

- Consider an infinite-horizon multi-stage lot sizing problem with substitution and joint service level constraints, which to the best of our knowledge is new to the literature.

- Propose a finite-horizon stochastic dynamic program for this problem and derive approximation-based rolling-horizon policies to define different decisions at each point of time. The approximations are based on the deterministic models and the two-stage stochastic programming, using sample average approximation.

- Apply a B&C algorithm to solve the model which explicitly considers the service level.

- Compare different policies including deterministic ones and a chance-constraint policy, using simulation, illustrate the value of substitution and provide other managerial insights under different settings.

The rest of this paper is organized as follows. In section 2, we survey the related literature. In section 3, we define the problem and the dynamics of decisions in the system. We also provide the dynamic programming formulation for the finite-horizon problem. In section 4, we explain the process of making different decisions at each stage, using different decision policies. In section 5, we present the B&C algorithm to solve the chance-constraint policy model, in which we explicitly consider the service level. In section 6, we illustrate the computational experiments, including the rolling-horizon implementation and simulation procedure, policy comparison, and insights. Finally, we conclude in section 7.

3.2 Literature review

The related literature of this work can be categorized in two streams. The first part is dedicated to the lot sizing and inventory models with substitution in both deterministic and stochastic versions and the second part is dedicated to the stochastic lot sizing problem with joint service level. To the best of our knowledge, no research has investigated stochastic lot sizing problem with substitution and joint service levels.

3.2.1 Lot sizing and inventory problems with substitution

In the literature, there are two types of substitution, the customer-driven substitution and supplier-driven substitution (Shin et al., 2015). In the customer-driven substitution, the customer decides which product to substitute (Zeppetella et al., 2017), while in the supplier-driven (firm-driven) case, it is the supplier, firm, or the vendor who makes the substitution decisions (Rao et al., 2004). The substitution possibility in inventory decisions is addressed in both deterministic and stochastic settings which are explained as follows.

Deterministic models

Hsu et al. (2005) study two different versions of the dynamic uncapacitated lot sizing problem with substitution, when there is a need for physical conversion before substitution, and when it does not require any conversion. The authors propose a mathematical model for this problem and solve it using a backward dynamic programming algorithm and a heuristic algorithm based on Silver-Meal heuristic to solve the problem. Lang and Domschke (2010) consider the uncapacitated lot sizing problem with general substitution in which a specific class of demand can be satisfied by different products based on a substitution graph. They model the problem as a mixed-integer linear program and propose a plant location reformulation in which the amount of production for an item is broken down into different amounts based on the period where they are used to satisfy the demand. The authors also propose some valid inequalities for the original model and solve the model using the CPLEX solver.

Stochastic models

Many studies in the field of the stochastic inventory planning have considered the possibility of substitution. While the majority of them investigated the customer-driven substitution, some research considered the supplier-driven substitutions. In the customer-driven substitution, the customer may choose another product, if the original item cannot be found. This is also known as “stock out substitution". Akçay et al. (2020) investigate a single-period inventory planning problem with substitutable products. Considering the stock out substitution, they propose an optimization based method, which jointly defines the stocking of each product, while satisfying a service level. Nagarajan and Rajagopalan (2008) consider the inventory problem with customer-driven substitution, and propose an optimal policy and heuristic algorithm for different versions of the problems in terms of planning periods and number of products.

In this research, we consider the supplier-driven substitution. In the same stream of research, Bassok et al. (1999) investigate the single-period inventory problem with random demand and downward substitution in which a lower-grade item can be substituted with the ones with a higher-grade. This model is an extension of the newsvendor problem and there is no setup cost in case of ordering. The sequence of decisions is as follows: first, they define the order quantity for each of the items, namely ordering decision. Second, when the demand is observed, they define the allocation decisions. The authors propose a profit maximization formulation and characterize the structure of the optimal policy for this problem. Using some decomposition ideas they propose

bounds on the optimal order amount and use them in an iterative algorithm to solve the model. Rao et al. (2004) also consider a single-period problem with stochastic demand and downward substitution, and model it as a two-stage stochastic program. In their model, they consider the initial inventory and the ordering cost as well. In addition to the extensive form of the model, the author propose two heuristic algorithms to solve this problem.

Another similar research stream considers the possibility of having multiple graded output items from a single input item, which is known as “co-production" (Ng et al., 2012). In these problems, there is a hierarchy in the grade of output items and it is possible to substitute a lower-grade item with the ones with higher-grade (Bitran and Dasu, 1992). (Hsu and Bassok, 1999) consider the single-period production system with random demand and random yields. They model the problem as a two-stage stochastic program which defines the production amount of a single item and the allocation of its different output items to different demand classes (Hsu and Bassok, 1999). They propose three different solution methods to solve the problem, including a stochastic linear model. In addition, two decomposition based methods in which the subproblems are network flow problems, are proposed for this problem. Bitran and Dasu (1992) study an infinite-horizon, multi-item, multi-period co-production problem with deterministic demand and random yields. As solving this problem in a infinite-horizon is intractable, they proposed two approximation algorithms to solve it. The first approximation is based on a rolling-horizon implementation of the finite-horizon stochastic model. For the second approximation, they consider a simple heuristic based on the optimal allocation policy, in a multi-period setting. This heuristic includes two modules; a module to determine the production quantities, and module to allocate produced items to the customers. This heuristic can be also applied in a rolling-horizon procedure. Bitran and Leong (1992) consider the same problem and propose deterministic near-optimal approximations within a fixed planning horizon. To adapt their model to the revealed information, they apply the proposed model using simple heuristics in a rolling planning horizon. Bitran and Gilbert (1994) consider the co-production and random yield in a semiconductor industry and propose heuristic methods to solve it.

3.2.2 Stochastic lot sizing problem and service level constraints

This section is dedicated to the stochastic lot sizing problem with random demand. Most of the research in this context consider a scenario set or a scenario tree to represent the stochasticity in demand and propose efficient methodologies to solve them. Haugen et al. (2001) consider the multi-stage uncapacitated lot sizing problem and propose a progressive hedging algorithm to solve it. Guan and Miller (2008) propose a dynamic programming algorithm for a similar version. Using the same algorithm, Guan (2011) study the capacitated version of the problem with the possibility of backlogging. Lulli and Sen (2004) propose a branch-and-price algorithm for multi-stage stochastic integer programming and apply their general method to the stochastic batch-sizing problem. In this problem, they consider that the demand, production, inventory and set up costs are uncertain. The difference between this problem and the lot sizing problem is that the production quantities are in batches and the production decisions are the number of batches that will be produced, as such integer-valued. This problem is a more general case of lot sizing problem. In another research, Lulli and Sen (2004) proposed a scenario updating method for the stochastic batch-sizing problem.

A common approach to deal with stochastic demand is using service levels. In this context the planners put a demand fulfillment criterion to mitigate the risk of stock outs. Stochastic lot sizing problems with service level constraints have been studied extensively (Tempelmeier, 2007) and many types of service levels exist in the literature. One of the main service levels is the α service level which is an event-oriented service level, and imposes limits on the probability of a stock out. This service level is represented as a chance-constraint and is usually defined for each period and product separately. Bookbinder and Tan (1988) investigate stochastic lot sizing problems with an α service level and propose three different strategies for this problem based on the timing of the setup and production decisions. These strategies are the *static*, *dynamic*, and *static-dynamic* strategies. In the *static* strategy, both the setup and production decisions are determined at the beginning of the planning horizon and they remain fixed when the demand is realized. In the *dynamic* strategy, both the setup and production decisions are dynamically changed with the demand realization throughout the planning horizon. The *static-dynamic* strategy is between these two strategies in which the setups are fixed at the beginning of the planning horizon and the production decisions

are updated when the demands are realized.

There are some studies which define the service level constraint jointly over different planning periods. Liu and Küçükyavuz (2018) consider the uncapacitated lot sizing problem with a joint service level constraint. They study the polyhedral structure of the problem and propose different valid inequalities and a reformulation for this problem. Jiang et al. (2017) consider the same problem with and without pricing decisions. Gicquel and Cheng (2018) investigate the capacitated version of the same problem. Jiang et al. (2017) and Gicquel and Cheng (2018) use a sample approximation method to solve their problems. This method is a variation of the sample average approximation method which is proposed by Luedtke and Ahmed (2008) to solve models with chance-constraints using scenario sets. All the mentioned studies consider single item models in which the joint service level is defined over all periods. There are few Studies that consider the service level jointly over all the products. Akçay et al. (2020) adapt the Type II service level or "fill rate" for each individual product and overall within a category of products, having customer-driven substitution assumption. This type of service level consider the expected value of backorder and it is not modeled as a chance constraints. Sereshti et al. study different type of aggregate service level for the lot sizing problem which are defined over multiple products, but they didn't consider substitution in their models. In this research, we consider supplier-driven substitution and a joint service level that is defined over all products.

Table 3.1 summarises the characteristics of the studied paper and illustrates their similarities and the differences to our work.

3.3 Problem definition and formulation

We consider a stochastic lot sizing problem with the possibility of substitution in an infinite time horizon which is discretized into planning periods. There are multiple types of products with random demand, and at each stage, we need to make decisions about the production setups, production and substitution amounts, and accordingly define the potential inventory and backlog levels. There is a production lead time of one, i.e., what is produced in the current stage is available in the next stages. These decisions are made sequentially at each stage, based on the available inventory and backlog in the system, random future demand, and the history of realized demands, such that a

Table 3.1: Review of the related papers

	Planning Horizon	Uncertainty	Problem	Substitution	Service level	Strategy	Methodology
Bitran and Leong (1992)	F	Yield	Co-production	V	J (Products)	S,D	A
Bitran and Dasu (1992)	I	Yield	Co-production	V	I	D	LP, H
Bassok et al. (1999)	S	Dem and Yield	Periodic review inventory model	V			G
Hsu and Bassok (1999)	S	Dem and Yield	Co-production	V			MILP, G
Rao et al. (2004)	S	Dem	Inventory planning + setup	V			H
Hsu et al. (2005)	F	Det	Lot sizing	V			DP
Nagarajan and Rajagopalan (2008)	S, F	Dem	Inventory planning	C		D	H
Lang and Domschke (2010)	F	Det	Lot sizing	V			MILP
Ng et al. (2012)	S	Dem	Co-production	C	M		LP
Zhang et al. (2014)	F	Dem	Inventory planning		J (Period)	D	B&C
Gicquel and Cheng (2018)	F, I	Dem	Lot sizing		J (Period)	S	SAA, MILP
Jiang et al. (2017)	F	Dem	production planning		J (Period)	S	SAA
Liu and Küçükyavuz (2018)	F	Dem	Lot sizing		J (Period)	S	B&C
Chen and Chao (2020)		Dem	Inventory control	C			Online learning
Akçay et al. (2020)	S	Dem		C	I		A
Our work	I	Dem	Lot sizing	V	J (Products)	D	MILP, B&C

Acronyms
Planning Horizon .. I: infinite F: Finite S: Single period
Uncertainty .. Det: Deterministic Dem: Random Demand Yield : Random Yiled
Substitution .. V: supplier-driven C: Customer driven
Service level .. I: Individual J (Period): Joint over multiple periods J (Products): Joint over multiple products M: Maximizing service level
Strategy .. S: Static D: Dynamic
Methodology .. MILP: Mixed integer linrear programming H: Heuristics SAA: Sample average approximation, DP: Dynamic programming
B&C: Branch-and-Cut G: Greedy algorithm A: Model approximation LP: Linear programming

joint service level over all products is to be satisfied in the following stage. This is in line with the "dynamic" strategy that is defined for the stochastic lot sizing problem (Bookbinder and Tan, 1988).

To provide a decision policy for this infinite-horizon problem we propose a rolling-horizon approach, where at each time period we solve a finite-horizon version of the problem and implement the first-period decision obtained from this problem, as illustrated in Figure 3.1. Being at period $\hat{t}$ as the actual time, for the T planning periods for the finite-horizon models, the actual time indices $(\hat{t},...,\hat{t}+T-1)$ are mapped into $(1,...,T)$ for convenience.

In this section, we provide the problem definition of the ideal finite-horizon problem we solve in each time period. This problem is a dynamic stochastic program with chance constraints to represent the service level constraints, and hence is intractable to solve exactly. In Section 3.4 we discuss our proposed approximate solution strategies.

In the finite-horizon problem, we have multiple types of products, whose index set is $\mathscr{K} = \{1,...,K\}$, and T planning periods indexed by $t \in \mathscr{T} = \{1,...,T\}$. We propose a multi-stage stochastic programming model with joint chance-constraints. Being at period $t = 1$ (which is equivalent to an actual decision-making period $\hat{t}$ in the infinite-horizon model), given the state of the system the model considers decisions for the T stages to guide the implementable first-stage ($t = 1$) decisions that would satisfy the joint service level in the next period, $t = 2$. It should be

noted that although we consider the service level in the next period, but by considering the rolling horizon framework we intend to satisfy the service level in all coming periods.

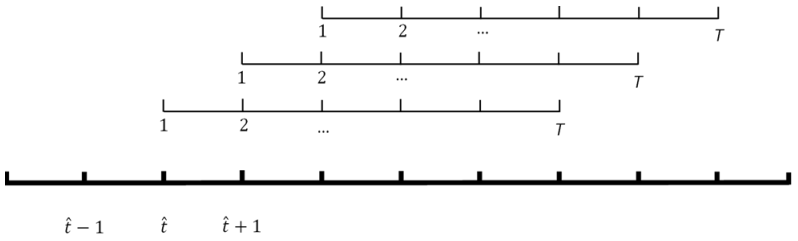


Figure 3.1: Rolling-horizon framework

Figure 3.2 illustrates the dynamics of decisions for the finite-horizon problem at each stage. At each point of time, t, the demand realization vector $\hat{D}_t = (\hat{D}_{t1}, \hat{D}_{t2}, ..., \hat{D}_{tK})$ is observed, and also given the initial state of the system, described by the vector of current on-hand inventory, $\hat{v}_{t-1} = (\hat{v}_{t-1,1}, \hat{v}_{t-1,2}, ..., \hat{v}_{t-1,K})$, and the backlog vector, $\hat{B}_{t-1} = (\hat{B}_{t-1,1}, \hat{B}_{t-1,2}, ..., \hat{B}_{t-1,K})$, two sets of decisions are made. The first set includes substitution, inventory, and backlog decisions denoted by $S_t = (S_{t11}, S_{t12}, \ldots, S_{tKK}), I_t = (I_{t1}, I_{t2}, ..., I_{tK}), B_t = (B_{t1}, B_{t2}, \ldots, B_{tK})$ vectors, respectively. The rest of the decisions in the current period are production, setup, and inventory level after production at the end of current period which are denoted by $x_t = (x_{t1}, x_{t2}, ..., x_{tK}), y_t = (y_{t1}, y_{t2}, ..., y_{tK}), v_t = (v_{t1}, v_{t2}, ..., v_{tK})$ vectors, respectively. It should be noted that all these decisions are made simultaneously, but having lead time of one period and assuming that demand in period t is observed at the beginning of the period, the production quantities made during period t can be used only in the next periods, i.e., they are not available to satisfy the same period demand or backlogged demand. Therefore, we defined two different inventory level vectors, namely, I_t as the inventory level immediately after demand satisfaction, but before production, and v_t as the inventory level at the end of the period, also taking into account the production in period t. The values of v_t and B_t will be the inputs for the next period, describing the next state of the system.

The inventory of a product can be used to satisfy its own demand or another product demand based on the substitution graph G with vertex set $\mathscr{K}$ and arc set $\mathscr{A}$. If $(k, j) \in \mathscr{A}$ then product k can fulfill demand of product j but a substitution cost of c^{sub}_{tkj} per unit is incurred at period t. Note that $(k,k) \in \mathscr{A}$ for all $k \in \mathscr{K}$, and S_{tkk} corresponds to the amount of product k which is used to satisfy its own demand. Demand for a specific product is met either from the inventory of that product

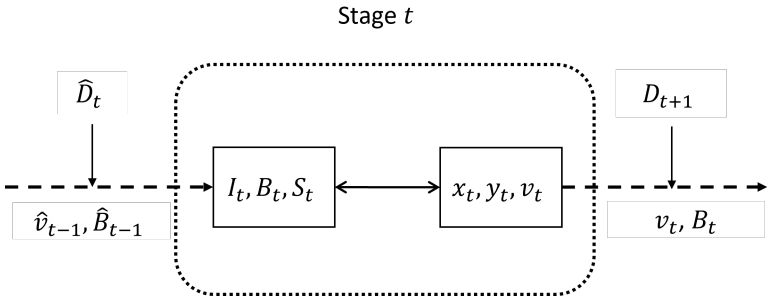


Figure 3.2: Dynamics of decisions at each stage

or from the inventory of another product through substitution, or else the demand is backlogged. In each period t, while insufficient inventory will lead to backlog denoted by B_{tk}, unnecessary stocks will increase the holding cost. An inventory holding cost of c_{tk}^{hold} per unit is charged for the quantity being stored after the demand satisfaction in each period, denoted by I_{tk}. Furthermore, in each period where production occurs, a setup has to be performed which incurs a fixed setup cost of c_{tk}^{setup}. We consider the trade-off between these costs while making decisions at each period, also ensuring that the random demand in the next period can be satisfied with high probability based on a predefined service level.

We model the underlying stochastic process of the demand as a scenario tree, for the finite model with T stages. D_{tk} is the random demand variable for product k in period t, whereas $\hat{D}_{tk}$ denotes its realization at time t. D_{tk}^{Hist} represents the random demand path from period 1 to period t for product k, and $\hat{D}_{tk}^{\text{Hist}}$ denotes its realization (the history) until period t.

We next present our proposed multi-stage stochastic programming model with chance-constraint for the finite-horizon variant of the considered lot sizing problem with substitution. Notation for different sets, parameters and the decision variables are presented in Table 3.2.

