Inventory and Service Optimization for Self-serve Kiosks

by

Gohram

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Ruxian Wang Associate Professor, Carey Business School, Johns Hopkins University
Supervisor(s):	Dr. Fatma Gzara Associate Professor, Management Sciences, University of Waterloo
Internal Member:	Dr. Samir Elhedhli Professor, Management Sciences, University of Waterloo
Internal Member:	Houra Mahmoudzadeh Assistant Professor, Management Sciences, University of Waterloo
Internal-External Member:	Ricardo Fukasawa Professor, Combinatorics and Optimization, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I am the sole author of Chapters 5 and 6.

Parts of Chapter 1 have been incorporated in two papers (Baloch and Gzara 2020a, 2021) that were co-authored by myself and my supervisor, Dr. Fatma Gzara.

Chapters 2 and 3 of this thesis are published in a paper (Baloch and Gzara 2020a) that was co-authored by myself and my supervisor, Dr. Fatma Gzara.

Chapter 4 of this thesis has been incorporated within a paper (Baloch and Gzara 2021) that is submitted for publication. The paper is co-authored by myself and my supervisor, Dr. Fatma Gzara.

Abstract

In the retail industry, labor costs constitute a big chunk of total operating costs and retailers are advancing towards process automation to minimize their operating costs and to provide reliable services to their customers. One such example of technological advancement is self-service kiosks that are becoming an integral part of our life, whether it be for cashing a cheque, self-checkout at retail stores, airports, hospitals, or checkout-free stores. Although self-serve kiosks are cost-effective due to low setup and operating costs, the technology is relatively new and poses new research questions that have not been studied before. This thesis explores and addresses strategic and operational challenges associated with self-serve kiosk technology.

The first part of the thesis is based on a collaboration with *MedAvail Technologies Inc.*, a Canada-based healthcare technology company, developing self-serve pharmacy kiosk technology to dispense over-the-counter and prescription drugs. MedAvail faces several challenges related to assortment and stocking decisions of medications in the kiosk due to its limited capacity and the thousands of drugs being ordered in various quantities. We address these challenges by analyzing pharmaceutical sales data and developing a data-driven stochastic optimization approach to determine optimized kiosk storage capacity and service levels and recommend assortment and stocking decisions under supplier-driven product substitution. A column-generation based heuristic approach is also proposed to solve the models efficiently.

The second part of the thesis extends the self-serve kiosk inventory planning problem to a robust optimization (RO) framework under fill rate maximization objective. We propose a data-driven approach to generate polyhedral uncertainty sets from hierarchical clustering and the resulting RO model is solved using column-and-constraint generation and conservative approximation solution methodologies. The proposed robust framework is tested on actual pharmacy sales data and randomly generated instances with 1600 products. The robust solutions outperform stochastic solutions with an increase in out-of-sample fill rate of 5.8%, on average, and of up to 17%.

Finally, the third part of the thesis deals with an application of pharmacy kiosks to improve healthcare access in rural regions. We present a mathematical function to model customer healthcare accessibility as the expected travel distance when multiple pharmacy location (store and kiosks) choices are available to customers. Customer choice behavior is modelled using

a multinomial logit (MNL) model where customer utility for a pharmacy location depends on travel distance which is not exactly known but rather depends on kiosk fill rate. We model the problem as a newsvendor problem with fill-rate dependent demand to decide on kiosk stock level (or capacity) to minimize the weighted sum of expected travel distance and total cost. Sensitivity analysis over modelling parameters is carried out to derive insights and to determine problem settings under which pharmacy kiosks improve customer accessibility.

Acknowledgements

First and foremost, I would like to extend my sincere gratitude and appreciation to my supervisor, Dr. Fatma Gzara. Her expertise, guidance, and tireless encouragement have transformed me into a better researcher. I would like to thank *MedAvail Technologies Inc.* for sharing pharmaceutical sales data that contributed significantly to this thesis.

Finally, I am grateful to my parents, siblings, friends, and family who have been of great support throughout my life. Most importantly, I am thankful to my wife, Sara, for her love and encouragement as I complete my Ph.D.

Dedication

To the world you may be one person; but to one person you may be the world.
— Dr. Seuss

For Ami, Abo, & Sara

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the retail industry, labor costs constitute a big chunk of total operating costs and retailers are advancing towards process automation to minimize their operating costs and to provide reliable services to their customers. One such example of technological advancement is *self-service* kiosks that are becoming an integral part of our life, whether it be for cashing a cheque, self-checkout at retail stores, airports, hospitals, and even checkout-free stores. Self-service kiosks have a global market of $16.9 billion and expect to reach $30.8 billion by 2024 (Ramanath 2019).

Figure 1.1 shows a variety of self-service kiosks sold at one of the largest online retail platforms, *Alibaba.com*, for pharmaceutical drugs, cell phone accessories, retail items, fruits & vegetables, and even adult products (Alibaba.com 2020). Retail prices for these kiosks range from $2,000 to $6,000 with the capacity to store 180 to 500 stock keeping units (SKUs). Several retail companies have installed self-service kiosks to provide 24/7 service to their customers at locations like airports. For instance, Best Buy, one of the largest electronics retailers in North America, has 30 Express Kiosks located across Canada in airports and ferry terminals and offers a variety of products including computer and laptop accessories, power chargers, and headphones (BestBuy 2020).

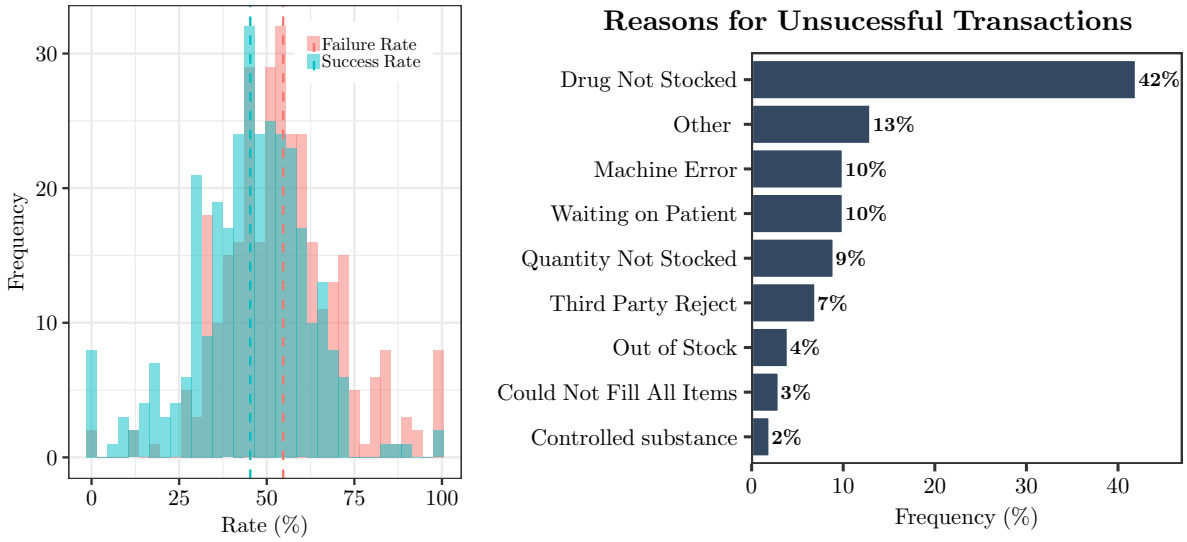
As in the retail industry, skyrocketing healthcare spending is a burden for officials in the health care industry. The total healthcare spending in the United States was $3.65 trillion in 2018 which is higher than the GDPs of U.K. and Canada (Sherman 2019). In Canada, healthcare

Figure 1.1: Self-service Kiosks available at Alibaba.com (Alibaba.com 2020)

expenditure reached $265 billion in 2019, equivalent to $7,064 per person representing 11.5% of country's total gross domestic product (GDP) (Canadian Institute for Health Information 2021). Healthcare executives are therefore looking for innovations that not only reduce healthcare costs but also improve patients' experience. A self-serve pharmacy kiosk is an example of such a recent innovation in healthcare technology aimed at reducing pharmaceutical operational costs. Several US and Canada-based companies including *MedAvail Technologies*, *MedifriendRx*, and *PharmaBox* have developed ATM-style self-dispensing pharmacy kiosks that store and dispense prescription and over-the-counter drugs (MedAvail 2017). When a customer inserts their prescription into the kiosk, he/she is connected to a remote pharmacist who after a careful review of the prescription, dispenses the prepackaged drug if it is stocked. Under the recent COVID-19 outbreak, Las Vegas airport and Greenville-Spartanburg International airport have installed vending kiosks offering personal protective equipment including facemasks, gloves, and sanitizers to travelers (Kelleher 2020, Lee 2020). Similarly, a retail store in the United States, Earthly Mist, installed kiosks to make their products available 24/7 during COVID-19 lockdown. It also offers up to 50% discount to customers using self-serve kiosk service (Maras 2020).

Self-serve kiosks are cost-effective due to low setup and operating costs. According to MedAvail, their pharmacy kiosk incurs an annual operating cost of $35,000 and can cover its total fixed and operating costs with as low as 25 dispenses a day (HealthcareConference 2017). Compared to a pharmacy store that has an upfront cost of around $1.5 million, a pharmacy kiosk can be purchased for only $100,000 (HealthcareConference 2017). Although the kiosk technology is cost-effective, it poses inventory challenges due to a limited storage capacity to stock a wide range of products. In addition to limited capacity, product demand at kiosks is often low. Anal-

ysis of pharmaceutical sales in Chapter 2 shows that thousands of distinct drugs are ordered annually each having low and erratic demand that does not fit any known theoretical probability distribution. Sporadic product demand coupled with limited storage capacity makes kiosk stocking decisions a challenging task resulting in poor service levels which may also affect future sales. In the case of MedAvail, it turns out that 55% of customer requests failed, mainly due to stocking issues as shown in Figure 1.2.



(a) (b)

Figure 1.2: The graphs depict the performance of the existing kiosk in meeting customer requests. Plot (a) illustrates daily success and failure rate distributions. Plot (b) summarizes the main reasons for failed transactions.

Lower service levels could adversely affect a firm's long-run profitability and sustainability. As such, during the early stages of the business cycle, firms often sacrifice their short-term business profits to boost customer satisfaction and loyalty, for the sake of business growth and long-term profitability (HBR 2009). One such great example is "Amazon.com" which was so focussed on customer service and experience that it took Amazon nine years after being founded in 1994 to make a profit (Hendricks 2014). Since self-serve kiosk technology is relatively new, it is critical that kiosk strategic and operational challenges are addressed from customer satisfaction perspective for long-term growth. This motivates us to make decisions such that customer

satisfaction is maximized. As opposed to the profit maximization objective, this thesis focuses on the fill rate maximization objective which is defined as the percentage of successful customer transactions.

To optimize fill rate, one possible solution is to increase the capacity. For instance, a startup company, *AiFi*, has recently introduced its fully autonomous retail nanostore with the ability to stock a large number of products and comes in various sizes ranging from 160 sq. ft to 300 sq. ft with the capacity to stock 300 to 720 SKUs (AiFi 2020). Similarly, MedAvail is developing a new kiosk with a refrigeration system and higher capacity to minimize missed opportunities (failed transactions). Chapter 3 of the thesis addresses inventory challenges from a capacity planning perspective for self-serve pharmacy kiosks that have an added complexity of products being ordered in various quantities. We model the problem as a data-driven stochastic optimization approach under supplier-driven substitution where the demand for higher quantities could be met by dispensing multiple packages of lower quantity.

The stochastic optimization approach proposed in Chapter 3 models the expected fill rate under the assumption that the empirical distribution is a true representation of the actual demand distribution. For the kiosk inventory planning problem, the demand is sporadic and does not follow a known theoretical distribution. On top of that, when limited historical data is available, a stochastic optimization approach may lead to poor out-of-sample performance, referred to as “Optimizer’s curse" in the Operations Research literature. This motivates us to follow a robust optimization framework that hedges against the worst-case realization of the demand within an uncertainty set. In Chapter 4, we present a robust optimization framework under fill rate maximization objective that has not been studied before. We construct a data-driven polyhedral uncertainty set using a hierarchical clustering algorithm to remove overly-conservative demand scenarios. The fill rate maximization objective in a robust framework poses computational challenges due to the non-convex adversarial problem and an exact robust counterpart does not exist. We therefore present an exact solution approach based on column-and-constraint generation (C&CG) and a conservative approximation approach where scenarios from uncertainty sets are generated from an adversarial problem and are dynamically added to the master problem.

Chapter 5 of the thesis examines the potential role of pharmacy kiosks from the perspective of improving healthcare accessibility in rural regions. In rural areas, due to low population

density, pharmacy stores are often located far from customer locations. Self-serve kiosk technology could be used to address the healthcare accessibility issue by placing multiple kiosks at various locations that are periodically replenished by a central pharmacy store which could be far from the customer location. This would allow customers to travel to a nearby kiosk to purchase their medications. However, in the case of stock-outs, which is likely to happen for kiosks due to their limited capacity, customers may end up travelling to the pharmacy store. As such, kiosks may even adversely affect customer accessibility if their capacity is limited. Eventually, customers may stop visiting kiosks in the long run if service levels are consistently low. This motivates us to model accessibility as a function of the expected distance which depends on kiosk fill rate as well as customer willingness to visit the kiosk. The latter is modelled using a multinomial logit (MNL) model where customer utility is in turn a function of the expected distance. The proposed accessibility function is used to model a newsvendor problem with fill rate-dependent demand that decides on stock levels for multiple kiosks such that the weighted sum of expected travel distance and the total cost is minimized.

1.1. Thesis Outline

The remainder of the thesis is organized as follows. In Chapter 2, we carry out extensive data analysis over pharmacy sales data to identify the critical factors that need to be considered in making assortment and stocking decisions. Motivated by the findings of the descriptive analysis, we address the capacity and assortment planning problem for pharmacy kiosks under one-way supplier driven substitution using a data-driven stochastic optimization approach in Chapter 3. In Chapter 4, we investigate the kiosk inventory problem in a robust setting under fill rate maximization objective. An integrated conservative approximation and column-and-constraint generation based solution methodology is proposed which is testing on real pharmacy data to determine problem settings where robust framework outperforms the stochastic approach. Chapter 5 studies the application of pharmacy kiosks in the context of improving healthcare accessibility in rural areas. Finally, some concluding remarks and future research directions are detailed in Chapter 6.

Chapter 2

Analytics of Demand

This thesis is motivated by an industry project completed in collaboration with MedAvail Technologies. We were provided with pharmacy store and self-serve pharmacy kiosk, MedCenter, sales transaction data for the year 2015. As a first step, we analyze one-year pharmaceutical sales data from seven stores with a data size of 18 million customer transactions. Our goal in this chapter is to derive demand characteristics and identify critical factors to model within an optimization framework with the ability to handle inventory planning for up to 30,000 drugs. A relational database was created in MS Access storing pharmacy data which is then linked to statistical software "R" to carry out extensive data analysis.

Pharmaceutical drugs are manufactured by various companies and are ordered by customers in various quantities. To evaluate drug demand distribution, it is important to understand how drugs are classified in general. In the US, each drug is assigned a unique 11-digit 3-segment numeric identifier called "National Drug Code (NDC)", denoting manufacturer code, product code, and the package code. Drugs are also assigned a 14-digit hierarchical classification scheme called "Generic Product Identifier (GPI)" that classifies drugs based on their therapeutic use, dosage form, and strength regardless of the manufacturer or package size. Drugs with same ingredients, dosage form, and strength but different manufacturers or package sizes share the same GPI code. At a pharmacy kiosk, drugs are stored in a specific quantity in a standardized package. This makes manufacturer's package size irrelevant in our context. Similarly, drugs with the same formula, dosage form, and strength but different manufacturers are pharmaceu-

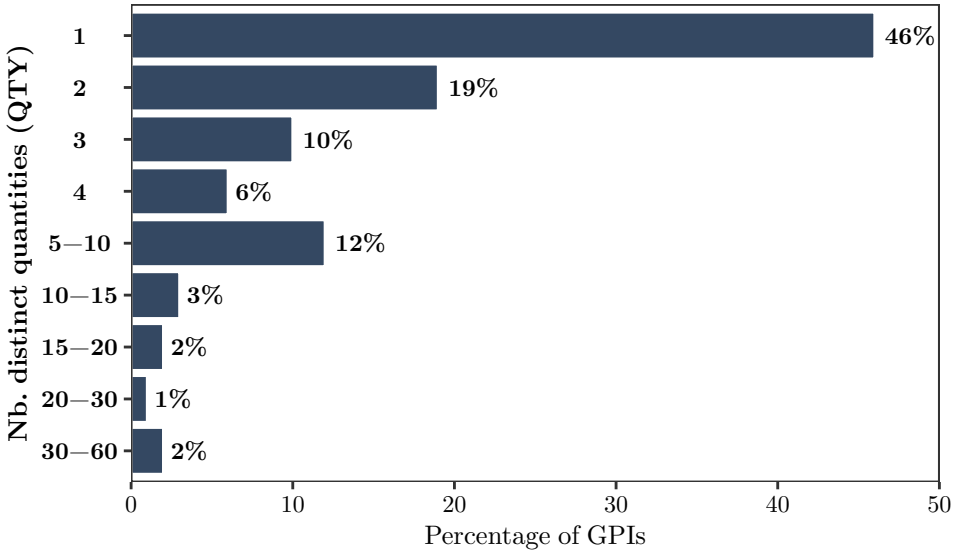


Figure 2.1: The graph depicts distribution of the GPIs' distinct quantities requested in the year 2015.

tically equivalent. We therefore consider GPI as a distinct drug identifier.

Analysis of pharmacy sales data shows that most of the GPIs are requested in multiple quantities (QTY). Figure 2.1 illustrates the distribution of GPIs' distinct quantities requested in the year 2015 over all stores. On average, each GPI is requested in four distinct quantities, while 46% of the GPIs are ordered in a single quantity. At a pharmacy kiosk, a medication is assumed to be in available only if it is stocked in the exact requested quantity. As such, a successful customer transaction requires the right drug with the right quantity to be in stock when ordered, so we use GPI-QTY to denote a distinct SKU in the rest of the analysis. We now analyze the significance of product substitution, demand distribution, and co-ordering of drugs using historical data and identify the critical factors to be modelled.

2.1. Product Substitution

Since GPIs are ordered in various quantities, multiple packages of the same GPI with different quantities may need to be stored resulting in higher capacity requirements. One possible so-

lution is to allow supplier-driven substitution between SKUs that share the same GPI code but have a different quantity. We explain the supplier-driven substitution effect using an illustrative example. Consider a GPI that is ordered in five different quantities: $\{20, 28, 40, 56, 60\}$. We may either stock five distinct packages, one of each quantity 20, 28, 40, 56, and 60 or, we may store only packages of 20 and 28 since 40 and 60 are multiples of 20 and 56 is a multiple of 28. As such, GPI-20 may substitute GPI-40 and GPI-60 while GPI-28 may substitute GPI-56. Optimal substitution decisions, however, depend on the demand for each quantity. For instance, if GPI-60 is frequently ordered, we should store it in quantities of 60 rather than 20, which would otherwise result in an increased number of packages. On the other hand, when GPI is rarely ordered in quantities of 60, it may be better to stock packages in quantities of 20 to satisfy sales in quantities of 20 and 60. This motivates us to consider inventory planning under supplier-driven substitution.

Another categorization of substitution is *customer-driven substitution* where customers decide on substitution when their preferred product is not available. For instance, if a customer wanted to buy his/her favorite brand of pain reliever that is not available at the pharmacy store, he/she may switch to another pain reliever. However, the data reveals that over-the-counter drugs constitute only 2.5% of the total pharmacy sales. At pharmacy stores, customer orders predominantly consist of prescribed drugs (97.5% of sales) which cannot be substituted by other drugs at the request of the customer. As such, customer-driven substitution is more appropriate in the context of retail items.

2.2. Demand distribution

We attempt to determine if demand follows a distribution that could be used in the modelling approach to make stocking and supplier-driven substitution decisions. Pharmacy sales data reveals that demand for the majority of drugs is low as shown in Figure 2.2. The latter illustrates the distribution of the number of days in a year GPI-QTYs are ordered where 40% of the GPI-QTYs appeared only one day and on average, the number of days GPI-QTYs are requested equals 11. Only 20% of the GPI-QTYs are requested in 10 days or more per year. Figure 2.3 plots the cumulative demand distribution and yearly demand of the GPI-QTYs. The top 14%

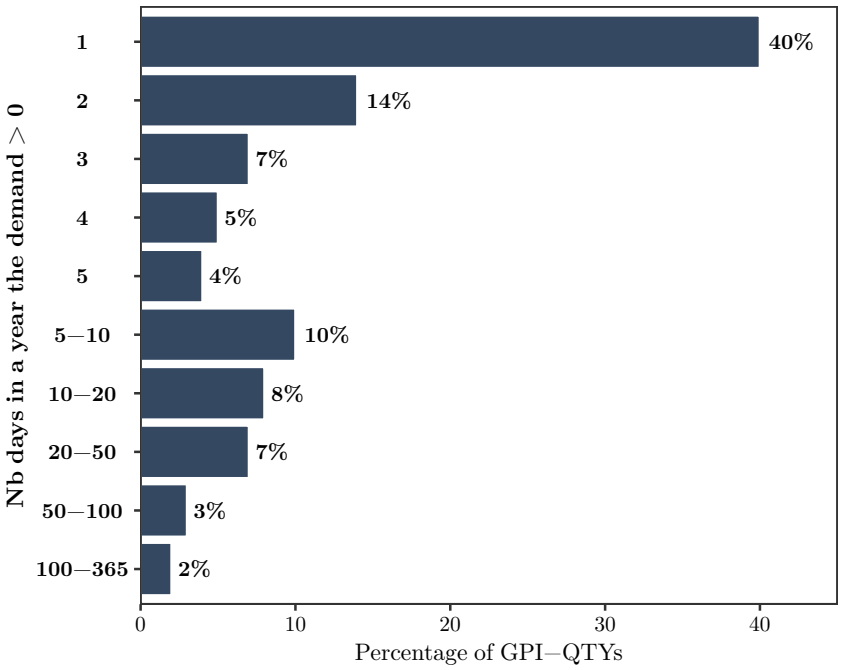


Figure 2.2: Distribution of the number of days drugs are ordered in a year

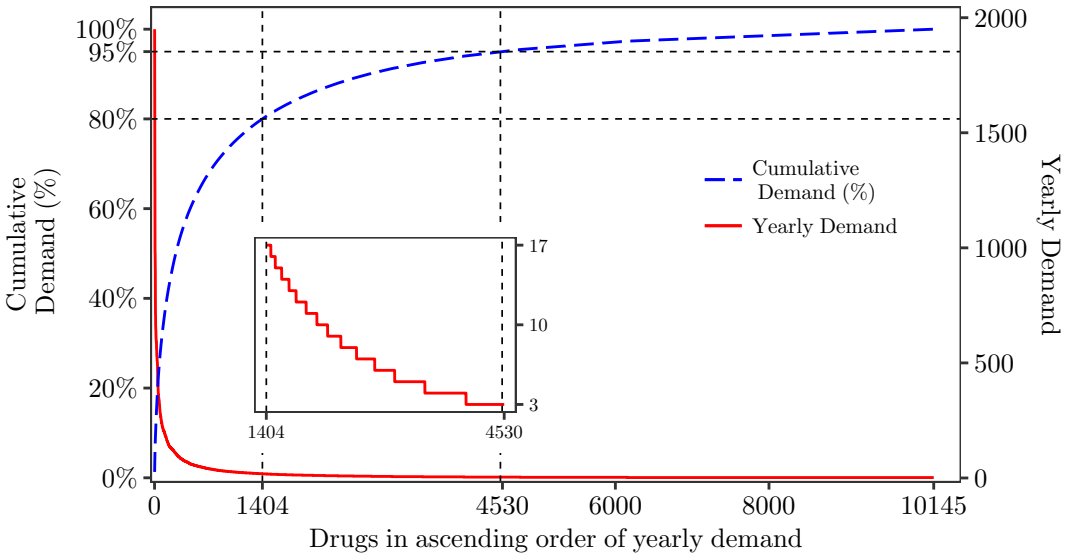


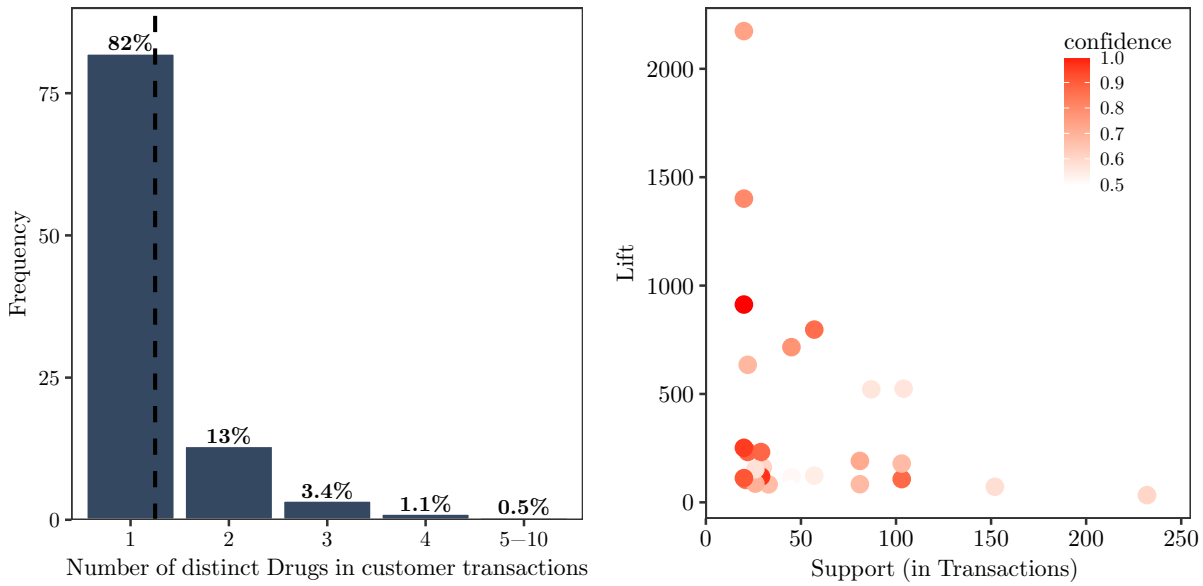
Figure 2.3: Cumulative demand distribution

(1404) of the GPI-QTYs capture 80% of the pharmacy sales. So to achieve a service level of 80%, it is sufficient to stock the top 14% of drugs. However, at higher service levels, the assortment problem is nontrivial as another 3126 drugs numbered from 1404 to 4530 in Figure 2.3 represent (31% of drugs) and capture only 15% of the sales. These drugs have yearly demand between 3 and 17 with no particular seasonal trends or patterns throughout the year. As MedAvail's target service level exceeds 90%, a large number of drugs with low and erratic demand must be considered in making the assortment decisions. Moreover, supplier-driven substitution is expected to have a significant impact on overall stock levels and required kiosk capacity. Due to such random and low demand, fitting theoretical distributions such as Normal and Poisson suffer from over or underestimation of the lead time demand leading to sub-optimal stocking decisions and consequently erroneous service levels. This motivates the use of a data-driven approach in making stocking decisions.

2.3. Co-ordering of Drugs

While making stocking decisions, one should also consider the possibility of co-ordering of drugs in a prescription. For prescriptions with multiple medications, a customer transaction is less likely to be successful if one of the prescribed drugs is not stocked. Figure 2.4(a) shows the co-ordering distribution of the transactions recorded in the year 2015 where 82% of the transactions record only one drug, and the average number of drugs in a transaction equals 1.25.

We use the Apriori association rule algorithm (Agrawal et al. 1994) to determine SKUs that frequently appear together in prescriptions. It proceeds by first identifying drugsets that frequently occur in the transactions. A drugset is a set containing one or more drugs. Frequent drugsets are determined using a minimum threshold known as *threshold support*. Support, $supp(X)$ of a drugset X is calculated as the number of times the drugset appears over the total number of transactions in the year 2015. If the support of a drugset is less than the threshold support, it is excluded from further analysis. We set threshold support to be $\frac{15}{D}$, where D is the number of transactions recorded in the year 2015. Threshold support of $\frac{15}{D}$ allows us to analyze association rules among top 20% of frequently ordered products that capture more than



(a) (b)

Figure 2.4: Figure (a) shows the drug co-ordering distribution and Figure (b) compares significance of association rules generated from Apriori association algorithm where threshold support is set to 15 and minimum confidence is 0.5

80% of the total sales. Once the frequently ordered drugsets are selected based on threshold support, the confidence for all pairs of drugsets is computed. The confidence, $conf\{X \Rightarrow Y\}$ is the probability of purchasing drugset Y when drugset X is purchased. In our case, we select a minimum confidence of 0.5 to ensure all rules that are likely to exist are selected i.e., probability is greater than 50%. The results of the algorithm are presented in Figure 2.4(b) where $\text{Lift}(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{supp(Y)}$ measures the significance of a rule. A total of 47 association rules between different drugs are found. For better decision-making, these association rules should be taken into account when making assortment and stocking decisions. We do not explicitly incorporate the effect of the association between drugs. However, we detail the justification in Chapter 3 where we note that at higher service levels, i.e., greater than or equal to 80%, all SKUs with yearly demand greater than or equal to 15 are selected. As such, all drugsets in 47 association rules found are already stocked.

Based on descriptive analysis of data, the following insights are derived

- Drug demand is sporadic and does not fit a known theoretical distribution. A data-driven optimization approach should therefore be used to make optimal operational decisions.
- Drugs are ordered in various quantities and as such, the substitution effect should be considered where multiple packages of a stocked GPI-QTY could be dispensed to satisfy the demand of another GPI-QTY as long as they share the same GPI code, and the quantities match.
- Around 97.5% of customer orders consists of prescription drugs that cannot be substituted at customer request. As such, modelling customer-driven substitution is not required in the context of pharmacy kiosks.
- The effect of co-ordering of drugs seems to be insignificant and therefore we do not model it to ensure tractability of the optimization model is maintained, allowing us to solve large-scale instances.

Chapter 3

Capacity & Assortment Planning under Supplier-driven Substitution

3.1. Introduction

MedAvail Technologies launched their first self-serve pharmacy kiosk, namely “MedCenter" (see Figure 3.1) in 2013 that have now been successfully deployed in U.S., Canada, and Switzerland where they are installed in pharmacies, retail stores, hospitals, community clinics, university campuses, and medical office buildings. MedCenter consists of multiple bins, each divided into several slots where a single slot can store various packages each containing a specific drug of a particular quantity. A customer is connected to a remote pharmacist who reviews scanned customer prescription and verifies if the drug is available in a package with the exact requested quantity. The customer is then instructed to make the payment and the pharmacist authorizes the release of the prescription. If medications are not stocked, a customer may request the pharmacist to call the physician for a substitute, to transfer the prescription to the home pharmacy, or to just cancel the order request.

MedCenter faces inventory challenges due to its limited storage capacity. The existing kiosk, developed to complement pharmacy operations, may store up to 1000 packages. In the MedCenter, a package is an SKU containing a specific drug of a specific quantity. Analysis of historical

Figure 3.1: MedAvail's MedCenter Kiosk (MedAvail 2017)

data in Chapter 2 shows that there are thousands of drugs, each ordered in four different quantities, on average. As such, inventory decisions are crucial in achieving high service levels when capacity is limited. Past data also shows that 60% of the failed transactions occur for three main reasons: (1) the drug is not stocked, (2) the drug is stocked but is currently out of stock, and (3) the drug is stocked but a package with the exact requested quantity is not available.

In this chapter, we address these stocking challenges by developing a modelling framework that determines the required capacity of the kiosk to achieve a desired level of service. Capacity is defined as the total number of packages stored which equals the sum of stock levels of all GPI-QTYs and therefore depends on the assortment of drugs to be stocked and corresponding stock levels. The stock level of a GPI-QTY is determined by its own demand and the demand of other GPI-QTYs it substitutes as well as the replenishment policy and the target service level. We develop three scenario-based stochastic optimization models that decide on optimal assortment, inventory, and supplier-driven substitution decisions. The proposed optimization models are tractable and CPLEX can solve large-scale instances with 30,000 GPI-QTYs. We also present a column-generation based heuristic solution approach that allows us to further reduce computational times by a factor of 3 at the expense of 1.1% optimality gaps, on average.

The remainder of the chapter is organized as follows. Section 3.2 reviews the related work in the literature. In particular, we review previous work on vending machine systems, newsvendor problems, and assortment problems under product substitution. Optimization models are formally defined in Section 3.3. In Section 3.4, we present a column-generation based heuristic approach to solve large-scale instances. In Section 3.5, we present model results for the capacity planning problem faced by MedAvail and analyze the effects of supplier-driven drug substitution and replenishment lead time on kiosk capacity. To further generalize model results, the product substitution is studied using randomly generated data and managerial insights are derived. We also compare the computational performance of the proposed column generation approach with CPLEX and Benders decomposition. Finally, some concluding remarks and future research directions are presented in Section 3.6.

3.2. Literature Review

The work in this chapter relates to the literature on vending machine systems, capacitated newsvendor problem, and the assortment problem under one-way substitution. In this section, we review previous work in each stream and position our work accordingly.

3.2.1 Vending Machine Systems

Internet of Things (IoT) is expected to be the next technological revolution aimed at creating a smart world through a network of interconnected smart devices (Kim et al. 2017). A smart vending machine or a self-serve kiosk is one such example of the use of IoT-enabled technology towards the smart world. Potential business gains from implementing smart vending machines were first studied by Shin et al. (2009) through a Beijing-based case study during the 2008 Olympics. The authors show that a digitized vending machine equipped with providing real-time information relating to game schedules, weather, and maps significantly improved revenues. For large-scale deployment of IoT-enabled vending machines, Solano et al. (2017) present a cost affordable solution using free Web services and technologies to minimize the total cost of ownership (TCO) for vending operators while enhancing customer experience.

Vending operators are also faced with operational challenges associated with managing vending machines including inventory planning, assortment problem, and replenishment interval. Rusdiansyah and Tsao (2005) are the first to address the operational planning problem for vending machines. The authors propose a multi-period inventory-routing problem for a network of single-product vending machines with constant demand rates. One of the earliest work on multi-product vending machine operation problem is by Poon et al. (2010). The authors consider a network of IoT-enabled kiosks connected to a cloud with sales and inventory level data available in real-time. They propose a replenishment index to formulate a simple feasible replenishment plan aimed at minimizing stock-out and transportation costs. Park and Yoo (2013) incorporate stock-out based product substitution to kiosk inventory management problem to decide on replenishment point and inventory up-to level for multiple products.

The mentioned studies are based on IoT-enabled kiosks and stochasticity in demand is ignored. On the other hand, Grzybowska et al. (2020) study the smart vending machine inventory planning under stochastic demand along with the optimal allocation of products within the machine using a simulation-optimization framework. These studies are however based on the assumption that demand distribution is either known or could be approximated through empirical data. However, analysis of pharmacy data reveals that it is hard to approximate product demand due to its sporadic nature. On top of that, during the early stages of kiosk operations, inventory planning has to be based on the limited amount of transactional data. To address this, Lin et al. (2011) propose a set-covering problem to select a set of products with maximal attributes to maximize kiosk's coverage. Later, when transactional data is recorded, a decision-tree based predictive model is used to decide on products to be stocked. The work by Lin et al. (2011) is mainly focussed on demand prediction and products are simply stocked in the order of their expected profits until the vending machine is full. In contrast, we follow a data-driven optimization approach where demand is not exactly known but is rather represented by the empirical distribution. To the best of our knowledge, our work is the first to address the inventory and assortment planning problem for vending machine systems with supplier-driven one-way substitution.

