Les Cahiers du GERAD

ISSN: 0711–2440

The value of aggregate service levels in stochastic lot sizing problem

N. Sereshti,
Y. Adulyasak, R. Jans

G–2019–68

September 2019

Citation suggérée : N. Sereshti, Y. Adulyasak, R. Jans (Septembre 2019). The value of aggregate service levels in stochastic lot sizing problem, Rapport technique, Les Cahiers du GERAD G–2019–68, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (https://www.gerad.ca/fr/papers/G-2019-68) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

Suggested citation: N. Sereshti, Y. Adulyasak, R. Jans (September 2019). The value of aggregate service levels in stochastic lot sizing problem, Technical report, Les Cahiers du GERAD G–2019–68, GERAD, HEC Montréal, Canada.

Before citing this technical report, please visit our website (https://www.gerad.ca/en/papers/G-2019-68) to update your reference data, if it has been published in a scientific journal.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2019
– Bibliothèque et Archives Canada, 2019

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2019
– Library and Archives Canada, 2019

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

The value of aggregate service levels in stochastic lot sizing problem

Narges Sereshti

Yossiri Adulyasak

Raf Jans

GERAD & Department of Logistics and Operations management, HEC Montréal, Montréal (Québec), Canada, H3T 2A7

narges.sereshti@hec.ca
yossiri.adulyasak@hec.ca
raf.jans@hec.ca

September 2019
Les Cahiers du GERAD
G–2019–68

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 programming models 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 a variety of products or components must be collectively satisfied. These extended models 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 models are proposed for this problem with different types of service levels. These models are a piece-wise linear approximation for the β, γ, and δ service levels and a quantile-based model for the α_c service level. Computational experiments are conducted to analyze the impact of aggregate service levels and demonstrate the value of the proposed models as opposed to standard service levels imposed on individual items.

Keywords: Stochastic lot sizing, Aggregate service level

Acknowledgments: The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (Grants 2014–03849 and PGSD3-504732–2017), the HEC Montréal Professorship in Operations Planning, and GERAD (Doctoral Fellowships Conference Fees). We would also like to thank professor Stefan Helber and professor Katja Schimmelpfeng for their valuable comments on an earlier version of this paper.

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 [13]. The lot sizing problem has been extended to include several practical cases such as multiple products, capacitated machines, or backlog costs [9].

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 [2]. 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 [12]. 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 [17]. 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 [4]. These service levels are typically defined for each product separately.

In this research, we investigate an aggregate service level which is defined aggregately for multiple products in addition to individual service levels when uncertainty in the demand is present [1]. 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 models are a piece-wise linear approximation and a quantile-based model. The contributions of this paper are as follows. First, we propose the idea of an aggregate service level in the stochastic lot sizing problem. 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). Some of the mathematical models proposed in this research are extensions of existing models, while others are originally proposed for these problems. Finally, we provide computational experiments and investigate the value of the aggregate service level.

This paper is structured into eight 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 proposed. Section 7 discusses the experimental results. Section 8 concludes the paper and discusses possible future research.

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 [11] 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. [23] 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 [23]. Akçay et al. [1] 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. We will also introduce a modified version of the *dollar fill rate* in this research which is aggregately defined for different products. Escalona et al. [5] 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 [14] 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. [7] 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 [10].

In the literature, several service level measures have been proposed when dealing with demand uncertainty [8]. 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 [6].

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 [8]. These service levels are defined for each of the products individually.

Bookbinder and Tan [3] 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. This strategy is the most common strategy used in stochastic lot sizing models. 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 set up 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.

Many papers studied the stochastic lot sizing problems and lot sizing problem with service levels using different strategies and service level constraints [17, 8, 19, 20, 22, 24]. Table 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.

Table 1: Overview of literature on lot sizing with service level constraints

Authors	Year	Strategy			Service level type				Individual	Aggregate	Capacity
		static	dynamic	static-dynamic	α	β	δ	γ			
Bookbinder and Tan [3]	1988	+	+	+	+				+		
Tarim and Kingsman [16]	2004			+	+				+		
Tempelmeier [17]	2007			+	+	+			+		
Tempelmeier and Herpers [18]	2010	+				+			+		+
Tempelmeier [19]	2011	+				+			+		+
Tempelmeier and Herpers [20]	2011	+				+			+		
Gade and Küçükyavuz [6]	2013							+	+		
Helber et al. [8]	2013	+					+		+		+
Tunc et al. [24]	2014			+	+				+		
Tempelmeier and Herpers [22]	2015	+				+			+		+
Gruson et al. [7]	2018		Deterministic		+	+	+		+	+	+
Meistering and Stadtler [15]	2019		Deterministic		+	+		+	+		+
Our Work	2019	+			+	+	+	+	+	+	+

3 Stochastic lot-sizing models with 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 lotsizing 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 fullfiled 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 [21]. Table 2 defines these different types of service levels in separate and aggregate format. Some of these service levels are defined over all planning horizon (globally) [21] and some others are defined per each planning per period. 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 aggregate β service level can be extended to consider the value of each product(v_k). This service level in the global case is shown in (1) and imposes that the value of backordered products should not exceed a certain percentage of the value of total demand.

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

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 [8]. 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 [8] 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 delta service level per period. This service level is 1 minus the expected backlog in each period divided by the maximum 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 [21]. In a company in which we have multiple products, managers may define an aggregate service level over all products. The aggregate service level guarantees that the weighted average of the resulting individual service levels is at least α_c^{agg} ($\alpha_c^{agg} \in [0, 1]$). 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 is shown in Table 2.

Before moving forward to the mathematical models of each service level, we will provide an example based on the β service level. Table 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. For example in the last column 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.

4 Models with aggregate β service level

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.

Table 2: Different types of service level and their separate and aggregate forms

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

Table 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%

4.1 Problem definition and mathematical model

The parameters and decision variables are presented in Table 4. The mathematical model for the stochastic capacitated lot sizing problem with aggregate β service level is as follows:

