Profit-Driven Network Redesign Through Value-Creation Services

by

Morgan Jessica DeHaan
Bachelor of Science in Business Administration, Supply Chain Management and Marketing,
University of Illinois Urbana-Champaign, 2018

and

Yujia Ke
Master of Science in Civil Engineering, University of California Berkeley, 2017
Bachelor of Engineering in Civil Engineering, Huazhong University of Science and Technology, 2016

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Signature of Author: _____
Morgan Jessica DeHaan
Department of Supply Chain Management
May 12, 2023

Signature of Author: _____
Yujia Ke
Department of Supply Chain Management
May 12, 2023

Certified by: _____
Dr. Milena Janjevic
Research Scientist, Center for Transportation and Logistics
Capstone Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

(This page is intentionally left blank)

Profit-Driven Network Redesign Through Value-Creation Services

by

Morgan Jessica DeHaan

and

Yujia Ke

Submitted to the Program in Supply Chain Management
on May 12, 2023 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Design of supply chain networks is a key strategic decision in supply chain management. Our capstone sponsor Armada is a supply chain service provider to restaurant chains across the United States. The company is envisioning a redesign of their current service network based on two components: (1) the addition of new distribution centers (referred to as iDCs), located closer to high volume service areas, (2) the deployment of value-creating services from the iDCs. In this capstone, we develop a series of models to support these decisions. First, we build a model that identifies demand dense areas as “urban clusters” where an iDC would best serve Armada’s clients. Second, we develop a model that minimizes costs in the network while maximizing revenues through the defined additional services. The results of these models provide us with two distinct network configurations based on cost-minimization and profit-maximization: one placing iDCs by key demand-dense areas and the other favoring high revenue generating regions. This study shows that layering revenue-maximization methodology with cost-minimization algorithms in mixed integer linear programming will alter results and favor the highest revenue generation location(s).

Capstone Advisor: Dr. Milena Janjevic
Title: Research Scientist, Center for Transportation and Logistics

ACKNOWLEDGMENTS

We would like to extend the biggest thank you to our capstone advisor, Milena Janjevic, for her consistent support, guidance, and insights as we worked through our research capstone. Thank you for challenging us to be the best researchers we could be and helping us achieve this major accomplishment.

Thank you to our capstone sponsor company, Armada, for presenting us with an engaging problem that challenged the way we typically approach network design. Thank you to Joe, Steve, and Phil for their consistent support throughout the academic year.

Finally, a huge thank you to our family and friends who supported us throughout our time at MIT. Whether you were near or far during out studies, your unconditional love and unwavering encouragement were felt each and every day. We could not have accomplished this achievement without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This research project required a strong understanding of the industry our sponsor company serves and the company's service offering. In this chapter, we dive into the motivation for the project. This includes an overview of our sponsor company, Armada, and the industry they serve, the restaurant food service industry. Then we cover the proposed problem statement and the research questions identified. Finally, we set forth our goals for the project and our expected outcomes.

1.1 Motivation

Armada is a supply chain service provider located in the United States. The primary solutions and services offered by Armada are transportation, logistics, supply chain planning and warehousing of goods. Specifically, Armada's expertise is in the restaurant industry, with their largest clients being fast-food and/or chain restaurants across the globe. Their mission is to continuously improve global supply chains while remaining true to their key values of simplicity, transparency, and extraordinary service to their clients (Armada, "Mission and Vision").

Armada currently operates a network of distribution centers, from which last-mile deliveries to their clients are organized. In order to improve client service, Armada is contemplating the implementation of additional distribution centers ("iDCs"). The distinction between iDC and traditional distribution centers lies in two features: 1) iDC is proximate to clients and is responsible for more frequent deliveries (daily), while a typical distribution center is in a suburban area and distributes about twice per week; 2) iDC can offer additional services to restaurants beyond deliveries such as reserving inventory, fresh food solutions, food preparation, and reverse materials logistics. Note that "iDC" refers both to these new types of facilities and the name of Armada's internal project which oversees their implementation.

The iDC project fits into Armada's vision statement of "there's a better way" because the project goal is to expand the current utilization of distribution centers and service frequency to better fit the needs of the clients (Wildes 2022).

Furthermore, the industry that Armada is in is critical to our society and economy. Armada's focus is the restaurant supply chain: their customers' value is driven by the food and service they provide. Armada must be mindful of this as they work on the network design of their large brands. If one of their clients has a value of "fresh, never frozen" foods, Armada must ensure that this value is incorporated throughout the supply chain, from the transportation to the storage and distribution of goods. Armada's industry is niche and demanding; therefore, the successful deployment of the iDC will be crucial to maintaining Armada's superior client service, while not disrupting the quality of the products their clients serve.

Although the iDC project is still in a conceptual planning phase, its distinctive functions, Armada's internal mission, and the current market state collaboratively drive Armada towards iDC network design and implementation. The company is attempting to transform from a conventional last mile delivery provider to a customer-oriented supply chain value creator.

1.2 Problem Statement and Research Questions

With the motivation to provide better restaurant-specific last mile deliveries, Armada's primary concern is determining how to design urban last-mile networks incorporating iDCs, to improve performance of existing distribution networks and to excavate potential value-creation services as efficiently as possible.

The initial hypothesis proposed by Armada, was to locate iDCs near urban restaurant clusters with high demand volume. This strategy would facilitate more frequent deliveries and more agile

responses to clients' requirements. However, it is not a quantitative approach that justifies why a certain location A instead of location B would be an ideal location for iDC. This leads to our first two research questions (RQ) in this capstone:

- RQ1: How can Armada quantitatively identify urban clusters in their current distribution network that will be better served by an iDC than a traditional DC?
- RQ2: What kind of quantitative method should Armada use to choose iDC locations and which are the resulting locations?

Moreover, Armada has limited insights into what cost and service levels to expect with this method. Therefore, Armada needed to utilize network mapping and cost optimization to evaluate the locations most appropriate for these futuristic distribution centers. This leads to the following research question:

- RQ3: What relationship exists between cost and distance between iDC and its location?

Armada would like to improve the efficiency and effectiveness of their distribution network with service levels and proximity and understand the value of constructing iDCs. In addition, they want to prioritize the value-creation services they can offer from the new iDCs. Armada has accumulated data on market size and value in dollars for each of the identified value-creation services from preliminary internal discussion. This data was a strong starting point for further analysis; however, service level agreements for these services will be necessary if Armada wants to proceed with opening iDCs to take on these tasks. Take a diner as an example: if an inventory reservation service costs $3,000 per year, we must know what service time and frequency (e.g., 60 minutes daily) Armada could offer to guarantee a contract. Armada will need to further evaluate the potential value-creation services they are interested in providing through the new distribution centers and determine quantitatively which services will be most feasible in the new model. All of this leads to our final research question:

- RQ4: How can Armada incorporate anticipated value-creation services into the quantitative method?

1.3 Project Goals and Expected Outcomes

The capstone project's overall goal is to provide Armada with a comprehensive, repeatable, and modularized decision methodology of iDC network design. It takes the research questions into account, as well as support simulation analysis and scenario planning analysis. Simulation analysis refers to the comparison between model results of different optimization objectives, while "what-if" scenario analysis indicates what would happen if some input parameters were changed.

The methodology consists of three main modules:

1. Network analysis focused on the identification of demand clusters that define urban areas and thus, potential iDC locations
2. Network design model focused on optimizing costs with Armada's current product-service offering and an iDC network
3. Network design model focused on maximizing revenue, considering an extended product-service offering (i.e., value-creation services) and an iDC network

The first module answers research question 1, where we hypothesize that some cluster algorithms would be useful for classification of urban areas. For example, a revised KMeans algorithm could be utilized to quantitatively address the identification of urban clusters in the current distribution network that will be better served by an iDC. Once urban areas are clearly defined, we can further narrow down the set of potential candidates for an iDC with the following steps.

The second module addresses research questions 2 and 3. We hypothesize that a data-driven optimization method would be a good fit to the network design problem, which includes but is not limited to mixed integer linear programming (MILP) and meta heuristics models. We reviewed literature in this area and refined our approach iteratively. Simulation and scenario analysis could be done by re-running our model multiple times with various objectives and inputs. Armada would ideally like to implement multiple iDCs, after evaluating multiple iterations of the approach.

Lastly, the third module addresses our research question 4; we further hypothesized that optimization methodology will again be used to evaluate capacity and labor constraints of new iDCs when evaluating which value-creation services should be recommended at each iDC location. We used Armada's revenue estimations by region and value-creation service, along with percentage of probability, to perform stochastic optimization. Combining the results of cost minimization and revenue maximization, our output identifies the top contending urban areas for iDC deployment.

