Title
Data-driven Approaches to Flexible Systems Design

Author
He, Long

Publication Date
2015

Peer reviewed|Thesis/dissertation

Data-driven Approaches to Flexible Systems Design

by

Long He

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Zuo-Jun Max Shen, Chair
Professor Phil Kaminsky
Professor Robert Leachman
Professor Terry Taylor

Spring 2015

Data-driven Approaches to Flexible Systems Design

Abstract

Data-driven Approaches to Flexible Systems Design

by

Long He

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun Max Shen, Chair

This dissertation studies the data-driven approaches to flexible systems design problems under uncertainty. We discuss real applications in various contexts with flexibility: the capability to satisfy different types of customer demands (e.g. one-way and round trips in the context of car sharing systems); the geographical demand distribution estimation and associated inventory allocation; and the freedom in production plans to fulfill uncertain customer demands (e.g. flexible recipes in continuous production process).

The problems we consider have different objectives and more importantly several degrees of richness in data availability. We develop data-driven optimization models accordingly. Specifically, in the case of new market expansion for example, the firm has to make one-shot decision with limited or side information. The focus of data-driven approach in this case is on the portability of information. Distributionally-robust optimization methodologies are applied to derive strategic decisions that hedge the risks. At the tactical level, e.g. resource planning, the firm deploys planning with ample historical data. For online retailers, geographical demand distributions need to be estimated from historical sales and serve as key input to their regular inventory allocation decisions. Furthermore, operational decisions generally require more detailed data, especially the continuous data for real-time decisions. We study the problem where routine production plans are chosen together with raw material investment decisions when periodic demand data may be available.

In the first part of the dissertation, we study the planning problem faced by urban electric vehicle (EV) sharing systems, that offer both one-way and round trips, in designing the geographical service region. This decision encompasses the trade-off between maximizing customer adoption by covering travel needs, and controlling fleet operations costs. We develop a mathematical programming model that incorporates details of both customer adoption behavior and fleet management (including EV repositioning and charging) operations under spatially-imbalanced and time-varying travel patterns. To address uncertainty in customer adoption, we employ a distributionally-robust optimization framework that in-

forms robust decisions to avoid possible ambiguity (or lack) of data. Mathematically, the problem is approximated by a mixed integer second-order cone program (MISOCP), which is computationally-tractable. Applying this approach to the case of Car2Go's service in San Diego, California, with real operations data, we investigate several planning questions and suggest potential for future development of the service.

To make better inventory allocation to distribution centers, understanding of the geographical demand distribution is essential to online retailers who possess historical sales data that might be contaminated and/or with missing data. The second part of the dissertation presents two models: the first model estimates the geographical demand distribution; the second model integrates the demand estimation together with inventory optimization. In the first model, we study the missing geo-demand data completion problem for a national online retailer. We formulate the problem as a low-rank tensor recover problem in a convex optimization framework. An alternating direction augmented Lagrangian (ADAL) method has been developed and tailored for solving the tensor recovery problem with partial observations. We first discuss efficiency and effectiveness of the algorithm via experiments with synthetic data. We then apply the framework with observed geo-demand from the online retailer. Finally, the benefits of the missing geo-demand data completion are summarized based on computational experiment results. We have shown that the recovered geo-demand distributions possesses more smoothness over time and rendered better generalization performance than the observed geo-demand upon integrated into the existing learning framework. We also integrate the missing data recovery with the data-driven newsvendor model which provides estimation of demands as well as optimal order quantity. A preliminary analysis shows that the proposed model preserves the condition for optimal order quantity as it is in the data-driven newsvendor model. Future work directions are also discussed.

The last part of this dissertation focuses on the inventory investment, recipe selection and resource allocation decisions in continuous process systems with flexible recipes under demand uncertainty. Due to variations in both raw material quality and market conditions, variations in the recipes are used in continuous production processes. Such flexibility is not on design but on the operation that allows adjustments of recipe items aiming to achieve better input utilization than traditionally fixed recipes. We develop a two-stage stochastic mixed integer program formulation and propose a heuristic to the second stage allocation optimization problem. In the first stage, the model determines inventory levels for each period based on past demand data. After demand arrivals are realized, the second stage recourse makes recipe selection and allocation decisions in production. With available historical demand data, a simulation-based approach based on SAA algorithm is developed to solve the stochastic program. The results of numerical study show the performance of the approach on various cost settings as well as the benefits of flexible recipes over fixed recipes. In the proposed approach, we focus on the application of the sample average approximation (SAA) algorithm and use Bootstrap sampling as the default in demand simulation. A direction of future improvement is to incorporate better techniques in the simulation of future demand

arrivals based on historical demand data. Those techniques may consider some properties of the demand, such as seasonality and autocorrelation. Also, with limited demand information, a robust optimization model might be developed that considers the worst cases. Moreover, since our model assumes any inventory leftover at the end of each period is disposed, the extension that relaxes this assumption and introduces inventory holding cost in multi-period setting should also be investigated.

To my fiancée Siyu

and my parents, Zhixiang He and Chunxia Guo

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Zuo-Jun Max Shen, for his continuous advice and support. Throughout the last five years, I have received his encouragement and guidance in all aspects of my life and study at Berkeley. He has taught me to not only think critically in research but also create effective learning environment in classroom. His open-mindedness has allowed me to explore the world and gain working experience in various industries.

I would like to thank my dissertation committee members: Professor Phil Kaminsky, Professor Robert Leachman, and Professor Terry Taylor for extending their time, effort and feedback on the parts of my dissertation. Thanks also to Professor Ying-Ju Chen for sharing his comments on my work and offering his help as I prepared for the job market. I am also grateful to Professor Pnina Feldman for the discussion I enjoyed in her seminar course.

I want to thank my collaborators: Professor Ho-Yin Mak, Professor Ying Rong, and Professor Simin Huang. I am touched by their efforts from weekly Skype meeting to discussion face to face with local commute and international travel. I am also grateful to my industry collaborators. Working with Dr. Zhiwei Qin at WalmartLabs has brought new ideas and data-driven methodologies. Without the resource and help they provided, I wouldn't have completed the research projects.

My colleagues and friends have left an indelible mark on my life at Berkeley. These includes Te Ke, Defeng Sun, Siyuan Sun, Wei Qi, Min Zhao, Yujia Wu, Zhao Ruan, Shiman Ding, Renyuan Xu, Jue Chen, Cheng Lu, Zhiwei Xu, Ruoyang Li, Xinyu Cao, Tianhu Deng, Yong Liang, Ye Xu, Mengshi Lu, and all others. I am very thankful for the joy and laughter we had over the years.

Finally, my deeply grateful acknowledgement goes to my family. My parents, Zhixiang He and Chunxia Guo, have patiently supported and encouraged me to pursuit my dream as a scholar. Thanks to my fiancée Siyu, who has made UC Berkeley the most romantic campus since we met. She has accompanied me through the journey, listened to my doubts and shared her devotion in research.

Chapter 1

Service Region Design for Urban Electric Vehicle Sharing Systems

1.1 Introduction

Sustainable transportation initiatives are gaining increasing attention in recent years as the public awareness of environmental issues grows. In 2012, the transportation sector accounted for 28% of total U.S. greenhouse gas (GHG) emissions [87]. Meanwhile, about 70% of U.S. oil consumption can be attributed to transportation activities [92]. To reduce emissions by transportation, innovative solutions in sustainable transportation have been gaining traction. Innovative technological solutions, such as those centered on energy-efficient electric vehicles (EVs), provide realistic alternatives to traditional modes of transportation based on internal combustion engine (ICE) vehicles, while reducing dependence on oil. EVs have no tailpipe emissions, and, when powered by efficient and more diverse sources of electricity (e.g., solar and wind power), can significantly improve on well-to-wheel energy efficiency and emission levels over ICE counterparts. The diversity of power sources also makes EV operations less sensitive to the depletion of fossil fuels as well as supply uncertainty of crude oil. From the consumer's viewpoint, EV enjoys low operational costs: the fuel cost per mile for passenger EVs is around 4 cents in the U.S., compared with 12 cents for ICE vehicles [29]. Despite the potential of EVs, the consumers are not ready to own EVs at a massive scale due to several major hurdles including the short driving ranges coupled with the insufficient charging facilities, the high upfront purchase cost and the possible higher depreciation rate due to faster technology development.

Interestingly, the combo of EVs with car sharing operations emerges globally as a viable alternative to car ownership for urban dwellers [27, 47]. Currently, Car2Go, a subsidiary of Daimler AG, is operating a car sharing system with a full EV fleet in San Diego (USA), Amsterdam (Netherlands) and Stuttgart (Germany). In several other cities including Austin (USA), Vancouver (Canada) and Berlin (Germany), Car2Go offers both EVs and ICEs to its

car sharing members [22]. DriveNow, operated by BMW, serves San Francisco Bay Area with all EVs, and also provides combination of EVs and ICEs in Berlin (Germany) and Munich (Germany) [36]. Autolib have deployed over 2,000 electric vehicles in Paris (France) through its EV sharing service [7]. This innovative operational model offers potential to overcome the major barriers against EV adoption. First, as car sharing systems operate in well-defined urban service areas, concerns over the range limitation are alleviated. The concentration of a sizable fleet within a dense urban area also makes charging infrastructure deployment more amendable. Second, car sharing effectively allows a pool of users to amortize the high fixed costs of purchasing EVs (and maintenance) to usage-based variable costs over their collective consumption of the service. By pooling their driving needs, EVs in sharing fleets can enjoy higher utilization, and thus the average costs can be reduced compared with the case of individual ownership. Third, by retaining ownership, the firm effectively eases the consumers' concerns over technological risks, future resale values, or depreciation.

In addition to introducing EVs to car sharing systems, Car2Go, DriveNow and Autolib also differ from those early car sharing systems, e.g., Zipcar and City CarShare, by allowing both round trips and one-way trips. Specifically, Car2Go allows customers to check out and return cars anywhere within the service region at any street parking slot, while DriveNow and Autolib allow customers to check out and return at any of their stations. This flexibility allows customers to use the service for regular trips with long stopover times (e.g., commuting to office or school) which are typically not economic feasible under Zipcar and City CarShare. Figure 1.1 shows the frequencies of trips classified by distances between origins and destinations (i.e., O-D distances), for Car2Go's operations in San Diego over a one month period. One can observe that the majority of trips are one way, i.e., the O-D distance is beyond walking distance (e.g., $\geq$ 2km).

Although the one-way car sharing system opens up a broader potential customer base, it makes fleet operations more difficult. One key strategic planning involved in this innovative car sharing system is to determine service region. On one hand, expanding geographical coverage entails significant operational challenges, such as the repositioning of cars to ensure availability under unbalanced demand and, in the case of EVs, the scheduling of recharge. On the other hand, customer adoption critically depends on service coverage, as travel needs can only be covered when both the trip origins and destinations are within the service region. Hence, a more extensive service region encourages adoption by covering higher proportions of travel needs, and thereby improves potential revenue.

Here, we address the strategic planning problem of service region design for one-way EV sharing systems. This problem encompasses several challenges. First, the travel pattern and adoption behavior of potential customers are highly uncertain to the firm at the planning stage. Moreover, before entering a new city, the firm does not possess accurate data to describe the uncertainty in terms of probability distributions, which further augments the planning challenge. As strategic commitments such as the acquisition of land for stations

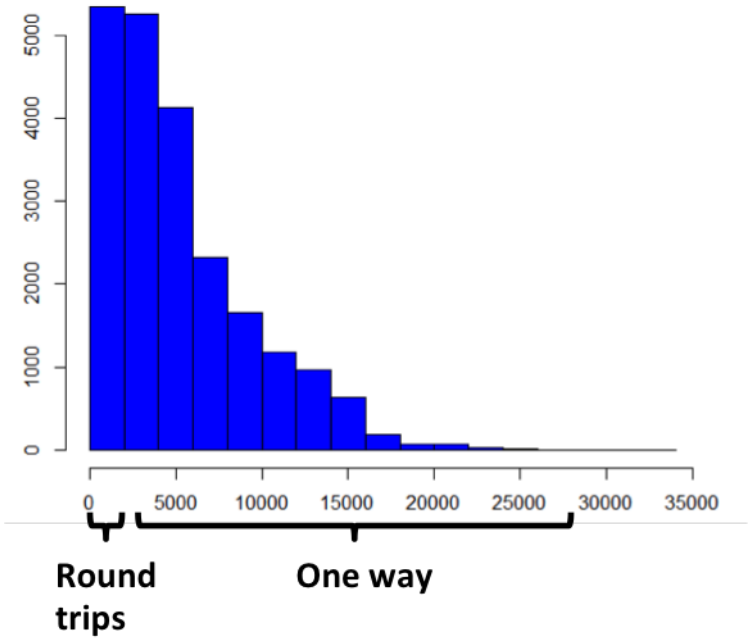


Figure 1.1: Histogram of Origin-Destination Distances (in meters)

and charging outlets are made in conjunction with service region design, a robust planning methodology is imperative. Secondly, the operation details of EV car sharing such as repositioning of EVs are dependent on not only the size but also the shape of service region. Hence, the service provider must also conscientiously account for operational cost drivers when determining service region in the face of limited data. In this work, we make the following contributions to the literature.

- We formulate an integrated service region planning model taking into account customers' satisficing behavior in service adoption, together with various operational characteristics of one-way EV sharing system. Our approach deliberately addresses data uncertainty and ambiguity with regard to customers' travel patterns. Using a distributionally-robust optimization framework, our model can be approximated by a computationally-efficient mixed integer second-order cone program (MISOCP).

- Using real operations data from Car2Go, travel characteristic data from the California Household Travel Survey and EV charging station deployment data from the U.S. Department of Energy, we perform a case study of Car2Go's service region design in San Diego. We address several planning questions and obtain the following findings:

    1. EV sharing systems bring more environmental benefits, e.g., savings in CO_2 emissions, than replacing personal gasoline cars with EV ownership.

2. Smaller regional variations in demographics, e.g., population and income levels, suggest more spread-out service region.
3. Charging technology advances that improve charging speed will lead to fleet size reduction and service region expansion, while the marginal impacts on service region coverage is diminishing.

Literature Review

Our work contributes to the expanding research on sustainable operations management that covers a wide range of topics [55, 73]. There are two streams of literature in sustainable operations related to ours: EV business models and vehicle sharing operations. Lim et al. (2014) [61] aim to evaluate performances of business practices toward the goal of mass adoption and study the impact of range and resale anxieties. Similarly, Avci et al. (2014) [8] highlight the key mechanisms driving adoption and use of EVs in a battery swapping system. Particularly, they build a behavioral model of motorist use and adoption. Calibrating with real data, they find that such system may not be beneficial to the environment. Besides the insights from business model analysis, infrastructure planning and charging coordination issues are also studied. Mak et al. (2013) [65] develop distributionally-robust optimization models that helps the planning process for deploying battery swapping infrastructure. Moreover, a couple of papers in transportation optimize the operations of charging station networks and coordinate the recharging scheme through area pricing or routing [41, 81]. Under the EV sharing system setting, we consider the EV charging operations together with customer adoption of the service instead of EV ownership.

While there are several major hurdles to achieve mass EV adoption, EV sharing is an alternative for customers to enjoy the benefits of EVs without ownership. By EV sharing, the high fixed costs of EV ownership is transformed to a usage-based cost of service. Researchers have used the terminology *servicizing* to describe a business model that offers the functionality of the products instead of selling the product itself. Agrawal and Bellos (2013) [3] assess the potential of servicizing business models as an environmentally sustainable strategy and draw insights into when and how servicizing is environmentally beneficial. Related to this study, Bellos et al. (2013) [9] determine the OEM's optimal pricing strategy and the optimal fleet size when it offers car sharing in conjunction with conventional sales. Their analysis reveals the discrepancy between profitability and environmental sustainability. Since the car sharing system in their model only allows round trips, they are able to focus the fleet operations of each station individually as a single server; whereas in our model, the EV sharing system is designed to support one-way trips and fleet repositioning is necessary in the presence of imbalanced trip flows.

The work closest to ours is Shu et al. (2013) [85] which consider the detailed bicycle sharing operations in a network context. They develop a network flow model with propor-

tionality constraints to estimate the flow of bicycles within the network and the number of trips supported. Using transit data from the train operator in Singapore, they examine the bicycle deployment, utilization and the value of bicycle redistribution. Due to the short range of bicycles, they restrict the trips within two transit stops and assume that bicycles are immediately available for next customers upon arrivals. However, in EV sharing systems, there is risk that the arriving EVs are at low battery level and need to be placed out for recharging. Our work considers redistribution of vehicles by modeling the repositioning (a.k.a. rebalancing) of fleet as a stochastic process while in Shu et al. (2013) [85] the system restores the bicycle distribution among all locations on a regular basis. Furthermore, their model assumes that the bicycle sharing station locations are given and demands follow known Poisson processes. In our problem, we aim to design a service region under incomplete information about the demands.

Since the service providers need to determine the service region before the system is in operation and customers join the membership, demand uncertainty becomes a big concern. It is therefore critical to make a robust service region design under various scenarios. The literature on robust optimization [10, 11, 12, 13] provides approaches to inform solutions that are robust with respect to perturbations in the model parameters. For problems where some limited distributional information, for example the mean and covariance of key parameters, may be available at the planning stage, it is possible to utilize the distributionally-robust optimization approaches discussed in Ghaoui et al. (2003) [39], Chen et al. (2007, 2010) [31, 30], Goh and Sim (2010) [44] and Natarajan et al. (2011) [70]. An advantage of this methodology is that it is often possible to preserve computational tractability using conic programming formulations. A recent application in EV infrastructure planning can be found in Mak et al. (2013) [65]. With some limited information, such as the moments of demand parameters, they develop distributionally-robust models for EV battery swapping station deployment. Their formulations are tightly approximated by mixed integer second-order cone programs (MISOCPs) which are readily solvable by commercial solvers. Several other applications include appointment scheduling in healthcare [57, 66], warehouse operations [6], supply chain management [67], inventory control [83] and portfolio management [69].

1.2 The Model

We consider an urban EV sharing service provider, e.g., Car2Go, that designs its service region in a metropolitan area, e.g., San Diego. An overview of the current service region of Car2Go San Diego in Figure 1.2 shows that it consists of the downtown San Diego, Chula Vista as well as San Diego State University (SDSU). An distinctive feature that differentiates Car2Go from other car sharing systems is that it allows one-way trips and offers free street parking. Customers can start trips anywhere inside the service region wherever there's a car available, and end trips wherever there is qualified parking space available. Customers can visit outside the service region during reservations but they are required to bring the car

back to the service region to end the trips [22].

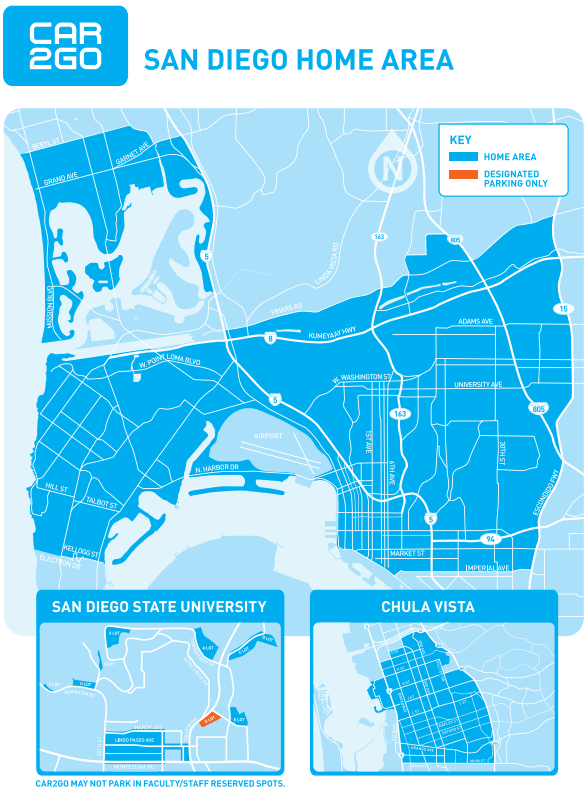


Figure 1.2: Current Service Region of Car2Go San Diego [24]

Due to the one-way nature of the service, a well-planned service region balances the goals of inducing more adoptions and maintaining cost-effective fleet operations. From the customers' perspective, it is more favorable to adopt and use the EV sharing service if the service region covers more of their preferred destinations. Nevertheless, a larger service region may result in more complex operations and thus higher operational costs to the service provider. Hence, it is crucial to model the interrelationships between customer adoption, fleet operations, and service region design. However, it is difficult in practice to obtain accurate estimations of individual valuations on coverage of destinations. In the model, we try to depict the aggregate customer adoption levels of the EV sharing service and propose an optimization model that strategically supports the service region design under uncertainty of customer travel patterns and preferences.

We consider the following *satisficing* model of service adoption. Each customer has a set of utilities of being able to travel to the set of destinations enabled by the service. For

instance, traveling to destination j brings a utility a_{ij} to a customer in region i. The values of a_{ij} are heterogeneous among customers. Hence, at an aggregate level, they are random variables from the service provider's point of view. Under satisficing behavior in service adoption [86], a customer adopts the service when the total utility from all served destinations exceeds his or her aspirational level. We consider customers to be categorized into K groups, indexed by $k = 1, \cdots, K$, by their aspirational levels b_k.

The firm plans the service region by selectng among pre-defined candidate regions. Such decisions can be modeled by binary decision variables x_j with value 1 denoting the covering of candidate region j. Given the service region design, the total utility provided to a customer in region i is $\sum_{j\in I} a_{ij}x_j$. The adoption decision for a customer is then expressed by the following indicator function, with a value of 1 representing the adoption of the service:

$$\mathbf{1}\big(\sum_{j\in I} a_{ij}x_j \geq b_k\big) = \begin{cases} 1, & \text{if } \sum_{j\in I} a_{ij}x_j \geq b_k \\ 0, & \text{otherwise.} \end{cases}$$

By taking expectation over the indicator function, we have the adoption rate q_{ik} of customer group k in region i:

$$\begin{aligned} q_{ik} &= \mathbb{E}\big[\mathbf{1}\big(\sum_{j\in I} a_{ij}x_j \geq b_k\big)\big] \\ &= \text{Prob}\big(\sum_{j\in I} a_{ij}x_j \geq b_k\big). \end{aligned}$$

The firm earns profit from two parts: membership revenue and operational profit. Each customer who sign up for the service has to pay a fixed membership fee f and is charged at c per minute of usage. There is also a fixed cost g_i of covering region i, which may include investments in charging infrastructure or payments to charging service providers, and payments to city governments for street parking. Our model maximizes the expected total profit in Equation (1.1), which is defined as total revenue less fixed coverage cost, operational costs such as charging cost, repositioning cost and fleet investment. For notational brevity, the operational profit is represented by a function $\Theta(x_i, q_{ik}, \alpha)$ with service level guaranteed to be α (i.e., customers will find available EVs with at least α probability). We will provide an explicit formulation for $\Theta(\cdot)$ in Section 1.2. The service region design model is formulated as follows:

$$\max_{q_{ik}, x_i} \sum_{i\in I}\sum_{k\in K} fQ_{ik}q_{ik} - \sum_{i\in I} g_i x_i + \Theta(x_i, q_{ik}, \alpha) \tag{1.1}$$

s.t.

$$q_{ik} \leq \text{Prob}\big(\sum_{j\in I} a_{ij}x_j \geq b_k\big), \forall i \in I, \forall k \in K \tag{1.2}$$

$$q_{ik} \leq x_i, \forall i \in I, \forall k \in K \tag{1.3}$$

$$x_i \in \{0, 1\}, \forall i \in I.$$

The membership revenue $\sum_{i\in I}\sum_{k\in K} fQ_{ik}q_{ik}$ in objective function (1.1) is computed from the membership fee and total customer adoption, where Q_{ik} is the size of customer group k in region i. Constraint (1.2) represents the adoption rate in probability constraints. Furthermore, constraint (1.3) states the fact that no customers will adopt the service if their origins are not in the service region. Appendix A.1 summarizes the notation used throughout the paper.

Adoption Rate Model

In this section, we focus on dealing with the probability constraint (1.2). In order to evaluate the exact adoption rate the firm needs the complete information on the joint distribution of a_{ij}. However, in reality, perfect information is often unavailable to the firm. Specifically, as the firm is in the planning stage with limited operations data (e.g., from pilot studies or household travel surveys), it is often difficult to fit the joint distribution of travel patterns with confidence. Furthermore, from the tractability standpoint, the term $\sum_{j\in I} a_{ij}x_j$ may still be hard to evaluate even for known distributions of a_{ij}, especially with correlations among a_{ij}'s. To this end, it is practical to consider a model that possesses both distributional robustness and computational tractability under limited information.

In particular, we relax the data requirement by assuming knowledge of only descriptive statistics of a_{ij}, i.e., their means and covariance matrix. We construct a robust model that delivers the worst-case adoption rate, i.e., the lowest adoption rate among all possible distributions $\mathcal{P}$ of the utility parameters a_{ij}'s that satisfy the known mean and covariance matrices:

$$q_{ik} \le \inf_{p\in\mathcal{P}} \text{Prob}\big(\sum_{j\in I} a_{ij}x_j \ge b_k\big). \tag{1.4}$$

The utility parameter a_{ij} is nonnegative by nature. Suppose the mean vector $\bar{\mathbf{a}}_i = [\bar{a}_{ij}]$ and covariance matrix $\Gamma_i = [cov(a_{ij_1}, a_{ij_2})]$ for each region i are known. We further assume the covariance matrix is positive definite: $\Gamma_i \succ 0$. The second moment matrix Σ_i is given by:

$$\Sigma_i := \mathbb{E}\begin{bmatrix}\mathbf{a}_i\\1\end{bmatrix}\begin{bmatrix}\mathbf{a}_i\\1\end{bmatrix}^T = \begin{bmatrix}S_i & \bar{\mathbf{a}}_i\\ \bar{\mathbf{a}}_i^T & 1\end{bmatrix}, \text{ where } S_i := \Gamma_i + \bar{\mathbf{a}}_i\bar{\mathbf{a}}_i^T.$$

Since $\Gamma_i \succ 0$, the covariance matrix Σ_i is also positive definite. With fixed x_i's, the worst-case adoption rate can be obtained by solving the convex optimization formulation with copositive constraints, as shown in Lemma 1.

Lemma 1. *In problem (1.1), given the mean vector $\bar{\mathbf{a}}_i$ and the covariance matrix Γ_i for each region $i \in I$, worst-case probability constraint (1.4) for each q_{ik} is equivalent to the following*

formulation with copositive constraints with known values of x_i's.

$$
\begin{aligned}
&\langle M_{ik}, \Sigma_i \rangle \leq 1 - q_{ik} \\
&M_{ik} \succeq_{co} 0 \\
&M_{ik} + \begin{bmatrix} 0 & \mathbf{d_{ik}} \\ \mathbf{d_{ik}}^T & -1 - 2\tau_{ik} b_k \end{bmatrix} \succeq_{co} 0 \\
&-\rho \mathbf{x} \leq \mathbf{d_{ik}} \\
&\mathbf{d_{ik}} \leq \rho \mathbf{x} \\
&\tau_{ik} \mathbf{e} + \rho(\mathbf{x} - e) \leq \mathbf{d_{ik}} \\
&\mathbf{d_{ik}} \leq \tau_{ik} \mathbf{e} + \rho(\mathbf{e} - \mathbf{x}) \\
&\tau_{ik} \geq 0
\end{aligned}
$$

where $\mathbf{x} = (x_j)$, a large scalar ρ, Σ_i the second moment matrix and the vector of ones $\mathbf{e}$ are known parameters; the symmetric matrix M_{ik}, τ_{ik}, $\mathbf{d_{ik}}$ are decision variables; and inner product $\langle A, B \rangle = trace(BA)$.

Proof. Proof of Lemma 1. Please see Appendix B.1. □

A matrix A is called a copositive matrix ($A \succeq_{co} 0$) if it satisfies $\mathbf{v}^T A \mathbf{v} \geq 0, \forall \mathbf{v} \in \mathbb{R}^n_+$. For more details on copositive matrices, please refer to Burer (2009) [18].

The formulation in Lemma 1 is not readily solvable by commercial solvers, due to the combination of copositive constraints and mixed integer decision variables. A natural approach to deal with copositive constraints is to approximate them by tractable convex relaxations, e.g., a series of linear and semidefinite constraints that can be further transformed into second-order cone constraints. We provide a lower bound formulation that is computationally tractable in Proposition 1.

Proposition 1. *The following formulation with second-order cone constraints provides a lower bound on the worst-case adoption rate q_{ik} in Lemma 1.*

$$
\begin{aligned}
&4X^T \Gamma_i X + (1 - q_{ik} - \sum_{(j_1, j_2) \in I \times I} \bar{a}_{ij_1} \bar{a}_{ij_2} z_{j_1 j_2} + 2b_k \sum_{j \in I} \bar{a}_{ij} x_j - b_k^2 - \sum_{(j_1, j_2) \in I \times I} \sigma_{ij_1 j_2} z_{j_1 j_2})^2 \\
&\leq (1 - q_{ik} + \sum_{(j_1, j_2) \in I \times I} \bar{a}_{ij_1} \bar{a}_{ij_2} z_{j_1 j_2} - 2b_k \sum_{j \in I} \bar{a}_{ij} x_j + b_k^2 + \sum_{(j_1, j_2) \in I \times I} \sigma_{ij_1 j_2} z_{j_1 j_2})^2 \qquad (1.5) \\
&z_{j_1 j_2} \in \mathcal{Z}(x_{j_1}, x_{j_2}), \forall j_1, j_2 \in I \qquad (1.6) \\
&(x_i, q_{ik}) \in \mathcal{X}_{ik} \qquad (1.7) \\
&q_{ik} \leq x_i
\end{aligned}
$$

where $\mathcal{Z}(x_{j_1}, x_{j_2})$ and $\mathcal{X}_{ik}$ are sets of linear constraints defined in Appendix B.2 Equations (B.4) and (B.5).