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}E[\overline{I}_{kt}]) \tag{2}$$

subject to:

$$\overline{I}_{k,t-1} + x_{kt} + \overline{B}_{kt} = \overline{I}_{kt} + \overline{D}_{kt} + \overline{B}_{k,t-1} \qquad \forall t \in T, \forall k \in K \tag{3}$$

$$x_{kt} \le M y_{kt} \qquad \forall t \in T, \forall k \in K \tag{4}$$

$$\sum_{k\in K}(st_{kt}y_{kt} + pt_{kt}x_{kt}) \le Cap_t \qquad \forall t \in T \tag{5}$$

$$
\begin{aligned}
E[\overline{BO}_{kt}] =& E[\max\{0, \sum_{j=1}^{t}(\overline{D}_{kj} - x_{kj}) - I_{k0}\}] \\
& - E[\max\{0, \sum_{j=1}^{t-1}\overline{D}_{kj} - \sum_{j=1}^{t} x_{kj} - I_{k0}\}] \qquad \forall t \in T, \forall k \in K \qquad (6)
\end{aligned}
$$

$$\frac{\sum_{t \in T}\sum_{k \in K} E[\overline{BO}_{kt}]}{\sum_{t \in T}\sum_{k \in K} E[\overline{D}_{kt}]} \leq 1 - \beta \qquad (7)$$

$$y_{kt} \in \{0, 1\} \qquad \forall t \in T, \forall k \in K \qquad (8)$$

$$x_{kt} \geq 0 \qquad \forall t \in T, \forall k \in K \qquad (9)$$

$$\overline{I}_{kt} \geq 0 \qquad \forall t \in T, \forall k \in K \qquad (10)$$

$$\overline{B}_{kt} \geq 0 \qquad \forall t \in T, \forall k \in K \qquad (11)$$

Table 4: Parameters and decision variables of the models with β service level

Sets	
T	Set of planning periods
K	Set of products
Parameters	
sc_{kt}	Setup cost for product k in period t
hc_{kt}	Inventory holding cost for product k in period t
st_{kt}	Setup time for product k in period t
pt_{kt}	Unit production time for product k in period t
cap_t	Production capacity in period t
M	A sufficiently large number
I_{k0}	The initial inventory for product k
β	Target fill rate as an aggregate service level
Random variables	
$\overline{D}_{kt}$	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
$\overline{B}_{kt}$	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
Decision variables	
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

The objective function of the model (2) minimizes the setup and expected inventory holding costs. Constraints (3) are the flow conservation constraint. Constraints (4) guarantee the setup forcing in case there is production. Constraints (5) enforce the capacity limitation. Constraints (6) calculate the expected backorder level for product k in period t [25]. 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 (7) ensures the aggregate β service level. Constraints (8) to (11) show the domain of the different variables in the model. In this model the expected value of inventory level ($E[\overline{I}_{kt}]$) can also be calculated by Equation (12) instead of having constraint (3).

$$E[\overline{I}_{kt}] = E[\max\{0, I_{k0} + \sum_{j=1}^{t}(x_{kj} - \overline{D}_{kj})\} \quad \forall t \in T, \forall k \in K \qquad (12)$$

As can be seen, in this mathematical formulation, $E[\overline{BO}_{kt}]$ and $E[\overline{I}_{kt}]$ are non-linear functions of the cumulative production which are shown in (6) and (12), respectively. In the next section, a mathematical model is presented to approximate this non-linear model.

4.2 Piece-wise linear approximation

The formulation presented here is an extension of the model proposed by Van Pelt and Fransoo [25], in which, the expected inventory, backlog, and backorder are calculated by (13), (14), and (15) 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 (6), (12), and (26) are equivalent to (15),(13), and (14), 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 [25] show that the expected backorder (15) 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 5.

$$E[\bar{I}_{kt}] = Q_{kt} - E[\overline{CD}_{kt}] + \Gamma^1_{\overline{CD}_{kt}}(Q_{kt}) \tag{13}$$

$$E[\overline{B}_{kt}] = \Gamma^1_{\overline{CD}_{kt}}(Q_{kt}) = E[max\{0, \overline{CD}_{kt} - Q_{kt}\}] \tag{14}$$

$$E[\overline{BO}_{kt}] = \Gamma^1_{\overline{CD}_{kt}}(Q_{kt}) - \Gamma^1_{\overline{CD}_{k,t-1}}(Q_{kt}) \tag{15}$$

Table 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.

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}(\Delta_{I_{0kt}} + \sum_{l\in L}\Delta_{I_{lkt}}w_{lkt})) \tag{16}$$

subject to constraints (4), (5), (8), (9), and:

$$x_{kt} = \sum_{l\in L} w_{lkt} - \sum_{l\in L} w_{lk,t-1} \qquad \forall t \in T, \forall k \in K \tag{17}$$

$$w_{l-1,kt} \geq (u_{l-1,kt} - u_{l-2,kt})\lambda_{lkt} \qquad \forall t \in T, \forall k \in K, \forall l \in L, l \geq 2 \tag{18}$$

$$w_{lkt} \leq (u_{lkt} - u_{l-1,kt})\lambda_{lkt} \qquad \forall t \in T, \forall k \in K, \forall l \in L \tag{19}$$

$$\sum_{l\in L} w_{lk,t-1} \leq \sum_{l\in L} w_{lkt} \qquad \forall t \in T, \forall k \in K \tag{20}$$

$$\frac{\sum_{t\in T}\sum_{k\in K}(\Delta_{BO_{0kt}} + \sum_{l\in L}\Delta_{BO_{lkt}}w_{lkt})}{\sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]} \leq 1 - \beta \tag{21}$$

$$w_{lkt} \geq 0 \qquad \forall t \in T, \forall k \in K, \forall l \in L \tag{22}$$

$$\lambda_{lkt} \in \{0, 1\} \qquad \forall t \in T, \forall k \in K, \forall l \in L \tag{23}$$

The objective function (16) is to minimize the setup cost plus the expected value of the holding costs. Constraints (17) calculate the production amount based on the selected segments. Constraints (18) to constraints (20) guarantee that the segments are selected in sequential order as proposed in [25]. Constraint (21) 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 (22) and (23) show the domain of the different variables in the model.

5 Models with aggregate γ and δ service level

In this section, we investigate an aggregate 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.

5.1 Problem definition and mathematical model

The mathematical model for this problem is presented as follows:

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}E[\overline{I}_{kt}]) \tag{24}$$

subject to constraints (3), (4), (5), (8), (9), and:

$$\frac{\sum_{t\in T}\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K} E[\overline{D}_{kt}]} \leq 1-\gamma \tag{25}$$

The objective function of the model (24) minimizes the setup and expected inventory holding costs. Constraint (25) guarantees the γ aggregate service level. In this model the expected value of backlog ($E[\overline{B}_{kt}]$) is calculated by constraint (26).

$$E[\overline{B}_{kt}] = E[\max\{0, \sum_{j=1}^{t}\overline{D}_{kj} - \sum_{j=1}^{t}x_{kj} - I_{k0}\}] \qquad \forall t\in T, \forall k\in K \tag{26}$$

It is also possible to define the aggregate γ service level in each planning period (γ_p). In this case, constraint (25) is substituted with constraints (27).

$$\frac{\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{k\in K} E[\overline{D}_{kt}]} \leq 1-\gamma_p \qquad \forall t\in T \tag{27}$$

It is also possible to use the δ service level instead of γ service level. In this case constraint (25) is substituted with constraint (28). Both γ and δ service levels work with the expected backlog for each product in each planning period.

$$\frac{\sum_{t\in T}\sum_{k\in K} E[\overline{B}_{kt}]}{\sum_{t\in T}\sum_{k\in K}(T-t+1)E[\overline{D}_{kt}]} \leq 1-\delta \tag{28}$$

In this mathematical formulation $E[\overline{I}_{kt}]$ and $E[\overline{B}_{kt}]$ are non-linear functions of cumulative production which are shown in (12) and (26), respectively. In the next section, a mathematical model is presented to approximate this non-linear model.