We wrote a report summarizing our insights and offering specific recommendations based on Armada's business preference. We used open-source geographical visualization packages or tools such as Echarts and OpenStreetMap to display our results.

The deliverables to the company include:

1. A methodology for a data-driven decision process for iDC network design, with support of multiple objectives and change of inputs; and
2. A report with location visualizations that explicitly demonstrates how the methodology improves performance.

This capstone is organized as follows: Chapter 2 is a state of the art, exploring relevant literature and methodologies related to the project. In Chapter 3 we meticulously describe our methodology and data analysis. We experimented with different approaches and consolidated our research to identify the best methodology to achieve our project goals. Beyond the methodology, we share our models for both cost-minimization and profit-maximization. Finally, in Chapter 4 we discuss the results of each of our models' outputs, and dive into a reflection on the results in the discussion chapter (5).

2. STATE OF THE ART

Armada's goal is to determine how to design urban last-mile networks incorporating iDCs in order to improve performance of existing distribution networks and to excavate potential value-creation services as efficiently as possible. Since the problem Armada would like to solve is context-specific, we can decompose the iDC case into general features/components, analyze and survey each of them, and put forward our methodology. Corresponding to the network analysis model focused on identifying demand clusters defining urban areas, we researched existing demand clustering models. For the module addressing network design around optimizing costs, we set out to understand basic facility location configuration and cost minimization while also appropriately accounting for missing cost data to build into the module. Similarly, the network design module focused on revenue maximization expanded upon network configuration and evaluated offering additional services that generate a profit and maximize results.

To synthesize, our problem set has the following research components:

1. Identification of various demand clusters and potential iDC locations. To address these components, a review of literature in this area is detailed in Section 2.1. We will review approaches to clustering analysis and assess how it ties with the model's scalability.
2. Facility location/network configuration with the current service-level offering and focus on cost minimization. Facility location decisions have been identified as strategic and having a profound effect on supply chain management, and there has been tremendous research in the area (Melo et al., 2009). To present literature in this area, we will proceed in several steps. In Section 2.2, we will first review literature on the basic facility location problem, which aims to determine location choices and covering choices of iDC, with considerations of both fixed and variable cost. In Section 2.3, we will review literature on the facility size problem, allowing us to determine the size/area of location candidates and the delivery

capacity constraints of the iDC and current warehouses. Given the potentially large-scale nature of the problem, we will also focus on methods to reduce the scale of our facility location problem to an acceptable solving time.

3. Identification of optimal network configuration with extended services and focus on profit maximization. A review of literature in this area is proposed in Section 2.4, where we will explore the limited literature on profit maximization with respect to multi-objective facility location models. We will draw from many pieces of literature around profit-maximization, but it is important to note that we will not directly apply these models, but rather forge our own approach to fit our model's requirements.

2.1 Demand Clustering

The number of demand locations and the number of facility location candidates are huge (~46,000 zip codes in the US), and thus some scale challenges must be addressed through individual analysis. The question is: how to prune nearly impossible location candidates to make the problem more solvable?

Clustering analysis is a common methodology for solving capacitated location-routing problems, and varying techniques exist to reach a solution, according to Miranda et al. (2011). A significant amount of research exists to demonstrate each of these techniques, but at the highest level exists an optimization model and specific decision variables, with the objective to optimize a performance indicator.

Given that Armada did not pre-determine the potential demand-dense locations, nor identify urban areas to focus the analysis on, demand clustering is a significant factor to build into our network design for the project. There are many clustering models to consider, each with unique qualities. Thus, our methodology will review three models and we will select the best performing model in

coordination with our optimization model. The three models we will test are K-Means, Mean Shift and DBSCAN. Given our methodological understanding and research on the capacitated facility and facility size location problem above, we propose a K-means clustering method that, under capacity restrictions, optimizes the location-allocation quality (Liao and Guo, 2008, p. 335).

It is important to consider demand clustering as it has a direct correlation to economies of scale and scope in the network design industry (Sheffi, 2013, p. 482). Most important to the iDC problem statement is the economies of scale, such that high density demand clusters are most cost effective for daily shipments. According to Mesa-Arango and Ukkusuri (2015), the definition of clusters has an added layer of complexity brought forth by revenues that is often missed. This is instrumental in evaluating demand clusters for Armada given that a major goal of the iDC project is revenue driven.

2.2 Basic facility location problem

Based upon Mak and Shen (2016), there are five classic facility location problems:

1. Set Covering Problem,
2. Maximum Covering Problem,
3. P-center Problem (Min-Max Problem),
4. P-dispersion Problem (Max-Min Problem), and
5. P-median Problem (Min-Sum Problem).

All these problems can be typically modeled as integer linear programming with some constraints. The set covering and P-median problem are most relevant in our context as explained below.

Set Covering Problem

Suppose that there are a set of demand locations as well as a set of facility location candidates, and service ranges of the candidates are known. The problem is to choose certain candidates to minimize total cost, subject to all demand locations being covered.

P-median Problem (Min-Sum Problem)

Suppose that there is a set of demand locations, and the number of customers of each location and the total number of facilities are known. The problem is to choose candidates from demand locations and their service assignment for demand locations to minimize the summation of distance between customers and facilities.

According to these basic facility location problems, we can further break down the problem into building blocks from aspects including decision variables, objectives, and constraints (Table 1).

Table 1. *Comparison of Set Covering and P-Median Problems*

Building blocks		Set covering	P-median	**Armada Problem**
Decision variables	Binary choice	√	√	√
	Covering choice		√	√
Objectives	Min Fixed cost	√		√
	Min Variable cost		√	√
Constraints	Location number		√	√
	Covering demand	√	√	√

From Table 1, it is clear that our problem is a hybrid of set covering and p-median problems, so we will utilize modeling techniques from both.

2.3 Capacitated facility location and facility size problem

From classical facility location problems to capacitated ones, there is a vast body of literature regarding their applications, modeling, and algorithms. We focus on the major review papers on the topic.

Sridharan (1995) formulates the capacitated plant location problem (CPLP) as a mathematical form and summarizes various algorithms to solve the problem. Specifically, the formulation is a mixed integer linear programming model, with binary decision variables on whether plants are built or not, and continuous decision variables on how much plants supply demand nodes. The way he treated capacity is to add constraints, in which the summation of supply from a plant should be no more than a given number. Wu et al. (2006) brings a flexible setup cost function to the objective function to allow for more complex cost structures than a fixed cost. However, they use the same methodology to deal with capacity limitations.

Melkote and Daskin (2001) introduces a more general capacity facility location problem (CFLP) and proposes adding valid constraints to improve solving efficiency. The authors subtly model the CFLP as a network flow problem; thus, there is no need to assume hierarchies in supply chains. Similarly with the previous two works, they use constraints to indicate capacity requirements.

Facility sizing problems are highly correlated with the facility capacity problems just discussed. Sankaran and Raghaven (1997) came up with a multi-type facility location model in the context of liquefied petroleum gas distribution, which is similar to CFLP except that there are K types of discrete sizes of facilities for every location (e.g., large, medium, small). Each type would have its own capacity constraint. Karatas and Dasci (2020) use similar treatment in a two-level facility location problem with non-linear constraints.

2.4 Profit-maximization facility location problem

Facility location problems are endogenously multi-objective, though much of the literature has focused on cost minimization. Current et al. (1990) reviews 45 related papers and summarizes them into 4 categories of objectives. The most frequent category is cost minimization or its proxy (e.g., distance and transportation time). The second most frequent category is demand coverage. Profit maximization and environmental concerns are the others. Similarly, Melo et al (2009) does an extensive review and concludes that the proportions of cost and profit objectives are 75% and 16%, respectively.

Profit objectives are commonly used in the private sector because most business activities are profit oriented. Typically, two kinds of profit equations have been discussed: 1) revenue minus costs, and 2) after-tax profit (Melo et al, 2009).

Note that if we assume static and deterministic demand, which must be fully fulfilled and uniform price, the revenue would be constant and the objective of profit maximization reduces to cost minimization (Klibi et al, 2010). Thus, revenue can be differentiated in several ways.

Various Prices

In this scenario, prices can be different. Ross and Soland (1980) and Soland (1983) assume prices are varied by location by demand point, and encompass profit into a minimization objective function, where cost terms have positive signs and revenue terms are negative. The way they model revenue is to multiply fulfilled demand and price of its assigned location. Mukundan and Daskin (1991) assumes that prices are different across locations and models the problem as maximizing profits (revenue minus variable and fixed cost). An interesting component that they add is to simultaneously determine facility sizes (investment levels).