Proof. Proof of Proposition 1. Please see proof in Appendix B.2. □

Although there are K customer groups, the average adoption rate in each region can be aggregated by population weighted sum of adoption rates q_{ik}. Let the weight of group k be $\psi'_{ik} = \frac{Q_{ik}}{\sum_{k \in K} Q_{ik}}$. *The population aggregated adoption rate* is then

$$q'_i = \sum_{k \in K} \psi'_{ik} q_{ik}, \forall i \in I.$$

Similarly, let ψ_{ik} be the proportion of outbound trips by group k among total outbound trips from region i. *The outbound trip weighted adoption rate* among total outbound trips is given by

$$q_i = \sum_{k \in K} \psi_{ik} q_{ik}, \forall i \in I.$$

Operational Profit Model

In the model (1.1), the operational profit is represented by a function $\Theta(x_i, q_{ik}, \alpha)$ with guaranteed EV availability (service level) α for each covered region. The EV availability is defined as the probability for customers to find EVs available at their origins. To improve the EV availability, repositioning is essential to even though it might be costly. In fact, Car2Go employs "street teams" to redistribute vehicles aiming to ensure even availability through the service region [20].

We model fleet operations in the EV sharing system as a closed queueing network where EVs go through queues (nodes in the network) that represent stochastic waiting times for customers and lead times for repositioning and recharging. For instance, consider an EV sharing system that serves only two regions: indexed by 1 and 2, which can be modeled as a closed queueing network in Figure 1.3. Nodes 1 and 2 represent EVs staying in the two regions, available (and waiting) for customer orders. Flows entering dummy nodes 1^r and 2^r are the EVs to be repositioned (incurring a stochastic delay) upon arrival at regions 1 and 2, respectively. Similarly, flows entering dummy nodes 1^c and 2^c are the EVs to be directed to charging stations (incurring a stochastic down time) upon arrival at regions 1 and 2 respectively.

We consider the arrivals of EVs in region i that are able to serve next customers (i.e., with sufficient battery levels and are not repositioned) as demand arrivals at queue i with effective rate Λ_i. The regions are acting as servers with service rates equal the customer demand rates; that is, when a customer arrives, the first EV in the queue finishes its delay at the server and departs the queue. Once the battery level of an EV falls below a prespecified level, e.g., 20% in Car2Go, the EV will need to be recharged. Consequently, an arriving EV has P_c probability to be re-directed to a charging facility and placed out of service until the car is

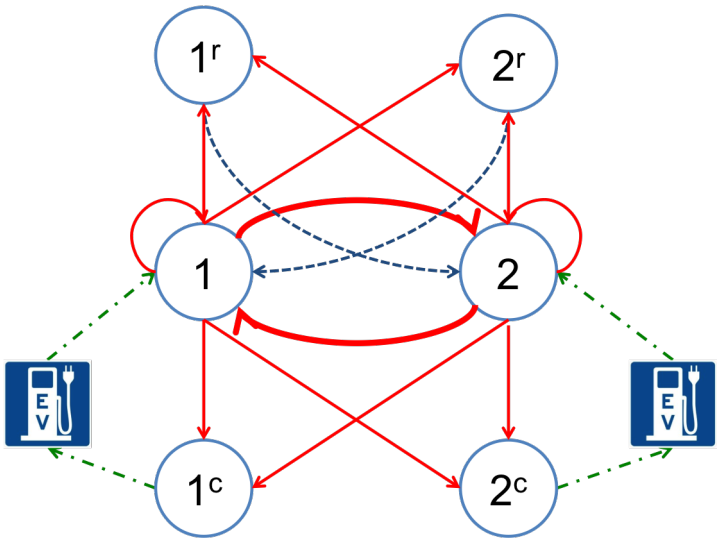


Figure 1.3: EV Sharing Operations as Closed Queueing Network

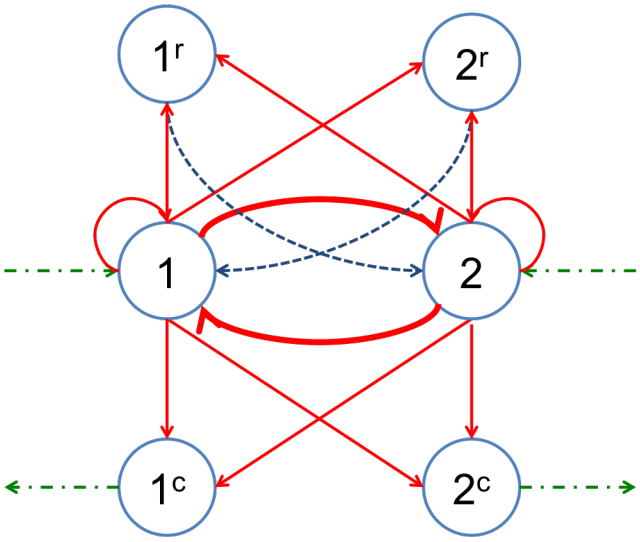


Figure 1.4: EV Sharing Operations as Open Queueing Network

fully charged. Given the service region design, a customer trip from i has $\hat{P}_{ij} = \frac{P_{ij}x_j}{\sum_{k\in I} P_{ik}x_k}$ probability of heading for destination j, where P_{ij} is the probability of an i-originated trip ending in j when all destinations are served. To ensure long-run availability of cars, balance of inflow and outflow rates of a region must be maintained by repositioning EVs as necessary. The repositioning policy is defined by γ_{jl} which is the probability of an arriving EV at j is repositioned to region l upon arrival. As the rate of customers driving from i to j is $\Lambda_i\hat{P}_{ij} = \Lambda_{ij}$, the rate of redirecting arriving EVs at j to region l that are originating from

i is then $\phi_{ijl} = \Lambda_{ij}\gamma_{jl}$. The external EV inflows to the system are denoted by λ_i for region i. The effective EV arrival rate Λ_i that are ready to serve next customers in steady state is described by the flow balance equations:

$$\Lambda_i = \lambda_i + \sum_{j\in I} \Lambda_{ji}(1 - P_c - \sum_{l\in I} \gamma_{il}) + \sum_{j\in I}\sum_{m\in I} \Lambda_{jm}\gamma_{mi}, \forall i \in I$$

$$\Leftrightarrow$$

$$\sum_{j\in I} \Lambda_{ij} = \lambda_i + \sum_{j\in I} \Lambda_{ji}(1 - P_c) - \sum_{j\in I}\sum_{l\in I} \phi_{jil} + \sum_{j\in I}\sum_{m\in I} \phi_{jmi}, \forall i \in I. \tag{1.8}$$

An EV has four possible states at any time: idle on street (i.e., available for customers), serving a customer, being recharged, and being repositioned. Hence, the inequality $1 - P_c - \sum_{l\in I} \gamma_{jl} \geq 0$ must hold. By multiplying both sides with Λ_{ji}, it is equivalent to

$$\sum_{l\in I} \phi_{jil} \leq \Lambda_{ji}(1 - P_c).$$

For the trips to destinations j and k from origin i, the trip distribution follows the rationing based on the travel pattern which is described in the constraint below.

$$\frac{\Lambda_{ij}x_k}{P_{ij}} = \frac{\Lambda_{ik}x_j}{P_{ik}}, \forall i \in I, j \in I, k \in I.$$

After recharge, the fully recharged EVs replenish the fleet. Over time, the inflow rates of recharged EVs equal the outflow rates of EVs to the charging stations. A conservative charging policy is to recharge P_c proportion of the maximum EV arrivals at all regions:

$$\lambda_i = \sum_{j\in J} \mu_j q_{ji} P_c \tag{1.9}$$

where $q_{ji} = q_j P_{ji} x_i$ is the adoption rate for trips from j to i.

To guarantee the EV availability, fleet size is an important consideration that needs to be determined at the cost of fleet investment. Generally, the larger service region requires more EVs. Using the fixed population mean (FPM) approximation introduced in Whitt (1984, 2002) [96, 97], we derive the population in the closed queueing network from the associated open queueing network. The key idea is to approximate the steady-state performance of a closed queueing network by the steady-state performance of an associated open queueing network in which the mean population is set to the specified population (which is fixed) in the closed network. In the case of the EV sharing system, the population in the closed queueing network is the fleet size. We approximate the closed queueing network with an associated open queueing network and each region works as a M/M/1 queue (i.e., assuming EVs being checked out by customers following first-come, first-served order). For instance,

the associated open queueing network in Figure 1.4 considers the flows to charging stations as departures of the system and the flows from charging stations as the external inflows. We obtain the fleet size by deriving the expected number of EVs with all possible status that can be calculated from Little's Law. There are $t_{ij}\Lambda_{ij}$ EVs en route from i to j with average travel time t_{ij}. Suppose the average charging time is t_c, then $\lambda_i t_c$ EVs are in charging facilities in region i. Lastly, let τ_{mj} be the travel time of repositioning trips from m to j, it is generally no larger than the travel time by customers t_{ij}, since the street team repositions EVs without intermediate stops. Therefore, coming from i to m, $\tau_{mj}\phi_{imj}$ EVs are in repositioning trip to region j upon arrival at m. With EV availability α, the expected number of available EVs awaiting for customers in region i is $L_i = \frac{\alpha}{1-\alpha}$ based on the M/M/1 queue assumption. The desired fleet size N must be no less than the expected number of EVs in the steady state:

$$\sum_{j\in I}\sum_{i\in I} t_{ij}\Lambda_{ij} + \sum_{i\in I} L_i x_i + \sum_{i\in I}\lambda_i t_c + \sum_{i\in I}\sum_{m\in I}\sum_{j\in I}\tau_{mj}\phi_{imj} \leq N. \tag{1.10}$$

We are now able to characterize the operational profit $\Theta(x_i, q_{ik}, \alpha)$ that consists of four parts: operational revenue, charging cost, repositioning cost as well as fleet investment. The major car sharing systems charge customers r per unit time of usage, e.g., per minute for Car2Go. The operational revenue is the total revenue from EV usage $\sum_{j\in I}\sum_{i\in I} r t_{ij}\Lambda_{ij}$ gained from all OD pairs. Similarly, suppose the charging cost is c per unit time and the charging time is t_c, the firm pays total charging cost $\sum_{i\in I} c\lambda_i t_c$. Moreover, it takes the street team τ_{jm} time units to reposition an EV from j to m with cost η per unit time. The corresponding total repositioning cost is then $\sum_{i\in I}\sum_{j\in I}\sum_{m\in I}\eta\tau_{jm}\phi_{ijm}$. Lastly, the annually amortized EV purchase cost is calculated as h, based on the price and typical life span in EV sharing fleet. Therefore, the operational profit is explicitly formulated as:

$$\Theta(x_i, q_{ik}, \alpha) = \sum_{j\in I}\sum_{i\in I} r t_{ij}\Lambda_{ij} - \sum_{i\in I} c t_c\lambda_i - \sum_{i\in I}\sum_{j\in I}\sum_{m\in I}\eta\tau_{jm}\phi_{ijm} - hN.$$

Combining the adoption rate and operational profit models with the model, the service region design problem is formulated as a mixed integer second-order cone program (MIS-

OCP):

$$\max_{q_{ik},x_i,N} \sum_{i\in I} fQ_i q_i' - \sum_{i\in I} g_i x_i + \sum_{j\in I}\sum_{i\in I} rt_{ij}\Lambda_{ij} - \sum_{i\in I} ct_c\lambda_i - \sum_{i\in I}\sum_{j\in I}\sum_{m\in I} \eta\tau_{jm}\phi_{ijm} - hN \quad (1.11)$$

s.t.

$$q_{ik} \le \text{Prob}\big(\sum_{j\in I} a_{ij}x_j \ge b_k\big), \forall i \in I, \forall k \in K \quad (1.12)$$

$$\sum_{j\in I} \Lambda_{ij} \ge \alpha\mu_i \sum_{j\in I} q_{ij}, \forall i \in I \quad (1.13)$$

$$\Lambda_{ij} \le \mu_i q_{ij}, \forall i \in I \quad (1.14)$$

$$q_i' = \sum_{k\in K} \psi_{ik}' q_{ik}, \forall i \in I$$

$$q_i = \sum_{k\in K} \psi_{ik} q_{ik}, \forall i \in I$$

$$q_i \le x_i, \forall i \in I$$

$$q_{ij} = q_i P_{ij} x_j, \forall i \in I, j \in J$$

$$\sum_{j\in I} \Lambda_{ij} = \lambda_i + \sum_{j\in I} \Lambda_{ji}(1-P_c) - \sum_{j\in I}\sum_{l\in I} \phi_{jil} + \sum_{j\in I}\sum_{m\in I} \phi_{jmi}, \forall i \in I$$

$$\sum_{l\in I} \phi_{jil} \le \Lambda_{ji}(1-P_c), \forall i \in I, j \in I$$

$$\lambda_i = \sum_{j\in J} \mu_j q_{ji} P_c, \forall i \in I$$

$$\sum_{j\in I}\sum_{i\in I} t_{ij}\Lambda_{ij} + \sum_{i\in I} L_i x_i + \sum_{i\in I} \lambda_i t_c + \sum_{i\in I}\sum_{m\in I}\sum_{j\in J} \tau_{mj}\phi_{imj} \le N$$

$$\frac{\Lambda_{ij}x_k}{P_{ij}} = \frac{\Lambda_{ik}x_j}{P_{ik}}, \forall i \in I, j \in I, k \in I$$

$$\Lambda_{ij} \ge 0, \forall i \in I, j \in I$$

$$\phi_{ikj} \ge 0, \forall i \in I, k \in I, j \in I$$

$$x_i \in \{0,1\}, \forall i \in I.$$

The objective function is the annual total profit including the membership revenue and operational profit. In computation, the worst-case probability constraint (1.12) is replaced by constraints (1.5) and (1.7) in Proposition 1. The service level constraint (1.13) guarantees at least α service level, while constraint (1.14) ensures the stationary condition of the queueing network by limiting the EV arrival rates not exceeding the customer request rates. The rest are either explained before or are nonnegativity and integrality constraints.

In practice, the customer travel patterns, including both the trip distribution P_{ij} and outbound trip rates μ_i, are time-varying. The average system performance can be captured

by the pointwise stationary approximation [46] approach as if the travel patterns are stationary at that point in time. We partition the 24 hours into T periods of a day with stationary travel patterns, by considering, e.g., P_{ij}^t and μ_i^t. The following proposition summarizes the formulation in the presence of time-varying travel pattern.

Proposition 2. *Given time-varying travel patterns, e.g. P_{ij}^t and μ_i^t, the service region design problem can be formulated as a mixed integer second-order cone program with constraints for multiple periods.*

Proof. Proof of Proposition 2. Please see proof in Appendix B.3. □

One further point that warrants some discussion is the integration of the adoption rate and operational profit submodels. Following Proposition 1, the worst-case adoption rate can be represented by (one minus) the q_{ik} variables, subject to a set of linear and second order conic constraints (1.5), in the absence of other constraints. However, when the constraints characterizing the queueing network dynamics, which involve the q_{ik} variables, are added, there is no guarantee that (1.5) is tight at the optimal solution. In light of this, we provide Proposition 3 that suggests a sufficient condition for the constraints to be tight.

Proposition 3. *Given the service region design decisions x_i's, the probability constraint (1.12) is tight for candidate region $i \in I$ if the following sufficient condition holds:*

$$r\sum_{j\in I} t_{ij}\hat{P}_{ij} - ct_cP_c - \eta\sum_{j\in I}\sum_{m\in I}\tau_{jm}\hat{P}_{ij}\gamma_{jm} - h(\sum_{j\in I} t_{ij}\hat{P}_{ij} + t_cP_c + \sum_{j\in I}\sum_{m\in I}\tau_{jm}\hat{P}_{ij}\gamma_{jm}) \geq 0.$$

That is, the marginal operational profit, which is marginal revenue less the marginal increase in charging cost, repositioning cost and fleet size investment, of outbound trips from all regions $i \in I$ are nonnegative.

Proof. Proof of Proposition 3. Please see proof in Appendix B.4. □

Therefore, as long as all the candidate regions brings non-negative marginal operational profits, tightness of constraint (1.5) is guaranteed.

1.3 Case Study: Car2Go in San Diego

We demonstrate the service region design framework with a case study of Car2Go in San Diego, the first city in North America where Car2Go operates all-EV fleet under the free floating model discussed before. We begin the case study with description of data used and estimation of key parameters that depict customer adoption behaviors and travel patterns.

Parameter Estimation

We conduct the computational experiments with cost parameters collected from Car2Go website and amortized to annual costs based on a 5-year planning horizon. The firm earns annual membership fee at $f = \$8$ and usage rate at $r = \$0.16$/min adjusted with variable costs. Based on the technical specifications of Smart Electric Drive, the EV model in Car2Go fleet, the cost to fully charge from 20% battery level is determined to be $c = \$0.5$/hr with charging time of 6 hours. Since Car2Go's policy requires the EVs to be recharged when the battery is below 20%, the probability that an arriving EV needs charging is set to $P_c = 0.2$. In the case of imbalanced flows, the street team has to reposition the EVs at the cost \$0.16/min. The total repositioning cost depends on the repositioning frequency and distance completed. Moreover, in our experiments, $\alpha = 80\%$ EV availability is guaranteed in each selected candidate regions. In the estimation of parameters and computational experiments, we use the following data sets.

1. *Car2Go San Diego operations data.* This data set contains one-month time stamp record of all idle EVs in the current EV sharing system of Car2Go San Diego at every 5-minute level. The record includes time, location, battery levels and charging status. Through preprocessing of the data, we identify 25,875 trips in total with the current fleet size of 379 EVs.

2. *San Diego geographic information and census data.* The travel distances and times between all OD pairs are provided by ArcGIS, a geographic information system, with road network map from SanGIS data warehouse [79]. The census data is from 2010 American Community Survey [1] with zip code level working population as well as per capita income.

3. *2010 California Household Travel Survey (CHTS).* The CHTS collects travel information from households in all of California's 58 counties [19]. All participating households were first recruited to record their travel in a diary for a pre-assigned 24-hour period. For our purposes, we focus on the households in San Diego county. We use the tables such as household, persons and places with interested attributes including age, income level, zip codes and modes of trips. In the sample of 1999 individuals at working age in San Diego county, we identify 6,562 trips out of which 5,335 trips were by car.

4. *EV charging station information.* We use the EV charging station data from [5] with attributes such as location, zip code, charger number and EV network. We focus on the charging stations in the current service region under EV charging network called Blink, the partner of Car2Go.

We observe time-varying travel patterns, e.g. total outbound trip rates, from the trips summarized from the operations data shown in Figure 1.5. We partition the 24 hours of a day into 2 periods: daytime from 7AM to 21PM and night from 21PM to 7AM, which minimizes the sum of squared errors of the outbound trip rates for all candidate regions. For

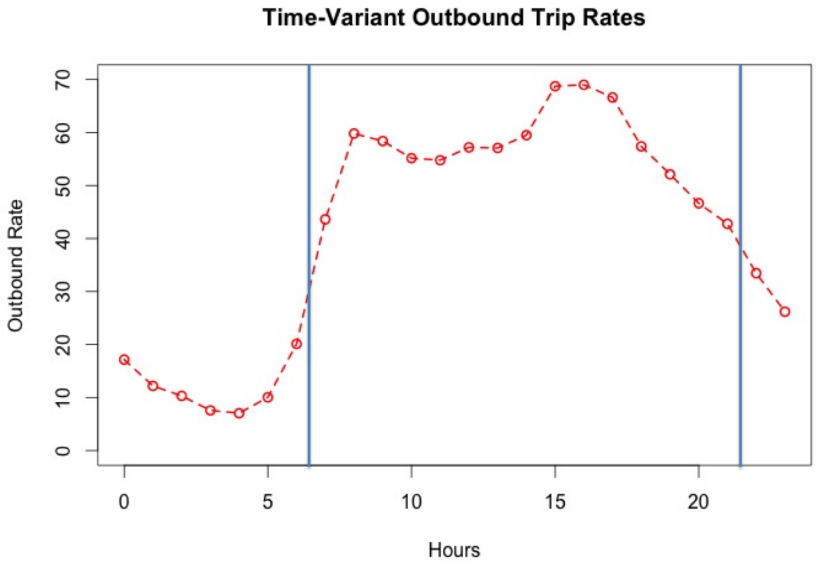


Figure 1.5: Partition of Time-varying Travel Patterns

each period, gravity models for trip distributions in transportation literature are applied to estimate the travel pattern and utilities of the potential customers, based on the sample trips from Car2Go operations data, census data and geographic information. The utility threshold to adopt is analyzed from the CHTS data with the trip modes of sampled population. From the current EV charging station information, we estimate the desired number of chargers for each candidate region and thus infer the corresponding region coverage cost. For details, please refer to Appendix C.

Optimal Service Region and Fleet Size

We solve the MISOCP in Equation (1.11) using CPLEX solver on Intel Core i5-3550 CPU at 3.30 GHz to obtain the optimal service region and fleet size for San Diego county with 61 candidate regions at zip code level. The results are shown in Figure 1.7 in comparison with the current Car2Go service region shown in Figure 1.6 that covers 32.57% working population. Both solutions agree to cover downtown San Diego. Although the region is geographically small, it is densely populated as central business and university districts. Car2Go data show that 49.88% of the trip observations happened within that region. The major discrepancy is to choose the north or south county. The current service region contains Chula Vista in the south. However, based on the Car2Go data, it is related to only 1.12% of the trip observations. The proposed solution tend to cover the north county, which is the second most populous region in the county and is well known for its tourism [71]. Based on the gravity models, the north county generates more trips because of higher population and income. In

practice, the firm needs to negotiate the free parking agreement with city governments for specific zones. The proposed service region may be grouped and rearranged into cluster of cities and exclude improper areas, e.g., mountains and forest. With additional clustering constraints for some adjacent regions that must be covered together, we result the optimal solution in Figure 1.8.

Furthermore, the optimal service region in Figure 1.7 suggests an expansion to 33 zip codes with 485 cars needed. Compared to the current fleet size, by adding only 28% more EVs, Car2Go can serve 62.60% more population. Finally, both the proposed service regions in Figure 1.7 and Figure 1.8 suggest future expansion opportunities to the north.

The Environmental Benefits

Based on the optimal solution in Figure 1.7.b, we conduct the analysis by estimating the savings of GHG emissions from operating an EV sharing system. For consistency with the related research, we choose to focus on CO_2 emissions as the measure of environmental impacts.

Observation 1. *Supporting customers' travel needs with zero emission, deploying EV sharing service with 485 EVs gains similar CO_2 emission savings from replacing 2312 gasoline cars with EV ownership. That is, each EV in the sharing fleet, on average, brings similar environmental benefits as converting 4.77 individually-owned gasoline cars into EVs.*

The EV sharing system supports 1,340,875 trips annually with total mileage of 26,147,915 miles. The U.S. Environmental Protection Agency determines an annual CO_2 emissions per mile to be 4.20×10^{-4} metric tons. Hence, the annual CO_2 emission savings from Car2Go is

Figure 1.6: Current Service Region

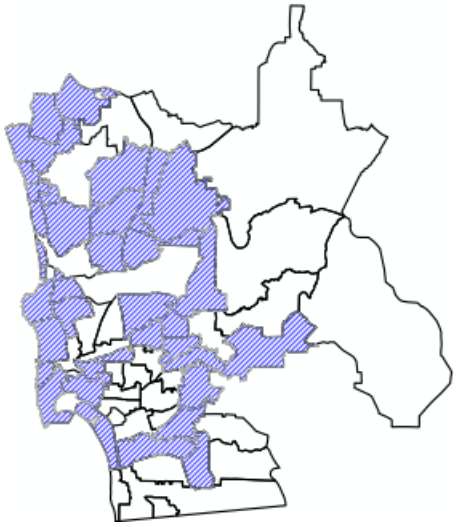


Figure 1.7: Optimal Service Region

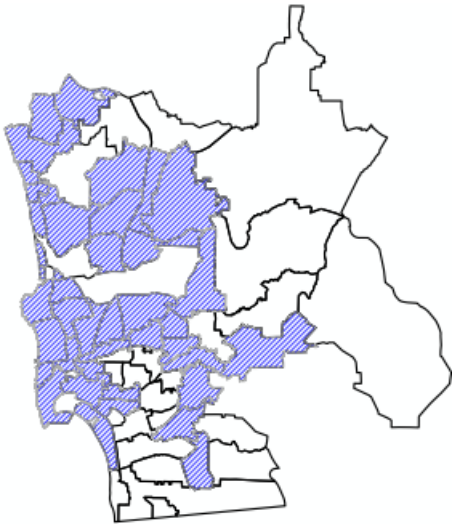


Figure 1.8: Optimal Service Region with Clustering

calculated as 10,982.12 metric tons. To visualize the savings, we find the number of gasoline cars that would cause equivalent CO_2 emissions. Given the annual passenger vehicle CO_2 emissions as 4.75 metric tons, the savings from Car2Go fleet of 485 EVs is similar to the savings from replacing 2312 individually-owned gasoline cars with EVs. The advantage of EV sharing over individual EV ownership mainly comes from the higher vehicle utilization in sharing fleet. In fact, the adoption of EV sharing is generally easier than mass adoption of individual EV ownerships. As a result, EV sharing systems will realize GHG emission savings earlier and create more cumulative environmental benefits through early adoption.

Impacts of Demographic Factors

We now examine how demographic changes provide may affect the optimal service region coverage. We aim to provide some insights to cities with different demographic configurations, e.g., New York city versus San Diego. Even for the same city, the demographic factors change over time due to factors such as migration and economic development. In the experiments, we consider two alternative scenarios: 1. the population is more evenly distributed among the regions; 2. the income disparity between regions is reduced. For both scenarios, we first generate new population and income levels as follows. Let pop_i be the population of region i, and $\bar{\text{pop}}$ be the mean over all regions. Then, we let the new population of region i be $\bar{\text{pop}} + 0.25(\text{pop}_i - \bar{\text{pop}})$. That is, we keep the average population unchanged and shrink the standard deviation by 75%. The new income level scenario is generated similarly. We then simulate the trip distributions through the same gravity models used in Section 1.3.

Observation 2. *Decreases in regional demographic variations lead to a more spread-out service region. Such impact is primarily attributed to customers' travel pattern change caused by smaller variations in destinations' attractiveness.*

With less regional variations in population and income levels, the proposed solutions in Figure 1.9, Figure 1.10 and Figure 1.11 suggest larger service region. The reason is that the trip distributions become more balanced as the population is more evenly spread and income disparity is reduced, following the gravity model.

From the customers' perspective, the attractiveness of destinations become more similar and thus the trip distributions become less concentrated, leading to the need to cover a larger service region. This finding suggests that the service region would be more concentrated in cities like New York, which is ranked as the metropolitan area of highest income inequality in the United States [94].

Implications of Charging Technology Advances

As the battery and charging technology improves, the charging speed for EVs are expected to be faster in the future. In fact, there exists fast charging technology in practice, e.g., supercharging for Tesla. We analyze how and to what extent the charging speed affects the optimal service region design. With faster charging speed, the charging time is equivalently reduced, if the battery capacity remains unchanged, which offers the potential to increase utilization of the EVs.

We vary

$$\rho = \frac{\text{current charging time} - \text{future charging time}}{\text{current charging time}},$$

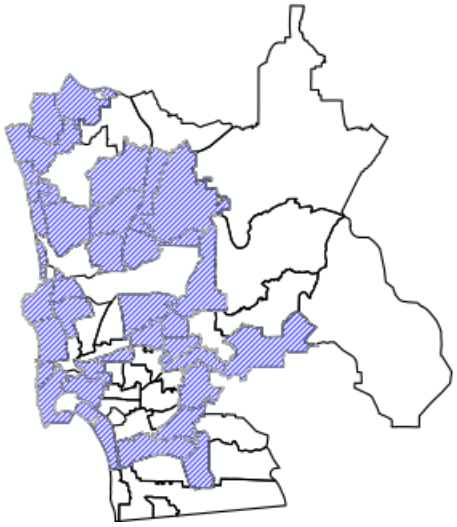


Figure 1.9: Optimal Service Region: Area coverage=37.31%; Selected zip codes= 33

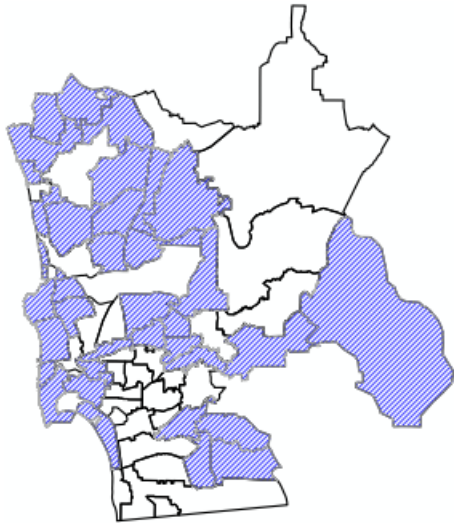


Figure 1.10: Reduced Population Variation: Area coverage=53.74%; Selected zip codes= 35

from 0.1 up to 1. Apparently, the larger ρ represents faster charging speed.

Observation 3. *Improvements in charging speed from the status quo brings significant benefits to the system with expanded service region and generally smaller fleet size. However, the increase in region coverage exhibits diminishing marginal effects.*

Not surprisingly, Figure 1.12 shows that more advanced charging technology enables the firm to serve a larger region with a smaller fleet. The major reduction in fleet size comes

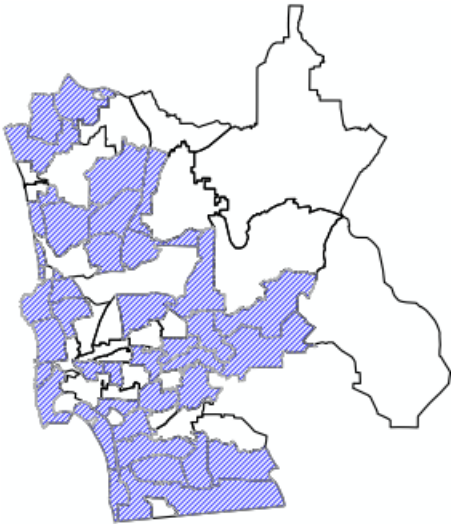


Figure 1.11: Reduced Income Disparity: Area coverage=43.87%; Selected zip codes= 37

from the decreasing queues of EVs at charging stations. In the extreme case of $\rho = 1$ when no charging time is required, the solution suggests the largest service region with the minimum fleet size. We notice that the change in optimal service region is diminishing as ρ becomes sufficiently high. This indicates that service region design is insensitive to charging technology advances as long as the charging speed is fast enough, e.g., 60% improvement in charging time from *status quo*. This suggests that charging speed is indeed one obstacle against service expansion, but as charging technology improves significantly, other obstacles such as fixed costs and demand imbalance will factor in to impede further region expansion. However, cost savings can still be achieved from reduction in fleet size.