3.2.2 Assortment & Inventory Decisions

MedAvail wants to determine optimal stock levels for GPI-QTYs with stochastic demand to minimize kiosk storage capacity while ensuring that desired service level is achieved. This problem is related to the well-known *Constrained Multi-Product Newsvendor Problem (CMPNP)* where a newsvendor wants to determine a single-period optimal stocking policy for multiple products with stochastic demand and resource or budget constraint(s). The literature that deals with stochastic modelling approaches for the newsvendor problem assume that the demand distribution is known. In this stream, Hadley and Whitin (1963) are the first to study a CMPNP and propose a Lagrangian-based method to solve the problem. Fractional stock levels are allowed and to obtain an integer solution, the optimal order quantity is approximated by rounding down to the nearest integer value. Such an approach, however, performs poorly when the demand for products is low. To overcome the issue, Hadley and Whitin (1963) propose a dynamic programming procedure that is computationally inefficient when the products size is large and the largest instance reported in the paper consists of three products only. Nahmias and Schmidt (1984) extends the work of Hadley and Whitin (1963) and propose multiple heuristic approaches to solve the problem efficiently. The approach is however only applicable for moderate-to-high demand items as the proposed solution methodologies use continuous decision variables. The authors argue that for low-demand items a discrete model would be more applicable. Lau and Lau (1996) observe that the methodology proposed by (Hadley and Whitin 1963) may lead to negative optimal order quantities when the capacity is tight. The authors present an extension of the procedure in (Hadley and Whitin 1963) to deal with general demand distributions including positive lower bounds. Abdel-Malek et al. (2004) propose a closed-form expression of optimal order quantities when the demand follows a uniform distribution and present a generic iterative method to find near-optimal solutions for other general distributions. To avoid the issue of negative order quantities, Abdel-Malek and Montanari (2005) suggest the use of thresholds to help decision makers remove products with low marginal utilities. A binary search method applicable to both continuous and discrete demand distribution is proposed by Zhang et al. (2009). The proposed solution approach, however, does not guarantee optimality for the discrete distribution. For a comprehensive review on uncapacitated and single newsvendor problems with known demand distribution, we refer the reader to Turken et al. (2012).

In stochastic models, the literature assumes a known distribution and could not be applied in our case where demand is highly erratic and low. Since the demand for each GPI-QTY is erratic and low, experimentation with fitting Negative binomial, Poisson, and Normal distributions reveals that demand does not follow any specific probability distribution. This is true for many real-life problems where the exact distribution is rarely known and is generally approximated based on historical data. This explains the issue of poor out-of-sample performance in stochastic optimization approaches. To address this, robust optimization approaches are proposed in the literature. In this stream, Vairaktarakis (2000) considers a robust CMPNP under the assumption that the demand distribution for each item is completely unknown and only a set of discrete demand scenarios are available. The author presents minmax regret formulations with the objective to minimize expected costs under the worst-case realization of demand. The scenario-based minmax modelling approach is often criticized for being overly conservative as outliers in the historical data are not excluded. Such a minmax approach could be used for the pharmacy kiosk problem but our results show that it performs poorly. The poor performance is not due to the overly conservative nature of the model but rather it is unable to provide robust solutions due to the fewer number of scenarios for the kiosk problem with thousands of SKUs. To deal with the issue of overly conservative solutions in minmax regret formulations, a standard approach is to assume that the demand for each item could deviate from its nominal demand while the total deviation for all items is controlled by a user-defined budget of uncertainty (see, for example, Bertsimas and Thiele (2006b), Lin and Ng (2011a)). However, under service level maximization objective, the adversarial problem in robust optimization is nonlinear and as such, a tractable robust counterpart formulation does not exist. In addition, mathematical formulations for such models are complex and difficult to understand for managers. We therefore adopt a scenario-based stochastic optimization framework where all values of demand for each GPI-QTY recorded in the past data are used. Such an approach does not require the probability associated with each scenario and is therefore appropriate in our case where the probability density functions of GPI-QTYs are not known. In order to obtain robust solutions, we generate robust scenarios using the maximum demand of each GPI-QTY over all stores data in a given time period.

In CMPNP literature, the objectives considered optimize costs, profits, or the probability to achieve a target profit under different criteria (Khouja 1999). Our objective is to deter-

Table 3.1: Literature on Newsvendor Problem

Paper	Multi Product	Capacitated	Uncertainty[1]	Distribution[2]	Service level	Objective[3]	Variable type[4]	Methodology[5]	Problem size
Our work	✓	✓	**S**	**E**	✓	**Max CP/S**	**D**	**MILP**	**30000**
Hadley and Whitin (1963)	✓	✓	S	K		Max P	C	L+DP	3
Nahmias and Schmidt (1984)	✓	✓	S	K		Max P	C	L+H	5000
Aardal et al. (1989)			S	K	✓	Min C	C	CF	***
Moon and Choi (1994)			S	F	✓	Min C	C	L+IA	***
Lau and Lau (1996)	✓	✓	S	K		Max P	C	L	1000
Erlebacher (2000)	✓	✓	S	K		Max P	C	CF	***
Vairaktarakis (2000)	✓	✓	S	E		Minmax R	D	DP	***
Chen and Chuang (2000)			S	K	✓	Min C	C	CF	***
Abdel-Malek et al. (2004)	✓	✓	S	K		Min C	C	GIM	6
Bertsimas and Thiele (2006b)		✓	R	I		Max P	D	MILP	1
Taleizadeh et al. (2008)	✓	✓	S	K	✓	Min C	D	GA	15
Zhang et al. (2009)	✓	✓	S	K		Max P	C/D	BSM	6
Taleizadeh et al. (2009)	✓	✓	S	K	✓	Max P+S	D	GA	15
Choi et al. (2011)	✓		S	K		Max PR	C	CF	10
Lin and Ng (2011a)	✓	✓	R	I		Minmax R	C	L	50
Waring (2012)			S	K	✓	Max P	C	L	1
Jammernegg and Kischka (2013)			S	K	✓	Max MDR	C/D	CF	1

Acronyms:
(1) S - stochastic, R - Robust; (2) K - known, U - unknown, E - Empirical, F - distribution free, I - interval data
(3) P - profit, C - Cost, R - Regret, PVDI - Value of perfect distribution information, MDR - mean deviation rule, PR - profits under risk, S - Service level, CP-Capacity
(4) C - continuous, D - Discrete
(5) L - lagrangian, H - heuristc, MILP - mixed integer linear programming, CF - closed form, DP - Dyanmic programming, GA - Genetic algorithm
(5) GIM - Generic iterative method, BSM - binary search method, IA - iterative algorithm

Table 3.2: Literature on Assortment Problem under one-way supplier-driven substitution

Paper	Multi Product	Capacitated	Uncertainty[1]	Distribution[2]	Service level	Objective[3]	Variable type[4]	Methodology[5]	Problem size
Our work	✓	✓	**S**	**E**	✓	**Max CP/S**	**D**	**MILP**	**30000**
Sadowski (1959)			S	K		Min L	C/D	DP	***
Pentico (1974)			S	K		Min C	D	DP	***
Pentico (1976)			D	K		Min C	D	MILP	***
Tryfos (1985)			S	K		Max P	C	CF	10
Leachman and Glassey (1987)	✓	✓	S	K		Min C	C	***	***
Bagchi and Gutierrez (1992)	✓		S	K	✓	Max P	C	CF	3
Wollmer (1992)	✓	✓	S	K		Max R	C	H	5
Chand et al. (1994)			D	K		Min C	D	MINLP + DP	***
Bassok et al. (1999)	✓		S	K		Max P	D	IA	2
Rajaram and Tang (2001)	✓		S	K		Max P	C	H	&
Rao et al. (2004)	✓		S	K		Max P	D	MILP+H	25
Dutta and Chakraborty (2010)	✓		F	U		Max P	C	NSP	2
Deflem and Van Nieuwenhuyse (2013)	✓		S	K		Min C	C	CF	2
Ahiska et al. (2017)	✓		S	K		Max P	C	LSA	3
Hsieh and Lai (2019)	✓		D	K		Max P	C	GTM	2

Acronyms:
(1) D - deterministic, S - stochastic, F - fuzzy (2) K - known, E - Empirical, U - unknown
(3) P - profits, C - Costs, L - loss, R - Revenue, CP - Capacity, S - Service level (4) C - continuous, D - Discrete
(5) H - heuristc, MILP - mixed integer linear programming, CF - closed form, DP - Dyanmic programming, IA - Iterative Algorithm
(5) MINLP - mixed integer nonlinear programming, NSP - Numerical search procedure, LSA - Local search algorithm, GTM - Game-theoretical model

mine minimum kiosk capacity under service level constraints. The modelling approaches in the CMPNP literature do not explicitly model service level constraints and understocking is penalized through shortage costs that are included in the objective function. Studies that do consider service levels (see Table 3.1) in CMPNP (Chen and Chuang 2000, Taleizadeh et al. 2008, 2009, Waring 2012, Abdel-Aal et al. 2017) include service level constraints for each item and use a well-defined cumulative distribution function of the demand to define the service level as the probability of meeting demand with a given stock level. However, such an approach is not applicable in our case since demand is erratic and low and does not follow a known distribution. We use fill rate to define service level as the proportion of successful transactions with given stock levels of GPI-QTYs over a planning horizon of one year. Moreover, the service level in our problem is defined for the kiosk rather than for each GPI-QTY.

3.2.3 Substitution Decisions

Another challenge is to make substitution decisions along with stocking decisions under stochastic demand. Product substitution, in general, is defined as the act of using one product to meet the demand of another product. In inventory and assortment planning literature, substitution is categorized as either *supplier-driven* or *customer-driven* (Shin et al. 2015). In customer-driven substitution, customers decide on substitution when their preferred product is not available. In such problems, customer behavior is modelled within the optimization framework, see, for example, (Gaur and Honhon 2006, Kök and Fisher 2007, Aydin and Porteus 2008). In this stream of literature, Gaur and Honhon (2006) consider an uncapacitated multiproduct assortment planning problem where the demand follows a known distribution and the goal is to decide on the stock level for each product such that the expected profits are maximized. A utility-based locational choice model is used to estimate the customer demand where substitution between the products is allowed based on the substitution rate. For each customer, the utility it derives from product j is calculated and it is assumed that a customer prefers the product that maximizes his/her utility. If such a product is not available, he/she may select the second highest utility product with a probability defined by substitution rates. Kök and Fisher (2007) model the assortment problem using an exogenous demand model where the demand and substitution rates are precomputed using regression models and are then used to decide on the number of facings

allocated to each product under a capacity constraint.

These models do not make substitution decisions but rather consider customers' substitution behavior to decide on assortment and stock levels. We do not incorporate such customer behavior in our modelling approach since customer orders at pharmacy stores predominantly consist of prescribed drugs (97.5% of sales) which cannot be substituted by other drugs at the request of the customer. However, incorporating customer substitution behavior within a model making supplier-driven substitution decisions is a promising future research work. We refer the reader to Kök et al. (2008) and Shin et al. (2015) for a comprehensive review of literature on customer-driven substitution. From here onward, the term "substitution" refers to supplier-driven substitution unless explicitly mentioned otherwise.

At a pharmacy kiosk, a pharmacist may dispense multiple packages of one GPI-QTY to satisfy the demand of another GPI-QTY as long as they share the same GPI code, and the quantities match. This is known as supplier-driven substitution where the supplier makes stocking decisions while taking into account product substitution (Shin et al. 2015). More specifically, such quantity based substitution is referred to in the literature as *one-way substitution* and is common in manufacturing and service industries such as semiconductor industry (Bassok et al. 1999), computer hardware industry (Leachman and Glassey 1987), and airline industry (Wollmer 1992). One-way substitution may improve the overall service level due to pooling. Potential benefits of one-way substitution in inventory management are detailed in Fuller et al. (1993).

The term assortment problem was first introduced by Sadowski (1959) who considers a problem of determining n steel beams of different strengths where the demand of a lesser strength beam is substitutable by a beam with greater strength. A similar problem in the apparel industry is considered by Tryfos (1985) where the manufacturer has to decide on the set of m sizes. In these two papers, demand patterns are described by continuous distributions. The modelling approach in these works only decides on whether a quantity is stocked or not. On the other hand, Pentico (1974) considers a single product ordered in different quantities following discrete probability distributions. The goal is to decide on the stock levels for each size while taking into account one-way substitution where a smaller stocked size can meet the demand of a larger unstocked size while incurring a substitution cost. The demand for each size is assumed to be probabilistic and some strong substitution assumptions are made in the paper. The author as-

sumes that to meet the demand for a larger stocked size, only the smallest stocked size could be used. It is also assumed that demand is realized in descending order of size. Moreover, capacity is incorporated implicitly as a fixed charge cost of stocking a given size. A dynamic programming approach is proposed to formulate and solve the problem. These assumptions greatly limit the applicability of the proposed model. Pentico (1976) relaxes the linear cost functions and substitution cost assumption in Pentico (1974) but considers deterministic demand. Chand et al. (1994) generalizes the problem in Pentico (1976) with infinite planning horizon. A different variant of demand uncertainty in the assortment problem is studied by Dutta and Chakraborty (2010) where the demand is fuzzy and lies within an interval data. Bassok et al. (1999), Rao et al. (2004), and Deflem and Van Nieuwenhuyse (2013) study multi-product assortment problem under downward substitution without incorporating storage or resource constraints. Bassok et al. (1999) present a two-stage profit maximization formulation with N products and N demand classes under full downward substitution. Rao et al. (2004) consider a similar problem but take into account setup costs while Deflem and Van Nieuwenhuyse (2013) derive optimality conditions where substitution outperforms separate stock levels for the two-item case. Ahiska et al. (2017) and Hsieh and Lai (2019) study one-way substitution for manufacturing industry problem where high-quality products substitute low-quality ones. Ahiska et al. (2017) formulate the problem using the Markov decision process while Hsieh and Lai (2019) use a game-theoretical modelling framework.

Pharmacy kiosk inventory & assortment problem poses new research questions within the assortment optimization literature that have not been studied before. As such, our work differs from existing literature in the following aspects.

1. Substitution rules considered in our work have not been studied before. The literature on supplier-driven substitution deals with problems where a high-quality product may substitute a lower quality one with one-to-one substitution i.e., to meet the demand of a single unstocked unit, only one unit of a higher quality item is dispatched. On the other hand, in our case, to meet the demand of a single unit, multiple packages must be dispensed to fulfill the demand while ensuring that the quantity dispensed is equal to the requested quantity. Such requirements are not handled by the models in the literature. From a modelling perspective, the exact requested quantity requirement leads to extremely complex mathematical models.

2. The models in the literature explicitly include substitution costs in the objective function. For instance, in a computer hardware industry, if a customer order of 4GB memory chip is not available, an 8GB memory chip may fulfill the demand with substitution cost equals to the difference between the prices of the two different memory chips. In our case, there are no explicit substitution costs. The latter is captured implicitly within the service level expression to avoid over-substitution that may lead to lower service levels. To the best of our knowledge, our work is the first to consider fill rate in assortment planning problems with one-way substitution. As shown in Table 3.2, other than (Bagchi and Gutierrez 1992), no work considers service level. In Bagchi and Gutierrez (1992), however, service level constraints are added for each item using a well-defined cumulative distribution function, and fractional stock levels are also allowed in the optimal solution.

3. A common assumption in assortment planning problems under one-way substitution is that demand for all items is realized at the same time. The problem is then formulated as a two-stage stochastic program. In the first stage, when demand is not realized, the formulation decides on the stock levels of each item while taking into account substitution. In the second stage when the demand is realized for all items, substitution decisions are made based on the given stock levels to meet the demand for all items. However, for a pharmacy kiosk, demand is realized in a dynamic fashion where customers arrive one at a time. Rao et al. (2004) correctly point out that dynamic substitution models are extremely complex. Such complex models are intractable for the large-scale capacity planning problem faced by MedAvail with around 30,000 GPI-QTYs. We therefore employ a stationary substitution policy i.e., the same substitution rules are employed throughout the planning horizon irrespective of the stock levels at any given time. However, to deal with the problem of dynamic customer arrivals, our models make robust substitution decisions which guarantee that the desired service level is always achieved irrespective of the sequence of demand realization for substitutable products.

4. Our proposed models are tractable for the pharmacy kiosk problem with 30,000 GPI-QTYs and could be solved using a commercial solver. Other models in the literature are too complex for the large-scale instances with thousands of GPI-QTYs.

3.3. Modelling Stocking and Assortment Decisions

The problem is to decide on the single period (replenishment lead time) stock level x_i, for each product $i \in I$ using the empirical distribution that is generated from historical data. When $x_i = 0$, product i is not stocked and the assortment is defined by $i \in I$ such that $x_i > 0$. We adopt a scenario-based stochastic optimization model that uses past data to generate T demand scenarios by dividing the planning horizon into $T = \left\lceil \frac{365}{h} \right\rceil$ lead time intervals, where h is the lead time. The demand A_{it} for $i \in I$, during time period $t \in \Theta = \{1, ..., T\}$, is calculated using historical sales data. Products are grouped in classes if they only differ by quantities. In the presence of substitution, The demand d_{it}, depends on the substitution variable s_{ij}, which equals 1 if product i substitutes product j. The latter is only possible if products i and j belong to the same product class and quantity q_j is a multiple of quantity q_i. A $0-1$ incidence matrix $\mathbf{b} = [b_{ij}]$ is computed where $b_{ij} = 1$ if product j is substitutable by i. As such, $d_{it} = \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} s_{ij}$, where $m_{ij} = \frac{q_j}{q_i}$ units of product i are required to meet the unit demand for product j. Note that we employ a stationary substitution policy where it is assumed that the demand is realized at once during the start of the period and substitution decisions s_{ij} do not change over time and stock level.

Substitution variables may be either predetermined or optimized within a mathematical model. In the pharmacy kiosk application, each GPI is a product class containing GPI-QTYs sharing the same GPI code. Multiple packages have to be dispensed to meet the demand for a higher quantity. Such a substitution arises for a variety of other industrial applications where a requested quantity could be substituted by multiple packages of smaller quantities. For instance, in the case of a Bank ATM, customer requests for $100 could be met by dispensing five currency notes of $20. Similarly, for a grocery store/vending machine, a customer may be willing to accept six 250ml bottles of Coke if a 1.5 liter family pack is not available.

Since there are no backorders, any unsatisfied demand is a lost sale. The lost sales for a product $i \in I$ during time period $t \in \Theta$ is $\max\{0, d_{it} - x_i\}$. Lost sales occur either because the drug is not stocked, i.e., $x_i = 0$, or observed demand in some period t exceeds the stock level, i.e., $x_i < d_{it}$. At a pharmacy kiosk, unsatisfied demand is lost because a customer is most likely going to use another pharmacy and not wait for the medication to be back-

ordered. The same applies to other kiosk applications such as Bank ATM and vending machines, etc. Since unsatisfied demand is lost, we model the problem with no backorders which also justifies single-period stock planning. The expected service level or fill rate is calculated as $1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}}{D}$ where D is the total yearly demand. Our goal is to determine the capacity such that the desired service level α is met.

We develop three optimization models to solve the capacity planning problem and address management's questions under three different substitution rules: (1) no substitution, i.e., $s_{ii} = 1$ and all other substitution variables take value 0, (2) management's substitution rule, (3) optimized substitution. In rules (1) and (2), substitution is predefined. We now discuss the models under predefined and optimized substitution.

3.3.1 Predefined substitution

The first model [M1] decides only on optimal stock levels for products using one of the predefined substitution rules, and minimizes the capacity under service level constraint. Given the substitution rule, demand scenarios d_{it} for each product $i \in I$ are precomputed and serve as input data to the model. The formulation is

$$\text{[M1]:} \quad \min \sum_{i \in I} x_i \tag{3.1}$$

$$\text{s.t.} \quad 1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}}{D} \geq \alpha \tag{3.2}$$

$$x_i \in \mathbb{Z}^+, \qquad \forall \quad i \in I, t \in \Theta, \tag{3.3}$$

where the objective function (3.1) minimizes the total number of packages stocked i.e., required capacity of the kiosk under the assumption that all packages occupy equal space. The latter could be easily relaxed by replacing objective function coefficients with products' space requirements. Constraint (3.2) ensures that the expected service level, $1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}}{D}$, is greater than or equal to the desired service level, α. Finally, constraint (3.3) is the nonneg-

ative integer requirement on x_i. The above formulation is nonlinear due to max functions in constraint (3.2). The latter may be linearized by introducing auxiliary variables (f_{it}, y_{it}) and replacing constraint (3.2) with the following set of constraints

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \geq \alpha, \tag{3.4}$$

$$f_{it} \geq 0, \qquad \forall \quad i \in I, t \in \Theta, \tag{3.5}$$

$$f_{it} \geq d_{it} - x_i, \qquad \forall \quad i \in I, t \in \Theta, \tag{3.6}$$

$$f_{it} \leq (d_{it} - x_i) + M \times y_{it}, \quad \forall \quad i \in I, t \in \Theta, \tag{3.7}$$

$$f_{it} \leq 0 + M \times (1 - y_{it}), \quad \forall \quad i \in I, t \in \Theta, \tag{3.8}$$

$$y_{it} \in \{0, 1\}, \qquad \forall \quad i \in I, t \in \Theta, \tag{3.9}$$

where M is a significantly large number. If $d_{it} > x_i$, y_{it} must be equal to 0 for the problem to be feasible. Constraint (3.6) is then $f_{it} \leq d_{it} - x_i$ and constraint (3.7) is $f_{it} \leq M$. As such, $f_{it} = d_{it} - x_i$. On the other hand, if $d_{it} < x_i$, $y_{it} = 1$ for the problem to be feasible and $f_{it} = 0$. The problem, however, becomes challenging to solve due to binary variables y_{it}. We therefore present a relaxed formulation [R1] where constraints (3.7) and (3.8) are dropped

$$\text{[R1]:} \quad \min \sum_{i \in I} x_i \tag{3.10}$$

$$\text{s.t.} \quad f_{it} \geq d_{it} - x_i \qquad \forall \quad i \in I, t \in \Theta, \tag{3.11}$$

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \geq \alpha, \tag{3.12}$$

$$x_i \in \mathbb{Z}^+, f_{it} \geq 0, \quad \forall \quad i \in I, t \in \Theta, \tag{3.13}$$

and prove in Lemma 1 that its optimal solution $\mathbf{x}^* = [x_i^*]$ is also optimal to the original model [M1].

Lemma 1. *An optimal solution* $\mathbf{x}^*$ *for model [R1] is also optimal to the original model [M1].*

Proof. Let $(\mathbf{x}^* = [x^*_{ij}], \mathbf{f}^* = [f^*_{it}])$ be an optimal solution to model [R1]. Rearranging constraint (3.12),

$$\sum_{i \in I} \sum_{t \in \Theta} f^*_{it} \leq (1 - \alpha) \times D$$

In model [R1], f^*_{it} may take a value greater than the max term $\max\{0, d_{it} - x^*_i\}$ in constraint (3.2). As such, $\sum_{i \in I} \sum_{t \in \Theta} f^*_{it} \geq \sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x^*_i\}$ and

$$(1 - \alpha) \times D \geq \sum_{i \in I} \sum_{t \in \Theta} f^*_{it} \geq \sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}$$

This implies $\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x^*_i\} \leq (1 - \alpha) \times D$ and constraint (3.2) holds for $\mathbf{x}^*$. This proves that solution $\mathbf{x}^*$ is feasible to the original model [M1].

Let z^*_{M1} and z^*_{R1} be the optimal objective function values for models [M1] and [R1], respectively. Since [R1] is a relaxed formulation of model [M1], $z^*_{\text{R1}} \leq z^*_{\text{M1}}$. Since the original model [M1] can not have a solution superior than z^*_{R1}, $\mathbf{x}^*$ is also optimal for [M1]. □

Note that f_{it} is simply an analysis variable used to linearize model [M1]. One could adjust its value after solving the model [R1] by setting $f^*_{it} = \max\{0, d_{it} - x^*_i\}$. Model [R1] is a new variant of the well-known single-period newsvendor problem under a service level constraint and could be applied to any inventory problem where the service level needs to be considered while making stocking decisions. In addition to the capacity minimization objective, the model is easily extendable for profit maximization or cost minimization objectives.

3.3.2 Optimized substitution

We develop two additional models that extend [M1] to optimize both stocking and substitution decisions. Model [M2] decides on substitution and stock levels to minimize storage capacity under a service level constraint. The parameter d_{it} in model [M1] is now a decision variable in [M2] as the model makes substitution decision s_{ij}. As such, model [M2] has two additional

decision variables : d_{it} and s_{ij}. The formulation is then as follows.

$$\text{[M2]:} \quad \min \sum_{i \in I} x_i \tag{3.14}$$

$$\text{s.t.} \quad d_{it} = \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} s_{ij} \qquad \forall i \in I, t \in \Theta, \tag{3.15}$$

$$\sum_{\substack{i \in I: \\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \tag{3.16}$$

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}}{D} \geq \alpha, \tag{3.17}$$

$$s_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in I, \tag{3.18}$$

$$x_i \in \mathbb{Z}^+, \; d_{it} \in \mathbb{Z}^+ \qquad \forall i \in I, t \in \Theta, \tag{3.19}$$

where the objective function (3.14) is the same as (3.1). Constraint (3.15) computes demand d_{it} of a product $i \in I$ in period $t \in \Theta$ taking into account the demand of products it substitutes. Constraint (3.16) ensures that each product $j \in I$ is substituted by exactly one product. If $s_{ii} = 1$, it implies that product $i \in I$ is not substituted by any other product. Constraint (3.18) is the binary requirement on variable s_{ij} and constraints (3.19) are nonnegative integer requirements on variables x_i and d_{it}. Constraint (3.17) defines the service level and is the same as constraint (3.2) in model [M1]. It may be linearized using the same approach discussed earlier for model [M1]. Note that substitution variables s_{ij} only change d_{it} to a decision variable and constraints (3.4) - (3.9) are valid for model [M2]. As such, the relaxed formulation [R2] for model

[M2] is

$$\text{[R2]:} \quad \min \sum_{i \in I} x_i \tag{3.20}$$

$$\text{s.t.} \quad (3.15), (3.16), (3.18), (3.19)$$

$$f_{it} \geq d_{it} - x_i, \qquad \forall i \in I, t \in \Theta \tag{3.21}$$

$$1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \geq \alpha, \tag{3.22}$$

$$f_{it} \geq 0, \qquad \forall i \in I, t \in \Theta. \tag{3.23}$$

Lemma 1 holds trivially and an optimal solution $(\mathbf{x}^*, \mathbf{s}^*)$ to model [R2] is also optimal for [M2].

Model [M3] is developed to maximize the expected service level of a kiosk under a capacity constraint. The decision variables are the same as in [M2], and the mathematical formulation is as follows:

$$\text{[M3]:} \quad \max \quad \alpha = 1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} \max\{0, d_{it} - x_i\}}{D} \tag{3.24}$$

$$\text{s.t.} \quad (3.15), (3.16), (3.18), (3.19),$$

$$\sum_{i \in I} x_i \leq C, \tag{3.25}$$

where the objective function (3.24) maximizes the expected service level α and constraint (3.25) ensures that the total number of packages stored is restricted to capacity, C. As in model [M2], [M3] is also nonlinear due to the max terms in the objective function. However, to linearize it,

we only introduce analysis variable f_{it}. The linear formulation is

$$\text{[R3]:} \quad \max \quad \alpha = 1 - \frac{\sum_{i\in I}\sum_{t\in\Theta} f_{it}}{D} \tag{3.26}$$

$$\text{s.t.} \quad (3.15), (3.16), (3.18), (3.19), (3.25),$$

$$f_{it} \geq d_{it} - x_i, \qquad \forall i \in I, t \in \Theta \tag{3.27}$$

$$f_{it} \geq 0, \qquad \forall i \in I, t \in \Theta \tag{3.28}$$

where constraint (3.27) along with nonnegativity constraint (3.28) ensure that $f_{it} \geq \max\{0, d_{it} - x_i\}$. At optimality, $f^*_{it} = \max\{0, d^*_{it} - x^*_i\}\ \forall i \in I, t \in \Theta$ and is proven in Lemma 2.

Lemma 2. *For model [R3], given an optimal solution* $(\mathbf{x}^*, \mathbf{f}^*, \mathbf{s}^*, \mathbf{d}^*)$, $f^*_{it} = \max\{0, d^*_{it} - x^*_i\}\ \forall i \in I, t \in \Theta$.

Proof. Note that constraints (3.27) and (3.28) ensure that $f^*_{it} \geq \max\{0, d^*_{it} - x^*_i\}\ \forall i \in I, t \in \Theta$.

We now prove by contradiction that at optimality, f^*_{it} can not take a value greater than the max term. Assume that $(\mathbf{x}^*, \mathbf{f}^*, \mathbf{s}^*, \mathbf{d}^*)$ is optimal with objective function value z^* and $f^*_{it} > \max\{0, d^*_{it} - x^*_i\}\ \exists i \in I, t \in \Theta$. Let $(\mathbf{x}^*, \mathbf{f^a}, \mathbf{s}^*, \mathbf{d}^*)$ be the adjusted solution with objective function value z^a where $f^a_{it} = \max\{0, d^*_{it} - x^*_i\}$. As such,

$$\sum_{i\in I}\sum_{t\in T} f^*_{it} > \sum_{i\in I}\sum_{t\in T} f^a_{it} \Longrightarrow \left(1 - \frac{\sum_{i\in I}\sum_{t\in T} f^*_{it}}{D}\right) < \left(1 - \frac{\sum_{i\in I}\sum_{t\in T} f^a_{it}}{D}\right) \Longrightarrow z^* < z^a$$

which contradicts the assumption that z^* is optimal. This proves that if $f^*_{it} > \max\{0, d^*_{it} - x^*_i\}\ \exists i \in I, t \in \Theta$, there always exists a better solution $f^a_{it} = \max\{0, d^*_{it} - x^*_i\}$ for which $z^a > z^*$. □

Models [R2] and [R3] are extensions of the capacitated newsvendor problem under supplier-driven substitution. Since 97.5% of customer orders consist of prescribed drugs that cannot be substituted by other drugs at the request of the customer, we only model one-way supplier-driven substitution where a pharmacist may dispense multiple packages of one GPI-QTY to

meet the demand of another sharing the same GPI code. As such, the proposed models are specific to supplier-driven substitution and do not readily handle customer-driven substitution.

Note that the proposed models are generic and apply to any demand values. However, when demand is less sporadic, a single period model that uses moments of the demand distribution may become useful. Under dynamic customer arrivals, our models make substitution decisions that are robust against the sequence of demand realization for substitutable products i.e., the desired service level is guaranteed irrespective of the order in which customers arrive. This is detailed next.

3.3.3 Substitution under dynamic customer arrivals

Models [R2] and [R3] make substitution decisions under the worst-case sequence of demand for substitutable products. We first explain this using an illustrative example and present a formal proof in Theorem 1. Consider two products i and j, in the same product class and let $q_i = 20$ and $q_j = 60$. Since q_j is a multiple of q_i, assume that product i substitutes j, and $m_{ij} = \frac{60}{20} = 3$. In a given period t, let $A_{it} = 10$, $A_{jt} = 1$, $D = 10 + 1 = 11$, and stock level $x_i = 10$. If product j is requested when less than three packages of product i are available, then the number of failed transactions equals 1 and demand for product i is fully met. As such, service level $\alpha = 1 - \frac{1}{11} = 91\%$. However, if product j is requested when at least three packages of product, i are available, the demand for product j is fulfilled and there is a shortage of three packages to meet the demand for product i. In this case, $\alpha = 1 - \frac{3}{11} = 73\%$. Depending on the sequence of demand realization, the service level either equals 91% or 73%. Proposed mathematical models calculate service level as $\alpha = 1 - \frac{f_{it}}{D} = 1 - \frac{3}{11} = 73\%$ if product i substitutes j. We now show that substitution decisions are robust against the sequence of demand realization.

Theorem 1. *Substitution decisions are robust against the sequence of demand realization for substitutable products and guarantee that desired service level is achieved.*

Proof. We first present the exact formula to compute the number of failures f_{it}^E. Then, we show that $f_{it} \geq f_{it}^E$ for any sequence of demand realization.

Given a solution $\mathbf{s}$, let $K_i = \{1, 2, ..., n-1, n\}$ be the set of products substituted by product $i \in I$ i.e., $s_{ij} = 1 \forall j \in K_i$. Without loss of generality, assume that the set K_i is ordered such that the sequence of demand realization is

$$A_{nt} \rightarrow A_{n-1,t} \rightarrow \cdots \rightarrow A_{2t} \rightarrow A_{1t} \tag{3.29}$$

Let f_{it}^j be the number of failures and x_i^j be the number of packages available for product $j \in K_i$. The exact formula for the number of failures is

$$f_{it}^j = \lceil \max\{0, A_{jt} - \frac{x_i^j}{m_{ij}}\} \rceil \tag{3.30}$$

where $\frac{x_i^j}{m_{ij}}$ computes the demand that could be met for product j using product i. Note that since $\frac{x_i^j}{m_{ij}}$ can take fractional values, the value $max\{0, A_{jt} - \frac{x_i^j}{m_{ij}}\}$ needs to be rounded up to the nearest integer value. Within an optimization model, one may linearize constraint (3.30) as

$$f_{it}^j \geq 0 \tag{3.31}$$

$$f_{it}^j \geq A_{jt} - \frac{x_i^j}{m_{ij}} \tag{3.32}$$

$$f_{it}^j \in \mathbb{Z} \tag{3.33}$$

Constraints (3.31) and (3.32) ensure that $f_{it}^j \geq \max\{0, A_{jt} - \frac{x_i^j}{m_{ij}}\}$ while integer requirement (3.33) rounds up f_{it}^j to the nearest integer value.

Given the demand sequence (3.29), $x_i^n = x_i$ and $x_i^j = x_i - \sum_{k=j+1}^{k=n} \left(A_{kt} - f_{it}^k\right) \times m_{ik}$ where

$\left(A_{kt} - f_{it}^k\right) \times m_{ik}$ is the number of packages of product i already used for product k. As such,

$$\begin{aligned}
f_{it}^n &\geq A_{nt} - \frac{x_i}{m_{in}} \\
f_{it}^j &\geq A_{j,t} - \frac{x_i - \sum_{k=j+1}^{k=n} \left(A_{kt} - f_{it}^k\right) \times m_{ik}}{m_{ij}} \qquad \forall \quad j \in K_i \setminus \{n\} \\
f_{it}^j &\in \mathbf{Z}^+ \qquad \forall \quad j \in K_i
\end{aligned} \tag{3.34}$$

The exact total number of failures is then

$$f_{it}^E = \sum_{j \in K_i} f_{it}^j \tag{3.35}$$

Given solution $s_{ij} = 1 \; \forall j \in K_i$, we rewrite constraint (3.15) as $d_{it} = \sum_{j \in K_i} m_{ij} A_{jt}$. Constraints (3.21) and (3.27) in models [R2] and [R3] are then

$$f_{it} \geq \sum_{j \in K_i} m_{ij} A_{jt} - x_i \tag{3.36}$$

We now show that $f_{it} \geq f_{it}^E$ for any sequence of demand realization. Let $\widetilde{f}_{it}^j = m_{ij} f_{it}^j$ and rearranging constraints (3.34),

$$\begin{aligned}
m_{in} f_{it}^n = \quad \widetilde{f}_{it}^n &\geq \quad m_{in} A_{nt} - x_i \\
m_{ij} f_{it}^j = \quad \widetilde{f}_{it}^j &\geq \quad m_{ij} A_{j,t} - \left(x_i - \sum_{k=j+1}^{k=n} \left(A_{kt} - f_{it}^k\right) \times m_{ik} \right) \qquad \forall \quad j \in K_i \setminus \{n\} \\
\widetilde{f}_{it}^j &\geq \quad 0 \qquad \forall \quad j \in K_i
\end{aligned} \tag{3.37}$$

Note that since $\mathbf{A}, \mathbf{m}, \mathbf{x}$ are integers, $\widetilde{f}_{it}^j$ always takes an integer value. The integrality require-

ment on $\widetilde{f}_{it}^{j}$ is therefore dropped. Since $m_{ij} \geq 1$,

$$\sum_{j \in K_i} \widetilde{f}_{it}^{j} \geq \sum_{j \in K_i} f_{it}^{j}. \tag{3.38}$$

Setting $x_i^j = x_i - \sum_{k=j+1}^{k=n} \left(A_{kt} - f_{it}^{k}\right) \times m_{ik} = 0 \; \forall j \in K_i \setminus \{n\}$, we have

$$\begin{aligned} \overline{f}_{it}^{n} &\geq m_{in} A_{nt} - x_i, \\ \overline{f}_{it}^{j} &\geq m_{ij} A_{j,t} && \forall j \in K_i \setminus \{n\}, \\ \overline{f}_{it}^{j} &\geq 0 && \forall j \in K_i, \end{aligned} \tag{3.39}$$

and

$$\sum_{j \in K_i} \overline{f}_{it}^{j} \geq \sum_{j \in K_i} \widetilde{f}_{it}^{j} \tag{3.40}$$

$$\sum_{j \in K_i} \overline{f}_{it}^{j} \geq \sum_{j \in K_i} m_{ij} A_{jt} - x_i \tag{3.41}$$

Since $f_{it} \geq \sum_{j \in K_i} m_{ij} A_{jt} - x_i$, then by inequalities (3.35), (3.38), (3.40), and (3.41)

$$f_{it} = \sum_{j \in K_i} \overline{f}_{it}^{j} \geq \sum_{j \in K^i} \widetilde{f}_{it}^{j} \geq \sum_{j \in K^i} f_{it}^{j} = F_{it}^{E} \Longrightarrow f_{it} \geq f_{it}^{E}$$

This shows that for any given sequence of demand realization, $f_{it} \geq f_{it}^{E}$. Let z^E be the service level achieved when the exact number of failures are computed while z^* be the service level using f_{it}. Then,

$$\sum_{i \in I} \sum_{t \in T} f_{it} \geq \sum_{i \in I} \sum_{t \in T} f_{it}^{E} \Longleftrightarrow \left(1 - \frac{\sum_{i \in I} \sum_{t \in T} f_{it}}{D}\right) \leq \left(1 - \frac{\sum_{i \in I} \sum_{t \in T} f_{it}^{E}}{D}\right) \Longleftrightarrow z^* \leq z^E \tag{3.42}$$

which proves that desired service level z^* is always achieved irrespective of the sequence of demand realization. □

3.4. A Column-Generation Based Heuristic Approach

In many practical problems, the optimization models are of large-scale and it may be impossible to explicitly include all variables in the initial formulation or it may consume too much memory. Column generation is a well-known procedure to solve such large-scale problems where columns are added at each iteration of the simplex method. The idea of column generation was first suggested by Ford Jr and Fulkerson (1958) for *multicommodity network flow* problem and have been successfully applied to many real-life problems including cutting stock problems (Gilmore and Gomory 1961, 1963), crew scheduling (Desaulniers et al. 1997), and vehicle routing (Agarwal et al. 1989). Oğuz (2002) show that column generation may even be efficient for some problems where the number of variables is low enough to be explicitly included in the model. In our problem, all variables can be explicitly included in the formulation but it consumes too much memory and thus slowing down CPLEX. In particular, the pharmacy kiosk problem consists of too many GPI-QTYs and only a few could be stocked due to limited capacity. As such, one may include variables x_i and s_{ij} only for products that are most likely to be stocked.