Partially Fulfilled Demand

In this scenario, demand can be partially met. If we choose to fulfill all the demand, the total fixed and variable cost would be high; on the other hand, if we avoid filling demand, the cash flow would be small, and the business cannot survive. The problem becomes to find the optimal fraction of demand to serve to maximize the profit. Kouvelis and Rosenblatt (2002) provides a complicated model on this case.

Service Sensitive Demand

In this scenario, demand can be a function of facility location and allocation. Klibi et al (2010) defines order winner attributes as attributes that have impact on demand, including but are not limited to lead time, product portfolio, and price. If we can describe the relationship between demand and these order winner attributes, a comprehensive model can be constructed to maximize profits.

To summarize, problems with profit maximization objectives have either various price or different demand properties, as shown in Table 2.

Table 2. *Comparison of Profit-Maximization Objective Literature*

	Demand property			Price property	
	Static	Dynamic	Fully Fulfilled	Uniform	Various
General problems with cost minimization objective	√		√	√	
Ross and Soland (1980)	√		√		√
Soland (1983)	√		√		√
Mukundan and Daskin (1991)	√		√		√
Kouvelis and Rosenblatt (2002)	√			√	
Klibi et al. (2010)		√		√	√

Table 2 demonstrates the importance of drawing from each of the research pieces represented for Armada's profit-maximizing model. However, in Armada's problem, the key distinction is that there are indeed two sources of demand rather than one. The first is its main business, which is static and needs to be fully fulfilled, while the second is additional value-creation services, which are facility location dependent. Thus, we cannot directly apply models from the literature.

3. DATA AND METHODOLOGY

Based on the literature reviewed, our overall problem of identifying demand-dense urban clusters and allocating iDCs to high impact locations was addressed by a hybrid approach:

In Section 3.1 Data Analytics, we explore data sets on hand and introduce how we cleaned them. Based on the cleaned data, we make some assumptions on specific data usage. Additionally, we propose and implement a K-Nearest-Neighbor based data inference method to fill missing transportation cost data of currently unused legs.

In Section 3.2 Demand Clustering, we first analyze demand distribution on zip code level and revise three different clustering algorithms to accommodate our case. Detailed comparisons between the original and revised algorithms are presented and highlighted.

In Section 3.3 The Models, we formulate the primary facility location models. First, we propose a cost minimization model as a baseline model. It combines p-median and set covering building blocks. Since there are no explicit capacity constraints, we focus on a multi-commodity (multi-category) model rather than a capacitated one. Upon that model, we build a profit maximization model. The distinction between the baseline model and the profit-maximization model is the introduction of potential revenues of value-creation services, and total profit (revenue minus cost) as its optimization objective.

We demonstrate how we approach the problem step by step, while in Chapter 4 Results and Analysis, our results are presented in the same structure.

3.1 Data Analytics

In this section we will cover the data analysis that went into each methodology. We first review and describe the data received from our sponsor company in Section 3.1.1. Section 3.1.2 covers the data preprocessing we did to prepare our data for modeling, followed by the statement of our data assumptions (3.1.3). Finally, Section 3.1.4 covers the missing cost data inference we made to fill in for any missing data in our datasets.

3.1.1 Data Description

We received data from Armada for four current clients and collected geographic data from the public source ListenData.com. The data information is summarized in Table 3.

Table 3. *Company Provided Data Sources*

Data name	*Source*	*Description*
iDC-Warehouse Sales (Four clients)	Armada	By supplier, warehouse, temperature, category, and product; Annual sales, unit cost, unit fee.
Store Sales (Four clients)	Armada	By warehouse, store, temperature, category, product; Annual sales.
Items and Current Pricing (Four clients)	Armada	By warehouse, store, temperature, category, product; Unit cost, unit fee.
iDC Profitability (Aggregated statistics)	Armada	By value-creation scenario, services, regions; Annual profit.
zip_to_lat_lon_North America	Public[1]	By zip code, city, county, state/province, country; Latitude, longitude.

[1]Note: https://www.listendata.com/2020/11/zip-code-to-latitude-and-longitude.html

The relationship among Armada's data sets is illustrated in Figure 1, depicting the two legs of deliveries.

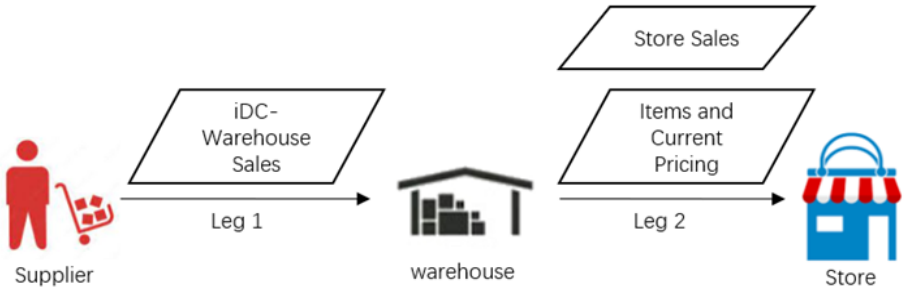

Fig. 1, *Data relationships*

The iDC Profitability data set is not included in Figure 1. It is an aggregated statistics of potential annual profit of different value-creation services. The data is aggregated at region level, which contains north-east, south-east, middle-west, and so on. Because it is not at the leg level, we process it in a different way. The processing steps include joining with region and zip code data, aggregating to get final annual results on scenario and zip code level.

The key dataset statistics overview is reported in Table 4, and the full reports can be retrieved in Appendix A.

Table 4. *Key Dataset Statistics*

Data name	*No. of vars.*	*No. of obs.*	*Missing cells*	*Missing cell (%)*	*Duplicate rows*	*Duplicate rows (%)*	*Categorical*	*Numeric*
iDC-Warehouse Sales (Four clients)	33	60,698	288,555	14.4%	44	0.1%	12	21
Store Sales (Four clients)	12	19,362,856	40,320,727	17.4%	427,710	2.2%	11	1
Items and Current Pricing (Four clients)	32	157,507	1,358,118	26.9%	5,949	3.8%	14	18

3.1.2 Data Preprocess

For the first three datasets, we implemented four steps to preprocess our data: combining data from various files, dealing with missing values, dealing with duplication, maintaining consistency of zip code. Detailed data preprocess reports are in Appendix A.

For the aggregated statistics of iDC profitability, we extended region-wise annual profits to zip code-wise and used it in a profit-based network design model in Section 3.3.2 and 4.3.

3.1.3 Assumptions

Based on the data sets, some assumptions had to be made to bridge the gap between what we knew and what the analytical models required:

1. Unit cost: Per Armada, unit delivery cost is the sum of pickup price, freight cost, and total markup. For the warehouse-to-store stage, we used cost columns with the “ReD” prefix.
2. Multiple commodities: In demand clustering analysis, we aggregate demand for various SKUs into the total number (annual cases), and derived demand-dense areas. In network design analysis, we aggregated demand by “TEMP” and developed a multicommodity network design model, which is not too trivial and widely considered in the storage and delivery process. All related costs of SKUs are also aggregated and averaged by “TEMP”. The storage temperature is categorized as three types: dry, cooler, and freezer.
3. Unknown cost data: Since we did not have delivery cost for any iDC candidate, we proposed the use of K Nearest Neighbor (KNN) algorithm to calculate average delivery costs for all iDC candidates and demand zip code pairs and used it as a proxy.

3.1.4 Missing Cost Data Inference

As illustrated in Figure 2, to infer cost data from the supplier to all iDC candidates and from iDC candidates to clients' restaurants, we used K Nearest Neighbors (KNN) algorithm to fill in missing data.

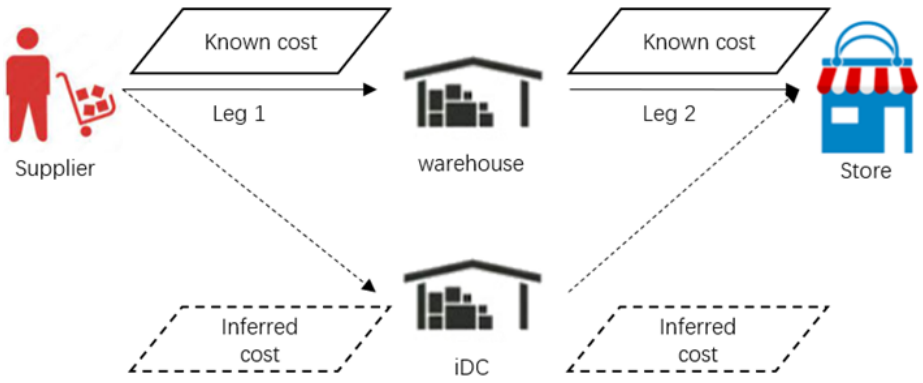

Fig. 2, *Missing Cost Data Inference*

Specifically, for Leg 1 we found the three nearest existing warehouses of each iDC candidate and calculated the average cost from supplier to those warehouses as a proxy for cost of each iDC.