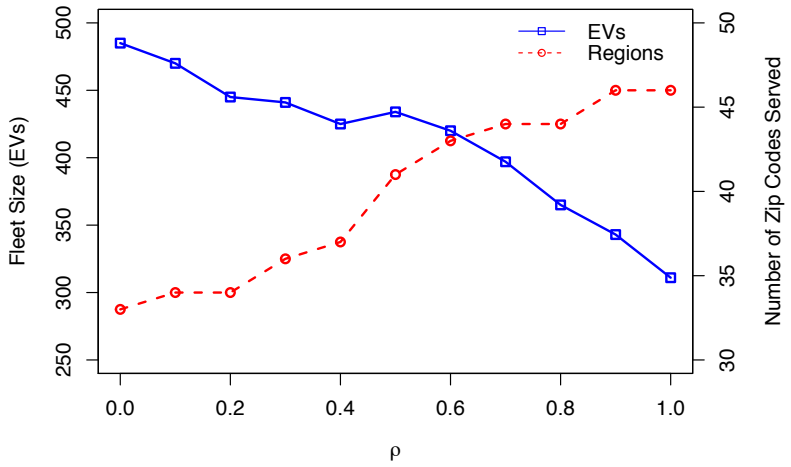


Figure 1.12: Service Region Design on Charging Speed

1.4 Summary

In this work, we study the service region design problem for an one-way EV sharing system. As customer adoption depends on the service coverage at their preferred destinations, we explicitly model the adoption decision, which primarily determines the firm's revenue, in a probabilistic form. Under limited information on the utility parameters of destinations, we first develop a distributionally-robust optimization model to evaluate the adoption rate aiming to maximize the expected profit. We further model the fleet operations, including repositioning and recharging, and determine the fleet size to guarantee the EV availability in the service region using queueing networks. We provide a lower bound on the expected profit by a computationally tractable MISCOP formulation. Several computational experiments are then conducted to demonstrate the model in a case study of Car2Go San Diego based on real operations data.

Our proposed solutions suggest expansion opportunities under properly selected service region and optimized fleet size. Because of higher vehicle utilization, EV sharing systems bring more environmental benefits, e.g., savings in CO_2 emissions, than replacing personal gasoline cars with EV ownership. We further examine how the service region changes along demographics, e.g., variations in population and income levels of the candidate regions. The recommended service region for a city with smaller regional variation in demographics is found to be more spread-out. Moreover, our results show that charging technology advances help to reduce the fleet size and expand the service region. While faster charging is always beneficial to fleet size reduction, it shows diminishing marginal impacts on service region

design.

Currently, Car2Go customers are able to book the cars on the website, smartphones or right on the street and allowed a 30-minute period to commence a trip after the vehicle is reserved without penalty [25]. Such reservation rule may be welcomed by the customers but may not be optimal to maintain EV utilization and availability. A future research direction will be to explore the dynamic grace period setting for reservation that helps to improve EV utilization, availability and profitability. For example, in peak hours, allowing 30-minute grace period may turn down many on-street demands that would have been able to utilize the cars immediately. Another possible research direction is dynamic pricing for vehicle sharing systems. To better matching supply and demand, Uber, a ridesharing service provider, is implementing "surge pricing" that increases rates to get more cars on the road and ensures reliability during the busiest times [95]. Despite the debatable performance of such pricing policy [74], it provides an idea of using dynamic pricing to coordinate the supply and demand in fleet operations. For the operations of EV sharing systems, such practice has potential to balance the trip distributions that leads to less repositioning activities and thus improves the service level.

Chapter 2

Demand Estimation and Inventory Allocation for Online Retailers

2.1 Introduction

Our motivating business problem comes from a major online retailer that makes inventory allocation decisions for a quarter million items among tens of distribution centers (DCs) across the country. It is important to optimize the allocation of inventory among the DCs in order to better utilize warehouse capacities as well as to save delivery costs for fulfilling customer orders from across the country. Inventory allocation (or positioning) decisions have to answer the following questions: 1) Which DCs within the network should fulfill the demand of a given item, and which customer zones should each DC be responsible for? 2) How much inventory should be on-hand in each DC in each planning period?

With the advances of information technology, firms are collecting more data and making decisions based on them. In particular, for online retailing, the firms rely on abundant sales data to understand customers and act on business plans. Therefore, such data-driven business decisions highly depend on the quality of sales data. In many cases, the sales data are incomplete and/or noisy with errors due to several reasons. For instance, a zero sales record may be erroneous and thus treated as missing data for reasons such as information system failures, product selling discontinued or inventory stock-outs. Hence, we are trying to make inventory decisions under poor data quality, e.g. with missing data.

In this chapter, we present models that facilitate inventory allocation for online retailers who possess abundant data of possibly poor quality. The first model aims to recover the missing data and remove outliers simultaneously to improve the data quality for data-driven business decision making. The second model provides ordering decisions for multiple products in the classic newsvendor setting.

Literature

Recent development in tensor-based multilinear data analysis has shown that tensor models are capable to provide better understanding and more precision from multilinear structures. Tensor decomposition is a type of multilinear data analysis method that commonly takes two forms: CANDECOMP/PARAFAC (CP) decomposition [26, 49] and Tucker decomposition [91]. Generally, tensor decomposition resembles Principal Component Analysis (PCA) for matrices. In particular, Tucker decomposition is also known as *higher-order SVD* (HOSVD) [34]. In many situations, even though the observed data may not be low-rank because of outliers and arbitrary errors, the underlying tensor data is often low-rank. That is, the variation in the data is greatly attributed to a relatively small number of latent factors. In view of this, robust tensor decompositions can be achieved from reconstructing the low-rank part of the observed sales data. Built upon Principal Component Pursuit (PCP) for Robust PCA [21] and Tensor Completion [43, 62, 90], robust low-rank tensor recovery has been formulated as a convex optimization model and efficient algorithms are discussed [45]. These methodologies have been widely implemented in image processing but are rarely seen in demand forecasting applications.

There are also work dealing with censored data as well as missing data in inventory literature that are related to our second model. Conrad (1976) [33] discuss the probability distribution of the demand from sales data with the Poisson arrival assumption. Nahmias (1994) [68] estimates Normal demand distribution and examine three estimators for the mean and standard deviation for lost sales inventory systems. Assuming the demand follows negative binomial distribution, Agrawal and Smith (1996) [2] develop parameter estimation methodology and demonstrate its effectiveness. Lu et al. (2008) [64] investigate the multiperiod inventory system of a perishable product and updates the demand distribution parameters periodically using the Bayesian approach based on the censored historical sales data. Most of the papers assume certain probability distributions of the demands. However, no one knows the exact forms of the demand distributions. Recent developments in data-driven approaches help to address the problems without such assumptions. Meanwhile, the tremendous scale of business data in industry brings the trend of "Big Data" in operations literature. In the stream of nonparametric "data-driven" approaches in newsvendor problems, Levi et al. (2007) [58] propose the Sample Average Approximation (SAA) based approach and establish the bounds on the number of samples required to guarantee the performance to be close to the scenario with known demand distributions. Other data-driven approaches in newsvendor settings have been discussed as well. Liyanage and Shanthikumar (2005) [63] introduce the operational statistics, a statistic of historical demand, for the newsvendor problem with ambiguous demand by integrating the estimation and optimization. In the context of censored demand data, Huh et al. (2011) [52] is the first application of the Kaplan-Meier estimator within an adaptive optimization algorithm. Besbes and Muharremoglu (2013) [15] show that the impact of censoring differs in the continuous and discrete demand cases and discuss that collecting even minimal information about lost sales can yield significant value.

Another stream of literature provides Bayesian perspectives. Li et al. (2014) [59] consider parameter estimation, parameter uncertainty characterization, and decision optimization for an inventory control problem with a demand model that includes customer choice with stochastically changing covariates, missing observations and auxiliary information. However, different from Bayesian models, e.g. Li et al. (2014) [59], our model is based on multilinear structure of the demands.

Furthermore, with the ambiguity of demand distribution, minimax approaches are considered to maximize the worst-case profit over possible demand distributions with known mean and variance in Scarf (1958) [80] and several extensions in Gallego and Moon (1993) [42]. With partial information available, e.g. mean, variance, symmetry and unimodality, Perakis and Roels (2008) [72] provide tractable formulations and derive the order quantities that minimize the newsvendor maximum regret of not acting optimally.

Recent ideas in machine learning have been brought into data-driven inventory studies as well. The work close to ours is Rudin and Vahn (2014) [77]. They propose both machine learning and kernel optimization approaches for optimal order quantity in newsvendor setting under "big data". Their work is an extension of the empirical model in He et al. (2012) [51], in which only two features, e.g. information on number of cases and information on types of cases, are made available. Their results demonstrate the impacts of data availability in newsvendor performance.

However, we consider multiple products so that the historical sales of products can provide "side information" on each others' demands that are usually highly correlated for same category products. Our model can be viewed as implicitly feature-based where side information might not be available and/or the firm is not sure what features the demands depend on.

2.2 Understand Geo-demand from Past Sales

In this section, we try to achieve better understanding of the demand distribution for various products across the country. To have a sound inventory allocation plan, item-location-time specific demand (geo-demand) distribution estimation from past sales data becomes critical. The geo-demand distributions serve as the guiding analytics for inventory allocation optimization. The sales data are pre-processed and organized in a tensor with three dimensions: item, location and time. However, the challenge of geo-demand estimation arises from the sales data that is noisy and sparse. The sales data acquired from the major online retailer have only 11.75% sales observations. The zero sales records are treated as missing data for the reasons discuss above. Moreover, the historical sales data is contaminated with noise coming from impulsive purchases and sporadic promotions. This kind of noise is, in general,

non-Gaussian, which renders traditional least-squares-based method inappropriate. Therefore, we aim to recover the underlying "true" geo-demand distributions that govern the sales observations via a convex optimization-based approach called robust low-rank tensor recovery [45].

Here, we formulate the missing geo-demand data completion problem as a robust low-rank tensor recovery problem in a convex optimization framework. We then develop an alternating direction augmented Lagrangian (ADAL) method that is easy to implement for solving the tensor recovery problem with partial observations. The algorithm efficiency and effectiveness are demonstrated with synthetic data. With a set of real-world sales data from a major online retailer, we investigate the performance of the framework both quantitatively and qualitatively through computational experiments. Finally, we conclude the benefits of the missing geo-demand data completion application for online retailing business.

Notation and Tensor Basics

We first introduce the mathematical notation and basic tensor operations following similar conventions in [45]. A *tensor* is denoted by boldface Euler script letters, e.g., $\mathcal{X}$, a matrix by boldface capital letters, e.g., $\mathbf{X}$, vectors by boldface lowercase letters, e.g., $\mathbf{x}$, and scalars by lowercase letters, e.g., x. The *order* N of a tensor is the number of dimensions (a.k.a. *ways* or *modes*). An Nth-order tensor is denoted by $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}$. A *fiber* is a column vector defined by fixing every index of $\mathcal{X}$ but one. The mode-i *unfolding* or matricization of the tensor $\mathcal{X}$ is denoted by the matrix $X_{(i)}$ that is a rearrangement of the mode-i fibers as the columns of the matrix in lexicographical order. The vectorization of $\mathcal{X}$ is denoted by $vec(\mathcal{X})$.

The inner product of two tensors in same dimensions $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is defined as $\langle \mathcal{X}, \mathcal{Y} \rangle := vec(\mathcal{X})^T vec(\mathcal{Y})$, and the Frobenius norm of $\mathcal{X}$ is defined as $\|\mathcal{X}\| := \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$. The nuclear norm (or trace norm) $\|\mathbf{X}\|_*$ of a matrix $\mathbf{X}$ is the sum of its singular values, i.e. $\|\mathbf{X}\|_* := \sum_i \sigma_i$, where the SVD of $\mathbf{X} = \mathbf{U}\text{diag}(\sigma)\mathbf{V}^T$. The L_1 norm of a vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|_1 := \sum_i |x_i|$. Likewise, for a matrix $\mathbf{X}$ and a tensor $\mathcal{X}$, $\|\mathbf{X}\|_1 := \|vec(\mathbf{X})\|_1$, and $\|\mathbf{X}\|_1 := \|vec(\mathcal{X})\|_1$.

We use symbol $\circ$ to denote vector outer product. The outer product of N vectors, $\mathbf{a}^{(n)} \in \mathbb{R}^{I_n}, n = 1, \cdots, N$ is an N-th-order tensor, defined as

$$\left(\mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}\right)_{i_1 i_2 \cdots i_N} := a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}$$

The multiplication of a tensor $\mathcal{X}$ of size $I_1 \times I_2 \times \cdots \times I_N$ with a matrix $\mathbf{A} \in \mathbb{R}^{J \times I_n}$ in mode n is denoted by $\mathcal{X} \times_n \mathbf{A} = \mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N}$, and is defined in terms of mode-n unfolding as $\mathbf{Y}_{(n)} := \mathbf{A}\mathbf{X}_{(n)}$.

We further use capital letters in calligraphic font to denote linear operators, e.g. $\mathcal{A}$, and $\mathcal{A}(\mathcal{X})$ as the result of applying the linear operator $\mathcal{A}$ to the tensor $\mathcal{X}$. $\mathcal{A}^*$ is the adjoint of

$\mathcal{A}$.

We then define a homogeneous tensor array (or tensor array for short) as the tensor obtained by stacking a set of component tensors of the same size along the first mode. An N-component tensor array is defined as $\bar{\mathcal{X}} := \begin{pmatrix} \mathcal{X}_1 \\ \vdots \\ \mathcal{X}_N \end{pmatrix} \in \mathbb{R}^{N \cdot I_1 \times \cdots \times I_N}$ is a "vector" of homogeneous tensor and written as $\text{TArray}(\mathcal{X}_1, \cdots, \mathcal{X}_N)$. A linear operator defined on a tensor array operates at component tensor level. For example, consider the linear (summation) operator $\mathcal{A} : \mathbb{R}^{N\dot{I}_1 \times \cdots \times I_N} \to \mathbb{R}^{I_1 \times \cdots \times I_N}$ such that $\mathcal{A}(\bar{\mathcal{X}}) := \sum_{i=1}^{N} \mathcal{X}_i$. Its adjoint is then the linear operator that reverts the operations $\mathcal{A}^* : \mathbb{R}^{I_1 \times \cdots \times I_N} \to \mathbb{R}^{N \cdot I_1 \times \cdots \times I_N}$ such that $\mathcal{A}^*(\mathcal{X}) := \text{TArray}(\mathcal{X}, \cdots, \mathcal{X})$. The non-calligraphic $\mathbf{A}$ denotes the matrix corresponding to the equivalent operation carried out by $\mathcal{A}$ on the mode-1 unfolding $\bar{\mathbf{X}}_{(1)}$ of $\bar{\mathbf{X}}$, where $\bar{\mathbf{X}}_{(1)} = \text{TArray}(\mathbf{X}_{1,(1)}, \cdots, \mathbf{X}_{N,(1)})$. Therefore, $\mathbf{A} = \begin{pmatrix} \mathbf{I} & \cdots & \mathbf{I} \end{pmatrix} \in \mathbb{R}^{I_1 \times N \cdot I_1}$ in this example.

Tensor Decompositions

The Tucker decomposition decomposes the tensor into the product of a small core tensor and a set of matrices. It approximates $\mathcal{X}$ as

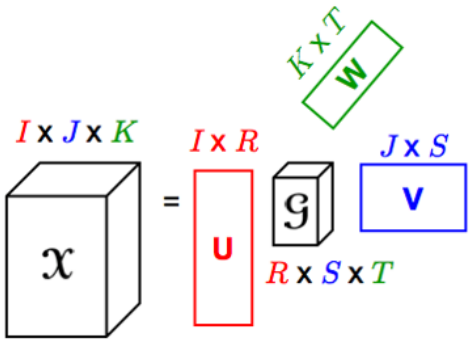
$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}$$

where $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_N}$ is the core tensor, and the factor matrices $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times r_n}, n = 1, \cdots, N$ are all column-wise orthonormal, where $(r_1 \times \cdots \times r_N)$ are given integers. The n-rank (or mode-n rank) of $\mathcal{X}$, denoted by $\text{rank}_n(\mathcal{X})$, is the column rank of $\mathbf{X}_{(n)}$. The set of N n-ranks of a tensor is also called the Tucker rank. If $\mathcal{X}$ is of rank-$(r_1, \cdots, r_N)$, then the approximation holds with equality, and for $n = 1, \cdots, N$, $\mathbf{U}^{(n)}$ is the matrix of the left singular vectors of $\mathbf{X}_{(n)}$. Figure 2.1 illustrates the procedure. The Tucker decomposition is formulated as a non-convex optimization problem. To compute the factor matrices, the higher-order orthogonal iteration (HOOI) [35] is usually deployed, which is essentially an alternating least-squares (ALS) algorithm [89] based on computing the dominant left singular vectors of each $\mathbf{X}_{(n)}$.

Robust PCA

In the context of matrices, PCA provides the optimal low-dimensional estimate with additive i.i.d. Gaussian noise. However, it is known to be susceptible to gross corruptions and outliers. To over come this challenge, robust PCA (RPCA) has been developed to robustify the solution to large errors and outliers. Candes et. al. [21] proposed a RPCA approach via Principal Component Pursuit (PCP) that decomposes a given observation (noisy) matrix $\mathbf{B}$ into a low-rank component $\mathbf{X}$ and a sparse component $\mathbf{E}$ by solving the optimization problem $\min_{\mathbf{X},\mathbf{E}} \left\{ \text{rank}(\mathbf{X}) + \lambda \|\mathbf{E}\|_0 \mid \mathbf{X} + \mathbf{E} = \mathbf{B} \right\}$. Since it is NP-hard, [21] uses the nuclear

Figure 2.1: Illustration of Tucker decomposition. [56]



norm and the L_1 norm to replace the rank and cardinality ($\|\cdot\|_0$) functions, and solves the following convex optimization problem:

$$\min_{\mathbf{X},\mathbf{E}} \left\{ \|\mathbf{X}\|_* + \lambda \|\mathbf{E}\|_1 \mid \mathbf{X} + \mathbf{E} = \mathbf{B} \right\}. \tag{2.1}$$

The optimal solution to problem (2.1) has been shown to exactly recover the low-rank matrix from sufficiently sparse errors $\mathbf{E}$ relative to the rank of $\mathbf{X}$, or more precisely, under the following condition [21]:

$$\text{rank}(\mathbf{X}) \leq \frac{\rho_r \max(n,m)}{\mu (\log \min(n,m))^2}, \quad \|\mathbf{E}\|_0 \leq \rho_s mn,$$

where ρ_r and ρ_s are positive constants, and μ is the incoherence parameter.

Higher-order RPCA

Robust tensor recovery, or higher-order RPCA (HoRPCA), is a generalization of RPCA to tensors that exploit the low-rank structure in all dimensions of the data. We regularize the Tucker rank Trank($\mathcal{X}$) and lead to the following tensor PCP optimization problem: $\min_{\mathcal{X},\mathcal{E}} \left\{ \text{Trank}(\mathcal{X}) + \lambda \|\mathcal{E}\|_0, s.t. \quad \mathcal{X} + \mathcal{E} = \mathcal{B} \right\}$. This problem is also NP-hard to solve and we replace Trank($\mathcal{X}$) by the convex surrogate CTrank($\mathcal{X}$), and $\|\mathcal{E}\|_0$ by $\|\mathcal{E}\|_1$ to make the problem tractable:

$$\min_{\mathcal{X},\mathcal{E}} \left\{ \text{CTrank}(\mathcal{X}) + \lambda \|\mathcal{E}\|_1 \mid \mathcal{X} + \mathcal{E} = \mathcal{B} \right\}. \tag{2.2}$$

The model (2.2) is called Higher-order RPCA (HoRPCA) [45]. In our model, the tensor rank regularization term is the sum of the N nuclear norms $\|\mathbf{X}_{(i)}\|_*$ of the mode-i unfoldings, $i = 1, \cdots, N$ of $\mathcal{X}$, i.e. $\text{CTrank}(\mathcal{X}) := \sum_i \|\mathbf{X}_{(i)}\|_*$.

Estimating Geo-demand Distribution

Our algorithm is motivated by the inventory allocation problem faced by a major U.S. online retailer that ships hundreds of thousand items across the country from tens of its DCs. The

guiding analytics for optimal inventory allocation (or positioning) is the geo-demand distribution of each item sold on the retailer's web-site. It is critical to learn the item-location-time specific demand (geo-demand) distribution collaboratively from historical sales data and to generalize well. Specifically, the geo-demand distribution provides estimation of the percentage of the demand, β_{rst} in each customer zone s relative to the total demand of a particular item r in time t: for example, in week 10, the demand of the Apple iPhone 6 in customer zone 20 counts for 5% of the total demand nationwide. The historical sales data, after it is normalized location-wise, is organized in tensor $\mathcal{B}$ with three dimensions: items, demand zones, and time (in weeks). However, only a small percentage of the entries in $\mathcal{B}$ have positive observations. The zero sales entries in $\mathcal{B}$ are not "trustful" and hence treated as missing data due to several reasons including system error, item stock out or zero demand. The goal of our algorithm is to recover the missing data by identifying the "true" values and noises to help estimate the geo-demand distributions $\{\beta_{rst}\}$.

Suppose $\mathcal{B}$ contains R items, S demand zones and T weeks of geo-demand records. Naturally, the tensor $\mathcal{B}$ is of dimension: $R \times S \times T$. Furthermore, since $\mathcal{B}$ is obtained by normalizing the past sales location-wise, we have $\sum_{s=1}^{S} \mathcal{B}_{(r,s,t)} = 1$, for all item r and week t.

Application of Robust Tensor Recovery

Suppose the observed geo-demand tensor $\mathcal{B}$ can be decomposed to the "true" geo-demand distribution $\mathcal{Y}$ and an error tensor ε: $\mathcal{B} = \mathcal{Y} + \varepsilon$. We then formulate the following optimization problem for robust tensor recovery.

$$\begin{aligned} \min_{\mathcal{Y},\varepsilon} \quad & \sum_{i=1}^{N} \lambda_i \|\mathbf{Y}_{(i)}\|_* + \lambda_e \|\varepsilon\|_1 \\ \text{s.t.} \quad & \\ & \mathcal{Y} + \varepsilon = \mathcal{B} \\ & \sum_{s \in S} y_{(r,s,t)} = 1, \forall r \in R, t \in T \\ & \mathcal{Y} \geq 0 \end{aligned} \tag{2.3}$$

where $N = 3$, λ_i is given penalty on the rank of mode-i unfolding of $\mathcal{Y}$ and λ_e is the penalty on l_1 norm of the error tensor ε. In the formulation, the objective function (2.3) is to minimize the rank of tensor $\mathcal{Y}$ in all modes together with the l_1 norm of the error tensor ε. The first constraint ensures that the resulting "true" tensor $\mathcal{Y}$ and error tensor ε are consistent with the observed tensor $\mathcal{B}$. The rest are simplex constraints so that the recovered $\mathcal{Y}$ is in probability space.

To take the advantage of the problem structure, we apply variable-splitting to $\mathcal{Y}$ and introduce three auxiliary variables $\mathcal{X}_1 = \cdots = \mathcal{X}_N = \mathcal{Y}$. Moreover, let the set Ω denote the the indices of positive observations. That is, the entry in $\mathcal{B}$ with index $(r, s, t) \in \Omega$ has positive value: $\mathcal{B}_{(r,s,t)} > 0, \forall (r, s, t) \in \Omega$. We then enforce the consistency on the observed data through linear projection operator $\mathcal{A}_\Omega$: $\mathbb{R}^{R\times S\times T} \to \mathbb{R}^m$ that selects the set of m elements of positive observations (Ω) from $\mathcal{B}$. The problem (2.3) can be reformulated into

$$\begin{aligned} \min_{\mathcal{X}_i,\mathcal{Y},\varepsilon} \quad & \sum_{i=1}^{N} \lambda_i \|\mathbf{X}_{i,(i)}\|_* + \lambda_e \|\varepsilon\|_1 \\ \text{s.t.} \quad & \\ & \mathcal{X}_i = \mathcal{Y}, \forall i = 1, \cdots, N \\ & \mathcal{A}_\Omega(\mathcal{Y} + \varepsilon) = \mathcal{B}_\Omega \\ & \sum_{s\in S} y_{(r,s,t)} = 1, \forall r \in R, t \in T \\ & \mathcal{Y} \geq 0. \end{aligned} \tag{2.4}$$

Following the same spirit in Goldfarb and Qin (2014) [45], we assume $[\varepsilon]_{\bar{\Omega}} = 0$, where $\bar{\Omega}$ is the complement of Ω. Otherwise, it is impossible to recover ε, since some of corrupted tensor elements are not observed. Similarly, we need to define several operations before we develop the optimization algorithm for solving problem (2.4). $\text{fold}_i(\mathcal{X})$ returns the tensor $\mathcal{Z}$ such that $\mathbf{Z}_{(i)} = \mathbf{X}$. $\mathcal{T}_\mu(\mathbf{X})$ is the matrix singular value thresholding operator: $\mathcal{T}_\mu(\mathbf{X}) := \mathbf{U}\text{diag}(\bar{\sigma})\mathbf{V}^T$, where $\mathbf{X} = \mathbf{U}\text{diag}(\sigma)\mathbf{V}^T$ is the SVD of $\mathbf{X}$ and $\bar{\sigma} := \max(\sigma - \mu, 0)$. We define $\mathcal{T}_{i,\mu}(\mathcal{X}) := \text{fold}_i(\mathcal{T}_\mu(\mathbf{X}_{(i)}))$. $\mathcal{S}_\mu(\mathcal{X})$ is the shrinkage operator on $\text{vec}(\mathcal{X})$ and returns the result as a tensor. The vector shrinkage operator is defined as $\mathcal{S}_\mu(\mathbf{x}) := \text{sign}(\mathbf{x}) \max(|\mathbf{x}| - \mu, 0)$, where the operations are all element-wise.

Solution Scheme

We adopt the alternating-direction augmented Lagrangian (ADAL) (or alternating-direction method of multipliers (ADMM)) [38, 17, 45] to solve the structured linearly-constrained optimization problem (2.4).

Define the simplex constraint set by $\Delta := \{\mathcal{Y} : \sum_{s\in S} y_{(r,s,t)} = 1, \mathcal{Y} \geq 0\}$. Keeping the simplex constraints for $\mathcal{Y}$, the partial augmented Lagrangian formulation for (2.4) is given by:

$$\begin{aligned} & \mathcal{L}(\mathcal{X}_i, \mathcal{Y} \in \Delta, \varepsilon, \Gamma_i, \theta) \\ & = \sum_{i=1}^{N} \lambda_i \|\mathbf{X}_{i,(i)}\|_* + \lambda_e \|\varepsilon\|_1 + \sum_{i=1}^{N} \left(\frac{1}{2\mu}\|\mathcal{X}_i - \mathcal{Y}\|^2 - \langle \Gamma_i, \mathcal{X}_i - \mathcal{Y}\rangle\right) \\ & \quad + \frac{1}{2\mu}\|\mathcal{A}_\Omega(\mathcal{Y} + \varepsilon) - \mathcal{B}_\Omega\|^2 - \langle \theta, \mathcal{A}_\Omega(\mathcal{Y} + \varepsilon) - \mathcal{B}_\Omega\rangle \end{aligned}$$

where $\Gamma_i \in \mathbb{R}^{R\times S\times T}$ and $\theta \in \mathbb{R}^m$, given m observations.

We start by solving the subproblem for $\mathcal{X}_i$:

$$\min \|\mathbf{X}_{i,(i)}\|_* + \frac{1}{2\mu\lambda_i}\|\mathcal{X}_i - \mathcal{Y}\|^2 - \langle \frac{\Gamma_i}{\lambda_i}, \mathcal{X}_i - \mathcal{Y}\rangle \tag{2.5}$$

Given $\mathcal{Y}$, the optimal $\mathcal{X}_i$ is obtained by solving the first-order condition (FOC) for the convex subproblem (2.5):

$$\mathcal{X}_i^* = \mathcal{T}_{i,\mu\lambda_i}(\mu\Gamma_i + \mathcal{Y})$$

We then proceed to solve the subproblem for ε:

$$\begin{aligned}&\min \lambda_e\|\varepsilon\|_1 + \frac{1}{2\mu}\|\mathcal{A}_\Omega(\mathcal{Y}+\varepsilon) - \mathcal{B}_\Omega\|^2 - \langle\theta, \mathcal{A}_\Omega(\mathcal{Y}+\varepsilon) - \mathcal{B}_\Omega\rangle\\ \equiv &\min \mu\lambda_e\|\varepsilon\|_1 + \frac{1}{2}\|\mathcal{A}_\Omega(\varepsilon) + \mathcal{A}_\Omega(\mathcal{Y}-\mathcal{B}) - \mu\theta\|^2\end{aligned}$$

Similarly, by taking the FOC, the optimal solution for ε is given by:

$$\varepsilon^* = \mathcal{S}_{\mu\lambda_e}(\mathcal{A}_\Omega^*(\mathcal{A}_\Omega(\mathcal{B}-\mathcal{Y}) + \mu\theta))$$

Given $\mathcal{X}_i$ and ε, the subproblem for $\mathcal{Y}$ can be rearranged as

$$\min_{\mathcal{Y}\in\Delta} \sum_{i=1}^{N} \frac{1}{2}\|\mathcal{C}(\mathcal{Y}) + \mathcal{D}\|^2 + \frac{1}{2}\|\mathcal{A}_\Omega(\mathcal{Y}) + \mathcal{A}_\Omega(\varepsilon - \mathcal{B}) - \mu\theta\|^2$$

where $\mathcal{C}(\mathcal{Y}) :=\text{TArray}(\mathcal{Y}, ..., \mathcal{Y})$ and $\mathcal{D} :=\text{TArray}(\mu\Gamma_1 - \mathcal{X}_1, ..., \mu\Gamma_N - \mathcal{X}_N)$.