We present a column-generation based heuristic approach (CGA) to solve model [R3] to near optimality by selecting only a subset of products in the initial formulation. Other products are then added iteratively. The approach is also applicable for the other two models, [R1] and [R2]. Let $\widetilde{I} \subseteq I$ be the set of products selected for the initial formulation. We rewrite model [R3] and drop integer requirements to formulate the restricted master problem [RMP] as

$$\text{[RMP]:} \quad \max \quad 1 - \frac{1}{D} \times \sum_{i \in I} \sum_{t \in \Theta} f_{it} \tag{3.4.43}$$

$$\text{s.t.} \quad \sum_{i \in I} x_i \leq C, \qquad [\lambda] \tag{3.4.44}$$

$$\sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} s_{ij} - x_i - f_{it} \leq 0 \quad \forall i \in I, t \in \Theta, \qquad [u_{it}] \tag{3.4.45}$$

$$\sum_{\substack{i \in I: \\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \qquad [\omega_j] \tag{3.4.46}$$

$$x_i, s_{ij} \geq 0 \qquad \forall i \in \widetilde{I}, j \in I : b_{ij} = 1, \tag{3.4.47}$$

$$f_{it} \geq 0, \qquad \forall i \in I, j \in I, t \in \Theta, \tag{3.4.48}$$

where [.] are dual variables for each constraint. Constraint (3.4.46) along with nonnegativity constraint (3.4.48) ensures $s_{ij} \leq 1$, and we therefore do not include this constraint to the model. We select a subset of products to initialize the algorithm. A subset should be selected such that it minimizes the number of iterations required to add the columns. To do so, we sort products in decreasing order of the number of substitution a product can make, and the yearly demand. We then select top $\frac{C}{2}$ products with highest yearly demand and the number of substitutions. This allows us to start off with products that are likely to be stocked due to higher demand and their ability to substitute the demand for other products. Variables x_i and s_{ij} $\forall j \in I$ are introduced for the selected products. Products that are not selected, we only include variable s_{ii} i.e., either its demand is met using one of the selected products or $s_{ii} = 1$, and the demand for such product is never met. Once model [RMP] is solved, its dual information is used to

determine potential products to be added to the model. Taking the dual,

$$\text{[RMP-D]:} \quad \min \quad 1 + C\lambda + \sum_{j \in I} \omega_j \tag{3.4.49}$$

$$\text{s.t.} \; -u_{it} \geq -\frac{1}{D} \qquad i \in I, t \in \Theta, \qquad [f_{it}] \tag{3.4.50}$$

$$\lambda - \sum_{t \in \Theta} u_{it} \geq 0 \qquad \forall i \in I, \qquad [x_i] \tag{3.4.51}$$

$$\sum_{t \in \Theta} m_{ij} A_{jt} u_{it} + \omega_j \geq 0 \qquad \forall i \in I, j \in I : b_{ij} = 1 \qquad [s_{ij}] \tag{3.4.52}$$

$$\lambda \geq 0, \; u_{it} \geq 0, \omega_j \rightarrow \text{urs} \qquad \forall i \in I, j \in I, t \in \Theta \tag{3.4.53}$$

Given λ and u_{it}, the reduced cost is $RC_i = \lambda - \sum_{t \in \Theta} u_{it}$ for product i. Let $\mathcal{I}$ be the set of products not included in the initial formulation. The pricing problem $\min_{i \in \mathcal{I}} \{\lambda - \sum_{t \in \Theta} u_{it}\}$ determines the product with most negative reduced costs which is then added to [RMP]. For most of the problems in the literature, enumerating over all possible columns is computational impractical and therefore a pricing subproblem is solved to determine the column to be added. In our case, however, one could easily calculate reduced costs for all products. Instead of selecting the product with most negative reduced cost, we select all products with reduced cost $RC_i < 0$ and columns x_i and s_{ij} $\forall j \in I$ are added to [RMP] which is solved again. This procedure terminates when $RC_i \geq 0 \forall i \in I$, and the latest [RMP] solution provides a lower bound to the original model [R3]. The other approach could be to first solve model [RMP] with all variables. Then, for products with positive RC_i, variables x_i and s_{ij} $\forall j \in I$ are removed. However, it turns out that such an approach is computationally inefficient compared to the proposed column generation approach.

To obtain a feasible solution, [RMP] is solved with integrality constraints on x_i and s_{ij} and its objective function value is an upper bound to model [R3]. Note that this approach does not guarantee optimality. To solve to optimality, one needs to apply the CGA at each node of the branch-and-bound tree. However, we implement CGA only at the root node and computational results in Section 3.5.3 show that optimality gap is 1.1%, on average.

3.5. Results

We perform numerical testing over several datasets including seven pharmacy store sales data and randomly generated instances. In Section 3.5.1, we use the proposed optimization models to determine the optimized storage capacity for MedAvail's pharmacy kiosk and recommend assortment and stocking guidelines using pharmacy sales data. To further generalize model results, we solve model [R2] using randomly generated instances in Section 3.5.2 and derive managerial insights. Finally, the proposed column generation solution approach is compared against CPLEX and Benders decomposition in Section 3.5.3.

3.5.1 The case of MedAvail

In this section, we first use models [R1] and [R2] to analyze the effects of substitution and replenishment lead time on the capacity of a kiosk using single pharmacy store data for the year 2015. The data records 2,355 GPIs (or product classes) and 10,145 GPI-QTYs (or products). The goal is to assess the savings in kiosk capacity through drug substitution and through reducing replenishment lead time from two days to one day. The management suggested that it is useful to explore the effect of capacity on the service level, as it may not be possible to build a machine of an optimized capacity. Therefore, we use model [R3] to determine the maximum service level achieved at different capacity levels as suggested by the management. We then perform several experiments using multiple datasets generated from seven 24/7 pharmacy store sales data to provide bounds on the service level that management should expect to achieve at a given capacity. All optimization models are coded in C++ and solved using CPLEX version 12.6.3 on a 64-bit Windows 10 with Intel(R) Core i5-5300U 2.30GHz processors and 4.00GB RAM. We solve all instances to an optimality gap of 0.5% since solving the problem to optimality may only reduce the required capacity by at most 57 ($11,497 \times 0.5\%$, see Table 3.3) for service levels of up to 99%. They were of the view that such an exact machine could not be built and the optimized capacity values be rounded off to the nearest 100. Finally, we evaluate the computational efficiency of the proposed column generation approach against solving model [R3] directly using CPLEX.

3.5.1.1 Effects of substitution

Service level, α	[R2]	[R1]-MedAvail's substitution		[R1]-no substitution	
	Capacity, C	Capacity, C	$\triangle\%$ to [R2]	Capacity, C	$\triangle\%$ to [R2]
80%	2,542	2,710	6.6%	2,618	3.0%
85%	3,261	3,449	5.8%	3,385	3.8%
90%	4,375	4,606	5.3%	4,583	4.8%
95%	6,485	6,856	5.7%	6,938	7.0%
96%	7,233	7,604	5.1%	7,686	6.3%
97%	7,980	8,506	6.6%	8,690	8.9%
98%	9,460	10,002	5.7%	10,186	7.7%
99%	10,956	11,497	4.9%	11,681	6.6%

Table 3.3: Kiosk storage capacity to achieve desired service level α under different substitution rules.

The management was inclined towards a predefined substitution criterion rather than a complex mathematical model. So we optimized capacity under various substitution strategies to see whether substitution plays a role in deciding on the capacity of the kiosk. We generate a dataset using one pharmacy store sales data for the year 2015 with replenishment lead time, $h = 2$. Model [R1] is solved under two distinct substitution rules: (1) MedAvail's substitution rule, and (2) no substitution. MedAvail's substitution rule was suggested by the management where a GPI-QTY i substitutes GPI-QTY j with the same GPI code if the quantity of j is twice that of i and its average lead time demand is less than 25% of that of i, or if the quantity of j is three times that of i and its average lead time demand is less than 15% of that of i. An iterative procedure is used to assign values to the substitution variables s_{ij} based on this rule, and d_{it} is calculated apriori. Model [R2] is solved to determine optimized substitution. Each model is solved repeatedly by varying the desired service level α between 80% and 99%. Table 3.3 summarizes the results.

At 95% service level, the capacity under optimized substitution is 6,485. It increases by 5.7% when MedAvail's rule is used and by 7.0% when substitution is not allowed. As the service level decreases, the effect of MedAvail's substitution rule decreases. In fact, the effect becomes negative relative to no substitution when the service level is 90% or lower. This is due to over substitution by MedAvail's substitution rule at lower service levels. At lower service levels, fewer GPI-QTYs should be substituted to optimize the capacity. Optimized substitution, as ex-

pected, is always better than both no substitution and apriori rules. This comes at the expense of larger solution times. Given the potential improvements in capacity under optimized substitution, model [R2] is used in subsequent analysis.

3.5.1.2 Effect of replenishment lead time

Before the start of the project, kiosks were being replenished every other day. MedAvail management wanted to investigate the effect of replenishment lead time on capacity and assortment decisions. A larger lead time is expected to increase capacity since lead time demand would be higher, so we experimented with 1 and 2 day lead times. Table 3.4 summarizes the results where [R2] is solved at eight different service levels. At 90% service level, the capacity is reduced by 14% when the replenishment lead time is reduced from two days to one day. Although a one-day lead time may increase the operating costs of the kiosk due to frequent replenishment, management believes that the significant reduction in capacity is much more important when taking into account the technical challenges in designing a kiosk with higher capacity. Testing in subsequent sections is based on daily replenishment.

To study the significance of co-ordering, we report the highest yearly demand among all GPI-QTYs that are not stocked in Table 3.4. For a one-day lead time ($h = 1$) and service level $\alpha = 80\%$, every GPI-QTY with a yearly demand greater than 15 is stocked. Recall that in the Section *Co-ordering of drugs*, we set threshold support to $\frac{15}{D}$ based on the idea to capture association rules among top 20% of drugs capturing more than 80% of total demand. The results in Table 3.4 show that the demand threshold is always less than or equal to the threshold support used in the Apriori association rule algorithm. As such, GPI-QTYs that frequently appear to-

Service level, α	Capacity			Threshold demand	
	C_1	C_2	$\triangle\%$	$h = 1$	$h = 2$
80%	2,093	2,542	18%	14	15
85%	2,743	3,261	16%	11	11
90%	3,769	4,375	14%	6	7
95%	5,745	6,485	11%	3	4
99%	10,161	10,956	7%	2	2

Table 3.4: We compare storage capacity C_h at one day ($h = 1$) and two day ($h = 2$) lead time. Threshold demand is the highest yearly demand among all GPI-QTYs that are not stocked.

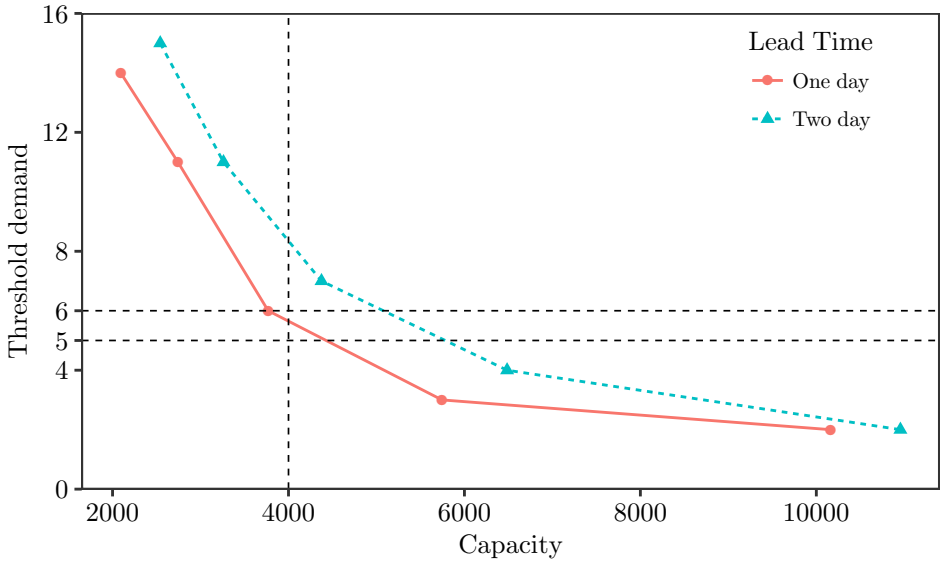


Figure 3.2: The graph plots threshold demand against storage capacity under different lead times.

gether are already stocked when the service level is $\alpha \geq 80\%$. As expected, threshold demand decreases as α increases. Therefore, we do not need to incorporate association rules explicitly in our modelling framework. Given the lead time and kiosk capacity, Figure 3.2 may be used as easy-to-use guidelines to decide on which medications to store without solving the assortment problem. For instance, if MedAvail decides on one day replenishment lead time for a kiosk with capacity $C = 4,000$, threshold demand is 5.6 based on Figure 3.2. As such, MedAvail should stock all GPI-QTYs with yearly demand greater or equal to 6.

3.5.1.3 Capacity planning over multiple pharmacies

A crucial question we faced in deriving demand distributions from the data is whether to use individual store data or multiple stores data and whether to use average or maximum observed demands in the latter case. Each of these approaches may have merits and drawbacks. We carried several tests to answer this question. At this point, management suggested that it is useful to explore the effect of limited capacity on the service level, as it may not be possible to build a machine of an optimized capacity. Hence, we modified the objective to service level

maximization and added a constraint that limits capacity to obtain model [R3]. The results presented next are based on service level maximization where capacity is varied between 2,000 and 7,000 with an increment of 1,000.

Individual store data (IAS) The IAS approach makes stocking and substitution decisions for each store individually using its yearly demand data. The expected service levels achieved at seven pharmacy stores are shown in Table 3.5a. On average, setting the capacity to 5,000 achieves a service level of 92.5%. The drawback of the IAS approach is that it may lead to overestimation of the service level due to over-fitting, also referred to as *optimizer's curse* in the Operations Research literature. Overfitting leads to stocking decisions that are susceptible to small changes in demand which could lead to much worse service levels. IAS approach therefore provides an upper bound on the service level achieved.

Most-active store data (MSD) To avoid overfitting, we make stocking and substitution decisions using the most active store data, i.e., the one with the highest yearly sales. The optimized decisions are then applied to all other stores to calculate their achieved service levels. The results in Table 3.5b highlight the problem of overfitting with the IAS approach. The expected service levels are substantially reduced when the optimal solution from the most active store is applied to other stores. On average, the service level achieved at capacity $C = 5,000$ is 84.3%.

Although the MSD approach addresses the problem of overfitting, it ignores GPI-QTYs ordered at other stores. The number of distinct GPIs recorded in the year 2015 at a store varies between 2,316 and 2,509. However, when the data is aggregated for all stores, the total number of distinct GPIs equals 3,579. Similarly, the number of distinct GPI-QTYs recorded at the most active store equals 12,014. This number increases to 29,626 when all store data is analyzed. As such, an optimal solution derived based on one store may be suboptimal for other stores and only provides a lower bound on the service level.

Average demand over all stores (ADS) Both IAS and MSD approaches use single-store data and ignore GPI-QTYs ordered at the other stores. To overcome this, we generate a new dataset by calculating the average demand of a GPI-QTY in time period $t \in \Theta$, over all stores. We use

Capacity C	Store ID							Average
	S1	S2	S3	S4	S5	S6	S7	
2,000	78.9%	78.3%	77.4%	77.1%	75.9%	75.9%	74.5%	76.9%
3,000	86.3%	86.0%	85.2%	85.1%	84.1%	84.0%	83.0%	84.8%
4,000	90.6%	90.4%	89.8%	89.7%	89.0%	88.9%	88.0%	89.5%
5,000	93.4%	93.4%	92.8%	92.7%	92.1%	92.0%	91.3%	92.5%
6,000	95.3%	95.4%	94.7%	94.1%	94.4%	94.2%	93.6%	94.5%
7,000	96.5%	96.5%	96.2%	96.1%	95.8%	95.7%	95.1%	96.0%

(a) Service level achieved at different stores at different capacities (IAS)

Capacity C	Store ID						
	S1	S2	S3	S4	S5	S6	S7
2,000	69.4%	69.6%	71.3%	70.1%	70.3%	71.1%	74.5%
3,000	76.5%	76.6%	78.1%	77.3%	77.8%	78.3%	83.0%
4,000	80.5%	80.7%	82.0%	81.4%	82.1%	82.3%	88.0%
5,000	83.8%	83.4%	84.6%	84.2%	85.0%	85.0%	91.3%
6,000	85.6%	85.1%	86.5%	86.0%	86.9%	86.9%	93.6%
7,000	87.1%	86.6%	88.1%	87.4%	88.6%	88.4%	95.1%

(b) Service level achieved using most-active store data to make stocking decisions (MSD)

Capacity C	Store ID							Average
	S1	S2	S3	S4	S5	S6	S7	
2,000	69.5%	69.9%	69.7%	68.7%	67.3%	67.3%	66.0%	68.3%
3,000	75.6%	76.4%	75.9%	75.2%	73.8%	73.7%	72.9%	74.8%
4,000	78.9%	79.7%	79.1%	78.7%	77.3%	77.3%	76.6%	78.2%
5,000	80.6%	81.5%	80.7%	80.8%	79.2%	79.4%	78.7%	80.1%
6,000	82.0%	82.5%	82.0%	82.1%	80.6%	81.2%	80.0%	81.5%
7,000	83.4%	84.1%	83.3%	83.7%	82.1%	82.6%	81.5%	83.0%

(c) Service level achieved using average demand over all stores to make stocking decisions (ADS)

Capacity C	Store ID							Average
	S1	S2	S3	S4	S5	S6	S7	
2,000	73.7%	74.6%	74.4%	73.8%	73.1%	71.9%	71.7%	73.3%
3,000	81.6%	82.2%	81.6%	81.5%	81.1%	79.6%	79.9%	81.1%
4,000	85.8%	86.6%	86.1%	85.9%	85.7%	84.4%	84.9%	85.6%
5,000	88.6%	89.1%	88.8%	88.6%	88.6%	87.2%	87.8%	88.4%
6,000	90.5%	91.0%	90.7%	90.6%	90.8%	89.7%	90.1%	90.5%
7,000	92.2%	92.6%	92.3%	92.4%	92.5%	91.6%	91.9%	92.2%

(d) Service level achieved using the highest demand across all stores in a given period $t \in \Theta$ to make stocking decisions (HDS)

Table 3.5: Capacity Planning using different demand prediction strategies

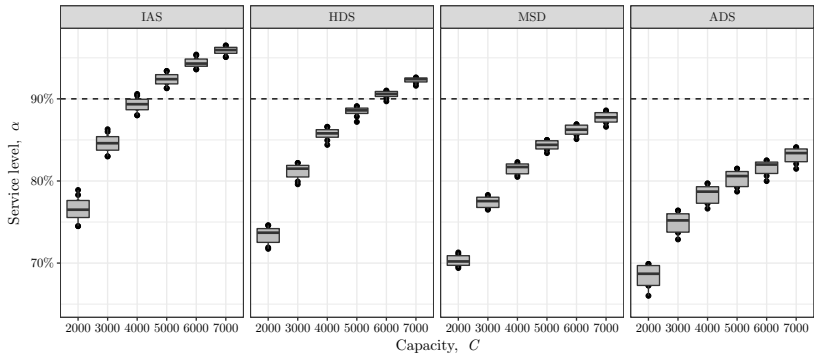
this new dataset containing all GPI-QTYs to make stocking and substitution decisions, which are then applied to stores data to calculate their achieved service levels. ADS approach results in poor stocking and substitution decisions as shown in Table 3.5c. When capacity $C = 5,000$, the average service level over all stores is 80.1%. This is due to the aggregation of demand which leads to reduced uncertainty. Consider a GPI-QTY i, with demand on a specific day at four stores as $\{0, 1, 0, 3\}$. If the stock level $x_i = 1$, then the number of failures at store 4 equals $3 - 1 = 2$. However, the average demand equals $\frac{0+1+0+3}{4} = 1$, and the calculated number of failures equals $1 - 1 = 0$. This example shows that averaging demand over all stores does not capture variability among stores, leading to suboptimal solutions and lower service levels.

Highest demand over all stores (HDS) Another approach is to use the maximum demand of each GPI-QTY in a given time period $t \in \Theta$ across all stores. The HDS approach provides better stock levels that are robust for all stores by making stocking and substitution decisions under the worst-case scenario. The results are summarized in Table 3.5d. At capacity $C = 5,000$, the service level achieved is 88.4% on average. The drawback of HDS is that it may overestimate stock levels for some SKUs as the decisions are made under the worst-case scenario. It is also possible that demand characteristics may vary from store to store and some GPI-QTYs ordered at one store may never be ordered at other stores.

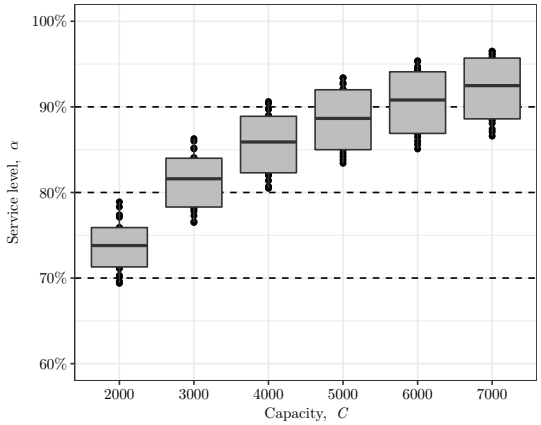
3.5.1.4 Recommendations

Computational results show that substitution and daily replenishment guidelines significantly reduce the capacity required to achieve the desired service level. We observe that MedAvail's substitution rule is not as effective as optimized substitution which may save up to 9% of capacity. On the other hand, daily replenishment saves up to 18% of capacity compared to two-day replenishment. The results also show that the marginal benefit of additional capacity decreases at higher capacities as illustrated in Figure 3.3(a) where the service level increases at a decreasing rate as the capacity increases. We also observe that the service level achieved at a fixed capacity is roughly the same across all stores as shown by the boxplots in Figure 3.3(a).

To present robust results, we perform several experiments at different service levels using four different demand prediction strategies. Other than ADS which results in suboptimal



(a) Comparative analysis of the stocking decision approaches



(b) Capacity planning using top 3 approaches (IAS,MSD,HDS)

Figure 3.3: Capacity Planning using different approaches

solutions, management may use any of the other three approaches discussed earlier. The approaches IAS and MSD provide upper and lower bounds on the service level, respectively. On the other hand, the HDS approach offers a more realistic expectation of the service level and gives stock levels that are robust against small changes in demand. Management may also make capacity decisions using a combination of the three approaches as illustrated in Figure 3.3(b). The boxplot represents the uncertainty in the service level achieved at a fixed capacity. For instance, when the capacity is set to 5,000, MedAvail should expect a service level between 84% and 93% depending on the store under consideration and the level of conservatism when making stocking decisions.

3.5.2 Numerical Analysis over Randomly Generated Instances

In this section, we solve model [R2] using randomly generated data instances to generalize the findings of the case study. Section 3.5.2.1 details the procedure employed to generate random data and in Section 3.5.2.2, we discuss model results and derive managerial insights.

3.5.2.1 Data Generation

To generate data instances, we consider 200 distinct product classes and randomly generate products for each class from a uniform distribution, $Unif[1, 10]$. To study the effect of substitution, three distinct substitution patterns are defined: (1) "None", QTYs = $\{2, 3, 5, 7, ...\}$, where product substitution is not possible as no product quantity is a multiple of another, (2) "Single", QTYs = $\{1, 2, 3, 5, ...\}$, where only the smallest quantity product is able to substitute all other quantities, and (3) "All", QTYs = $\{1, 2, 4, 8, ...\}$, where all smaller quantity products can substitute larger quantity product. To generate demand values, we randomly generate yearly demand for each product from an exponential distribution $Exp(\frac{1}{\mu})$ where μ is varied between 10 and 50 with increments of 10. Mean daily demand μ_i for product i is calculated as $\mu_i = \frac{Exp(\frac{1}{\mu})}{365}$ which is used to generate 200 demand scenarios from Poisson distribution, $Poi(\mu_i)$. Figure 3.4 plots the cumulative distribution of yearly demand for different values of μ. As μ increases, the product's probability of having high yearly demand increases. As such, increasing μ reduces the number of products with low yearly demand. Sensitivity analysis over μ allows us to study the effect of substitution under different demand settings where low values of μ imply low and erratic demand while setting higher values for μ implies less sporadic demand. Service level α is also set at eight different levels between 80% to 99%. For a given μ, substitution pattern, and service level, 5 random instances are generated, resulting in a total of 600 instances.

3.5.2.2 Results on Random Instances

Tables 3.6 and 3.7 summarize computational results for substitution patterns "Single" and "All", respectively. Average values over 5 randomly generated instances are reported in the tables. Column "None" under "Capacity" records the minimum capacity required to achieve desired

Mean	Service	Capacity			Product Substitution			Product Coverage		
Demand	level	None	Optimal	Δ% imp	Possible	Optimized	% Substituted	NbProducts	Covered	% Covered
10	80%	463	461	0.3%	872	21	2.4%	1072	460	42.9%
	85%	534	532	0.4%	872	30	3.4%	1072	529	49.4%
	90%	626	623	0.5%	872	42	4.9%	1072	616	57.4%
	95%	757	751	0.7%	872	71	8.2%	1072	730	68.1%
	96%	789	781	1.1%	872	73	8.3%	1072	760	70.8%
	97%	846	833	1.6%	872	113	13.0%	1072	796	74.2%
	98%	904	891	1.5%	872	121	13.9%	1072	819	76.4%
	99%	962	949	1.4%	872	115	13.2%	1072	849	79.1%
Average		734	728	0.9%	872	73	8.4%	1072	695	64.8%
20	80%	510	508	0.3%	872	15	1.7%	1072	502	46.8%
	85%	593	590	0.5%	872	25	2.9%	1072	580	54.1%
	90%	701	696	0.7%	872	36	4.1%	1072	676	63.0%
	95%	867	856	1.3%	872	69	7.9%	1072	796	74.2%
	96%	918	902	1.8%	872	80	9.2%	1072	834	77.8%
	97%	977	960	1.7%	872	93	10.6%	1072	860	80.2%
	98%	1047	1022	2.4%	872	99	11.3%	1072	904	84.3%
	99%	1164	1136	2.4%	872	128	14.7%	1072	942	87.8%
Average		846	834	1.4%	872	68	7.8%	1072	762	71.0%
30	80%	562	559	0.5%	872	18	2.0%	1072	546	50.9%
	85%	653	648	0.7%	872	25	2.9%	1072	626	58.3%
	90%	774	766	1.1%	872	37	4.2%	1072	725	67.6%
	95%	958	944	1.5%	872	58	6.6%	1072	846	78.9%
	96%	1016	999	1.7%	872	74	8.5%	1072	869	81.1%
	97%	1086	1063	2.1%	872	79	9.0%	1072	907	84.6%
	98%	1174	1151	2.0%	872	91	10.5%	1072	948	88.5%
	99%	1317	1283	2.6%	872	119	13.7%	1072	977	91.1%
Average		941	927	1.5%	872	63	7.2%	1072	806	75.1%
40	80%	590	587	0.5%	872	16	1.9%	1072	565	52.7%
	85%	690	685	0.7%	872	24	2.7%	1072	647	60.3%
	90%	823	815	1.0%	872	39	4.5%	1072	746	69.5%
	95%	1030	1013	1.6%	872	60	6.9%	1072	871	81.2%
	96%	1090	1073	1.7%	872	68	7.8%	1072	897	83.6%
	97%	1168	1146	1.9%	872	74	8.5%	1072	924	86.1%
	98%	1269	1241	2.2%	872	90	10.3%	1072	955	89.1%
	99%	1422	1381	3.0%	872	118	13.5%	1072	989	92.2%
Average		1008	993	1.6%	872	61	7.0%	1072	824	76.9%
50	80%	621	619	0.3%	872	14	1.7%	1072	583	54.4%
	85%	726	723	0.5%	872	19	2.2%	1072	666	62.1%
	90%	866	860	0.7%	872	32	3.7%	1072	762	71.0%
	95%	1087	1075	1.1%	872	51	5.9%	1072	877	81.8%
	96%	1154	1139	1.3%	872	58	6.7%	1072	909	84.8%
	97%	1240	1221	1.6%	872	69	7.9%	1072	939	87.6%
	98%	1353	1327	1.9%	872	80	9.2%	1072	970	90.4%
	99%	1519	1481	2.6%	872	94	10.8%	1072	999	93.2%
Average		1069	1056	1.3%	872	52	6.0%	1072	838	78.2%

Table 3.6: Numerical results for Random Instances under Substitution Pattern “Single"

Mean	Service	Capacity			Product Substitution			Product Coverage		
Demand	level	None	Optimal	Δ% imp	Possible	Optimized	% Substituted	NbProducts	Covered	% Covered
10	80%	463	457	1.4%	872	59	6.8%	1072	456	42.5%
	85%	534	525	1.6%	872	87	10.0%	1072	524	48.9%
	90%	626	612	2.3%	872	110	12.6%	1072	610	56.9%
	95%	757	732	3.3%	872	154	17.6%	1072	725	67.6%
	96%	789	761	3.7%	872	159	18.3%	1072	752	70.1%
	97%	846	796	6.3%	872	159	18.3%	1072	781	72.9%
	98%	904	853	6.0%	872	153	17.6%	1072	822	76.7%
	99%	962	911	5.6%	872	154	17.6%	1072	866	80.8%
Average		733	706	3.8%	872	129	14.8%	1072	692	64.5%
20	80%	510	501	1.7%	872	55	6.3%	1072	499	46.5%
	85%	593	579	2.4%	872	79	9.0%	1072	576	53.7%
	90%	701	678	3.3%	872	106	12.2%	1072	672	62.6%
	95%	867	826	5.0%	872	158	18.1%	1072	802	74.8%
	96%	918	865	6.1%	872	160	18.4%	1072	828	77.2%
	97%	977	920	6.1%	872	185	21.2%	1072	865	80.7%
	98%	1047	979	6.9%	872	202	23.2%	1072	902	84.1%
	99%	1164	1075	8.3%	872	186	21.3%	1072	939	87.6%
Average		843	803	5.0%	872	141	16.2%	1072	760	70.9%
30	80%	562	552	1.8%	872	59	6.7%	1072	546	50.9%
	85%	653	637	2.5%	872	79	9.0%	1072	625	58.3%
	90%	774	747	3.6%	872	109	12.5%	1072	724	67.5%
	95%	958	910	5.4%	872	147	16.8%	1072	849	79.1%
	96%	1016	955	6.4%	872	160	18.4%	1072	878	81.9%
	97%	1086	1014	7.1%	872	179	20.5%	1072	914	85.3%
	98%	1174	1095	7.2%	872	177	20.3%	1072	946	88.3%
	99%	1317	1207	9.1%	872	182	20.9%	1072	974	90.8%
Average		937	889	5.4%	872	136	15.6%	1072	807	75.3%
40	80%	590	580	1.7%	872	58	6.6%	1072	566	52.8%
	85%	690	672	2.6%	872	78	8.9%	1072	649	60.5%
	90%	823	794	3.6%	872	112	12.9%	1072	750	69.9%
	95%	1030	978	5.3%	872	162	18.5%	1072	876	81.7%
	96%	1090	1033	5.6%	872	163	18.7%	1072	905	84.4%
	97%	1168	1098	6.3%	872	183	21.0%	1072	932	86.9%
	98%	1269	1180	7.5%	872	180	20.6%	1072	961	89.6%
	99%	1422	1301	9.3%	872	190	21.8%	1072	988	92.2%
Average		1005	955	5.2%	872	141	16.1%	1072	828	77.3%
50	80%	621	610	1.8%	872	58	6.6%	1072	589	55.0%
	85%	726	707	2.7%	872	82	9.4%	1072	674	62.8%
	90%	866	836	3.6%	872	117	13.4%	1072	777	72.4%
	95%	1087	1035	5.0%	872	156	17.8%	1072	894	83.4%
	96%	1154	1094	5.5%	872	162	18.6%	1072	920	85.8%
	97%	1240	1167	6.2%	872	176	20.2%	1072	951	88.7%
	98%	1353	1262	7.2%	872	185	21.2%	1072	978	91.2%
	99%	1519	1403	8.3%	872	202	23.2%	1072	1005	93.7%
Average		1066	1014	5.1%	872	142	16.3%	1072	848	79.1%

Table 3.7: Numerical results for Random Instances under Subsitution Pattern: “ALL"

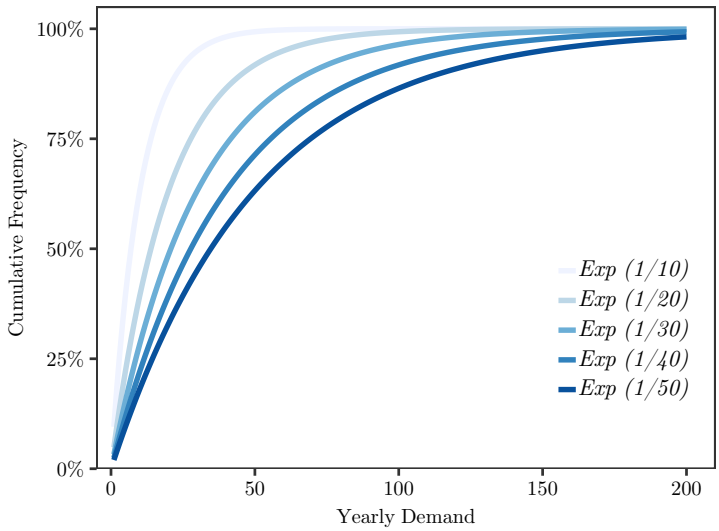


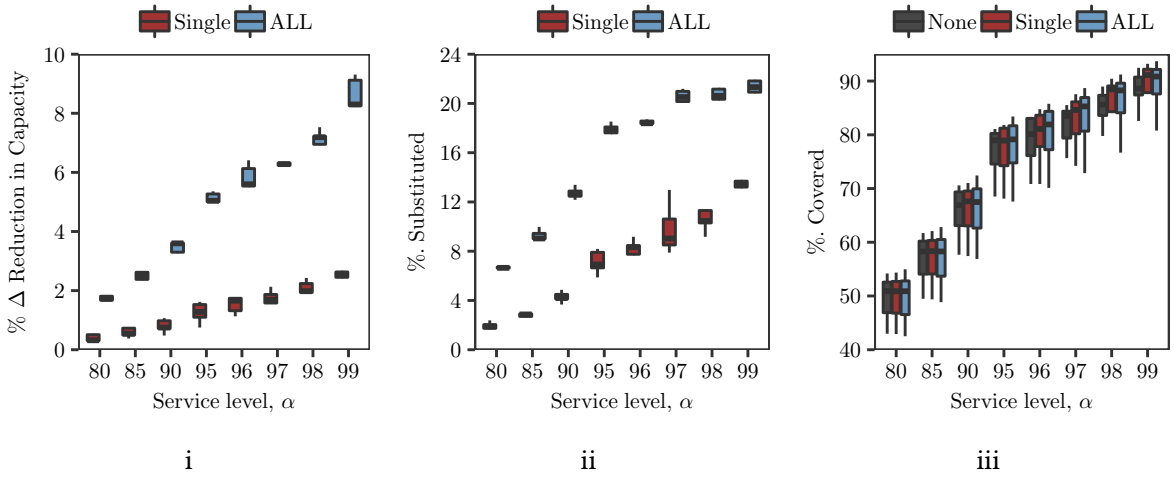
Figure 3.4: The graph plots exponential distribution under different mean values.

service level α without substitution and column “optimal" refers to the required capacity with substitution (“Single" or “All"). Column “$\Delta\%$ imp" denotes percentage reduction in required capacity due to substitution. The latter is calculated as the percentage difference in optimal capacity under a given substitution pattern (“Single" or “All") and substitution pattern “None". Column “Possible" counts the total number of products that can be substituted by other products, while Column “Optimized" counts the number of products substituted by other products in the optimal solution, i.e., $\sum_{\substack{i\in I:\\ b_{ij}=1}} \sum_{\substack{j\in I:\\ i\neq j}} s_{ij}$. Column “% Substituted" is the ratio of “Optimized" to “Possible". The total number of products considered are given in column “Nb. Products" out of which, “Nb. Covered" number of products are stocked or substituted by stocked products in the optimal solution. Column “%. Covered" is the percentage of products covered in each instance.