For Leg 2, we first tried to estimate the correlation between existing unit delivery cost and distance. The scatter plot in Figure 3 indicates there is no significant linear correlation. The most likely reason is that consumed SKUs in different areas are distinct. Thus, we further investigated ways to address this issue.

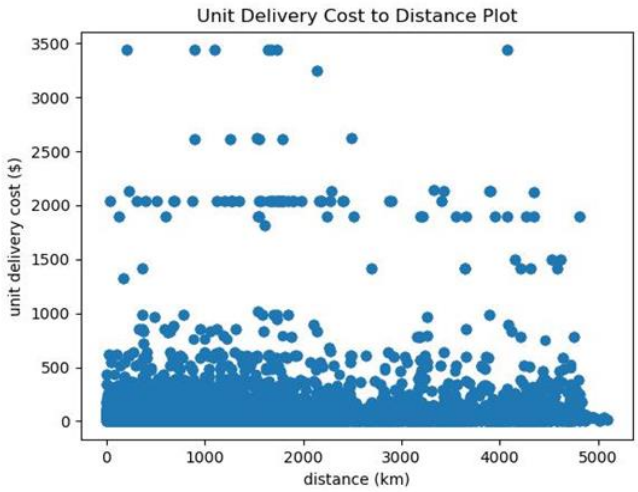

Fig. 3, *Overall Correlation Between Unit Delivery Cost and Distance*

We found two causes for this situation:

1. In our cost structure of pickup price, freight cost, and total markup, only freight cost is relevant to logistics and distance;
2. For different origin-destination pairs, the number of items/products varies dramatically; thus, we cannot aggregate the cost and then calculate the correlation between transportation distance and freight cost. Otherwise, the aggregated cost will be inaccurate because of outliers (some products have few samples which are not representative, see Table 5). Instead, it needs to be done in reverse.

Table 5. *Current DC with Serving Zip Codes and Item Count*

DC Zip Code	*Demand Zip Code*	*Item Count*
32809	32806	501
13748	99999	496
60133	99999	474
92507	99999	473
33844	99999	461
75160	99999	455
…	…	…
28027	60455	1
75071	77833	1
43137	07108	1
43137	7470	1
75071	77437	1
60446	62024	1

As a result, we iterated over the SKU set. For some high-volume SKUs, the correlation between transportation distance and freight cost per case is significant (Figure 4).

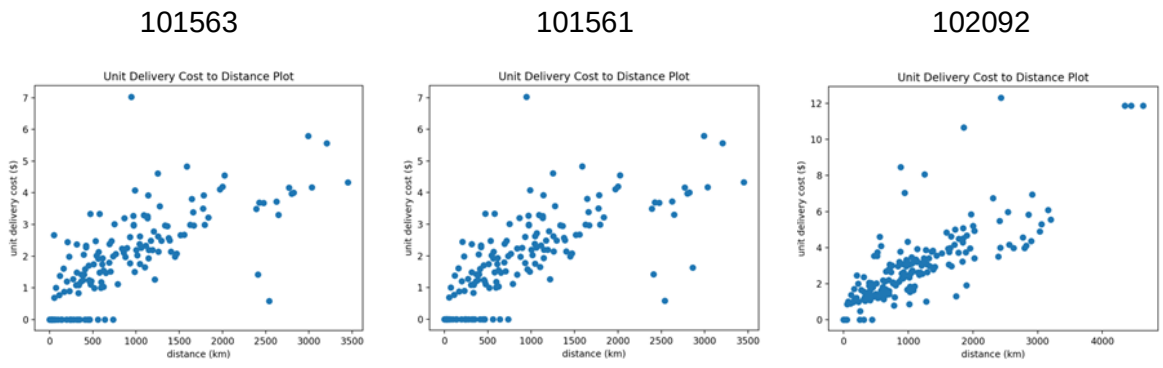

Fig. 4, *Significant Correlation of Sample High-Volume SKUs*

However, for some other SKUs the correlation is unclear (Figure 5).

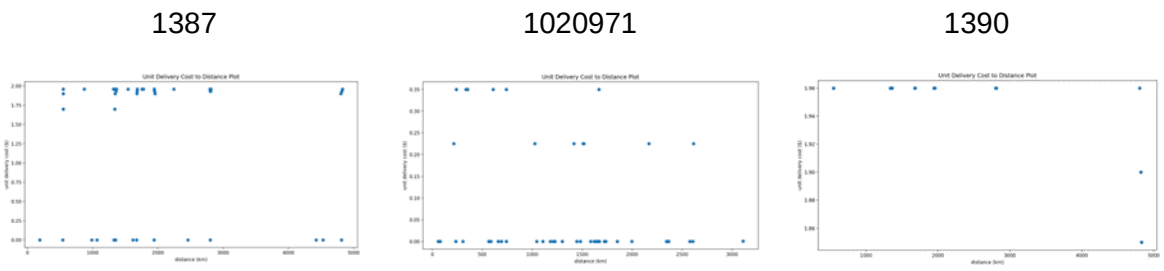

Fig. 5, *Insignificant Correlation of Sample SKUs*

Confronted with this situation, several methods can be used to find the aggregated freight cost per case per mile. The first method is to group SKUs with similar cost coefficients and estimate a common cost coefficient among these products. We tried using grouping criteria like temperature and product categories, but the result was not ideal. Every time we combined data points of grouped SKUs, the overall shapes of scatter plots became irregular. The alternate method is to estimate cost coefficients of SKUs one by one, and to test the significance level. High significance level indicates to use cost coefficients, and vice versa. The flow chart in Figure 6 shows the method we adopted, the process of estimating cost coefficients of SKUs one by one and testing the significance level, and the output is freight unit cost per case per mile by product storage temperature.

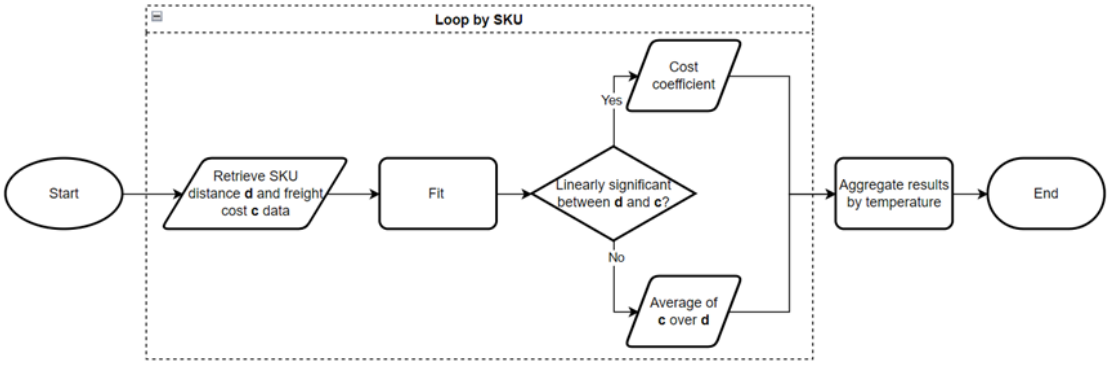

Fig. 6, *Freight Cost Inference Flowsheet*

The process shown in Figure 6 only deals with the freight cost. In terms of pickup price and total markup, we directly apply KNN (where k also equals 3) to infer them. The output is the inferred pickup price and total markup for all legs from iDCs to stores. The last step for unknown cost between iDC and stores would be the sum of the two pieces together (Table 6).

Table 6. *Missing Cost Data Inference Summary*

	Unknown cost from suppliers to iDCs	*Unknown cost from iDCs to stores*		
		Freight cost	*Pickup price*	*Total markup*
Inference methods	KNN	Linear fit Significance test	KNN	KNN

3.2 Demand Clustering

In this section, we first analyze and understand the demand characters of Armada, and then propose three different but compelling clustering algorithms on the demand.

3.2.1 Demand Distribution

In our analysis, we wanted to answer the question: among all zip codes, what are the urban areas or, in other words, demand-dense areas? Plotting the Pareto cumulative curve between the demand and zip codes revealed that the demand distribution is skew and has a long tail. The first 23% percentage of top zip codes contribute to 50% percentage of total annual demand, and the first 52% percentage contribute to 80% percentage (Figure 7). The demand heatmap also indicates the distribution is unbalanced, and we can see from the distinct colors that it is feasible to identify demand-dense areas (zip codes) in Figure 8.