Its FOC is then

$$\begin{aligned}&0 \in \mathcal{C}^*\mathcal{C}(\mathcal{Y}) + \mathcal{C}^*(D) + \mathcal{A}_\Omega^*\mathcal{A}_\Omega(\mathcal{Y}) + \mathcal{A}_\Omega^*\mathcal{A}_\Omega(\varepsilon - \mathcal{B}) - \mathcal{A}_\Omega^*(\mu\theta)\\ \equiv &0 \in N\mathcal{Y} + \sum_{i=1}^{N}(\mu\Gamma_i - \mathcal{X}_i) + \mathcal{A}_\Omega^*\mathcal{A}_\Omega(\mathcal{Y}) + \mathcal{A}_\Omega^*\mathcal{A}_\Omega(\varepsilon - \mathcal{B}) - \mathcal{A}_\Omega^*(\mu\theta)\end{aligned}$$

Given Ω, we can find the closed form expression for the elements of y in two cases: If $(r, s, t) \in \Omega$, then we have

$$y^{*(r,s,t)} = \big(\mathcal{B}^{(r,s,t)} - \varepsilon^{(r,s,t)} + \mathcal{A}_\Omega^*(\mu\theta)^{(r,s,t)} - \sum_{i=1}^{N}(\mu\gamma_i^{(r,s,t)} - X_i^{(r,s,t)})\big)/(N+1) \tag{2.6}$$

If $(r, s, t) \notin \Omega$, then we have

$$y^{*(r,s,t)} = \frac{\sum_{i=1}^{N}(X_i^{(r,s,t)} - \mu\gamma_i^{(r,s,t)})}{N} \tag{2.7}$$

The results obtained from (2.6) and (2.7) are then projected onto the simplex Δ. The Euclidian projection onto the simplex can be solved for efficiently by an $\mathcal{O}(n \log n)$ algorithm proposed in [37].

By integrating the projection method developed in Duchi et al. (2008) [37], we summarize the HoRPCA-GD algorithm for geo-demand estimation.

Experiments

We investigate the performance of the proposed algorithm through experiments with both synthetic and past sales data. The experiments are implemented in R with package **rTensor** [60] which contains basic tensor operations, e.g. folding/unfolding, multiplication of a tensor and a matrix.

Synthetic Data

Based on the approach in Tomioka et al. (2010) [90], we generate a random rank-(5,5,5) tensor of size (50,50,20) by drawing the core tensor of size (5,5,5) from the uniform distribution $\mathcal{U}(5, 15)$ and multiplying each mode of the core tensor by an orthonormal matrix of appropriate dimensions. The generated tensor is verified to have the desired Tucker rank. 10% of the tensor elements are randomly corrupted by additive i.i.d. noise from the uniform distribution $\mathcal{U}(-5, 5)$. We then randomly selected a fraction 50% of the noisy tensor elements to be the given observations $\mathcal{B}_\Omega$. Let the penalty parameters be $\lambda_1 = \sqrt{50}, \lambda_2 = \sqrt{50}, \lambda_3 = \sqrt{20}$ and $\lambda_e = 1$.

Given the output $\hat{\mathcal{Y}}$ and the "true" tensor $\mathcal{X}$, we define the relative error as a performance measure below:

$$\text{relative error} = \frac{\|\hat{\mathcal{Y}} - \mathcal{X}\|}{\|\mathcal{X}\|} \tag{2.8}$$

The algorithm reports a relative error at 0.34% after 100 iterations. Figure 2.2 shows the algorithm converges quickly, e.g. at 80 iterations. The algorithm effectively and efficiently delivers good recovery results in small number of iterations for the simplex-constrained low-rank tensor recovery problem.

Algorithm 1 HoRPCA-GD

Given $\mathcal{B}, \lambda, \mu$. Initialize $\mathcal{X}_i^{(0)} = \varepsilon^{(0)} = \Gamma_i^{(0)} = 0 \ \forall i \in \{1, ..., N\}$, $\mathcal{Y} = 0$ and $\theta = 0$.
for $k = 0, 1, ...$ **do**
  for $i = 1 : N$ **do**
    $\mathcal{X}_i^{(k+1)} \leftarrow \mathcal{T}_{i,\mu\lambda_i}(\mu\Gamma_i + \mathcal{Y}^{(k)})$
  end for
  $\varepsilon^{(k+1)} \leftarrow \mathcal{S}_{\mu\lambda_e}(\mathcal{A}_\Omega^*(\mathcal{A}_\Omega(\mathcal{B} - \mathcal{Y}^{(k)}) + \mu\theta))$
  Update $\mathcal{Y}$ element-wise:
    If $(r, s, t) \in \Omega$, then

$$y^{(k+1)^{(r,s,t)}} = \big(\mathcal{B}^{(r,s,t)} - \varepsilon^{(k+1)^{(r,s,t)}} + \mathcal{A}_\Omega^*(\mu\theta^{(k)})^{(r,s,t)} - \sum_{i=1}^{N}(\mu\gamma_i^{(k)^{(r,s,t)}} - X_i^{(k+1)^{(r,s,t)}})\big)/(N+1)$$

    If $(r, s, t) \notin \Omega$, then

$$y^{(k+1)^{(r,s,t)}} = \frac{\sum_{i=1}^{N}(X_i^{(k+1)^{(r,s,t)}} - \mu\gamma_i^{(k)^{(r,s,t)}})}{N}$$

  Projection to the simplex:
    For each item r and week t,
  sort vector $y(r, :, t)$ into $v : v_1 \geq v_2 \geq ... \geq v_S$.
    Find $\rho = \max\left\{s : v_s - \frac{1}{s}\left(\sum_{n=1}^{s} v_n - 1\right) > 0\right\}$.
    $\theta = \frac{1}{\rho}\left(\sum_{m=1}^{\rho} v_m - 1\right)$.
    Update $\mathcal{Y}$ s.t. $y(r, s, t) = \max\{v_i - \theta, 0\}$.
  for i=1:N **do**
    $\Gamma_i^{(k+1)} \leftarrow \Gamma_i^{(k)} - \frac{1}{\mu}(\mathcal{X}_i^{(k+1)} - \mathcal{Y}^{(k+1)})$
  end for
  $\theta^{(k+1)} \leftarrow \theta^{(k)} - \frac{1}{\mu}(\mathcal{A}_\Omega(\mathcal{Y}^{(k+1)} + \varepsilon^{(k+1)}) - \mathcal{B}_\Omega)$
end for
return $(\frac{1}{N}(\sum_{i=1}^{N} \mathcal{X}_i^{(k)}), \varepsilon_i^{(k)})$

Figure 2.2: Convergence with Synthetic Data

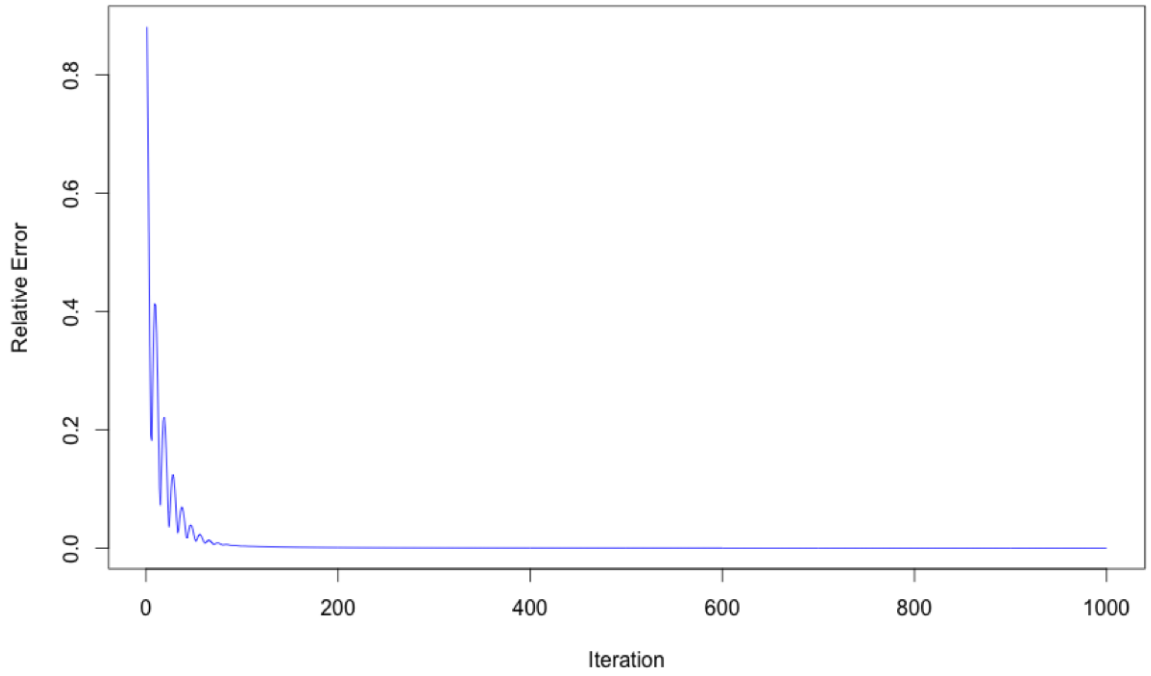
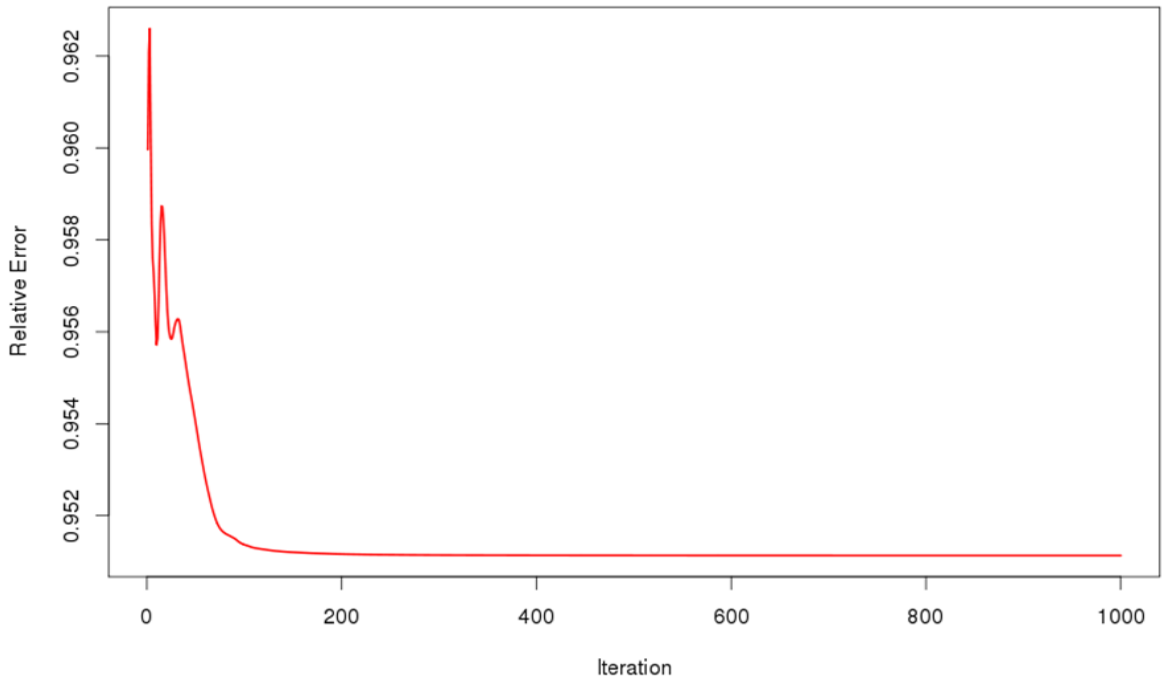


Figure 2.3: Convergence with Sample Geo-Demand



Real Sales Data

The algorithm is then applied to past sales data of the top 100 best selling items. The identities of the items, weeks, and customer locations have been masked and replaced by numerical indices. There are 57.9% of entries with positive observations in the sales data tensor with dimensions: 100 items, 125 zones and 26 weeks respectively. Since we are interested in geo-demand distribution, the sales data tensor is normalized along the zone dimension and leads to sample geo-demand tensor $\mathcal{B}$. In reality, we do not know the "true" geo-demand distribution. Therefore, with the output $\hat{\mathcal{Y}}$, we modify the stopping criterion measure as

$$\text{relative distance} = \frac{\|\hat{\mathcal{Y}} - \mathcal{B}\|}{\|\mathcal{B}\|} \tag{2.9}$$

We set the penalty parameters as $\lambda_1 = \sqrt{100}, \lambda_2 = \sqrt{125}, \lambda_3 = \sqrt{26}$ and $\lambda_e = 1$. Figure 2.3 shows that the output $\hat{\mathcal{Y}}$ converges after 100 iteration.

Qualitative Analysis

We discuss the performance of our framework by comparing the recovered geo-demand distribution $\hat{\mathcal{Y}}$ for year 2012 and the observed geo-demand distribution from year 2013. We start by picking item 1 and zone 1 where the data for first two weeks in 2012 and the majority of 2013 data are missing. Figure 2.4 shows the recovered geo-demand from 2012 correctly predict the geo-demand in 2013. Moreover, the recovered geo-demand for 2013 shows similar pattern as the observed geo-demand in 2012.

Moreover, as shown in Figure 2.5, there are smoothing effect on the geo-demand distribution. The recovered geo-demands for both years agree with each other and represent smaller volatility over time compared to the observed geo-demands.

Consider a demand zone where usually has relatively small geo-demand, a sudden peak of the demand in previous year can not be the signal of large demand in this year. Figure 2.6 shows that high demands are observed for several weeks in 2012. However, the recovered geo-demand still suggests low geo-demand that coincides with the observed geo-demand in 2013. One of the benefits of geo-demand recovery is to identify "false" demand peak and avoid mistakenly allocate more inventory to distribution centers serving that demand zone.

Quantitative Analysis

The resulting tensor $\hat{\mathcal{Y}}$ was then multiplied to the item-week aggregation of the original sales tensor $\mathcal{B}$ to obtain 're-distributed' sales tensor $\hat{\mathcal{B}}$, which was then input into the multinomial Bayesian framework [75] as the *evidence* data. In Qin et al. (2014) [75], the geo-demand distributions estimation problem is approached as learning the probability distribution $\{\beta_{rst}\}$

Figure 2.4: Geo-Demand of Item 1 in Zone 1

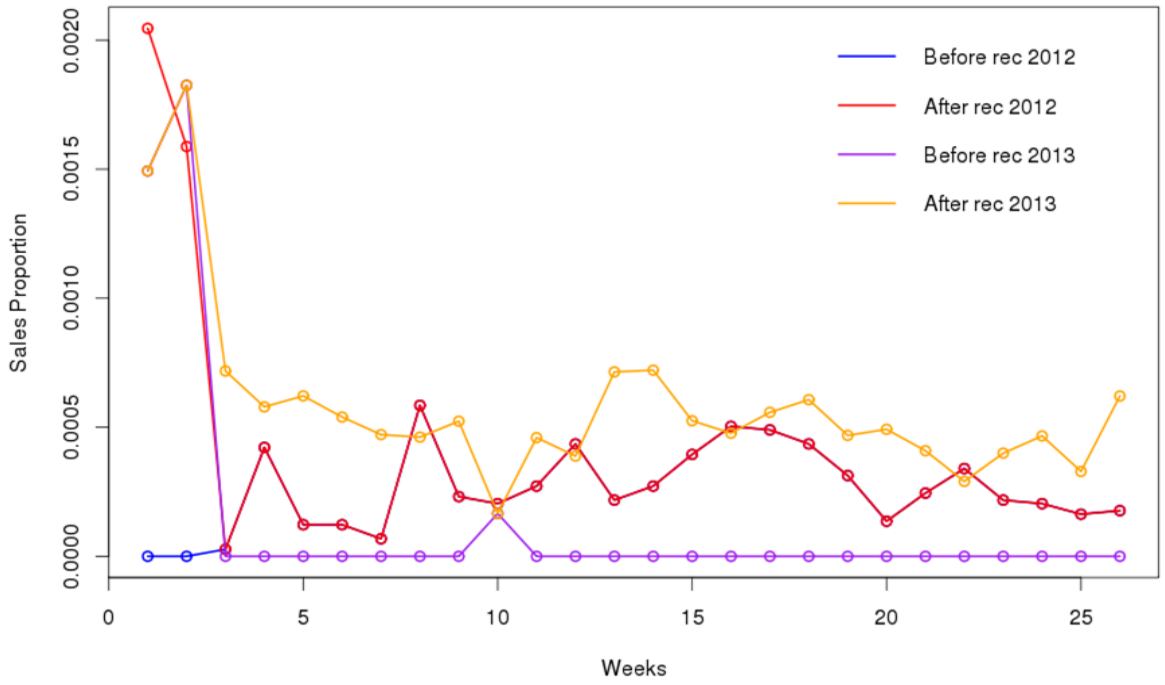


Figure 2.5: Geo-Demand of Item 10 in Zone 10

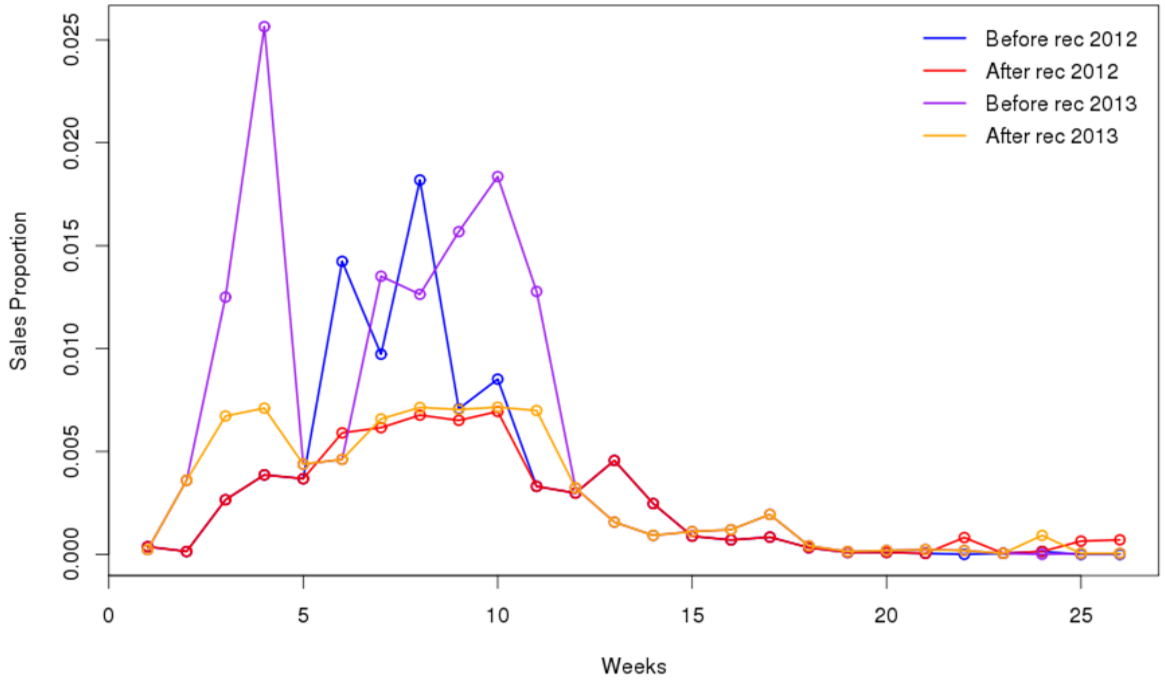
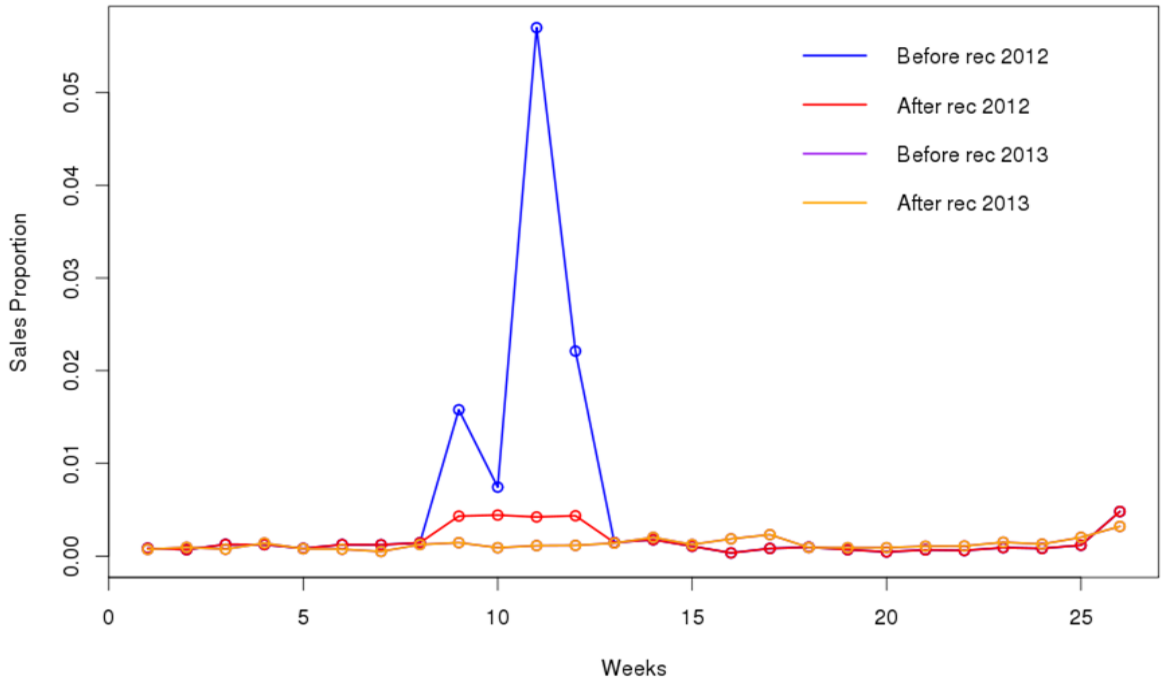
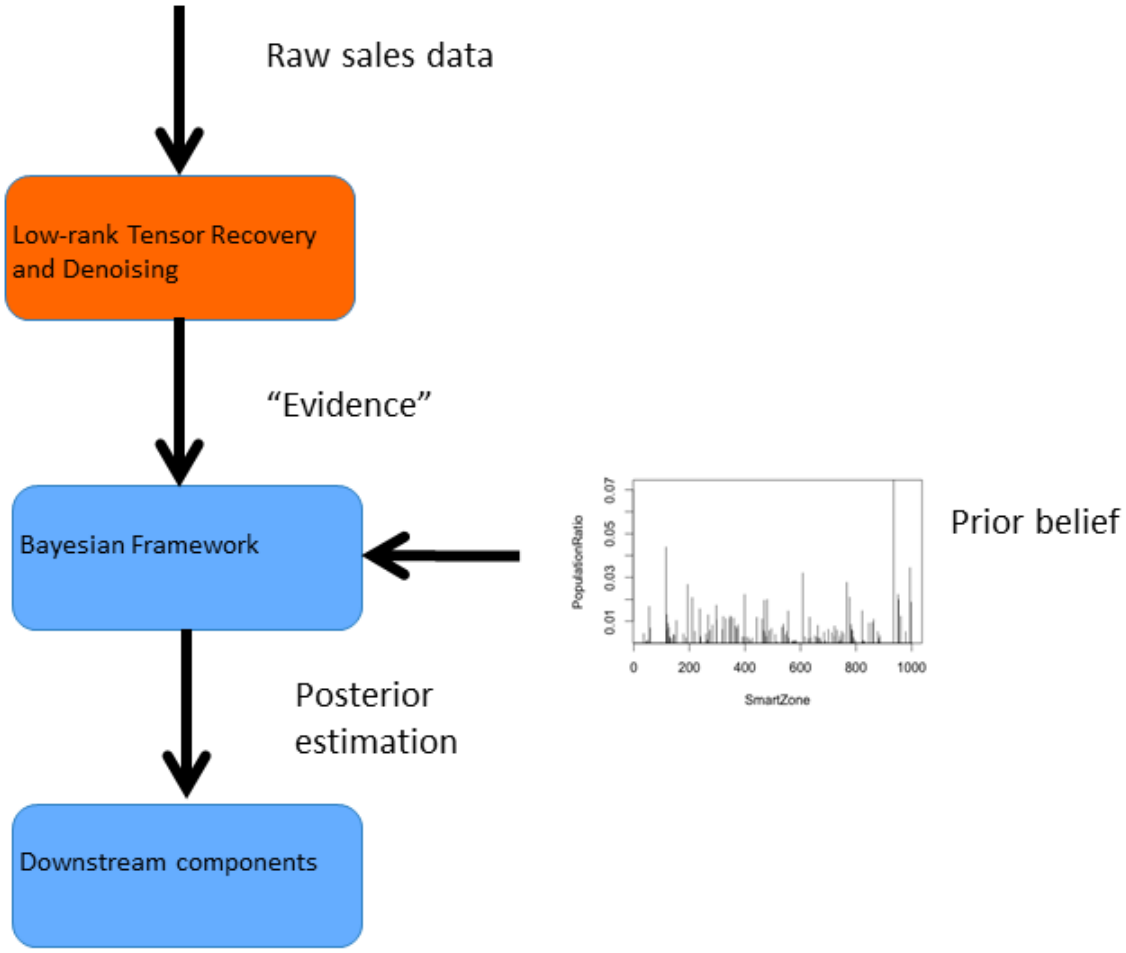


Figure 2.6: Geo-Demand of Item 3 in Zone 11



for an order of item r arising from demand zone s at time t. The set of sales data for item r at time t across all demand zones is assumed to follow the Multinomial distribution with parameters $\{\beta_{rst}\}$. The high-level work flow is show in Figure 2.7. The final geo-demand distributions were estimated by the posterior multinomial distributions. It is well known that the quality of the evidence has strong impact on the posterior of a Bayesian framework. By using the 'denoised' evidence data $\hat{\mathcal{B}}$ of 2012, we were able to improve the estimation for 2013 by various extents. Figures 2.8, 2.9, and 2.10 compare the generalization results of using $\hat{\mathcal{B}}$ as the evidence data to those using the raw historical sales data. The values being plotted are the difference between the two, with negative values indicating improvements. The mean absolute error (MAE) by week (Figure 2.8) was reduced for all the weeks except one within the test horizon of 26 weeks. The MAE's were also improved for the majority of the items under consideration according to Figure 2.9. From the inventory positioning and control perspective, it is also beneficial to have smaller week-over-week volatility of the geo-demand distributions, in addition to smaller MAE's. The reason comes in two folds: 1) The distribution center operations require smooth ramp-up and ramp-down. A smaller week-over-week volatility would result in a smoother allocation plan of the inventory across time. 2) In the classical EOQ paradigm [28], a smaller demand volatility generally results in lower safety stock requirement, and hence, a lower inventory level. We measured the week-over-week volatility by the average demand distribution variation of two consecutive weeks for a given customer zone for each item, same as the penalty function in the fused-Lasso [88]. Figure 2.10 shows that the fused-Lasso norm decreases for most of the items under

Figure 2.7: Integration of HoRPCA-GD with existing framework



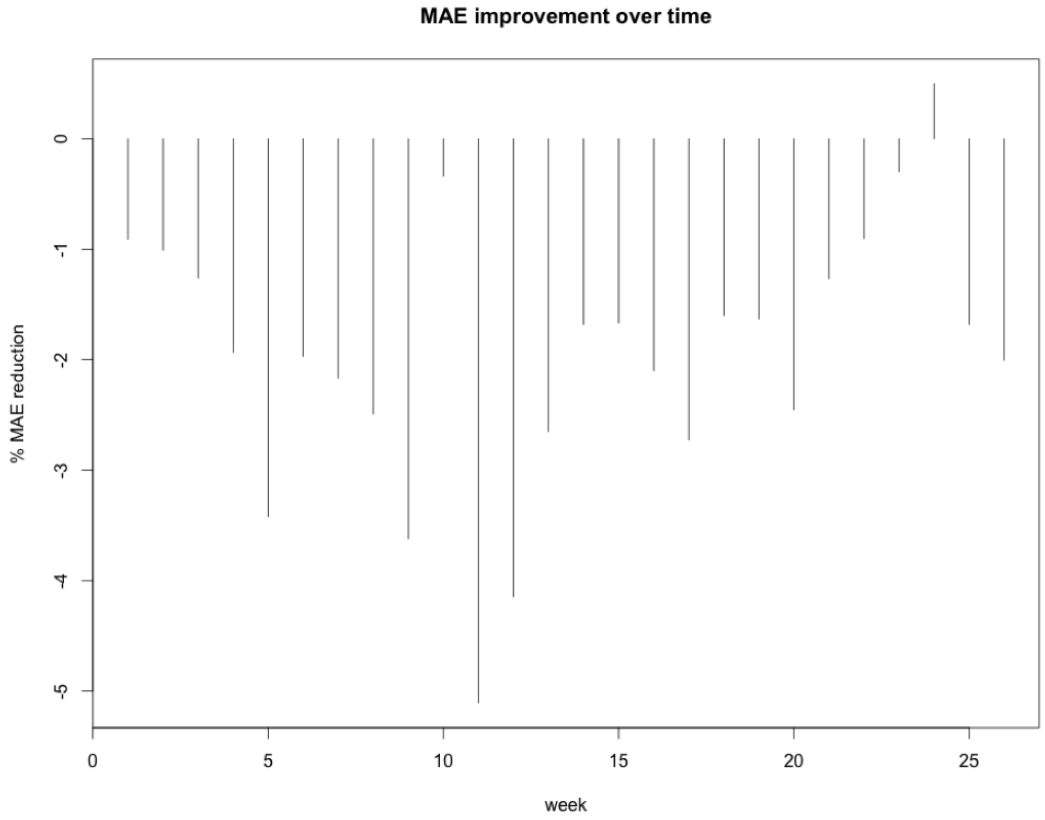
consideration, suggesting a smoother evolution of the demand distributions.