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 loss function of the normal distribution. As the non-linear functions for the expected inventory and expected backlog are convex, different segments on the piece-wises 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 6 indicates the new parameters and decision variables of this model.

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}(\Delta_{I_{0kt}} + \sum_{l\in L}\Delta_{I_{lkt}}w_{lkt})) \tag{29}$$

subject to constraints (4), (5), (8), (9), (17), (22), and:

$$w_{lkt} \le u_{lkt} - u_{l-1,kt} \qquad \forall t\in T, \forall k\in K, \forall l\in L \tag{30}$$

$$\frac{\sum_{t\in T}\sum_{k\in K}(\Delta_{B_{0kt}} + \sum_{l\in L}\Delta_{B_{lkt}}w_{lkt})}{\sum_{t\in T}\sum_{k\in K}E[\overline{D}_{kt}]} \le 1-\gamma \tag{31}$$

Table 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

The objective function (29) is to minimize the setup cost plus the expected value of the holding costs. Constraints (30) define the maximum amount that the production quantity associated with segment l can take in period t. Constraint (31) 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 (31) should be substituted with constraint (32) in which the total average backlog divided by the total maximum expected backlog is less than or equal to $1-\delta$.

$$\frac{\sum_{t\in T}\sum_{k\in K}(\Delta_{B_{0it}} + \sum_{l\in L}\Delta_{B_{lkt}}w_{lkt})}{\sum_{t\in T}\sum_{k\in K}(T-t+1)E[\overline{D}_{kt}]} \le 1-\delta \tag{32}$$

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.

6.1 Problem definition and mathematical model

This model is an extension of the model presented by Tempelmeier [17] 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 [3]. 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 7.

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}E[\overline{I}_{kt}]) \tag{33}$$

subject to constraints (3), (4), (5), (8), (9), and:

$$pr(I_{k0} + \sum_{j=1}^{t}(x_{kj} - \overline{D}_{kj}) \ge 0) \ge \alpha_c^{min} \qquad \forall k\in K, \forall t\in T \tag{34}$$

$$\sum_{k\in K} w_k \min_{t\in T}(pr(I_{k0} + \sum_{j=1}^{t}(x_{kj} - \overline{D}_{kj}) \ge 0)) \ge \alpha_c^{agg} \tag{35}$$

Table 7: Parameters and decision variables of the basic models

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

The objective function (33) minimizes the setup and expected holding cost. The minimum service level for each product is imposed through the chance constraints (34). These constraints guarantee that the probability of a stock out is not larger than $(1-\alpha_c^{min})$. Constraint (35) 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.

6.2 Quantile-based approximation

In this section, we approximate the model with aggregate α_c service level using a quantile approach. 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 [17]. 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 8. The mathematical model is as follows:

$$\text{Min} \sum_{t\in T}\sum_{k\in K}(sc_{kt}y_{kt} + hc_{kt}I_{kt}) \tag{36}$$

subject to constraints (4), (5), (8), (9), and:

$$I_{kt} = I_{k0} + \sum_{j=1}^{t}(x_{kj} - E[\overline{D}_{kj}]) \qquad \forall t \in T, \forall k \in K \tag{37}$$

$$I_{k0} + \sum_{t=1}^{j} x_{kt} \geq F^{-1}_{\overline{CD}_{kj}}(\alpha_c^{kn})s_{kn} \qquad \forall j \in T, \forall k \in K, \forall n \in N \tag{38}$$

$$\sum_{n\in N} s_{kn} = 1 \qquad \forall k \in K \tag{39}$$

$$\sum_{k\in K}\sum_{n\in N} w_k \alpha_c^{kn} s_{kn} \geq \alpha_c^{agg} \tag{40}$$

$$s_{kn} \in \{0,1\} \qquad \forall k \in K, \forall n \in N \tag{41}$$

$$I_{kt} \geq 0 \qquad \forall k \in K, \forall t \in T \tag{42}$$

Table 8: Parameters and decision variables of the basic models

Sets	
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 such that $P\{\overline{CD}_{kt} \leq cd\} \geq \alpha_c$
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

The objective function (36) minimizes the sum of set up and holding costs. Constraints (37) 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 (38) 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 (43) and for cases that the demand follows a normal distribution the value of $F^{-1}_{CD_{kj}}(\alpha_c^{kn})$ 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 (44) proposed for the single item problem [21]. Constraints (39) guarantee that exactly one service level for each item is selected. Constraint (40) 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.

$$pr(I_{k0} + \sum_{j=1}^{t} x_{kj} - \sum_{j=1}^{t} \overline{D}_{kj} \geq 0) \geq \alpha_c^{kn} s_{kn} \qquad \forall t \in T, \forall k \in K, \forall n \in N \tag{43}$$

$$I_{k0} + \sum_{t=1}^{j} x_{kt} \geq F^{-1}_{CD_{kj}}(\alpha_c^{k}) \qquad \forall j \in T, \forall k \in K \tag{44}$$

It is also possible to use constraint (45) as an alternative aggregate constraint for the service levels and substitute it with constraint (40). In this constraint, es_{kn} is the expected shortfall for product k based on service level n. Considering that c_k is the cost of shortfall for product k, constraint (45) guarantees that the cost of average unmet demand should be less than a certain value (DV). This value can be the maximum acceptable lost profit over the planning horizon.

$$\sum_{k\in K}\sum_{n\in N} c_k es_{kn} s_{kn} \leq DV \tag{45}$$

7 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 result of different service level based on an initial data set. Forth section is dedicated to extensive sensitivity analysis on the value of 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.

7.1 Instance generation

In this section, we explain the data which are used to test the models with different aggregate service levels. 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. [8] with some modification.

The set A is used to investigate the difference between the aggregate and separate service level for all types of service levels. As the original data set which was proposed by Helber et al. [8] 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 9 shows the parameters used for the generation of these test instances. VC^{ip} is the parameter which is used to generate dynamic time series and defines the average demand for each product in each planning period. 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. The standard deviation of the demand is equal to the average demand multiplied by the VC^{d}. Note that in practice, if the forcasted demand is used, one can calculate VC^{d} of forcasting 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 [8]. In set A there are 432 instances with all the combinations of parameters and the size of $\{|K| = 5, |T| = 5\}$, $\{|K| = 10, |T| = 5\}$, and $\{|K| = 5, |T| = 10\}$.

Table 9: Parameters of the test instances

Parameters	
Number of products	$|K|$ = *5 , 10 , 20*
Number of periods	$|T|$ = *5 , 10 , 20*
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	$w_k = 1/|K|$

The second data set, set B, is used for the sensitivity analyses. To this end, 10 base instances 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 10. In the sensitivity analysis section, we will explain how different scenarios for each of the parameters are generated.

Table 10: 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 setups 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.