We present a dynamic programming formulation where $F_t(.)$ denotes the cost-to-go function at each period $t = 1, 2, ..., T$ and is defined as follows:

$$F_t(v_{t-1}, B_{t-1}, \hat{D}_t^{\text{Hist}}) = \min \sum_{k \in \mathscr{K}} \left(c_{tk}^{\text{setup}} y_{tk} + c_{tk}^{\text{prod}} x_{tk} + c_{tk}^{\text{hold}} I_{tk} + \sum_{j \in \mathscr{K}_k^+} c_{tkj}^{\text{sub}} S_{tkj} \right) +$$

$$\mathbb{E}_{D_{t+1}} \left[F_{t+1}(v_t, B_t, D_{t+1}^{\text{Hist}} | D_t^{\text{Hist}} = \hat{D}_t^{\text{Hist}}) \right] \tag{3.1a}$$

$$\text{s.t. } x_{tk} \leq M_{tk} y_{tk} \qquad \forall k \in \mathscr{K} \tag{3.1b}$$

Table 3.2: Notation for the mathematical model

Sets	**Definition**
$\mathscr{T}$	Set of planning periods, indexed by $1,...,T$
$\mathscr{K}$	Set of products, indexed by $1,...,K$
$G=(\mathscr{K},\mathscr{A})$	Substitution graph
$\mathscr{A}\subseteq\mathscr{K}\times\mathscr{K}$	Directed arcs of substitution graph denoting feasible substitutions, which include self loops
$\mathscr{K}_k^+=\{j\mid(k,j)\in\mathscr{A}\}$	Set of products whose demand can be fulfilled by product k
$\mathscr{K}_k^-=\{j\mid(j,k)\in\mathscr{A}\}$	Set of products that can fulfill the demand of product k
Parameters	**Definition**
c_{tk}^{setup}	Setup cost for product k in period t
c_{tk}^{hold}	Inventory holding cost for product k in period t
c_{tkj}^{sub}	Substitution cost if product k is used to fulfill the demand of product j in period t
c_{tk}^{prod}	Production cost for product k in period t
c_{tk}^{back}	Backlog cost for product k in period t
α	Minimum required joint service level
M_{tk}	A sufficiently large (Big-M) number (to model the logical constraint)
D_{tk}	Random demand variable for product k in period t
D_{tk}^{Hist}	Random demand history from period 1 to period t for product k
$\hat{v}_k$	The amount of initial inventory level for product k
$\hat{B}_k$	The amount of initial backlog for product k
$\mathbb{P}$	The probability distribution of the demand process
Decision variables	**Definition**
y_{tk}	Binary variable which is equal to 1 if there is a setup for product k at period t, 0 otherwise
x_{tk}	Amount of production for product k at period t
S_{tkj}	Amount of product k used to fulfill the demand of product j at period t
I_{tk}	Amount of physical inventory for product k immediately after the demand satisfaction for period t
B_{tk}	Amount of backlog for product k at the end of period t
v_{tk}	The inventory level after production for product k at the end of period t

$$\sum_{j\in\mathscr{K}_k^-} S_{tjk}+B_{tk}=\hat{D}_{tk}+B_{t-1,k} \qquad \forall k\in\mathscr{K} \qquad (3.1c)$$

$$\sum_{j\in\mathscr{K}_k^+} S_{tkj}+I_{tk}=v_{t-1,k} \qquad \forall k\in\mathscr{K} \qquad (3.1d)$$

$$v_{tk}=I_{tk}+x_{tk} \qquad \forall k\in\mathscr{K} \qquad (3.1e)$$

$$\mathbb{P}_{D_{t+1}}\{(v_t,B_t)\in Q(D_{t+1})|D_t^{\text{Hist}}=\hat{D}_t^{\text{Hist}}\}\geq\alpha \qquad (3.1f)$$

$$x_t,v_t,I_t,B_t\in\mathbb{R}_+^K, S_t\in\mathbb{R}_+^{|\mathscr{A}|}, y_t\in\{0,1\}^K \qquad (3.1g)$$

The objective function of the model at time t shown as (3.1a). $F(\cdot)$ represents the optimal objective value from period t to the end of the horizon given the initial inventory level vector and backlog vector. More specifically it minimizes the current stage total cost plus the expected cost-to-go function, which includes the total setup cost, production cost, holding cost, and substitution cost. It should be noted that $F_{T+1}(\cdot)=0$. Constraints (3.1b) are the set up constraints which guarantee that when there is production, the setup variable is forced to take the value 1. Constraints

(3.1c) show that the demand of each product is satisfied by its own production and the substitution by other products or it will be backlogged to the next period. Constraints (3.1d) show that the inventory of product k at the beginning of the current period may be used to satisfy its own demand or other products demand through substitution or it will be stored as inventory for future periods. Constraints (3.1e) define the inventory level after production at the end of the current period which is equal to the amount of inventory (immediately after demand satisfaction) plus the amount of production during the current period.

Constraint (3.1f) is to ensure the joint service level for period $t+1$ which is modeled as a chance-constraint. In this constraint, $Q(D_{t+1})$ is the set of inventory levels after production and backlog quantities such that customer demands given by D_{t+1} can all be met and there is no stock out for any of the products.

$$Q(D_{t+1}) := \{(v_t, B_t) \in \mathbb{R}_+^{2K} : \exists \overline{S} \in \mathbb{R}_+^{|\mathscr{A}|} \text{ s.t. } \sum_{j \in \mathscr{K}_k^-} \overline{S}_{jk} = D_{t+1,k} + B_{tk} \; \forall k \in \mathscr{K} \quad \text{and} \quad \sum_{j \in \mathscr{K}_k^+} \overline{S}_{kj} \leq v_{tk} \; \forall k \in \mathscr{K}\} \tag{3.2}$$

The service level constraint guarantees that the probability of having no stock out in the next period is greater than or equal to α. This probability is defined over D_{t+1}, the demand distribution until period $t+1$, having that part of the demand history until period t is realized and known. Lastly, constraints (3.1g) define the domains of different variables in the model. In addition to these constraints it is possible to add different types of constraints such as capacity constraints to the model.

The main goal of our finite-horizon model is to compute $F_1(\cdot)$, which is the cost-to-go function for the period $t = 1$. In this case, v_{t-1}, B_{t-1} are equal to $\hat{v}_0, \hat{B}_0$ which indicate the vectors for the initial state of the system, and the service level constraint is defined for the second period in finite-horizon model.

It should be noted that in this model, the feasibility of the next stage service level should be guaranteed. This can be satisfied if we have at least one uncapacitated production option for each product whether by its own production or substitution. As there is no capacity constraint in our model, this feasibility is guaranteed.

3.4 Approximate solution policies

In this section, we explain how to make different decisions at each period using different policies. These policies map the state of the system to different decisions and can be used in a rolling-horizon framework. More specifically, we define a type of policies, namely, “production substitution policy" to be applied at each period, guided by our proposed multi-stage stochastic programming model, and inspired by the dynamics of the decision-making process, as shown in Figure 3.2. A “production substitution policy"aims to make the setup, production, and substitution decisions such that the service level in the next period can be satisfied. Ideally, we would solve the multi-stage stochastic programming model, apply its optimal solution for $t = 1$, update the state of the system based on the observed demand, and repeat this process as we move forward in the rolling-horizon framework. However solving this model is challenging due to its complexity and recursive nature. We hence propose different alternative policies in which we use mixed-integer programming (MIP) models as an approximation of the multi-stage model. We propose two deterministic approximations (average and quantile policies), and a two-stage approximation (chance-constraint policy).

In our multi-stage model and similarly in its approximations, we consider the next period demand where we have the uncertainty, due to one period delay in the order arrivals, whether by considering its expected value, quantile value, or the service level. Therefore, it is possible that in period 1, based on the solution of the model we do not meet all demands, or even reach the service level even though our model would generally prioritize meeting demands when possible. In other words, model (3.1) and its approximation policy models do not impose any constraints to minimize the amount of backorder or satisfy the service level in the first period after the demand is observed. To resolve this challenge, in each issue, we first apply an initial step in which we focus on satisfying the current period observed demand. In this step, we define which product observed demand has the potential to be fully satisfied without any backlog, and based on this result extra constraints will be added to the first planning period of the policy model. This initial step and different types of “production substitution policy" are explained in detail in the following sections.

3.4.1 Backlog determination in the first period

The backlog determination step is a prerequisite for a “production substitution policy" which focuses on substitution, inventory and backlog decisions only for the first period to satisfy period $t = 1$ demand (or rather to guarantee the feasibility of the current-stage demand satisfaction). To this end, we solve the linear programming (LP) model (3.3) which minimizes the total backlog in the current period (corresponding to $t = 1$ in the model). This is also applicable in reality, as the companies try to satisfy their available orders simultaneously as much as possible. The result of this model is a subset of products for which it is possible to fully satisfy their demand without any backlog. We then force the backlog for this subset equal to zero in period 1 in “production substitution policy" model. It should be noted that in this step, we do not define how the demand should be satisfied, and this will be later defined in the “production substitution policy". The mentioned LP model is presented as follows:

$$\min \sum_{k \in \mathscr{K}} B_{1k} \tag{3.3a}$$

$$\text{s.t.} \sum_{j \in \mathscr{K}_k^-} S_{1jk} + B_{1k} = \hat{D}_{1k} + \hat{B}_{0k} \qquad \forall k \in \mathscr{K} \tag{3.3b}$$

$$\sum_{j \in \mathscr{K}_k^+} S_{1kj} + I_{1k} = \hat{v}_{0k} \qquad \forall k \in \mathscr{K} \tag{3.3c}$$

$$I_1, B_1 \in \mathbb{R}_+^K, S_1 \in \mathbb{R}_+^{|\mathscr{A}|} \tag{3.3d}$$

The objective function (3.3a) is to minimize the total backlog. Constraints (3.3b) guarantee that the demand and backlog is either satisfied in the current period or it will be backlogged. Constraints (3.3c) show that the available inventory is either used to satisfy the demand of different products, or will be stored as an inventory. If the optimal value of B_{1k} is equal to zero, we add product k to set $\hat{\mathscr{K}}$. Constraints $B_{1k} = 0$ will be added to the model of the “production substitution policy" for all $k \in \hat{\mathscr{K}}$. How the demand of this product is satisfied is also defined in the policy model based on different cost parameters.

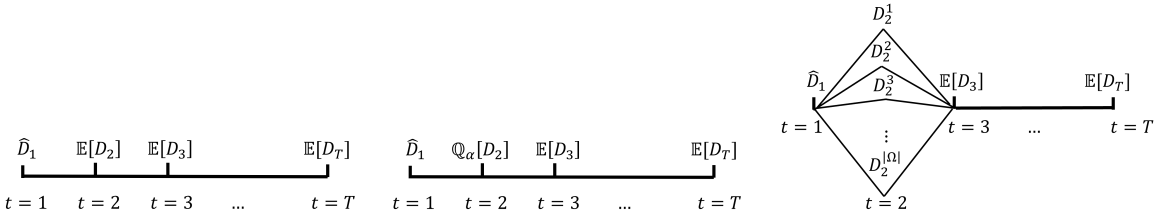
It should be noted that it is also possible to use other models in this step. For example, minimizing the backlog for each product separately. However, it is not possible to consider substitution options as the products should not have a link together. Another option is to add some constraints

in the policy model to satisfy the service level; however, we may have an infeasibility issue, or in the case of using soft constraints, we have the challenge of parameter tuning.

The advantage of our proposed method is its simplicity and the fact that it can be easily applied to different versions of “production substitution policy" by adding extra constraints. In addition, minimizing the total backlog is in line with the notion of joint service level and the possibility of substitution. It is also possible to minimize the weighted sum of backlog, in case the company has priority in satisfying different products demand.

3.4.2 Decision policy

At period $t = 1$, based on the current state of the system we apply a “production substitution policy", which takes $\hat{v}_0, \hat{B}_0$, $\hat{D}_1$, and the set $\hat{\mathscr{K}}$ from the backlog determination step as inputs. In this section, we explain two policies with deterministic models (average and quantile policies) and a policy with a chance-constrained two-stage stochastic programming model (chance-constraint policy) as an approximation for the multi-stage model. We later show that while the deterministic policy models have the advantage of faster execution time, the chance-constraint policy model results in more accurate solutions.



(a) Average policy (b) Quantile policy (c) Chance-constraint policy

Figure 3.3: Demand approximation in different decision policies

Deterministic policies

In these policies, we represent the future demand by a single scenario, and we propose two different deterministic policies based on that. These policies’ model approximate the dynamic programming model by eliminating the chance-constraint and substituting the stochastic demand with the deterministic value. In the first policy, namely the “average policy", the stochastic demand for all the products and in all periods is substituted by its expected value. The second policy which is called

the “quantile policy" differs from the “average policy" by substituting the stochastic demand for the next immediate period (which corresponds to $t = 2$) by the quantile of the future demand distribution, which is defined based on the service level. Figure 3.3 illustrates the demand pattern for the average and quantile policies in sub-figures (a) and (b), respectively. For both of these policies, we assume that there is no backlog for $t > 1$, which means that we should at least satisfy the expected demand in all future periods. Model (3.4) represents the “average policy".

$$\min \sum_{t\in\mathscr{T}}\sum_{k\in\mathscr{K}}\left(c_{tk}^{\text{setup}}y_{tk}+c_{tk}^{\text{prod}}x_{tk}+c_{tk}^{\text{hold}}I_{tk}+\sum_{j\in\mathscr{K}_k^+}c_{tkj}^{\text{sub}}S_{tkj}\right) \tag{3.4a}$$

$$\text{s.t. } x_{tk}\leq M_{tk}y_{tk} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{K} \tag{3.4b}$$

$$\sum_{j\in\mathscr{K}_k^-}S_{1jk}+B_{1k}=\hat{D}_{1k}+\hat{B}_{0k} \qquad \forall k\in\mathscr{K} \tag{3.4c}$$

$$\sum_{j\in\mathscr{K}_k^-}S_{tjk}=\mathbb{E}[D_{tk}]+B_{t-1,k} \qquad \forall t\in\mathscr{T}\setminus\{1\},\forall k\in\mathscr{K} \tag{3.4d}$$

$$\sum_{j\in\mathscr{K}_k^+}S_{1kj}+I_{1k}=\hat{v}_{0k} \qquad \forall k\in\mathscr{K} \tag{3.4e}$$

$$\sum_{j\in\mathscr{K}_k^+}S_{tkj}+I_{tk}=v_{t-1,k} \qquad \forall t\in\mathscr{T}\setminus\{1\},\forall k\in\mathscr{K} \tag{3.4f}$$

$$v_{tk}=I_{tk}+x_{tk} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{K} \tag{3.4g}$$

$$B_{1k}=0 \qquad \forall k\in\hat{\mathscr{K}} \tag{3.4h}$$

$$x_t,v_t,I_t,B_t\in\mathbb{R}_+^K,S_t\in\mathbb{R}_+^{|\mathscr{A}|},y_t\in\{0,1\}^K \qquad \forall t\in\mathscr{T} \tag{3.4i}$$

The objective function (3.4a) minimizes the total cost of setup, production, holding and substitution cost. Constraints (3.4b) guarantee that in each planning period, when there is a production, there will be a setup. Tight big-M values for these constraints can calculated as follows:

$$M_{tk}=\sum_{j\in\mathscr{K}_k^+}\left(\hat{B}_{0j}+\hat{D}_{1j}+\sum_{t=2}^{T}\mathbb{E}[D_{tj}]\right) \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{K} \tag{3.5}$$

Constraints (3.4c) to (3.4f) are the inventory, backlog, and substitution balance constraints, which are defined for $t = 1$ and $t > 1$ separately. In constraints (3.4d) there is no backlog variable for period t, which guarantee the average demand satisfaction in periods $t > 1$. In this model, we may have backlog only in period 1. Constraints (3.4g), define the available inventory after production.

Constraints (3.4h) are defined based on the result of backlog determination step and force the backlog equal to 0 in the first period for all the products in set $\hat{\mathscr{K}}$.