Results in Tables 3.6 and 3.7 show that substitution plays an important role in reducing the storage capacity at higher service levels. For instance, under substitution pattern “All" and $\mu = 40$, substitution is able to reduce the storage capacity by 9.3% when desired service level $\alpha = 99\%$. On the other hand, when $\alpha = 80\%$, the capacity is reduced by only 1.7%. This is further illustrated in Figure 3.5(a)(i) that plots a boxplot for the percentage reduction in capacity (PRC) at given a service level under each substitution pattern. At higher service levels,

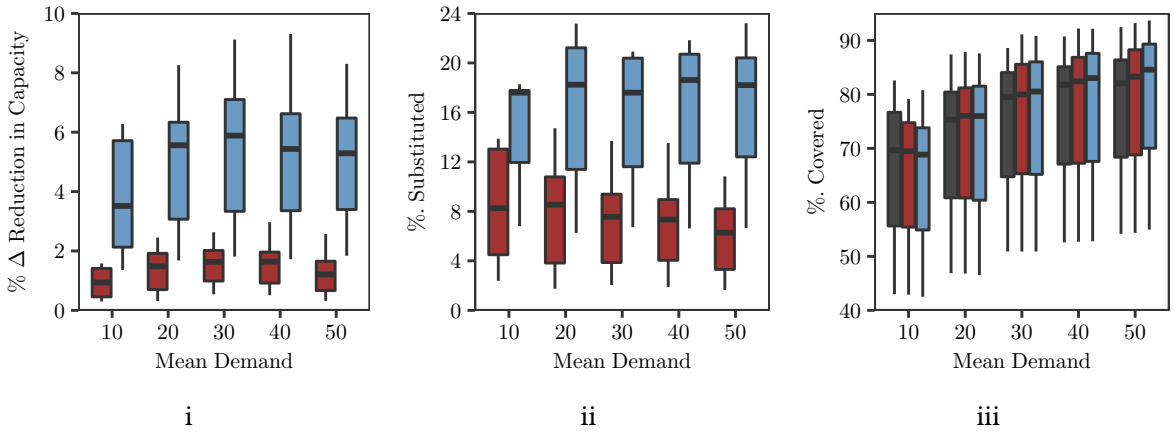
more products are substituted as shown in Figure 3.5(a)(ii). We observe that the effect of substitution is significant when more products are able to substitute. As shown in Figure 3.5(a)(ii), the percentage of products substituted is higher under substitution pattern “All" compared to “Single", and as such, PRC is significantly higher for “All". For instance, when $\mu = 30$, the average PRCs under patterns “Single and “All" are 1.5% and 5.4%, respectively. However, when product demand is generated from an exponential distribution with $\mu = 10$, PRC starts decreasing at higher service levels. Under substitution pattern “All" and $\alpha = 97\%$, PRC is 6.3% which decreases to 5.6% for $\alpha = 99\%$. This is because substitution negatively impacts the service level of the products substituting other products and at higher service levels, the negative effect outweighs the positive effect of substitution in improving the service level of the unstocked products. As such, product substitution is less preferred as shown in Table 3.7 where the number of products substituted decreases from 159 to 154 when the desired service level is increased from 97% to 99%.

Sensitivity analysis over mean demand μ shows that when the number of products with low demand is high, the effect of substitution on PRC is low. The effect of product demand is illustrated in Figure 3.5(b)(i) where PRC increases at a decreasing rate as the product demand increases. Under pattern “All" and $\mu = 10$, the average PRC is 3.8% which increases to 5.0% when $\mu = 20$. When all or most of the products have low demand, products' stock levels are low and cannot substitute higher quantity products that require multiple packages to be dispensed. In contrast, when μ changes from 30 to 40, PRC decreases from 5.4% to 5.2%. This is due to the fact at higher values of μ, product demand is high and it is preferred to stock a product rather than substituting it which would result in multiple packages to be dispensed, whenever it is ordered. This is illustrated in Figure 3.5(b)(ii) where product substitution does not increase with increasing μ. Figures 3.5(a)(iii) and 3.5(b)(iii) illustrate how product coverage is affected by service level and mean demand μ under each substitution pattern, respectively. The plots show that the effect of substitution in improving product coverage is not significant. For $\mu = 30$, the percentage of products covered is 75.1%, on average, under pattern “Single" which increases slightly to 75.3% under pattern “All". Analysis over demand shows that product substitution is preferred when there is a right balance between the number of products with low demand and the ones with high demand. Product substitution does not have a significant effect when most of the products have either high or low demand.



i ii iii

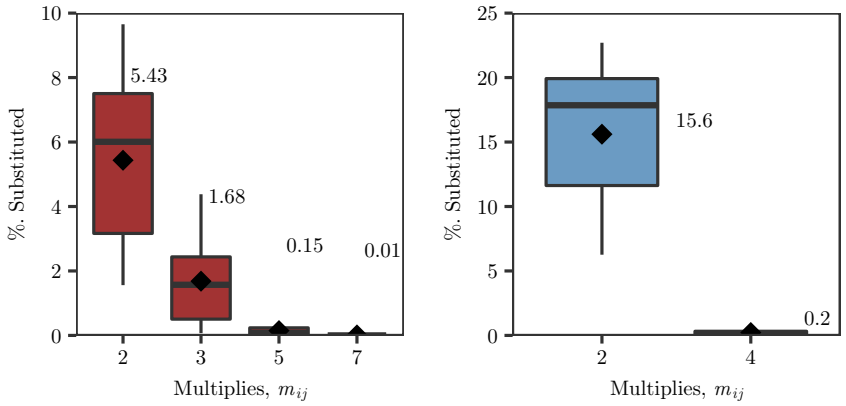
(a) Effect of Service level



i ii iii

(b) Effect of Product Demand

Figure 3.5: Figures (a) and (b) illustrate the effect of service level and product demand on percentage reduction in capacity (PRC), products substituted, and product coverage, respectively.



(a) Single (b) All

Figure 3.6: Figures (a) and (b) plot the percentage of products substituted with multiples m_{ij} under substitution pattern “Single" and “All", respectively.

We also study the effect of multiplies on product substitution as shown in Figures 3.6(a) and 3.6(b). Under substitution pattern “Single", the smallest quantity product only substitutes products with multiples m_{ij} less than or equal 7, where 5.3% of the products substituted have multiplies $m_{ij} = 2$ only 0.01% of the products are substituted with $m_{ij} = 7$. For pattern “All", where all smaller quantity products are able to substitute higher quantity products, only the products with $m_{ij} \leq 4$ are substituted. The results show that the cost of substitution is implicitly captured by m_{ij} and the product substitution is less preferred when m_{ij} increases. This is due to multiple packages being dispensed for the substituted products which would lead to fewer packages available for substituting product.

3.5.3 Analysis of Solution Approach: CGA

To test the computational efficiency of the column generation approach, we generate five instances using HDS data with 29,626 products by varying the kiosk capacity between 1,000 and 7,000. For each instance, we solve model [R3] using CPLEX and Benders decomposition, and compare their performances against the CGA approach. Column generation and Benders decomposition algorithms are coded in C++ Visual Studio 2013 and all optimization problems are solved using CPLEX version 12.6.1 on a 64-bit Windows 10 with Intel(R) core i7-4790 3.60GHz

processors and 8.00 GB RAM. For CPLEX and CGA, all optimization models are executed to an optimality gap of 1e-09 with no time limit.

Computational results are summarized in Table 3.8a where the column generation is compared against the CPLEX solution. Column “RMP linear Sol" is the optimal objective function value to model [RMP] and column “Best found Sol" denotes the objective function value for the best integer solution found by adding integer constraint to [RMP]. Gap is calculated as $\frac{\text{RMP linear Sol-Best found Sol}}{\text{Best found Sol}}$. Column “CG time" is the time spent to generate all columns while “RMP-MIP time" denotes the time spent to solve model [RMP] with integer constraint. The total CPU time (in seconds) spent by column generation approach and CPLEX are denoted by “TCGA time" and “CPLEX time", respectively. Finally, Time Ratio in Table 3.8a is calculated as $\frac{\text{CPLEX time}}{\text{TCGA time}}$. Overall, CGA is able to solve all instances in less than one hour with optimality gaps of less than 2%. At capacity $C = 7,000$, the gap value shows that the best solution obtained from CGA can only be improved by at most 1.94% if the original model [R3] is solved to optimality. In fact, for all instances, optimal solutions obtained by directly solving model [R3] equals the solution obtained by CGA. This signifies the effectiveness of the CGA in obtaining solutions that are close to optimal while reducing the computational effort by a factor of three.

We also compare our proposed column generation approach against the L-shaped Benders decomposition approach generally applied in stochastic programming where the master problem decides on first stage decision variables while the subproblem decides on second-stage decision variables. The overall Benders procedure is based on the general framework in Carøe and Tind (1998) and is detailed in A.1. All instances are solved with a time limit of 3600 seconds. Computational results are summarized in Table 3.8b where Columns "UB" and "LB" denote upper bound and lower bound obtained from respective solution approaches, respectively, while column "Iterations" refers to the number of iterations between the master problem and subproblems in Benders decomposition. Computational results show that the proposed column generation approach outperforms Benders decomposition. The latter fails to solve any instance to optimality within a one-hour time window and reports an average optimality gap of 97.6%.

Capacity	Column Generation Approach					Time Comparison		
C	[RMP] linear Sol	Best found Integer Sol	Gap	CG time	RMP-MIP time	TCGA time(s)	CPLEX time (s)	Time ratio
1000	38.85%	38.84%	0.02%	82.92	568.35	651.27	3651.55	5.61
2000	58.68%	58.41%	0.47%	143.35	1162.32	1304.67	3968.64	3.04
3000	70.45%	69.63%	1.18%	143.77	1728.56	1872.33	4300.48	2.30
5000	82.95%	81.39%	1.91%	195.48	1754.58	1949.62	4906.26	2.52
7000	89.06%	87.36%	1.94%	268.98	2854.92	3123.90	5880.13	1.88
		Avg	**1.10%**				**Avg**	**3.07**

(a) Column Generation Approach vs CPLEX

Capacity, C	Benders				Column Generation Approach			
	UB	LB	Gap	Iterations	UB	LB	Gap	Time(s)
1000	41.16%	26.00%	58.31%	71	38.85%	38.84%	0.02%	651.27
2000	67.78%	36.70%	84.67%	52	58.68%	58.41%	0.47%	1304.67
3000	88.17%	42.47%	107.60%	38	70.45%	69.63%	1.18%	1872.33
5000	114.67%	53.13%	115.83%	31	82.95%	81.39%	1.91%	1949.62
7000	1.33629	60.31%	121.56%	28	89.06%	87.36%	1.94%	3123.9
		Average	97.59%			**Average**	1.10%	

(b) Column Generation Approach vs L-shaped Benders decomposition

Table 3.8: Computational efficiency of the proposed column generation against CPLEX and L-shaped Benders Decomposition

3.6. Conclusions

In this chapter, we addressed the strategic capacity and assortment planning problem faced by MedAvail through extensive descriptive and prescriptive analytics. We developed three optimization models that decide on stock levels and product substitution. In addition, we developed a column-generation based heuristic solution methodology that is able to obtain near-optimal solutions within 1.1% of the optimality gap while reducing computational times by a factor of 3. Computational experiments over real and randomly generated data show that product substitution reduces kiosk's capacity requirements by up to 9%. We also show that the effect of product substitution depends on desired service level and the nature of demand data. As an outcome of this work, MedAvail expects to improve its service levels by 30% using a larger capacity kiosk. MedAvail also expects a 10% improvement in service levels of the existing kiosks by optimizing assortment and stocking decisions using the suggested optimization models.

A possible extension could be to model exact substitution. As discussed earlier in Theorem 1, the proposed model obtains a lower bound on the service level due to conservative approximation of the number of failures. As such, some of the substitution rules that could improve the solution are not selected. Also, the model allows only one substitute for each quantity. Modelling the problem with exact substitution and multiple substitutes would be computationally difficult to solve, and developing an efficient solution methodology for such a model is another promising future work.

Chapter 4

Robust Inventory Planning

4.1. Introduction & Literature Review

Analysis of pharmacy data shows that product demand for self-serve kiosks is sporadic in nature and may not follow a known distribution. On top of that, historical data is available only after a kiosk is operational. As such, during the early stage of kiosk operations, we may not have access to enough data points to make good stocking decisions. In such circumstances, a data-driven stochastic approach as used in the previous chapter may fail to obtain good stocking decisions. This motivates us to model the problem under a robust framework. We extend the modelling approach presented in Chapter 3 to a robust setting. Unlike existing newsvendor problem literature, we use a fill rate maximization objective that has not been studied before. We propose a RO framework that aims to maximize the worst-case fill rate over all demand scenarios in a polyhedral uncertainty set. The latter is constructed using a hierarchical clustering algorithm and is based on the idea that cluster demand has less variability compared to individual product demand. We propose an integrated column-and-constraint generation and conservative approximation approach to optimally solve the nonlinear RO formulation where nonlinearity arises due to the fill rate objective. The main contributions of our work are as follows

- First, our work is the first to study a robust inventory planning problem for self-serve

kiosks with rare product demand under fill rate maximization objective. We numerically show that the fill rate objective is preferred over the more commonly used profit objective function for self-serve kiosks when products have similar profit margins and capacity is limited.

- Second, we show that robust optimization modelling with a single budgeted uncertainty constraint fails to distinguish between stocking decisions. It leads to a case where every feasible solution is optimal. To address this, we propose a novel approach to define the uncertainty set using clustering. The variability in demand for each cluster is relatively low compared to individual product demand variability and the bounds are therefore tighter. Clusters of negatively correlated products are generated using an agglomerative hierarchical clustering algorithm. This removes demand scenarios that are highly unlikely and avoids the issue of overly conservative solutions. In contrast to the literature, where a level of conservatism (or budget of uncertainty) has to be set by the practitioner, the proposed uncertainty set in this work does not require any user-defined threshold.

- Third, we propose an exact solution methodology that integrates both conservative approximation and column-and-constraint generation approaches. Under fill rate maximization objective, the adversarial problem of the robust model is nonconvex. We propose a fast iterative approach that solves the nonlinear adversarial problem in a few iterations. Our computational results show that the model with fill rate objective is inherently difficult to solve. Our proposed methodology is however generalizable and could be applied to other objectives, for instance, profit or weighted fill rate. Testing with the profit objective shows that the proposed solution methodology is able to solve it exactly within few seconds.

- Fourth, we test the proposed robust modelling approach using real pharmacy data with around 1600 drugs. Computational testing shows that our approach improves fill rates by 5.8%, on average, and up to 17% compared to other benchmark approaches in the literature. Compared to the profit objective, the fill rate objective function yields a 17% higher fill rate by compromising 20% in profits, on average.

From a modelling perspective, our work relates to *Capacitated Multiproduct Newsvendor*

Problem (CMPNP) where a newsvendor determines optimal single-period stock levels for multiple products with stochastic demand under a budget constraint. One way to deal with demand uncertainty is to employ a stochastic optimization framework as discussed in Chapter 3 to maximize the expected performance under the assumption that demand distribution is exactly known. In situations where true distribution is not known, data-driven solution approaches such as sample average approximation (SAA), could be employed where empirical data is used as input to the optimization model. Huber et al. (2019) show superiority of data-driven SAA approach over standard stochastic models for service levels of up to 90%. At higher service level requirements, empirical data fed to SAA need to grow exponentially. Our computational results also confirm this phenomenon where we observe that SAA performance improves only when the size of the training dataset is increased. Under limited transactional data, both SAA and standard stochastic optimization approach may lead to poor out-of-sample performance. This motivates us to follow a robust optimization (RO) framework that attempts to maximize the worst-case expected performance over an uncertainty set which we construct using hierarchical clustering algorithm. We now review related work on CMPNP where stochasticity in demand is incorporated via robust optimization.

Robust Optimization

To address the issue of poor out-of-sample performance in stochastic optimization, the RO modelling framework hedges against the worst-case realization of the demand within an "uncertainty set". The concept of RO was first introduced by Soyster (1973) and has been an active research area since the work of Ben-Tal and Nemirovski (1999). In the last two decades, RO gained a lot of attention due to its simplicity in modelling and tractability of the robust counterpart formulations. RO is used in a broad spectrum of applications including portfolio optimization (Ghaoui et al. 2003, Tütüncü and Koenig 2004, Olivares-Nadal and DeMiguel 2018), statistics and machine learning, e.g., regression (El Ghaoui and Lebret 1997) and classification (Xu et al. 2009), logistics and supply chain management such as routing (Montemanni et al. 2007), schedulling (Hazır et al. 2010), facility location (Baron et al. 2011), inventory management (Bertsimas and Thiele 2006b, See and Sim 2010), and revenue management(Gao et al. 2009, Rusmevichientong and Topaloglu 2012). The construction of uncertainty set is however

critical which if not properly defined, may lead to overly conservative solutions. The latter is addressed either by restricting the total deviation from nominal demand (Bertsimas and Sim 2004) or a distributionally robust optimization (DRO) framework (Scarf 1958) is used where the expected performance is maximized against the worst-case distribution within a set of distributions, referred to as ambiguity set. Ambiguity set could be defined either using moment-based distributional information such as mean, variance, and covariance (Gallego and Moon 1993, Gallego et al. 2001, Alfares and Elmorra 2005, Roels 2006, Yue et al. 2006, Perakis and Roels 2008), or by using statistical distance measures (phi-divergence, Wasserstein-distance) controlling the deviation from nominal distribution, for instance, empirical distribution (Ben-Tal et al. 2013, Gao and Kleywegt 2016). DRO approaches are however computationally difficult to solve for large-scale instances as considered in this thesis with thousands of products. For instance, Gao and Kleywegt (2016) and Ben-Tal et al. (2013) were able to solve instances with up to 12 products only.

In this chapter, we address the conservatism issue by constructing a polyhedral uncertainty set using hierarchical clustering algorithm to remove overly-conservative demand scenarios. In the literature, several types of uncertainty sets are used including the scenario-based, box, polyhedral, and ellipsoidal sets. Vairaktarakis (2000) presents a scenario-based RO formulation where a set of discrete demand scenarios are used to make stocking decisions such that the expected cost under the worst-case realization of demand is minimized. Computational results in Section 4.4 show that such a scenario-based RO approach fails to present robust solutions due to the limited number of scenarios relative to a large number of products. RO with box uncertainty set was introduced by Soyster (1973) where the uncertain parameter may take any value between interval data. This leads to overly conservative solutions as all uncertain parameters take worst-case values. To control the level of conservatism, Bertsimas and Sim (2004) propose a polyhedral uncertainty set and introduce the concept of budgeted uncertainty constraint to control scaled-deviation from a nominal value using a user-defined budget of uncertainty. Bertsimas and Thiele (2006a) extends the uncertainty set proposed by (Bertsimas and Sim 2004) in a multi-period setting where budgeted uncertainty constraint is defined for each period k. As opposed to scaled-deviations, we restrict variations in total demand by using a set of linear inequalities similar to ones used by Simchi-Levi et al. (2018) based on central limit theorem (CLT) proposed by Bandi and Bertsimas (2012). The main difference however lies in the fact that we

do not explicitly define the linear inequalities in the uncertainty set but rather use hierarchical clustering algorithm to derive them purely from data. The clustering algorithm creates clusters of negatively correlated products and does not require covariance-based uncertainty sets as used commonly in the literature (Pachamanova 2002). A similar approach is employed by Qiu et al. (2019) who use support vector clustering (SVC) to construct a data-driven uncertainty set for CMPNP under the profit maximization objective. We also show that under fill rate objective, an uncertainty set with a single budgeted uncertainty set fails to distinguish between stocking decisions and every solution is optimal when product demand is low.

Our work also differs from existing literature due to the fill rate maximization objective which is a function of piece-wise linear *max* functions. The resulting inner maximization adversarial problem is non-convex which when linearized, results in a mixed-integer linear optimization problem. This issue frequently appears in robust newsvendor problems where the robust counterpart based on the duality of the adversarial problem is a conservative approximation approach and does not guarantee the optimality of the original robust problem. Ardestani-Jaafari and Delage (2016) are the first to present an exact linear formulation for the robust (and distributionally robust) multi-product newsvendor problem under the profit objective based on the total unimodularity of the adversarial problem where the demand lies in a polyhedral uncertainty set defined by interval data and a budget of uncertainty. Unfortunately, the totally unimodular property does not hold under the fill rate objective. We therefore present an exact solution approach based on column-and-constraint generation (C&CG) and conservative approximation approaches where scenarios from uncertainty sets are generated from an adversarial problem and are dynamically added to the master problems. C&CG solution methodology for RO problems was first examined by Zeng and Zhao (2013) and have been applied in various RO settings such as location-transportation problem (Ardestani-Jaafari and Delage 2020), regret minimization problems (Poursoltani and Delage 2019), and facility location problems (An et al. 2014). Our solution approach differs from Zeng and Zhao (2013) due to the non-linear adversarial problem and we propose a fast iterative approach to solve the adversarial problem. In addition, we show that under fill rate objective, the C&CG algorithm fails to close the gap and requires integration of conservative approximation which is used to warm-start the algorithm.

The rest of the chapter is organized as follows. We formally define the self-serve kiosk inventory planning problem with fill rate maximization and rare demand in Section 4.2, We

develop an exact solution methodology based on the column-and-constraint generation and approximation approaches in Section 4.3. Section 4.4 details numerical results on a test case for pharmacy kiosks and on randomly generated data. Concluding remarks and future research directions are given in Section 4.5.

4.2. Self-serve Kiosk Inventory Planning Problem

In this section, we propose a robust optimization framework for the inventory planning problem faced at self-serve kiosks. Specifically, we model this problem as a newsvendor problem with fill rate objective where the demand for each product is low and sporadic. Formally, we define low demand product as

Definition 1. *Let P be the true demand distribution of a product and $S = (s_1, s_2, ..., s_m)$ be a set of m demand values sampled iid from P. The demand for an item k is defined as low if $\frac{1}{m}|\{i : s_i = 0\}| > \frac{1}{2}$.*

Let I be the set of products and d_i be the demand for product $i \in I$. We assume the replenishment lead time for all products to be the same and fixed. Product demand $\mathbf{d} = [d_i]$ is not exactly known and lies within a polyhedral uncertainty set $\mathcal{D}$. We present a modelling framework to decide on the lead time stock level x_i for each product $i \in I$ under a resource constraint C, such that the fill rate α is maximized over all possible realizations of demand $\mathbf{d} \in \mathcal{D}$. We define the fill rate α as the percentage of successful transactions during the replenishment lead time.

The lost sales for a product i is $\max\{0, d_i - x_i\}$. Lost sales occur either because the product is not stocked, i.e., $x_i = 0$, or observed demand exceeds the stock level, i.e., $x_i < d_i$. The fill rate is then calculated as α =1-$\frac{\sum_{i\in I} \max\{0,d_i-x_i\}}{\sum_{i\in I} d_i}$, where $\sum_{i\in I} d_i$ denotes the total number of customer requests during the replenishment lead time and $\sum_{i\in I} \max\{0, d_i - x_i\}$ is the total

number of failed transactions. The robust problem [RO] is

$$\text{[RO]:} \quad \max \quad \alpha \tag{4.2.1}$$

$$\text{s.t.} \quad \sum_{i \in I} x_i \leq C \tag{4.2.2}$$

$$\alpha \leq 1 - \frac{\sum_{i \in I} \max\{0, d_i - x_i\}}{\sum_{i \in I} d_i} \qquad \forall \mathbf{d} \in \mathcal{D} \tag{4.2.3}$$

$$\alpha \in [0,1],\ x_i \in \mathbb{Z}_+ \qquad \forall i \in I, \tag{4.2.4}$$

where constraint (4.2.2) limits the total stock to available capacity, C. The objective function (4.2.1), along with constraint (4.2.3), maximizes the fill rate α for all demand realizations $\mathbf{d} \in \mathcal{D}$. Constraint (4.2.4) enforces nonnegativity and integer requirement on the stock level x_i and bounds fill rate α between 0 and 1.

Robust optimization is often criticized for its conservative solutions. Our goal is to use robust optimization framework such that the expected fill rate is maximized and as such, conservative solutions cannot be afforded. It is therefore critical to construct a well-defined uncertainty set that does not allow conservative solutions. In the literature, several types of uncertainty sets are used including scenario-based (Vairaktarakis 2000), box (Ben-Tal and Nemirovski 1999, Lin and Ng 2011b), polyhedral (Bertsimas and Thiele 2006b, Simchi-Levi et al. 2018), and ellipsoidal sets (Pachamanova 2002). We first show that under a mild assumption, robust optimization with box or polyhedral uncertainty sets with single budgeted uncertainty constraint fail to make good stocking decisions under the problem setting being studied in this chapter. For model [RO] with single budgeted constraint, the uncertainty set is defined as

$$\mathcal{D}^{\text{single}} = \left\{ \mathbf{d} : \begin{array}{ll} l_i \leq d_i \leq u_i & \forall i \in I, \\ \underline{N} \leq \sum_{i \in I} d_i \leq \overline{N}, & \\ d_i \in \mathbb{Z}_+ & \end{array} \right. \tag{4.2.5}$$

where the first inequality denotes that demand d_i lies within interval $[l_i, u_i]$ while the second

inequality defines the budget of uncertainty as an interval that lies between $[\underline{N}, \overline{N}]$. The uncertainty set (4.2.5) is similar to the partial-sum uncertainty set introduced by Mamani et al. (2017) that allows asymmetric uncertainty as opposed to symmetric uncertainty in (Bertsimas and Thiele 2006b) where demand $d_i \in [\overline{d}_i - \hat{d}_i, \overline{d}_i - \hat{d}_i]$. We show that when demand is low and capacity is limited, every feasible solution to [RO] with uncertainty set $\mathcal{D}^{\text{single}}$ is optimal.

Theorem 2. *Let $H = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n\}$ be the set of all feasible solutions i.e. $\sum_{i \in I} x_i^h \leq C, \ \forall h \in [H] = \{1, 2, ..., n\}$. If $l_i = 0, u_i \geq 1 \ \forall i \in I$, and $\underline{N} + C \leq |I|$, every solution $h \in [H]$ is optimal to [RO] with $\mathcal{D} = \mathcal{D}^{single}$.*

Proof. For $h \in [H]$, let S_1^h and S_0^h be sets of stocked and unstocked products, respectively. Since $\forall h \in [H], \sum_{i \in I} x_i^h \leq C$,

$$|S_1^h| \leq C, \quad \forall h \in [H], \tag{4.2.6}$$

$$|S_0^h| = |I| - |S_1^h| \quad \forall h \in [H], \tag{4.2.7}$$

which implies

$$|S_0^h| \geq |I| - C \quad \forall h \in [H]. \tag{4.2.8}$$

Since $\underline{N} \leq |I| - C$,

$$\underline{N} \leq |S_0^h| \quad \forall h \in [H]. \tag{4.2.9}$$

Let $\underline{S}_0^h$ be the set of any $\underline{N}$ products in S_0^h, and $\mathbf{d}^h : d_i = 1 \forall \ i \in \underline{S}_0^h, d_i = 0 \forall \ i \notin \underline{S}_0^h$. Since $\underline{S}_0^h \subseteq S_0^h$ and $S_1^h \cap S_0^h = \emptyset$, it implies that $S_1^h \cap \underline{S}_0^h = \emptyset$. This shows that $\forall h \in [H]$, there exists $\mathbf{d}^h \in \mathcal{D}^{\text{single}}$ such that $\alpha = 0$ and therefore every h is optimal. □

Remark 1. *Theorem 2 is based on three conditions. First, lower bound $l_i = 0$ for all products which is an appropriate assumption for self-serve kiosks where most of the products have low demand as defined in Definition 1. It is trivial to observe that the median and mode of a low-demand product sample is 0 which is also the lower bound on demand. In the context of robust optimization, our problem could be thought of as the case where the demand lies between a nominal value (which*

equals the lower bound) and an upper bound. Second, $u_i \geq 1$ because if $u_i < 1$, product is not ordered at all and its stock level is set to 0. Third, $\underline{N} + C \leq |I|$ holds in problem settings where the number of products offered $|I|$ is sufficiently large compared to capacity C, and $\underline{N}$. Intuitively, $\underline{N}$ is the lower bound on lead time demand. The latter assumption holds in several problem settings such as vending machines, pharmacy kiosks, and online shopping constrained by the number of products.

4.2.1 Clustering-based Uncertainty Set

We address the issue of every feasible solution being optimal under single-budgeted uncertainty set by using the idea of product clustering to define a data-driven uncertainty set. Intuitively, the budgeted uncertainty constraint in $\mathcal{D}^{\text{single}}$ (4.2.5) adds bounds to the joint distribution of demand for a cluster of all products while intervals $[l_i, u_i]$ provide bounds obtained based on the marginal distribution of clusters containing a single product. The motivation is to see how demand for products behaves jointly and separately. Our methodology is motivated by the fact that such behavior could be incorporated for different clusters of products instead of just considering clusters with single or all products. Let B be the set of all clusters of products where each cluster $b \subseteq I$. The total possible number of clusters one may consider equals $|B| = 2^{|I|}$. However, solving a robust problem with $2^{|I|}$ clusters to define an uncertainty set could be computationally intractable. We present an illustrative example in Figure 4.1 to show that for some clusters, interval uncertainty is sufficient to capture bounds on the clusters' demand distribution. Consider case 1 in Figure 4.1(a) where demand $\mathbf{d} = (d_1, d_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$. As such, the total demand $d_1 + d_2$ lies in interval $[0, 2]$ and budgeted uncertainty is defined as $0 \leq d_1 + d_2 \leq 2$ and is denoted by dotted lines. Note that budgeted uncertainty constraint is redundant and interval constraints (illustrated by solid lines) on each product's demand are sufficient to capture the joint distribution. In contrast, under case 2, $\mathbf{d} = (d_1, d_2) \in \{(0,0), (0,1), (1,0)\}$ and total demand $d_1 + d_2$ lies in interval $[0, 1]$, the budgeted uncertainty constraints for cluster containing product 1 and 2 reduce the size of uncertainty region $\mathcal{D}$ by half.

We propose the use of clustering algorithms to derive such relevant clusters. We focus on problem settings where replenishment lead time demand is sufficiently low compared to

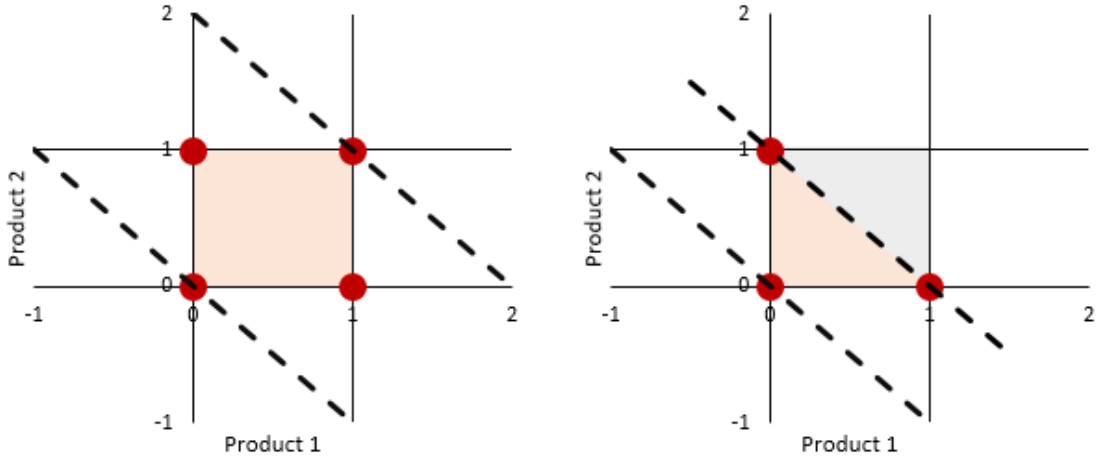


Figure 4.1: The figure illustrates the effect of bounds on joint demand under two cases.

the total number of products, i.e., $\underline{N}, \overline{N} << |I|$. When demand is low, most of the products do not appear together and the likelihood of positive correlation is low. As such, our focus is on the formation of clusters of products that infrequently occur together. Specifically, we cluster products with negatively correlated demand, i.e., when one product is ordered, it is unlikely that the other products in the cluster are ordered. We define a polyhedral uncertainty set $\mathcal{D}$ consisting of budgeted constraint for each cluster formed such that the distributional information derived from historical data is preserved. The uncertainty set $\mathcal{D}$ is defined in a generalized fashion as

$$\mathcal{D} = \left\{ \mathbf{d} : \begin{array}{ll} l_i \leq d_i \leq u_i & \forall i \in I, \\ \underline{N} \leq \sum_{i \in I} d_i \leq \overline{N}, & \\ \underline{\Gamma}_b \leq \sum_{i \in b} d_i \leq \overline{\Gamma}_b & \forall b \in B, \\ d_i \in \mathbb{Z}_+ & \end{array} \right. \tag{4.2.10}$$

where the first two sets of constraints define $\mathcal{D}_{\text{single}}$ and the third set denotes budgeted uncertainty constraint for each cluster $b \in B$.

To generate clusters B to preserve distributional information and avoid overly conservative

solutions, we consider two types of clusters: (1) clusters based on ordering frequency, and (2) clusters based on demand correlation. The first two sets of constraints in $\mathcal{D}$ do not distinguish products based on their ordering frequencies. Consider two products j and k with lead time demand between [0,1]. Let us assume that the yearly demand for j and k equals 100 and 10, respectively. Clearly, while making stocking decisions, j should be preferred over k due to the higher ordering frequency. This crucial information is not captured by the first two constraints. To incorporate the ordering frequency, one possible solution could be to include the time dimension within the uncertainty set and add bounds to the total demand for each product over the planning horizon. The problem could then be setup as to maximize the worst expected fill rate over the planning horizon. This results in a distributionally robust optimization problem that is complex to solve for real-life instances. Instead, we use a novel approach by clustering products based on their total demand $\mathbf{Y} = [Y_i]$ over the training sample. Several heuristic approaches could be used to obtain product clusters. For cluster analysis, k-means clustering is widely used and several heuristic approaches have been proposed to obtain clusters quickly. However, optimal clusters could be obtained for one-dimensional data using the dynamic programming algorithm proposed by Wang and Song (2011). Since in our case clusters need to be obtained using one-dimensional total demand vector $\mathbf{Y}$, we use the dynamic programming algorithm in Wang and Song (2011) to cluster products based on ordering frequency.