We utilized clustering methods to isolate demand-dense areas from all areas. The resulting areas would serve as new iDC candidates in further network design analysis. Our methods

provide a byproduct, a set of clusters of market demand in Armada's business. We think this information will be useful for Armada when designing their future market segmentation and strategy.

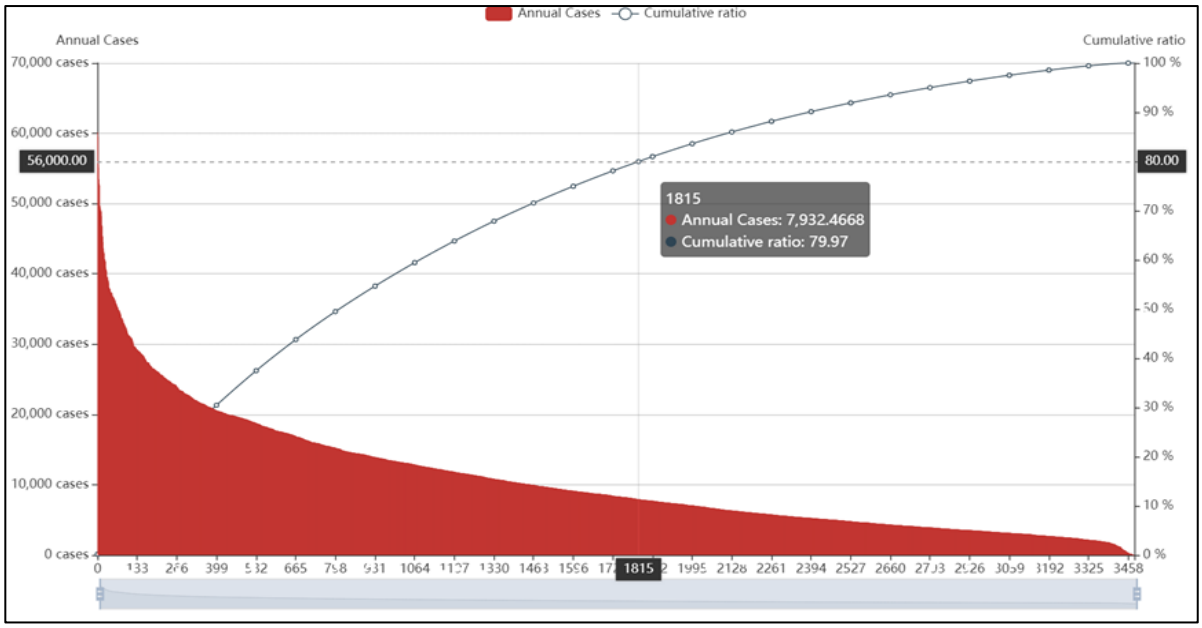

Fig. 7, *Cumulative Demand Illustration*

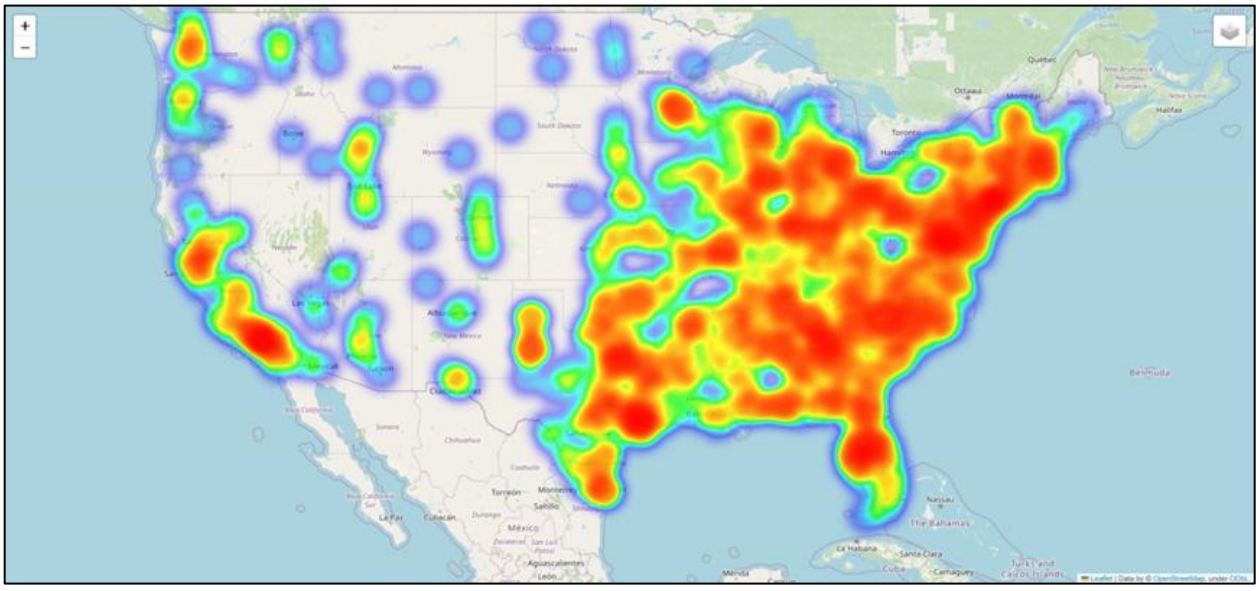

Fig. 8, *Demand Heat Map*

3.2.2 Revised KMeans Algorithm

KMeans clustering is a distance-based iterative algorithm to find k best clusters and their centers. However, it does not take sample weight and filtering into account. In this analysis, we modify the classic KMeans to tackle weight and filtering. The original and revised pseudo-code are presented in Table 7.

Table 7. *Revised KMeans Algorithm Summary*

KMeans	*Revised KMeans*
1. Initialize k cluster centers randomly; 2. Assign each demand zip code to its nearest center; 3. Update locations of centers; 4. Repeat Step 2 and 3 until convergence.	1. Order demand nodes by descending demand, initialize k cluster centers with top k nodes; 2. Assign each demand zip code to its nearest center; 3. Update locations of centers, each node is **weighted by its demand**; 4. Repeat step 2 and 3 until convergence; 5. Set demand **coverage threshold**, and within each cluster, keep nodes below it.

Another component of using KMeans is that we need to determine a good parameter k to achieve the expected results. We decided to use the Elbow Curve Method to find an appropriate k. The method involves iterating over a range of k (e.g., from 2 to 10), recording total distance to centers of all demand nodes, and plotting every k and its associated total distance to find the k whose total distance starts to fall slowly. This specific result is discussed in Section 4.1.1.

3.2.3 Revised Mean Shift Algorithm

Mean Shift clustering is a non-parameter iterative method to find clusters and their centers. It focuses on moving towards denser areas and dealing with multi-feature spaces. Nevertheless, Mean Shift has similar limitations as KMeans; thus, we also modified it to accommodate our

scenario. The original and revised pseudo-code are presented in Table 8. The result is discussed in Section 4.1.2.

Table 8. *Revised Mean Shift Algorithm Summary*

Mean Shift	*Revised Mean Shift*
1. Initialize a cluster center randomly; 2. Assign each demand zip code within predefined bandwidth to the center, and increment the probability of belonging by 1; 3. Update the location of the center by sum all vectors between nodes in step 2 and the center; 4. Repeat Step 2 and 3 until convergence; 5. If the current center is close to some existed center, combine them; 6. Repeat Step 1 to 5 until all nodes are visited; 7. Reassign each node to the center with its largest probability of belonging.	1. Initialize a cluster center randomly; 2. Assign each demand zip code within predefined bandwidth to the center, and increment the probability of belonging by 1; 3. Update the location of the center by sum all vectors between nodes in Step 2 and the center, each vector is **weighted by the demand** of its node; 4. Repeat Step 2 and 3 until convergence; 5. If the current center is close to some existed center, combine them; 6. Repeat Step 1 to 5 until all nodes are visited; 7. Reassign each node to the center with its largest probability of belonging; 8. Set demand **coverage threshold**, and within each cluster, keep nodes below it.

3.2.4 Revised DBSCAN Algorithm

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It is different from our previous methods, for it defines a cluster as a maximum set of nodes that connect with each other due to both proximity and high density. The algorithm does not require us to specify the number of clusters, and it can capture clusters in random shapes. However, the algorithm in its original form still does not take sample weight and demand filtering into consideration. Thus, we proposed a modified DBSCAN. The original and its pseudo code are presented in Table 9.