The quantile policy model is the same as model (3.4), except for one set of constraints. In this policy, for $t = 2$ constraints (3.4d) are replaced by constraints (3.6). The difference between the average policy and the quantile policy is that in the second period, where we need to consider the service level, the average demand is replaced by the α quantile of the demand, denoted by $\mathbb{Q}_\alpha$. In the quantile policy the M values in constraints (3.4b) are calculated by equations (3.7).

$$\sum_{j \in \mathscr{K}_k^-} S_{2jk} = \mathbb{Q}_\alpha(D_{2k}) + B_{1k} \qquad \forall k \in \mathscr{K} \tag{3.6}$$

$$M_{tk} = \sum_{j \in \mathscr{K}_k^+} \left(\hat{B}_{0j} + \hat{D}_{1j} + \mathbb{Q}_\alpha(D_{2j}) + \sum_{t=3}^{T} \mathbb{E}[D_{tj}] \right) \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \tag{3.7}$$

Chance-constraint policy

The deterministic policies that we explained in the previous section did not consider the service level constraint explicitly in their models. To consider the service level we add a chance constraint to the deterministic policy models in the second period. To this end, we consider a set of demand scenarios, Ω, for period $t = 2$. D_{tk}^{ω} denotes the demand of product $k \in \mathscr{K}$ in period $t \in \mathscr{T}$ under scenario $\omega \in \Omega$. The drawback of this idea is that it fails to capture the cost of substitution implicit in the chance constraint. More specifically, the chance constraint ensures that with high probability there exists substitutions that can meet all demands, but ignores the cost of the substitutions which result in choosing low order quantity so that significant substitution is necessary to meet demands. To solve this issue, we consider different cost elements for each single scenario $\omega \in \Omega$, in addition to defining the service level constraint using these scenarios. For periods $t \geq 3$, the stochastic demand is substituted by its expected value, similar to the deterministic policies. The demand pattern used in this policy in depicted in sub-figure (c) in Figure 3.3. We next present the mathematical model for this policy. In addition to previous decision variables, new decision variables are defined for each of the scenarios in the period 2. In this model, $I_{tk}^{\prime\omega}$ and $B_{tk}^{\prime\omega}$ denote the inventory and backlog at period t for product k under scenario ω, respectively. $S_{tkj}^{\prime\omega}$ represents the substitution amount of product k for product j at period t under scenario ω. The mentioned variables are then connected to the related variable in period 3 by using additional constraints in

the model. This model is presented as follows:

$$\min \sum_{k\in\mathscr{K}}\left(c_{1k}^{\text{setup}}y_{1k}+c_{1k}^{\text{prod}}x_{1k}+\sum_{j\in\mathscr{K}_k^+}c_{1kj}^{\text{sub}}S_{1kj}+c_{1k}^{\text{hold}}I_{1k}\right)+$$
$$\sum_{k\in\mathscr{K}}\left(c_{2k}^{\text{setup}}y_{2k}+c_{2k}^{\text{prod}}x_{2k}+c_{2k}^{\text{hold}}I_{2k}+c_{2k}^{\text{back}}B_{2k}+\frac{1}{|\Omega|}\sum_{\omega\in\Omega}\sum_{j\in\mathscr{K}_k^+}c_{2kj}^{\text{sub}}S_{2kj}^{\prime\omega}\right)+$$
$$\sum_{t=3}^{T}\sum_{k\in\mathscr{K}}\left(c_{tk}^{\text{setup}}y_{tk}+c_{tk}^{\text{prod}}x_{tk}+\sum_{j\in\mathscr{K}_k^+}c_{tkj}^{\text{sub}}S_{tkj}+c_{tk}^{\text{hold}}I_{tk}\right) \tag{3.8a}$$

$$\text{s.t. } x_{tk}\le M_{tk}y_{tk} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{K} \tag{3.8b}$$
$$\sum_{j\in\mathscr{K}_k^-}S_{1jk}+B_{1k}=\hat{D}_{1k}+\hat{B}_{0k} \qquad \forall k\in\mathscr{K} \tag{3.8c}$$
$$\sum_{j\in\mathscr{K}_k^-}S_{tjk}=\mathbb{E}[D_{tk}]+B_{t-1,k} \qquad \forall t\in\mathscr{T},t\ge 3,\forall k\in\mathscr{K} \tag{3.8d}$$
$$\sum_{j\in\mathscr{K}_k^+}S_{1kj}+I_{1k}=\hat{v}_{0k} \qquad \forall k\in\mathscr{K} \tag{3.8e}$$
$$\sum_{j\in\mathscr{K}_k^+}S_{tkj}+I_{tk}=v_{t-1,k} \qquad \forall t\in\mathscr{T},t\ge 3,\forall k\in\mathscr{K} \tag{3.8f}$$
$$v_{tk}=I_{tk}+x_{tk} \qquad \forall t\in\mathscr{T},\forall k\in\mathscr{K} \tag{3.8g}$$
$$B_{1k}=0 \qquad \forall k\in\hat{\mathscr{K}} \tag{3.8h}$$
$$\sum_{j\in\mathscr{K}_k^-}S_{2jk}^{\prime\omega}+B_{2k}^{\prime\omega}=D_k^{\omega}+B_{1k} \qquad \forall k\in\mathscr{K},\forall\omega\in\Omega \tag{3.8i}$$
$$\sum_{j\in\mathscr{K}_k^+}S_{2kj}^{\prime\omega}+I_{2k}^{\prime\omega}=v_{1k} \qquad \forall k\in\mathscr{K},\forall\omega\in\Omega \tag{3.8j}$$
$$\frac{1}{|\Omega|}\sum_{\omega\in\Omega}I_{2k}^{\prime\omega}=I_{2k} \qquad \forall k\in\mathscr{K} \tag{3.8k}$$
$$\frac{1}{|\Omega|}\sum_{\omega\in\Omega}B_{2k}^{\prime\omega}=B_{2k} \qquad \forall k\in\mathscr{K} \tag{3.8l}$$
$$\sum_{\omega\in\Omega}\mathbb{1}\{(v_1,B_1)\in Q(D_{2k}^{\omega})\}\ge\lceil\alpha|\Omega|\rceil \tag{3.8m}$$
$$x_t,v_t,I_t,I_2^{\prime},B_t,B_2^{\prime}\in\mathbb{R}_+^K,S_t,S_2^{\prime}\in\mathbb{R}_+^{|\mathscr{A}|},y_t\in\{0,1\}^K \tag{3.8n}$$

To have a more clear description, the objective function of the extensive form model (3.8a) has broken into three parts, the cost of period 1, the cost of period 2, and the cost of periods 3 to T. In period 2, the substitution cost is defined for each of the scenarios separately, and the average substitution cost over all scenarios is used as the total substitution cost. Only in this period, we also consider backlog cost so that the cost function matches with the service level constraint. Constraints (3.8b) are production setup constraints for which the value of M is calculated using equation 3.9. Constraints (3.8c) to (3.8f) are the inventory, backlog, and substitution balance constraints. Constraints (3.8c) and (3.8e) are for period 1 period and constraint (3.8d) and (3.8f) are for the periods 3 to T. Constraints (3.8g) define the inventory after production.

$$M_{tk} = \sum_{j \in \mathscr{K}_k^+} \left(\hat{B}_{0j} + \hat{D}_{1j} + \max_{\omega \in \Omega} D_{2j}^{\omega} + \sum_{t=3}^{T} \mathbb{E}[D_{tj}] \right) \qquad \forall t \in \mathscr{T}, \forall k \in \mathscr{K} \tag{3.9}$$

Constraints (3.8i) and (3.8j) are the inventory, backlog, and substitution balance constraints for period 2 and each individual scenario. Constraints (3.8k) and (3.8l) recombine the scenarios in period 2 and link the average inventory and backlog over all scenarios in this period to the inventory and backlog amount in the same period. These averages are used in the inventory and backlog balance constraints for period 3. Constraint (3.8m) is the service level constraint in which the sum of feasible scenarios based on the values of v_1 and B_1 should be greater than α percent of the number of scenarios. This model is challenging to solve mostly due to the constraints for the joint service level. In the next section we propose an efficient B&C algorithm to solve this model.

3.5 Solving the chance-constrained model

We now discuss how to solve the proposed model (3.8). We first present a MIP formulation for the service level constraint (3.8m) based on the demand scenario set Ω, and then substitute it with the service level constraint in model (3.8).

We define the binary variables z_ω, where $z_\omega = 0$ indicates the available inventory is adequate to meet demands in scenario ω without backlogging, and $z_\omega = 1$ otherwise. $\overline{B}^{\omega}$ represents the backlog vector and $\overline{S}^{\omega}$ represents the substitution vector for scenario ω. The joint chance-constraint (3.8m)

in model (3.8) is then replaced by constraints (3.10a)-(3.10d). It should be noted that it is possible to use the previously defined decision variables for each of the scenarios (namely B'^{ω}_{tkj} and S'^{ω}_{tkj}) in the resulting extensive formulation. However, as the B&C algorithm is used to only deal with the joint chance constraint MIP formulation, we need to define a separate set of decision variables for this set of constraints. To match the service level constraint and the objective function cost for each scenario and the service level constraint, we need to tune the backlog cost parameter, c^{back}_{2k}, in (3.8a).

$$\overline{B}^{\omega}_k + \sum_{j \in \mathscr{K}^-_k} \overline{S}^{\omega}_{jk} = D^{\omega}_k + B_{1k} \qquad \forall \omega \in \Omega, \forall k \in \mathscr{K} \tag{3.10a}$$

$$\sum_{j \in \mathscr{K}^+_k} \overline{S}^{\omega}_{kj} \leq v_{1k} \qquad \forall \omega \in \Omega, \forall k \in \mathscr{K} \tag{3.10b}$$

$$\overline{B}^{\omega}_k \leq \overline{M}^{\omega}_k z_{\omega} \qquad \forall \omega \in \Omega, \forall k \in \mathscr{K} \tag{3.10c}$$

$$\sum_{\omega \in \Omega} z_{\omega} \leq \lfloor (1-\alpha)|\Omega| \rfloor \tag{3.10d}$$

Constraints (3.10a) and (3.10b) define the backlog and substitution for each scenario ω. Constraints (3.10c) guarantee that when there is a backlog for any of the product in scenario ω, the indicator variable is turned on, i.e., $z_{\omega} = 1$. In these constraints, the $\overline{M}$ values are defined using equation (3.10e). Constraint (3.10d) is the service level constraint.

$$\overline{M}^{\omega}_k = D^{\omega}_k + \hat{D}_{1k} + \hat{B}_{0k} \qquad \forall \omega \in \Omega, \forall k \in \mathscr{K} \tag{3.10e}$$

Constraints (3.10c) make the mathematical model challenging to solve, therefore, we propose a B&C algorithm to solve the extensive form of the two-stage chance constrained formulation. This method is based on the algorithm proposed by Luedtke (2014) for joint chance constraints, which is tailored for our problem. Considering variable z_{ω} for each $\omega \in \Omega$, when $z_{\omega} = 0$, we should enforce that v_1, B_1 lie within the $Q(D^{\omega})$ polyhedron.

$$Q(D^{\omega}) := \{(v,B) \in \mathbb{R}^{2K}_+ : \exists \overline{S}^{\omega} \in \mathbb{R}^{|\mathscr{A}|}_+ \text{ s.t. } \sum_{j \in \mathscr{K}^-_k} \overline{S}^{\omega}_{jk} = D^{\omega}_k + B_{1k} \; \forall k \in \mathscr{K} \text{ and } \sum_{j \in \mathscr{K}^+_k} \overline{S}^{\omega}_{kj} \leq v_{1k} \; \forall k \in \mathscr{K}\}$$

To solve the model, we consider a master problem where we eliminate constraints (3.10a)-(3.10c) from the extensive form of the model. Assume we have solved a master problem and obtained a solution $(\hat{z}_1, \hat{v}_1, \hat{B}_1)$ for period 1. Note that this solution may or may not satisfy the integrality constraints (e.g., if we have solved an LP relaxation of the master problem). Given a demand scenario $\omega \in \Omega$ with $\hat{z}_1^\omega < 1$, our task is to assess if $(\hat{v}_1, \hat{B}_1) \in Q(D^\omega)$, and if not, attempt to generate a cut to remove this solution. In the case of an integer feasible solution, we will always be able to do so when $(\hat{v}_1, \hat{B}_1) \notin Q(D^\omega)$.

We can test if given $(\hat{v}_1, \hat{B}_1) \in Q(D^\omega)$ by solving the following LP:

$$
\begin{aligned}
V_\omega(\hat{v}_1, \hat{B}_1) := \min_{w, \overline{S}^\omega} \quad & \sum_{k \in \mathscr{K}} w_k \\
\text{s.t.} \quad & \sum_{j \in \mathscr{K}_k^-} \overline{S}^\omega_{jk} + w_k = D^\omega_k + \hat{B}_{1k} && \forall k \in \mathscr{K} && (\pi_k) \\
& - \sum_{j \in \mathscr{K}_k^+} \overline{S}^\omega_{kj} \geq -\hat{v}_{1k} && \forall k \in \mathscr{K} && (\beta_k) \\
& w \in \mathbb{R}^K_+, \overline{S}^\omega \in \mathbb{R}^{|\mathscr{A}|}_+
\end{aligned}
$$

By construction, $(\hat{v}_1, \hat{B}_1) \in Q(D^\omega)$ if and only if $V_\omega(\hat{v}_1, \hat{B}_1) \leq 0$, which means that there is no backlog for this scenario. Furthermore, if $(\hat{\pi}, \hat{\beta})$ is an optimal dual solution, then by weak duality, the cut

$$\sum_{k \in \mathscr{K}} \hat{\pi}_k (D^\omega_k + B_{1k}) - \sum_{k \in \mathscr{K}} \hat{\beta}_k v_{1k} \leq 0$$

is a valid inequality for $Q(D^\omega)$. Rearranging this, it takes the form:

$$\sum_{k \in \mathscr{K}} \hat{\beta}_{1k} v_{1k} - \sum_{k \in \mathscr{K}} \hat{\pi}_k B_{1k} \geq \sum_{k \in \mathscr{K}} \hat{\pi}_k D^\omega_k.$$

If $V_\omega(\hat{v}_1, \hat{B}_1) > 0$ then the corresponding cut will be violated by $(\hat{v}_1, \hat{B}_1)$.

The inequality derived above is only valid when $z_\omega = 0$. We thus need to modify it to make it valid for the master problem. To derive strong cuts based on this base inequality, we can solve an additional set of subproblems once we have the coefficients $(\hat{\pi}, \hat{\beta})$. In particular, for every scenario ω', we can solve:

$$\min_{v_1, B_1, S_1, \overline{S}^{\omega'}} \sum_{k \in \mathscr{K}} \hat{\beta}_k v_{1k} - \sum_{k \in \mathscr{K}} \hat{\pi}_k B_{1k}$$

$$
\begin{aligned}
\text{s.t.} \quad & \sum_{j \in \mathscr{K}_k^-} \overline{S}_{jk}^{\omega'} = D_k^{\omega'} + B_{1k} && \forall k \in \mathscr{K} \\
& \sum_{j \in \mathscr{K}_k^+} \overline{S}_{kj}^{\omega'} \leq v_{1k} && \forall k \in \mathscr{K} \\
& \sum_{j \in \mathscr{K}_k^-} S_{1jk} + B_{1k} = D_{1k} + \hat{B}_{0k} && \forall k \in \mathscr{K} \\
& \sum_{j \in \mathscr{K}_k^+} S_{1kj} \leq \hat{v}_{0k} && \forall k \in \mathscr{K} \\
& v_1, B_1 \in \mathbb{R}_+^K,\ S_1, \overline{S}^{\omega'} \in \mathbb{R}_+^{|\mathscr{A}|}
\end{aligned}
$$

Note that in this problem we consider substitution variables both for the period 1 and for each scenario in period 2 under consideration, ω'. The substitution variables for the scenario ω' are to enforce that $(v_1, B_1) \in Q(D^{\omega'})$. The substitution variables for the period 1 are to enforce that B satisfies this period constraints.

Considering $h_{\omega'}(\hat{\pi}, \hat{\beta})$ as any lower bound on the optimal objective value, $V_{\omega'}(\hat{v}_1, \hat{B}_1)$, and given the structure of this problem, we can obtain potentially weaker cuts, but saving significant work. In particular, we use $h_{\omega'}(\hat{\pi}, \hat{\beta}) = \sum_{k \in \mathscr{K}} \hat{\pi}_k D_k^{\omega'}$ for each $\omega' \in \Omega$, instead of solving the above defined LP and using its optimal solution. This should be valid because the dual feasible region of the set $Q(D^{\omega'})$ is independent of ω', so a dual solution from one ω can be used to define an inequality valid for any other ω'.

After evaluating $h_{\omega'}(\hat{\pi}, \hat{\beta})$ for each $\omega' \in \Omega$, we then sort the values to obtain a permutation σ of Ω which satisfies:

$$
h_{\sigma_1}(\hat{\pi}, \hat{\beta}) \geq h_{\sigma_2}(\hat{\pi}, \hat{\beta}) \geq \cdots \geq h_{\sigma_{|\Omega|}}(\hat{\pi}, \hat{\beta})
$$

Then, letting $p = \lfloor (1-\alpha)|\Omega| \rfloor$, the following inequalities are valid for the master problem (Luedtke, 2014):

$$
\sum_{k \in \mathscr{K}} \hat{\beta}_k v_1 k - \sum_{k \in \mathscr{K}} \hat{\pi}_k B_1 k + \Big(h_{\sigma_1}(\hat{\pi}, \hat{\beta}) - h_{\sigma_i}(\hat{\pi}, \hat{\beta}) \Big) z_{\sigma_1} + \Big(h_{\sigma_i}(\hat{\pi}, \hat{\beta}) - h_{\sigma_{p+1}}(\hat{\pi}, \hat{\beta}) \Big) z_{\sigma_i} \geq h_{\sigma_1}(\hat{\pi}, \hat{\beta}),
$$

$\forall i = 1, \ldots, p$

Any of these inequalities could be added, if violated by the current solution $(\hat{z}_1, \hat{v}_1, \hat{B}_1)$.

The final step for obtaining strong valid inequalities is to search for *m*ixing inequalities, which have the following form. Given a subset $T = \{t_1, t_2, \ldots, t_\ell\} \subseteq \{\sigma_1, \sigma_2, \ldots, \sigma_p\}$, where $h_{t_{l+1}} = h_{\sigma_{p+1}}$,

the inequality

$$\sum_{k\in\mathscr{K}} \hat{\beta}_k v_{1k} - \sum_{k\in\mathscr{K}} \hat{\pi}_k B_{1k} + \sum_{i=1}^{\ell} \left(h_{t_i}(\hat{\pi},\hat{\beta}) - h_{t_{i+1}}(\hat{\pi},\hat{\beta}) \right) z_{t_i} \geq h_{t_1}(\hat{\pi},\hat{\beta})$$

is valid for the master problem. Although the number of such inequalities grows exponentially with p, there is an efficient algorithm for finding a most violated inequality (Günlük and Pochet, 2001) for given $(\hat{z}_1, \hat{B}_1, \hat{v}_1)$, which is provided in Algorithm 1.

Algorithme 1 : Finding the most violated inequality

OUTPUT: A most violated mixing inequality defined by ordered index set t ;

INPUT: $\hat{z}_{\sigma_i}, \sigma_i, h_{\sigma_i}(\hat{\pi},\hat{\beta})$, $i = 1,\ldots,p+1$;

Sort the $\hat{z}_{\sigma_i}$ values to obtain permutation ρ of the indices satisfying:

$\hat{z}_{\rho_1} \leq \hat{z}_{\rho_2} \leq \cdots \leq \hat{z}_{\rho_{p+1}}$;

$v \leftarrow h_{\sigma_{p+1}}(\hat{\pi},\hat{\beta})$;

$T \leftarrow \{\}$;

$i \leftarrow 1$;

while $v < h_{\sigma_1}(\hat{\pi},\hat{\beta})$ **do**

  if $h_{\rho_i}(\hat{\pi},\hat{\beta}) > v$ **then**

    $T \leftarrow T \cup \{\rho_i\}$;

    $v \leftarrow h_{\rho_i}(\hat{\pi},\hat{\beta})$;

  end

  $i \leftarrow i+1$;

end

3.6 Computational experiments

In this section, we explain the instance generation, and the rolling-horizon framework that we use to conduct the numerical experiments.