7.2 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 shows the average solution time and average accuracy of the solution over all solved instances. The accuracy measures are the cost accuracy (46) and service level accuracy (47). 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. Considering the trade-off between 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. [8] who used 18 segments for their piece-wise linear model for the δ service level.

$$Cost\ Accuracy(\%) = \frac{|Evaluated\ total\ cost - Model\ objective\ function|}{Model\ objective\ function} \tag{46}$$

$$Service\ Level\ Accuracy(\%) = |Evaluated\ service\ level - Target\ service\ level| \tag{47}$$

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 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.

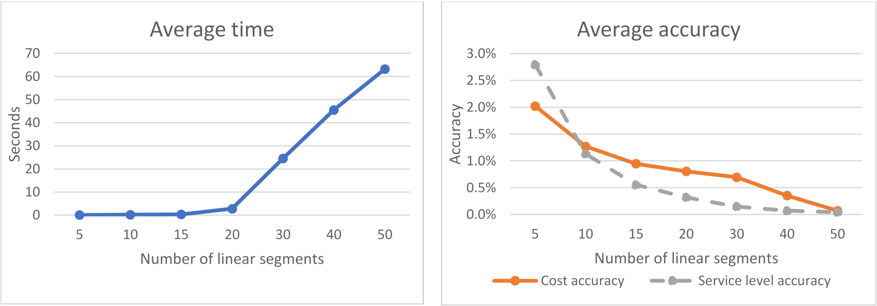

Figure 1: Execution time and accuracy of the piece-wise linear model for the γ service level based on the number of linear segments

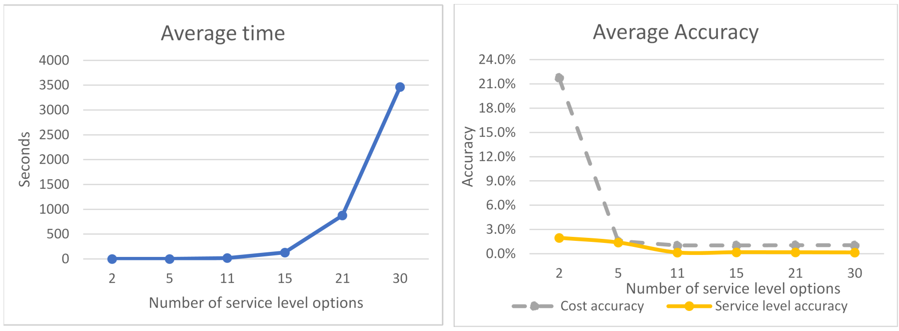

Figure 2: Execution time and accuracy of the quantile model for the α_c service level based on the number of service level options

7.3 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. Table 11 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 cost accuracy and service level accuracy, we analyze the average cost from the evaluation (Average Cost), the deviation of the actual service level from the defined target (Service Level Deviation)(48), the average percentage difference between the evaluated cost of the model objective and the model objective function (Cost Deviation)(49), 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 11 and shows the advantage of the aggregate service level over the separate one based on the total cost increase percentage (50).

$$Service\ Level\ Deviation(\%) = Evaluated\ service\ level - Target\ service\ level \tag{48}$$

$$Cost\ Deviation(\%) = \frac{Evaluated\ total\ cost - Model\ objective\ function}{Model\ objective\ function} \tag{49}$$

$$\Delta Cost(\%) = \frac{Cost\ of\ separate\ service\ level\ - Cost\ of\ aggregate\ service\ level}{Cost\ of\ aggregate\ service\ level} \tag{50}$$

Table 11: Results of the approximation models for different types of service level

Service Level	Version	Average Cost	Service Level Deviation (%)	Service level Accuracy (%)	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	

The first service level is the β service level. To analyze the results, the two versions of the model, 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. 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% $\Delta Cost$. 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.

7.4 Sensitivity analysis

To have a better understanding on the effect on 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 12. A low level for *Util* shows the loose capacity and a high level for *Util* pertains to the tight one. Table 13 shows different cases for the holding cost and their variance.

Table 12: Parameter values for the sensitivity analysis

Parameter	Values
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%*

Figure 3 presents the results of the sensitivity analysis of the γ service level. The capacity constraint 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 $\Delta Cost$.

Table 13: 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

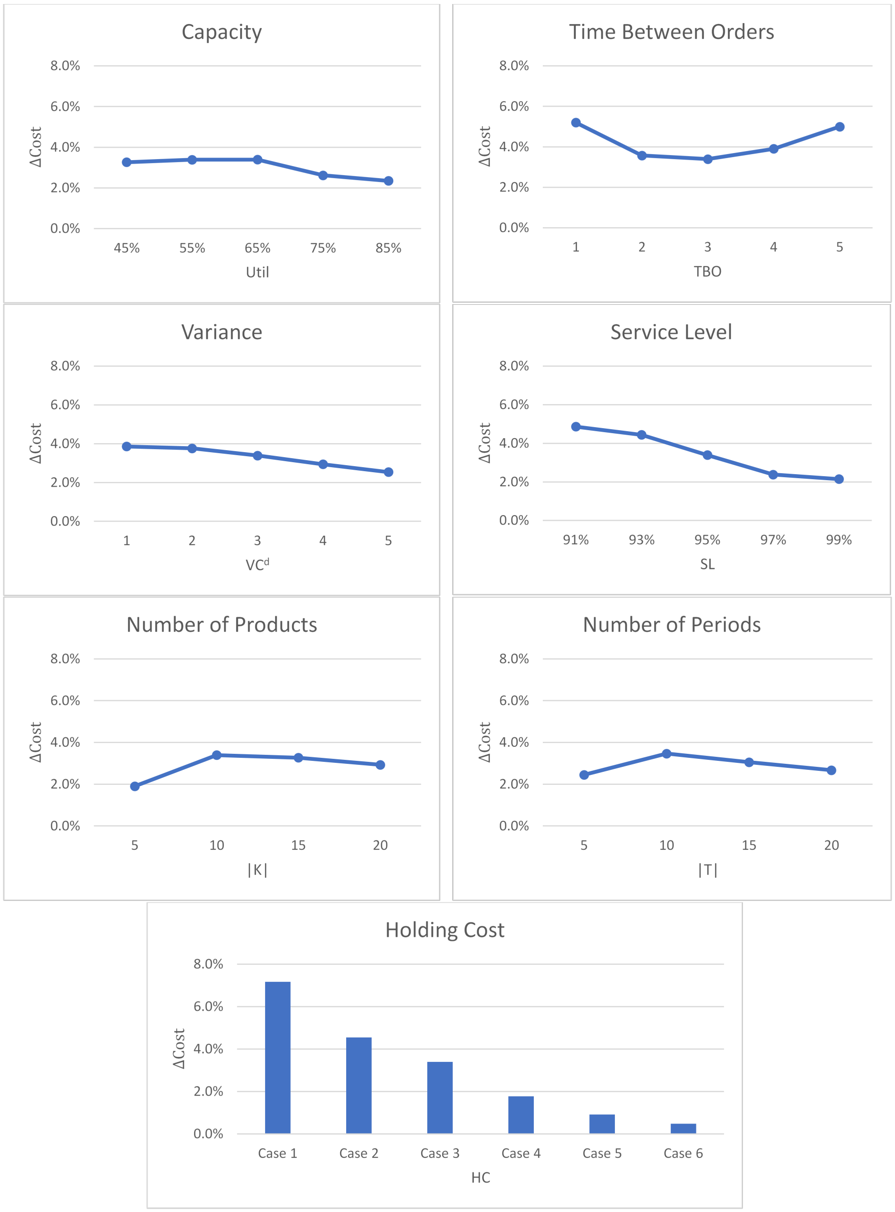

Figure 3: Sensitivity analysis plots for γ service level

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.