Table 9. *Revised DBSCAN Algorithm Summary*

DBSCAN	*Revised DBSCAN*
1. Initialize at an unvisited demand node whose bandwidth contains at least prespecified minimal points (neighbors), and mark it as a core point in the current cluster; 2. Check along the neighbors one by one; for each neighbor, if there are at least minimal points within its bandwidth, it is included in the cluster and marked as a core point, if not, it becomes a boundary point of the current cluster; 3. Repeat Step 2 until the current cluster is bounded by boundary points; 4. Repeat Step 1 to 3 until all nodes are visited and clustered.	1. Initialize at an unvisited demand node whose bandwidth contains at least prespecified minimal points (neighbors), and mark it as a core point in the current cluster; 2. Check along the neighbors one by one; for each neighbor, if there are at least minimal points within its bandwidth (the number of points is the sum of **weights/demand**), it is concluded in the cluster and marked as a core point, if not, it becomes a boundary point of the current cluster; 3. Repeat Step 2 until the current cluster is bounded by boundary points; 4. Repeat Step 1 to 3 until all nodes are visited and clustered; 5. Set demand **coverage threshold**, and within each cluster, keep nodes below it.

The DBSCAN algorithm starts with one core point in a new cluster and extends it by including "close and dense neighbors" until the cluster is surrounded by the boundary of the dense neighborhood. This specific result is discussed in Section 4.1.3.

3.3 The Models

Using the identified urban zip code clusters, we wanted to conduct Armada's iDC network design in our models. We addressed the challenge by the following approach:

1. *Infer missing cost data*

   Since we only have cost data from suppliers to existing warehouses and from existing warehouses to customers, the data of iDC candidates should be inferred. The method we propose is K Nearest Neighbors algorithm (see Section 3.1.4);

2. *Solve cost minimization model*

We first designed an Armada delivery network with an objective to minimize total delivery cost, subject to some constraints;

3. *Solve profit maximization model*

   Since iDCs are expected to provide other value-added services according to Armada's long-term vision, we will redesign Armada delivery network with an objective to maximize total profit, subject to certain constraints. The revenues that are earned by iDCs will be estimated by market forecast.

3.3.1 Cost Minimization Model

Our initial model focuses on minimization of cost. The full model is presented below.

Indices:

I Aggregated client (zip code) index

J_1 iDC Candidate facility location index

J_2 Existing facility location index

$J_1 \cup J_2$ Candidate facility location (iDC) index

K SKU storage temperature category index, namely dry, cooler, and freezer

Parameters:

s_{jk} Unit variable cost of delivering product k from supplier to location j

c_{ijk} Unit variable cost of delivering product k from location j to client i

d_{ik} Demand of product k of client i

m_{ij} Estimated routing distance in mile from location j to client i

m Maximum delivery distance in mile from iDCs to clients

t Maximum number of iDC

Decision variables:

x_{ijk} 1 if the good k of client i is assigned to location j, 0 otherwise

y_{jk} 1 if location j with good k is chosen, 0 otherwise

z_j 1 if location j is chosen, 0 otherwise

$$\min \quad \sum_{j\in J}\sum_{k\in K}\left(s_{jk}\sum_{i\in I}x_{ijk}\right)+\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}c_{ijk}d_{ik}x_{ijk} \tag{1}$$

$$\text{s.t.} \quad \sum_{j\in J}x_{ijk}\geq 1,\ \forall i\in I,\ \forall k\in K \tag{2}$$

$$x_{ijk}\leq y_{jk},\ \forall i\in I,\ \forall j\in J,\ \forall k\in K \tag{3}$$

$$y_{jk}\leq z_j,\ \forall j\in J,\ \forall k\in K \tag{4}$$

$$\sum_{j\in J_1}z_j\leq t \tag{5}$$

$$y_{jk}=z_j=1,\ \forall j\in J_2,\ \forall k\in K \tag{6}$$

$$(m-m_{ij})\,x_{ijk}\geq 0,\ \forall i\in I,\ j\in J_1 \tag{7}$$

$$x_{ijk},\ y_{jk},\ z_j\in\{0,1\} \tag{8}$$

Equation (1) is the objective function of our cost-minimization model, while equations (2-8) are the respective constraints. Each of these equations are explained below:

1. The sum of delivery cost of leg 1 and 2;
2. For any product demand of any client, it must be fulfilled by at least one distribution center;
3. Linking constraint, if a certain distribution center is assigned to fulfill demand, it must be opened;
4. Linking constraint, if a certain distribution center is assigned to fulfill demand, it must be opened;
5. The number of total opened iDCs cannot exceed a management requirement;
6. For any existing distribution center, it must be open;
7. We cannot assign stores to iDCs if distance between them exceeds a management requirement;
8. In our model, the decision models are binary.

3.3.2 Profit Maximization Model

Similarly with the cost minimization model, we proposed the profit maximization model with several revenue scenarios, each of which is associated with a certain probability.

Indices:

I Aggregated client (zip code) index

J_1 iDC Candidate facility location index

J_2 Existing facility location index

$J_1 \cup J_2$ Candidate facility location (iDC) index

K SKU storage temperature category index, namely dry, cooler, and freezer

S Possible profit scenarios of value-added service

Parameters:

a_{jk} Unit variable cost of delivering product k from supplier to location j

b_{ijk} Unit variable cost of delivering product k from location j to client i

c_{iks} Yearly profit of value-creation service at client i of product k under scenario s

d_{ik} Demand of product k of client i

p_s Scenario probability of s

m_{ij} Estimated routing distance in mile from location j to client i

m Maximum delivery distance in mile from iDCs to clients

t Maximum number of iDC

Decision variables:

x_{ijks} 1 if the good k of client i is assigned to location j in scenario s, 0 otherwise

y_{jk} 1 if location j with good k is chosen, 0 otherwise

z_j 1 if location j is chosen, 0 otherwise

$$\min \quad \sum_{s\in S}\sum_{k\in K} p_s \left(\sum_{j\in J} \left(a_{jk} \sum_{i\in I} x_{ijks} \right) + \sum_{i\in I}\sum_{j\in J} b_{ijk} d_{ik} x_{ijks} - \sum_{i\in I}\sum_{j\in J_1} c_{iks} x_{ijks} \right) \tag{9}$$

$$\text{s.t.} \quad \sum_{j\in J} x_{ijks} = 1,\ \forall i \in I,\ \forall k \in K,\ \forall s \in S \tag{10}$$

$$x_{ijks} \le y_{jk},\ \forall i \in I,\ \forall j \in J,\ \forall k \in K,\ \forall s \in S \tag{11}$$

$$y_{jk} \le z_j,\ \forall j \in J,\ \forall k \in K \tag{12}$$

$$\sum_{j\in J_1} z_j \le t \tag{13}$$

$$y_{jk} = z_j = 1,\ \forall j \in J_2,\ \forall k \in K \tag{14}$$

$$(m - m_{ij})\, x_{ijks} \ge 0,\ \forall i \in I,\ j \in J_1,\ \forall s \in S \tag{15}$$

$$x_{ijks},\ y_{jk},\ z_j \in \{0,1\} \tag{16}$$

Equation (9) is the objective function of our cost-minimization model, while equations (10-16) are the respective constraints. Each of these equations are explained below:

9. Minimization objective: The sum of delivery cost of leg 1 and 2, minus the sum of profits of value-creation services; note that the model captures potential profits from zip codes that served by iDCs only;
10. For any product demand of any client, it must be fulfilled by at least one distribution center;
11. Linking constraint: if a certain distribution center is assigned to fulfill demand, it must be opened;
12. Linking constraint: if a certain distribution center is assigned to fulfill demand, it must be opened;
13. The number of total opened iDCs cannot exceed a management requirement;
14. For any existing distribution center, it must be open;
15. We cannot assign stores to iDCs if distance between them exceeds a management requirement;
16. In our model, the decision models are binary.

To summarize, our proposed methodologies answer the initial business questions, and are worth exploring further. We present our computational results in the next chapter.

Table 10. *Review of Business Questions and Proposed Solutions*

Question	*Solution*
What quantitative methods to use?	A hybrid model with clustering and mixed integer linear programming
How to identify urban clusters?	Revised clustering algorithm with emphasis on demand weights
How to deal with incomplete transportation data?	An original missing cost data inference framework (KNN and linear fit)
How to choose iDC location from many candidates?	Mixed integer linear programming

4. RESULTS AND ANALYSIS

In this chapter, which follows the same structure as Chapter 3 Data and Methodology, we present our findings and analysis.

4.1 Demand Clustering

As we revised the three different clustering algorithms, their implemented results were different as well. We first present these results, then compare them, and give our recommendation.