3.6.1 Instance generation

To evaluate different policies and algorithms, we generate a variety of instances based on Rao et al. (2004) and Hsu et al. (2005) for the substitution part, and Helber et al. (2013), for the lot sizing

related parameters, with some justifications for our problem. Table 3.4 illustrates different cost parameters in the model and how to define them based on the data generation parameters. Table 3.3 summarizes the data generation parameters, their base value and their variation for sensitivity analysis. Considering 10 products, in the base case, one way substitution is available for four consecutive products ordered based on their values. It should be noted that τ should start from 1 to include the production cost difference in the substitution cost. In other words, more than average production of a product means lower than average production of some other products due to the substitutions and these variation from averages is considered in the substitution cost. Due to this cost structure, we can ignore the production cost in the objective function as it constant under different policies.

Table 3.3: Parameters for the base case and the sensitivity analysis

Parameters	Base Case	Variation
T	6	6, 8, 10
K	10	
η	0.2	0.1, 0.2, 0.5
τ	1.5	1, 1.25, 1.5, 1.75, 2, 2.5
ρ	0.05	0.02, 0.05, 0.1, 0.2 ,0.5
TBO	1	1, 1.25, 1.5, 1.75, 2
α	95%	80%, 90%, 95%, 99%

Table 3.4: Data generation for the cost parameters

c_{tk}^{prod}	$1+\eta\times(K-k)$
c_{tkj}^{sub}	$\max(0,\tau\times(c_{tk}^{\text{prod}}-c_{tj}^{\text{prod}}))$
c_{tk}^{hold}	$\rho\times c_{tk}^{\text{prod}}$
c_{tk}^{setup}	$E[\overline{D_{tk}}]\times TBO^2\times c_{tk}^{\text{hold}}/2$

In addition to the mentioned cost parameters, in the chance-constraint policy formulation we have a backlog cost which needs to be tuned. This parameter is calculated by equation (3.11), which is equal to the maximum possible cost of substitution.

$$c_{2k}^{back}=\max_{l\in\mathscr{K},j\in\mathscr{K}_k^-} c_{2jl}^{\text{sub}} \qquad \forall k\in\mathscr{K} \tag{3.11}$$

To generate the random demand we used autoregressive process model (Jiang et al., 2017) which considers the correlation in different stage demand as

$$D_{t+1,k} = C + AR_1 \times D_{kt} + AR_2 \times \varepsilon_{t+1,k} \qquad \forall k \in \mathscr{K}, \forall t \in \mathscr{T} \tag{3.12}$$

where C, AR_1, and AR_2 are parameters of the model, and $\varepsilon_{t+1,k}$ is a random noise with normal distribution with the mean of 0 and standard deviation of 1. In our data sets, $C = 20$, $AR_1 = 0.8$, and $AR_2 = 0.1 \times 100$. With these data, the expected demand for each product in each period is equal to 100. To apply this autoregressive process, we generated a random noise set and used it for both observed demand and the stochastic scenario sets generation for different policy formulations. For the observed demand, we choose one random number out of the noise set for each product and we define the new observed demand based on that and the last observed demand. For the random demand in the next period we generate a scenario set based on all the elements in the random noise set and the observed demand. The scenarios are assigned equal probabilities.

As we have no production in the first period, without loss of generality, we assume that the demand in the first period is equal to zero, otherwise, if there is no initial inventory, the service level constraint will not be satisfied. In the AR data generation procedure, we start with the defined average in the first period and then follow the procedure for the rest of the periods. Then to make the first period demand equal to 0 we subtract this average from the first period demand.

The algorithms are implemented in Python and MIP models are solved using IBM ILOG CPLEX 12.8. All the experiments are performed on a 2.4 GHz Intel Gold processor with only one thread on the Compute Canada computing grid.

3.6.2 Rolling-horizon framework

The original problem we consider in this research is an infinite-horizon problem and the decision policies proposed based on the finite-horizon version of this problem are applied and evaluated in a rolling-horizon framework. The infinite-horizon time period is indexed by $\hat{t} \in \{1, 2, ..., T_{\text{Sim}}\}$ and the finite-horizon model periods are indexed by $t \in \{1, 2, ..., T\}$. In this framework, at each period $\hat{t}$, a model with T planning periods is solved, and only the decisions corresponding to the first period are implemented. Based on these decisions and the observed demand, the state of the system is updated and the next T planning period model is solved, as we roll the horizon. We simulate the

decision-making process in a rolling-horizon framework using Algorithm (2) in which at each time period $\hat{t}$ we first execute the backlog determination model, and then solve a finite-horizon model depending on the selected policy. We use two different measures to evaluate different policies, the actual total cost and the actual service level. At each period $\hat{t}$, we calculate the total cost based on the observed demand and implemented solution values, in this period. We consider the average of this cost over all simulated time periods as the total cost of a policy. It should be noted that we ignore the production cost in this calculation as it is constant under different policies. For the service level we consider the percentage of the periods in which the joint service level is satisfied by checking if there is any backlog after observing the demand in each period. For calculating the confidence intervals on these measures, we use batch-based estimations, with a batch size of 25 periods. The simulation time horizon (T_{Sim}) is 4000 periods, and we ignore 10 initial periods as the warm-up periods.

Algorithme 2 : Rolling-horizon implementation

OUTPUT: The confidence interval of the total cost and the service level

INPUT: Demand simulation over T_{Sim} periods, Production policy

$\hat{t} = 1, \hat{v}_{t_0} = 0, \hat{B}_{t_0} = 0, \mathscr{O} = \emptyset, \mathscr{Z} = \emptyset$

while $\hat{t} \leq T_{Sim}$ **do**

Solve the backlog determination model (3.3) for period $\hat{t}$ and let $B^*_{\hat{t}}$ be the optimal backlog solution.

Let $\hat{\mathscr{K}} = \{k \in \mathscr{K} : B^*_{\hat{t}k} = 0\}$

Solve the model (3.1) approximation based on selected policy, $\hat{v}_{\hat{t}-1,k}, \hat{B}_{\hat{t}-1,k}$, the observed demands $\hat{D}_{\hat{t}}$, and set $\hat{\mathscr{K}}$.

Let $x^*_{\hat{t}}, y^*_{\hat{t}}, S^*_{\hat{t}}, B^*_{\hat{t}}$ be the resulting solution for period $\hat{t}$, i.e., the first-period solution of model (3.1), and the $Obj_{\hat{t}}$ be the total cost of period $\hat{t}$ based on this solution.

if $\exists k \in \mathscr{K}, B^*_{\hat{t}k} \geq 0$ **then**

$Z_{\hat{t}} = 1$

else

$Z_{\hat{t}} = 0$

end

Add $Obj_{\hat{t}}$ to the set $\mathscr{O}$

Add $Z_{\hat{t}}$ to the set $\mathscr{Z}$

$\hat{t} \leftarrow \hat{t} + 1$

end

Build confidence intervals for the cost and service level using $\mathscr{O}$ and $\mathscr{Z}$, respectively.

3.6.3 Methodology evaluation

In this section, we analyze the efficiency of the proposed B&C algorithm used to solve the chance-constraint policy model against the extensive form formulation (Model (3.8) in which the joint chance-constraint (3.8m) is replaced by constraints (3.10a)-(3.10d). To this end, we generate some instances with different parameters and solve each instance for 5 different periods of the problem with two different methods. Based on some preliminary analysis we select 24 different challenging instances as follows. $T \in \{6, 8, 10\}$, $K = 10$, $\eta = 0.2$, $\tau = 0.5$, $\rho = 0.1$, $TBO \in \{1, 2\}$, and $\alpha \in \{80\%, 90\%, 95\%, 99\%\}$. We consider partial substitution in which a product can be substituted by

three consecutive higher-grade products. The time limit is set to 7200 seconds. We used the stages after the warm-up periods with similar initial state for all the methods. For the B&C algorithm we use the faster version instead of the stronger version, as the preliminary results do not show significant difference between the two versions. More specifically, the faster version shows slightly better performance.

We analyze the performance of the two methods using three measures, the average CPU time in second (Time), the average integrality gap (Gap), and the average number of optimal solutions found out of 5 (# OPT) over all the instances in one group. The results are given in Table 3.5 .

Table 3.5: Comparison of methodologies to solve the model with the service level

	B&C			Extensive form		
$\|\Omega\|$	Time	Gap (%)	# Opt	Time	Gap(%)	# Opt
100	10.3	0.0	5.0	74.6	0.0	5.0
200	34.6	0.0	5.0	400.8	0.0	5.0
300	60.6	0.0	5.0	1152.5	0.0	4.8
500	206.1	0.0	5.0	3356.3	0.4	4.0
1000	990.1	0.0	4.9	6170.3	1.4	1.3
α (%)						
80	417.2	0.0	5.0	2486.6	0.3	3.9
90	299.5	0.0	5.0	2690.7	0.4	3.8
95	202.2	0.0	5.0	2166.3	0.4	3.9
99	122.5	0.0	5.0	1579.9	0.3	4.5
T						
6	131.4	0.0	5.0	1780.3	0.4	4.2
8	261.8	0.0	5.0	2040.1	0.3	4.2
10	387.9	0.0	5.0	2872.3	0.8	3.7
Average	260.4	0.1	5.0	2230.9	0.5	4.0

The first analysis is based on the size of the scenario set. The B&C hold its superiority in all the scenario set sizes. By increasing the number of scenarios the solution time is increased significantly, but the B&C algorithm finds the optimal solution in most cases in a reasonable amount of time. When we have 1000 scenarios, the extensive form finds the optimal solution in about 25% of the cases and considering the time limit the average gap is about 1.4%. Considering the service level, we can see that increasing the service level results in faster solution time in both methods. In the last analysis, we can see that by increasing the number of periods in the finite-

horizon model, the solution time is also increased, but this increase is such that the B&C method is still a reasonable method to use.

In general, the B&C algorithm has a significantly better performance compared to the extensive form, and hence we use this method for the rest of the experiments. In the following experiments, due to the large number of simulation iterations (4000) the instances are solved with 100 branches. However, based on the reported solution time of the B&C algorithm, we can easily scale up the number of scenario set, when we want to solve the problem for one or couple of stages ahead.

3.6.4 Policy evaluation

We first compare the three policies, namely, average, quantile, and chance-constraint policies against each other. This comparison is based on the objective function and the joint service level, using the procedure explained in section 3.6.2. Table 3.6 compares the three policies using these measures and their 95% confidence interval at two different TBOs and four different service levels. TBO or time between orders is the parameter which defines the trade off between the setup cost and inventory holding cost. Among the three policies the chance-constraint policy is the only policy which respects the service level in all the instances. In all the instances with acceptable service level the chance-constraint policy has the lower cost. Among the three policies the average policy is not sensitive to the service level and it has poor performance in this measure, in which at TBO equal to 1 the joint service level is about 21%. When the TBO is equal to 2 the service level is slightly more than 78% for this policy. This result shows that this policy is not a reliable policy, and it will not be used in the rest of experiments. The quantile policy has an acceptable performance in both measures.

The rest of this section is dedicated to the comparison between the chance-constraint and quantile policies using the additional instances presented in Table 3.3. To this end, we use two measures, the joint service level and the relative cost change, ΔCost, which is defined as:

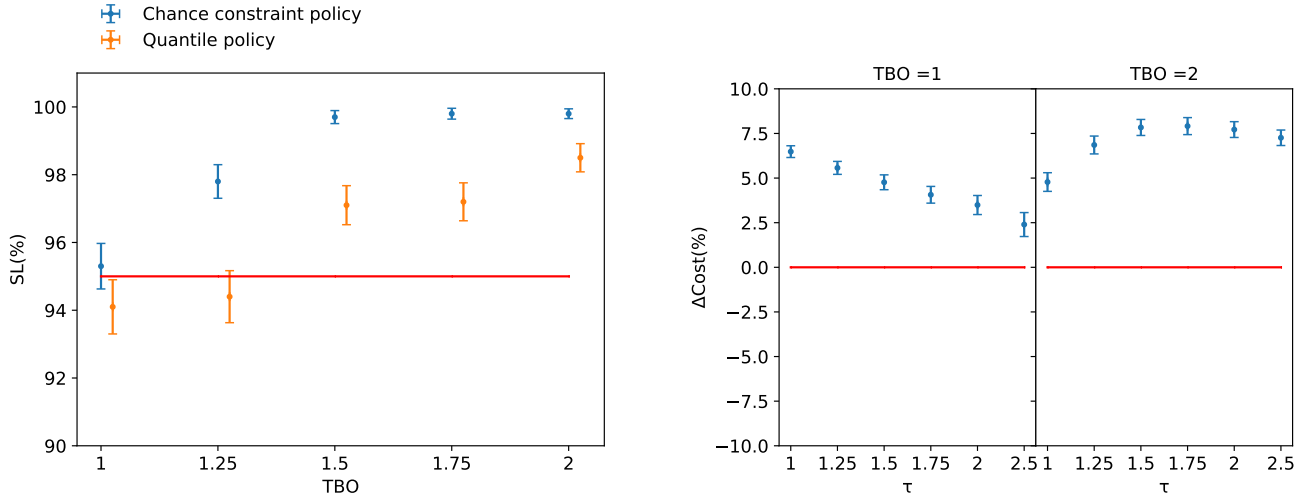
$$\Delta\text{Cost}(\%) = \frac{\text{Total Cost}_{\text{Quantile}} - \text{Total Cost}_{\text{chance-constraint}}}{\text{Total Cost}_{\text{Quantile}}} \times 100 \tag{3.13}$$

Figure 3.4 shows the comparison of the quantile policy and the chance-constraint policy under different values of TBO. Figure 3.4-(a) shows the service level, labeled as SL, and its 95% confidence interval for each of the policies at different values of TBO. Figure 3.4-(b) illustrate the

Table 3.6: Policy comparison based on total cost and service level

TBO	α (%)	Total cost			Service level (%)		
		Average	Quantile	Chance-constraint	Average	Quantile	Chance-constraint
1	80	74.4 ± 0.4	66.4 ± 0.2	66.7 ± 0.2	21.4 ± 1.4	76.5 ± 1.4	84.9 ± 1.3
	90	74.4 ± 0.4	68.2 ± 0.1	67.1 ± 0.2	21.4 ± 1.4	90.6 ± 1.1	90.3 ± 1.1
	95	74.4 ± 0.4	71.0 ± 0.1	67.6 ± 0.2	21.4 ± 1.4	94.1 ± 0.9	95.3 ± 0.7
	99	74.4 ± 0.4	76.1 ± 0.1	69.1 ± 0.2	21.4 ± 1.4	99.1 ± 0.3	99.1 ± 0.3
2	80	204.3 ± 0.6	204.3 ± 0.5	191.2 ± 0.6	78.1 ± 2.7	93.2 ± 1.2	98.7 ± 0.4
	90	204.3 ± 0.6	207.2 ± 0.4	192.4 ± 0.6	78.1 ± 2.7	97.4 ± 0.6	99.5 ± 0.3
	95	204.3 ± 0.6	210.0 ± 0.4	193.5 ± 0.6	78.1 ± 2.7	98.5 ± 0.5	99.8 ± 0.2
	99	204.3 ± 0.6	215.3 ± 0.4	195.4 ± 0.6	78.1 ± 2.7	99.7 ± 0.2	100.0 ± 0.0

ΔCost for each value of TBO. The positive percentages show the superiority of chance constraint policy. In all cases, the chance-constraint policy has a better performance in terms of service level. The chance-constraint policy has a lower cost in all cases in which both policies have an acceptable service levels. When TBO is more than 1 the service level is over satisfied. This is mostly because of the backlog determination step. We later discuss the necessity of this modification and if it imposes any extra costs.

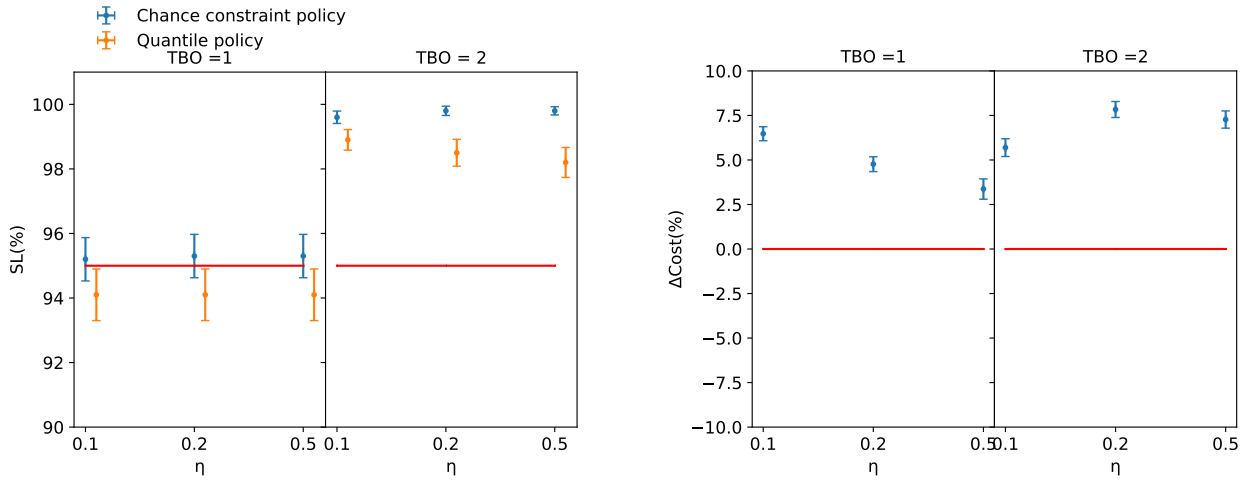


(a) Service level

(b) Total cost

Figure 3.4: Comparison based on TBO

Figure 3.5 shows the comparison based on different values of η under two different values of TBO, 1 and 2. This parameter defines the production cost of different item. Higher η means higher variety in products and the lower ones refer to the situations with higher similarities. When TBO



(a) Service level

(b) Total cost

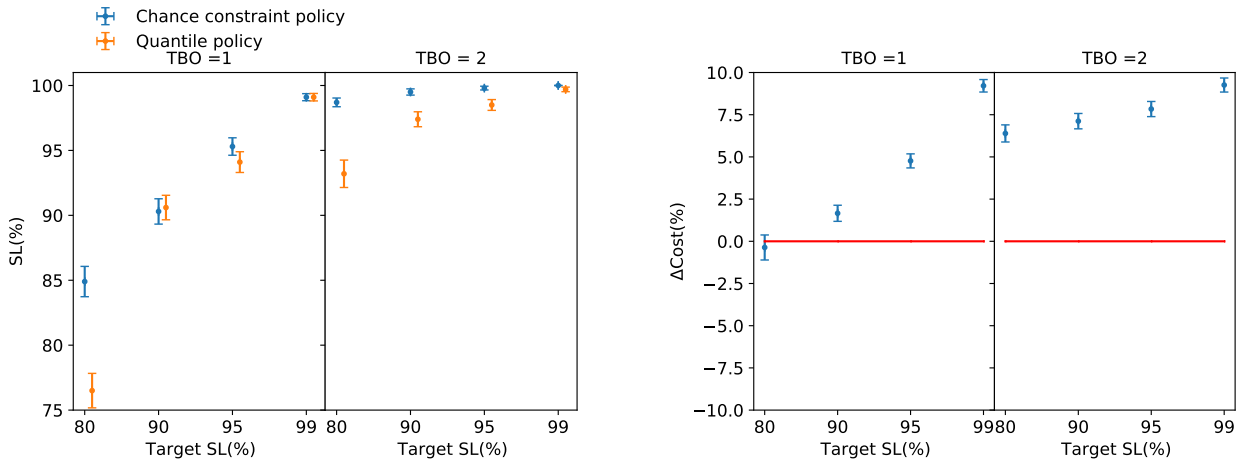
Figure 3.5: Comparison based on η

is equal to 1, quantile policy service level is lower than the target service level. In all cases, the chance-constraint policy has a better performance in terms of joint service level and the total cost.