The last plot in Figure 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 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 in γ service level.

Similar to the case of 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 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 A 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.

7.5 The effect of minimum individual service level

In the previous section, there was no a minimum individual service level imposed in the aggregate models except for the α_c service level since this model requires 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 5 shows the plots for the γ service level.

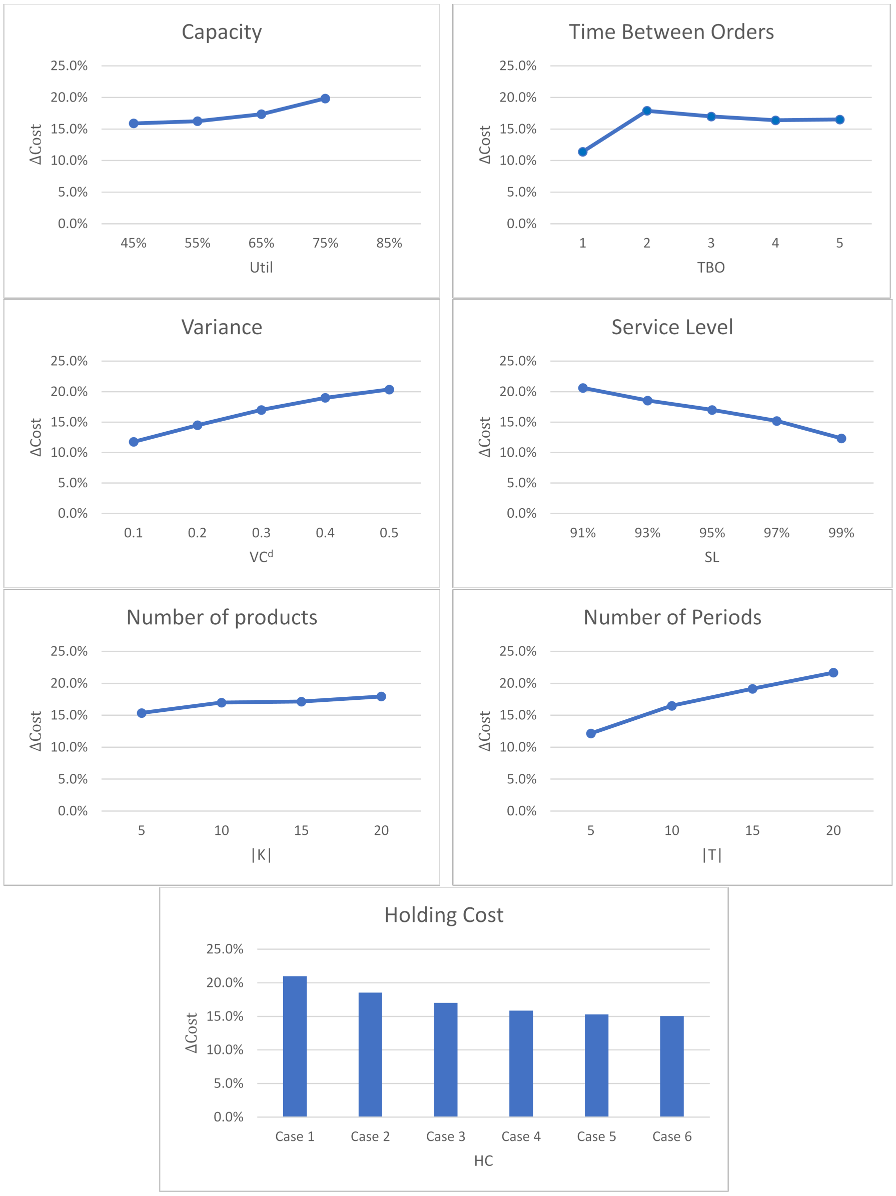

Figure 4: Sensitivity analysis plots for γ_p service level

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.

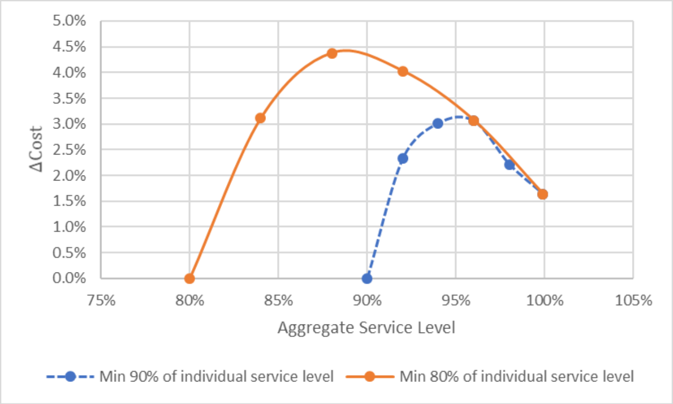

Figure 5: Effect of individual service levels (γ)

Appendix B 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 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.

8 Conclusion and future directions

In this research, different aggregate service levels have been 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. 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. The numerical experiments show that using the aggregate service level provides flexibility to the problem which result in overall cost reductions. This cost reduction varies based on different service levels and parameters. The numerical experiments show that the cost reduction is more in the case where the aggregate service level is imposed in each period compared to global case, and in quantity and time oriented service levels (β, γ, δ) compared to the event oriented one (α_c). The next step of this research is to approximate the problem using scenario-based models which can be used in dealing with general service levels and demand distributions. Such modeling scheme, albeit general, may not be scalable and this justifies hence further research on the development of efficient solution frameworks.

A Appendix A: Sensitivity analysis

In this appendix the diagrams of sensitivity analysis for the β, β_p, δ, and δ_p service levels are presented. The β and δ service levels have similar trends to 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 in δ_p and its lowest value in γ service level, if the models are feasible. What is common in all the diagram 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.