4.1.1 Revised KMeans Algorithm

We developed and performed the revised KMeans algorithm and Elbow Curve Method. The elbow curve indicates that both 6 and 15 are applicable for k. However, 15 is more practical considering the geographic features in the United States, as the number can capture multiple metropolitan areas. According to the Pareto cumulative curve, we set the threshold to 50%, and the corresponding clustering result is also illustrated in Figure 11. As a result, the range of the heatmap of demand-dense areas is much smaller, which would be the expected “urban” areas.

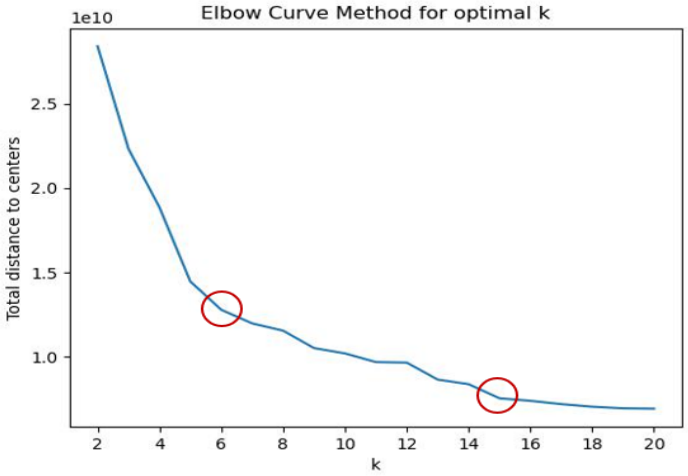

Fig. 9, *Elbow Curve with Circled k Values*

Fig. 10, *Clustering Results of Revised KMeans*

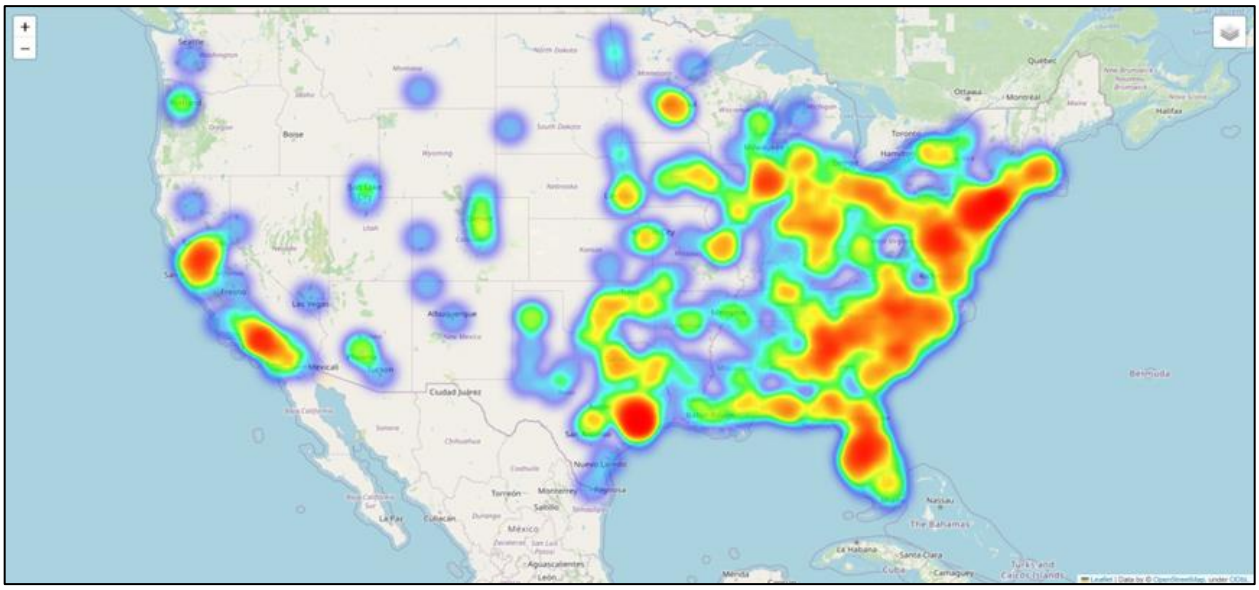

Fig. 11, *Filtered Demand Heat Map of Revised KMeans*

4.1.2 Revised Mean Shift Algorithm

We developed and performed the revised Mean Shift algorithm. The clustered result and filtered heatmap are presented in Figures 12 and 13. The cluster result is different from revised KMeans in detail, but it selects a large portion of common areas as KMeans.