Figure 3.6 shows the comparison based on different service level values under two different values of TBO, 1 and 2. In all cases, the chance-constraint policy respects the service level and in cases where both policies have acceptable service level, the chance-constraint policy has better performance in terms of the total cost. It should be noted that when the service level increases, the performance of the chance-constraint policy against the quantile policy improves. Figure 3.7 is complementary to Figure 3.6 and illustrates the trend of the total cost for different values of service level. As can be seen in this figure, the total cost of the quantile policy increases significantly with an increase in the target service level, which is not the case in the chance-constraint policy.

Figure 3.8 shows a similar comparison based on τ values. This parameter define the substitution cost based on the difference in the production cost of two products. In all cases, the chance-constraint policy has better performance compared to the quantile policy.

We can conclude that although the quantile policy has an acceptable performance in general and under different parameter settings, the proposed chance-constraint policy has consistent superiority against it. In other words, using the chance-constrained policy results in lower cost, at the same or better service level values.



(a) Service level (b) Total cost

Figure 3.6: Comparison based on the target service level

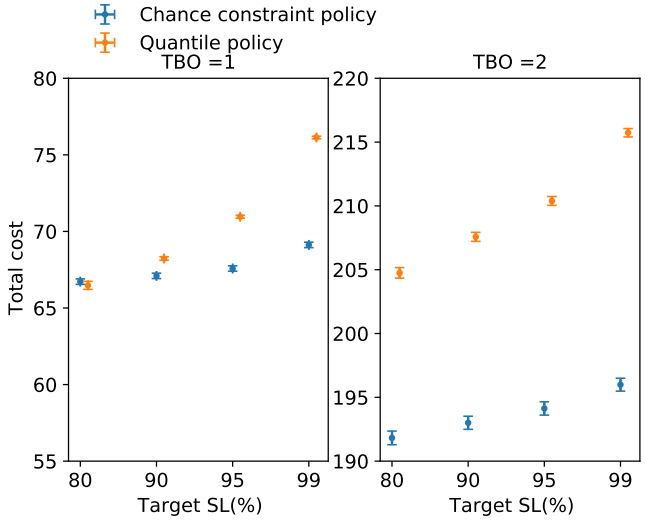
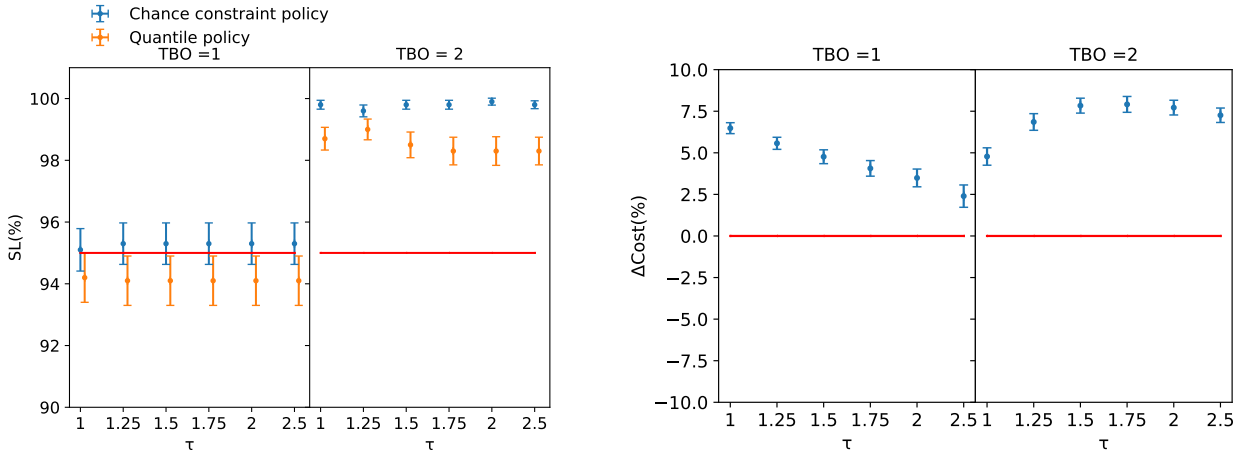


Figure 3.7: Total cost trend comparison based on α

3.6.5 Sensitivity analysis

In this section, we perform some sensitivity analysis for different elements of the cost function, namely the setup cost, inventory holding cost and the substitution cost. Figure 3.9 shows the cost change based on different values for TBO. It is intuitive that in the lot sizing problem, by increase in TBO, there will be an increase in setup cost plus the inventory holding cost. In addition to this increase, we can see a constant increase in the substitution cost, which means an increase in the



(a) Service level

(b) Total cost

Figure 3.8: Comparison based on τ

amount of substitution.

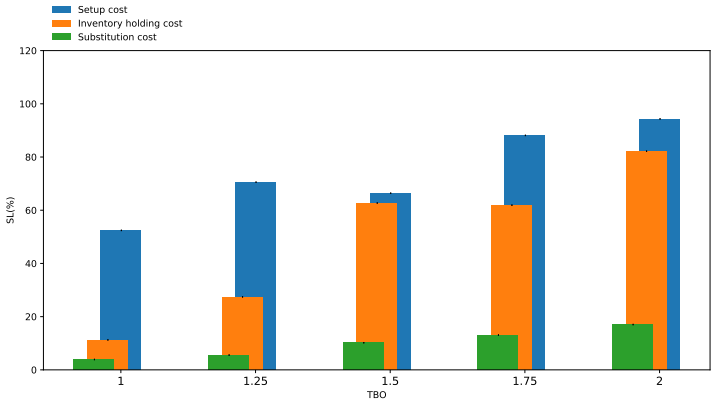


Figure 3.9: Cost analysis based on TBO

Figure 3.10 illustrates different cost changes based on changes in parameter τ for two different TBO values. When TBO is equal to 1, by increasing the substitution cost, the inventory cost slightly increases, and the substitution cost does not increase. We can conclude that by increasing the substitution cost, the amount of inventory will increase and the amount of substitution will decrease. This is more obvious when TBO is equal to 2. In this case, the increase in the substitution cost per unit results in total substitution cost reduction, total holding cost increase, and a slight setup cost increase.

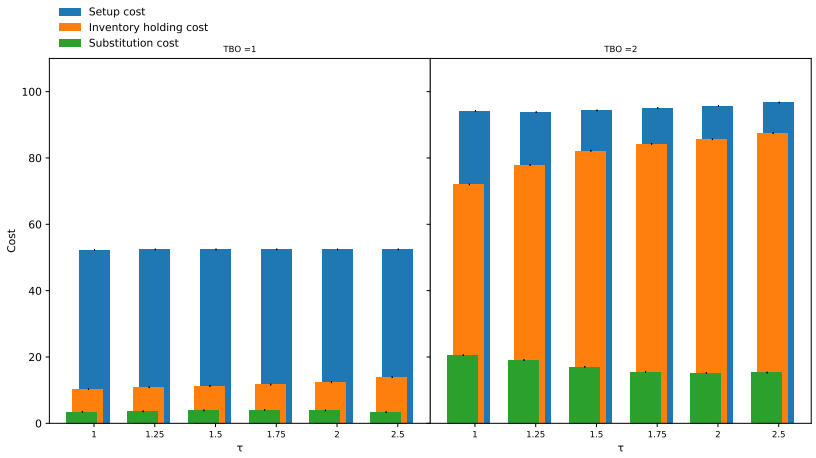
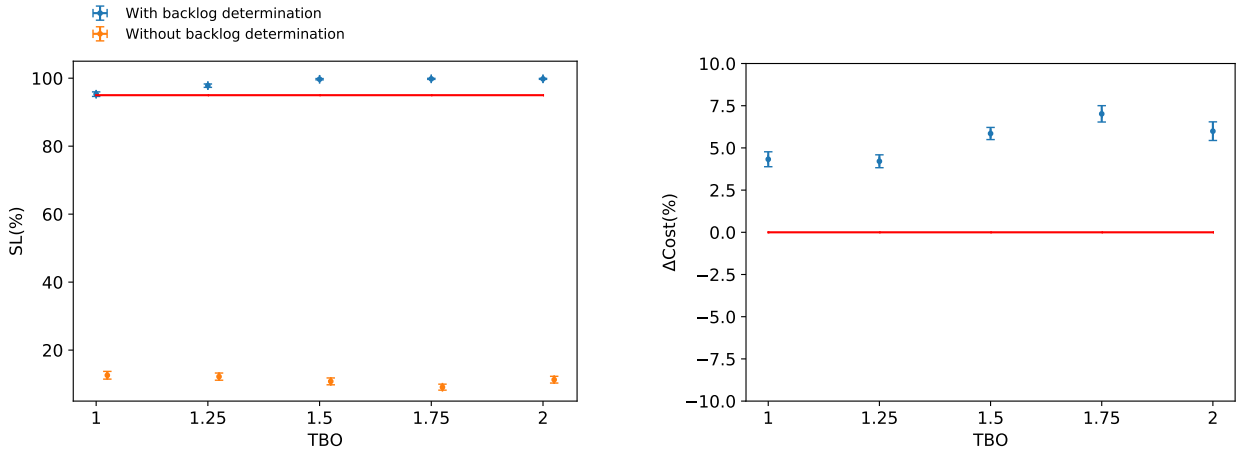


Figure 3.10: Cost analysis based on τ

3.6.6 The necessity of backlog determination step

When TBO is greater than 1 the service level is over satisfied (See Figure 3.4-(a)). This is due to the fact that to save up on the setup cost the production amount will be higher than the average demand of one period. With this higher production level, many of the demands can be satisfied in the current period when we apply the backlog determination step. In this section, we discuss the necessity of the backlog determination step, without which it is not possible to satisfy the target service level. In these experiments, we cancel the backlog determination step, and calculate the total cost and service level. Figure 3.11-(a) illustrates the service level and Figure 3.11-(b) shows the relative cost decrease without and with backlog determination step. Without this step the service level falls under 20%. We see that, even with very low service levels, there is a small cost reduction compared to the case in which the model tries to minimize the backlog in the current stage as much as possible. This means that the over satisfaction of the service level will not impose a huge cost on the model.



(a) Service level

(b) Total cost

Figure 3.11: The necessity of backlog determination

3.6.7 Effect of substitution

In this section, we investigate the effect of substitution. To this end, we run some experiments and eliminate the possibility of substitution. We compare this case with the possibility of partial and full substitution, for different target service levels. In full substitution a product can be substituted by all the higher-grade products. Figure 3.12 illustrates the percentage of cost decrease when adding the possibility of substitution to the model, fully and partially. The value of substitution is more at the higher values of the target service level and larger TBO. We can see that having the option of substitution can result in substantial cost savings, about 7% to 25% in our experiments. We can also see that considering the possibility of full substitution does not result in more cost savings compared to the partial flexibility option.

3.7 Conclusion

We study an infinite-horizon stochastic lot sizing problem with a supplier-driven product substitution option and the service level constraint which is defined jointly over different products. To solve this problem, we consider a finite-horizon version of this problem and apply it in a rolling-horizon framework. We propose different MIP-based policies for decision-making in each period. The mathematical models in these policies are approximations of the finite-horizon multi-stage

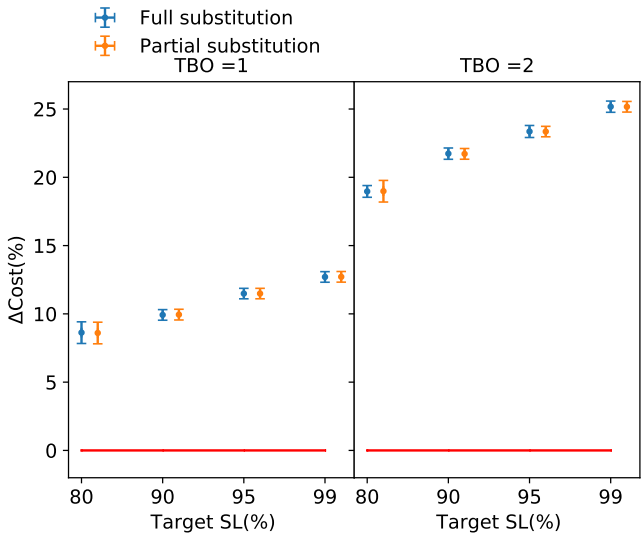


Figure 3.12: Effect of substitution (Relative cost decrease)

model. We propose two deterministic policies and a policy based on the two-stage approximation, namely, the chance-constraint policy. While the deterministic policy models are very efficient to solve, the chance-constraint extensive formulation is very challenging. To solve this model we proposed a B&C algorithm.

To compare different policies and evaluate different solutions, we simulate the rolling framework. The random demand is generated through an autoregressive process and stochasticity is considered as a discrete scenario set. We show that while the deterministic policy based on the quantile value of the scenario set is very efficient, our proposed chance-constraint policy results in more reliable and accurate decisions. In addition, we show that limited levels of substitution possibility results in noticeable cost savings.

References

Akçay, Y., Li, Y., and Natarajan, H. P. (2020). Category inventory planning with service level requirements and dynamic substitutions.

Bassok, Y., Anupindi, R., and Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research*, 47(4):632–642.

Bitran, G. R. and Dasu, S. (1992). Ordering policies in an environment of stochastic yields and substitutable demands. *Operations Research*, 40(5):999–1017.

Bitran, G. R. and Gilbert, S. M. (1994). Co-production processes with random yields in the semiconductor industry. *Operations Research*, 42(3):476–491.

Bitran, G. R. and Leong, T.-Y. (1992). Deterministic approximations to co-production problems with service constraints and random yields. *Management science*, 38(5):724–742.

Bookbinder, J. H. and Tan, J.-Y. (1988). Strategies for the probabilistic lot-sizing problem with service-level constraints. *Management Science*, 34(9):1096–1108.

Chen, B. and Chao, X. (2020). Dynamic inventory control with stockout substitution and demand learning. *Management Science*.

Gicquel, C. and Cheng, Jianqiang, A. (2018). joint chance-constrained programming approach for the single-item capacitated lot-sizing problem with stochastic demand. *Annals of Operations Research*, 264(1-2):123–155.

Guan, Y. (2011). Stochastic lot-sizing with backlogging: computational complexity analysis. *Journal of Global Optimization*, 49(4):651–678.

Guan, Y. and Miller, A. J. (2008). Polynomial-time algorithms for stochastic uncapacitated lot-sizing problems. *Operations Research*, 56(5):1172–1183.

Günlük, O. and Pochet, Y. (2001). Mixing mixed-integer inequalities. *Mathematical Programming*, 90(3):429–457.

Haugen, K. K., Løkketangen, A., and Woodruff, D. L. (2001). Progressive hedging as a meta-heuristic applied to stochastic lot-sizing. *European Journal of Operational Research*, 132(1):116–122.

Helber, S., Sahling, F., and Schimmelpfeng, K. (2013). Dynamic capacitated lot sizing with random demand and dynamic safety stocks. *OR Spectrum*, 35(1):75–105.

Hsu, A. and Bassok, Y. (1999). Random yield and random demand in a production system with downward substitution. *Operations Research*, 47(2):277–290.

Hsu, V. N., Li, C.-L., and Xiao, W.-Q. (2005). Dynamic lot size problems with one-way product substitution. *IIE transactions*, 37(3):201–215.

Jiang, Y., Xu, J., Shen, S., and Shi, C. (2017). Production planning problems with joint service-level guarantee: a computational study. *International Journal of Production Research*, 55(1):38–58.

Lang, J. C. and Domschke, W. (2010). Efficient reformulations for dynamic lot-sizing problems with product substitution. *OR spectrum*, 32(2):263–291.

Liu, X. and Küçükyavuz, Simge, A. (2018). polyhedral study of the static probabilistic lot-sizing problem. *Annals of Operations Research*, 261(1-2):233–254.

Luedtke, J. (2014). A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1):219–244.

Luedtke, J. and Ahmed, Shabbir, A. (2008). sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699.

Lulli, G. and Sen, Suvrajeet, A. (2004). branch-and-price algorithm for multistage stochastic integer programming with application to stochastic batch-sizing problems. *Management Science*, 50(6):786–796.

Nagarajan, M. and Rajagopalan, S. (2008). Inventory models for substitutable products: Optimal policies and heuristics. *Management Science*, 54(8):1453–1466.