In addition, we propose a hierarchical clustering-based approach to derive additional clusters in B for products with high (negative) correlation. Specifically, we cluster products with negative correlations, i.e., when one product is ordered, it is unlikely that other products in the cluster are ordered. Let $[m] = \{1, 2, ..., m\}$ be the set of indices for a training sample of size m. we first calculate a simple matching matrix to determine the association between two products i and j as

$$dist(i, j) = \frac{|\forall t \in [m] : d_{it} = d_{jt}|}{m}, \quad \forall i, j \in I, \tag{4.2.11}$$

and use it as a distance matrix which is fed to an agglomerative hierarchical clustering algorithm (Kaufman and Rousseeuw 2009) to generate a set of clusters. Agglomerative hierarchical clustering starts with each product as a cluster. In subsequent steps, clusters are merged until one big cluster containing all products is created. Set B in uncertainty set $\mathcal{D}$ refers to clusters

both from ordering frequency and hierarchical clustering. Once clusters are obtained, we restrict lead time demand for each cluster $b \in B$ between $[\underline{\Gamma}_b, \overline{\Gamma}_b]$ derived from data. Parameters $\underline{\Gamma}_b$ and $\overline{\Gamma}_b$ are calculated using training sample as

$$\underline{\Gamma}_b = \min_{t\in[m]} \left\{ \sum_{i\in b} d_{it} \right\}, \quad \overline{\Gamma}_b = \max_{t\in[m]} \left\{ \sum_{i\in b} d_{it} \right\} \tag{4.2.12}$$

4.3. Solution Approach

Solving model [RO] is computationally challenging. Constraint (4.2.3) is defined for all demand realizations in $\mathcal{D}$ and as such, the model may contain infinitely many constraints. Secondly, constraint (4.2.3) is defined by sums of piece-wise linear functions due to the max term. Although these challenges are quite common in inventory problems and have been extensively studied in the literature, our problem has an added complexity of fractional objective function due to the fill rate maximization objective.

To solve the nonlinear model, we propose an exact column-and-constraint generation solution approach and extend it to develop a conservative approximation model. The two approaches are then integrated where conservative approximation solution is used to generate demand scenarios to warm start the column-and-constraint generation approach and to provide tight bounds on the best found solution.

4.3.1 An Exact Column and Constraint Generation Approach

Since there could be infinitely many demand scenarios $\mathbf{d} \in \mathcal{D}$, we use a column-and-constraint generation (CCG) approach to solve the problem by generating demand scenarios iteratively. Let $\mathbf{d}^k$, $k \in K$ be a subset of demand scenarios from the uncertainty set $\mathcal{D}$. We define the

master problem $[\text{MP}]_{\text{RO}}$ as a relaxation of [RO]

$$[\text{MP}_{\text{RO}}]: \quad \max \quad \alpha \tag{4.3.1}$$

$$\text{s.t.} \quad (4.2.2), (4.2.4),$$

$$\alpha \leq 1 - \frac{\sum_{i \in I} \max\{0, d_i^k - x_i\}}{\sum_{i \in I} d_i^k} \quad \forall \quad k \in K \tag{4.3.2}$$

$[\text{MP}]_{\text{RO}}$ decides on stock levels $\mathbf{x} = [x_i]$ to maximize the fill rate α given demand scenarios $\mathbf{d}^k$, $k \in K$. Model $[\text{MP}]_{\text{RO}}$ is nonlinear due to the max function in constraint (4.3.2). However, it may be linearized by introducing a new decision variable f_i^k as the number of failed transactions for product $i \in I$. The linearized formulation $[\text{LMP}]_{\text{RO}}$ is

$$[\text{LMP}_{\text{RO}}]: \quad \max \quad \alpha \tag{4.3.3}$$

$$\text{s.t.} \quad (4.2.2), (4.2.4),$$

$$f_i^k \geq d_i^k - x_i \qquad \forall \quad i \in I, k \in K, \tag{4.3.4}$$

$$\alpha \leq 1 - \frac{\sum_{i \in I} f_i^k}{\sum_{i \in I} d_i^k} \quad \forall \quad k \in K, \tag{4.3.5}$$

$$f_i^k \geq 0 \qquad \forall \quad i \in I, k \in K, \tag{4.3.6}$$

An optimal solution to model $[\text{LMP}_{\text{RO}}]$ is also optimal for $[\text{MP}]_{\text{RO}}$. This is proven in Lemma 3. Each demand scenario results in a nonlinear constraint (4.3.2) and to linearize it, constraints (4.3.4) and (4.3.5) and variables (or columns) f_i^k need to be introduced. Hence, our solution approach is a column and constraint generation approach.

Lemma 3. *An optimal solution to linearized model $[\text{LMP}_{\text{RO}}]$ is also optimal for the original problem $[\text{MP}]_{\text{RO}}$.*

Proof. Proof. Let $(\overline{\mathbf{x}}, \overline{\mathbf{f}})$ be an optimal solution to $[\text{LMP}_{\text{RO}}]$ with objective value $\overline{\alpha}$. For solution $(\overline{\mathbf{x}}, \overline{\mathbf{f}})$ to be optimal for model $[\text{MP}]_{\text{RO}}$, $\overline{f}_i^k$ must be equal to the max term $\max\{0, d_i^k - \overline{x}_i\}$. We

prove that $\overline{f}_i^k$ always equals the max term.

Constraints (4.3.5) and (4.3.6) ensure that $\overline{f}_i^k \geq \max\{0, d_i^k - x_i^*\}$ which proves that at optimality, $\overline{f}_i^k$ cannot take a value less than the max term. We now show by contradiction that $\overline{f}_i^k$ cannot take a value greater than the max term. Assume that $\exists\ k \in K, i \in I$, $\overline{f}_i^k > \max\{0, d_i^k - x_i^*\}$. Let $(\overline{\mathbf{x}}, \mathbf{f^a}, \alpha^a)$ be the adjusted solution where $f_i^a = \max\{0, d_i^k - \overline{x}_i\}$. Then,

$$\sum_{i\in I} \overline{f}_i^k \geq \sum_{i\in I} f_i^a \iff \left(1 - \frac{\sum_{i\in I} \overline{f}_i^k}{\sum_{i\in I} d_i^k}\right) < \left(1 - \frac{\sum_{i\in I} f_i^a}{\sum_{i\in I} d_i^k}\right) \tag{4.3.7}$$

Since the objective is to maximize the fill rate, constraints (4.3.2) and (4.3.5) are always binding at optimality and as such, $\overline{\alpha} = 1 - \frac{\sum_{i\in I} \overline{f}_i^k}{\sum_{i\in I} d_i^k}$ and $\alpha^a = 1 - \frac{\sum_{i\in I} f_i^a}{\sum_{i\in I} d_i^k}$. Then by (4.3.7),

$$\left(1 - \frac{\sum_{i\in I} \overline{f}_i^k}{\sum_{i\in I} d_i^k}\right) < \left(1 - \frac{\sum_{i\in I} f_i^a}{\sum_{i\in I} d_i^k}\right) \iff \overline{\alpha} < \alpha^a \tag{4.3.8}$$

which contradicts the assumption that $\overline{\alpha}$ is optimal. □

4.3.1.1 Adversarial Problem

The master problem solution $(\overline{\alpha}, \overline{\mathbf{x}})$ provides an upper bound to the original problem [RO] and is optimal if and only if

$$\overline{\alpha} \leq 1 - \frac{\sum_{i\in I} \max\{0, d_i - \overline{x}_i\}}{\sum_{i\in I} d_i} \quad \forall \mathbf{d} \in \mathcal{D}.$$

Since the master problem $[\text{LMP}]_{\text{RO}}$ considers only a subset of demand scenarios in $\mathcal{D}$, we solve an adversarial problem to find a vector $\mathbf{d} \in D$ that minimizes the fill rate α for a given solution

$\overline{\mathbf{x}}$. The adversarial problem $[\text{AP}_{\text{RO}}]$ is

$$[\text{AP}_{\text{RO}}]: \quad \min \quad \alpha_{\min} \tag{4.3.9}$$

$$\text{s.t.} \quad \mathbf{Rd} \leq \mathbf{r} \tag{4.3.10}$$

$$\alpha_{\min} \geq 1 - \frac{\sum_{i \in I} \max\{0, d_i - \overline{x}_i\}}{\sum_{i \in I} d_i}, \tag{4.3.11}$$

$$\alpha_{\min} \geq 0, \; d_i \in \mathbb{Z}_+ \qquad i \in I, \tag{4.3.12}$$

Constraint (4.3.10) defines the uncertainty set $\mathcal{D}$ where the coefficient matrix $\mathbf{R}$ and right hand side vector $\mathbf{r}$ describe inequality constraints in set (4.2.10). Constraint (4.3.11) defines the minimum fill rate $\alpha_{\min}$ while constraint (4.3.12) enforces the nonnegativity and integer requirement on $\alpha_{\min}$ and d_i, respectively. Note that $[\text{AP}_{\text{RO}}]$ aims to find a vector $\mathbf{d} \in \mathcal{D}$ that minimizes the fill rate $\alpha_{\min}$ achieved for the current master problem solution $\overline{\mathbf{x}}$. If $\alpha_{\min} \geq \overline{\alpha}$, the current master solution $\overline{x}$ is optimal to the original problem (4.2.1). On the other hand, if $\alpha_{\min} < \overline{\alpha}$, this confirms that there exists a demand scenario $\mathbf{d}$ for which the current master problem solution $(\overline{\mathbf{x}}, \overline{\alpha})$ is infeasible to the original problem. If such a scenario exists, it is added to the set of scenarios K. This procedure continues until $\alpha_{\min} = \alpha$, and an optimal solution to the original problem [RO] is obtained. We now detail the procedure to solve the nonlinear adversarial problem $[\text{AP}_{\text{RO}}]$.

Rearranging Constraint (4.3.11)

$$\sum_{i \in I} \left(\max\left\{0, d_i - \overline{x}_i\right\} - (1 - \alpha_{\min}) d_i\right) \geq 0 \tag{4.3.13}$$

Constraint (4.3.13) is nonlinear due to the max function and quadratic term $\alpha_{\min} d_i$. We propose an iterative procedure to deal with the quadratic term by fixing $\alpha_{\min} = \overline{\alpha}$ and maximizing the slack of Constraint (4.3.13). We refer to this problem as master adversarial problem $[\text{MAP}_{\text{RO}}]$

$$[\text{MAP}_{\text{RO}}]: \quad \max \quad \sum_{i \in I} \left(\max\left\{0, d_i - \overline{x}_i\right\} - (1 - \overline{\alpha}_{\min}) d_i\right) \tag{4.3.14}$$

$$\text{s.t.} \quad \mathbf{Hd} \leq \mathbf{h}, \tag{4.3.15}$$

$$d_{ij} \in \mathbb{Z}_+ \qquad i \in I, \tag{4.3.16}$$

The master adversarial problem solution $\overline{\mathbf{d}}$ is then used to minimize α_{min} in the sub-adversarial problem [SAP$_{\text{RO}}$] as

$$[\text{SAP}_{\text{RO}}]: \quad \min \quad \alpha_{\text{min}} \tag{4.3.17}$$

$$\text{s.t.} \quad \alpha_{\text{min}} \geq 1 - \frac{\sum_{i \in I} \max\{0, \overline{d}_i - \overline{x}_i\}}{\sum_{i \in I} \overline{d}_i}. \tag{4.3.18}$$

which is trivial to solve and the optimal solution is calculated as $\alpha^*_{\text{min}} = 1 - \frac{\sum_{i \in I} \max\{0, \overline{d}_i - \overline{x}_i\}}{\sum_{i \in I} \overline{d}_i}$.

We update $\overline{\alpha}_{\text{min}} = \alpha^*_{\text{min}}$ in model MAP$_{\text{RO}}$ and resolve. This procedure continues until the optimal objective function value of MAP$_{\text{RO}}$ $z^*_{\text{MAP}} = 0$ which implies that the surplus of constraint (4.3.13) cannot increase further. Proof of optimality to the original model [AP] is given in Lemma 4.

Lemma 4. *The solution obtained from the iterative procedure is an optimal solution to the original problem [AP].*

Proof. Proof. Let $(\overline{\mathbf{d}}, \overline{\alpha}_{\text{min}})$ be the solution obtained in the last iteration where $z^*_{\text{MAP}} = 0$. We first prove that $(\overline{\mathbf{d}}, \overline{\alpha}_{\text{min}})$ is a feasible solution to the original model [AP]

$$z^*_{\text{MAP}} = \sum_{i \in I} \left(\max\left\{0, \overline{d}_i - \overline{x}_i\right\} - (1 - \overline{\alpha}_{\text{min}})\overline{d}_i \right) = 0 \tag{4.3.19}$$

Rearranging equation (4.3.19)

$$\overline{\alpha}_{\text{min}} = 1 - \frac{\sum_{i \in I} \max\{0, \overline{d}_i - \overline{x}_i\}}{\sum_{i \in I} \overline{d}_i} \tag{4.3.20}$$

which does not violate constraint (4.3.11) while Constraint (4.3.10) is already contained in master adversarial problem [MAP$_{\text{RO}}$]. This proves that solution $(\overline{\mathbf{d}}, \overline{\alpha}_{\text{min}})$ is feasible for model [AP].

We prove by contradiction that solution $(\overline{\mathbf{d}}, \overline{\alpha}_{\text{min}})$ is optimal to the original model [AP]. Let $(\mathbf{d}^*, \alpha^*_{\text{min}})$ be the optimal solution to the original model [AP] where $\alpha^*_{\text{min}} < \overline{\alpha}_{\text{min}}$. This implies

$$\overline{\alpha}_{\text{min}} > 1 - \frac{\sum_{i \in I} \max\{0, d^*_i - \overline{x}_i\}}{\sum_{i \in I} d^*_i} = \alpha^*_{\text{min}}. \tag{4.3.21}$$

Rearranging inequality (4.3.21)

$$\sum_{i \in I} \left(\max\left\{0, d^*_i - \overline{x}_i\right\} - (1 - \overline{\alpha}_{\text{min}}) d_i\right) > 0. \tag{4.3.22}$$

Using $\overline{\alpha}_{\text{min}}$, model [MAP] maximized the left hand side of equation (4.3.22) with optimal objective value $z^*_{\text{MAP}} = 0$. This confirms that there does not exist any solution $\mathbf{d}$ for which inequality (4.3.22) holds and contradicts the assumption that $(\mathbf{d}^*, \alpha^*_{\text{min}})$ is an optimal solution to the original model [AP]. □

Solution $\overline{\alpha}_{\text{min}}$ obtained in the last iteration is a lower bound to the original problem [RO]. The proposed iterative procedure adds multiple demand scenarios in a single iteration between the master and adversarial problem. If $\alpha^*_{\text{min}} < \overline{\alpha}$ for some $\overline{\mathbf{d}}$, it is added to the set of demand scenarios K.

Solving the nonconvex master adversarial problem (MAP) The master adversarial problem [MAP_{RO}] is a nonconvex optimization problem due to the max max objective function. We linearize it by introducing auxilary binary variables z_i and continuous variables Δ_i. The complete linearized version of [MAP_{RO}] is

$$\text{[MAP-L]:} \quad \max \quad \sum_{i \in I} (\Delta_i - \overline{x}_i z_i - (1 - \overline{\alpha}_{\text{min}}) d_i) \tag{4.3.23}$$

$$\text{s.t. } \mathbf{Hd} \leq \mathbf{h}, \tag{4.3.24}$$

$$\Delta_i \leq u_i z_i \qquad \forall \quad i \in I, \tag{4.3.25}$$

$$\Delta_i \leq d_i \qquad \forall \quad i \in I, \tag{4.3.26}$$

$$\Delta_i \geq 0,\ z_i \in \{0, 1\}, d_i \in \mathbb{Z}_+ \qquad \forall \quad i \in I, \tag{4.3.27}$$

where $\max\{0, d_i - \overline{x}_i\} = 0$ if $z_i = 0$, else $\max\{0, d_i - x_i\} = d_i - \overline{x}_i \geq 0$.

The column and constraint generation algorithm iterates between the master problem [MP_{RO}] and the adversarial problem [AP_{RO}]. Model [MP_{RO}] provides an upper bound to the original problem [RO] given a set of demand scenarios K. Using the master solution $(\overline{\alpha}, \overline{\mathbf{x}})$, model [$\text{AP}_{\text{RO}}$] attempts to find a demand scenario $\mathbf{d} \in \mathcal{D}$ that minimizes the fill rate α_{min} achieved for the current master solution $\overline{\mathbf{x}}$ and provides a lower bound to [RO].

4.3.2 A Conservative Approximation Approach

Computational results show that the column and constraint generation (CCG) approach proposed earlier fails to converge in reasonable time due to weak lower bounds. To improve the solution obtained from CCG, we suggest an approximation model that provides a tighter lower bound and set of demand scenarios to warm-start CCG algorithm. Recall Constraint (4.2.3) in model [DRO]

$$\alpha \leq 1 - \frac{\sum_{i \in I} \max\{0, d_i - x_i\}}{\sum_{i \in I} d_i} \qquad \forall \quad \mathbf{d} \in \mathcal{D}$$

which is equivalent to

$$\max_{\mathbf{d} \in \mathcal{D}} \left\{ \sum_{i \in I} \left(\max\{0, d_i - x_i\} - (1 - \alpha) d_i \right) \right\} \leq 0 \tag{4.3.28}$$

The inner maximization problem in Constraint (4.3.28) is equivalent to the master adversarial problem [MAP_{RO}]. As such, one may replace the right hand side in Constraint (4.3.28) with dual of [MAP_{RO}] to develop the robust counterpart for [RO]. Unfortunately, [MAP_{RO}] is a nonconvex optimization problem and its linear version [MAP-L_{RO}] consists of binary variables. As such, an exact robust counterpart does not exist for [RO]. However, we relax the integer requirement in [MAP-L_{RO}] to develop a conservative approximation robust counterpart. Consider the relaxed

primal problem

$$\text{[MAP-LR]:} \quad \max \quad \sum_{i\in I}(\Delta_i - \overline{x}_i z_i - (1-\overline{\alpha}_{\min})d_i) \tag{4.3.29}$$

$$\text{s.t.} \quad \Delta_i - d_i \le 0 \qquad i\in I, \quad [\lambda_i] \tag{4.3.30}$$

$$\Delta_i - u_i z_i \le 0 \qquad i\in I, \quad [\eta_i] \tag{4.3.31}$$

$$z_i \le 1 \qquad i\in I, \quad [\pi_i] \tag{4.3.32}$$

$$d_i \le u_i \qquad i\in I, \quad [\gamma_i^u] \tag{4.3.33}$$

$$-d_i \le -l_i \qquad i\in I, \quad [\gamma_i^l] \tag{4.3.34}$$

$$\sum_{i\in I} d_i \le \overline{N} \qquad [\nu^u] \tag{4.3.35}$$

$$-\sum_{i\in I} d_i \le -\underline{N} \qquad [\nu^l] \tag{4.3.36}$$

$$\sum_{i\in b} d_i \le \overline{\Gamma}_b \qquad b\in B, \quad [\omega_b^u] \tag{4.3.37}$$

$$-\sum_{i\in b} d_i \le -\underline{\Gamma}_b \qquad b\in B, \quad [\omega_b^l] \tag{4.3.38}$$

$$\Delta_i \ge 0,\ z_i \ge 0,\ d_i \ge 0 \qquad i\in I, \tag{4.3.39}$$

where $[\cdot]$ are the dual variables for each constraint. Model [MAP-LR] is a linear relaxation of [MAP-L] as the integrality requirement is dropped. Taking the dual of [MAP-LR]

$$\min \quad \sum_{i\in I}\left(\pi_i + u_i\gamma_i^u - \gamma_i^l\right) + \overline{N}\nu^u - \underline{N}\nu^l + \sum_{b\in B}\left(\overline{\Gamma}_b\omega_b^u - \underline{\Gamma}_b\omega_b^u\right) \tag{4.3.40}$$

$$\text{s.t.} \quad \lambda_i + \eta_i \ge 1 \qquad \forall \quad i\in I, [\Delta_i] \tag{4.3.41}$$

$$\pi_i - u_i\eta_i \ge -x_i \qquad \forall \quad i\in I, \ [z_i] \tag{4.3.42}$$

$$-\lambda_i + \gamma_i^u - \gamma_i^l + \nu^u - \nu^l + \sum_{b\in B: i\in b}\left(\omega_b^u - \omega_b^l\right) \ge -(1-\alpha) \qquad \forall \quad i\in I, \ [d_i] \tag{4.3.43}$$

$$\lambda_i, \eta_i, \pi_i, \gamma_i^u, \gamma_i^l, \nu^u, \nu^l, \omega_b^u, \omega_b^l \ge 0 \qquad \forall \quad i\in I, \tag{4.3.44}$$

Let Z^*_{model} be the optimal objective function value to problem "model". Since the integrality constraint is dropped, $Z^*_{\text{dual}} \ge Z^*_{\text{MAP-L}}$. Replacing the right hand side of Constraint (4.3.28) with

the dual of [MAP-LR]

$$\min_{\text{s.t.}(4.3.41)-(4.3.44)} \sum_{i\in I}\left(\pi_i + u_i\gamma_i^u - \gamma_i^l\right) + \sum_{i\in I} +\overline{N}\nu^u - \underline{N}\nu^l + \sum_{b\in B}\left(\overline{\Gamma}_b\omega_b^u - \underline{\Gamma}_b\omega_b^u\right) \leq 0 \quad (4.3.45)$$

The resulting robust counterpart [RC] is

$$\begin{aligned}
\text{[RC]:}\quad \max\quad & \alpha && && (4.3.46)\\
\text{s.t.}\quad & \sum_{i\in I}\left(\pi_i + u_i\gamma_i^u - \gamma_i^l\right) + \overline{N}\nu^u - \underline{N}\nu^l + \sum_{b\in B}\left(\overline{\Gamma}_b\omega_b^u - \underline{\Gamma}_b\omega_b^u\right) \leq 0 && && (4.3.47)\\
& \lambda_i + \eta_i \geq 1 && \forall \quad i\in I, && (4.3.48)\\
& \pi_i - u_i\eta_i \geq -x_i && \forall \quad i\in I, && (4.3.49)\\
& \gamma_i^u - \gamma_i^l + \delta_i^u - \delta_i^l + \nu^u - \nu^l - \lambda_i + \sum_{\substack{b\in B:\\ i\in b}}\left(\omega_b^u - \omega_b^l\right) \geq \alpha - 1 \;\forall \quad i\in I, && && (4.3.50)\\
& \sum_{i\in I} x_i \leq C, && && (4.3.51)\\
& \lambda_i, \eta_i, \pi_i, \gamma_i^u, \gamma_i^l, \nu^u, \nu^l, \omega_b^u, \omega_b^l \geq 0,\ x_i \in \mathbb{Z}_+ && \forall \quad i\in I, b\in B. && (4.3.52)
\end{aligned}$$

Model [RC] is a conservative approximation of the original problem [RO] since $Z^*_{\text{dual}} \geq Z^*_{\text{MAP-L}}$ which results in an underestimation of the fill rate α.

4.3.3 An Integrated Approach

We propose an integrated approach that makes use of both column-and-constraint generation and conservative approximation methodologies. The overall integrated column-and-constraint generation and conservative approximation approach is detailed in Algorithm 1. The algorithm starts by solving the approximate model [RC] to obtain stock levels $\overline{\mathbf{x}}_{CA}$. Since the solution to [RC] is a conservative approximation of the original model [RO] and underestimates the fill rate, we use $\overline{\mathbf{x}}_{CA}$ to solve the adversarial problem $[\text{AP}_{\text{RO}}]$. The solution to $[\text{AP}_{\text{RO}}]$ is the fill rate α^*_{CA} achieved for the stock levels $\overline{\mathbf{x}}_{CA}$ under the worst-case realization of the demand. This serves as the initial lower bound $LB = \alpha^*_{\text{CA}}$ to original model [RO] with upper bound $UB = 1$. The adversarial problem $[\text{AP}_{\text{RO}}]$ generates multiple demand scenarios K_{CA} which we use to

initialize the column-and-constraint generation.

The master problem $[\text{MP}]_{\text{RO}}$ is a mixed integer program and it is computationally expensive to solve it to optimality at each iteration. Therefore, we drop integrality constraint from $[\text{MP}]_{\text{RO}}$ (refer to line 6 in Algorithm 1) and column-and-constraint generation is initialized. At each iteration of the algorithm, $[\text{MP}]_{\text{RO}}$ solution serves as input to the adversarial problem $[\text{AP}_{\text{RO}}]$. Master problem $[\text{MP}]_{\text{RO}}$ provides an upper bound to the original problem [RO] while $[\text{AP}_{\text{RO}}]$ obtains the worst-case fill rate for the current master problem solution $\overline{\alpha}_{\text{min}}$. Demand Scenarios $\mathbf{d}$ generated from $[\text{AP}_{\text{RO}}]$ are added to $[\text{MP}]_{\text{RO}}$ and the procedure repeats. Note that when $\{\mathbf{x} \in \mathbb{Z}\} \notin [\text{MP}_{\text{RO}}]$, the adversarial problem solution $\overline{\alpha}_{\text{min}}$ is a lower bound to the relaxed original problem and LB is updated only if $\{\mathbf{x} \in \mathbb{Z}\} \in [\text{MP}_{\text{RO}}]$ as conditioned by line 25 in Algorithm 1. When $\overline{\alpha}_{\text{min}} = UB$ for the first time, the original problem without the integrality constraint has been solved to optimality. Therefore, we add integrality constraint to $[\text{MP}]_{\text{RO}}$ (see line 27, Algorithm 1) and the algorithm continues. We terminate the algorithm when $UB = LB$ and the latest master problem solution $\overline{\mathbf{x}}$ is optimal to original problem [RO].

4.4. Computational Results

In this section, we present computational results using actual pharmacy sales data and generalized randomly generated instances. In Section 4.4.1, we use pharmacy sales data to compare our proposed RO approach against other benchmark approaches including stochastic and maxmin approaches. Section 4.4.2 compares the fill rate objective against the conventional profit objective while Section 4.4.3 extends testing to show effectiveness of the proposed approach over randomly generated instances. Finally, in Section 4.4.4, we detail discussion on computational efficiency of the integrated column-and-constraint generation and conservative approximation solution methodology. Statistical software *R* is used to split data into training and testing datasets, and to generate clusters using hierarchical clustering algorithm. The optimization models are coded in C++ and solved using CPLEX version 12.6.3 on a 64-bit Windows 10 with Intel(R) Core i7-4790 3.60 GHz processors and 8.00GB RAM. The conservative approximation model [RC] is solved to an optimality gap of 1e-09 with a time limit of 3600 seconds. For column-and-constraint generation, a time limit of 7200 seconds is used where at each iteration,

Algorithm 1 Pseudo code for Integrated Column-and-Constraint Generation and Conservative Approximation Approach

Require: Uncertainty Set $\mathcal{D}$ parameters, Capacity C

Initialization

Solve Conservative robust counterpart [RC] to obtain $(\alpha^*_{\text{CA}}, \mathbf{x}_{\text{CA}})$
Use $\mathbf{x}_{\text{CA}}$ solve Adversarial Problem $[\text{AP}_{\text{RO}}]$ and obtain demand scenarios K_{CA}
$K \leftarrow K_{\text{CA}}$ ▷ initial set of demand scenarios
$UB \leftarrow 1$ ▷ upper bound
$LB \leftarrow \alpha^*_{\text{CA}}$ ▷ lower bound
$[\text{MP}_{\text{RO}}] \leftarrow [\text{MP}_{\text{RO}}] - \{\mathbf{x} \in \mathbb{Z}\}$ ▷ drop integrality constraint from master problem

Main Loop

while $UB \neq LB$ **do**
    Solve $[\text{MP}_{\text{RO}}]$ given K ▷ obtain solution $(\overline{\alpha}, \overline{\mathbf{x}})$
    $UB \leftarrow \overline{\alpha}$ ▷ update upper bound

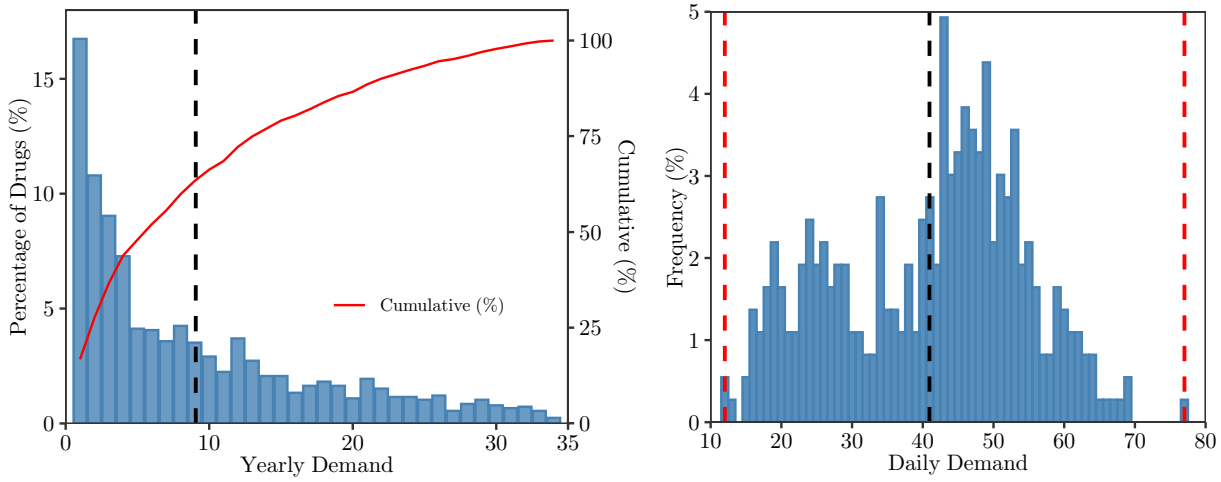
    Adversarial Problem $[\text{AP}_{\text{RO}}]$ Starts...
    $\overline{\alpha}_{\text{min}} \leftarrow \overline{\alpha}$ ▷ $\overline{\alpha}_{\text{min}}$ could be set to any value
    $z_{\text{MAP}} \leftarrow \infty$ ▷ Obj. Value of master adversarial problem $[\text{MAP}_{\text{RO}}]$
    while $z_{\text{MAP}} \neq 0$ **do**
        Solve $[\text{MAP}_{\text{RO}}]$ given $\overline{\alpha}_{\text{min}}, \overline{\mathbf{x}}$ ▷ obtain solution $\overline{\mathbf{d}}$
        Update z_{MAP}
        Solve $[\text{SAP}_{\text{RO}}]$ given $\overline{\mathbf{d}}, \overline{\mathbf{x}}$ ▷ obtain α^*_{min}
        if $\alpha^*_{\text{min}} < \overline{\alpha}$ **then** ▷ current $[\text{MP}_{\text{RO}}]$ solution is infeasible for $\overline{\mathbf{d}}$
            $K \leftarrow K \cup \overline{\mathbf{d}}$
        end if
        Update $\overline{\alpha}_{\text{min}} \leftarrow \alpha^*_{\text{min}}$
    end while
    Adversarial Problem $[\text{AP}_{\text{RO}}]$ Ends...

    if $LB < \overline{\alpha}_{\text{min}}$ & $\{\mathbf{x} \in \mathbb{Z}\} \in [\text{MP}_{\text{RO}}]$ **then**
        $LB \leftarrow \overline{\alpha}_{\text{min}}$ ▷ Update lower bound
    end if
    if $\overline{\alpha}_{\text{min}} = UB$ **then**
        $[\text{MP}_{\text{RO}}] \leftarrow [\text{MP}_{\text{RO}}] \cup \{\mathbf{x} \in \mathbb{Z}\}$ ▷ Add integrality constraint to $[\text{MP}_{\text{RO}}]$
    end if
end while
    return stock levels $\overline{\mathbf{x}}$

master problem [MP] is solved to an optimality gap of 0.01 while the adversarial problem [AP] is solved to optimality.

4.4.1 The Case of Pharmacy Kiosks

In this section, we consider pharmacy sales datasets with around 2,000 distinct drugs (GPI) are ordered annually with the majority of drugs having low and erratic demand. In contrast to Chapter 3 where GPI-QTY is used as an SKU, in this chapter, we consider GPI as an SKU. We run computational experiments over 70% of the drugs with low demand to exclude stable demand drugs. Figure 4.2(a) illustrates yearly demand distribution for one of the pharmacy stores data with 1650 low demand products. On average, drugs have yearly demand of 9 units with a maximum daily demand of 3. Around 17% of drugs appear once in the year and 33% of the drugs have yearly demand of 10 or more. Figure 4.2(b) illustrates the daily demand distribution where daily customer arrivals vary between 12 and 77 with average arrivals equal 41.



(a) Drugs Yearly Demand Distribution

(b) Distribution of Daily Customer Arrivals

Figure 4.2: Demand Distribution

We assume that a pharmacy kiosk is replenished daily which results in a total of 365 time periods. We split one-year data into training and testing datasets by randomly selecting a fraction of the days as a training dataset. From here onward, we refer to the percentage of training

data as SplitRatio. We use the training dataset to decide on stock levels $\mathbf{x}$. The testing dataset is used to determine out-of-sample (test) average fill rate α_{test}. For each dataset, we generate training datasets at three levels of SplitRatio $\in \{0.1, 0.2, 0.3\}$. Each instance is then solved at five capacity levels, $Cap = \{50\%, 60\%, 70\%, 80\%, 90\%\}$, defined as a percentage of the total number of drugs.

Note that $Cap < 100\%$ is set to ensure that there is limited capacity and not all drugs could be stocked. Recall Theorem 2 where we show that conventional robust optimization with single budgeted uncertainty constraint results in every feasible solution being optimal with $\alpha = 0$, if $\underline{N} + C \leq |I|$. Since capacity $C \leq 0.9 \times |I|$, inequality $\underline{N} + C \leq |I|$ holds as long as $\underline{N} \leq |I| - 0.9|I| \Rightarrow \underline{N} \leq 0.1 \times |I|$. For pharmacy sales data discussed earlier, $|I| = 1650$ and Theorem 2 holds when $\underline{N} \leq 165$. On the other hand, daily customer arrivals vary between 12 and 77 and therefore $\underline{N}$ cannot take a value greater than 77. This shows that conventional robust approach fails for problem settings considered in this thesis.

We compare our proposed RO approach with stochastic and maxmin models along with the expected value of perfection information. The formulation for stochastic model [SO] is given in Appendix A.2.1, it maximizes the expected fill rate α given demand scenarios in the training dataset. Maxmin model [Maxmin] is detailed in Appendix A.2.2 and it maximizes the minimum fill rate over all demand scenarios in the training dataset. For each instance, we estimate the expected value of perfect information (EVPI) defined as the fill rate achieved if the demand distribution is exactly known. To estimate EVPI, we solve the stochastic model using complete data (training + testing) to decide on stock levels which are then tested on the testing dataset.