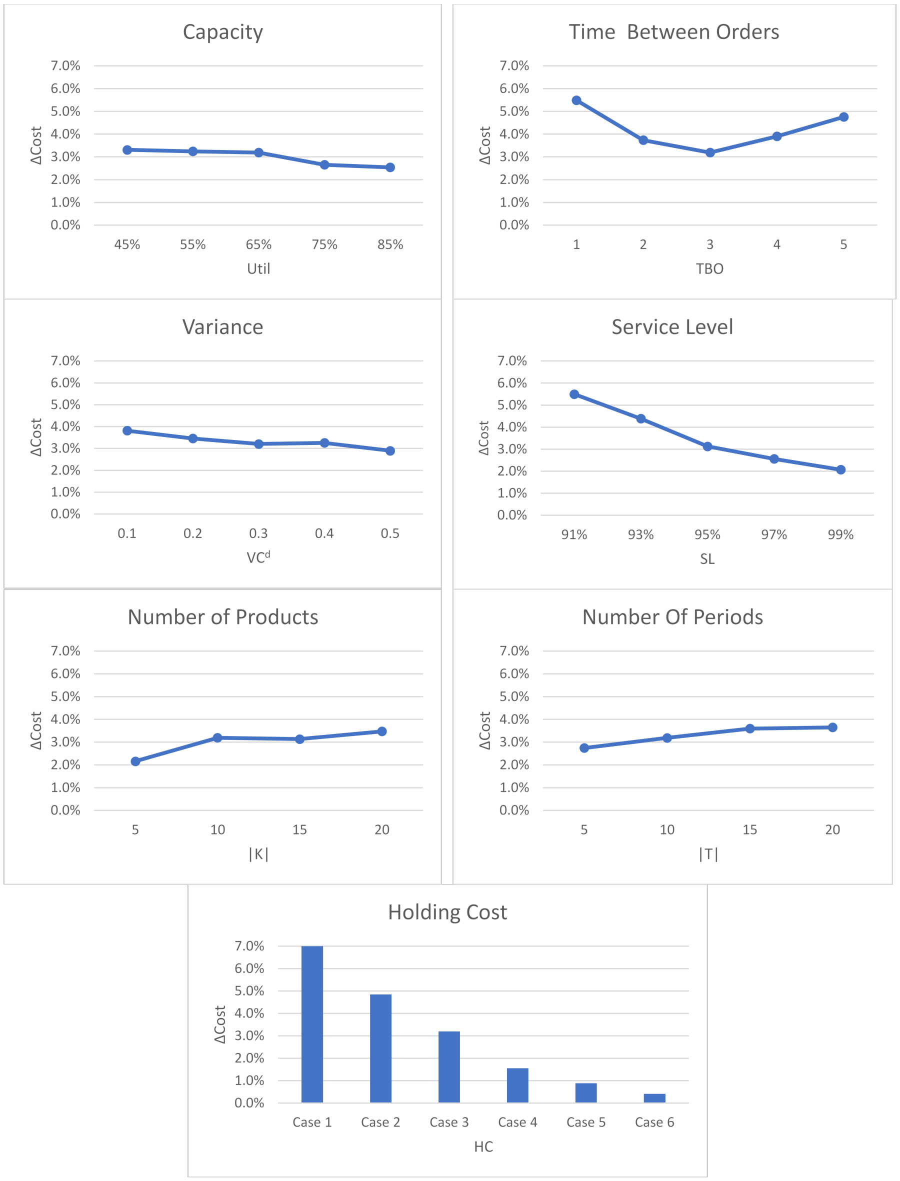

Figure 6: Sensitivity analysis plots for β service level

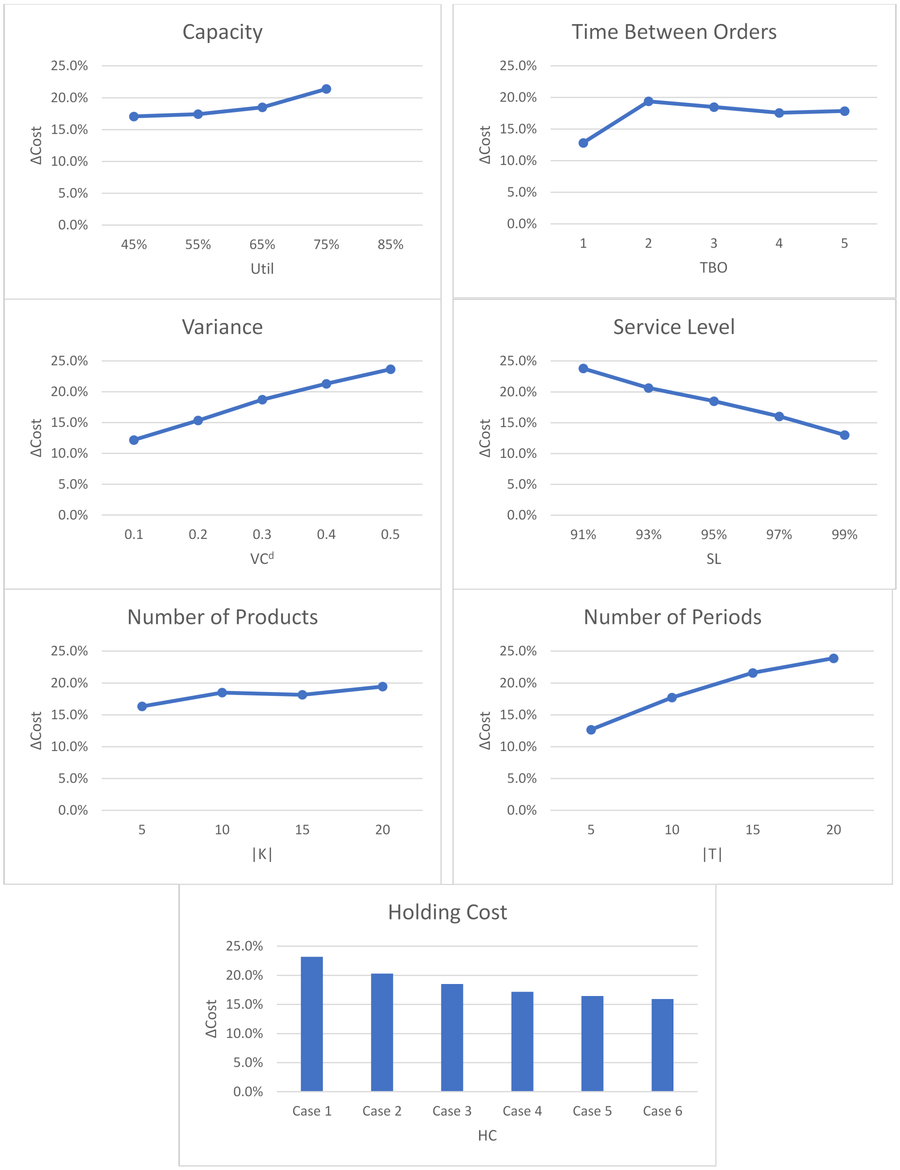

Figure 7: Sensitivity analysis plots for β_p service level

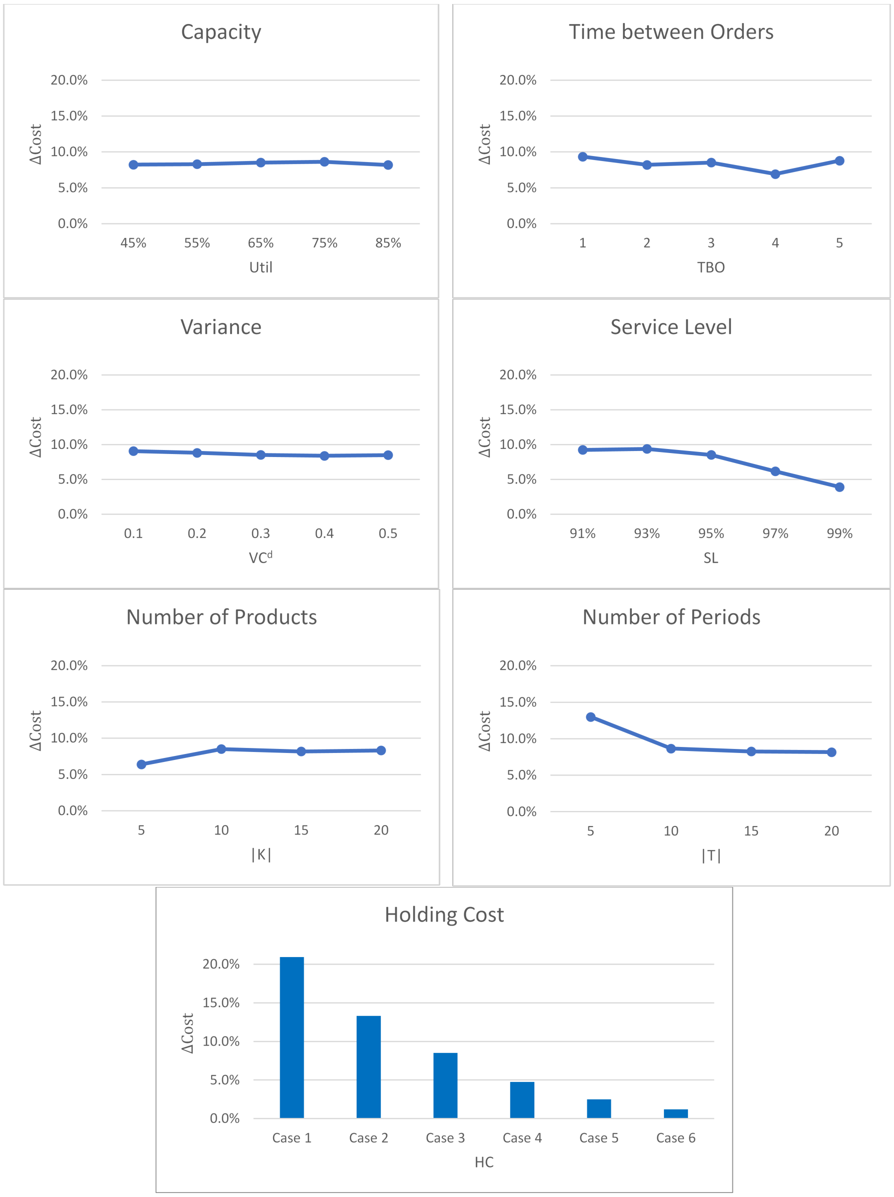

Figure 8: Sensitivity analysis plots for δ service level

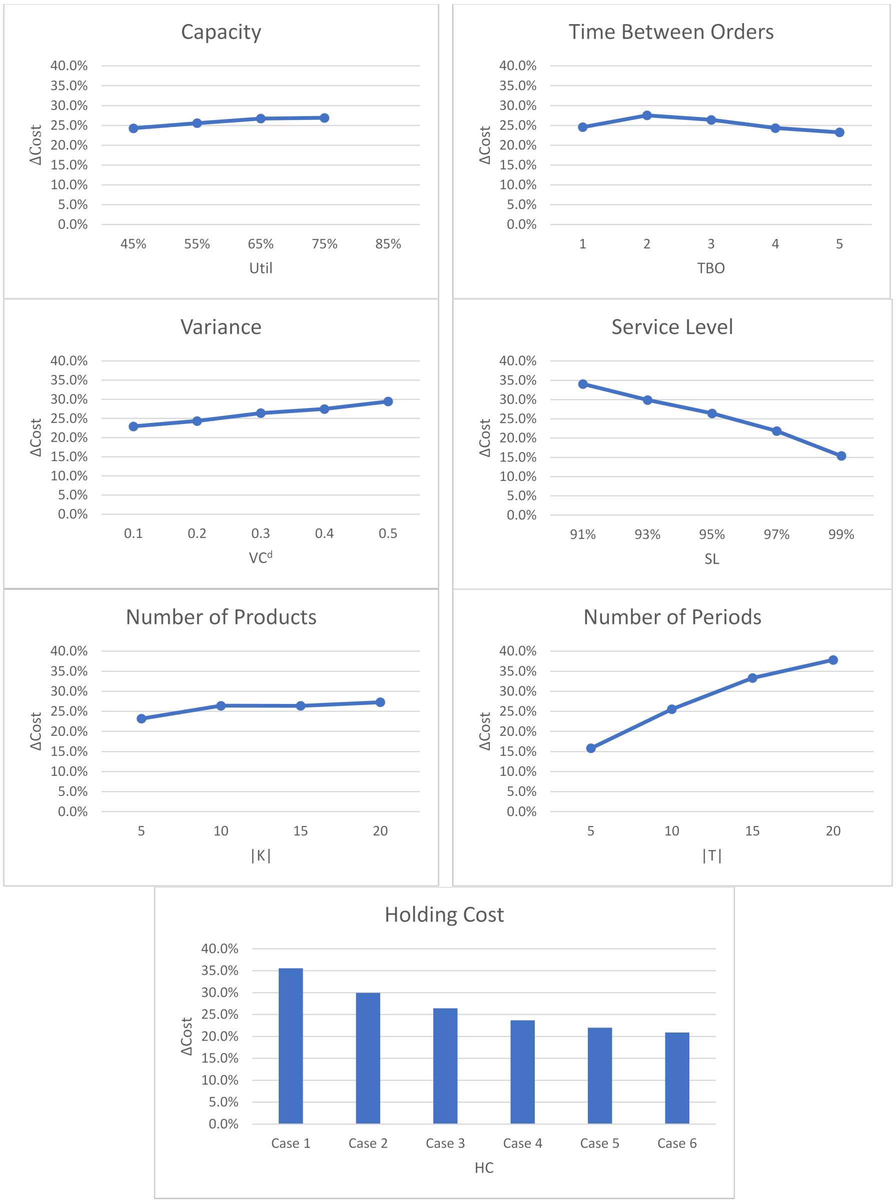

Figure 9: Sensitivity analysis plots for δ_p service level

A.1 Appendix B: Effect of minimum individual service levels

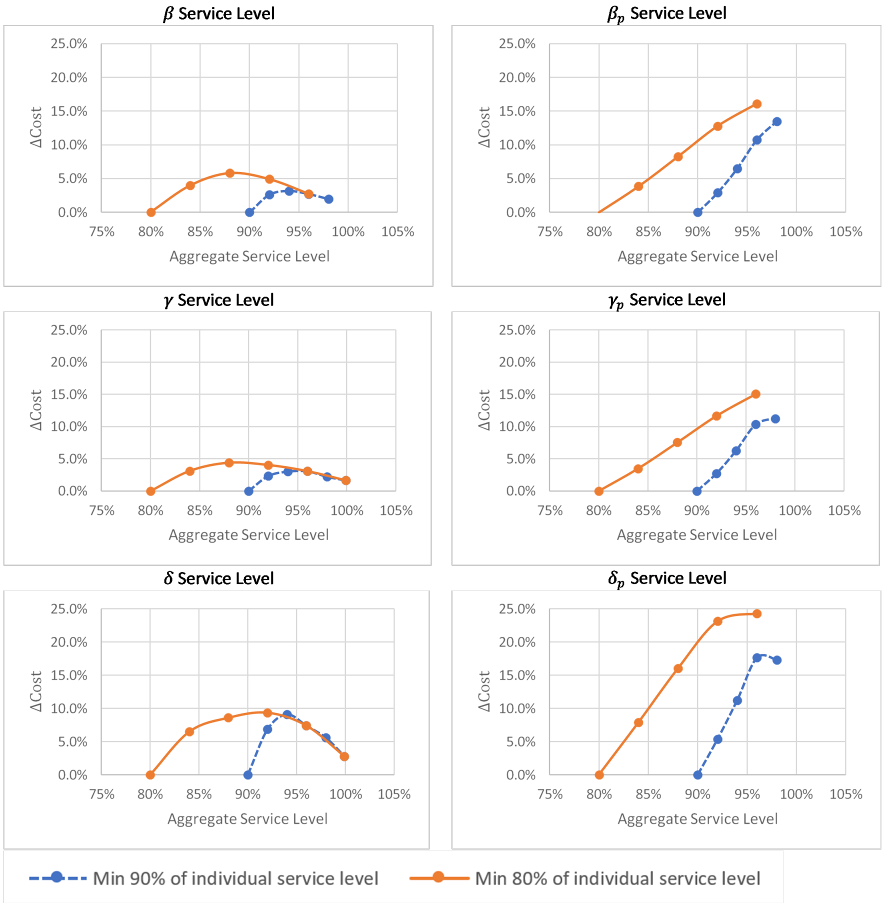

Figure 10: Effect of individual service levels

References

[1] Akçay, Alp and Biller, Bahar and Tayur, Sridhar R, Beta-Guaranteed aggregate Service Levels, Available at SSRN: https://ssrn.com/abstract=2709937 or http://dx.doi.org/10.2139/ssrn.2709937, (2016)

[2] Ben-Tal, Aharon and El Ghaoui, Laurent and Nemirovski, Arkadi, Robust Optimization, Prinston University Press, (2009)

[3] Bookbinder, James H and Tan, Jin-Yan, Strategies for the probabilistic lot-sizing problem with service-level constraints, Management Science, 34(9), 1096–1108, (1988)

[4] Brahimi, Nadjib and Absi, Nabil and Dauzère-Pérès, Stéphane and Nordli, Atle, Single-item dynamic lot-sizing problems: An updated survey, European Journal of Operational Research, 263(3), 838–863, (2017)

[5] Escalona, P and Angulo, A and Weston, J and Stegmaier, R and Kauak, I, On the effect of two popular service-level measures on the design of a critical level policy for fast-moving items, Computers & Operations Research, (forthcoming), (2019). https://doi.org/10.1016/j.cor.2019.03.011

[6] Gade, Dinakar and Küçükyavuz, Simge, Formulations for dynamic lot sizing with service levels, Naval Research Logistics, 60(2), 87–101, (2013)