Ng, T. S., Fowler, J., and Mok, I. (2012). Robust demand service achievement for the co-production newsvendor. *IIE Transactions*, 44(5):327–341.

Rao, U. S., Swaminathan, J. M., and Zhang, J. (2004). Multi-product inventory planning with downward substitution, stochastic demand and setup costs. *IIE Transactions*, 36(1):59–71.

Shin, H., Park, S., Lee, E., and Benton, W. C., A. (2015). classification of the literature on the planning of substitutable products. *European Journal of Operational Research*, 246(3):686–699.

Tempelmeier, H. (2007). On the stochastic uncapacitated dynamic single-item lotsizing problem with service level constraints. *European Journal of Operational Research*, 181(1):184–194.

Zeppetella, L., Gebennini, E., Grassi, A., and Rimini, B. (2017). Optimal production scheduling with customer-driven demand substitution. *International Journal of Production Research*, 55(6):1692–1706.

Zhang, M., Küçükyavuz, S., and Goel, Saumya, A. (2014). branch-and-cut method for dynamic decision making under joint chance constraints. *Management Science*, 60(5):1317–1333.

General Conclusion

We conclude with a general overview of this study, mentioning the contributions of the thesis and the future research avenues. We study the stochastic lot sizing problem with random demand, where we wish to define the optimal timing and production quantities to minimize the total expected costs while considering a service level to mitigate the risk of stock outs.

Considering the main focus of this research as the stochastic lot sizing problem with service level constraints, we study three extensions of this problem. Our main focus is on multi-product and multi-period problems under various types of service levels. To this end, we investigate different solution methodologies to find the solutions. Strategies in stochastic lot sizing refer to the level of adaptability with respect to adjusting the setup and quantity decisions when the demand is observed. We first present a panoramic view of the three studies and chapters in this thesis which is followed by a more detailed discussion of the contributions of these studies and the notable results and the possible future research.

The first study, presented in Chapter 1, extends the stochastic lot sizing problem with service level constraints by investigating aggregate service levels. The service levels are commonly defined for each product separately, while the aggregate service levels are defined based on the averages over several products. We investigate various service levels in both individual and aggregate versions. We use the static strategy, in which both setups and production decisions are defined at the beginning of the planning horizon, and they remain unchanged even when the demands are observed.

The second research, presented in Chapter 2, is dedicated to the stochastic multi-level capacitated lot sizing problem in which we have a bill-of-material (BOM). The service level used in this problem is a time and quantity-oriented service level. We address a general setting in which, in addition to the end items, the components can also have independent demand. In this research,

we study the value of having an adaptive strategy at different levels of the BOM. In the adaptive strategy, some or all items follow a static-dynamic strategy, in which the production amounts may be updated when the demand is observed and we compare it to the case in which all the items in BOM follow the static strategy.

The third research, presented in Chapter 3, is an extension to our main problem by adding the possibility of product substitution. We consider an infinite time horizon in which different decisions, including production setup, and the amount of production and substitution are dynamically updated when the demand is realized, hence we follow the dynamic strategy. We use the α service level defined jointly over multiple products. More specifically, this study presents the multi-stage stochastic lot sizing problem with substitution and a joint service level.

Contributions

This research presents several contributions in operations and inventory management. In this thesis, we introduce several extensions to the stochastic lot sizing problem with service level constraints, which can be considered as the main contribution of this thesis. These extensions investigate the value of adding various types of flexibility, which are interesting both scholastically and in practice. This thesis provides mathematical models and solution methodologies for these new problems and their approximations. Finally, this thesis provides extensive numerical experiments by generating many test instances, simulation, and sensitivity analyses to test and evaluate the methodologies and derive managerial insights. The detailed contributions of this thesis are as follows.

I. Investigating the value of aggregate service levels in stochastic lot sizing

Most of the research in the literature of stochastic lot sizing problems with service level constraints consider the service levels individually for each SKU. We address this gap in the literature, by investigating the value of aggregate service level. Through extensive numerical experiments, we show that the aggregate service levels will provide some flexibility which results in cost reductions, as opposed to the traditional service levels imposed independently on each item. This cost reduction varies based on the type of service level and under different parameters settings. Using the same scenario set for all the variations of service levels, we provide a fair comparison environment, which helps the decision makers to choose an appropriate service level and assign a

reasonable value to that. The results show that, at the same value for all the service levels, considering the aggregate version of the δ service level provides the highest cost decrease, and for the α service level the decrease is at its lowest value. In general, a lower service level, higher variance in random demand, higher capacity, and higher variability in the holding cost will increase the benefit of an aggregate service level compared to separate service levels. Using a tight aggregate service level in combination with less strict minimum individual service levels, is a reasonable approach which will enable companies to benefit from the value of aggregate service level, while guaranteeing a minimum service level for each individual product.

II. Investigating the value of a static strategy in a receding horizon framework

In the first research, we choose to use the static strategy due to the capacity constraint. One of the disadvantages of the static strategy is the lack of responsiveness to the demand realization which leads to large inventory levels and costs in the system. We show that when some levels of planning flexibility is allowed in the system, we can still use static models in a rolling/receding horizon environment to overcome this inherent limitation. In this research, we illustrate that this implementation leads to a plan with lower inventory levels and total cost.

III. Investigating the value of flexibility in stochastic multi-level lot sizing

We consider a stochastic multi-level lot sizing problem, and we investigate the value of having an adaptive strategy compared to a fully static strategy for all the items. In the adaptive strategy, the production level for some of the items in the BOM can be updated when the demand is observed. Three BOM structures, serial, assembly, and general, are considered, and we also address a more general setting in which in addition to the end items some of the components may also have independent demand. We numerically show that adding flexibility to the system by applying an adaptive strategy results in cost savings depending on where we add the flexibility in the BOM. Adding the flexibility only for the end item will result in about 20% of cost savings. Extensive numerical experiments show that the cost savings depend on some parameters, such as the service level, holding cost structure in the BOM, and time between orders. In situations with a higher service level and lower time between orders, we can benefit more by adding some levels of flexibility in the system. While controlling the variation in the plans is very important in the multi-level system, this research shows that even having a small degree of flexibility whenever possible may result in a significant amount of cost savings.

IV. Investigating the value of substitution in stochastic lot sizing

We study an infinite-horizon stochastic lot sizing problem with a supplier-driven product substitution option and a service level constraint which is defined jointly over multiple products. To solve this problem, we consider a finite-horizon version of this problem and apply it in a rolling horizon framework. We propose MIP-based policies for decision-making in each period. The mathematical models in these policies are approximations of the finite-horizon multi-stage model, in which the stochasticity is represented in a scenario set. We propose two deterministic policies and a policy based on the two-stage approximation, namely, the chance constraint policy. While the deterministic policy models are very efficient to solve, the chance constraint extensive formulation is very challenging. To compare the policies and evaluate the solutions, we simulate the rolling framework. The random demand is generated through an autoregressive process. We show that while the deterministic policy based on the quantile value of the scenario set is very efficient with respect to the solution time, the proposed chance constraint policy results in more reliable and accurate decisions. In addition, we show that the substitution possibility results in noticeable cost savings of about 10% to 25% at high service levels.

V. Solution methodologies

The stochastic nature of the problem in this thesis makes the proposed mathematical models challenging to solve. The problems in the first two studies, i.e., the aggregate service level and stochastic multi-level lot sizing, are modeled as two-stage stochastic programming problems. In the first research, due to the non-linearity of the objective function, we apply piece-wise linear approximations. In the second research, the stochasticity is represented as a scenario set and we apply the sample average approximation technique. In the third research where we investigate the product substitution option, we model the problem as a multi-stage stochastic problem, in which the random demand is represented as a scenario set. Different types of policies are used to make the decisions at each stage. For the policy which explicitly considers the service level, we apply a branch-and-cut algorithm to solve the extensive form of the joint service level constraint. The branch-and-cut algorithm finds the solution in a significantly lower time compared to the extensive form.

VI. Extensive numerical experiments

All three studies include extensive numerical experiments, from instance generation to sensi-

tivity analysis. As the problems are new to the literature, we generate new instances or modify existing data sets for lot sizing problems to be able to address and evaluate new assumptions. One of the unique strengths of the first study is that all the service levels in aggregate and individual versions under both static and receding horizon implementations are compared under the same set of scenarios, which enables a fair comparison between them. In the second research, we address different demand profiles, different levels of flexibility, and different parameter settings under three BOM structures. For each structure, we test more than 1000 instances at each flexibility level. In the third research, we generate new test instances by modifying and merging related instances. We apply and evaluate different policies in a rolling horizon framework to simulate the dynamic decision-making and derive acceptable confidence intervals for the performance measures. We perform extensive sensitivity analyses to show the superiority of our proposed policy under various settings.

Future work

We will conclude this research by highlighting some potential future research directions. We classify potential extensions to the current research into two main categories, i.e., modelling perspective, and solution techniques, and present them separately as follows:

I. Modeling perspective

All the studied problems in this research were exploring new problems as an extension to the lot sizing problem. Therefore, adding transportation decisions such as routing to these problems is a new research avenue that needs further investigation. Except for the first research which addresses different types of service levels, the last two studies consider one type of service level. Investigating other types of service levels for the last two problems is another stream for future research. For example, considering the α service level, which imposes limits on the probability of stock-outs, for stochastic multi-level lot sizing is one of such streams, which may result in new insights and is worthwhile to investigate.

II. Solution techniques

Stochastic lot sizing problems are challenging to solve and increasing the size of the problems requires faster algorithms to find a high-quality solution in a reasonable amount of time. Applying decomposition-based algorithms such as the ones developed for stochastic programming problems and proposing heuristic algorithms for larger size instances of the mentioned problems are interesting research avenues.

As shown in the first research, it is not possible to apply the same methodology for problems with various types of service levels. Considering different types of service levels for the mentioned problem does not only change the problem formulation and insights but also requires different solution methodologies.

Bibliography

Afentakis, P. and Gavish, B. (1986). Optimal lot-sizing algorithms for complex product structures. *Operations Research*, 34(2):237–249.

Afentakis, P., Gavish, B., and Karmarkar, U. (1984). Computationally efficient optimal solutions to the lot-sizing problem in multistage assembly systems. *Management Science*, 30(2):222–239.

Akartunalı, K. and Miller, Andrew J., A. (2009). heuristic approach for big bucket multi-level production planning problems. *European Journal of Operational Research*, 193(2):396–411.

Akçay, A., Biller, B., and Tayur, S. R. (2016). *Beta-Guaranteed aggregate Service Levels*. Available at SSRN.

Akçay, Y., Li, Y., and Natarajan, H. P. (2020). Category inventory planning with service level requirements and dynamic substitutions.

Almeder, C. (2010). A hybrid optimization approach for multi-level capacitated lot-sizing problems. *European Journal of Operational Research*, 200(2):599–606.

Alvarez, A., Cordeau, J.-F., Jans, R., Munari, P., and Morabito, R. (2020). Inventory routing under stochastic supply and demand. *Omega*, pages 102–304.

Bassok, Y., Anupindi, R., and Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research*, 47(4):632–642.

Ben-Tal, A., Laurent, E. G., and Arkadi, N. (2009). *Robust Optimization*. Prinston University Press.

Bitran, G. R. and Dasu, S. (1992). Ordering policies in an environment of stochastic yields and substitutable demands. *Operations Research*, 40(5):999–1017.

Bitran, G. R. and Gilbert, S. M. (1994). Co-production processes with random yields in the semiconductor industry. *Operations Research*, 42(3):476–491.

Bitran, G. R. and Leong, T.-Y. (1992). Deterministic approximations to co-production problems with service constraints and random yields. *Management science*, 38(5):724–742.

Bookbinder, J. H. and Tan, J.-Y. (1988). Strategies for the probabilistic lot-sizing problem with service-level constraints. *Management Science*, 34(9):1096–1108.

Brahimi, N., Absi, N., Dauzère-Pérès, S., and Nordli, A. (2017). Single-item dynamic lot-sizing problems: An updated survey. *European Journal of Operational Research*, 263(3):838–863.

Chen, B. and Chao, X. (2020). Dynamic inventory control with stockout substitution and demand learning. *Management Science*.

Contreras, I., Cordeau, J.-F., and Laporte, G. (2011). Stochastic uncapacitated hub location. *European Journal of Operational Research*, 212(3):518–528.

Dural-Selcuk, G., Rossi, R., Kilic, O. A., and Tarim, S. A. (2019). The benefit of receding horizon control: Near-optimal policies for stochastic inventory control. *Omega*, 97(10209):1.

Escalona, P., Angulo, A., Weston, J., Stegmaier, R., and Kauak, I. (2019). On the effect of two popular service-level measures on the design of a critical level policy for fast-moving items. *Computers & Operations Research*, 107:107–126.

Gade, D. and Küçükyavuz, S. (2013). Formulations for dynamic lot sizing with service levels. *Naval Research Logistics*, 60(2):87–101.

Gicquel, C. and Cheng, Jianqiang, A. (2018). joint chance-constrained programming approach for the single-item capacitated lot-sizing problem with stochastic demand. *Annals of Operations Research*, 264(1-2):123–155.

Gruson, M., Cordeau, J.-F., and Jans, R. (2018). The impact of service level constraints in deterministic lot sizing with backlogging. *Omega*, 79:91–103.

Gruson, M., Cordeau, J.-F., and Jans, R. (2021). Benders decomposition for a stochastic three-level lot sizing and replenishment problem with a distribution structure. *European Journal of Operational Research*, 291(1):206–217.

Guan, Y. (2011). Stochastic lot-sizing with backlogging: computational complexity analysis. *Journal of Global Optimization*, 49(4):651–678.

Guan, Y. and Miller, A. J. (2008). Polynomial-time algorithms for stochastic uncapacitated lot-sizing problems. *Operations Research*, 56(5):1172–1183.

Günlük, O. and Pochet, Y. (2001). Mixing mixed-integer inequalities. *Mathematical Programming*, 90(3):429–457.

Haugen, K. K., Løkketangen, A., and Woodruff, D. L. (2001). Progressive hedging as a meta-heuristic applied to stochastic lot-sizing. *European Journal of Operational Research*, 132(1):116–122.

Helber, S., Sahling, F., and Schimmelpfeng, K. (2013). Dynamic capacitated lot sizing with random demand and dynamic safety stocks. *OR Spectrum*, 35(1):75–105.

Hsu, A. and Bassok, Y. (1999). Random yield and random demand in a production system with downward substitution. *Operations Research*, 47(2):277–290.

Hsu, V. N., Li, C.-L., and Xiao, W.-Q. (2005). Dynamic lot size problems with one-way product substitution. *IIE transactions*, 37(3):201–215.

Hung, Y.-F. and Chien, Kuo-Liang, A. (2000). multi-class multi-level capacitated lot sizing model. *Journal of the Operational Research Society*, 51(11):1309–1318.

Jans, R. and Degraeve, Z. (2008). Modeling industrial lot sizing problems: a review. *International Journal of Production Research*, 46(6):619–1643.

Jiang, Y., Xu, J., Shen, S., and Shi, C. (2017). Production planning problems with joint service-level guarantee: a computational study. *International Journal of Production Research*, 55(1):38–58.

Kelle, P. (1989). ,optimal service levels in multi-item inventory systems. *Engineering Costs and Production Economics*, 15:375–379.

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

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

Lang, J. C. and Domschke, W. (2010). Efficient reformulations for dynamic lot-sizing problems with product substitution. *OR spectrum*, 32(2):263–291.

Liu, X. and Küçükyavuz, Simge, A. (2018). polyhedral study of the static probabilistic lot-sizing problem. *Annals of Operations Research*, 261(1-2):233–254.

Luedtke, J. (2014). A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1):219–244.

Luedtke, J. and Ahmed, Shabbir, A. (2008). sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699.

Lulli, G. and Sen, Suvrajeet, A. (2004). branch-and-price algorithm for multistage stochastic integer programming with application to stochastic batch-sizing problems. *Management Science*, 50(6):786–796.

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

Mousavi, K., Bodur, M., and Roorda, M. (2021). Stochastic last-mile delivery with crowd-shipping and mobile depots.

Mula, J., Poler, R., Garcia-Sabater, J. P., and Lario, F. C. (2006). Models for production planning under uncertainty: A review. *International Journal of Production Economics*, 103(1):271–285.

Nagarajan, M. and Rajagopalan, S. (2008). Inventory models for substitutable products: Optimal policies and heuristics. *Management Science*, 54(8):1453–1466.

Ng, T. S., Fowler, J., and Mok, I. (2012). Robust demand service achievement for the co-production newsvendor. *IIE Transactions*, 44(5):327–341.

Pochet, Y. and Wolsey, L. A. (2006). *Production Planning by Mixed Integer Programming*. Springer, Science & Business Media, New York.

Quezada, F., Gicquel, C., Kedad-Sidhoum, S., and Vu, Dong Quan, A. (2020). multi-stage stochastic integer programming approach for a multi-echelon lot-sizing problem with returns and lost sales. *Computers & Operations Research*, 116(10486):5.

Rao, U. S., Swaminathan, J. M., and Zhang, J. (2004). Multi-product inventory planning with downward substitution, stochastic demand and setup costs. *IIE Transactions*, 36(1):59–71.

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

Rossi, R., Tarim, S. A., Prestwich, S., and Hnich, B. (2014). Piecewise linear lower and upper bounds for the standard normal first order loss function. *Applied Mathematics and Computation*, 231:489–502.

Sahling, F., Buschkühl, L., Tempelmeier, H., and Helber, S. (2009). Solving a multi-level capacitated lot sizing problem with multi-period setup carry-over via a fix-and-optimize heuristic. *Computers & Operations Research*, 36(9):2546–2553.

Seeanner, F., Almada-Lobo, B., and Meyr, H. (2013). Combining the principles of variable neighborhood decomposition search and the fix&optimize heuristic to solve multi-level lot-sizing and scheduling problems. *Computers & Operations Research*, 40(1):303–317.