2.3 A Multi-product Newsvendor Model with Missing Data

The Newsvendor Problem

Consider a firm that sells perishable goods with both understock backordering and overstock holding costs denoted by b and h respectively. In the single period setting, the firm's objective is to minimize its expected costs $\mathbf{C}(q)$ by ordering q quantity in the face of uncertain demand D that follows some distributions. Given the realization of demand D, the total costs is

Figure 2.8: Comparison of MAE across time



provided by:

$$\mathbf{C}(q, D) = b(D - q)^+ + h(q - D)^+ \tag{2.10}$$

Thus, the expected costs is as follows:

$$\min_{q \geq 0} \mathbb{E}\mathbf{C}(q, D)$$

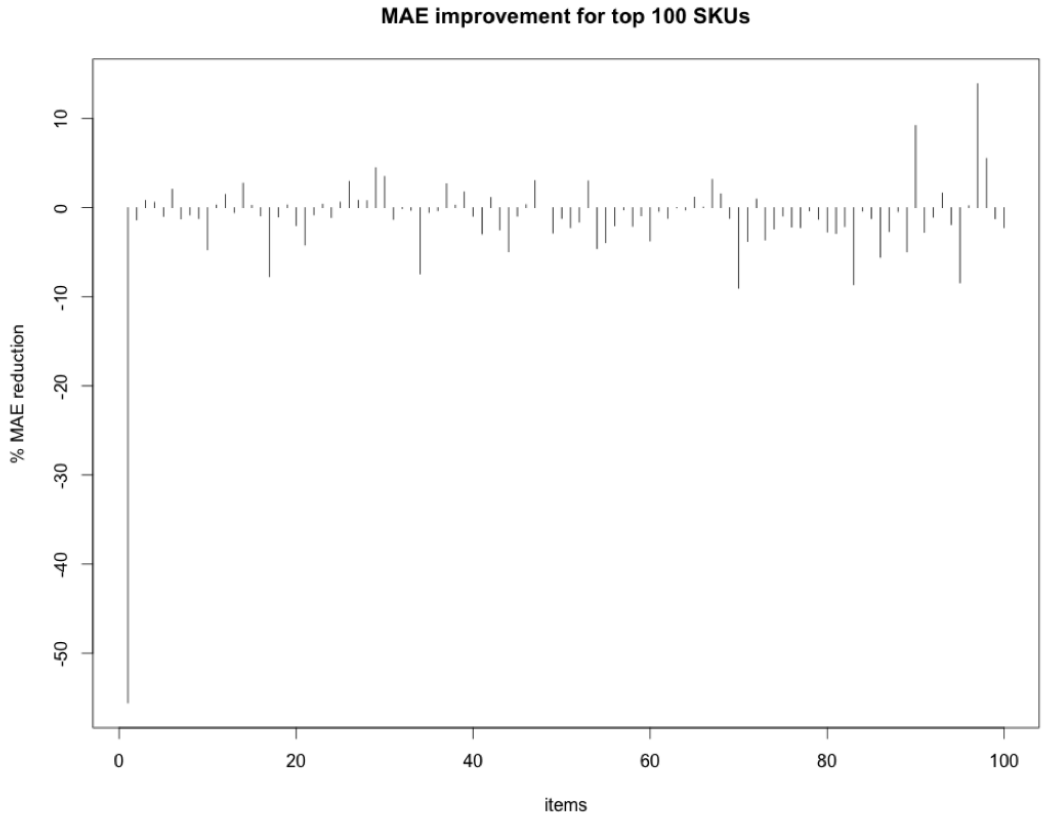
When the demand distribution F is provided, it is well known that the optimal order quantity is given by the critical fractile, e.g. $b/(b + h)$, that is

$$q^* = \inf\{y : F(y) \geq \frac{b}{b + h}\}$$

The Data-driven Newsvendor Problem

In reality, the firm usually does not have access to the demand distribution. The traditional data-driven formulation assumes that the firm has only access to well-maintained historical

Figure 2.9: Comparison of MAE across items

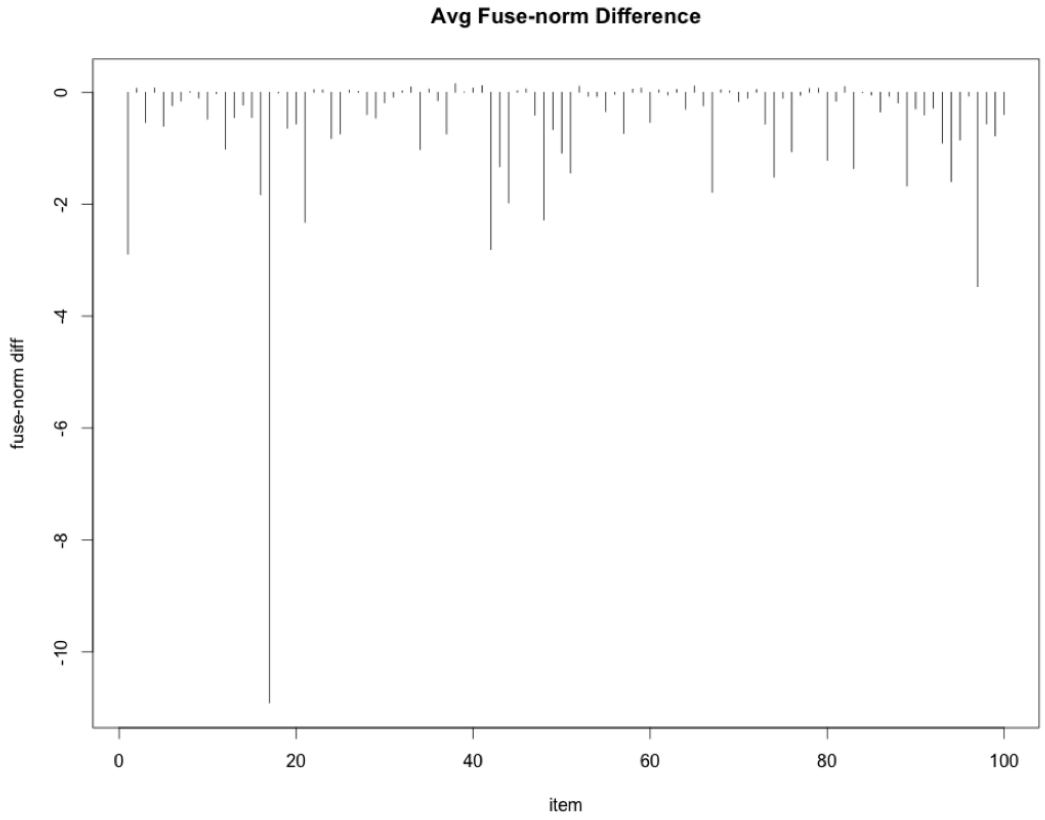


demand observations $\mathbf{d}(T) = [d_1, ..., d_T]$ regarding to a specific item for T periods. The firm is then able to minimize the sample average cost $\hat{\mathbf{C}}(q; \mathbf{d}(T))$ by choosing the right quantity q based on $\mathbf{d}(T)$ only:

$$\begin{aligned} \min_{q \geq 0} \hat{\mathbf{C}}(q; \mathbf{d}(T)) &= \frac{1}{T} \sum_{t=1}^{T} [b(d_t - q)^+ + h(q - d_i)^+] \\ &\equiv \\ \min_{u_t, o_t, q \geq 0} \hat{\mathbf{C}}(q; \mathbf{d}(T)) &= \frac{1}{T} \sum_{t=1}^{T} [bu_t + ho_t] \\ &\text{s.t.} \\ &u_t \geq d_t - q, \forall t \in T \\ &o_t \geq q\_d_t, \forall t \in T \\ &u_t, o_t \geq 0, \forall t \in T \end{aligned} \tag{2.11}$$

With the order statistics of historical demand observations $d(t)$, Bertsimas and Thiele

Figure 2.10: Comparison of fuse-Lasso norm across items



(2005) [14] shows that the optimal order quantity satisfies:

$$q^* = d_{(j)}$$

where

$$j = \lceil \frac{b}{b+h} T \rceil.$$

However, when the historical demand data contains missing data, if the firm treats missing data as zero demands, the above approach provides a lower ordering quantity and putting the the firm at a higher risk of stock out.

Basic Model

The demands of items in the same category are usually highly correlated. For example, in the category of tablets, the demands of Apple iPad and Microsoft Surface are influenced by several same factors: holidays, promotions and technology advances, etc. However, due to

various reasons, such as system down, stock out and zero demand, there will be no demand observations (sales) for some items and some periods. When there is missing data of a item, the demand observations of other items can provides some hints of the the missing demands. Meanwhile, impulsive purchase and sporadic promotions of some items brings noises in the observations. In the following model, we jointly identify the noise, recover the missing data and optimize the order quantity.

Given sales observation $\mathcal{S}$ for I items in the same category over T periods and is organized as a matrix below:

$$\mathcal{S} = \begin{pmatrix} s_{11} & \dots & s_{1T} \\ \vdots & s_{it} & \vdots \\ s_{I1} & \dots & s_{IT} \end{pmatrix}$$

where s_{it} is the sales observation of item i in period t.

Following the convention in geo-demand model, let Ω denote the the indices of positive sales. Similarly, let $\mathcal{D}$ be the "true" underlying demand matrix. We can write

$$\mathcal{D} = \begin{pmatrix} d_{11} & \dots & d_{1T} \\ \vdots & d_{it} & \vdots \\ d_{I1} & \dots & d_{IT} \end{pmatrix}$$

where d_{it} is the "true" demand of item i in period t.

The difference between the "truth" and observation can be summarized by the noise matrix ε and the following equation holds:

$$\mathcal{D} + \varepsilon = \mathcal{S}$$

Specifically, for the positive sales observations $s_{it} > 0$, we are able to identify the noise e_{it} such that

$$d_{it} + e_{it} = s_{it}, \forall (i,t) \in \Omega$$

By introducing the linear projection operator $\mathcal{A}_\Omega$: $\mathbb{R}^{I \times T} \to \mathbb{R}^m$ that selects the set of m elements of positive sales (Ω) from $\mathcal{S}$, we can simply write:

$$\mathcal{A}_\Omega(\mathcal{D} + \varepsilon) = \mathcal{S}_\Omega$$

Following Candès et al. (2011) [21], we use the nuclear norm and the L_1 norm to replace the rank and cardinality functions. That is, the low rankness for the "true" demand matrix $\mathcal{D}$ is enforced by $\|\mathcal{D}\|_* \leq \lambda_r$, where λ_r is the low rankness parameter and the regularization over noise is expressed by $\lambda_e \|\varepsilon\|_1$ where λ_e is the corresponding regularization parameter.

We are then able to write the data-driven newsvendor problem with missing data as:

$$\min_{\mathcal{D},\varepsilon,q_i,u_{it},o_{it}} \sum_{i=1}^{I} \frac{1}{T} \sum_{t=1}^{T} [bu_{it} + ho_{it}] + \lambda_e \|\varepsilon\|_1 \tag{2.12}$$

$$\begin{aligned} \text{s.t.} \\ & \mathcal{A}_\Omega(\mathcal{D} + \varepsilon) = \mathcal{S}_\Omega \\ & \|\mathcal{D}\|_* \leq \lambda_r \\ & u_{it} \geq d_{it} - q_i, \forall i, t \\ & o_{it} \geq q_i - d_{it}, \forall i, t \\ & u_{it}, o_{it} \geq 0, \forall i, t \\ & \mathcal{D} \geq 0 \end{aligned}$$

To solve the optimization problem in Equation (2.12), an alternating direction multiplier method (ADMM) may be developed in the light of Candès et al. (2011) [21]. Through careful examination of the formulation, we see that the minimization objective in the expected profit is in line with the low rankness of the "true" demand where the demand are assumed to be smoother than the sales we observed. The following proposition depicts the fact that the data-driven model with missing data in (2.12) preserved the optimal order quantity in the classic model in (2.11).

Proposition 4. *For any feasible solution to the recovered demand $\mathcal{D} \geq 0$, the optimal order quantity q_i^* for item i satisfies the critical fractile $\frac{b}{b+h}$ of the ordered statistics $d_{(it)}$ for each i. That is,*

$$q_i^* = d_{(it^*)}, \textit{where } t^* = \lceil \frac{b}{b+h} T \rceil.$$

When the regularization parameter $\lambda_e \to \infty$, the sales observations are treated as "true" demand observation without recovery and the noise matrix ε is set to zero. As a result, the model (2.12) becomes equivalent to the classic data-driven newsvendor model (2.11).

2.4 Summary

In this chapter, we present two models that deals with sales data in the context of online retailer. An ADAL method for robust tensor recovery in high dimensional data and ADMM in matrix completion are proposed in the two models respectively.

In the first model, we study the missing geo-demand data completion problem for a national online retailer. We formulate the problem as a low-rank tensor recover problem in a convex optimization framework. An alternating direction augmented Lagrangian (ADAL)

method has been developed and tailored for solving the tensor recovery problem with partial observations. We first discuss efficiency and effectiveness of the algorithm via experiments with synthetic data. We then apply the framework with observed geo-demand from the online retailer. Finally, the benefits of the missing geo-demand data completion are summarized based on computational experiment results. We show that the recovered geo-demand distributions possesses more smoothness over time and rendered better generalization performance than the observed geo-demand upon integrated into the existing learning framework.

We also integrate the missing data recovery with the data-driven newsvendor model which provides estimation of demands as well as optimal order quantity. A preliminary analysis shows that the proposed model preserves the condition for optimal order quantity as it is in the data-driven newsvendor model. An ADMM algorithm may be developed as a solving solution to the formulation. As a future work, the choices of key parameters, e.g. λ_r and λ_e, can be discussed.

Chapter 3

Continuous Process Systems with Flexible Recipes

3.1 Introduction

Oil consumption has been escalating in the past decades, especially in emerging economies regions, such as Asia. Meanwhile, the dramatically increasing oil price is impeding the growth of the world economy. Despite its increasing trend, oil price also exhibits high volatility. After it reached the record peak US$ 145 in July 2008, it fell significantly to US$ 30.28 a barrel on December 23, 2008. Such increasing trend together with jumps of prices also prevails in other commodities over the past decades as shown in Figure 3.1. This phenomenon leads to higher manufacturing costs as well as more difficulties in supply chain management under price uncertainty among many industries.

Facing such challenges, joint inventory investment and allocation decision making becomes an important tool that makes the manufacturing systems robust. Consider an oil refinery that converts crude oil into profitable petroleum products such as gasoline, diesel, kerosene, heating oil and asphalt. Those products are actually inputs for further manufacturing processes. Generally, it operates in 3 phases: crude oil unloading and blending, fractionation and reaction processes and product blending and shipping. In the first phase, crude oil of different grades is transported by crude oil marine vessels from different regions. Since the properties of crude oil highly depend on its origins, there are usually dedicated storage tanks for crude oil of different grades. In many situations, before crude oil enters distillation, the first step of production, different grades are blended to achieve certain properties, such as viscosity and density, in order to meet the production requirements.

The manufacturing process presented above belongs to *continuous process* (a.k.a. *batch process*) that primarily schedules short production runs of products [32]. Continuous process industries often obtain their raw materials from mining or from agricultural industries. These

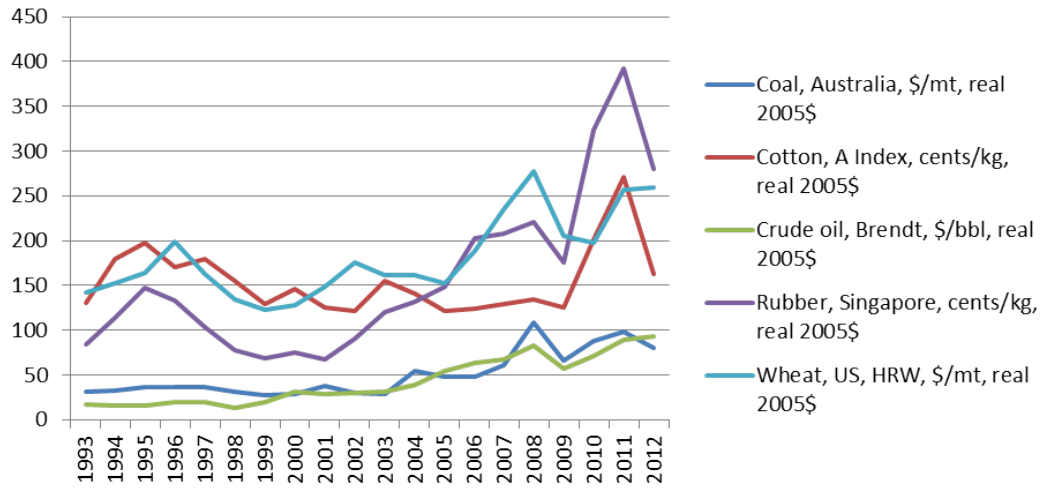


Figure 3.1: Selected Commodity Prices in Past 20 Years.

raw materials have natural variations in quality [78]. Some common continuous processes can be found in fields such as oil refining, agricultural, chemicals and fertilizers. Schuster and Allen (1998) [82] illustrates how Welch's Inc manages grape-processing among plants using linear program models. In that case, grapes are usually processed in plants located near growing areas. To maintain national consistency, Welch's often transfers juice for blending between plants. The selection of recipes is a key decision that affects the profitability via both operational costs and production capacity. The nature of variations in both raw material quality and market conditions often lead to the variations in the recipes. Such recipe flexibility is not on design but on the operation that allows adjustments of recipe items aiming to achieve better performance than traditionally fixed recipes. Here, *flexible recipe* refers mainly to the adjustments of recipe items as input of continuous process in response to market conditions, i.e. demand arrivals.

In this work, we simplify the system by considering three types of goods: raw materials, ingredients and final products. In the grape-processing case, we regard grapes from different growing areas as raw materials, intermediate juice of different concentration as ingredients and packaged juice on market as final products. The continuous process is simplified into two phases: separation and blending. Since different raw material grades have various concentration of desired ingredients, in the separation stage, those ingredients are separated first from raw materials. Then a combination of ingredients are blended into final products that meet certain specifications.

Literature

There are mainly two streams of literature related to our work. In the first stream, given the structure of the system, optimal investment decisions are analyzed under demand uncertainty and/or inventory procurement cost variability. Fine and Freund (1990) [40] investigate the optimal capacity investment problem in single period by two-stage stochastic programming with discrete demand distribution. Their study focuses on the case with two products, two dedicated resources, one flexible resource. While in our study, we allow the demand distribution to be unknown and the system handles multiple raw materials and final products. Following the lead of Fine and Freund (1990) [40], under a two-product firm, Van Mieghem (1998) [93] analyze the optimal investment in flexible resources as a function of margins, costs and multivariate demand uncertainty. Contrary to the previous work, they show that it can be advantageous to invest in flexible resource even with perfectly positively correlated product demands. Harrison and Van Mieghem (1999) [48] study the optimal investment strategy with a multi-dimensional newsvendor model and conclude a critical fractile property for the optimal investment levels. Given the structure of an assemble-to-order system, Akçay and Xu (2004) [4] formulate the join inventory replenishment and component allocation problem into a two-stage stochastic program and propose an order-based component allocation rule for the second stage problem.

In the second stream, the applications of flexible recipes in continuous processes are mostly studied. Rutten and Bertrand (1998) [78] study the balancing of safety stock costs and recipe flexibility costs for continuous industries with high customer service requirements. They conclude that under certain circumstances the use of recipe flexibility can lead to lower costs when compared to using fixed recipes. Keesman (1993) [54] investigate the application of flexible recipes for continuous process optimization and applies an adaptive feedforward control strategy for a priori known disturbances in the process inputs. Furthermore, a new framework that fully exploits the inner flexibility of continuous processes at the plant level is developed by Romero et al. (2003) [76]. Their framework considers a continuous recipe model that interacts with a plant-wide model to constitute the flexible recipe model. The most related work to ours, under continuous process manufacturing, is done by Karmarkar and Rajaram (2001) [53]. They formulate the grade selection and blending problem as a nonlinear mixed-integer program with fixed cost for grade selection and inventory holding cost. However, they assume the annual demand is known and constant for each final products.

Our work is different from the literature mainly in two ways. First, in our model, the recipe flexibility is embedded in the operations of continuous processes, rather than the system design as seen in literature on process flexibility. Second, we study the decisions of inventory investment, recipe selection and resource allocation in an integrated model.

3.2 An Application Example

In this section, we briefly illustrate the application of flexible recipes in continuous process. Consider a manufacturer, i.e. refinery or food processing factory, whose operations can be categorized as separation and blending stages. There are 3 final products made from 3 raw materials. The raw material inventory is given as $Z = (z_1 = 200, z_2 = 300, z_3 = 400)$ units, where z_i is the inventory of raw material i. The raw material cost is $C = (c_1 = 6, c_2 = 4, c_3 = 3)$ dollars per unit, where c_i is the purchase cost of raw material i per unit. In the continuous process, raw materials are separated first into 3 ingredients, depending on their concentration in raw materials. The ingredient concentration matrix for raw materials is

$$P = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

where row i represents raw material i and column j represents ingredient j.

The element on row i and column j, denoted by p_{ij}, is the proportion of ingredient j contained in a unit of raw material i. Then in the blending stage, final products are blended from those ingredients. The ingredient requirement matrix for final products is

$$A = \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.7 & 0.2 & 0.1 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$

where row k represents final product k and column j represents ingredient j. The element on row k and column j, denoted by α_{kj}, is the proportion of ingredient j required in a unit of final product k.

In a system that implements fixed recipes, it determines the optimal fixed recipe before any demand arrivals. The resulting optimal fixed recipe in this case is:

$$B = \begin{pmatrix} 0.53 & 0.44 & 1.03 \\ 0.40 & 0.18 & 1.29 \\ 0.32 & 0.11 & 1.22 \end{pmatrix}$$

where row k represents final product k and column i represents raw material i. This matrix is similar to BOM. That is, the element on raw k and column i, denoted by b_{ki}, is the amount of raw material i required in a unit of final product k.

For simplicity, we assume that the demand for each final product follows Bernoulli distribution with 0.5 probability equals 100 units and 0.5 probability equals 200 units. When the system sees demand arrivals, it determines the optimal flexible recipes that maximize its revenue with $R = (r_1 = 10, r_2 = 8, r_3 = 6)$ dollars, where r_k represents the revenue of a unit

Table 3.1: Expected Profit Summary (in dollars)

	Flexible Recipes ($R1$)	Fixed Recipes ($R2$)	Improvements= $\frac{R1-R2}{R2}$
Current Setting	3525	2904.6	21.36%
Demand $\times 2$	4350	3199.8	35.95%
Demand $\times 0.8$	2880	2643.7	8.94%

final product k.

Table 3.1 summarizes the computational results of expected profit and shows that the flexible recipes enable the system to achieve higher profit via better resource utilization, especially under the presence of large demand variance.

3.3 The Generic Model

As shown in Figure 3.2, the continuous manufacturing process consists of two stages: in the separation stage, raw materials are processed into a set of ingredients(or intermediate products); in the blending stage, a selection of ingredients are blended into final products to fulfill the demands. In oil refinery industry, for instance, the raw materials are different crude oil grades. The three most quoted oil grades are North America's West Texas Intermediate crude (WTI), North Sea Brent Crude, and the UAE Dubai Crude.[1] Depending on the mixture of hydrocarbon molecules, crude oil varies in color, composition and consistency. Different oil-producing areas yield significantly different varieties of crude oil.[2] The ingredients are the intermediate products such as light ends, naphtha, kerosene, distillate, atmospheric residua, vacuum gas oil and vacuum residua. Final products are various gasoline types, lubricants, petrochemicals, diesel, asphalt, etc,.

Our model considers a single period inventory investment problem. Hence, there is no inventory cost. Moreover, there is no salvage value of leftover raw materials. We also assume linear cost structure. There are basically three types of costs: raw material procurement cost, grade selection cost and final product revenue. Similar to the definition by Karmarkar and Rajaram(2001) [53], the grade selection cost incurs when a grade is selected by a recipe into separation stage and varies by grades. Hence, the assessment of grade selection cost highly depends on the categorization of separation stage. In the oil refinery case, the grade selection cost can be the fixed set-up cost for the machine to process particular grade or operating cost of pre-process blending and transportation if applicable. Raw material procurement cost can be the price from real option or spot market. On the other hand, in grape-processing case, if we see the concentrated juice (intermediate products) in different growing areas as raw materials, the selection cost can be defined as the transfer cost among

[1] http://en.wikipedia.org/wiki/List_of_crude_oil_products

[2] http://pascagoula.chevron.com/home/abouttherefinery/whatwedo/typesofcrudeoil.aspx

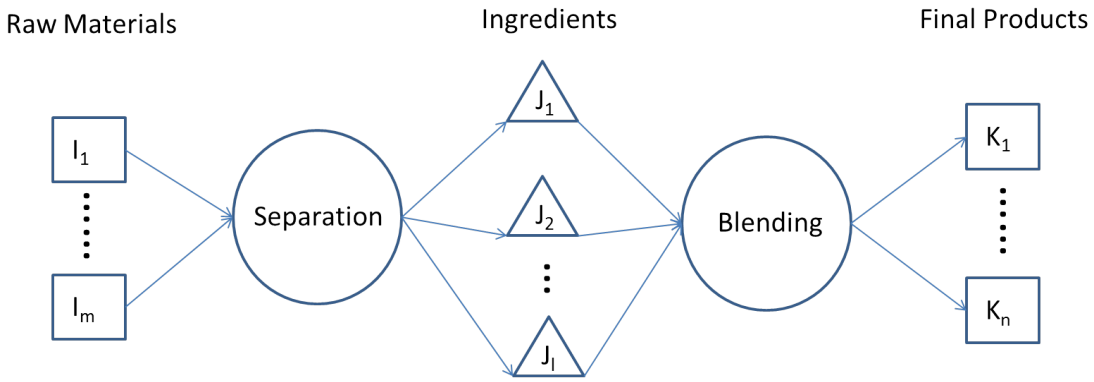


Figure 3.2: Simplified Continuous Process System

those areas. The model we propose is able to optimize investment on raw materials together with flexible recipe selection for production when the system sees demands. It can be easily extend to the systems with many separation and blending stages in serial structure.

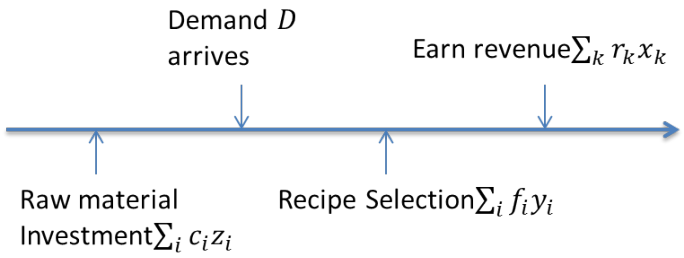


Figure 3.3: Event Sequence Diagram

The event sequence is illustrated in Figure 3.3. The system first invests in raw material inventory with total procurement cost $\sum_i c_i z_i$, where c_i is the unit cost of raw material i and z_i is the amount of raw material i purchased. After the system sees the demand arrivals $D = (d_k)$, where d_k is the demand of final product k, it then selects the optimal recipe to satisfy the demands. Under the presence of grade selection cost, it is not always profitable to satisfy as much demand as possible. As discussed above, the recipe selection cost is the sum of grade selection costs, given by $\sum_i f_i y_i$, where f_i is the grade selection cost of raw material i and y_i is the selection decision of raw material i. Lastly, the fulfillment of demands generates total revenue $\sum_k r_k x_k$, where r_k is the unit revenue of final product k and x_k is the fulfilled demand of final product k. Another assumption is that the material loss in separation and

blending stages are negligible. Indeed, if the loss is consistent and fractional in the process, we can introduce some discount factors in the formulation so that the structure of the model is still preserved.

The decision making process for the system with flexible recipes differentiates itself from the newsvendor model with the postponement in final product blending. In the newsvendor model, the final products are produced ahead and the total manufacturing cost can be assessed before demand arrivals. As a result, in the newsvendor model, the critical fractiles and subsequently the raw material inventory levels can be determined in advance. While in a continuous process with flexible recipes, the system makes tradeoffs between revenue and recipe selection cost after demand arrivals.

We formulate the generic model as a two-stage stochastic program:

$$\Pi = \max \mathbb{E}_D \pi(Z, D) - \sum_i c_i z_i \tag{3.1}$$

where

$$\pi(Z, D) = \max \sum_k r_k x_k - \sum_i f_i y_i \tag{3.2}$$

s.t.

$$A(x_1, x_2, ..., x_K) \leq P(z_1 y_1, z_2 y_2, ..., z_I y_2) \tag{3.3}$$

$$x_k \leq d_k, \forall k \in K \tag{3.4}$$

$$y_i \in \{0, 1\}, \forall i \in I \tag{3.5}$$

Uppercase is for vectors while lowercase is for scalars. The first stage of the formulation maximizes the expected total profit of the system. Given raw material inventory vector $Z = (z_1, z_2, ..., z_I)$ and demand arrival vector $D = (d_1, d_2, ..., d_K)$, the second stage with recourse maximizes the total revenue minus total grade inclusion cost as shown in subproblem (3.2). In constraint (3.3), $A(x_1, x_2, ..., x_K)$ is a material transformation function that calculates the ingredient requirement vector given the demand fulfillment vector $X = (x_1, x_2, ..., x_K)$. $P(z_1 y_1, z_2 y_2, ..., z_I y_2)$ is another material transformation function that calculates ingredient supply vector given the recipe selection $Y = (y_1, y_2, ..., y_I)$ and raw material inventory Z. Because the initial inventory investment is sunk cost, $(z_1 y_1, z_2 y_2, ..., z_I y_I)$ shows that the system chooses to use up all selected raw materials. This constraint states that the supply of each ingredient must be no less than the ingredient requirement for producing X amount final products. Constraint (3.4) represents that demand fulfillment can not exceed the demand arrivals. The recipe selection is determined by the binary decision vector Y expressed in constraint (3.5).

3.4 Analysis of Special Cases

The generic model (3.1) is a Stochastic Mixed Integer Program (SMIP) which may not be easy to solve. In this section, we study some special cases to explore structural properties of the system.