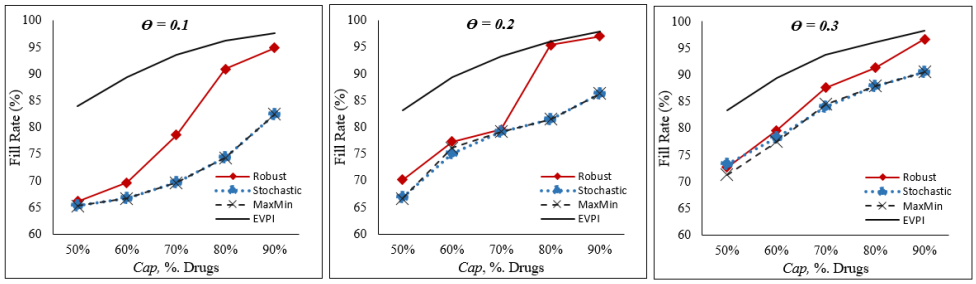
Table 4.1 compares out-of-sample fill rate α_{test} obtained from the proposed robust approach with [SO], [Maxmin], and EVPI. On average, RO solutions outperform stochastic and maxmin solutions by 5.75% and 5.84%, respectively. As shown in Figure 4.3, RO outperforms the stochastic model by up to 16.63% at higher capacity levels and when there is limited data available. For SplitRatio $= 0.1$, only 36 days of data are available, RO approach improves average out-of-sample fill rate α_{test} by 8.36% , on average. When SplitRatio $= 0.3$, average improvement is 2.78%. As training data size SplitRatio increases, the empirical distribution is a good approximation of the true demand distribution, and therefore improvements over the stochastic model decrease. Surprisingly, the maxmin model, which is quite often used in the literature to obtain robust solutions performs poorly. This is because the model has access to only a few demand

Split Ratio	Cap, % Drugs	Testing Fill Rate, α_{test}				Difference from RO		
		Robust	Stochastic	MaxMin	EVPI	Stochastic	Maxmin	EVPI
0.1	50%	66.19	65.30	65.30	83.99	0.89	0.89	-17.81
	60%	69.64	66.68	66.68	89.42	2.95	2.95	-19.79
	70%	78.56	69.64	69.64	93.55	8.92	8.92	-15.00
	80%	90.89	74.26	74.26	96.16	16.63	16.63	-5.28
	90%	94.84	82.42	82.42	97.66	12.42	12.42	-2.82
					Average	8.36	8.36	-12.14
0.2	50%	70.11	66.79	66.63	83.16	3.33	3.49	-13.05
	60%	77.27	75.11	76.11	89.30	2.17	1.16	-12.03
	70%	79.51	79.12	79.12	93.28	0.40	0.40	-13.76
	80%	95.37	81.49	81.49	96.00	13.87	13.87	-0.64
	90%	97.01	86.31	86.31	97.87	10.70	10.70	-0.86
					Average	6.09	5.92	-8.07
0.3	50%	72.67	73.24	71.27	83.44	-0.58	1.39	-10.77
	60%	79.61	78.23	77.53	89.43	1.38	2.08	-9.82
	70%	87.59	84.03	84.50	93.75	3.57	3.09	-6.15
	80%	91.34	87.88	87.88	96.13	3.47	3.47	-4.79
	90%	96.66	90.58	90.58	98.29	6.08	6.08	-1.62
					Average	2.78	3.22	-6.63
					Average	5.75	5.84	-8.95

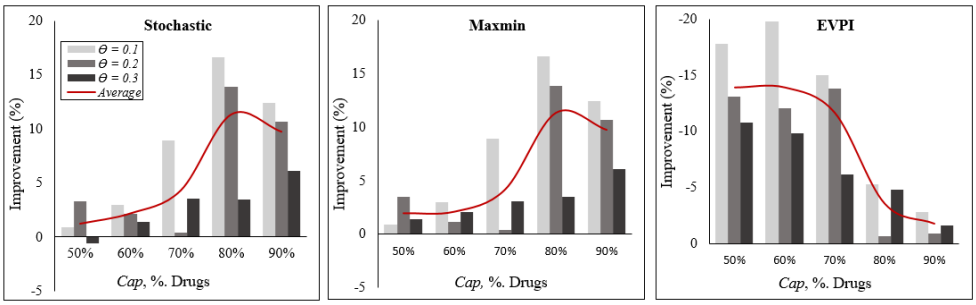
Table 4.1: Comparing out-of-sample fill rate α_{test} achieved from proposed robust approach against other approaches

scenarios in the training dataset, which does not provide sufficient information relating to the worst case demand scenario for the large number of drugs. As shown in Table 4.1 and illustrated in Figure 4.3(b), our solutions are on average, only 8.95% inferior to EVPI which implies that if any other modelling framework is used, it can not outperform our model by more than 8.95%, on average. Note that no modelling approach can ever achieve EVPI as it assumes that the demand distribution is exactly known.

Our results suggest that the RO approach is preferred when pharmacy kiosks are installed at new locations and there is limited data available. Once sufficient demand data is collected, the decision-maker may switch to a stochastic model. Our modelling approach is generalizable and may be useful for new products in the retail and clothing industries.



(a) Out-of-sample Fill rate α_{test}



(b) Difference from RO

Figure 4.3: Out-of-sample Performance of Robust Approach against Stochastic, Maxmin, and EVPI

4.4.2 Fill rate under Profit Objective

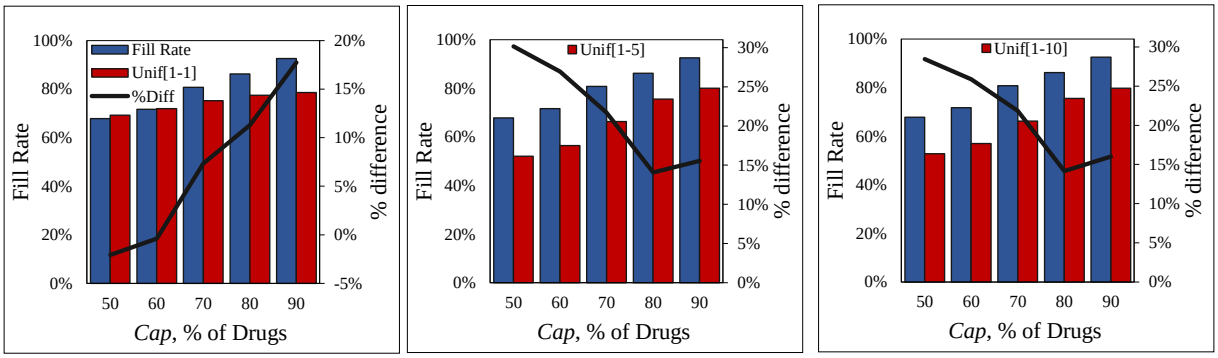
In this section, we compare the proposed fill rate objective against the more commonly used profit maximization approach using a pharmacy store data with 1650 drugs. Let p_i be the profit margin for product $i \in I$ and the profit objective is

$$\max Z = \sum_{i \in I} \left(p_i \left(d_i - \max\{0, d_i - x_i\} \right) \right) \tag{4.4.1}$$

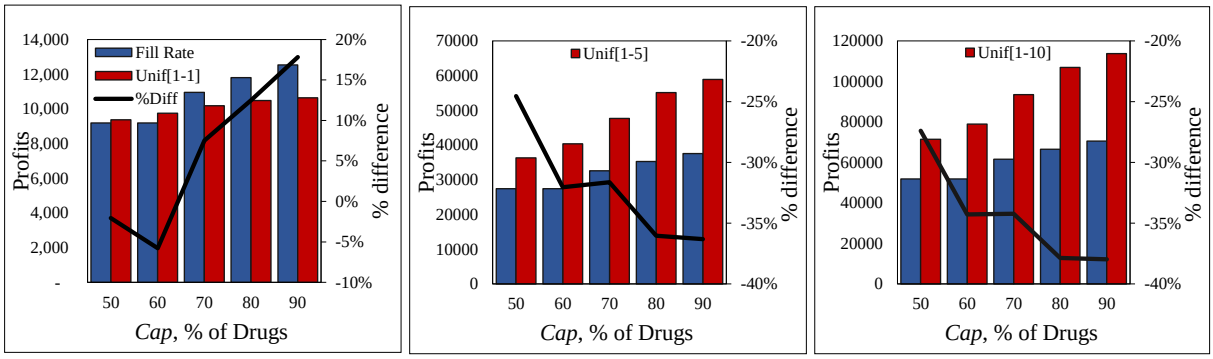
where $d_i - \max\{0, d_i - x_i\}$ denotes the number of successful transactions. Our proposed solution methodology is generalizable allowing us to solve the problem under the profit maximization objective as well.

We randomly generate product profit margins $\mathbf{p} = [p_i]$ between 1 and M from a uniform distribution $unif[1, M]$, where $M \geq 1$. For split ratio 0.1, we solve each instance at three

values of $M \in \{1, 5, 10\}$, to examine how out-of-sample fill rate α_{test} and profit change with increasing variability in profit margins for the two objective functions. Results are illustrated in Figure 4.4. Figure 4.4(a) compares out-of-sample fill rate α_{test} achieved under the fill rate objective denoted by blue bars and that achieved under the profit objective shown by red bars. As expected, the profit maximization approach leads to a lower out-of-sample fill rate compared to the fill rate objective. On average, the fill rate objective yields a 17% higher out-of-sample fill rate. The Profit function with $M = 1$ is equivalent to maximizing the number of successful transactions. The latter leads to a 7% reduction in fill rate, on average, compared to the fill rate objective. This difference increases to 21% for $unif[1, 10]$.



(a) Effect on out-of-sample fill rate



(b) Effect on out-of-sample profit

Figure 4.4: Comparative Analysis of Fill Rate vs Profit Objectives

We now examine the effect of fill rate objective on profits. Figure 4.4(b) plots out-of-sample profits achieved under both objectives. On average, the fill rate objective results in 20% decrease in profits. However, for instances with $M = 1$, the fill rate objective outperforms the profit maximization approach by 6%, on average. This suggests that when products have comparable profit margins, the fill rate maximization objective is preferable as it would not only result in a higher fill rate but also leads to higher profits. In contrast, when product profit margins vary significantly, for instance when profit margins are generated from $unif[1, 10]$, the fill rate objective leads to a 34% decrease in profits, on average. We also observe these differences to be large when capacity is high. For instance, for $Cap = 50\%$, profits under fill rate objective are 27% lower and the difference is amplified to 38% for capacity $Cap = 90\%$.

Our analysis shows that maximizing fill rate objective as opposed to the traditional profit maximization approach is preferred for settings where capacity is limited and products have similar profit margins. Self-serve Kiosks are a great example of such a setting as it has limited capacity and it stocks similar products with roughly the same profit margins. The fill rate objective is even more appropriate during the early stages of kiosk deployment when the decision-maker is willing to compromise short-run profits to build long-run customer confidence.

4.4.3 Testing on Random Instances

We carry out comparative analysis over randomly generated instances to identify settings under which the RO approach would lead to better decisions compared to the stochastic approach. We consider a total of 2000 products and 365 time periods. We consider low-demand products satisfying Definition 1. It follows that the lower bound l_i on daily demand is set to 0. To generate random demand $\mathbf{d} = [d_{it}]$, we first define interval $[0, \mu_i]$ by randomly generating upper bound μ_i for each product i from probability distribution P^{μ}. Demand values d_{it} for $\mathcal{T}_i$ number of days are then randomly generated from interval $[1, u_i]$ using probability distribution P_i^D where $\mathcal{T}_i$ is randomly generated from distribution $P^{\mathcal{T}}$. We assume that $d_{it} > 0$ on $\mathcal{T}_i$ days and $d_{it} = 0$ for the remaining days.

As such, we require to generate two random parameters: (1) upper bound μ_i on daily demand, and number of days $\mathcal{T}_i$ when daily demand $d_{it} > 0$ for each product i. The two random parameters are generated from probability distributions P^{μ} and $P^{\mathcal{T}}$, respectively. Func-

tion $f_r(y) = 1/(\beta_r)^y$ is used to define probability density function for distribution P^r as $p_r(y) = f_r(y)/\sum_y f_r(y)\ \forall y \in \{1, ..., \max^r\}$ where $r = \{\mu, \mathcal{T}\}$, β^r is a nonnegative constant parameter and $\max^r$ is the maximum value that random parameter r can take.

For the base case, we consider $\max^\mu = 3$ i.e., upper bound $\mu_i \leq 3\ \forall i \in I$, and $\beta^\mu = 7.5$. We consider three distinct values of $\beta^\mu \in \{0.1, 1.0, 7.5\}$ and resulting distributions when $\max^\mu = 3$ are illustrated in Figure 4.5. When $\beta^\mu = 0.1$, most of the products have upper bound value $\mu = 3$ while for $\beta^\mu = 7.5$, the probability shifts towards upper bound value 1. To randomly generate the number of days $\mathcal{T}_i$, we use probability distribution $P^\mathcal{T}$ with base case values $\max^\mathcal{T} = 36$ i.e., $\mathcal{T}_i \leq 36\ \forall i \in I$, and $\beta^\mathcal{T} = 1.2$. Figure 4.6 plots different probability distributions derived by varying $\beta^\mathcal{T}$ at three levels $\{0.8, 1.0, 1.2\}$. Once $\mathcal{T}_i$ is generated, we randomly generate $\mathcal{T}_i$ demand values in interval $[1, \mu_i]$ for each product i using probability distribution P_i^D which is defined in a similar fashion as P^r with parameters β^D and $\max^D = \mu_i$. We consider three different values of $\beta^D \in \{0.5, 1.0, 1.5\}$ with 0.5 being the base case value. Each generated demand value is then randomly assigned to a time period between 1 and 365 while for all other time periods, demand d_{it} is set to 0.

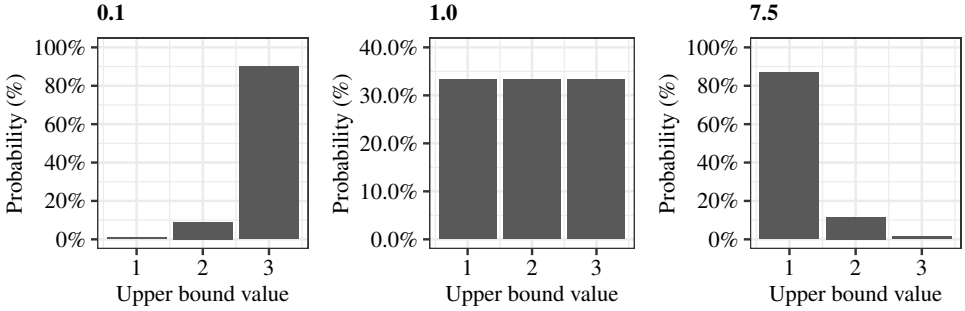


Figure 4.5: Effect of constant parameter β^μ on upper bound probability distribution P^μ

Note that the base case values are tuned to match actual store data distribution. Once demand data $\mathbf{d} = [d_{it}]$ is generated, we divide it into training and testing data based on SplitRatio at two levels $\{0.1, 0.2\}$ with 0.1 being the base case value. Training data is used to make stocking decisions using robust and stochastic models while testing data is used to estimate out-of-sample performance. A total of 12 demand instances are generated, each solved at 5 different capacity levels resulting in 60 distinct instances. Capacity is set as $Cap \times \sum_{i=1}^{i=2000} \mu_i$ where

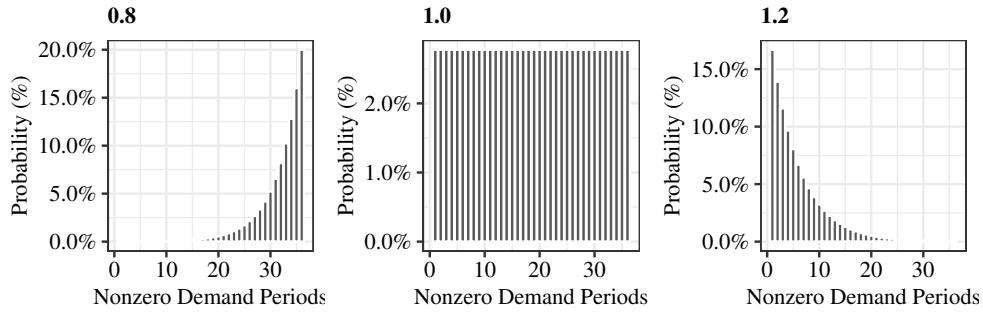


Figure 4.6: Effect of constant parameter $\beta^{\mathcal{T}}$ on nonzero demand days probability distribution $P^{\mathcal{T}}$

Cap is varied between 0.4 and 0.8. All parameter values used for random data generation are summarized in Table 4.2 where the first value denoting the base case instance.

Parameter	Description	Values
NbProducts	Nb. of products	$\{2000\}$
NbPeriods	Nb. of time periods	$\{365\}$
Cap	Capacity as percentage of $\sum UB$	$\{0.4, 0.5, 0.6, 0.7, 0.8\}$
SplitRatio	Training data size (%)	$\{0.1, 0.2\}$
$\max^{\mu}$	Maximum upper bound value	$\{3, 5, 7\}$
β^{μ}	Constant parameter to define probability distribution P^{μ}	$\{7.5, 1.0, 0.5\}$
$\max^{\mathcal{T}}$	Maximum nonzero demand days	$\{36, 72, 182\}$
$\beta^{\mathcal{T}}$	Constant parameter to define probability distribution $P^{\mathcal{T}}$	$\{1.2, 1.0, 0.8\}$
β^{D}	Constant parameter to define probability distributions P_i^D	$\{0.5, 1.0, 1.5\}$

Table 4.2: Testing Parameters

Results are summarized in Table 4.3 where we report the difference between out-of-sample fill rate α_{test} achieved through robust approach and stochastic approach. Our proposed RO approach improves the fill rate by up to 10.59%, and on average, fill rates are 4.02% higher compared to the stochastic approach. However, there are instances where the RO approach does not perform well. Our results indicate that RO outperforms stochastic approach when β^{μ}, $\beta^{\mathcal{T}}$, and β^{D} are high as shown in Figure 4.7. Note that β^{μ} affects upper bound μ_i while β^{D}

Parameters	Value	Capacity Level, Cap					Stats		
		0.4	0.5	0.6	0.7	0.8	Min	Max	Avg
SplitRatio	0.1	0.33	2.17	2.90	9.89	10.10	0.10	10.10	5.08
	0.2	-1.85	-0.65	-0.07	0.38	-2.50	-2.50	0.38	-0.94
β^{μ}	0.5	1.52	4.20	0.28	4.74	7.94	0.28	7.94	3.74
	1	0.21	4.89	5.34	1.47	4.12	0.21	5.34	3.21
	7.5	0.33	2.17	2.90	9.89	10.10	0.33	10.10	5.08
$\max^{\mu}$	3	0.33	2.17	2.90	9.89	10.10	0.33	10.10	5.08
	5	0.31	3.15	3.49	9.82	10.06	0.31	10.06	5.37
	7	0.12	2.67	3.26	9.82	10.06	0.12	10.06	5.19
$\beta^{\mathcal{T}}$	0.8	-0.29	-0.75	0.06	-2.00	-1.08	-2.00	0.06	-0.81
	1	1.36	-0.24	-1.26	0.15	2.87	-1.26	2.87	0.58
	1.2	0.33	2.17	2.90	9.89	10.10	0.33	10.10	5.08
$\max^{\mathcal{T}}$	36	0.33	2.17	2.90	9.89	10.10	0.10	10.10	5.08
	72	0.27	3.70	3.41	9.48	10.29	0.20	10.29	5.43
	182	0.18	2.40	3.37	9.48	10.29	0.18	10.29	5.14
β^{D}	0.5	0.33	2.17	2.90	9.89	10.10	0.33	10.10	5.08
	1	0.59	3.48	2.67	9.88	10.33	0.59	10.33	5.39
	1.5	0.81	3.31	2.74	10.04	10.59	0.81	10.59	5.50
	Average	0.31	2.30	2.39	7.21	7.86	-0.09	8.17	4.02

Table 4.3: Difference in out-of-sample fill rates α_{test} between Robust and stochastic approach

determines daily demand between 1 and upper bound μ_i. When these parameters are set at higher values, the magnitude of demand gets smaller, suggesting that RO performs well when demand is low. On the other hand, $\beta^{\mathcal{T}}$ effects the number of days when $d_{it} > 0$. As $\beta^{\mathcal{T}}$ increases, the number of days with nonzero demand decreases which suggests that RO performs well when demand is rare. On the other hand, as the size of training data increases, the stochastic approach outperforms RO.

The RO approach is preferred over the stochastic approach when there is limited information available relating to demand distribution and under problem settings where demand is low and sporadic. Analysis over capacity reveals that the RO approach performs well when capacity is high as shown in Figure 4.8. This is because when capacity is low, the decision is more straightforward where high-demand products with relatively stable demand are stocked.

4.4.4 Solution Quality

Computational results for the pharmacy case are summarized in Table 4.4 where the first column "# Iter" denotes the total number of iterations between the master problem and the adversarial problem, column "RC" is the time spent to solve the conservative approximation model [RC], and the total time spent during column-and-constraint generation is denoted by "CCG". Column "Gap(%)" reports optimality gap to the original problem [RO] for each instance and is calculated as $UB - LB$. On average, model [RO] is solved to an optimality gap of 40.3% under the fill rate objective. We observe that optimality gaps decrease at higher capacity and lower training data. When we set SplitRatio $= 0.1$, i.e., 36 days are used as training data, the average gap is 23.6% which increases to 62.1% for SplitRatio $= 0.3$ or 110 days of training data. The results indicate that our proposed solution methodology is unable to prove optimality of the solution for real-life instances, this is mainly due to the upper bound which fails to converge in most cases. However, under the profit objective with $M \in \{1, 5, 10\}$, the proposed solution approach is able to solve each instance to optimality within 22 seconds, on average. In fact, under the profit objective, the conservative approximation solution is optimal. For all instances, we find the adversarial problem solution to be the same as conservative approximation. This suggests that the RO model with fill rate objective is inherently difficult to solve compared to the profit objective function.

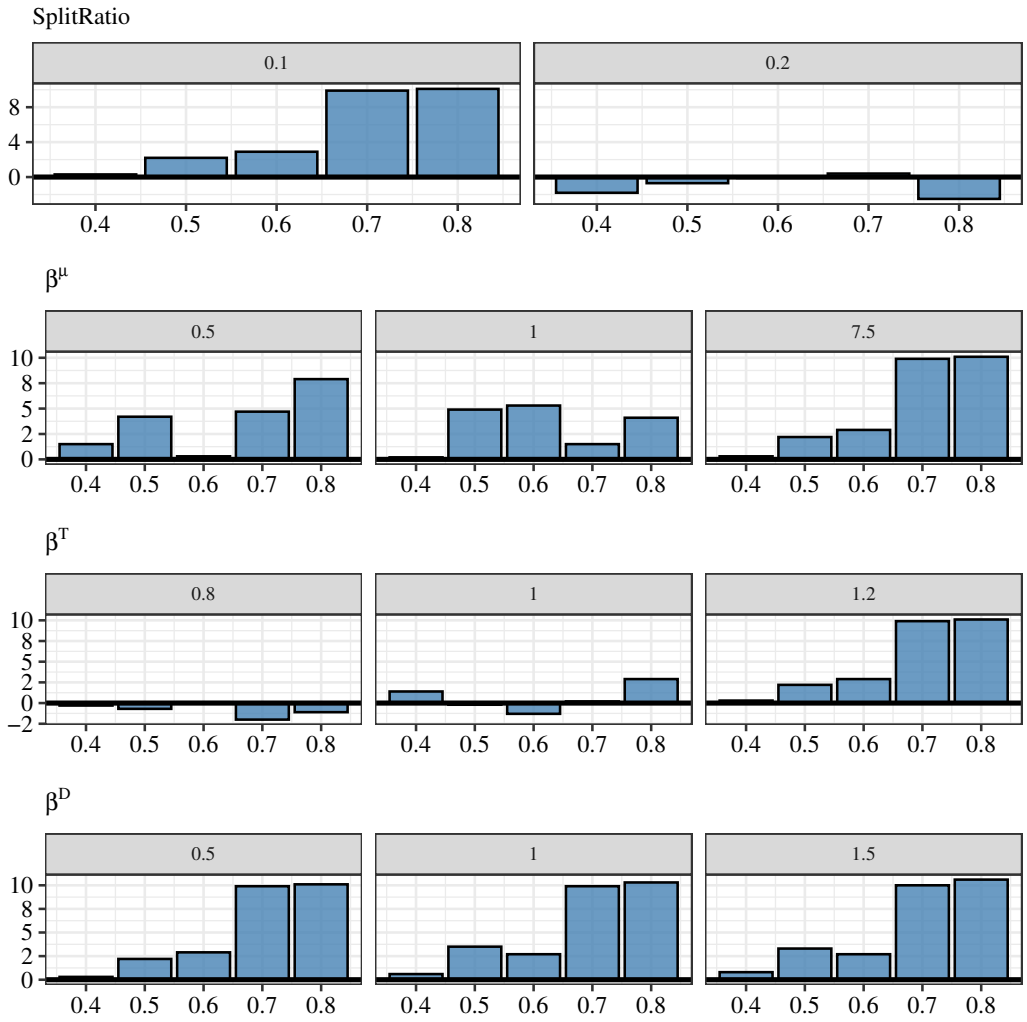


Figure 4.7: Effect of Split Ratio, β^{μ}, $\beta^{\mathcal{T}}$, and β^{D} on RO improvements

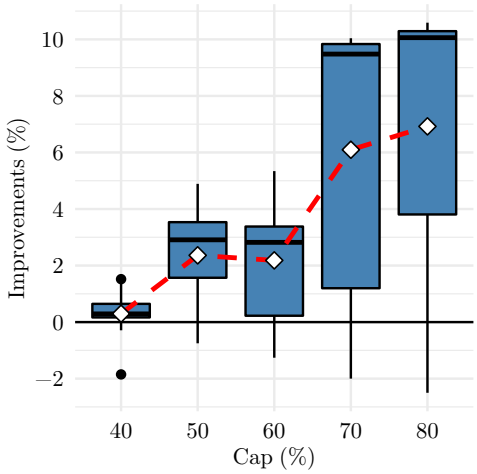


Figure 4.8: RO improvements at different capacity levels

Split	*Cap*,	Fill Rate Objective					Profit Objective				Robust vs Stochastic		
		# Iter	CPU time (s)			Gap	CPU time (s)			Gap	Out-of-sample fill rate α_{test} (%)		
Ratio	% Drugs		RC	CCG	Total	(%)	RC	CCG	Total	(%)	Robust	Stochastic	Difference
0.1	50%	71	186.2	7251.7	7437.8	25.35	11.6	11.6	23.2	0.00	66.19	65.30	0.89
	60%	80	2098.5	7213.6	9312.0	23.68	11.3	11.3	22.6	0.00	69.64	66.68	2.96
	70%	81	3601.5	7211.1	10812.6	19.76	11.1	11.1	22.2	0.00	78.56	69.64	8.92
	80%	81	3600.9	7179.0	10779.8	30.21	10.1	10.1	20.2	0.00	90.89	74.26	16.63
	90%	80	3600.9	7181.7	10782.7	19.04	10.4	10.4	20.8	0.00	94.84	82.42	12.42
	Average	79	2617.6	7207.4	9825.0	23.61	10.9	10.9	21.8	0.00	80.02	71.66	8.36
0.2	50%	15	3600.6	7203.5	10804.1	47.05	14.0	14.0	28.0	0.00	70.11	66.79	3.32
	60%	28	583.3	7201.5	7784.7	35.88	12.1	12.1	24.1	0.00	77.27	75.11	2.16
	70%	81	3601.0	7200.1	10801.1	35.24	14.7	14.7	29.3	0.00	79.51	79.12	0.39
	80%	83	3600.8	7249.6	10850.4	32.70	10.2	10.2	20.4	0.00	95.37	81.49	13.88
	90%	86	3600.9	7207.6	10808.5	22.03	8.5	8.5	17.0	0.00	97.01	86.31	10.70
	Average	59	2997.3	7212.4	10209.8	34.58	11.9	11.9	23.8	0.00	83.86	77.76	6.09
0.3	50%	76	3601.4	7200.3	10801.7	73.68	11.8	11.8	23.6	0.00	72.67	73.24	-0.57
	60%	90	3601.1	7234.9	10836.0	69.05	10.4	10.4	20.9	0.00	79.61	78.23	1.38
	70%	76	3600.9	7195.2	10796.1	61.11	10.1	10.1	20.2	0.00	87.59	84.03	3.56
	80%	47	3603.1	7201.2	10804.3	53.90	9.8	9.8	19.7	0.00	91.34	87.88	3.46
	90%	60	3602.1	7297.8	10899.9	56.00	8.5	8.5	17.1	0.00	96.66	90.58	6.08
	Average	70	3601.7	7225.9	10827.6	62.75	10.1	10.1	20.3	0.00	85.58	82.79	2.78

Table 4.4: Summary of Computational Results for Integrated Conservative Approximation & Column-and-Constraint Generation Appraoch using pharmacy data

In Section 4.4.1, the best found solution from RO approach is often superior compared to other models. We now show statistically that there there is no evidence that the optimality gap has an effect on RO improvements over the stochastic approach. We conduct bivariate analysis to determine statistical relationships between RO improvements and optimality gap, capacity level Cap, and SplitRatio as shown in Figure 4.9. We observe statistically significant negative relationship between RO improvements and optimality gap with $P - value = 0.046 < 0.05$. Note however that bivariate analysis in Figure 4.9 does not accurately depict the relationship as suggested by low $R^2 = 0.27$. One should consider "clustered" linear regression as shown in Figure 4.10 where we cluster data points into two clusters and fit linear line for each set of points separately. Clustered regression improves R^2 to 0.88 and shows no negative relationship between RO improvements and gap. Further analysis of clusters reveals that cluster A consists of data points where $SplitRatio \leq 0.2$ and $Cap \geq 0.7$. This suggests RO improvements not only depend on the optimality gap but it is also affected by capacity level Cap and SplitRatio. Figure 4.9(b) illustrates a strong positive correlation between RO improvements and cap. We observe no statistically significant relationship between RO improvements and SplitRatio. Bivariate analysis shows that RO improvements depend on multiple variables and as

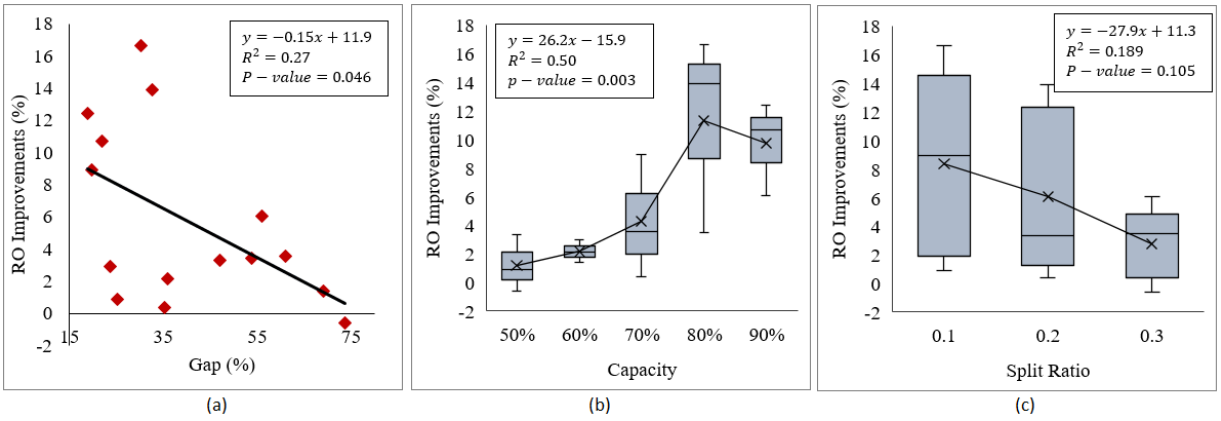


Figure 4.9: RO improvements as a function of optimality gap, cap, and SplitRatio

such, multivariate regression analysis should be carried out to confirm correlation between improvements and optimality gap. To do so, we use RO improvement as predictor variable and optimality gap, cap, and SplitRatio are used as independent variables. Table 4.5 summarizes regression results where independent variable gap has $P - value = 0.184$ which is significantly

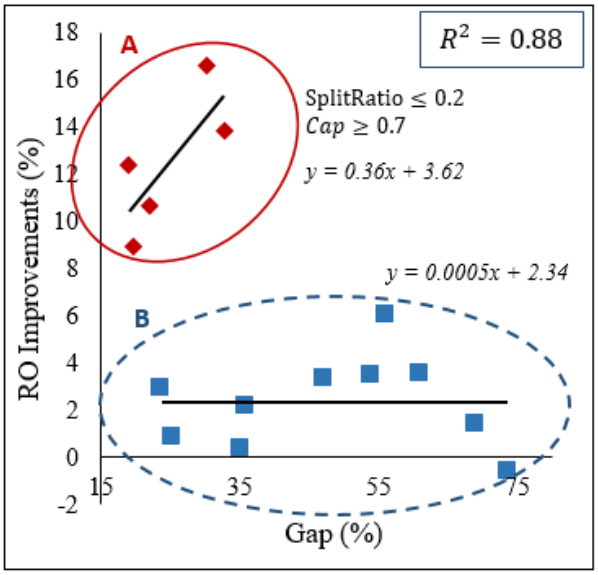


Figure 4.10: Effect of Gap revisited

high. As such, there is no statistically significant evidence even at 90% confidence level to claim that RO improvements decrease as gap increases.

Variable	Coefficients	Standard Error	t Stat	P-value
Intercept	-12.18	5.82	-2.091	0.061
Cap	33.26	7.59	4.384	0.00109*
Gap	0.193	0.14	1.418	0.184
Split Ratio	-65.66	28.40	-2.312	0.0411*

(*) significant at 95% confidence level
$R^2 = 0.737$

Table 4.5: Mulitvariate regression analysis

Statistical analysis suggests that high gaps do not effect RO improvements which may be due to best found solutions being close to optimality. We now explain it using an illustrative example. Consider three products with uncertainty set defined as $d_i \in [0, 2]$ and total demand $1 \leq d_1+d_2+d_3 \leq 4$. Assume that capacity $C = 5$ and we start with conservative approximation solution $\mathbf{x} = (x_1, x_2, x_3) = (2, 2, 1)$ which is optimal to the original problem. The adversarial

problem [AP] outputs demand $\mathbf{d}$ that minimizes α for given solution $\mathbf{x}$. In iteration 1, fill rate is minimized when $d_3 = 2$ while $d_1 = d_2 = 0$ and $\alpha = 0.5$. In the second iteration, an optimal master problem [MP] solution is $(1, 2, 2)$ with $\alpha = 1$ and AP outputs solution $\mathbf{d} = [2, 0, 0]$ with $\alpha = 0.5$. Note that this procedure continues providing optimal solutions by assigning the same stock levels but for different products resulting in no change in gap.

Iteration	Master Problem Solution				Adversarial Problem Solution				Gap
#	x_1	x_2	x_3	UB	d_1	d_2	d_3	LB	
1	2	2	1	1	0	0	2	0.5	0.5
2	1	2	2	1	2	0	0	0.5	0.5
3	2	1	2	1	0	2	0	0.5	0.5
4	2	2	1	0.5	0	0	2	0.5	0

Table 4.6: A Small Illustrative Example

4.5. Conclusions

We studied a multiproduct capacitated newsvendor problem with fill rate maximization objective and presented a robust optimization framework to deal with low and sporadic product demand. We proposed a novel approach to define the uncertainty set for robust optimization using a hierarchical clustering algorithm where negatively correlated products are clustered together. We presented an exact solution methodology to deal with nonlinear fill rate objective using an integrated column-and-constraint generation and conservative approximation approach. We showed that our solution approach is able to solve 20% of the small-sized problem instances to optimality, and on average, the optimality gap is 21%. Numerical testing for the case of pharmacy kiosks confirmed the effectiveness of the proposed modelling approach, resulting in a 5.8% improvement in average daily fill rate, compared to stochastic and maxmin approaches. In addition, we carried out comparative analysis between our proposed robust approach and stochastic modelling through randomly generated instances. The results suggested that the RO approach outperforms the stochastic approach when product demand is low and rare, and limited demand information is available. The proposed RO approach could be used

for newly installed pharmacy kiosks with no or little demand information. Our work lay the foundation for further exploration of robust newsvendor problems under fill rate objective. One future research direction is to incorporate supplier-driven product substitution where similar products can substitute each other. Another future research area worth exploring is to build a machine learning model e.g., support vector regression and quantile regression, to construct a minimum-sized uncertainty set that may allow us to solve real-life instances to optimality.

Chapter 5

Kiosk Location-Inventory Problem with Accessibility Considerations

5.1. Introduction

Accessibility refers to people's access to services and commodities that are essential in improving their quality of life (Kwan 2013). In the context of healthcare, accessibility is an important subject as it directly relates to the overall health of a population. According to a report by National Quality Forum, the problem of accessibility in healthcare services arises mainly due to long distances to care sites and lack of transportation (HealthLeaders 2018). Self-serve pharmacy kiosks may partially address the issue of healthcare accessibility by not only providing medications in close proximity to customers but it can also be used for other health-related clinical services that do not require in-person visits.

In this chapter, we study the potential role of self-serve pharmacy technology in improving accessibility, particularly in rural regions. Our goal is to model accessibility as a function of spatial locations of pharmacy kiosks that would provide foundations for optimally placing multiple kiosks in a given region. However, placing kiosks in close proximity to customers does not completely address the issue of accessibility. In fact, accessibility is also affected by the unavailability of healthservices and customer acceptability to these services (Cabrera-Barona

et al. 2017).

In order to improve healthcare accessibility, one must therefore consider spatial accessibility (Guagliardo 2004) which refers to the combination of both accessibility and availability. Accessibility dimension refers to the travel distance or travel time to a healthcare facility while the availability dimension captures the number of healthcare facilities that an individual may choose from as well as whether services are actually available (Penchansky and Thomas 1981). As such, individual access to healthcare is a function of four distinct factors, (1) number of healthcare facilities, (2) distances to these facilities, (3) customer willingness to visit each facility, and (4) availability of services.

We model accessibility as a function of these factors by considering a multi-kiosk inventory problem where each kiosk is periodically replenished by a central pharmacy store that is farther away. We model the problem as a newsvendor problem with fill-rate dependent demand where the goal is to decide on the stock level or capacity at the kiosks such that the weighted sum of total cost and expected travel distance is minimized. The latter depends on the kiosk fill rate as well as customer willingness to visit kiosks. Kiosk fill rate is a function of its capacity as well as customer demand (willingness) which is modelled using a multinomial logit (MNL) model where the utility derived from a pharmacy location is, in turn, a function of the expected distance. Locational decisions are not explicitly modelled and are captured by capacity decisions. Kiosks with an optimal capacity greater than zero are located.