Sereshti, N., Adulyasak, Y., and Jans, R. (2020). The value of aggregated service levels in stochastic lot sizing problems. *Omega*, 102335.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

Shin, H., Park, S., Lee, E., and Benton, W. C., A. (2015). classification of the literature on the planning of substitutable products. *European Journal of Operational Research*, 246(3):686–699.

Shivsharan, C. T. (2012). *Optimizing the Safety Stock Inventory Cost Under Target Service Level Constraints*. Master of science), University of Massachusetts Amherst.

Stadtler, H. (2003). Multilevel lot sizing with setup times and multiple constrained resources: Internally rolling schedules with lot-sizing windows. *Operations Research*, 51(3):487–502.

Stadtler, H. and Meistering, M. (2019). Model formulations for the capacitated lot-sizing problem with service-level constraints. *OR Spectrum*, 41(4):1025–1056.

Tarim, S. A. and Kingsman, B. G. (2004). The stochastic dynamic production/inventory lot-sizing problem with service-level constraints. *International Journal of Production Economics*, 88(1):105–119.

Taş, D., Gendreau, M., Jabali, O., and Jans, Raf, A. (2019). capacitated lot sizing problem with stochastic setup times and overtime. *European Journal of Operational Research*, 273(1):146–159.

Tavaghof-Gigloo, D. and Minner, S. (2021). Planning approaches for stochastic capacitated lot-sizing with service level constraints. *International Journal of Production Research*, 59(17):5087–5107.

Tempelmeier, H. (2007). On the stochastic uncapacitated dynamic single-item lotsizing problem with service level constraints. *European Journal of Operational Research*, 181(1):184–194.

Tempelmeier, H. (2011). A column generation heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. *Omega*, 39(6):627–633.

Tempelmeier, H. (2013). Stochastic lot sizing problems. *Handbook of Stochastic Models and Analysis of Manufacturing System Operations*, pages 313–344.

Tempelmeier, H. and Derstroff, Matthias, A. (1996). Lagrangean-based heuristic for dynamic multilevel multiitem constrained lotsizing with setup times. *Management Science*, 42(5):738–757.

Tempelmeier, H. and Herpers, Sascha, A. B. C. (2010). β–a heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. *International Journal of Production Research*, 48(17):5181–5193.

Tempelmeier, H. and Herpers, S. (2011). Dynamic uncapacitated lot sizing with random demand under a fillrate constraint. *European Journal of Operational Research*, 212(3):497–507.

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

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

Thevenin, S., Adulyasak, Y., and Cordeau, J.-F. (2021). Material requirements planning under demand uncertainty using stochastic optimization. *Production and Operations Management*, 30(2):475–493.

Toledo, C. F. M. and da Silva Arantes (2015). Márcio and hossomi, marcelo yukio bressan and França, paulo morelato and akartunalı, kerem, a relax-and-fix with fix-and-optimize heuristic applied to multi-level lot-sizing problems. *Journal of Heuristics*, 21(5):687–717.

Tunc, H., Kilic, O. A., Tarim, S. A., and Eksioglu, Burak, A. (2014). reformulation for the stochastic lot sizing problem with service-level constraints. *Operations Research Letters*, 42(2):161–165.

Tunc, H., Kilic, O. A., Tarim, S. A., and Eksioglu, B. A. (2013). simple approach for assessing the cost of system nervousness. *International Journal of Production Economics*, 141(2):619–625.

Tunc, H., Kilic, O. A., Tarim, S. A., and Rossi, R. (2018). An extended mixed-integer programming formulation and dynamic cut generation approach for the stochastic lot-sizing problem. *INFORMS Journal on Computing*, 30(3):492–506.

Van Pelt, T. D. and Fransoo, Jan C., A. (2018). note on “linear programming models for a stochastic dynamic capacitated lot sizing problem". *Computers & Operations Research*, 89:13–16.

Verweij, B., Ahmed, S., Kleywegt, A. J., Nemhauser, G., and Shapiro, A. (2003). The sample average approximation method applied to stochastic routing problems: a computational study. *Computational optimization and applications*, 24(2):289–333.

Wagner, S. M. and Lindemann, Eckhard, A. (2008). case study-based analysis of spare parts management in the engineering industry. *Production Planning & Control*, 19(4):397–407.

Wu, T., Shi, L., Geunes, J., and Akartunalı, K. (2011). An optimization framework for solving capacitated multi-level lot-sizing problems with backlogging. *European Journal of Operational Research*, 214(2):428–441.

Xiao, Y., Zhang, R., Zhao, Q., Kaku, I., and Xu, Yuchun, A. (2014). variable neighborhood search with an effective local search for uncapacitated multilevel lot-sizing problems. *European Journal of Operational Research*, 235(1):102–114.

You, M., Xiao, Y., Zhang, S., Zhou, S., Yang, P., and Pan, X. (2019). Modeling the capacitated multi-level lot-sizing problem under time-varying environments and a fix-and-optimize solution approach. *Entropy*, 21(4):377.

Zeppetella, L., Gebennini, E., Grassi, A., and Rimini, B. (2017). Optimal production scheduling with customer-driven demand substitution. *International Journal of Production Research*, 55(6):1692–1706.

Zhang, M., Küçükyavuz, S., and Goel, Saumya, A. (2014). branch-and-cut method for dynamic decision making under joint chance constraints. *Management Science*, 60(5):1317–1333.

Appendix A – Proofs for service level weights

In this appendix we prove that for specific weights, the weighted sum of the individual service levels is equal to the aggregate service level.

For the β service level the Left Hand Side (LHS) of the individual service level (β_k) constraint (See Table 1.2) is $\frac{\sum_{t\in T} E[\overline{BO}_{kt}]}{\sum_{t\in T} E[\overline{D}_{kt}]}$. When we take the weighted sum of these individual LHSs using the proposed weights of $w_k = \frac{\sum_{t\in T} E[\overline{D}_{kt}]}{\sum_{t\in T}\sum_{l\in K} E[\overline{D}_{lt}]}$, we obtain:

$\sum_{k\in K}(\frac{\sum_{t\in T} E[\overline{D}_{kt}]}{\sum_{t\in T}\sum_{l\in K} E[\overline{D}_{lt}]}\frac{\sum_{t\in T} E[\overline{BO}_{kt}]}{\sum_{t\in T} E[\overline{D}_{kt}]}) = \frac{\sum_{k\in K}\sum_{t\in T} E[\overline{BO}_{kt}]}{\sum_{t\in T}\sum_{l\in K} E[\overline{D}_{lt}]}$. The latter term is the LHS of the aggregate β service level constraint (See Table 1.2).

For the γ service level the LHS of the individual service level (γ_k) constraints (See Table 1.2) is $\frac{\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T} E[\overline{D}_{kt}]}$. When we take the weighted sum of these individual LHSs using the proposed weights of $w_k = \frac{\sum_{t\in T} E[\overline{D}_{kt}]}{\sum_{t\in T}\sum_{l\in K} E[\overline{D}_{lt}]}$, we obtain:

$\sum_{k\in K}(\frac{\sum_{t\in T} E[\overline{D}_{kt}]}{\sum_{t\in T}\sum_{l\in K} E[\overline{D}_{lt}]}\frac{\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T} E[\overline{D}_{kt}]}) = \frac{\sum_{k\in K}\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{l\in K} E[\overline{D}_{lt}]}$. The latter term is the LHS of the aggregate γ service level constraint (See Table 1.2).

For the δ service level the LHS of the individual service level (δ_k) constraints (See Table 1.2) is $\frac{\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T}(T-t+1)E[\overline{D}_{kt}]}$. When we take the weighted sum of these individual LHSs using the proposed weights of $w_k = \frac{\sum_{t\in T}(T-t+1)E[\overline{D}_{kt}]}{\sum_{t\in T}\sum_{l\in K}(T-t+1)E[\overline{D}_{lt}]}$, we obtain:

$\sum_{k\in K}(\frac{\sum_{t\in T}(T-t+1)E[\overline{D}_{kt}]}{\sum_{t\in T}\sum_{l\in K}(T-t+1)E[\overline{D}_{lt}]}\frac{\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T}(T-t+1)E[\overline{D}_{kt}]}) = \frac{\sum_{k\in K}\sum_{t\in T} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{l\in K}(T-t+1)E[\overline{D}_{lt}]}$. The latter term is the LHS of the aggregate δ service level constraint (See Table 1.2).

For the β_p service level the LHS of the individual service level (β_{pk}) constraints (See Table 1.2)

is $\frac{E[\overline{BO}_{kt}]}{E[\overline{D}_{kt}]}$. When we take the weighted sum of these individual LHSs using the proposed weights of $w_k = \frac{E[\overline{D}_{kt}]}{\sum_{l\in K} E[\overline{D}_{lt}]}$, we obtain:

$\sum_{k\in K}(\frac{E[\overline{D}_{kt}]}{\sum_{l\in K} E[\overline{D}_{lt}]}\frac{E[\overline{BO}_{kt}]}{E[\overline{D}_{kt}]}) = \frac{\sum_{k\in K} E[\overline{BO}_{kt}]}{\sum_{l\in K} E[\overline{D}_{lt}]}$. The latter term is the LHS of the aggregate β_p service level constraints (See Table 1.2).

For the γ_p service level the LHS of the individual service level (γ_{pk}) constraints (See Table 1.2) is $\frac{E[\overline{B}_{kt}]}{E[\overline{D}_{kt}]}$. When we take the weighted sum of these individual LHSs using the proposed weights of $w_k = \frac{E[\overline{D}_{kt}]}{\sum_{l\in K} E[\overline{D}_{lt}]}$, we obtain:

$\sum_{k\in K}(\frac{E[\overline{D}_{kt}]}{\sum_{l\in K} E[\overline{D}_{lt}]}\frac{E[\overline{B}_{kt}]}{E[\overline{D}_{kt}]}) = \frac{\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{l\in K} E[\overline{D}_{lt}]}$. The latter term is the LHS of the aggregate γ_p service level constraints (See Table 1.2).

For the δ_p service level the LHS of the individual service level (δ_{pk}) constraints (See Table 1.2) is $\frac{E[\overline{B}_{kt}]}{\sum_{j=1}^{t} E[\overline{D}_{kj}]}$. When we take the weighted sum of these individual LHSs using the proposed weights of $w_{kt} = \frac{\sum_{j=1}^{t} E[\overline{D}_{kj}]}{\sum_{l\in K}\sum_{j=1}^{t} E[\overline{D}_{lj}]}$, we obtain:

$\sum_{k\in K}(\frac{\sum_{j=1}^{t} E[\overline{D}_{kj}]}{\sum_{l\in K}\sum_{j=1}^{t} E[\overline{D}_{lj}]}\frac{E[\overline{B}_{kt}]}{\sum_{j=1}^{t} E[\overline{D}_{kj}]}) = \frac{\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{l\in K}\sum_{j=1}^{t} E[\overline{D}_{lj}]}$. The latter term is the LHS of the aggregate δ_{pk} service level constraints (See Table 1.2).

Appendix B – Aggregate service level for product families

It is possible to define different aggregate service levels for different product families. In this case, the service levels are aggregately defined over all products within a family of products. Table 1 shows the new sets and parameters for these models.

Table 1: Parameters and decision variables of the models with aggregate service level over product family

Sets	
F	Set of product families
K_f	Set of products in product family f, $K_f \in K$
Parameters	
α_{cf}^{agg}	Target aggregate α_c service level for product family f
β_f	Target fill rate as an aggregate service level for product family f
γ_f	Target aggregate γ service level for product family f
δ_f	Target aggregate δ service level for product family f

$$\frac{\sum_{t\in T}\sum_{k\in K_f} E[\overline{BO}_{kt}]}{\sum_{t\in T}\sum_{k\in K_f} E[\overline{D}_{kt}]} \leq 1-\beta_f \qquad \forall f \in F \quad (14)$$

$$\frac{\sum_{t\in T}\sum_{k\in K_f} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K_f} E[\overline{D}_{kt}]} \leq 1-\gamma_f \qquad \forall f \in F \quad (15)$$

$$\frac{\sum_{t\in T}\sum_{k\in K_f} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K_f} (T-t+1)E[\overline{D}_{kt}]} \leq 1-\delta_f \qquad \forall f \in F \quad (16)$$

$$\sum_{k\in K_f} w_k \min_{t\in T}(pr(I_{k0}+\sum_{j=1}^{t}(x_{kj}-\overline{D}_{kj})\geq 0)) \geq \alpha_{cf}^{agg} \qquad \forall f \in F \quad (17)$$

The objective function and all the constraints except the aggregate service level constraints of all the models will remain the same. In the case of product families, for the models with β, γ,

δ, and α_c service levels the constraints (1.6), (1.25), (1.27), and (1.34) will change to constraints (14), (15), (16), and (17), respectively.

Appendix C – Sensitivity analysis

In this appendix the diagrams of sensitivity analysis for the $\beta, \beta_p, \delta,$ and δ_p service levels are presented. The β and δ service levels have similar trends as γ service level. β_p and δ_p service levels are similar to γ_p service level which was explained in the section of sensitivity analysis. Despite these similarities, the value of $\Delta Cost$ differ for different types of service level. In general, at the same service level, the $\Delta Cost$ has its highest value for the δ_p and its lowest value for the γ service level, if the models are feasible. What is common in all the diagrams is that the $\Delta Cost$ is more sensitive to the holding cost and it increases when the variation in holding cost increases. In addition to that, in all the diagrams, the $\Delta Cost$ decreases when the service level increases.