Case 1: System with Zero Grade Selection Cost

In the absence of grade selection cost, we can eliminate the binary decision variables in the formulation as all available raw materials will be considered in the recipe and result in the following linear program.

$$\Pi = \max_{Z \geq 0} \mathbb{E}_D \pi(Z, D) - C^T Z \tag{3.6}$$

where

$$\begin{aligned} \pi(Z, D) = \max & \sum_{k \in K} r_k x_k \\ & \text{s.t.} \\ & \sum_{k \in K} x_k \alpha_{kj} \leq \sum_{i \in I} z_i p_{ij}, \forall j \in J \\ & x_k \leq D_k, \forall k \in K \end{aligned} \tag{3.7}$$

Lemma 2. *Given the raw material inventory Z and demand realization D, the subproblem $\pi(Z, D)$ is a concave function in Z.*

Proof. Consider the subproblem (3.7), suppose X^1 and X^2 are the optimal solutions to $\pi(Z^1, D)$ and $\pi(Z^2, D)$ respectively. Then, $\alpha X^1 + (1-\alpha)X^2$ is a feasible solution to $\pi(\alpha Z^1 + (1-\alpha)Z^2, D)$. As a result, $\alpha\pi(Z^1, D) + (1-\alpha)\pi(Z^2, D) \leq \pi(\alpha Z^1 + (1-\alpha)Z^2, D)$, since $\pi(\alpha Z^1 + (1-\alpha)Z^2, D)$ is a maximization problem. □

In fact, the flexible recipe system with zero grade inclusion cost can be transformed to the multi-dimensional newsvendor model presented in Van Mieghem (1999) [48]. Their model optimizes the capacity investment when the firm is facing random demands. If we view the ingredient amounts as the capacity investment L for final product manufacturing, let

$$P^T Z = L$$

Problem (3.6) can then be rewritten as:

$$\Pi = \max_{L \geq 0} \mathbb{E}_D \pi(L, D) - C(L)$$

where

$$\begin{aligned} \pi(L, D) = \max\ & R^T X \\ & \text{s.t.} \\ & A^T X \leq L \\ & X \leq D \end{aligned}$$

and

$$\begin{aligned} C(L) = \min\ & C^T Z \\ & \text{s.t.} \\ & P^T Z \geq L \\ & Z \geq 0 \end{aligned}$$

It is well-known that $C(L)$ is convex. Though the results in Van Mieghem (1999) [48] are derived from linear investment cost function, they can be directly generalized to any convex function $C(L)$.

Problem (3.6) is essentially a two-stage stochastic linear program. Therefore, by Lemma 2, the optimality condition for problem (3.6) is:

$$C \in \nabla \mathbb{E}_D \pi(Z, D)$$

where $\nabla \mathbb{E}_D \pi(Z, D)$ is the subgradient of $\mathbb{E}_D \pi(Z, D)$.

Since $\pi(Z, D) < \infty$ by the finiteness of demands, the problem has relatively complete recourse. Use the argument (Corollary 12, pp.96) in [16], we can then interchange differentiation and integration to rewrite the optimality condition as:

$$C \in \mathbb{E}_D \nabla \pi(Z, D)$$

where $\nabla \pi(Z, D)$ is the subgradient of $\pi(Z, D)$.

The subgradient $\nabla \pi(Z, D)$ for given Z and D is simply the optimal shadow prices given by the dual of constraint 3.7:

$$\begin{aligned} \min\ & \lambda^T P^T Z + \mu^T D \\ & \text{s.t.} \\ & A\lambda + \mu \geq R \end{aligned} \tag{3.8}$$

The subgradient of the optimal profit function $\pi(Z, D)$ with respect to Z is:

$$\nabla \pi(Z, D) = P\lambda(Z, D)$$

where $\lambda(Z, D)$ is the optimal solution to the dual problem (3.8).

Following the statement in Harrison and Van Mieghem (1999) [48], we have a similar result below.

Proposition 5. *Let the demand D is continuous and finite with probability 1, the expected profit function $\mathbb{E}_D\pi(Z, D)$ is differentiable. Therefore, the optimality condition is given by:*

$$P\mathbb{E}_D\lambda(Z, D) = C \tag{3.9}$$

Case 2: Single Final Product

When the firm only sells single final product, we only need to consider raw material investment decision for that final product. We call a system decentralized if the system is decomposed into several flexible recipe systems with single final product. The total optimal raw material investment is simply the sum of investment for each decentralized system. As a result, though it does not provide full flexibility, the decentralized system can be viewed as an approximation to the original system with flexible recipe. Apparently, such ignorance leads to higher raw material investment requirement to maintain same service level as that of original system.

Consider the flexible recipe system with single final product (FRSF), the corresponding optimal expected profit is:

$$\Pi = \max_Z \mathbb{E}_d\pi(Z, d) - C^T Z \tag{3.10}$$

where

$$\pi(Z, d) = \max rx - \sum_{i\in I} f_i y_i \tag{3.11}$$

$$\text{s.t.}$$

$$\alpha_j x \leq \sum_{i\in I} p_{ij} z_i y_i, \forall j \in J \tag{3.12}$$

$$x \leq d \tag{3.13}$$

$$y_i \in \{0, 1\}, \forall i \in I$$

Proposition 6. *The second stage allocation optimization in FRSF is NP-complete.*

Proof. First, we argue that the FRSF is in NP, since given a solution $(x, y_1, ..., y_I)$, a certifier can efficiently check that x is no greater than d and $\alpha_j x$ is no greater than $\sum_{i\in I} p_{ij} z_i y_i$ for all $j \in J$ in $|J| + 1$ time. Moreover, set cover problem is a special instance of the FRSF by setting the revenue per unit to a large number (i.e.1000), the grade selection cost and

demand to one, and $\beta_{ij} = \frac{p_{ij}z_i}{\alpha_j}$ to either zero or one. The second stage formulation is reduced to the set cover problem:

$$\begin{aligned} &\min \sum_{i\in I} y_i \\ &\text{s.t.} \\ &\quad 1 \le \sum_{i\in I} \beta_{ij} y_i, \forall j \in J \\ &\quad y_i \in \{0,1\}, \forall i \in I \end{aligned}$$

It is known that set cover problem is NP-complete. The reduction indicates that second stage allocation optimization in FRSF is NP-complete. □

Lemma 3. *In a flexible recipe system with single final product, let $I' = \{i \in I | z_i > 0\}$, for any $I'' \subseteq I'$, there is $r * \min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I''} f_i \le r * \min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I'} f_i$. That is, in second stage allocation optimization, the maximum profit that I' generates is always larger than the maximum profit that its subset I'' generates.*

Proof. Proof by contradiction. Suppose $r*\min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I''} f_i \ge r*\min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I'} f_i$. Since $I'' \subseteq I'$, we have $\min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\} \le \min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\}$. If $d \le \min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\}$, we use at most I'' to satisfy the demand. If $d \in [\min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\}, \min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\}]$, we still use I'' to satisfy the demand, because $r * d - \sum_{i\in I'} f_i \le r * \min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I'} f_i \le r * \min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I''} f_i$. If $d > \min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\}$, the set I'' is again preferable over I', because $r * \min_j\{\frac{\sum_{i\in I'} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I'} f_i \le r * \min_j\{\frac{\sum_{i\in I''} p_{ij}z_i}{\alpha_j}\} - \sum_{i\in I''} f_i$. This implies that we never select raw material $i \in I' \setminus I''$. We should then set $z_i = 0, \forall i \in I' \setminus I''$. This reduces I' to I''. Contradiction. Therefore, we conclude the result above. □

Lemma 3 implies that if the system chooses or has to not fully satisfy the demand, it selects a set of raw materials that outperforms all its subset.

Two Raw Materials and One Final Product

We start with the most simple case to investigate the role of flexible recipes in continuous process. The system we consider here consists of only 2 kinds of raw materials and 1 final product. The two-stage stochastic program can be explicitly expressed as:

$$\Pi = \max \mathbb{E}_d \pi(z_1, z_2, d) - c_1 z_1 - c_2 z_2$$

where

$$\begin{aligned}\pi(z_1, z_2, d) = &\max rx - f_1y_1 - f_2y_2\\ &\text{s.t.}\\ &\alpha_j x \le \sum_{i=1,2} p_{ij}(z_i y_i), \forall j \in J\\ &x \le d,\\ &y_i \in \{0,1\}, i = 1,2\end{aligned}$$

Let the upper bound of the demand be $\bar{d}$ and define $\rho_1 = \min_j\{\frac{p_{1j}}{\alpha_j}\}$ and $\rho_2 = \min_j\{\frac{p_{2j}}{\alpha_j}\}$. Without loss of generality, we further assume that $f_1 \le f_2$. Based on Lemma 3, we have

$$\max\{r\rho_1 z_1 - f_1, r\rho_2 z_2 - f_2\} \le r * \min_j\{\frac{p_{1j}z_1 + p_{2j}z_2}{\alpha_j}\} - f_1 - f_2 \tag{3.14}$$

This result enables us to further summarize the optimal selection of raw materials in the second stage allocation optimization based on the regions of z_1 and z_2 and the realized demand intervals.

Table 3.2: Optimal Selection in System with Two Raw Materials and One Final Product

Regions	Boundary	Optimal Selection for demand intervals
I	$\{r\rho_1 z_1 - f_1 < 0,$ $r\rho_2 z_2 - f_2 > 0\}$	$d \in [0, \frac{f_2}{r}] \leftarrow \emptyset$ $d \in [\frac{f_2}{r}, \rho_2 z_2 + \frac{f_1}{r}] \leftarrow \{2\}$ $d \in [\rho_2 z_2 + \frac{f_1}{r}, \bar{d}] \leftarrow \{1,2\}$
II	$\{r\rho_1 z_1 - f_1 > 0,$ $r\rho_2 z_2 - f_2 < 0\}$	$d \in [0, \frac{f_1}{r}] \leftarrow \emptyset$ $d \in [\frac{f_1}{r}, \rho_1 z_1 + \frac{f_2}{r}] \leftarrow \{1\}$ $d \in [\rho_1 z_1 + \frac{f_2}{r}, \bar{d}] \leftarrow \{1,2\}$
III	$\{r\rho_1 z_1 - f_1 > 0,$ $r\rho_2 z_2 - f_2 > 0,$ $r\rho_1 z_1 - f_1 > r\rho_2 z_2 - f_2\}$	$d \in [0, \frac{f_1}{r}] \leftarrow \emptyset$ $d \in [\frac{f_1}{r}, \rho_1 z_1 + \frac{f_2}{r}] \leftarrow \{1\}$ $d \in [\rho_1 z_1 + \frac{f_2}{r}, \bar{d}] \leftarrow \{1,2\}$
IV	$\{r\rho_1 z_1 - f_1 > 0,$ $r\rho_2 z_2 - f_2 > 0,$ $r\rho_1 z_1 - f_1 < r\rho_2 z_2 - f_2\}$	$d \in [0, \frac{f_1}{r}] \leftarrow \emptyset$ $d \in [\frac{f_1}{r}, \rho_1 z_1 + \frac{f_2 - f_1}{r}] \leftarrow \{1\}$ $d \in [\rho_1 z_1 + \frac{f_2 - f_1}{r}, \rho_2 z_2 + \frac{f_1}{r}] \leftarrow \{2\}$ $d \in [\rho_2 z_2 + \frac{f_1}{r}, \bar{d}] \leftarrow \{1,2\}$
V	$\{r\rho_1 z_1 - f_1 \le 0,$ $r\rho_2 z_2 - f_2 < 0\}$	$d \in [0, \frac{f_1 + f_2}{r}] \leftarrow \emptyset$ $d \in [\frac{f_1 + f_2}{r}, \bar{d}] \leftarrow \{1,2\}$

In Table 3.2, we can see different roles that each raw material plays in optimal selection. In region I, raw material 2 acts as a primary resource while raw material 1 is supplementary that is only selected together with raw material 1 when the demand is large. Similarly, in

region II and III, raw material 1 acts as a primary resource while raw material 2 is supplementary. In region IV, raw material 1 and 2 serves as primary resource alternatively. In region V, it is profitable only when raw material 1 and 2 are selected together. We conclude that only the raw material with positive value of $r\rho_i z_i - f_i$ can become primary. That is, the primary resource should provide positive profit when it is selected as the only source. When both raw materials give positive value of $r\rho_i z_i - f_i$, the one with smaller f_i serves as primary for small demand and the one with larger $r\rho_i z_i - f_i$ is primary for larger demand before both raw materials are selected.

Given the parameters ρ_1, ρ_2, the region partition is illustrated in Figure 3.4. Here, different regions are separated by lines: $z_1 = \frac{f_1}{r\rho_1}$, $z_2 = \frac{f_2}{r\rho_2}$ and $r\rho_1 z_1 - f_1 = r\rho_2 z_2 - f_2$.

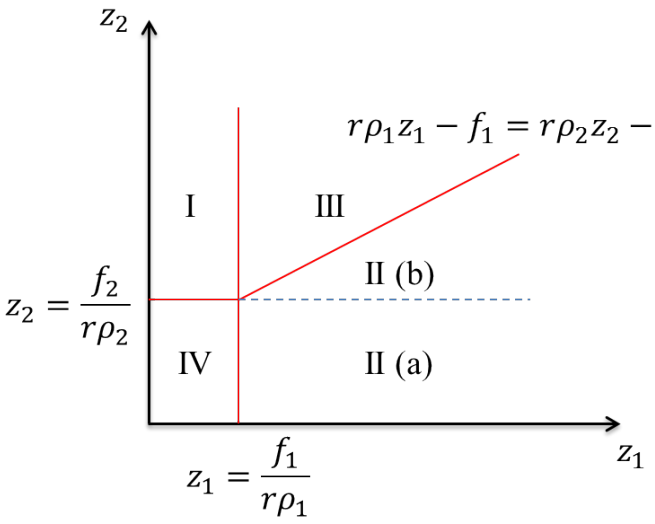


Figure 3.4: Region Partition

Figure 3.5 shows an example of the performance of different investment in raw material 1 and 2. We then try to find the extreme point with highest expected profit.

Suppose that demand is continuously distributed with pdf $\phi(d)$, concave cdf $\Phi(d)$ and support $[0, \bar{d}]$,i.e. truncated Normal distribution. Upon the knowledge of region partition, we try to find the optimal z_1 and z_2. In the analysis, we first define a new decision(control) variable $\rho(z_1, z_2) = \min_j\{\frac{p_{1j}z_1 + p_{2j}z_2}{\alpha_j}\}$ which is concave in z_1 and z_2. Define $\frac{\partial\rho}{\partial z_1}$ and $\frac{\partial\rho}{\partial z_2}$ as the subgradients associated with z_1 and z_2.

Lemma 4. *The subgradients satisfy:* $\frac{\partial\rho}{\partial z_1} \geq \rho_1$ *and* $\frac{\partial\rho}{\partial z_2} \geq \rho_2$.

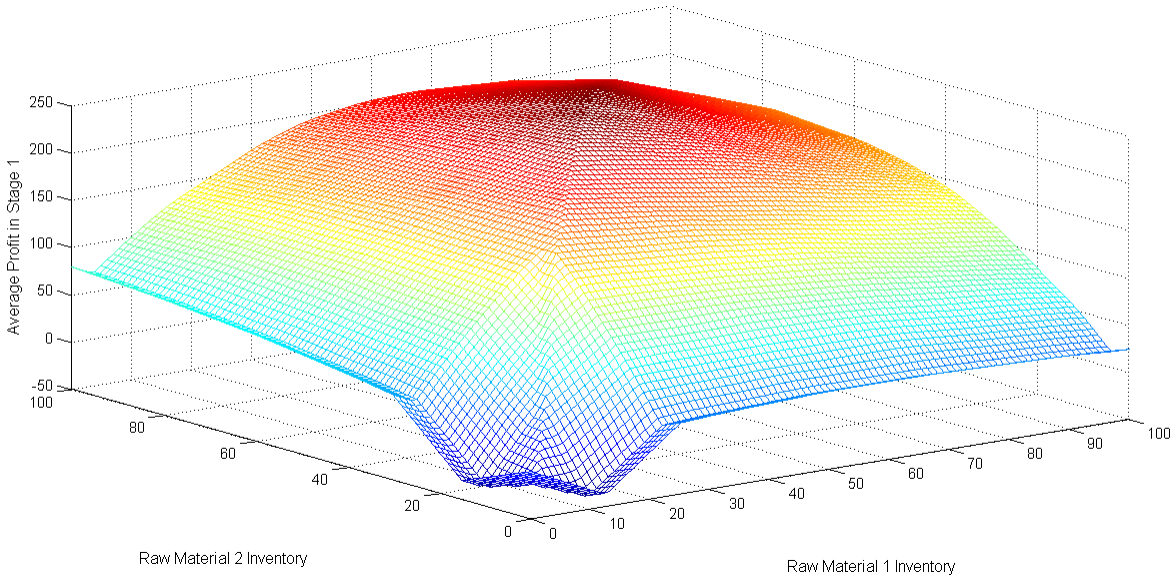


Figure 3.5: Average Profit over All Regions

Proof. Let Δz_1 be a small increase in raw material 1 inventory, $j' = \arg\min_j\{\frac{p_{1j}z_1+p_{2j}z_2}{\alpha_j}\}$ and $j^* = \arg\min_j\{\frac{p_{1j}(z_1+\Delta z_1)+p_{2j}z_2}{\alpha_j}\}$. Then we have

$$\rho(z_1+\Delta z_1, z_2) = \frac{p_{1j^*}(z_1+\Delta z_1)+p_{2j^*}z_2}{\alpha_{j^*}} \geq \frac{p_{1j'}z_1+p_{2j'}z_2}{\alpha_{j'}} + \frac{p_{1j^*}}{\alpha_{j^*}}\Delta z_1 \geq \rho(z_1+z_2)+\rho_1\Delta z_1$$

As a result, $\frac{\partial\rho}{\partial z_1} \geq \frac{\rho(z_1+\Delta z_1,z_2)-\rho(z_1,z_2)}{\Delta z_1} \geq \rho_1$ Similarly, we have $\frac{\partial\rho}{\partial z_2} \geq \rho_2$. □

Lemma 5. Π *is concave in each region.*

Proof. Given the concavity of $\Phi(d)$ and $\rho(z_1, z_2)$, one can write Π explicitly and check the concavity in each region by second order subgradients. □

A Search Algorithm to First Stage Investment Optimization

For the system with two raw materials and one final product, we can determine the optimal solution by solving local optimal solutions in all regions and then determine the global optimal solution simply by comparison. Since the number of region partition grows exponentially in the number of raw materials, it is inefficient to find every optimal solution in each region and then determine the global optimal solution. It is important to determine in which region(s) the global optimal solution can be found. We start the heuristic with sorting of grade selection cost and individual profitability of each raw materials:

Step 1 Sorting the grade selection costs in ascending order, e.g. $f_1 < f_2$;

Step 2 Start from the point $(\frac{f_1}{r\rho_1}, \frac{f_2}{r\rho_2})$. If $\frac{c_1}{\rho_1} < \frac{c_2}{\rho_2}$, search optimal solution $\hat{Z}$ in Region II and III; If $\frac{c_1}{\rho_1} > \frac{c_2}{\rho_2}$, search optimal solution $\hat{Z}$ in Region IV.

Step 3 Compare $\hat{Z}$ with $(z_1^*, 0)$ and $(0, z_2^*)$, where $(z_1^*, 0)$ is the optimal solution assuming use raw material 1 only and $(0, z_2^*)$ is the optimal solution assuming use raw material 2 only and determine the optimal solution.

By Lemma 5, the optimal solution in each region can be solved by its concavity. The explicit expected profit function and its subgradients in each region are derived as below:

1. Region II and III:

$$\begin{aligned}
\Pi(z_1, z_2) =& r\int_{\frac{f_1}{r}}^{\rho_1 z_1} t\phi(t)dt + r\int_{\rho_1 z_1}^{\rho_1 z_1 + \frac{f_2}{r}} \rho_1 z_1 \phi(t)dt \\
&+ r\int_{\rho_1 z_1 + \frac{f_2}{r}}^{\rho(z_1,z_2)} t\phi(t)dt + r\int_{\rho(z_1,z_2)}^{\bar{d}} \rho(z_1, z_2)\phi(t)dt \\
&- f_1(1 - \Phi(\frac{f_1}{r})) - f_2(1 - \Phi(\rho_1 z_1 + \frac{f_2}{r})) - c_1 z_1 - c_2 z_2 \\
\frac{\partial \Pi}{\partial z_1} =& r\rho_1[\Phi(\rho_1 z_1 + \frac{f_2}{r}) - \Phi(\rho_1 z_1)] + r\frac{\partial \rho}{\partial z_1}[1 - \Phi(\rho)] - c_1 \\
\approx& r\rho_1[\Phi(\rho_1 z_1 + \frac{f_2}{r}) - \Phi(\rho_1 z_1)] + r\frac{\rho(z_1 + \epsilon, z_2) - \rho(z_1, z_2)}{\epsilon}[1 - \Phi(\rho(z_1, z_2))] - c_1 \\
\frac{\partial \Pi}{\partial z_2} =& r\frac{\partial \rho(z_1, z_2)}{\partial z_2}[1 - \Phi(\rho(z_1, z_2))] - c_2 \\
\approx& r\frac{\rho(z_1, z_2 + \epsilon) - \rho(z_1, z_2)}{\epsilon}[1 - \Phi(\rho(z_1, z_2))] - c_2,
\end{aligned}$$

where ϵ is a small step size.

2. Region IV:

$$\begin{aligned}
\Pi(z_1,z_2) =& r\int_{\frac{f_1}{r}}^{\rho_1 z_1} t\phi(t)dt + r\int_{\rho_1 z_1}^{\rho_1 z_1+\frac{f_2-f_1}{r}} \rho_1 z_1\phi(t)dt + r\int_{\rho_1 z_1+f_2-\frac{f_1}{r}}^{\rho_2 z_2} t\phi(t)dt \\
&+ r\int_{\rho_2 z_2}^{\rho_2 z_2+\frac{f_1}{r}} \rho_2 z_2\phi(t)dt + r\int_{\rho_2 z_2+\frac{f_1}{r}}^{\rho(z_1,z_2)} t\phi(t)dt + r\int_{\rho(z_1,z_2)}^{\bar{d}} \rho(z_1,z_2)\phi(t)dt \\
&- f_1[\Phi(\rho_1 z_1 + \frac{f_2-f_1}{r}) - \Phi(\frac{f_1}{r})] - f_1(1-\Phi(\rho_2 z_2\frac{f_1}{r})) \\
&- f_2(1-\Phi(\rho_1 z_1 + \frac{f_2-f_1}{r})) - c_1 z_1 - c_2 z_2 \\
\frac{\partial \Pi}{\partial z_1} =& r\rho_1[\Phi(\rho_1 z_1 + \frac{f_2-f_1}{r}) - \Phi(\rho_1 z_1)] + r\frac{\partial \rho}{\partial z_1}[1-\Phi(\rho)] - c_1 \\
\approx& r\rho_1[\Phi(\rho_1 z_1 + \frac{f_2-f_1}{r}) - \Phi(\rho_1 z_1)] + r\frac{\rho(z_1+\epsilon,z_2)-\rho(z_1,z_2)}{\epsilon}[1-\Phi(\rho)] - c_1 \\
\frac{\partial \Pi}{\partial z_2} =& r\rho_2[\Phi(\rho_2 z_2 + \frac{f_1}{r}) - \Phi(\rho_2 z_2)] + r\frac{\partial \rho}{\partial z_2}[1-\Phi(\rho)] - c_2 \\
\approx& r\rho_2[\Phi(\rho_2 z_2 + \frac{f_1}{r}) - \Phi(\rho_2 z_2)] + r\frac{\rho(z_1,z_2+\epsilon)-\rho(z_1,z_2)}{\epsilon}[1-\Phi(\rho)] - c_2
\end{aligned}$$

where ϵ is a small step size.

3. Use raw material i only: The first stage expected total profit

$$\Pi(z_i) = r\int_{\frac{f_i}{r}}^{\rho_i z_i} t\phi(t)dt + r\int_{\rho_i z_i}^{\bar{d}} \rho_i z_i\phi(t)dt - f_i(1-\Phi(\frac{f_i}{r})) - c_i z_i$$

and the the optimal z_i is

$$z_i^* = \max\{0, \frac{\Phi^{-1}(1-\frac{c_i}{r\rho_i})}{\rho_i}\}$$

We compare the results of the heuristic with the solutions obtained by numerical search. In the simulation, we assume that demand is uniformly distributed, $P = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$ and $A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \end{pmatrix}$. The results are summarized as follows:

In Table 3.3, we notice that the gap between the average profit obtained from the heuristic and that from numerical search is small. Moreover, decrease of fixed cost improves the average profit significantly by comparison of Run 1 and Run 3. When the revenue is small, as shown in Run 4, the system's optimal decision is not to satisfy demand at all. Moreover, the fixed costs f are generally positively correlated to purchase costs c. As a result, the case of Run 2 will rarely happen in reality.

Table 3.3: Average Profit of Selected Runs

Parameters:	Run 1	Run 2	Run 3	Run 4
C=(c_1, c_2)	(4,6)	(6,4)	(4,6)	(4,6)
R	20	20	20	10
F=(f_1, f_2)	(50,100)	(50,100)	(20,50)	(50,100)
Heuristic	344.33	320.46	417.20	0
Numerical Search	349.39	343.39	419.37	0
Optimality Gap	1.4%	6.7%	0.5%	0

3.5 The Linear Model

When material transformation processes are linear, the generic model (3.1) can be simplified as a stochastic mixed-integer linear program by writing constraint (3.3) with matrices A and P.

$$\Pi = \max \mathbb{E}_D \pi(Z, D) - \sum_{i \in I} c_i z_i \tag{3.15}$$

where

$$\pi(Z, D) = \max \sum_{k \in K} r_k x_k - \sum_{i \in I} f_i y_i \tag{3.16}$$

s.t.

$$\sum_{k \in K} \alpha_{kj} x_k \leq \sum_{i \in I} p_{ij} z_i y_i, \forall j \in J \tag{3.17}$$

$$x_k \leq d_k, \forall k \in K \tag{3.18}$$

$$y_i \in \{0, 1\}, \forall i \in I \tag{3.19}$$

Similar to the generic model, the objective of the basic model is also to achieve the maximum expected profit expressed in objective function (3.15). Given Z and D, the system maximizes its profit in the second-stage subproblem (3.16). Under the two assumptions, constraint (3.17) is an explicit expression of constraint (3.3). Here α_{kj} is the amount of ingredient j required to produce a unit of final product k and p_{ij} is the amount of ingredient j contained in a unit of raw material i. The remaining constraints are the same as those in the generic model.

A Heuristic to Second Stage Allocation Optimization

According to Proposition 6, the second stage allocation problem is NP-hard. Hence, a heuristic solution might be needed when the system is of large scale. In the following, we present two heuristics for the mixed-integer subproblem.

Lagrangian Relaxation Algorithm

The second stage is a fixed-charge problem with fixed cost of selection f. We first propose a lagrangian relaxation algorithm. Relaxing constraint (3.17) provides the resulting Lagrangian subproblem:

$$\begin{aligned}\min_{\lambda_j \geq 0, \forall j \in J} \max & \sum_{k \in K} r_k x_k - \sum_{i \in I} f_i y_i + \sum_{j \in J} \lambda_j (\sum_{i \in I} w_{ij} y_i - \sum_{k \in K} x_k \alpha_{kj}) \\ & = \sum_{k \in K} (r_k - \sum_{j \in J} \alpha_{kj} \lambda_j) x_k + \sum_{i \in I} (\sum_{j \in J} \lambda_j w_{ij} - f_i) y_i \\ \text{s.t.} & \\ & x_k \leq d_k, \forall k \in K \\ & y_i \in \{0,1\}, \forall i \in I\end{aligned}$$

where $w_{ij} = p_{ij} z_i$.

By observation, we are able to identify the opportunity to separate the problems for x_k and y_i respectively into the following subproblems:

1. Subproblem 1:

$$\begin{aligned}\max & \sum_{k \in K} (r_k - \sum_{j \in J} \alpha_{kj} \lambda_j) x_k \\ \text{s.t.} & \\ & x_k \leq d_k, \forall k \in K\end{aligned}$$

   Optimal solution: if $r_k - \sum_{j \in J} \alpha_{kj} \lambda_j \geq 0$, set $x_k = d_k$; Otherwise, set $x_k = 0$.

2. Subproblem 2:

$$\begin{aligned}\max & \sum_{i \in I} (\sum_{j \in J} \lambda_j w_{ij} - f_i) y_i \\ \text{s.t.} & \\ & y_i \in \{0,1\}, \forall i \in I\end{aligned}$$

   Optimal solution: if $\sum_{j \in J} \lambda_j w_{ij} - f_i \geq 0$, set $y_i = 1$; Otherwise, set $y_i = 0$.

The remaining part of the algorithm follows the standard lagrangian relaxation algorithm to update the lagrangian multipliers.

"Greedy Add" Algorithm

The "greedy add" algorithm iteratively improves the solution by gradually adding a raw material into production and achieve profit increments by improving demand fulfillments.

The algorithm starts with no selection of raw materials. In each iteration, it does a what-if analysis to find the selection of raw material that increases the profit the most. The termination condition is that all demands are satisfied or all raw materials are selected or no further improvement can be obtained. The detailed algorithm is as follows:

- Iteration 0: initialize $t = 0$, $S_t = I$ and $y_i = 0, \forall i \in S_t$; Since there is no raw material included into production, $\pi^0(Z, \mathbb{D}) = 0$.
- Iteration $t = t + 1$: set $S_i^t = S^{t-1} \cup \{i\}$ where $i \in I \setminus S^{t-1}$; For all S_i^t, solve the continuous knapsack problem,

$$\begin{aligned} \pi_i^t(Z, \mathbb{D}) = \max & \sum_{k \in K} r_k x_k - \sum_{i \in S_i^t} f_i \\ & \text{s.t.} \\ & \sum_{k \in K} x_k \alpha_{kj} \leq \sum_{i \in S_i^t} z_i p_{ij}, \forall j \in J \\ & x_k \leq D_k, \forall k \in K \end{aligned}$$

- If $\max\{\pi_i^t(Z, \mathbb{D}), \forall i \in I \setminus S^{t-1}\} \geq \pi^{t-1}(Z, \mathbb{D})$, set $\pi^t(Z, \mathbb{D}) = \max\{\pi_i^t(Z, \mathbb{D}), \forall i \in S^{t-1}\}$ and $S^t = S^{t-1} \cup \{\arg\max_{i \in I \setminus S^{t-1}} \pi_i^t(Z, \mathbb{D})\}$; Otherwise, stop and set $\pi^c(Z, \mathbb{D}) = \pi^{t-1}(Z, \mathbb{D})$ and $S = S^{t-1}$.