[7] Gruson, Matthieu and Cordeau, Jean-François and Jans, Raf, The impact of service level constraints in deterministic lot sizing with backlogging, Omega, 79, 91–103, (2018)

[8] Helber, Stefan and Sahling, Florian and Schimmelpfeng, Katja, Dynamic capacitated lot sizing with random demand and dynamic safety stocks, OR spectrum, 35(1), 75–105, (2013)

[9] Jans, Raf and Degraeve, Zeger, Modeling industrial lot sizing problems: a review, International Journal of Production Research, 46(6), 619–1643, (2008)

[10] Jiang, Yuchen and Xu, Juan and Shen, Siqian and Shi, Cong, Production planning problems with joint service-level guarantee: a computational study, International Journal of Production Research, 55(1), 38–58, (2017)

[11] Kelle, Peter, Optimal service levels in multi-item inventory systems, Engineering Costs and Production Economics, 15, 375–379, (1989)

[12] Mula, Josefa and Poler, Raul and Garcia-Sabater, JP and Lario, Francisco Cruz, Models for production planning under uncertainty: A review, International journal of production economics, 103(1), 271–285, (2006)

[13] Pochet, Yves and Wolsey, Laurence A, Production planning by mixed integer programming, Springer Science & Business Media, (2006)

[14] Shivsharan, Chetan T, Optimizing the Safety Stock Inventory Cost Under Target Service Level Constraints, (Master of science), University of Massachusetts Amherst, (2012)

[15] Stadtler, Hartmut and Meistering, Malte, Model formulations for the capacitated lot-sizing problem with service-level constraints. OR Spectrum, 1–32, (forthcoming), (2019). https://doi.org/10.1007/s00291-019-00552-1