The resulting multi-kiosk inventory problem is computationally difficult to solve and we therefore approximate it as a dynamic multi-stage game that is solved using a simple iterative heuristic procedure. Sensitivity analysis is carried over modelling parameters using an illustrative example to derive insights.

The rest of this chapter is organized as well follows. In Section 5.2, we review related work on modelling competition, inventory models with endogenous demand, and inventory-location problems. In Section 5.3, we formally define the accessibility function and model it within a multi-kiosk inventory problem. The latter is approximated by a dynamic multi-stage game in Section 5.4. Sensitivity analysis over modelling parameters are presented in Section 5.5 and some concluding remarks are given in Section 5.6.

5.2. Literature Review

Our work essentially relates to inventory problems with endogenous demand, location-inventory problems under stochastic demand, and competitive location problems. We now review literature under each stream and position our work accordingly.

5.2.1 Inventory Planning with Endogenous Demand

Retailers' future demand and long-term sustainability are likely to depend on customer past experience. For kiosks with limited capacity, poor service levels may eventually lead to customers switching to traditional in-store shopping. To address this, researchers have suggested adding service level constraints or stock-out costs (Chen and Chuang 2000, Taleizadeh et al. 2008, 2009, Waring 2012, Abdel-Aal et al. 2017). These approaches are based on exogenous demand assumption and stock-out cost in terms of customer goodwill loss is often difficult to measure (Schwartz 1966). Rather than explicitly incorporating stock-out costs, we model demand as an endogenous variable of the fill rate which is determined by inventory decisions.

In the literature, endogenous demand is modelled as a function of inventory and service levels. For models with inventory-dependent demand, two types of modelling frameworks are used: (1) "initial", where demand is a function of initial inventory level (Gerchak and Wang 1994, Dana Jr and Petruzzi 2001, Wang and Gerchak 2001), and (2) "instantaneous", where demand at a given time is a function of the inventory level at that time (Baker and Urban 1988, Datta and Pal 1990, Balkhi and Benkherouf 2004, Goyal and Chang 2009)

Schwartz (1966) introduced the idea of perturbed demand where the demand rate $\lambda = \frac{\lambda_0}{1+\gamma I}$ is modelled as a function of disappointment factor γ which is the proportion of the demand backlogged and I is a penalty parameter. Initial results of Schwartz (1966) are extended by Schwartz (1970) and Caine and Plaut (1976) to stochastic demand case. Ernst and Cohen (1992) extends the modelling approach to a coordinated distribution system where the demand is a function of fill rate and is modelled as $D(X) = (1 + v(\alpha - \alpha_0))D_0$ where α_0 is the current fill rate with demand D_0 while v captures the rate of change in demand per unit deviation from the current fill rate α_0.

These essentially model future demand as a function of fill rate and are more appropriate under a single-firm setting. Our problem deals with a competitive environment where the total demand is fixed and is allocated to multiple firms/locations based on customer preference. The competitive setting makes the problem much more complicated. In this stream, Wang and Gerchak (2001) use an initial-inventory dependent demand model in the context of competition between two retailers where demand is a function of shelf space allocated. Balakrishnan et al. (2004) evaluate the finite horizon inventory model under instantaneous inventory-dependent demand. The authors show that applying the EOQ model iteratively by updating the demand rate leads to a unique equilibrium solution that may not necessarily be optimal. We use a similar iterative heuristic approach to solve the optimization model proposed in this chapter. Our computational results show that optimality gaps obtained from the proposed heuristic approach are small for most of the instances.

The assumption that customers can observe firm's inventory level is more applicable in the context of in-store shopping where items are displayed. In other settings, customers do not have access to actual inventory levels but rather perceive it based on their past experience. In this stream, Hall and Porteus (2000) considers a multi-period dynamic model with two firms competing based on the capacity that measures customer service level. The authors allow service level (measured in terms of capacity) to vary over time. Customer behavior is modelled such that a service failure leads to an immediate response to switch firms. In addition, Hall and Porteus (2000) assume that a customer switches its first-choice firm based on an exogenous loyalty factor. In contrast, we consider the change in customer choice based on a utility model. A similar approach is followed by Gans (2002) and extended by Gaur and Park (2007). Gans (2002) models customer choice in response to random variation in quality offered by competing firms. A customer picks a firm that maximizes his/her expected utility in each period. We use a multinomial logit (MNL) market share model similar to the one used by Gaur and Park (2007). The authors model asymmetric learning where customers' perceived fill rate may differ from the actual fill rate offered by the firm. Dana Jr and Petruzzi (2001) extends the newsvendor problem where demand is a function of both fill rate and price, and customers decide to visit the retailer or an outside option based on maximizing their utility. Bernstein and Federgruen (2004) develop a general inventory model in an oligopoly setting with aggregate demand. The latter is divided among competing firms based on price and target fill rates using an MNL model.

Netessine et al. (2006) considers a multi-period inventory model where two firms compete on product availability under the assumption that a firm's demand is its first choice customers as well as the first choice customers from the second firm who may switch due to unavailability of product at the second firm. The authors evaluate various situations including lost sales and backlogging.

The proposed multi-kiosk inventory problem could be easily extended to a location problem and it is one of our future research works. We therefore present a brief review of the literature on competitive location problems.

5.2.2 Competitive Facility Location

Competition in location models is considered in three ways: (1) static competition, (2) dynamic competition, and (3) competition with foresight. The static competition focuses on locating facilities in a competitive market under the assumption that existing competitors' features do not change. There have been several research publications on static competition but none of them considers inventory decisions (Wu and Lin 2003, Fernández et al. 2007, Aboolian et al. 2007, Baloch and Gzara 2020b, Lin et al. 2020). In competition with foresight, where leader's entrance to market result in new competitors entering the market. (Drezner and Drezner 1998, Rhim et al. 2003, Aboolian et al. 2009).

We study a dynamic competition where competition between kiosks arises due to customer choice behavior and change in the fill rate at given kiosk affects the demand for all other kiosks. Under dynamic competition, competitor's operational features may change when a firm enters the market such as pricing and service level (Tsay and Agrawal 2000, Boyaci and Gallego 2004, Bernstein and Federgruen 2004, Boyaci and Gallego 2004, Meng et al. 2009). Such models mainly deal with inventory decisions with customer utility defined as a function of operational features with no consideration of location decisions. Meng et al. (2009) is, however, an exception that considers a competitive facility location with pricing as a decision variable. We refer the reader to Farahani et al. (2014) and Wang et al. (2015) for a comprehensive review on competitive facility location problems. The existing literature on dynamic competition however does not study the location-inventory problem.

Paper	Location	Inventory	Objective	Operational Features		Fill	Capacity	Distance	Competition	Customer	Demand	Max. Problem size	
				Price	Service level	Rate			Type	Choice	Type	Location	Products
Wu and Lin (2003)	*		max FC					*	Static	MNL	Deterministic	20	1
Fernández et al. (2007)	*		max P					*	Static	MNL	Deterministic	2	1
Baloch and Gzara (2020b)	*		max P					*	Static	MNL	Deterministic	50	1
Aboolian et al. (2007)	*		max MS					*	Static	MNL	Deterministic	320	1
Lin et al. (2020)	*		max MS					*	Static	MNL	Deterministic	400	1
Drezner and Drezner (1998)	*		max MS					*	Foresight	MNL	Deterministic	7	1
Rhim et al. (2003)	*		max P				*	*	Foresight	Linear	Deterministic	2	1
Aboolian et al. (2009)	*		max P				*	*	Foresight	max utility	Poisson	100	1
Tsay and Agrawal (2000)		*	max P	*	*				Dynamic	Linear	Deterministic	2	1
Boyaci and Gallego (2004)		*	max P		*	*			Dynamic	MNL (α)	General Cont.	2	1
Bernstein and Federgruen (2004)		*	max P	*	*	*			Dynamic	MNL (α)	General Cont.	2	1
Meng et al. (2009)	*		Max P	*			*		Dynamic	Linear	Deterministic	10	1

C - cost, P - Profit, MS - market share, FC - flow capture

Table 5.1: Literature on Competitive models in location and inventory problems

5.2.3 Location-Inventory Models

Location-inventory problems (LIP) are widely studied with little attention given to customer choice. Table 5.2 provides a summary of location-inventory problems explored in the literature. Shen et al. (2003) consider a location-inventory problem for a single supplier with multiple retailers under the assumption that demand for each retailer is random and follows a normal distribution. The objective is to minimize fixed location costs and inventory costs which is a function of mean demand and variance for retailers. To hedge against variability in demand, a safety stock level be maintained to ensure the given service level is achieved under the assumption that demand is normally distributed. Shen and Daskin (2005) extends the cost-based location-inventory problem in Shen et al. (2003) to customer service using a (Q, r) inventory model with type I service level requirement. Atamtürk et al. (2012) considers several extensions of the work by Shen et al. (2003) to incorporate capacity constraints, multiple products, correlated demand, and stochastic lead times. The authors present conic quadratic mixed-integer reformulations for nonlinear optimization problems arising from location-inventory problems. Benjaafar et al. (2008) study a location-inventory problem with random Poisson demand and exponential production times with fixed capacity constraints while Gzara et al. (2014) and Wheatley et al. (2015) consider location-inventory problems with time-based service level constraints.

All of the above papers assume direct assignment of customers to the facility by firms such that profits are maximized without taking into account customer choice. In contrast, Berman et al. (2016) study how customer choice in selecting their favorable facility result in deviation

in optimal costs. The authors assume an assignment rule that assigns each customer to the facility maximizing his/her utility. As opposed to incorporate assignment rules, we allow each customer demand to be allocated among various pharmacy locations based on customer choice behavior.

Paper	Objective	Service level	Capacity	Customer Choice	Demand Distribution	Max. Problem size	
						Location	Products
Shen et al. (2003)	min. C	*			Normal	150	1
Atamtürk et al. (2012)	min C	*	*		Normal	25	15
Shen and Daskin (2005)	min C	*			Poisson	263	1
Berman et al. (2016)	min C/UD			max Utility	Normal	50	1
Benjaafar et al. (2008)	min C		*		Poisson	50	1
Gzara et al. (2014)	min C	*			Poisson	100	1
Wheatley et al. (2015)	min C	*			Poisson	120	20
Berman et al. (2016)	min C		*		Poisson	100	1

C - Cost, UD - uncovered demand

Table 5.2: Literature on Location-Inventory Problems

5.3. Multi-kiosk Inventory Planning

Our goal is to model accessibility to pharmacy services as a function of distance customers have to travel to buy medications. Let $j \in J$ be kiosk locations and let 0 be a single pharmacy store located in the region. Demand for products $p \in P$ originates from a set of customer zones $l \in L$. When customers have to buy medications, they may drive to the pharmacy store or to one of the pharmacy kiosks. To model accessibility, it is important to understand customer behavior. If a customer travels to the store, the distance travelled is δ_{0l} as shown in Figure 5.1. On the other hand, if a customer in zone l decides to travel to a kiosk j, the distance travelled may equal δ_{jl} with probability equal to kiosk's fill rate α_j. However, in previous chapters, we showed that kiosks do not enjoy high service levels and as such, the product may not be available. If the product is not available, the customer travels to a store that is located at distance δ_{j0} from kiosk j. The total distance travelled then equals $\delta_{jl} + \delta_{j0}$ with probability $1 - \alpha_j$. Due to the

uncertainty in product availability, it is unclear whether a customer would prefer a store or a nearby pharmacy kiosk to make a purchase.

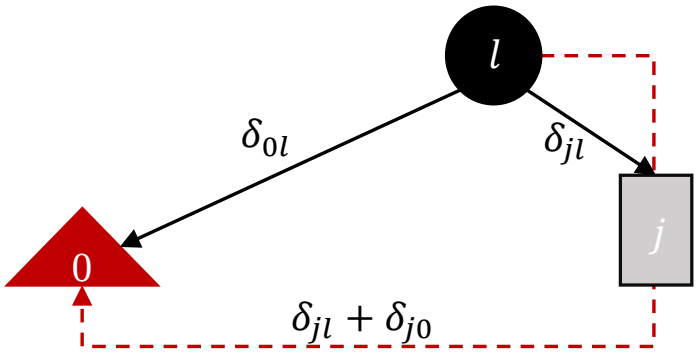


Figure 5.1: Modelling Customer choice behavior

This motivates us to incorporate customer choice behavior in our modelling framework. We use a multinomial logit (MNL) model to estimate the probability of a customer visiting a given pharmacy location (store/kiosk). Let $\overline{\delta}_{jl}$ be the expected distance a customer in zone l has to travel if he/she decides to visit a pharmacy location. We define $\overline{\delta}_{jl}$ as

$$\overline{\delta}_{jl} = \alpha_j \delta_{jl} + (1 - \alpha_j)(\delta_{jl} + \delta_{j0}) \tag{5.3.1}$$

We define customer utility for a particular location j as $U_{jl} = \upsilon_{jl} + \xi_{jl}$, where υ_{jl} refers to intrinsic utility while ξ_{jl} is a random term that takes into account variations among various customers. The intrinsic utility $\upsilon_{jl} = -\beta\overline{\delta}_{jl}(\alpha_j)$ is defined as a function of customer sensitivity parameter β and expected travel distance $\overline{\delta}_{jl}(\alpha_j)$ which depends on fill rate offered by kiosk j. Under standard MNL model, it is assumed that ξ_{jl}'s are i.i.d random variables following a Gumbel distribution (Wang and Wang 2017). Based on random utility maximization scheme, the probability that a customer in zone l decides to visit pharmacy $j \in J_l$ is then estimated as

$$p_{jl} = \Pr(U_{jl} > U_{j'l}, \forall j' \in J_l \setminus \{j\}) = \frac{\exp[-\beta\overline{\delta}_{jl}(\alpha_j)]}{\sum_{j' \in J_l} \exp[-\beta\overline{\delta}_{j'l}(\alpha_{j'})]} \tag{5.3.2}$$

Using customer choice behavior, customer accessibility is measured in terms of the expected

distance travelled to make a purchase and is modelled as

$$A_l = \sum_{j \in J_l} p_{jl} \overline{\delta}_{jl} \tag{5.3.3}$$

$$= \sum_{j \in J_l} \frac{\exp[-\beta \overline{\delta}_{jl}(\alpha_j)]}{\sum_{j' \in J_l} \exp[-\beta \overline{\delta}_{j'l}(\alpha_{j'})]} \times \overline{\delta}_{jl} \tag{5.3.4}$$

Given that the demand originating from zone l is D_l, our goal is to decide on kiosks' stock levels c_j such that the weighted sum of the total expected distance defined by $\sum_{l \in L} A_l D_l$ and the stocking cost is minimized. Let K be the unit cost per stock level and θ be scaling parameter. The optimization problem is

$$\min_{c_j \geq 0} \quad \sum_{l \in L} \sum_{j \in J_l} \frac{\exp[-\beta \overline{\delta}_{jl}(\alpha_j)]}{\sum_{j' \in J_l} \exp[-\beta \overline{\delta}_{j'l}(\alpha_{j'})]} \times \overline{\delta}_{jl}(\alpha_j) E[D_l] + \theta K \sum_{j \in J} c_j \tag{5.3.5}$$

where α_j is defined as

$$\alpha_j = 1 - \frac{E[(\sum_{l \in L} p_{jl}(\alpha_j) D_l - c_j)^+]}{\sum_{l \in L} p_{jl}(\alpha_j) E[D_l]} \tag{5.3.6}$$

It is worth noting that equation (5.3.6) does not have a closed-form solution but rather it needs to be derived numerically.

5.4. Solution Approach

Due to the inherent difficulty of solving the proposed optimization model, we reformulate it as a multi-stage dynamic game. Let $t = 0, 1, 2, ...$ be the index for game stages and let α_j^t be the fill rate observed in period t at kiosk j. Customer choice behavior in period $t + 1$ depends on the fill rate achieved in the previous period t. Initially, customers have no information relating to the kiosk service level at $t = 0$ and perceive it to be equal to 1.0. As time progresses, fill rate is realized and customer choice changes based on fill rate observed in the past. Let υ_{jl}^t be the utility customer expects to derive if he/she decides to visit pharmacy $j \in J_l$. Since customer

utility changes over time, the probability of visiting a kiosk is also time dependent

$$p_{jl}^t = \frac{\upsilon_{jl}^t}{\sum_{j' \in J_l} \upsilon_{j'l}^t} \tag{5.4.1}$$

$$= \frac{\exp(-\beta(\overline{\delta}_{jl}^{t-1}))}{\sum_{j' \in J_l} \exp(-\beta(\overline{\delta}_{j'l}^{t-1}))} \tag{5.4.2}$$

where $\overline{\delta}_{jl}^{t-1}$ is the expected distance travelled by customers during stage $t-1$. The demand q_j^t for kiosk j is then given by

$$q_j^t = \sum_{l \in L} p_{jl}^t \times D_l \tag{5.4.3}$$

with probability density function $f_j^t(q_j^t)$ and a non-decreasing continuous cumulative function $F_j^t(q_j^t)$. The expected fill rate α_j^t achieved is defined as

$$\alpha_j^t = 1 - \frac{E[(q_j^t - c_j^t)^+]}{E[q_j^t]} \tag{5.4.4}$$

where $E[(q_j^t - c_j^t)^+]$ is the expected unfulfilled demand and $E[q_j^t] = \mu_j^t = \sum_{l \in L} p_{jl}^t \times E[D_1]$ is the expected total demand at kiosk j in stage t. The fill rate is

$$\alpha_j^t = 1 - \frac{1}{\mu_j^t} \int_{q_j^t = c_j^t}^{\infty} (q_j^t - c_j^t) f_j^t(q_j^t) dq_j^t \tag{5.4.5}$$

The expected distance travelled by a customer in zone l in order to make a purchase is

$$A_l^t = \sum_{j \in J_l} p_{jl}^t \overline{\delta}_{jl}^t \tag{5.4.6}$$

$$= \sum_{j \in J_l} \frac{\exp[-\beta \overline{\delta}_{jl}^{t-1}]}{\sum_{j' \in J_l} \exp[-\beta \overline{\delta}_{j'l}^{t-1}]} \times \overline{\delta}_{jl}^t \tag{5.4.7}$$

Our objective is then

$$\min_{c_j^t \geq 0} \quad \lim_{t\to\infty} \sum_{l\in L}\sum_{j\in J_l} \frac{E[D_l] \times e^{-\beta \overline{\delta}_{jl}^{t-1}}}{\sum_{j'\in J_l} e^{-\beta \overline{\delta}_{j'l}^{t-1}}} \times \overline{\delta}_{jl}^{t} + \theta K \sum_{j\in J} c_j^t \tag{5.4.8}$$

where the capacity for each kiosk j is $c_j^* = \lim_{t\to\infty} c_j^t$. Our numerical results show that the proposed iterative procedure converges to a steady equilibrium solution. However, we are unable to theoretically prove its convergence and is left for future research.

Given pharmacy locations $J^0 = J \cup \{0\}$ we need to decide on capacity c_j^t that minimizes the expected travel distance at each stage t until a steady equilibrium state is achieved. At each stage t, the subproblem is

$$\min_{c_j^t \geq 0} \quad \pi^t = \sum_{l\in L}\sum_{j\in J_l} \frac{E[D_l] \times e^{-\beta \overline{\delta}_{jl}^{t-1}}}{\sum_{j'\in J_l} e^{-\beta \overline{\delta}_{j'l}^{t-1}}} \times \overline{\delta}_{jl}^{t}(c_j^t) + \theta K \sum_{j\in J} c_j^t \tag{5.4.9}$$

Note that the capacity not only effects the total cost but also the fill rate α_j^t based on which $\overline{\delta}_{jl}^{t}$ is calculated. Simplifying equation (5.4.9),

$$\min_{c_j^t \geq 0} \quad \pi^t = \sum_{l\in L}\sum_{j\in J_l} p_{jl}^t E[D_l] \times (\delta_{jl} + \delta_{0j}) - \sum_{l\in L}\sum_{j\in J_l} p_{jl}^t E[D_l] \times (\alpha_j^t \delta_{0j}) + \theta K \sum_{j\in J} c_j^t \tag{5.4.10}$$

Taking first derivative with respect to c_j^t

$$\frac{\partial \pi^t}{\partial c_j^t} = -\frac{\partial \alpha_j^t}{\partial c_j^t}\left(\delta_{0j}\sum_{l\in L} p_{jl}^t E[D_l]\right) + \theta K = 0 \tag{5.4.11}$$

$$\Rightarrow -\frac{1}{\mu_j^t}(1 - F_{jt}(c_j^t))\left(\delta_{0j}\sum_{l\in L} p_{jl}^t E[D_l]\right) + \theta K = 0 \tag{5.4.12}$$

$$\Rightarrow c_j^t = F_j^{-1}\left(1 - \frac{\theta K}{\delta_{0j}}\right) \tag{5.4.13}$$

Given the stock level c_j^t, we estimate α_j^t and this procedure continues until a steady state is

achieved.

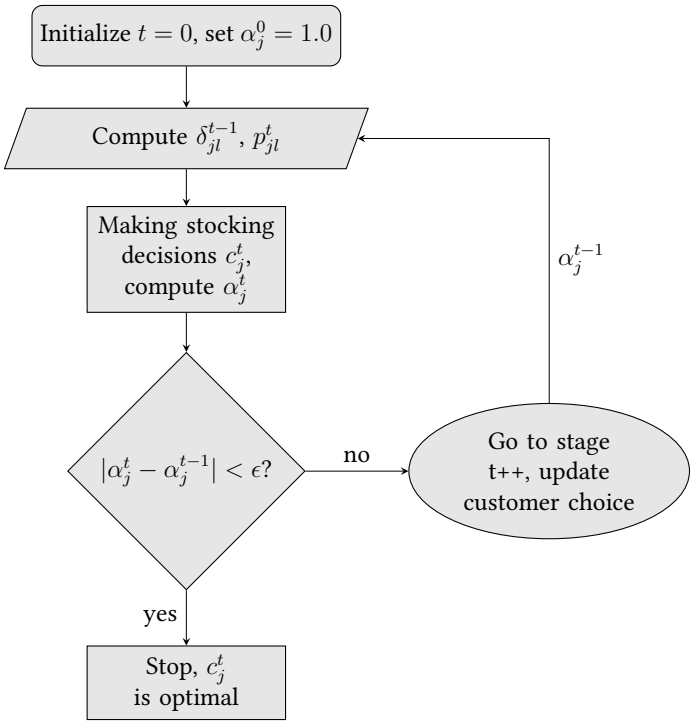


Figure 5.2: Multi-stage Iterative Heuristic Solution Approach

The overall iterative heuristic procedure is summarized in Figure 5.2 where given α_j^{t-1}, the algorithm makes stocking decisions c_j^t and calculates α_j^t. This procedure continues until $|\alpha_j^t - \alpha_j^{t-1}| < \epsilon \; \forall j \in J^0$ where ϵ is a user-defined error tolerance. The latter could be set based on the level of accuracy a manager is interested to achieve. Numerical testing however shows that our proposed algorithm is quite fast even for ϵ=1e-09 and a manager may therefore choose highest level accuracy.

5.5. An Illustrative Example

In this section, we present numerical results for an illustrative example with a single product and single customer zone where customers can either go to the kiosk or pharmacy store. Base

case network is illustrated in Figure 5.3 and data used to carry out the analysis is summarized in Table 5.3. The demand originates from a single customer zone and follows a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. For customer choice behavior, sensitivity to distance β is set to 0.90. The pharmacy store and kiosk are located at distances of $\delta_{0l} = 1.0$ km and $\delta_{jl} = 0.1$ km from the customer zone, respectively. The distance between the store and the kiosk is $\delta_{0j} = 1.1$. Finally, the multi-stage game continues until $|\alpha_j^t - \alpha_j^{t-1}| < \epsilon$ where ϵ is set to 1e-09.

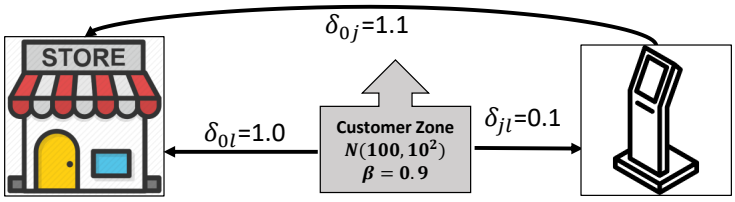


Figure 5.3: Base Case Network

Parameters	Values
Demand characteristics:	$\mu = 100, \sigma = 10$
Customer Choice Behavior:	$\beta = 0.90, \gamma = 0.9$
Distances:	
Customer-Store Distance	$\delta_{0l} = 1.0$
Customer-Kiosk Distance	$\delta_{jl} = 0.1$
Kiosk-Store Distance	$\delta_{0j} = 1.1$
Allowable Error	$\epsilon = 1e - 09$

Table 5.3: Base case Data

5.5.1 Analysis of the Iterative Solution Approach

As discussed in previous sections, the recursive function (5.3.5) could be solved by playing a multi-stage game. We first consider the case where capacity c_j is fixed and the remaining problem is to solve the recursive function (5.3.6) such that the left-hand side of the equation equals the right-hand side. We consider four distinct capacity levels $C \in \{1, 10, 30, 90\}$ while initial fill rate α_j^0 is varied between 0.0 and 1.0. The results are summarized in Figure 5.4 where

we observe that irrespective of the initial value α_j^0, the function converges to the same value. We also note that the convergence rate is fast and all instances are solved within 15 iterations. Figure 5.4 also illustrates how average expected travel distance changes as the game progresses. Note that for capacity $C = 1$, locating a kiosk leads to worse travel distances. This is because some customers may take a chance expecting that the product may be available. However, due to limited stock, the fill rate is low and they often have to travel further to purchase the item from the store. This signifies that the benefits of a kiosk may be realized if there is sufficiently high capacity. For $C = 90$, the kiosk stocks sufficient inventory to achieve an expected fill rate of 100%. Even then, the expected travel distance does not decrease to $\delta_{jl} = 0.1$. This is because some customers may still prefer to visit the store over the kiosk as a result of the market share model. This is one of the drawbacks of the multinomial market share model (5.4.2) where even when the kiosk dominates the store, some demand is still allocated to the store. To address this, one possible solution could be to add dominance rules using a threshold Luce model (Luce 2012) where a kiosk with significantly low fill rate is dominated by other customer choices.

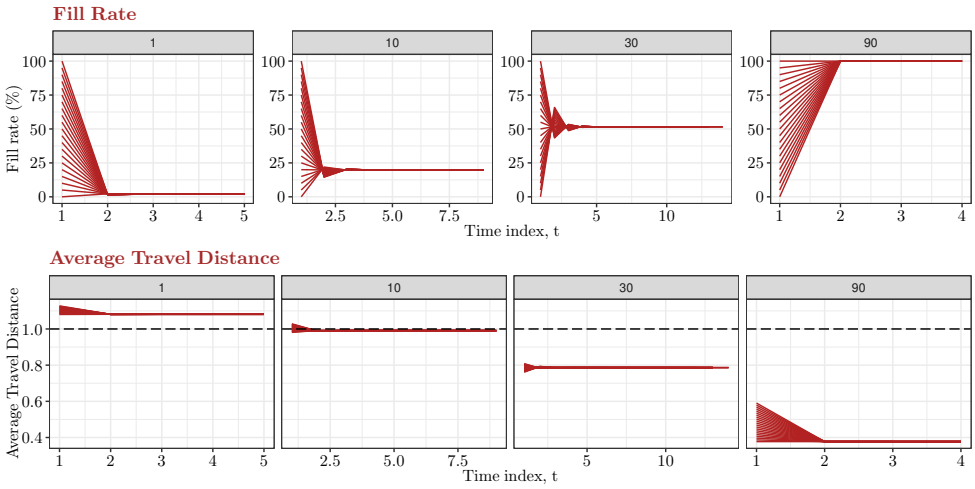


Figure 5.4: Fill rate & travel distance at various capacities during different game stages

5.5.2 Iterative Approach under Optimized Capacity

In this section, we analyze the multi-stage game when capacity c_j is not fixed, but rather updated iteratively as the game progresses. We consider three distinct values of $\theta K \in \{0.2, 0.5, 1.09\}$ and the results are shown in Figure 5.5. As θK increases, the cost of increasing capacity outweighs the reduction in expected travel distance. As such, the capacity decreases with increasing θK. Similar to the earlier case with fixed capacity, the algorithm is fast and converges to a steady-state within a few iterations.

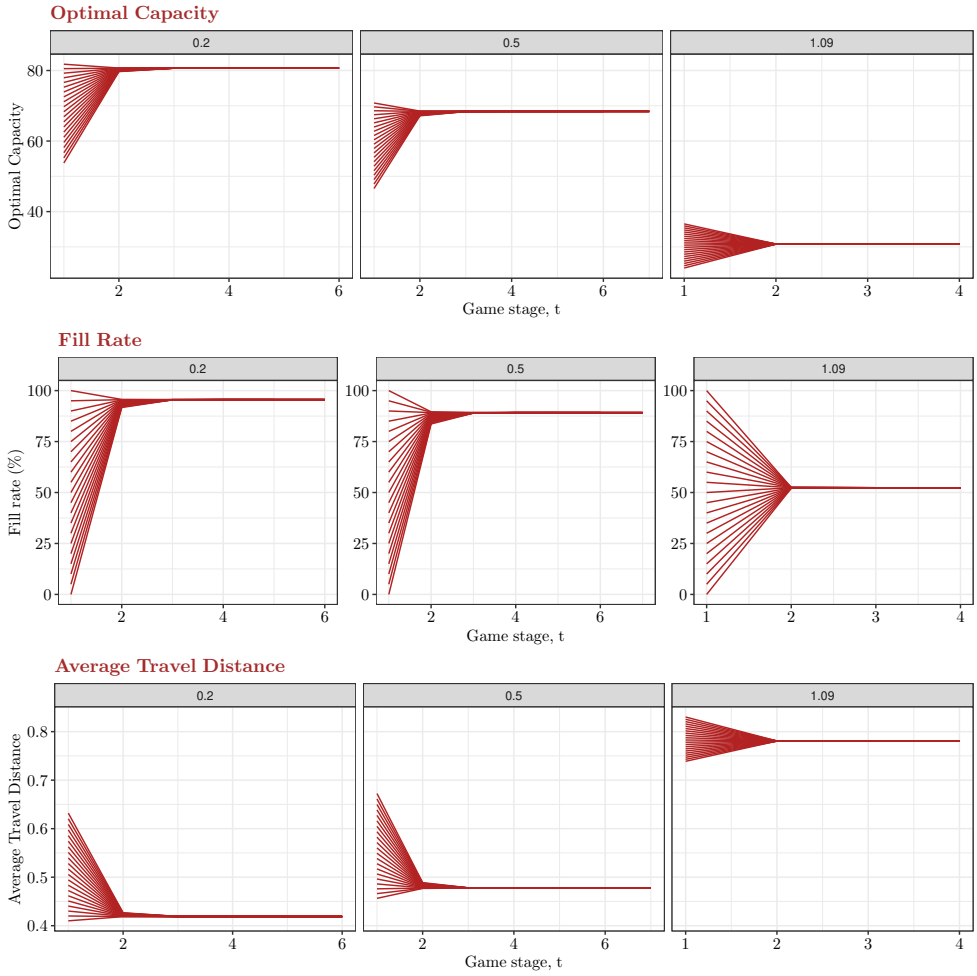


Figure 5.5: Game progression under optimized capacity

It is however in unclear whether the solutions obtained are in fact optimal. We now analyze solution quality by varying model parameters including $\theta K \in \{0.2, 0.5, 1\}$, $\sigma \in \{10, 30, 50\}$, and $\beta \in \{0.1, 0.9, 3.0\}$. To find the optimal solution under each instance, we enumerate over all possible capacity levels. Results are summarized in Table 5.4 where columns “Optimal" and “Heuristic" report optimal and iterative procedure solutions, respectively. It turns out that the iterative solution approach leads to sub-optimal solutions with optimality gaps of 1.50%, on average, and up to 5.70%. These gaps increase as variability in demand increases. As the standard deviation, σ is increased from 10 to 50, the average optimality gap increases from 0.69% to 2.21%. Customer sensitivity to travel distance β also affects solution quality. For instance, when $\beta = 0.1$, average gap is 1.8% which increases to 2.8% for $\beta = 3.0$. It turns out that gaps are high when θK is set too small or too large.

σ	θK	β	Optimal		Heuristic		Optimality
			Obj Value	x_j	Obj Value	x_j	Gap(%)
10	0.2	0.1	65.8	61	66.4	57	0.87
		0.9	54.4	79	54.8	75	0.67
		3	37.4	103	37.5	102	0.08
	0.5	0.1	83.4	56	83.5	53	0.12
		0.9	77.0	71	77.1	69	0.11
		3	67.2	95	67.2	94	0.03
	1	0.1	106.5	40	107.0	45	0.44
		0.9	108.0	50	108.4	57	0.35
		3	107.1	0	110.8	78	3.50
						Average	0.69
30	0.2	0.1	70.6	80	72.3	66	2.46
		0.9	60.1	100	61.3	86	1.90
		3	43.5	123	43.6	118	0.27
	0.5	0.1	92.1	63	92.4	53	0.35
		0.9	86.5	76	86.8	68	0.39
		3	76.5	98	76.7	94	0.14
	1	0.1	109.0	16	110.5	31	1.35
		0.9	109.1	0	111.3	35	2.00
		3	107.1	0	113.2	47	5.73
						Average	1.62
50	0.2	0.1	75.4	98	78.3	75	3.86
		0.9	65.8	120	67.8	97	3.03
		3	49.6	143	49.8	134	0.46
	0.5	0.1	100.7	70	101.3	54	0.55
		0.9	96.0	81	96.7	66	0.73
		3	86.1	102	86.4	94	0.35
	1	0.1	111.6	0	114.0	17	2.15
		0.9	110.8	0	114.1	17	3.03
		3	108.3	0	114.5	18	5.70
						Average	2.21

Table 5.4: Iterative Heuristic Procedure Solution Quality

5.5.3 Effect of Demand Variability

In this section, we evaluate the effect of capacity on fill rate and expected travel distance at different levels of demand variability. Demand variability is captured through standard deviation σ which is varied between $\sigma \in [10, 50]$. Results are summarized in Figure 5.6 where fill rate increases at a decreasing rate as capacity is increased. A higher fill rate leads to lower expected travel distances as shown by the right panel in Figure 5.6. Results show that locating a kiosk with low capacity may result in customer travel distances being even greater than their distance from the store.

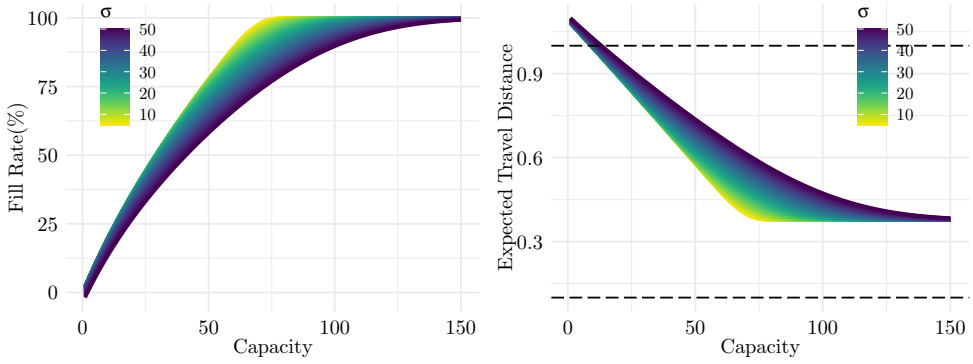


Figure 5.6: Effect of standard deviation σ

5.5.4 Effect of Customer Sensitivity to Travel Distance β

We analyze the effect of customer sensitivity to travel distance by varying parameter β between 0.2 and 3.0 as shown in Figure 5.7. At higher capacities, travel distance decreases exponentially with increasing sensitivity β. In contrast, when capacity is limited, the effect of sensitivity parameter β on the expected travel distance is not significant. As such, accessibility not only depends on kiosk capacity but is also a function of customer sensitivity to travel distance. Kiosk benefits are maximized for distance-sensitive customers. The top dotted line on the right panel of Figure 5.7 denotes customer distance to the central store. Note that when customer sensitivity is low, placing a kiosk with limited capacity may lead to adverse effects. This is shown by the region above the top dotted line where the expected travel distance is higher than customer

distance from the store. This suggests that for customers with low sensitivity to distance, one should place a kiosk with higher capacity to ensure higher fill rates. If such a kiosk is not available, then customers are better off in terms of accessibility when the kiosk is not located.