Performance of "Greedy Add" Algorithm for 2nd Stage Allocation Optimization
We test the heuristic on a system consists of 20 raw materials, 3 ingredients and 5 final products. Given the raw material inventory, the system makes raw material selection decisions and further allocation decisions. In the experiment, we try two sets of grade selection costs: the homogeneous and the heterogeneous selection costs. In the first case, all selection costs are set to be 100; while in the latter case, the selection costs are set to be 50 for odd indexed raw materials and 100 for even indexed raw materials. The inventory of each raw material is equally set to be 100. Moreover, the ingredient concentration matrix and ingredient requirement matrix are specified as inputs. The demands for final products are independent and equally likely to be {100, 200, 300, 400, 500}. We solved the system for all 3125 demand scenarios by the heuristic and compare the sample average profit with exact solution provided by CPLEXMILP solver. The results are summarized in Table 3.4:

Table 3.4: Performance of "Greedy Add" Algorithm

	$f = 100$	$f_{odd} = 50, f_{even} = 100$
Expected revenue by exact solution (dollars)	7453.6	7853.5
Expected revenue by heuristic (dollars)	7412.6	7806.7
Gap	0.55%	0.60%

3.6 Simulation-based Optimization

In this section, we focus on solving the linear model with exact solutions to the second-stage allocation problem. Our solution approach to the proposed two-stage stochastic mixed-integer program consists of two modules: demand simulator and SAA optimizer, as shown in Figure 3.6. Given the available historical demand data, the demand simulator generates simulated demand arrivals. There are several ways to simulate or forecast demands based on previous information. Here, we use Bootstrap sampling in the demand simulator.

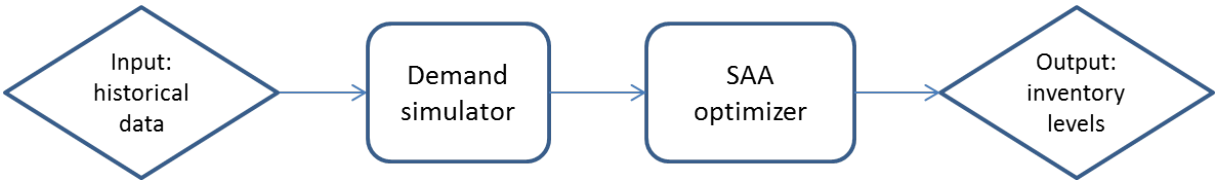


Figure 3.6: Solution Approach

With the simulated demand arrivals, the second module finds the optimal inventory levels by *Sample Average Approximation* (SAA), an simulation-based approach. The algorithm is modified from the SAA method provided in Akçay and Xu (2004) [4]. A detailed introduction of the SAA method can be found in Shapiro and Homem-de Mello (1998) [84].

Let $\Pi(Z^*)$ be the optimal solution to the two-stage stochastic program in the linear model (3.15) and Z^* be the associated optimal inventory levels. We start with generating M independent samples of random vector $\mathbf{D}$ from the demand simulator, each of size N. That is, $\mathbf{D}^l = (D^{l,1}, D^{l,2}, ..., D^{l,N})$ is the realization of the lth sample, where $D^{l,h} = (d_1^{l,h}, d_2^{l,h}, ..., d_K^{l,h})$ with $d_k^{l,h}$ as the realization of demand for final product k in the hth vector of the lth sample realization. We solve the SAA problem referring to each sample, as below:

$$\max_{Z^l} \left\{ \Pi^N(Z^l) = \frac{1}{N} \sum_{h=1}^{N} \pi(Z^l, D^{l,h}) - \sum_{i \in I} c_i z_i^l \right\} \tag{3.20}$$

s.t.

$$\text{Constraints (3.17)} - \text{(3.19) for each } D^{l,h}, h = 1, ..., N \tag{3.21}$$

where $Z^l = (z_1^l, z_2^l, ..., z_I^l)$

For $l = 1, ..., M$, let $\Pi^N(\hat{Z}^l)$ be the corresponding optimal solution to the above SAA problem and $\hat{Z}^l$ be the associated optimal inventory investment. Since Z^* is always a feasible solution to problem (3.20), we have $\Pi^N(\hat{Z}^l) \geq \Pi(Z^*)$ for all $l = 1, ..., M$. We then have,

$$E[\bar{\Pi}^N] \geq \Pi(Z^*), \text{ where } \bar{\Pi}^N = \frac{1}{M} \sum_{l=1}^{M} \Pi^N(\hat{Z}^l)$$

Therefore, $E[\bar{\Pi}^N]$ is used as the estimate of an upper bound of $\Pi(Z^*)$.

In order to have an unbiased estimator of $\Pi(\hat{Z}^l)$, we again generate one large number of independent sample from the demand simulator, $\mathbf{D}^{N'} = (D^1, D^2, ..., D^{N'})$. Then, for each inventory level vector $\hat{Z}^l$, compute the estimate of $\Pi(\hat{Z}^l)$ by

$$\hat{\Pi}^{N'}(\hat{Z}^l) = \frac{1}{N'} \sum_{h=1}^{N'} \pi(\hat{Z}^l, D^h) - \sum_{i \in I} c_i \hat{z}_i^l$$

where $\pi(\hat{Z}^l, D^h)$ is the optimal solution to the second stage allocation optimization with inventory vector $\hat{Z}^l$ and demand realization D^h.

The estimated optimal inventory vector $\hat{Z}^*$ is then determined by choosing the one that gives the largest $\hat{\Pi}^{N'}(\hat{Z}^l)$ among all candidate inventory vectors $\hat{Z}^l$ with sampled demand realizations, as follows:

$$\hat{Z}^* \in \arg\max\{\hat{\Pi}^{N'}(\hat{Z}^l), l = 1, ..., M\}$$

Since $\hat{Z}^*$ is a feasible solution to the linear model (3.15), we further have

$$E[\hat{\Pi}^{N'}(\hat{Z}^*)] \leq \Pi(Z^*)$$

As a result, $E[\hat{\Pi}^{N'}(\hat{Z}^*)]$ can serve as a lower bound of $\Pi(Z^*)$. Thus, the difference between $E[\bar{\Pi}^N]$ and $E[\hat{\Pi}^{N'}(\hat{Z}^*)]$ is an estimate of the optimality gap of SAA solution. In brief, the SAA method works in the following procedure:

Step 0 Determine appropriate values for N, M and N', and initialize $l = 0$;

Step 1 Set $l = l + 1$ and generate an independent sample $\mathbf{D}^l = (D^{l,1}, D^{l,2}, ..., D^{l,N})$; Solve the SAA problem (3.20)-(3.21) for $\hat{Z}^l$ and $\Pi^N(\hat{Z}^l)$; If $l < M$, go to Step 1; otherwise, go to Step 2;

Step 2 Generate an independent sample $\mathbf{D}^{N'} = (D^1, D^2, ..., D^{N'})$; Initialize $l = 0$;

Step 3 Set $l = l + 1$ and solve the SAA problem with $\mathbf{D}^{N'}$ and $\hat{Z}^l$ for $\hat{\Pi}^{N'}(\hat{Z}^l)$; If $l < M$, go to Step 2; otherwise go to Step 4;

Step 4 Choose $\hat{Z}^* \in \arg\max\{\hat{\Pi}^{N'}(\hat{Z}^l), l = 1, ..., M\}$;

The quality of the solution, measured by the optimality gap, improves as the sample sizes N and N' grow. However, larger sample sizes require higher computational capacity. Therefore, tradeoff between sample sizes and computational effort need to be considered.

3.7 Numerical Study

In this section, we apply the proposed approach to a real-world flour manufacturing system with one-year demand data and some scaled cost values. The system produces 18 kinds of flour "A" to "R" for different uses from 3 grades of wheat numbered "1" to "3" from different origins. The ingredients are mainly starch, protein and fiber. The wheat with higher protein concentration costs more. The costs, ingredient concentration, and requirement matrices are summarized in Table 3.5.

Table 3.5: Parameters in Numerical Study

	c (\$/kg)	f (10^3 \$)	r (\$/kg)	Starch (100%)	Protein (100%)	Fiber (100%)
Wheat 1	2.10	2.0		0.80	0.10	0.10
Wheat 2	1.83	2.5	N/A	0.60	0.15	0.25
Wheat 3	1.78	3.0		0.50	0.30	0.20
Flour A			2.63	0.98	0	0.02
Flour B			2.43	0.95	0	0.05
Flour C			2.21	0.72	0.1	0.18
Flour D			2.38	0.88	0	0.12
Flour E			2.40	0.88	0.02	0.1
Flour F			2.19	0.8	0	0.2
Flour G			2.15	0.68	0.15	0.17
Flour H			2.41	0.93	0	0.07
Flour I			2.06	0.75	0	0.25
Flour J	N/A	N/A	1.79	0.65	0	0.35
Flour K			1.77	0.2	0.7	0.1
Flour L			2.33	0.76	0.1	0.14
Flour M			2.13	0.57	0.2	0.23
Flour N			2.37	0.57	0.33	0.1
Flour O			2.23	0.33	0.52	0.15
Flour P			2.41	0.3	0.65	0.05
Flour Q			2.36	0.09	0.87	0.04
Flour R			1.48	0	1	0

The sample demand arrivals are generated by the demand simulator module, which implements Bootstrap sampling in the current setting. The SAA optimizer module is realized via CPLEX solver with $N = 100$, $M = 30$ and $N' = 500$. We compute the average profit, inventory investment in dollar value and gaps for various parameter settings listed in Table 3.6. As mentioned in the SAA algorithm, the gap is defined as the difference between the upper and lower bounds. The upper bound is estimated by $\bar{\Pi}^N = \frac{1}{M}\sum_{l=1}^{M}\Pi^N(\hat{Z}^l)$ and the lower bound is estimated by $\frac{1}{N'}\sum_{h=1}^{N'}\pi(\hat{Z}^*, D^h) - \sum_{i\in I} c_i\hat{z}_i^*$. For current parameter

setting, the optimal solution given by our approach is $\$2721.74\times10^6$ with inventory levels [5718.194;0;4041.851] for wheat 1, 2 and 3 respectively. We summarize the computational results of run 1 to 5 for flexible recipe system in Table 3.7.

Table 3.6: Experiment Setting of Selected Runs

	Wheat costs	Grade selection costs
Run 1	[1.680;1.464;1.424]	[2;2.5;3]
Run 2	[1.890;1.647;1.602]	[2;2.5;3]
Run 3	[2.100;1.830;1.780]	[2;2.5;3]
Run 4	[2.310;2.013;1.958]	[2;2.5;3]
Run 5	[2.520;2.196;2.136]	[2;2.5;3]

Table 3.7: Results for Flexible Recipes

	Avg. profit (10^6\$)	Inv. (10^6 \$)	Gap
Run 1	6750.69	17631.84	1.82%
Run 2	4666.52	18948.49	0.91%
Run 3	2721.74	19275.04	0.05%
Run 4	938.55	18000.56	0.41%
Run 5	30.22	3310.80	0.15%

For the selected run 1 to 5, all gaps are less than 2%. This suggests the good performance of the simulation-based approach with our choice of N, M and N'. If the gap is big, the number of samples N and N' should be increased accordingly. Besides, both the average profit and total inventory investment decrease convexly as the raw material cost increases. This implies that the system tends to stock less inventory when raw materials cost is high. Recall that in newsvendor model, the critical fractile decreases linearly with production cost and thus the inventory level also decreases convexly if the demand distribution is concave, i.e. Normal distribution.

The impact of large grade selection cost is illustrated in Figure 3.7. The increasing grade selection costs decrease the average profit at a mild rate. Meanwhile, grade selection cost increase does not change the inventory investment significantly. Since the grade selection costs are only scaled by multipliers, their relative ranking among different raw materials are not changed. As a result, the preference among wheat are not much affected. This makes the inventory investment decision and average revenue almost unchanged. Therefore, the average profit, which equals average revenue less inventory investment and grade selection costs, decreases linearly in the multipliers of grade selection costs.

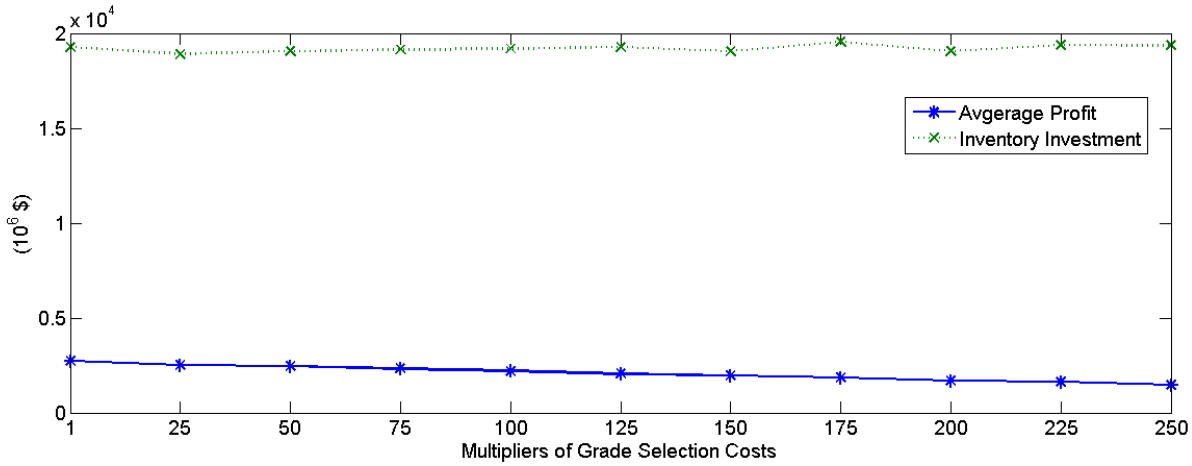


Figure 3.7: Average Profit and Inventory Investment for Different Grade Selection Costs

Table 3.8: Comparison of Flexible Recipes and Fixed Recipes

		Run 1	Run 2	Run 3	Run 4	Run 5
Flexible	Avg. Profit (10^6\$)	6750.69	4666.52	2721.74	938.55	30.22
recipe	Inv. (10^6\$)	17631.84	18948.49	19275.04	18000.56	3310.80
Flexible	Avg. Profit (10^6\$)	4869.70	3244.55	1695.32	448.34	0
recipe:	Ratio	72.14%	69.53%	62.29%	47.77%	0%
wheat 1	Inv. (10^6\$)	14188.77	14384.60	14035.47	11636.23	0
Flexible	Avg. Profit (10^6\$)	5834.94	3551.48	1595.87	85.50	0
recipe:	Ratio	86.43%	76.11%	58.63%	9.11%	0%
wheat 2	Inv. (10^6\$)	18140.89	18677.88	19062.30	7330.68	0
Flexible	Avg. Profit (10^6\$)	5039.26	3151.62	1786.61	695.27	29.62
recipe:	Ratio	74.65%	67.54%	65.64%	74.08%	98.02%
wheat 3	Inv. (10^6\$)	16165.77	14579.06	12421.80	10315.34	3256.07

We consider three simple fixed recipes. That is, fixed recipes with single source: wheat 1, 2 or 3. Table 3.8 summarizes the computational results of the average profit, inventory investment as well as the ratio between average profit of the flexible recipes and those of the fixed recipes. For all experiments, flexible recipes can always achieve larger average profit than fixed recipes. For example, in run 1, the "best" fixed recipe can at most generate 86.43% of the average profit of flexible recipe. Meanwhile, in run 5, which is an extreme case, wheat 1 and 2 are too costly to be used as raw materials. This leaves wheat 3 as the only choice as raw material for the flexible recipe. Therefore, in run 5, the flexible recipe is equivalent to the fixed recipe with wheat 3. In addition, as we see from run 1 to 3, the optimal solution of the flexible recipes requires not much more or even significantly less inventory investment than the "best" fixed recipes. That is, flexible recipes achieve higher average profit with

lower inventory investment than fixed recipes.

3.8 Summary

In this chapter, we propose a two-stage stochastic mixed-integer program to an inventory management problem in continuous process with flexible recipes. In the first stage, the model determines inventory levels for each period based on past demand data. After demand arrivals are realized, the second stage recourse makes recipe selection and allocation decisions in production. With available historical demand data, a simulation-based approach based on SAA algorithm is developed to solve the stochastic program. The results of numerical study show the performance of the approach on various cost settings as well as the benefits of flexible recipes over fixed recipes.

In the proposed approach, we focus on the application of the SAA algorithm and use Bootstrap sampling as the default in demand simulation. A direction of future improvement is to incorporate better techniques in the simulation of future demand arrivals based on historical demand data. Those techniques may consider some properties of the demand, such as seasonality and autocorrelation. Also, with limited demand information, a robust optimization model might be developed that considers the worst cases. Moreover, since our model assumes any inventory leftover at the end of each period is disposed, the extension that relaxes this assumption and introduces inventory holding cost in multi-period setting should also be investigated.

Appendix A

Tables

A.1 Summary of the Notation

Table A.1: Notation

Parameters	Symbol	Units	Definition
	a_{ij}	[0,1]	Utility of serving destination j for a customer in region i
	b_k	[0,1]	Utility threshold for a customer in group k to adopt the service
	f	$	fixed membership fee
	r	$ per unit of time	Usage based price of the service
	η	$ per unit of time	Repositioning cost
	c	$ per unit of time	Charging cost
	α	(0,1)	EV availability (service level)
	Q_{ik}	Customers	Population of customer group k in region i
	g_i	$	Fixed coverage cost of region i
	$\bar{a}_{ij}$	[0,1]	Expected value of a_{ij}
	Γ_i		Covariance matrix of a_{ij} for region i
	μ_i	Trips per unit of time	Outbound trip rate from region i
	P_c	[0,1]	Probability of an arrival EV needs recharge
	P_{ij}	[0,1]	Proportion of customer flows to destination j from origin i
	t_c	minute	Average charging time
	t_{ij}	minute	Travel time to destination j from origin i
	τ_{ij}	minute	Reposition time to destination j from origin i
	L_i	EVs	Expected available EVs in region i
	ψ'_{ik}	[0,1]	Population weight of customer group k in region i
	ψ_{ik}	[0,1]	Proportion of outbound trips by customer group k in region i

Decision Variables	Symbol	Units	Definition
	x_i	$\{0,1\}$	1 if region i is served; 0 otherwise.
	q_{ik}	[0,1]	Expected adoption rate of customer group k in region i
	Λ_i	EVs per unit of time	Arrival rate of EVs available for customers in region i
	λ_i	EVs per unit of time	External EV arrival rate to region i from Charging Stations
	γ_{jl}	[0,1]	Probability of an EV arrival at j is repositioned to l
	ϕ_{ijl}	EVs per unit of time	Repositioning trip rate of EV arrivals from i at j to l
	N	EVs	Fleet size

Appendix B

Proofs of Analytical Results

B.1 Proof of Lemma 1

We begin the proof with the computation of worst-case probability constraint (1.4). For the ease of computation, it is safe to temporarily drop the index $i \in I$ and $k \in K$. Given the service region decision $\mathbf{x}$, the worst-case probability constraint $V(\mathbf{x}) = \sup \text{Prob}(\sum_{j\in J} a_{ij}x_j \leq b_k)$ can be obtained by solving:

$$\begin{aligned} &\max \mathbb{E}[I(\mathbf{a})] \\ &\text{s.t.} \\ &\int_{\mathbb{R}^n_+} \begin{bmatrix} \mathbf{a} \\ 1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ 1 \end{bmatrix}^T p(\mathbf{a}) d\mathbf{a} = \Sigma \end{aligned}$$

where $p \in \mathcal{P}$ is the probability density function and $I(\mathbf{a})$ is the indicator function defined as

$$I(a) = \begin{cases} 1, & \text{if } \sum_{j\in J} a_j x_j \leq b \\ 0, & \text{otherwise} \end{cases}$$

We write the Lagrange function with symmetric multiplier matrix $M \in \mathcal{S}_{n+1}$

$$\begin{aligned} L(p, M) &= \int_{\mathbb{R}^n_+} I(\mathbf{a}) p(\mathbf{a}) d\mathbf{a} + \left\langle M, \Sigma - \int_{\mathbb{R}^n_+} \begin{bmatrix} \mathbf{a} \\ 1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ 1 \end{bmatrix}^T p(\mathbf{a}) d\mathbf{a} \right\rangle \\ &= \langle M, \Sigma \rangle + \int_{\mathbb{R}^n_+} \big(I(\mathbf{a}) - l(\mathbf{a})\big) p(\mathbf{a}) d\mathbf{a} \end{aligned}$$

where $l(\mathbf{a}) = [\ \mathbf{a}^T \quad 1\ ] M [\ \mathbf{a} \quad 1\ ]^T$. Since $\Sigma \succ 0$, strong duality holds. Therefore, we have

$$V(\mathbf{x}) = \inf_{M=M^T} \sup_{p\in\mathcal{P}} L(p, M)$$

where

$$\begin{aligned}\sup_{p\in\mathcal{P}} L(p,M) &= \langle M,\Sigma\rangle + \sup_{p\in\mathcal{P}} \int_{\mathbb{R}^n_+} \big(I(\mathbf{a}) - l(\mathbf{a})\big)p(\mathbf{a})d\mathbf{a} \\ &= \begin{cases} \langle M,\Sigma\rangle, & \text{if } I(\mathbf{a}) - l(\mathbf{a}) \le 0, \forall \mathbf{a}\in\mathbb{R}^n_+ \\ +\infty, & \text{otherwise.} \end{cases}\end{aligned}$$

$V(\mathbf{x})$ is finite if and only if $I(\mathbf{a}) - l(\mathbf{a}) \le 0, \forall \mathbf{a}\in\mathbb{R}^n_+$. There are two cases:

1. $l(\mathbf{a}) \ge 0, \forall \mathbf{a}\in\mathbb{R}^n_+$. Equivalently, $M \succeq_{co} 0$.
2. $l(\mathbf{a}) \ge 1, \forall \mathbf{a}\in\mathbb{R}^n_+$ such that $\sum_{j\in J} a_j x_j \le b$. That is, there exist a scalar $\tau \ge 0$ such that, $l(\mathbf{a}) \ge 1 - 2\tau(\mathbf{a}^T\mathbf{x} - b)$. Equivalently, $M + \begin{bmatrix} 0 & \tau\mathbf{x} \\ \tau\mathbf{x}^T & -1-2\tau b \end{bmatrix} \succeq_{co} 0$.

The worst-case probability constraint $V(\mathbf{x})$ is then the solution to the following copositive program (CP):

$$\begin{aligned} V(\mathbf{x}) = &\min \langle M,\Sigma\rangle \\ &\text{s.t.} \\ &\quad \tau \ge 0 \\ &\quad M \succeq_{co} 0 \\ &\quad M + \begin{bmatrix} 0 & \tau\mathbf{x} \\ \tau\mathbf{x}^T & -1-2\tau b \end{bmatrix} \succeq_{co} 0. \end{aligned}$$

We then complete the proof by restoring the indices $i \in I$ and $k \in K$ to the probability constraint and replacing the worst-case probability constraint $V(\mathbf{x}) \le q$ with the above CP:

$$\begin{aligned} \max_{q_{ik},x_i,N} &\sum_{i\in I}\sum_{k\in K} fQ_{ik}q_{ik} - \sum_{i\in I} g_i x_i + \Theta(x_i, q_{ik}, \alpha) \\ &\text{s.t.} \\ &\quad \langle M_{ik},\Sigma_{ik}\rangle \le 1 - q_{ik}, \forall i\in I, \forall k\in K \\ &\quad \tau_{ik} \ge 0, \forall i\in I, \forall k\in K \\ &\quad M_{ik} \succeq_{co} 0, \forall i\in I, \forall k\in K \\ &\quad M_{ik} + \begin{bmatrix} 0 & \tau_{ik}\mathbf{x} \\ \tau_{ik}\mathbf{x}^T & -1-2\tau_{ik}b_k \end{bmatrix} \succeq_{co} 0, \forall i\in I, \forall k\in K \\ &\quad q_{ik} \le x_i, \forall i\in I, \forall k\in K \\ &\quad x_i \in \{0,1\}, \forall i\in I. \end{aligned}$$

The last step is to linearize the term $\tau_i \mathbf{x}$. Since $\mathbf{x}$ is a binary vector and τ_i is continuous, we can replace the term $\tau_i X$ by vector $\mathbf{d_i}$ with the following constraints:

$$\begin{aligned} -\rho\mathbf{x} &\leq \mathbf{d_i} \\ \mathbf{d_i} &\leq \rho\mathbf{x} \\ \tau_i\mathbf{e} + \rho(\mathbf{x} - e) &\leq \mathbf{d_i} \\ \mathbf{d_i} &\leq \tau_i\mathbf{e} + \rho(\mathbf{e} - \mathbf{x}) \end{aligned}$$

where ρ is a large scalar and $\mathbf{e}$ is the vector of ones. This completes the proof.

B.2 Proof of Proposition 1

By restricting the copositive constraints with semidefinite constraints, we obtain a lower bound to the optimal solution to the formulation with copositive constraints in Lemma 1. When $\Theta(x_i, q_{ik}, \alpha)$ is linear, the following formulation is a mixed integer semidefinite program (MISDP):

$$\max_{q_{ik}, x_i} \sum_{i \in I} \sum_{k \in K} f Q_{ik} q_{ik} - \sum_{i \in I} g_i x_i + \Theta(x_i, q_{ik}, \alpha)$$

s.t.

$$\begin{aligned} &\langle M_{ik}, \Sigma_{ik} \rangle \leq 1 - q_{ik}, \forall i \in I, \forall k \in K \qquad \text{(B.1)} \\ &M_{ik} \succeq 0, \forall i \in I, \forall k \in K \\ &M_{ik} + \begin{bmatrix} 0 & \mathbf{d_{ik}} \\ \mathbf{d_{ik}}^T & -1 - 2\tau_{ik} b_k \end{bmatrix} \succeq 0, \forall i \in I, \forall k \in K \\ &-\rho\mathbf{x} \leq \mathbf{d_{ik}}, \forall i \in I, \forall k \in K \\ &\mathbf{d_{ik}} \leq \rho\mathbf{x}, \forall i \in I, \forall k \in K \\ &\tau_{ik}\mathbf{e} + \rho(\mathbf{x} - e) \leq \mathbf{d_{ik}}, \forall i \in I, \forall k \in K \\ &\mathbf{d_{ik}} \leq \tau_{ik}\mathbf{e} + \rho(\mathbf{e} - \mathbf{x}), \forall i \in I, \forall k \in K \\ &\tau_{ik} \geq 0, \forall i \in I, \forall k \in K \qquad \text{(B.2)} \\ &q_{ik} \leq x_i, \forall i \in I, \forall k \in K \\ &x_j \in \{0, 1\}, \forall j \in J. \end{aligned}$$

El Ghaoui et al. (2003)[39] provide a closed-form expression to the semedifinite constraints above. Let $\mathcal{P}$ be the set of probability distributions with mean μ and covariance matrix $\Gamma \succ 0$. Let $\epsilon \in (0, 1]$ and $\gamma \in \mathbf{R}$ given. The following propositions are equivalent.

1. $\sqrt{\frac{1-\epsilon}{\epsilon}} \sqrt{\mathbf{x}^T \Gamma_i \mathbf{x}} - \mu^T \mathbf{x} \leq \gamma$

2. There exist a symmetric matrix M and $\tau \in \mathbf{R}$ such that

$$\begin{aligned} \langle M, \Sigma \rangle &\leq \tau\epsilon \\ M &\succeq 0 \\ M + \begin{bmatrix} 0 & \mathbf{x} \\ \mathbf{x}^T & -\tau + 2\gamma \end{bmatrix} &\succeq 0 \\ \tau &\geq 0 \end{aligned}$$

where Σ is the second-moment matrix.

We are then able to rewrite the semidefinite constraint (B.1) to (B.2) into

$$\begin{aligned} &\sqrt{\frac{q_{ik}}{1-q_{ik}}}\sqrt{X^T\Gamma_i X} - \bar{\mathbf{a}}_i^T\mathbf{x} + b_i \leq \quad 0 \\ &\equiv \begin{cases} 1 - q_{ik} \geq \frac{\mathbf{x}^T\Gamma_i\mathbf{x}}{(\bar{\mathbf{a}}_i^T\mathbf{x} - b_k)^2 + \mathbf{x}^T\Gamma_i\mathbf{x}} \\ \bar{\mathbf{a}}_i^T\mathbf{x} - b_k \geq 0. \end{cases} \end{aligned} \tag{B.3}$$

In the case of $\bar{\mathbf{a}}_i^T\mathbf{x} - b_k \leq 0$, the worst case probability constraint is 1, because there exists a two-point distribution of $\bar{\mathbf{a}}_i^T\mathbf{x}$ with both value less than b_k. To allow such possibility, we further introduce a set of disjunctive constraints. When $\bar{\mathbf{a}}_i^T\mathbf{x} - b_k \leq 0$, we set q_{ik} to 0. Therefore, we have either $\bar{\mathbf{a}}_i^T\mathbf{x} \geq b_k$ or $1 - q_{ik} \geq 1$. By introducing new variables, we can express the disjunctive constraints as the feasible set

$$\mathcal{X}_{ik} = \left\{ (x_i, q_{ik}) : \begin{array}{l} b_k u_{ik}^1 \leq \bar{\mathbf{a}}_i^T \mathbf{s}_{ik} \\ u_{ik}^2 \leq o_{ik} \\ u_{ik}^1 + u_{ik}^2 = 1 \\ \mathbf{s}_{ik} \leq \mathbf{x} \\ o_{ik} \leq 1 - q_{ik} \\ u_{ik}^1, u_{ik}^2 \geq 0 \end{array} \right\}. \tag{B.4}$$

Moreover, we linearize the term $x_{j_1}x_{j_2}$ with $z_{j_1j_2}$ defined in

$$\mathcal{Z}(x_{j_1}, x_{j_2}) = \left\{ z_{j_1j_2} : \begin{array}{l} z_{j_1j_2} \leq x_{j_1} \\ z_{j_1j_2} \leq x_{j_2} \\ x_{j_1} + x_{j_2} - 1 \leq z_{j_1j_2} \\ z_{j_1j_2} \geq 0 \end{array} \right\}. \tag{B.5}$$

Lastly, the first inequality in constraint (B.3) can be expressed as $(1 - q_{ik} + (\bar{\mathbf{a}}_i^T\mathbf{x} - b_k)^2 + \mathbf{x}^T\Gamma_i\mathbf{x})^2 \geq 4\mathbf{x}^T\Gamma_i\mathbf{x} + (1 - q_{ik} - (\bar{\mathbf{a}}_i^T\mathbf{x} - b_i)^2 - \mathbf{x}^T\Gamma_i\mathbf{x})^2$. Since $z_{j_1j_2} = x_{j_1}x_{j_2}$ in (B.5), we further linearize the terms $(\bar{\mathbf{a}}_i^T\mathbf{x})^2$ and $\mathbf{x}^T\Gamma_i\mathbf{x}$ as $\sum_{(j_1,j_2)\in I\times I} \bar{a}_{ij_1}\bar{a}_{ij_2}z_{j_1j_2}$ and $\sum_{(j_1,j_2)\in I\times I} \sigma_{ij_1j_2}z_{j_1j_2}$ respectively and results the second-order cone constraint (1.5).