[16] Tarim, S Armagan and Kingsman, Brian G, The stochastic dynamic production/inventory lot-sizing problem with service-level constraints, International Journal of Production Economics, 88(1), 105–119, (2004)

[17] Tempelmeier, Horst, On the stochastic uncapacitated dynamic single-item lotsizing problem with service level constraints, European Journal of Operational Research, 181(1), 184–194, (2007)

[18] Tempelmeier, Horst and Herpers, Sascha, ABC 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, (2010)

[19] Tempelmeier, Horst, A column generation heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint, Omega, 39(6), 627–633, (2011)

[20] Tempelmeier, Horst and Herpers, Sascha, Dynamic uncapacitated lot sizing with random demand under a fillrate constraint, European Journal of Operational Research, 212(3), 497–507, (2011)

[21] Tempelmeier, Horst, Stochastic lot sizing problems, Handbook of Stochastic Models and Analysis of Manufacturing System Operations, 313–344, Springer, (2013)

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

[23] Teunter, Ruud H and Babai, M Zied and Syntetos, Aris A, ABC classification: service levels and inventory costs, Production and Operations Management, 19(3), 343–352, (2010)

[24] Tunc, Huseyin and Kilic, Onur A and Tarim, S Armagan and Eksioglu, Burak, A reformulation for the stochastic lot sizing problem with service-level constraints, Operations Research Letters, 42(2), 161–165, (2014)

[25] Van Pelt, Thomas D and Fransoo, Jan C, A note on “Linear programming models for a stochastic dynamic capacitated lot sizing problem”, Computers & Operations Research, 89, 13–16, (2018)