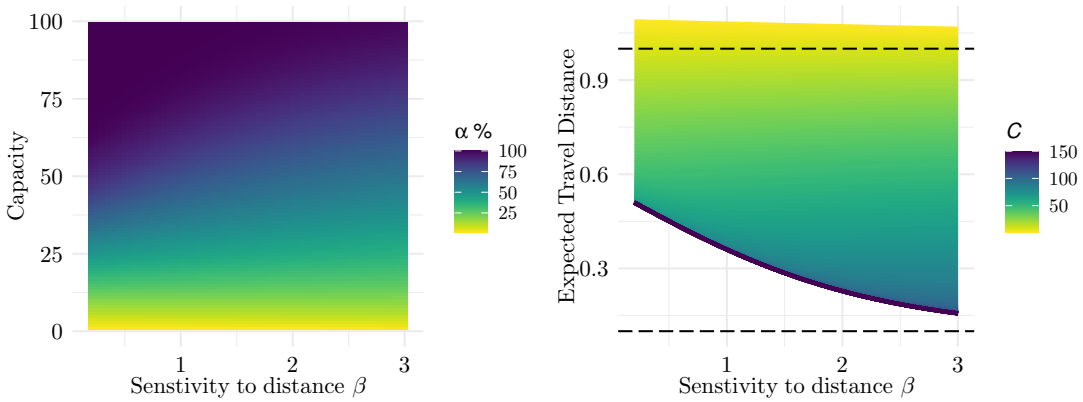


Figure 5.7: Effect of customer sensitivity to travel distance, β

5.5.5 Effect of Distances

Finally, we analyze the effects of distances between kiosk, store, and customer location on travel distances at different capacity levels. Figure 5.8(a) plots expected travel distance as a function of distance between kiosk and customer location δ_{jl}. As expected, the expected travel distance increases as the kiosk is located farther from the customer location. Accessibility could be improved by placing kiosks closer to the customer location only if there is sufficient kiosk capacity. Figure 5.8(b) plots expected travel distance against customer distance from store. In contrast to the kiosk where the expected distance increases almost linearly with δ_{jl}, the store distance δ_{0l} increases expected travel distances at a decreasing rate. Finally, Figure 5.8 examines the effect of distance between store and kiosk δ_{0j} that does not impact travel distances at higher capacity levels. At lower levels, increasing δ_{0j} leads to higher travel distances for customers due to low fill rates.

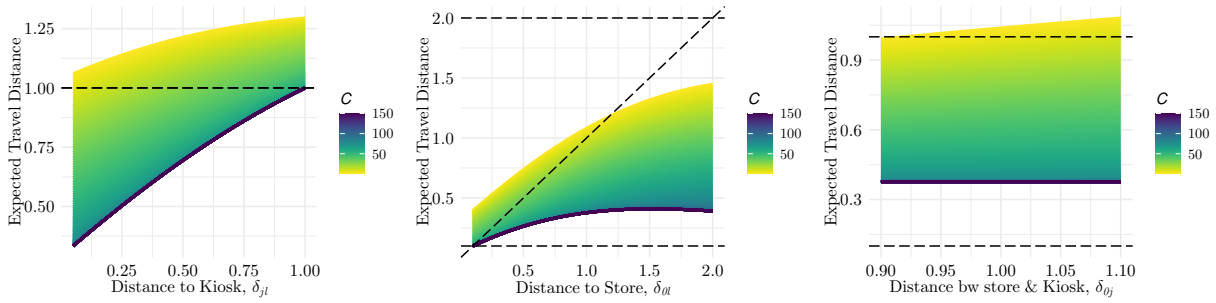


Figure 5.8: Effect of distances

5.6. Conclusions

In this chapter, we proposed a function to model spatial accessibility that accounts for both accessibility and availability dimensions in the context of pharmacy kiosks. The proposed accessibility function is then extended to a multi-kiosk newsvendor problem to optimally decide on the capacity for each kiosk to minimize both stocking cost and the expected distance travelled by the customers. The problem is approximated as a multi-stage game that provides solutions within reasonable optimality gaps of 1.5%, on average. Our results showed that spatial accessibility is only improved if kiosks have sufficient capacity. In fact, in some settings, kiosk service may adversely affect customer accessibility due to limited availability.

For future research, we plan to extend the multi-kiosk inventory problem to a location-inventory problem where one needs to optimally locate p kiosks at potential locations $j \in J$ and to decide on their capacities. We plan to devise a Logic-based Benders decomposition solution methodology that could solve the location-inventory problem optimally. Another promising future research direction is to consider a threshold Luce MNL model to exclude dominated choices for customers.

Chapter 6

Conclusions and Future Research

Self-serve kiosk technology is expected to grow and will play an important role in futuristic smart city logistics, bringing commodities in close proximity to customers. This thesis extends the traditional work on inventory and assortment planning in the context of large retail stores to self-serve kiosks with limited capacity and rare demand.

Motivated by an industry project, this thesis began with addressing the strategic capacity planning problem for pharmacy kiosks through a comprehensive analysis of pharmaceutical sales data which was used to build computationally tractable optimization models. Data analysis revealed that product demand for pharmacy kiosks is low and erratic in nature. As such, a data-driven stochastic optimization approach is used to make stocking and assortment decisions to determine optimized capacity. The issue of limited capacity is partially addressed through supplier-driven substitution where drug demand for higher quantities could be fulfilled by dispensing multiple packages of lower quantities only if it is a multiple of higher quantity. Such a substitution rule has not been previously studied in the literature. A mathematical formulation is proposed that is not only computationally fast to solve large-scale instances but also guarantees robustness against the sequence of demand realization. The proposed modelling approach however allows only one substitute for each quantity. This limits the potential of substitution to improve fill rate as many potential substitution rules that could improve service level cannot be selected. In the future, one may model the exact substitution that would also require developing a sophisticated solution approach to solve the complex problem.

The second part of the thesis addressed the issue of stochasticity of demand using a robust optimization framework under fill rate maximization objective. The proposed framework posed additional challenges in terms of its computational performance and overly-conservative solutions produced by traditional uncertainty sets. In this thesis, the issue of conservative solutions is addressed by constructing a data-driven uncertainty set that is derived purely based on data using a hierarchical clustering algorithm. The resulting RO formulation is solved using an exact solution approach based on the column-and-constraint generation and conservative approximation. A possible future research direction could be to use a constrained version of the support vector regression or quantile regression to estimate bounds on constraints in the polyhedral set such that the size of a polyhedral set is compact while percentage deviation from empirical distribution or training sample is controlled by user-defined threshold.

Finally, an application of self-serve kiosks in the context of improving healthcare accessibility is examined. Accessibility is modelled as a function of the expected distance and customer demand, both of which are fill-rate dependent. A multi-kiosk inventory planning problem is proposed which is solved using an iterative heuristic approach. The latter approximates the inventory problem with fill rate-dependent demand to a dynamic multi-stage game that provides solutions within optimality gaps of 1.5%, on average.

In the context of smart city logistics, one possible extension of the work could be to design an integrated drone-kiosk network where drones are used to replenish pharmacy kiosks that are placed at designated locations. The modelling framework is to decide on the location of pharmacy kiosks, which drugs to stock and in what quantities, and where a pharmacy warehouse be established from where drones make direct deliveries to kiosks. Another promising future research direction could be to consider a network of IoT-enabled kiosks, remote pharmacists, and customers, all connected through a cloud. Customers will have real-time inventory information based upon which they decide whether to visit the kiosk or the traditional brick-and-mortar store.

References

Aardal K, Jonsson Ö, Jönsson H (1989) Optimal inventory policies with service-level constraints. *Journal of the operational research society* 40(1):65–73.

Abdel-Aal MA, Syed MN, Selim SZ (2017) Multi-product selective newsvendor problem with service level constraints and market selection flexibility. *International Journal of Production Research* 55(1):96–117.

Abdel-Malek L, Montanari R, Morales LC (2004) Exact, approximate, and generic iterative models for the multi-product newsboy problem with budget constraint. *International Journal of Production Economics* 91(2):189–198.

Abdel-Malek LL, Montanari R (2005) An analysis of the multi-product newsboy problem with a budget constraint. *International Journal of Production Economics* 97(3):296–307.

Aboolian R, Berman O, Krass D (2007) Competitive facility location and design problem. *European Journal of operational research* 182(1):40–62.

Aboolian R, Sun Y, Koehler GJ (2009) A location–allocation problem for a web services provider in a competitive market. *European Journal of Operational Research* 194(1):64–77.

Agarwal Y, Mathur K, Salkin HM (1989) A set-partitioning-based exact algorithm for the vehicle routing problem. *Networks* 19(7):731–749.

Agrawal R, Srikant R, et al. (1994) Fast algorithms for mining association rules. *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, 487–499.

Ahiska SS, Gocer F, King RE (2017) Heuristic inventory policies for a hybrid manufacturing/remanufacturing system with product substitution. *Computers & Industrial Engineering* 114:206–222.

AiFi (2020) Nanostore - a fully autonomous, retail-ready, and portable store solution. URL https://www.aifi.io/nanostore.

Alfares HK, Elmorra HH (2005) The distribution-free newsboy problem: Extensions to the shortage penalty case. *International Journal of Production Economics* 93:465–477.

Alibabacom (2020) Vending machines, service equipment suppliers and manufacturers. URL https://www.alibaba.com/catalog/vending-machines_cid282905?spm=a2700.details.debelsubf.4.3bb361adrh2meJ.

An Y, Zeng B, Zhang Y, Zhao L (2014) Reliable p-median facility location problem: two-stage robust models and algorithms. *Transportation Research Part B: Methodological* 64:54–72.

Ardestani-Jaafari A, Delage E (2016) Robust optimization of sums of piecewise linear functions with application to inventory problems. *Operations research* 64(2):474–494.

Ardestani-Jaafari A, Delage E (2020) Linearized robust counterparts of two-stage robust optimization problems with applications in operations management. *INFORMS Journal on Computing* .

Atamtürk A, Berenguer G, Shen ZJ (2012) A conic integer programming approach to stochastic joint location-inventory problems. *Operations Research* 60(2):366–381.

Aydin G, Porteus EL (2008) Joint inventory and pricing decisions for an assortment. *Operations Research* 56(5):1247–1255.

Bagchi U, Gutierrez G (1992) Effect of increasing component commonality on service level and holding cost. *Naval Research Logistics (NRL)* 39(6):815–832.

Baker RA, Urban TL (1988) A deterministic inventory system with an inventory-level-dependent demand rate. *Journal of the Operational Research Society* 39(9):823–831.

Balakrishnan A, Pangburn MS, Stavrulaki E (2004) "stack them high, let'em fly": lot-sizing policies when inventories stimulate demand. *Management Science* 50(5):630–644.

Balkhi ZT, Benkherouf L (2004) On an inventory model for deteriorating items with stock dependent and time-varying demand rates. *Computers & Operations Research* 31(2):223–240.

Baloch G, Gzara F (2020a) Capacity and assortment planning under one-way supplier-driven substitution for pharmacy kiosks with low drug demand. *European Journal of Operational Research* 282(1):108–128.

Baloch G, Gzara F (2020b) Strategic network design for parcel delivery with drones under competition. *Transportation Science* 54(1):204–228.

Baloch G, Gzara F (2021) Inventory planning for fill rate maximization in self-serve kiosks. *Submitted to INFORMS Journal on Optimization* .

Bandi C, Bertsimas D (2012) Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical programming* 134(1):23–70.

Baron O, Milner J, Naseraldin H (2011) Facility location: A robust optimization approach. *Production and Operations Management* 20(5):772–785.

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

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. *Operations research letters* 25(1):1–13.

Benjaafar S, Li Y, Xu D, Elhedhli S (2008) Demand allocation in systems with multiple inventory locations and multiple demand sources. *Manufacturing & Service Operations Management* 10(1):43–60.

Berman O, Krass D, Menezes MB (2016) Directed assignment vs. customer choice in location inventory models. *International Journal of Production Economics* 179:179–191.

Bernstein F, Federgruen A (2004) A general equilibrium model for industries with price and service competition. *Operations research* 52(6):868–886.

Bertsimas D, Sim M (2004) The price of robustness. *Operations research* 52(1):35–53.

Bertsimas D, Thiele A (2006a) Robust and data-driven optimization: modern decision making under uncertainty. *Models, methods, and applications for innovative decision making*, 95–122 (INFORMS).

Bertsimas D, Thiele A (2006b) A robust optimization approach to inventory theory. *Operations research* 54(1):150–168.

BestBuy (2020) Best buy express kiosk. URL https://www.bestbuy.ca/en-ca/about/best-buy-express-kiosk/blt84327369f33d27f2.

Boyaci T, Gallego G (2004) Supply chain coordination in a market with customer service competition. *Production and operations management* 13(1):3–22.

Cabrera-Barona P, Blaschke T, Kienberger S (2017) Explaining accessibility and satisfaction related to healthcare: a mixed-methods approach. *Social indicators research* 133(2):719–739.

Caine G, Plaut R (1976) Optimal inventory policy when stockouts alter demand. *Naval Research Logistics Quarterly* 23(1):1–13.

Canadian Institute for Health Information (2021) National health expenditure trends. *Canadian Institute for Health Information* URL https://www.cihi.ca/en/national-health-expenditure-trends.

Carøe CC, Tind J (1998) L-shaped decomposition of two-stage stochastic programs with integer recourse. *Mathematical Programming* 83(1-3):451–464.

Chand S, Ward JE, Weng ZK (1994) A parts selection model with one-way substitution. *European Journal of Operational Research* 73(1):65–69.

Chen M, Chuang C (2000) An extended newsboy problem with shortage-level constraints. *International Journal of Production Economics* 67(3):269–277.

Choi S, Ruszczyński A, Zhao Y (2011) A multiproduct risk-averse newsvendor with law-invariant coherent measures of risk. *Operations Research* 59(2):346–364.

Dana Jr JD, Petruzzi NC (2001) Note: The newsvendor model with endogenous demand. *Management Science* 47(11):1488–1497.

Datta T, Pal A (1990) A note on an inventory model with inventory-level-dependent demand rate. *Journal of the Operational Research Society* 41(10):971–975.

Deflem Y, Van Nieuwenhuyse I (2013) Managing inventories with one-way substitution: A newsvendor analysis. *European Journal of Operational Research* 228(3):484–493.

Desaulniers G, Desrosiers J, Dumas Y, Solomon MM, Soumis F (1997) Daily aircraft routing and scheduling. *Management Science* 43(6):841–855.

Drezner T, Drezner Z (1998) Facility location in anticipation of future competition. *Location Science* 6(1-4):155–173.

Dutta P, Chakraborty D (2010) Incorporating one-way substitution policy into the newsboy problem with imprecise customer demand. *European Journal of Operational Research* 200(1):99–110.

El Ghaoui L, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications* 18(4):1035–1064.

Erlebacher SJ (2000) Optimal and heuristic solutions for the multi-item newsvendor problem with a single capacity constraint. *Production and Operations Management* 9(3):303–318.

Ernst R, Cohen MA (1992) Coordination alternatives in a manufacturer/dealer inventory system under stochastic demand. *Production and Operations Management* 1(3):254–268.

Farahani RZ, Rezapour S, Drezner T, Fallah S (2014) Competitive supply chain network design: An overview of classifications, models, solution techniques and applications. *Omega* 45:92–118.

Fernández J, Pelegrı B, Plastria F, Tóth B, et al. (2007) Solving a huff-like competitive location and design model for profit maximization in the plane. *European Journal of operational research* 179(3):1274–1287.

Ford Jr LR, Fulkerson DR (1958) A suggested computation for maximal multi-commodity network flows. *Management Science* 5(1):97–101.

Fuller JB, O'Conor J, Rawlinson R (1993) Tailored logistics: the next advantage. *Harvard Business Review* 71(3):87–98.

Gallego G, Moon I (1993) The distribution free newsboy problem: review and extensions. *Journal of the Operational Research Society* 44(8):825–834.

Gallego G, Ryan JK, Simchi-Levi D (2001) Minimax analysis for finite-horizon inventory models. *Iie Transactions* 33(10):861–874.

Gans N (2002) Customer loyalty and supplier quality competition. *Management Science* 48(2):207–221.

Gao C, Johnson E, Smith B (2009) Integrated airline fleet and crew robust planning. *Transportation Science* 43(1):2–16.

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Gaur V, Honhon D (2006) Assortment planning and inventory decisions under a locational choice model. *Management Science* 52(10):1528–1543.

Gaur V, Park YH (2007) Asymmetric consumer learning and inventory competition. *Management Science* 53(2):227–240.

Gerchak Y, Wang Y (1994) Periodic-review inventory models with inventory-level-dependent demand. *Naval Research Logistics (NRL)* 41(1):99–116.

Ghaoui LE, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations research* 51(4):543–556.

Gilmore PC, Gomory RE (1961) A linear programming approach to the cutting-stock problem. *Operations research* 9(6):849–859.

Gilmore PC, Gomory RE (1963) A linear programming approach to the cutting stock problem—part ii. *Operations research* 11(6):863–888.

Goyal SK, Chang CT (2009) Optimal ordering and transfer policy for an inventory with stock dependent demand. *European Journal of Operational Research* 196(1):177–185.

Grzybowska H, Kerferd B, Gretton C, Waller ST (2020) A simulation-optimisation genetic algorithm approach to product allocation in vending machine systems. *Expert Systems with Applications* 145:113110.

Guagliardo MF (2004) Spatial accessibility of primary care: concepts, methods and challenges. *International journal of health geographics* 3(1):1–13.

Gzara F, Nematollahi E, Dasci A (2014) Linear location-inventory models for service parts logistics network design. *Computers & Industrial Engineering* 69:53–63.

Hadley G, Whitin TM (1963) Analysis of inventory systems. Technical report.

Hall J, Porteus E (2000) Customer service competition in capacitated systems. *Manufacturing & Service Operations Management* 2(2):144–165.

Hazır Ö, Haouari M, Erel E (2010) Robust scheduling and robustness measures for the discrete time/cost trade-off problem. *European Journal of Operational Research* 207(2):633–643.

HBR (2009) *Harvard Business Review* ISSN 0017-8012, URL https://hbr.org/2009/03/strike-a-balance-between-custo.

HealthcareConference (2017) *MedAvail Technologies Inc.* URL http://medavail.com/media/Cowen-Healthcare-Conference-Investor-presentation.pdf.

HealthLeaders (2018) 3 ways providers can improve healthcare access in rural areas. URL https://www.healthleadersmedia.com/clinical-care/3-ways-providers-can-improve-healthcare-access-rural-areas.

Hendricks D (2014) 5 successful companies that didn't make a dollar for 5 years. URL https://www.inc.com/drew-hendricks/5-successful-companies-that-didn-8217-t-make-a-dollar-for-5-years.html.

Hsieh CC, Lai HH (2019) Pricing and ordering decisions in a supply chain with downward substitution and imperfect process yield. *Omega* .

Huber J, Müller S, Fleischmann M, Stuckenschmidt H (2019) A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research* 278(3):904–915.

Jammernegg W, Kischka P (2013) Risk preferences of a newsvendor with service and loss constraints. *International Journal of Production Economics* 143(2):410–415.

Kaufman L, Rousseeuw PJ (2009) *Finding groups in data: an introduction to cluster analysis*, volume 344 (John Wiley & Sons).

Kelleher SR (2020) Las vegas' airport vending machines are selling ppe. URL https://www.forbes.com/sites/suzannerowankelleher/2020/05/18/las-vegas-airport-vending-machines-are-selling-ppe/.

Khouja M (1999) The single-period (news-vendor) problem: literature review and suggestions for future research. *Omega* 27(5):537–553.

Kim Th, Ramos C, Mohammed S (2017) Smart city and iot.

Kök AG, Fisher ML (2007) Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* 55(6):1001–1021.

Kök AG, Fisher ML, Vaidyanathan R (2008) Assortment planning: Review of literature and industry practice. *Retail supply chain management*, 99–153 (Springer).

Kwan MP (2013) Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and giscience. *Annals of the Association of American Geographers* 103(5):1078–1086.

Lau HS, Lau AHL (1996) The newsstand problem: A capacitated multiple-product single-period inventory problem. *European Journal of Operational Research* 94(1):29–42.

Leachman R, Glassey R (1987) Preliminary design and development of a corporate level production planning system for the semiconductor industry. *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, volume 4, 710–710 (IEEE).

Lee A (2020) Gsp airport installs ppe vending machines for travelers. URL https://greenvillejournal.com/news/gsp-airport-installs-ppe-vending-machines-greenville-sc/.

Lin FC, Yu HW, Hsu CH, Weng TC (2011) Recommendation system for localized products in vending machines. *Expert Systems with Applications* 38(8):9129–9138.

Lin J, Ng TS (2011a) Robust multi-market newsvendor models with interval demand data. *European Journal of Operational Research* 212(2):361–373.

Lin J, Ng TS (2011b) Robust multi-market newsvendor models with interval demand data. *European Journal of Operational Research* 212(2):361–373.

Lin YH, Wang Y, Lee LH (2020) Parcel locker location problem under threshold luce model. *arXiv preprint arXiv:2002.10810* .

Luce RD (2012) *Individual choice behavior: A theoretical analysis* (Courier Corporation).

Mamani H, Nassiri S, Wagner MR (2017) Closed-form solutions for robust inventory management. *Management Science* 63(5):1625–1643.

Maras E (2020) How self-service helps a healing products retailer weather covid-19. URL https://www.retailcustomerexperience.com/articles/hows-self-service-helps-healing-products-retailer-weather-covid-19-2/.

MedAvail (2017) medavail.com. URL http://medavail.com/.

Meng Q, Huang Y, Cheu RL (2009) Competitive facility location on decentralized supply chains. *European Journal of Operational Research* 196(2):487–499.

Montemanni R, Barta J, Mastrolilli M, Gambardella LM (2007) The robust traveling salesman problem with interval data. *Transportation Science* 41(3):366–381.

Moon I, Choi S (1994) The distribution free continuous review inventory system with a service level constraint. *Computers & industrial engineering* 27(1-4):209–212.

Nahmias S, Schmidt CP (1984) An efficient heuristic for the multi-item newsboy problem with a single constraint. *Naval Research Logistics (NRL)* 31(3):463–474.

Netessine S, Rudi N, Wang Y (2006) Inventory competition and incentives to back-order. *IIE Transactions* 38(11):883–902.

Oğuz O (2002) Generalized column generation for linear programming. *Management Science* 48(3):444–452.

Olivares-Nadal AV, DeMiguel V (2018) A robust perspective on transaction costs in portfolio optimization. *Operations Research* 66(3):733–739.

Pachamanova DA (2002) *A robust optimization approach to finance*. Ph.D. thesis, Massachusetts Institute of Technology.

Park YB, Yoo JS (2013) The operation problem of smart vending machine systems. *LISS 2012*, 1013–1017 (Springer).

Penchansky R, Thomas JW (1981) The concept of access: definition and relationship to consumer satisfaction. *Medical care* 127–140.

Pentico DW (1974) The assortment problem with probabilistic demands. *Management Science* 21(3):286–290.

Pentico DW (1976) The assortment problem with nonlinear cost functions. *Operations Research* 24(6):1129–1142.

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Operations Research* 56(1):188–203.

Poon T, Choy K, Cheng C, Lao S (2010) A real-time replenishment system for vending machine industry. *2010 8th IEEE International Conference on Industrial Informatics*, 209–213 (IEEE).

Poursoltani M, Delage E (2019) *Adjustable robust optimization reformulations of two-stage worst-case regret minimization problems* (GERAD HEC Montréal).

Qiu R, Sun Y, Fan ZP, Sun M (2019) Robust multi-product inventory optimization under support vector clustering-based data-driven demand uncertainty set. *Soft Computing* 1–17.

Rajaram K, Tang CS (2001) The impact of product substitution on retail merchandising. *European Journal of Operational Research* 135(3):582–601.

Ramanath P (2019) Tillster releases 2019 self-service kiosk index. URL https://www.tillster.com/press-news/2019/7/16/self-service-kiosk-index.

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

Rhim H, Ho TH, Karmarkar US (2003) Competitive location, production, and market selection. *European journal of operational Research* 149(1):211–228.

Roels G (2006) *Information and decentralization in inventory, supply chain, and transportation systems*. Ph.D. thesis, Massachusetts Institute of Technology.

Rusdiansyah A, Tsao Db (2005) An integrated model of the periodic delivery problems for vending-machine supply chains. *Journal of Food Engineering* 70(3):421–434.

Rusmevichientong P, Topaloglu H (2012) Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations Research* 60(4):865–882.

Sadowski W (1959) A few remarks on the assortment problem. *Management Science* 6(1):13–24.

Scarf H (1958) A min max solution of an inventory problem. *Studies in the mathematical theory of inventory and production* .

Schwartz BL (1966) A new approach to stockout penalties. *Management Science* 12(12):B–538.

Schwartz BL (1970) Optimal inventory policies in perturbed demand models. *Management Science* 16(8):B–509.

See CT, Sim M (2010) Robust approximation to multiperiod inventory management. *Operations research* 58(3):583–594.

Shen ZJM, Coullard C, Daskin MS (2003) A joint location-inventory model. *Transportation science* 37(1):40–55.

Shen ZJM, Daskin MS (2005) Trade-offs between customer service and cost in integrated supply chain design. *Manufacturing & service operations management* 7(3):188–207.

Sherman E (2019) U.s. health care costs skyrocketed to $3.65 trillion in 2018. URL https://fortune.com/2019/02/21/us-health-care-costs-2/.

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

Shin JH, Min D, Wan L, Kim JH (2009) It service 2.0: A case study of smart vending machines in beijing. *Journal of Service Science* 1(2):227–243.

Simchi-Levi D, Wang H, Wei Y (2018) Increasing supply chain robustness through process flexibility and inventory. *Production and Operations Management* 27(8):1476–1491.

Solano A, Duro N, Dormido R, González P (2017) Smart vending machines in the era of internet of things. *Future Generation Computer Systems* 76:215–220.

Soyster AL (1973) Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations research* 21(5):1154–1157.

Taleizadeh AA, Akhavan Niaki ST, Hoseini V (2009) Optimizing the multi-product, multi-constraint, bi-objective newsboy problem with discount by a hybrid method of goal programming and genetic algorithm. *Engineering Optimization* 41(5):437–457.

Taleizadeh AA, Niaki STA, Hosseini V (2008) The multi-product multi-constraint newsboy problem with incremental discount and batch order. *Asian Journal of Applied Sciences* 1(2):110–122.

Tryfos P (1985) On the optimal choice of sizes. *Operations Research* 33(3):678–684.

Tsay AA, Agrawal N (2000) Channel dynamics under price and service competition. *Manufacturing & Service Operations Management* 2(4):372–391.

Turken N, Tan Y, Vakharia AJ, Wang L, Wang R, Yenipazarli A (2012) The multi-product newsvendor problem: Review, extensions, and directions for future research. *Handbook of newsvendor problems*, 3–39 (Springer).

Tütüncü RH, Koenig M (2004) Robust asset allocation. *Annals of Operations Research* 132(1-4):157–187.

Vairaktarakis GL (2000) Robust multi-item newsboy models with a budget constraint. *International Journal of Production Economics* 66(3):213–226.

Wang H, Song M (2011) Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R journal* 3(2):29.

Wang R, Wang Z (2017) Consumer choice models with endogenous network effects. *Management Science* 63(11):3944–3960.

Wang Y, Gerchak Y (2001) Supply chain coordination when demand is shelf-space dependent. *Manufacturing & Service Operations Management* 3(1):82–87.

Wang Y, Wallace SW, Shen B, Choi TM (2015) Service supply chain management: A review of operational models. *European Journal of Operational Research* 247(3):685–698.

Waring AC (2012) Risk-averse selective newsvendor problems. *PhD Dissertation* .

Wheatley D, Gzara F, Jewkes E (2015) Logic-based benders decomposition for an inventory-location problem with service constraints. *Omega* 55:10–23.

Wollmer RD (1992) An airline seat management model for a single leg route when lower fare classes book first. *Operations research* 40(1):26–37.

Wu TH, Lin JN (2003) Solving the competitive discretionary service facility location problem. *European Journal of Operational Research* 144(2):366–378.

Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *Journal of Machine Learning Research* 10(Jul):1485–1510.

Yue J, Chen B, Wang MC (2006) Expected value of distribution information for the newsvendor problem. *Operations research* 54(6):1128–1136.

Zeng B, Zhao L (2013) Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters* 41(5):457–461.

Zhang B, Xu X, Hua Z (2009) A binary solution method for the multi-product newsboy problem with budget constraint. *International Journal of Production Economics* 117(1):136–141.

Appendix A

APPENDICES

A.1. L-shaped Benders Decomposition

In this section, we solve model [M3] using L-shaped Benders decomposition based on the general framework by Carøe and Tind (1998) where the master problem decides on first stage decision variables while the subproblem decides on second stage decision variables. For model [M3], $\mathbf{x} = [x_i]$ and $\mathbf{s} = [s_{ij}]$ are the first-stage variables, and second stage consists of variables $\mathbf{f} = [f_{it}]$. The master problem [MP] is

$$\text{[MP]:} \quad \max \quad 0 + z(\mathbf{x}, \mathbf{s}) \tag{A.1.1}$$

$$\text{s.t.} \quad \textit{Benders Optimality Cuts}$$

$$\sum_{\substack{i \in I: \\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \tag{A.1.2}$$

$$\sum_{i \in I} x_i \leq C \tag{A.1.3}$$

$$x_i \in \mathbb{Z}^+, s_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in I \tag{A.1.4}$$

The optimal solution to [MP] is an upper bound to the original problem [M3] and $z(\mathbf{x}, \mathbf{s})$ is the optimal solution to the subproblem [SP] given $(\overline{\mathbf{x}}, \overline{\mathbf{s}})$

$$\text{[SP]:} \quad \max \quad 1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f_{it}}{D} \tag{A.1.5}$$

$$\text{s.t.} \quad f_{it} \geq \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i \quad i \in I, t \in \Theta, \qquad [\mu_{it}] \tag{A.1.6}$$

$$f_{it} \geq 0, \qquad \forall i \in I, t \in \Theta. \tag{A.1.7}$$

where [.] corresponds to dual variable for constraint (A.1.6). The optimal solution to [SP] provides an upper bound to the lower bound to the original problem [M3]. Note that subproblem [SP] further into sub subproblems for each GPI-QTY $i \in I$ and scenario $t \in \Theta$ as

$$[\text{SP}]_{it} \quad \min \quad f_{it} \tag{A.1.8}$$

$$\text{s.t.} \quad f_{it} \geq \sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i \qquad [\mu_{it}] \tag{A.1.9}$$

$$f_{it} \geq 0, \tag{A.1.10}$$

Let f^*_{it} be the optimal solution to $[\text{SP}]_{it}$, then the optimal solution to [SP] is $1 - \frac{\sum_{i \in I} \sum_{t \in \Theta} f^*_{it}}{D}$. To solve sub subproblem $[\text{SP}]_{it}$, we take its dual

$$[\text{DSP}]_{it} \quad \max \left(\sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i \right) \mu_{it} \tag{A.1.11}$$

$$\text{s.t.} \quad \mu_{ijt} \leq 1, \tag{A.1.12}$$

$$\mu_{it} \geq 0 \tag{A.1.13}$$

which is trivial to solve. The optimal solution $\mu^*_{it} = 1$ if $\sum_{\substack{j \in I: \\ b_{ij}=1}} m_{ij} A_{jt} \overline{s}_{ij} - \overline{x}_i > 0$, else $\mu^*_{it} = 0$. Note that since the subproblem [SP] is always feasible for a given $(\mathbf{x}, \mathbf{s})$, we do not need to add

feasibility cuts (extreme rays) to the master problem [MP]. Let $\mathcal{E}_{it}$ be the set of the extreme points to $[\text{DSP}]_{it}$. The master problem could be written as

$$[\text{MP}]: \quad \max \quad 1 - \frac{\sum_{i\in I}\sum_{t\in\Theta} z_{it}}{D} \tag{A.1.14}$$

$$\text{s.t.} \quad z_{it} \geq \left(\sum_{\substack{j\in I:\\ b_{ij}=1}} m_{ij} A_{jt} s_{ij} - x_i \right) \overline{\mu}^e_{it} \quad i \in I, t \in \Theta, e \in \mathcal{E}_{it}, \tag{A.1.15}$$

$$\sum_{\substack{i\in I:\\ b_{ij}=1}} s_{ij} = 1 \qquad \forall j \in I, \tag{A.1.16}$$

$$\sum_{i\in I} x_i \leq C \tag{A.1.17}$$

$$x_i \in \mathbb{Z}^+, s_{ij} \in \{0,1\} \qquad \forall i \in I, j \in I \tag{A.1.18}$$

Note that the set of extreme points $\mathcal{E}_{it} = \{0,1\}$. For $e = 0$, $\overline{\mu}_{ijt} = 0$, and Constraint (A.1.15) is $z_{it} \geq 0$ which corresponds to the nonnegativity constraint (3.28) in the original formulation [M3]. On the other hand, when $e = 1, \overline{\mu}_{ijt} = 1$ and Constraint (A.1.15) is $z_{ijt} \geq \sum_{\substack{k\in J_i:\\ b_{ijk}=1}} m_{ijk} A_{ikt} s_{ijk} - x_{ij}$ corresponding to constraint (3.27). The approach is equivalent to a cutting plane algorithm where constraints (3.27) and (3.28) in the original model [M3] are dropped and added iteratively.

To warm-start the algorithm, nonnegativity constraints (A.1.15) corresponding to $e = 0$ are included in [MP]. To tighten the relaxation, we also add a set of valid inequality constraints

$$x_i \leq \sum_{\substack{i\in I:\\ b_{ij}=1}} d^{\max}_j s_{ij} \quad \forall \quad i \in I \tag{A.1.19}$$

where $d^{\max}_j$ is the maximum daily demand recorded for GPI j in the sales data. Constraint (A.1.19) ensures that GPI-QTY i is not stocked if $s_{ij} = 0 \forall j \in I$.

A.2. Benchmark Models

A.2.1 Stochastic Model

The stochastic model is

$$\text{[SO]:} \quad \max \quad \alpha \tag{A.2.1}$$

$$\text{s.t.} \quad \sum_{i \in I} x_i \leq C \tag{A.2.2}$$

$$\alpha \leq 1 - \frac{\sum_{i \in I} \sum_{t \in T} \max\left\{0, d_i^t - x_i\right\}}{\sum_{i \in I} \sum_{t \in T} d_i^t} \tag{A.2.3}$$

$$x_i \in \mathbb{Z}_+ \qquad \forall i \in I, \tag{A.2.4}$$

$$0 \leq \alpha \leq 1 \tag{A.2.5}$$

where d_i^t is the demand for product i in training sample $t \in T$. Constraint (A.2.2) is the capacity constraint while constraint (A.2.3) computes fill rate α. The resulting formulation is, however, nonlinear due to max function in Constraint (A.2.3). To linearize, we add auxiliary variables f_i^t to the model as

$$\text{[SO-LR]:} \quad \max \quad \alpha \tag{A.2.6}$$

$$\text{s.t.} \quad \sum_{i \in I} x_i \leq C, \tag{A.2.7}$$

$$\alpha \leq 1 - \frac{\sum_{i \in I} \sum_{t \in T} f_i^t}{\sum_{i \in I} \sum_{t \in T} d_i^t}, \tag{A.2.8}$$

$$f_i^t \geq d_i^t - x_i, \qquad \forall i \in I, t \in T, \tag{A.2.9}$$

$$x_{ij} \in \mathbb{Z}_+, f_i^t \geq 0, \qquad \forall i \in I, t \in T, \tag{A.2.10}$$

$$0 \leq \alpha \leq 1 \tag{A.2.11}$$

A.2.2 Maxmin Model

Maxmin model maximizes the minimum fill rate over all demand scenarios T and is

$$\text{[Maxmin]:} \quad \max \quad \alpha \tag{A.2.12}$$

$$\text{s.t.} \quad \text{(A.2.2)}, \text{(A.2.4)}, \text{(A.2.5)}$$

$$\alpha \leq 1 - \frac{\sum_{i \in I} \max\left\{0, d_i^t - x_i\right\}}{\sum_{i \in I} \sum_{t \in T} d_i^t} \quad \forall t \in T, \tag{A.2.13}$$

where constraint (A.2.13) defines α as the minimum fill rate achieved for all demand scenarios in T. Model [Maxmin] can also be linearized as model [SO].