B.3 Time-varying Travel Pattern

Suppose there are T periods in a day within which the travel patterns are approximately stationary, e.g. P_{ij}^t and μ_i^t, then we formulate the problem as

$$\max_{q_{ik},x_i,N} \sum_{i\in I} fQ_i q_i' - \sum_{i\in I} g_i x_i + \frac{1}{T}\sum_{t\in T}\sum_{j\in I}\sum_{i\in I} rt_{ij}\Lambda_{ij}^t - \frac{1}{T}\sum_{t\in T}\sum_{i\in I} ct_c\lambda_i^t - \frac{1}{T}\sum_{t\in T}\sum_{i\in I}\sum_{j\in I}\sum_{m\in I}\eta\tau_{jm}\phi_{ijm}^t - hN$$

s.t.

$$q_{ik} \le \text{Prob}\big(\sum_{j\in I} a_{ij}x_j \ge b_k\big), \forall i\in I, k\in K$$

$$\sum_{j\in I}\Lambda_{ij}^t \ge \alpha\mu_i \sum_{j\in I} q_{ij}^t, \forall i\in I, t\in T$$

$$\Lambda_{ij}^t \le \mu_i^t q_{ij}^t, \forall i\in I, t\in T$$

$$q_i' = \sum_{k\in K}\psi_{ik}' q_{ik}, \forall i\in I$$

$$q_i = \sum_{k\in K}\psi_{ik} q_{ik}, \forall i\in I$$

$$q_i \le x_i, \forall i\in I$$

$$q_{ij}^t = q_i P_{ij}^t x_j, \forall i\in I, j\in J, t\in T$$

$$\sum_{j\in I}\Lambda_{ij}^t = \lambda_i^t + \sum_{j\in I}\Lambda_{ji}^t(1-P_c) - \sum_{j\in I}\sum_{l\in I}\phi_{jil}^t + \sum_{j\in I}\sum_{m\in I}\phi_{jmi}^t, \forall i\in I, t\in T$$

$$\sum_{l\in I}\phi_{jil}^t \le \Lambda_{ji}^t(1-P_c), \forall i\in I, j\in I, t\in T$$

$$\lambda_i^t = \sum_{j\in J}\mu_j^t q_{ji}^t P_c, \forall i\in I, t\in T$$

$$\sum_{j\in I}\sum_{i\in I} t_{ij}\Lambda_{ij}^t + \sum_{i\in I} L_i x_i + \sum_{i\in I}\lambda_i^t t_c + \sum_{i\in I}\sum_{m\in I}\sum_{j\in J}\tau_{mj}\phi_{imj}^t \le N, \forall t\in T$$

$$\frac{\Lambda_{ij}^t x_k}{P_{ij}^t} = \frac{\Lambda_{ik}^t x_j}{P_{ik}^t}, \forall i\in I, j\in I, k\in I, t\in T$$

$$\Lambda_{ij}^t \ge 0, \forall i\in I, j\in I, t\in T$$

$$\phi_{ikj}^t \ge 0, \forall i\in I, k\in I, j\in I, t\in T$$

$$x_i \in \{0,1\}, \forall i\in I.$$

B.4 Proof of Proposition 3

From charging flow constraint (1.9), it is straightforward that λ_i is linearly increasing in q_n for $n\in I$. With any fixed repositioning probability γ_{il}, the flow balance constraint (1.8) can

be rewritten as

$$\Lambda_i = \lambda_i + \sum_{j\in I} \Lambda_j[\hat{P}_{ji}(1 - P_c - \sum_{l\in I} \gamma_{il}) + \sum_{m\in I} \hat{P}_{jm}\gamma_{mi}], \forall i \in I. \tag{B.6}$$

Since $(1 - P_c - \sum_{l\in I} \gamma_{il}) \geq 0$, Λ_i is increasing in λ_i and subsequently increasing in q_n. By summing up the flow balance equations (B.6) for all $i \in I$, we get

$$\begin{aligned}
\sum_{i\in I} \Lambda_i &= \sum_{i\in I} \lambda_i + \sum_{j\in I} \Lambda_j(1 - P_c) \sum_{i\in I} \hat{P}_{ji} - \sum_{j\in I} \Lambda_j \sum_{i\in I} \sum_{l\in I} \hat{P}_{ji}\gamma_{il} + \sum_{j\in I} \Lambda_j \sum_{m\in I} \sum_{i\in I} \hat{P}_{jm}\gamma_{mi} \\
&= \sum_{i\in I} \lambda_i + \sum_{j\in I} \Lambda_j(1 - P_c) \\
\Rightarrow P_c \sum_{i\in I} \Lambda_i &= \sum_{i\in I} \lambda_i.
\end{aligned}$$

Furthermore, under the profit maximization objective, fleet size constraint (1.10) is always binding. Given the service region decisions x_i's, a sufficient condition for the probability constraints to be binding is to have a non-decreasing operational profit $\Theta(x_i, q_{ik}, \alpha)$.

$$\Theta(x_i, q_{ik}, \alpha) = \sum_{j\in I} \sum_{i\in I} (r - h)t_{ij}\Lambda_i\hat{P}_{ij} - (c + h)t_c \sum_{i\in I} \lambda_i - \sum_{i\in I} \sum_{j\in I} \sum_{m\in I} (\eta + h)\tau_{jk}\Lambda_i\hat{P}_{ij}\gamma_{jm} - h \sum_{i\in I} L_i x_i.$$

Apparently, a neccessary condition for Θ to be non-decreasing is $r \geq h$. To derive a sufficient condition, by taking derivative regarding to q_n, we have

$$\begin{aligned}
\frac{\partial\Theta}{\partial q_n} =& \sum_{j\in I} \sum_{i\in I} (r - h)t_{ij}\hat{P}_{ij} \frac{\partial\Lambda_i}{\partial q_n} - (c + h)t_c \frac{\partial \sum_{i\in I} \lambda_i}{\partial q_n} - \sum_{i\in I} \sum_{j\in I} (\eta + h)(\sum_{m\in I} \tau_{jm}\hat{P}_{ij}\gamma_{jm}) \frac{\partial\Lambda_i}{\partial q_n} \\
=& \sum_{j\in I} \sum_{i\in I} (r - h)t_{ij}\hat{P}_{ij} \frac{\partial\Lambda_i}{\partial q_n} - (c + h)t_c P_c \frac{\partial \sum_{i\in I} \Lambda_i}{\partial q_n} - \sum_{i\in I} \sum_{j\in I} (\eta + h)(\sum_{m\in I} \tau_{jm}\hat{P}_{ij}\gamma_{jm}) \frac{\partial\Lambda_i}{\partial q_n} \\
=& \sum_{i\in I} \big[(r - h) \sum_{j\in I} t_{ij}\hat{P}_{ij} - (c + h)t_c P_c - (\eta + h)(\sum_{j\in I} \sum_{m\in I} \tau_{jm}\hat{P}_{ij}\gamma_{jm})\big] \frac{\partial\Lambda_i}{\partial q_n} \\
=& \sum_{i\in I} \big[r \sum_{j\in I} t_{ij}\hat{P}_{ij} - ct_c P_c - \eta \sum_{j\in I} \sum_{m\in I} \tau_{jm}\hat{P}_{ij}\gamma_{jm} - h(\sum_{j\in I} t_{ij}\hat{P}_{ij} + t_c P_c + \sum_{j\in I} \sum_{m\in I} \tau_{jm}\hat{P}_{ij}\gamma_{jm})\big] \frac{\partial\Lambda_i}{\partial q_n} \\
\geq& 0
\end{aligned}$$

when the sufficient condition holds.

Appendix C

Estimation of Key Parameters

C.1 Adoption Requirement

The minimum utility threshold to adopt b_k for group k is evaluated based on the mode choice data in CHTS. We summarize the mode choice distribution in 4 categories: non-moto (including walking and cycling, etc,.), private car (including driver, passenger, car rental and carpool, etc,.), bus (including bus and shuttle, etc,.), and rail (including subways and light rail, etc,.). We focus on motorized trips and group bus and rail modes to public transportation. By K-means clustering approach in Hartigan and Wong (1979)[50], we group the individuals into 5 clusters with different mode choice distributions as shown in Table C.1.

Table C.1: Mode Choice Distribution

Group	Car	Public	b_k	ψ'_k	ψ_k
1	0	1.00	0	3.95%	0%
2	0.33	0.67	0.33	0.45%	0.13%
3	0.52	0.48	0.52	0.70%	0.21%
4	0.72	0.28	0.72	0.50%	0.49%
5	1.00	0	1	94.40%	99.17%

For instance, the potential car buyers from group 2 will choose to drive for 33% of their trips after they acquire cars. We hence assume that group 2 customers will switch to Car2Go if the service region covers at least $b_2 = 33\%$ of their destinations. After adoption, customers will use Car2Go service regularly with destinations that are in the service region. Group 2 has population weight $\psi'_2 = 0.45\%$ and is accounted for $\psi_2 = 0.13\%$ of total trips by car of the entire population. Specially, group 1 is a group that is not a target of Car2Go because those customers do not drive. In addition, group 5 is also not the target group since they rely on cars so heavily that requires 100% service coverage to adopt and they might already own private cars. As a result, the target groups 2, 3 and 4 count 1.65% of the entire population with relative ratio among them $2 : 3 : 2$. Due to the limited sample sizes at zip code level,

we assume all zip codes share the same mode choice distribution. In the following sections, we set the default market size to 1.65%.

C.2 Utility Parameters

We use the trip distributions that describe customers preferences over destinations as utility parameters a_{ij}. Our study focuses on the 61 candidate zip codes in San Diego county, excluding remote and military areas. We estimate the trip distributions for the 61 candidate zip codes from Car2Go San Diego operations data and the customer group information in previous section. We partition the current service region into 18 zip codes. The sample daily trips for each OD pair in the 18 zip codes are counted and the outbound (inbound) trips from (to) each zip code are summarized in Table C.2.

Table C.2: Sample Daily Outbound(Inbound) Trips for Current Car2Go Service Region

91910	**91911**	**92101**	**92102**	**92103**	**92104**
7.92 (8.19)	3.19 (3.27)	334.04 (335.23)	55.69 (54.77)	144.04 (144.04)	81.69 (80.5)
92105	**92106**	**92107**	**92108**	**92109**	**92110**
11.23 (10.35)	45.89 (45.65)	54.23 (54.92)	50.77 (51.92)	84.08 (83.96)	44.92 (46)
92111	**92113**	**92115**	**92116**	**92120**	**92123**
0.35 (0.35)	1.23 (1.19)	14.62(14.85)	60.08 (58.73)	1.12 (1.19)	0.12 (0.08)

Table C.2 suggests a large variation in trip demands of the 18 zip codes. The majority of the trips were generated in the downtown San Diego with zip codes 92101 and 92103 while few trip demands observed from zip codes 92111 and 92123. To better capture the demand pattern, we exclude zip codes with very low demands, e.g., 91911, 92111, 92113, 92120 and 92123, in the regression analysis.

Since the travel patterns are time-varying, we partition the 24 hours of a day into 2 periods: daytime from 7AM to 21PM and night from 21PM to 7AM, which minimizes the sum of squared errors of the outbound trip rates. To simulate the trip distributions for both day and night between all OD pairs of the 61 candidate regions, we first apply the classic gravity model for trip distributions. Besides the population factor in the classic gravity model, we also test other socialeconomic factors, such as per capita income, business establishments, students enrollments and workplace population, that may affect the trip distributions. The only statistical significant factor we find is per capita income. Similarly, Wills (1986)[98] integrates income as a destination-attribute variable in his trip distribution models. Hence, we fit the gravity model for the trips with destinations different from the origins as follows:

$$T_{ij} = a\frac{P_i P_j Inc_i^b Inc_j^c}{dist_{ij}^d} \tag{C.1}$$

where P_i is the working population and Inc_i is the per capita income in i. The number of trips generated from i is proportional to P_i while P_j and Inc_j are indicators of the attractiveness of destination j.

Normalized with the market size 1.65% estimated in Section C.1, we compute the sample daily trip distribution for 18 zip codes. We then apply log-linear regression and obtain the gravity model in (C.1) for the aggregated daily trip distribution below with adjusted R-squared 0.7515 and residual standard error 0.9227.

$$T_{ij} = \exp(-62.212)\frac{P_i P_j Inc_i^{2.253} Inc_j^{2.249}}{dist_{ij}^{2.013}}. \tag{C.2}$$

For the trips within a zip code, we fit the following model to take same socialeconomic factors into account:

$$T_{ii} = aP_i Inc_i^b.$$

We use the sample daily trip distribution adjusted with the market size to obtain the coefficients a and b. The regressed model has adjusted R-squared 0.7845 and residual standard error 0.707 as below:

$$T_{ii} = \exp(-43.194)P_i Inc_i^{3.413}. \tag{C.3}$$

With the residual standard errors provided by the regressions, we randomly generate 1000 sample trip distributions for thw 61 candidate zip codes using the gravity models (C.2) and (C.3). In each sample k, the utility parameter is estimated by the trip proportion from normalization on outbound trips:

$$\hat{a}_{ij}^k = \frac{\hat{T}_{ij}^k}{\sum_{j\in I}\hat{T}_{ij}^k}.$$

The mean of utility parameter $\bar{a}_{ij}$ is then estimated by its sample average of $\hat{a}_{ij}^k$. We also construct the estimated diagonal covariance matrix Γ_i for each i with σ_{ij}^2 to be sample variance of $\hat{a}_{ij}^k$.

We further apply similar method to get the trip distributions for the time-varying travel pattern case with 2 periods defined as day and night.

C.3 Coverage Costs

The fixed coverage cost associated with serving region i includes investments in infrastructure such as partnership with charging service provider. As planned in 2011, Car2Go's fleet of 300 Smart Fortwo plug-ins can be recharged at 1000 Blink EV charging stations [23]. We use the EV charging station data from U.S. Department of Energy (DOE)[5] to estimate

the number of chargers desired in each region to support the EV sharing system. For each zip code in the current service region, we compute the charger density (CD) by dividing the number of chargers (CG) with the land area (LA). By fitting linear regression with various socialeconomic factors, we find that number of business establishment (BE) is the only significant factor to CD with the fitted model as below

$$CD_i = 3.19 \exp(-11) BE_i$$

and the number of chargers needed for each candidate zip code is then approximated as

$$CG_i = \max\{CD_i, 0\} \times LA_i.$$

Suppose the investment in each charger by Car2Go through partnership with Blink is h_c (e.g., \$800), then the coverage cost for region i is approximated by

$$g_i = h_c CG_i$$

Bibliography

[1] *2010 American Community Survey"*. 2010.

[2] Narendra Agrawal and Stephen A Smith. "Estimating negative binomial demand for retail inventory management with unobservable lost sales". In: *Naval Research Logistics (NRL)* 43.6 (1996), pp. 839–861.

[3] Vishal Agrawal and Ioannis Bellos. "The Potential of Servicizing as a Green Business Model". In: *Georgetown McDonough School of Business Research Paper* (2013).

[4] Y. Akçay and S.H. Xu. "Joint inventory replenishment and component allocation optimization in an assemble-to-order system". In: *Management Science* 50.1 (2004), pp. 99–116.

[5] *Alternative Fuels Data Center*. 2014.

[6] Marcus Ang, Yun Fong Lim, and Melvyn Sim. "Robust Storage Assignment in Unit-load Warehouses". In: *Management Science* 58.11 (2012), pp. 2114–2130.

[7] *Autolib*. 2014.

[8] Buket Avci, Karan Girotra, and Serguei Netessine. "Electric Vehicles with a Battery Switching Station: Adoption and Environmental Impact". In: *Management Science* (2014).

[9] Ioannis Bellos, Mark Ferguson, and L Beril Toktay. "To Sell and to Provide? The Economic and Environmental Implications of the Auto Manufacturer's Involvement in the Car Sharing Business". In: *Working Paper* (2013).

[10] Aharon Ben-Tal and Arkadi Nemirovski. "Robust Convex Optimization". In: *Mathematics of Operations Research* 23.4 (1998), pp. 769–805.

[11] Aharon Ben-Tal and Arkadi Nemirovski. "Robust Solutions of Uncertain Linear Programs". In: *Operations research letters* 25.1 (1999), pp. 1–13.

[12] Dimitris Bertsimas and Melvyn Sim. "Robust Discrete Optimization and Network Flows". In: *Mathematical programming* 98.1-3 (2003), pp. 49–71.

[13] Dimitris Bertsimas and Melvyn Sim. "The Price of Robustness". In: *Operations research* 52.1 (2004), pp. 35–53.

[14] Dimitris Bertsimas and Aurélie Thiele. "A Data-Driven Approach To Newsvendor Problems". In: (2005).

[15] Omar Besbes and Alp Muharremoglu. "On implications of demand censoring in the newsvendor problem". In: *Management Science* 59.6 (2013), pp. 1407–1424.

[16] J.R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Verlag, 1997.

[17] Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine Learning* 3.1 (2011), pp. 1–122.

[18] Samuel Burer. "On the Copositive Representation of Binary and Continuous Nonconvex Quadratic Programs". In: *Mathematical Programming* 120.2 (2009), pp. 479–495.

[19] *"California Household Travel Survey"*. 2010.

[20] *Can Car2Go Transform New York Into a City of Drivers?* 2015.

[21] Emmanuel J Candès et al. "Robust principal component analysis?" In: *Journal of the ACM (JACM)* 58.3 (2011), p. 11.

[22] *Car2Go*. 2014.

[23] *Car2go brings North America's first all-electric carsharing program to San Diego*. 2014.

[24] *Car2Go San Diego*. 2014.

[25] *Car2Go TRIP PROCESS AGREEMENT (U.S.)* 2014.

[26] J Douglas Carroll and Jih-Jie Chang. "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition". In: *Psychometrika* 35.3 (1970), pp. 283–319.

[27] Robert Cervero and Yuhsin Tsai. "City CarShare in San Francisco, California: Second-year Travel Demand and Car Ownership Impacts". In: *Transportation Research Record: Journal of the Transportation Research Board* 1887.1 (2004), pp. 117–127.

[28] Stephen N Chapman. *The fundamentals of production planning and control*. Pearson College Division, 2006.

[29] *Charging Plug-In Electric Vehicles at Home*. 2014.

[30] Wenqing Chen et al. "From CVaR to Uncertainty Set: Implications in Joint Chance-Constrained Optimization". In: *Operations research* 58.2 (2010), pp. 470–485.

[31] Xin Chen, Melvyn Sim, and Peng Sun. "A Robust Optimization Perspective on Stochastic Programming". In: *Operations Research* 55.6 (2007), pp. 1058–1071.

[32] Susan J Connor. "Process Industry Thesaurus". In: American Production & Inventory Control Society. 1986.

[33] SA Conrad. "Sales data and the estimation of demand". In: *Operational Research Quarterly* (1976), pp. 123–127.

[34] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "A multilinear singular value decomposition". In: *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[35] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "On the best rank-1 and rank-(R 1, R 2,..., Rn) approximation of higher-order tensors". In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1324–1342.

[36] *DriveNow*. 2014.

[37] John Duchi et al. "Efficient projections onto the l 1-ball for learning in high dimensions". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 272–279.

[38] Jonathan Eckstein and Dimitri P Bertsekas. "On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators". In: *Mathematical Programming* 55.1-3 (1992), pp. 293–318.

[39] Laurent El Ghaoui, Maksim Oks, and Francois Oustry. "Worst-case Value-at-risk and Robust Portfolio Optimization: A Conic Programming Approach". In: *Operations Research* 51.4 (2003), pp. 543–556.

[40] C.H. Fine and R.M. Freund. "Optimal investment in product-flexible manufacturing capacity". In: *Management Science* 36.4 (1990), pp. 449–466.

[41] Christoph M Flath et al. "Improving electric vehicle charging coordination through area pricing". In: *Transportation Science* 48.4 (2013), pp. 619–634.

[42] Guillermo Gallego and Ilkyeong Moon. "The distribution free newsboy problem: review and extensions". In: *Journal of the Operational Research Society* (1993), pp. 825–834.

[43] Silvia Gandy, Benjamin Recht, and Isao Yamada. "Tensor completion and low-n-rank tensor recovery via convex optimization". In: *Inverse Problems* 27.2 (2011), p. 025010.

[44] Joel Goh and Melvyn Sim. "Distributionally Robust Optimization and Its Tractable Approximations". In: *Operations Research* 58.4-part-1 (2010), pp. 902–917.

[45] Donald Goldfarb and Zhiwei Qin. "Robust low-rank tensor recovery: Models and algorithms". In: *SIAM Journal on Matrix Analysis and Applications* 35.1 (2014), pp. 225–253.

[46] Linda Green and Peter Kolesar. "The pointwise stationary approximation for queues with nonstationary arrivals". In: *Management Science* 37.1 (1991), pp. 84–97.

[47] *Green Benefits*. 2014.

[48] J.M. Harrison and J.A. Van Mieghem. "Multi-resource investment strategies: Operational hedging under demand uncertainty". In: *European Journal of Operational Research* 113.1 (1999), pp. 17–29.

[49] Richard A Harshman. "Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis". In: (1970).

[50] John A Hartigan and Manchek A Wong. "Algorithm AS 136: A k-means Clustering Algorithm". In: *Applied statistics* (1979), pp. 100–108.

[51] Biyu He et al. "The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem". In: *Manufacturing & Service Operations Management* 14.1 (2012), pp. 99–114.

[52] Woonghee Tim Huh et al. "Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator". In: *Operations Research* 59.4 (2011), pp. 929–941.

[53] U.S. Karmarkar and K. Rajaram. "Grade selection and blending to optimize cost and quality". In: *Operations Research* 49.2 (2001), pp. 271–280.

[54] Karel J Keesman. "Application of flexible recipes for model building, batch process optimization and control". In: *AIChE journal* 39.4 (1993), pp. 581–588.

[55] Paul R Kleindorfer, Kalyan Singhal, and Luk N Wassenhove. "Sustainable Operations Management". In: *Production and operations management* 14.4 (2005), pp. 482–492.

[56] Tamara G Kolda and Brett W Bader. "Tensor decompositions and applications". In: *SIAM review* 51.3 (2009), pp. 455–500.

[57] Qingxia Kong et al. "Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones". In: *Operations Research* 61.3 (2013), pp. 711–726.

[58] Retsef Levi, Robin O Roundy, and David B Shmoys. "Provably near-optimal sampling-based policies for stochastic inventory control models". In: *Mathematics of Operations Research* 32.4 (2007), pp. 821–839.

[59] Baiyu Li, Andrew EB Lim, and Tong Wang. "Estimation and Optimization of Logit Demand Model with Covariates, Missing Data, and Auxiliary Information". In: (2014).

[60] James Li, Jacob Bien, and Martin Wells. *Package 'rTensor', http://jamesyili.github.io/rTensor/.* July 2014.

[61] Michael K Lim, Ho-Yin Mak, and Ying Rong. "Toward Mass Adoption of Electric Vehicles: Impact of the Range and Resale Anxieties". In: *Manufacturing & Service Operations Management* 17.1 (2015), pp. 101–119.

[62] Ji Liu et al. "Tensor completion for estimating missing values in visual data". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.1 (2013), pp. 208–220.

[63] Liwan H Liyanage and J George Shanthikumar. "A practical inventory control policy using operational statistics". In: *Operations Research Letters* 33.4 (2005), pp. 341–348.

[64] Xiangwen Lu, Jing-Sheng Song, and Kaijie Zhu. "Analysis of perishable-inventory systems with censored demand data". In: *Operations Research* 56.4 (2008), pp. 1034–1038.

[65] Ho-Yin Mak, Ying Rong, and Zuo-Jun Max Shen. "Infrastructure Planning for Electric Vehicles with Battery Swapping". In: *Management Science* 59.7 (2013), pp. 1557–1575.

[66] Ho-Yin Mak, Ying Rong, and Jiawei Zhang. "Appointment Scheduling with Limited Distributional Information". In: *Management Science* (2014).

[67] Ho-Yin Mak and Zuo-Jun Max Shen. "Pooling and Dependence of Demand and Yield in Multiple-Location Inventory Systems". In: *Manufacturing & Service Operations Management* 16.2 (2014), pp. 263–269.

[68] Steven Nahmias. "Demand estimation in lost sales inventory systems". In: *Naval Research Logistics* 41.6 (1994), pp. 739–758.

[69] Karthik Natarajan, Melvyn Sim, and Joline Uichanco. "Tractable Robust Expected Utility and Risk Models for Portfolio Optimization". In: *Mathematical Finance* 20.4 (2010), pp. 695–731.

[70] Karthik Natarajan, Chung Piaw Teo, and Zhichao Zheng. "Mixed 0-1 Linear Programs under Objective Uncertainty: A Completely Positive Representation". In: *Operations research* 59.3 (2011), pp. 713–728.

[71] *"North County (San Diego area)"*. 2014.

[72] Georgia Perakis and Guillaume Roels. "Regret in the newsvendor model with partial information". In: *Operations Research* 56.1 (2008), pp. 188–203.

[73] Erica L Plambeck. "OM Forum-Operations Management Challenges for Some Clean-tech Firms". In: *Manufacturing & Service Operations Management* 15.4 (2013), pp. 527–536.

[74] *Pricing the Surge: The Microeconomics of Uber?s Attempt to Revolutionise Taxi Markets*. 2014.

[75] Zhiwei Qin, John Bowman, and Jagtej Bewli. "A Bayesian Framework for Large-scale Geo-demand Estimation in On-line Retailing". In: *Proceedings of 2014 INFORMS Data Mining and Analytics Workshop*. INFORMS. 2014.

[76] Javier Romero et al. "A new framework for batch process optimization using the flexible recipe". In: *Industrial & Engineering Chemistry Research* 42.2 (2003), pp. 370–379.

[77] Cynthia Rudin and Gah-Yi Vahn. "The big data newsvendor: Practical insights from machine learning". In: (2014).

[78] WGMM Rutten and JWM Bertrand. "Balancing stocks, flexible recipe costs and high service level requirements in a batch process industry: A study of a small scale model". In: *European Journal of Operational Research* 110.3 (1998), pp. 626–642.

[79] *SanGIS/SANDAG Data Warehouse"*. 2014.

[80] Herbert Scarf, KJ Arrow, and S Karlin. "A min-max solution of an inventory problem". In: *Studies in the mathematical theory of inventory and production* 10 (1958), pp. 201–209.

[81] Michael Schneider, Andreas Stenger, and Dominik Goeke. "The electric vehicle-routing problem with time windows and recharging stations". In: *Transportation Science* 48.4 (2014), pp. 500–520.

[82] E.W. Schuster and S.J. Allen. "Raw Material Management at Welchs, Inc." In: *INTERFACES* 28.5 (1998), pp. 13–24.

[83] Chuen-Teck See and Melvyn Sim. "Robust Approximation to Multiperiod Inventory Management". In: *Operations research* 58.3 (2010), pp. 583–594.

[84] Alexander Shapiro and Tito Homem-de Mello. "A simulation-based approach to two-stage stochastic programming with recourse". In: *Mathematical Programming* 81.3 (1998), pp. 301–325.

[85] Jia Shu et al. "Models for Effective Deployment and Redistribution of Bicycles within Public Bicycle-Sharing Systems". In: *Operations Research* 61.6 (2013), pp. 1346–1359.

[86] Herbert A Simon. *Models of Man: Social and Rational*. Wiley, 1957.

[87] *Sources of Greenhouse Gas Emissions*. 2014.

[88] Robert Tibshirani et al. "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1 (2005), pp. 91–108.

[89] Giorgio Tomasi and Rasmus Bro. "A comparison of algorithms for fitting the PARAFAC model". In: *Computational Statistics & Data Analysis* 50.7 (2006), pp. 1700–1734.

[90] Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. "Estimation of low-rank tensors via convex optimization". In: *arXiv preprint arXiv:1010.0789* (2010).

[91] Ledyard R Tucker. "Some mathematical notes on three-mode factor analysis". In: *Psychometrika* 31.3 (1966), pp. 279–311.

[92] *U.S. Emissions*. 2014.

[93] J.A. Van Mieghem. "Investment strategies for flexible resources". In: *Management Science* 44.8 (1998), pp. 1071–1078.

[94] D Weinberg. "US Neighborhood Income Inequality in the 2005–2009 Period". In: *American Community Survey Reports. US Census Bureau, Washington* (2011).

[95] *What Is Surge Pricing And How Does It Work?* 2014.

[96] Ward Whitt. "Open and closed models for networks of queues". In: *AT&T Bell Laboratories Technical Journal* 63.9 (1984), pp. 1911–1979.

[97] Ward Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, 2002.

[98] Michael J Wills. "A Flexible Gravity-Opportunities Model for Trip Distribution". In: *Transportation Research Part B: Methodological* 20.2 (1986), pp. 89–111.