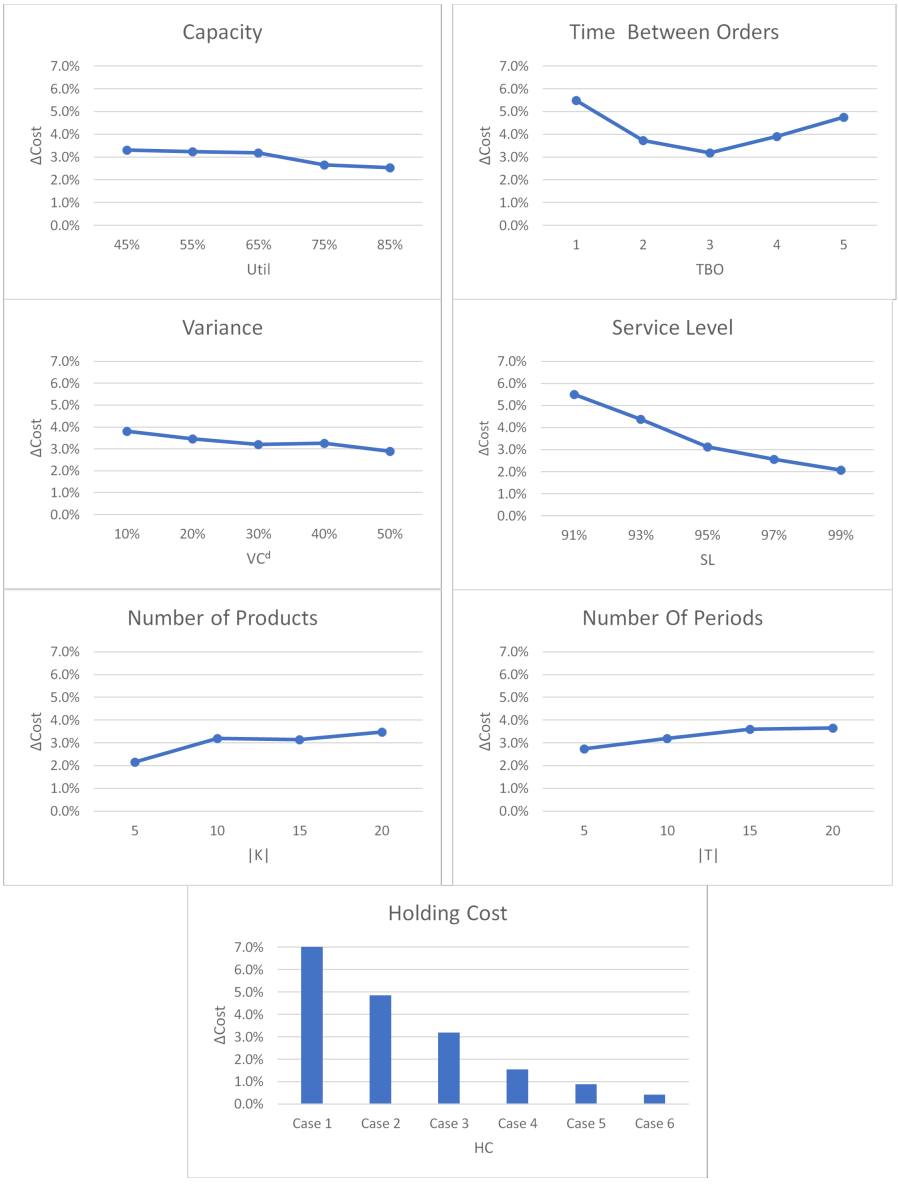


Figure 1: Sensitivity analysis plots for β service level

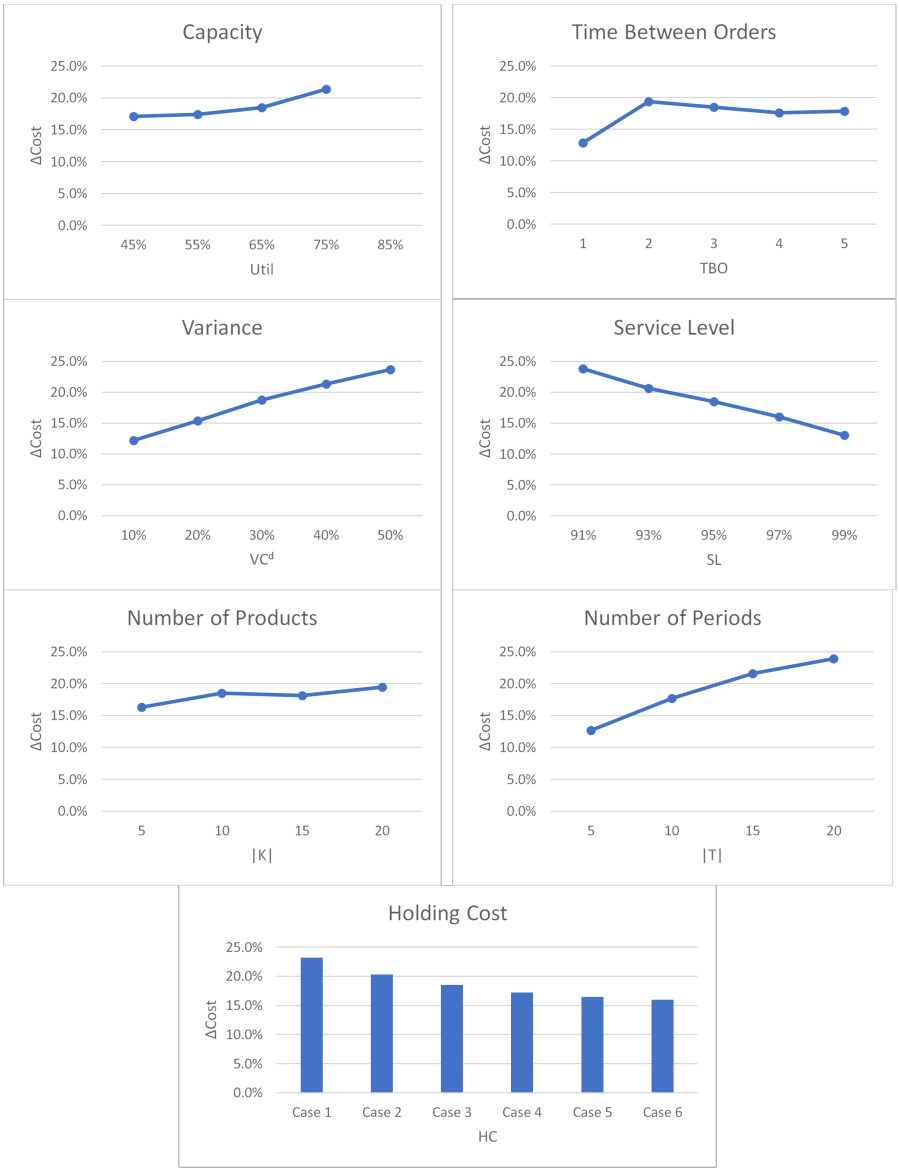


Figure 2: Sensitivity analysis plots for β_p service level

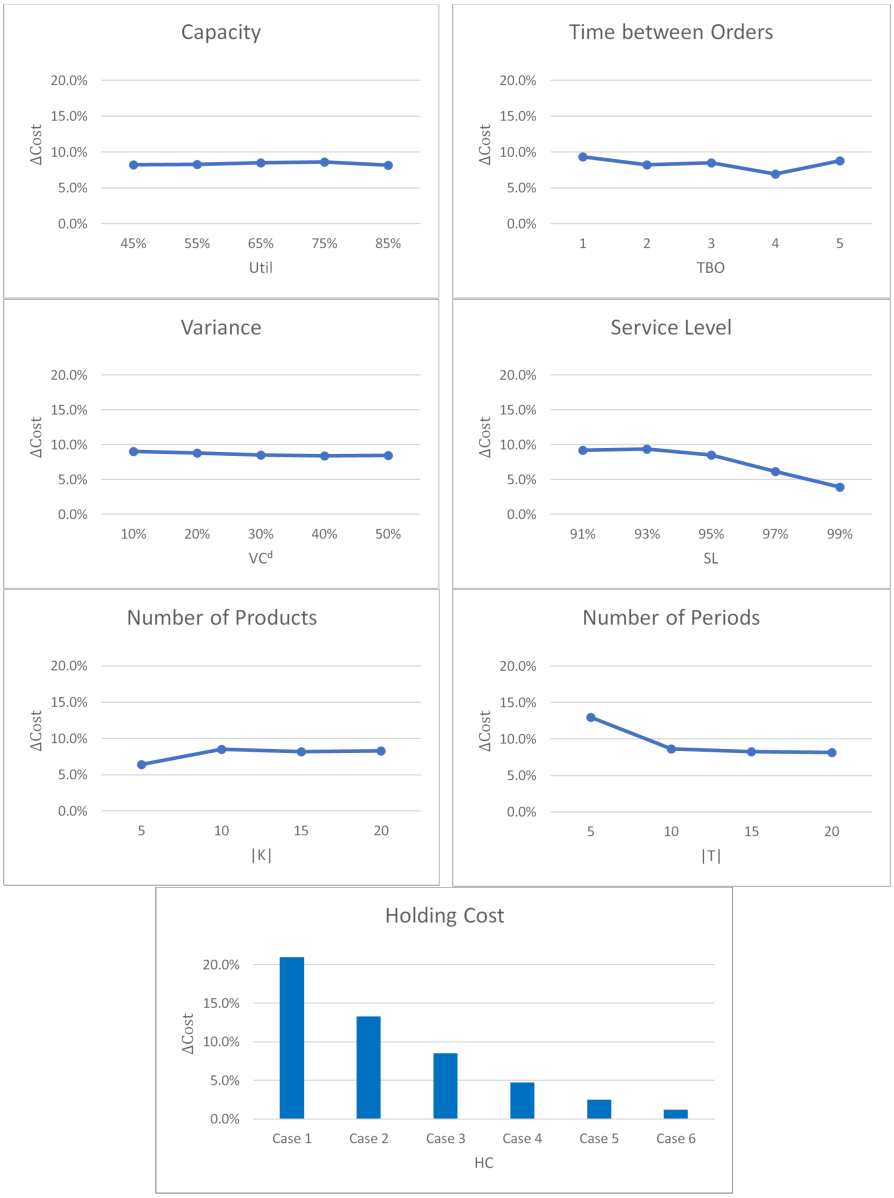


Figure 3: Sensitivity analysis plots for δ service level

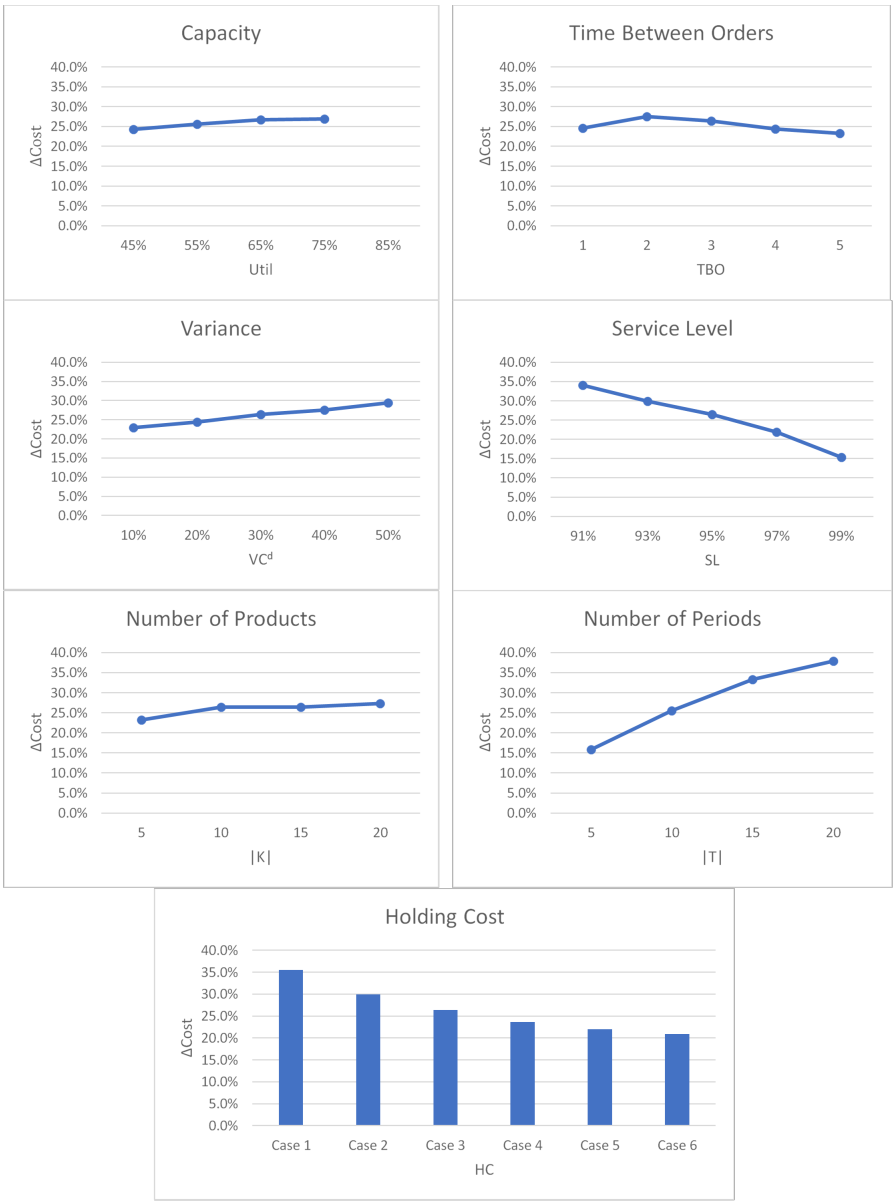


Figure 4: Sensitivity analysis plots for δ_p service level

Appendix D – Partial flexibility

Partial flexibility

In the previous sections, the flexibility was added to the system level by level. Having multiple items per level in assembly and general structures, we may add flexibility to some of the items in each level. In this section, we study the value of flexibility when it is added to different single items at each level. To this end, we calculate the cost decrease percentage when we add the flexibility to only one item compared to the case where we do not have any flexibility in that specific level. Tables 2 and 3 illustrate the cases with partial flexibility in assembly and general structures, respectively. The highlighted rows are the base case at the specific level based on which we calculate the cost decrease when we add the flexibility to a single item at the same level.

Table 2: Levels of flexibility for assembly structure

# Flexibility	# product									
	1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	0	0	0	0	0	0
1-2	1	1	0	0	0	0	0	0	0	0
1-3	1	0	1	0	0	0	0	0	0	0
1-4	1	0	0	1	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0	0
2-5	1	1	1	1	1	0	0	0	0	0
2-6	1	1	1	1	0	1	0	0	0	0
2-9	1	1	1	1	0	0	0	0	1	0
2-10	1	1	1	1	0	0	0	0	0	1
Level	0	1	1	1	2	2	2	2	2	2

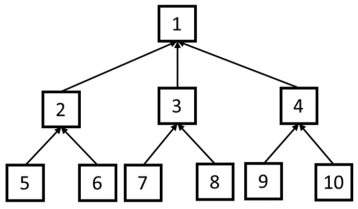


Figure 5 shows the diagram of adding flexibility to different items at level 1 and level 2 based on different external demand patterns for the assembly structure. We can see that adding flexibility to any item regardless of its place in BOM and the external demand result in cost decrease, but the amount of this decrease is affected by the mentioned two factors. When there is no external

demand for any of the components (external demand profile 1), the cost decrease is the same for all the products as there is no difference between different items. In external demand profile 3 adding flexibility to item 5 and 6 results in higher cost reduction compared to item 9 and 10. We can conclude that adding flexibility to the items whose parent have external demand results in higher cost decrease compared to the items whose parent does not have external demand.

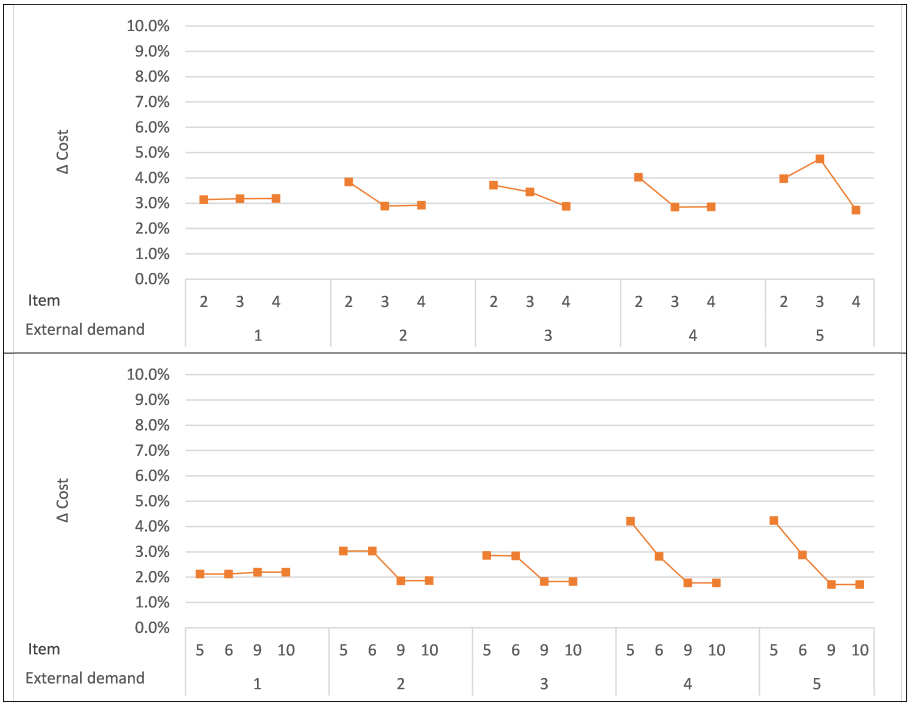


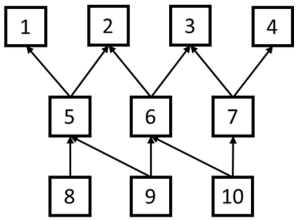
Figure 5: Analysis of adding partial flexibility at different levels (Assembly structure)

Figure 6 shows the diagram of adding flexibility to different items at levels 0, 1, and 2 based on different external demand patterns for the general structure. We can conclude that even adding flexibility to a single item can decrease the total cost of the system. In our experiments, the minimum decrease is about 2% for general structure. The cost decrease depends on the external demand of an item, the demand of its components or parents, and its position in the BOM. Having many factors, it makes it difficult to address the effect of each individually. Here are some general, conclusion that we can draw regarding partial flexibility.

Considering the flexibility for an item with external demand, result in a higher cost decrease compared to the item without any external demand, with other similar characteristics. Adding flexibility to the item with a higher number of direct or indirect components has a higher impact

Table 3: Levels of flexibility for general structure

# Flexibility	# product									
	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0
0-1	1	0	0	0	0	0	0	0	0	0
0-2	0	1	0	0	0	0	0	0	0	0
0-3	0	0	1	0	0	0	0	0	0	0
0-4	0	0	0	1	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0	0
1-5	1	1	1	1	1	0	0	0	0	0
1-6	1	1	1	1	0	1	0	0	0	0
1-7	1	1	1	1	0	0	1	0	0	0
2	1	1	1	1	1	1	1	0	0	0
2-8	1	1	1	1	1	1	1	1	0	0
2-9	1	1	1	1	1	1	1	0	1	0
2-10	1	1	1	1	1	1	1	0	0	1
Level	0	0	0	0	1	1	1	2	2	2



compared to the items with a lower number of components. For example, in our experiments for the general structure, adding flexibility to item 2 which has the highest number of components, result in the highest cost decrease compared to other items at level 0.

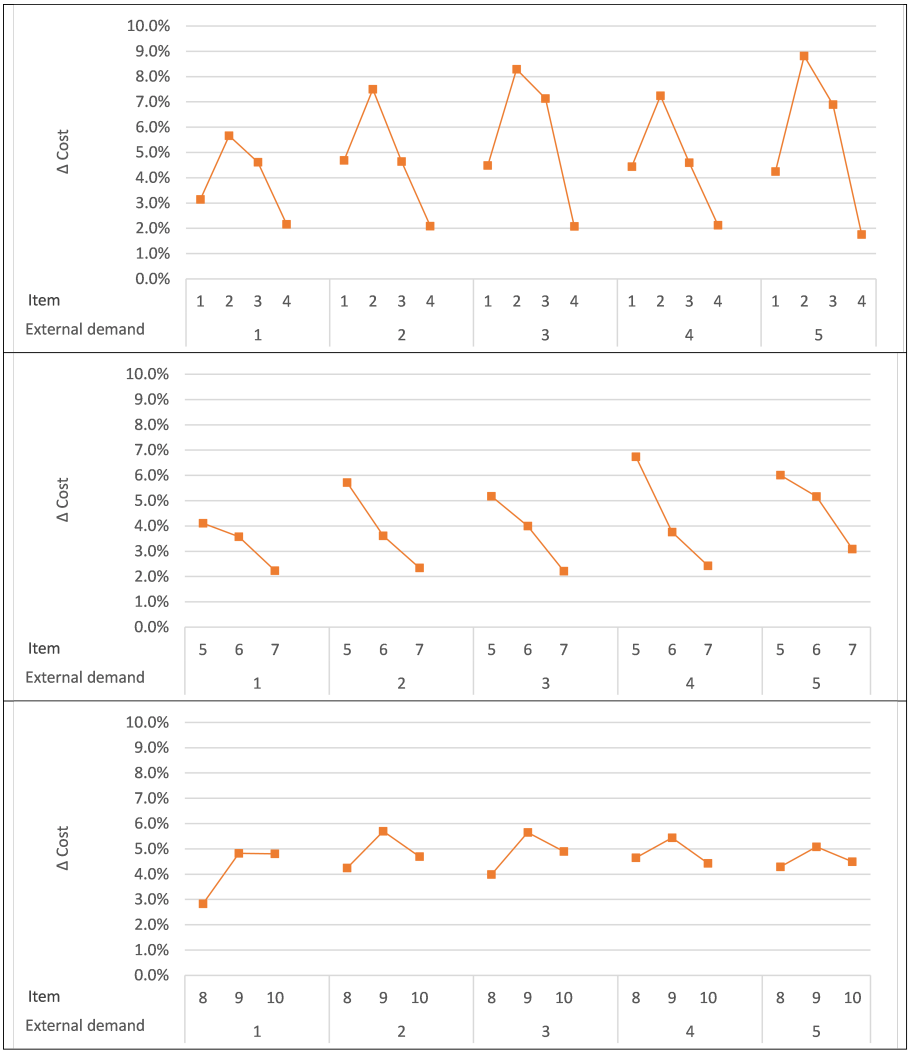


Figure 6: Analysis of adding partial flexibility at different levels (General structure)