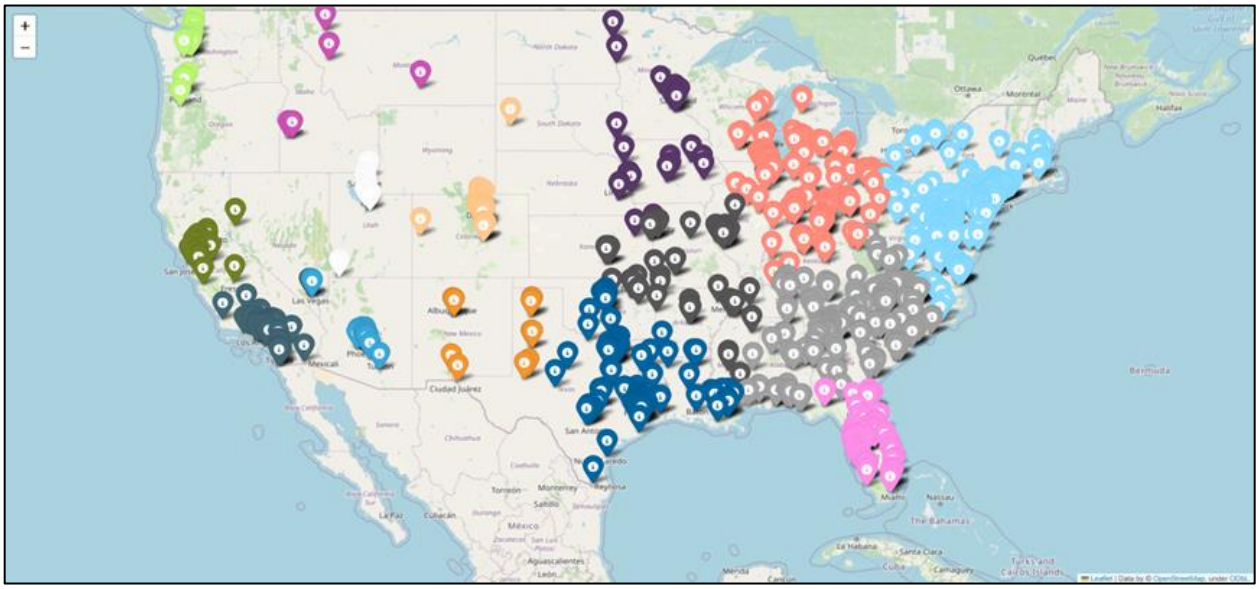

Fig. 12, *Clustering Results of Revised Mean Shift*

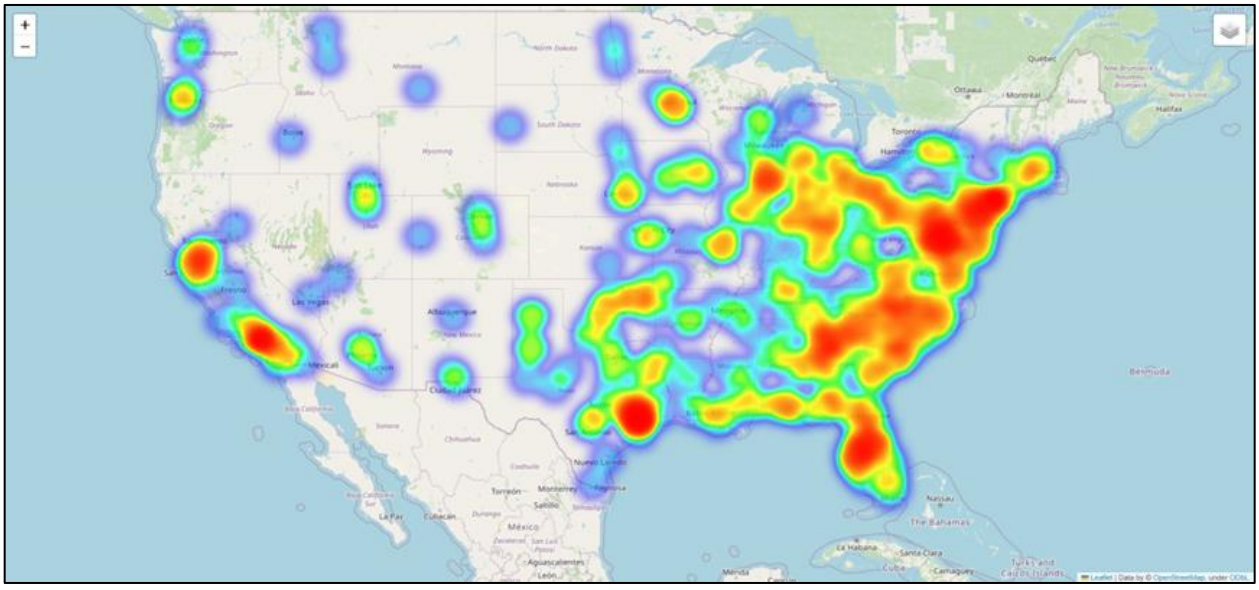

Fig. 13, *Filtered Demand Heat Map of Revised Mean Shift*

4.1.3 Revised DBSCAN Algorithm

We developed and performed the revised DBSCAN algorithm. The clustering and filtering results are shown in the figures below. It is obvious the method can isolate expected urban areas; however, the clustering performance is not as good as with the previous methods: the zip codes in middle and east regions are categorized into only one cluster. This outcome reveals one drawback of DBSCAN: it relies on fixed but not adaptive parameters (bandwidth and minimal point threshold), so it is difficult to identify clusters with various densities. In the middle and east regions, there should have been more than one cluster, but they are not well labeled because they are closer to each other compared to clusters in the Midwest region.

Fig. 14, *Clustering Results of Revised DBSCAN*

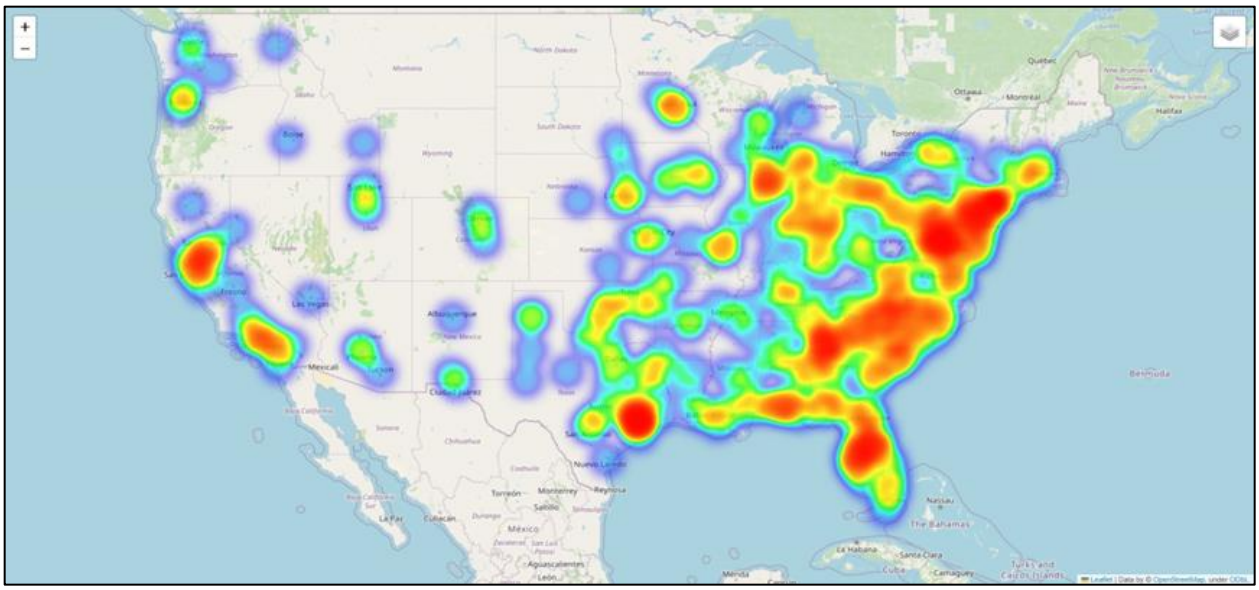

Fig. 15, *Filtered Demand Heat Map of Revised DBSCAN*

4.1.4 Comparison and Summary

We further compared the three algorithms about their advantages and drawbacks. In addition, we analyzed common zip codes and their percentages under different algorithms.

Table 11. *Comparison of the Three Clustering Algorithms*

Algorithms	*Pros*	*Cons*	*Common Zip Codes and Percentages*
Revised KMeans	- Easy to implement; - Fast and scalable to large dataset; - Interpretable with decent performance.	- Need to specify number of clusters; - Sensitive to outliers; Sensitive to initially selected centers.	693, 84.2%
Revised Mean Shift	- No need to specify number of clusters; - Easy to tune with only bandwidth parameters.	- Sensitive to bandwidth; - Computational complex.	693, 84.2%
Revised DBSCAN	- No need to specify number of clusters; - Robust to outliers; Applicable to non-convex shapes.	- Sensitive to parameters (bandwidth and minimal points); - Difficult to identify clusters with various densities.	693, 86.0%

As shown in Table 11, Revised KMeans and Mean Shift perform similarly, while DBSCAN gives fewer urban zip codes. We built further upon the result of revised KMeans, due to its good interpretability and categorization in Armada's dataset.

4.2 Cost Minimization Model

By running the cost-based model, we obtained the initial result of iDC location selection. In the map in Figure 16, we use a gray icon to denote selected iDC locations. Currently, the maximum delivery distance is 100 miles, while the maximum number of iDCs is 100.

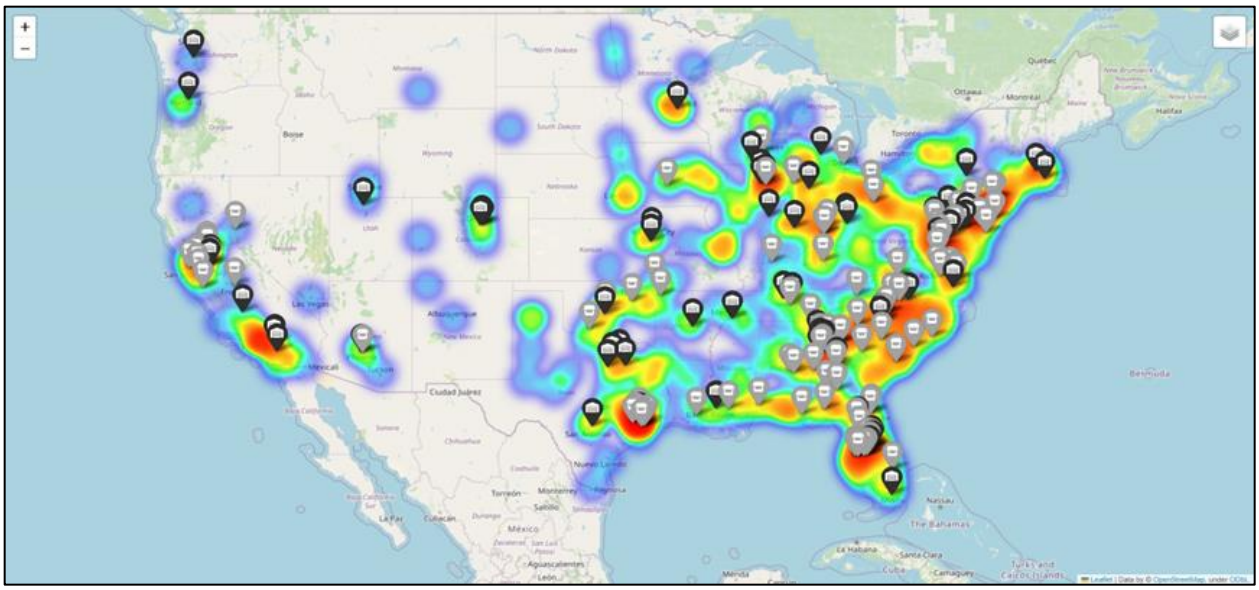

Fig. 16, *iDC Locations of Cost Minimization Model*

By summarizing the number of selected iDCs in different states, we list the states having the most iDCs in the cost minimization model in Table 12. Appendix A provides detailed coverage results between facilities (current warehouses and selected iDCs) and clients.

Table 12. *Top 5 iDC Allocation States - Cost Model*

State Code	*No. of iDCs*
FL	15
GA	12
VA	11
CA	10
TX	7

4.3 Profit Maximization Model

By running the profit-based model, we have the result of iDC location selection. In the map in Figure 17, we use a gray icon to denote selected iDC locations. Currently, the maximum delivery distance is 100 miles, while the maximum number of iDCs is 100.

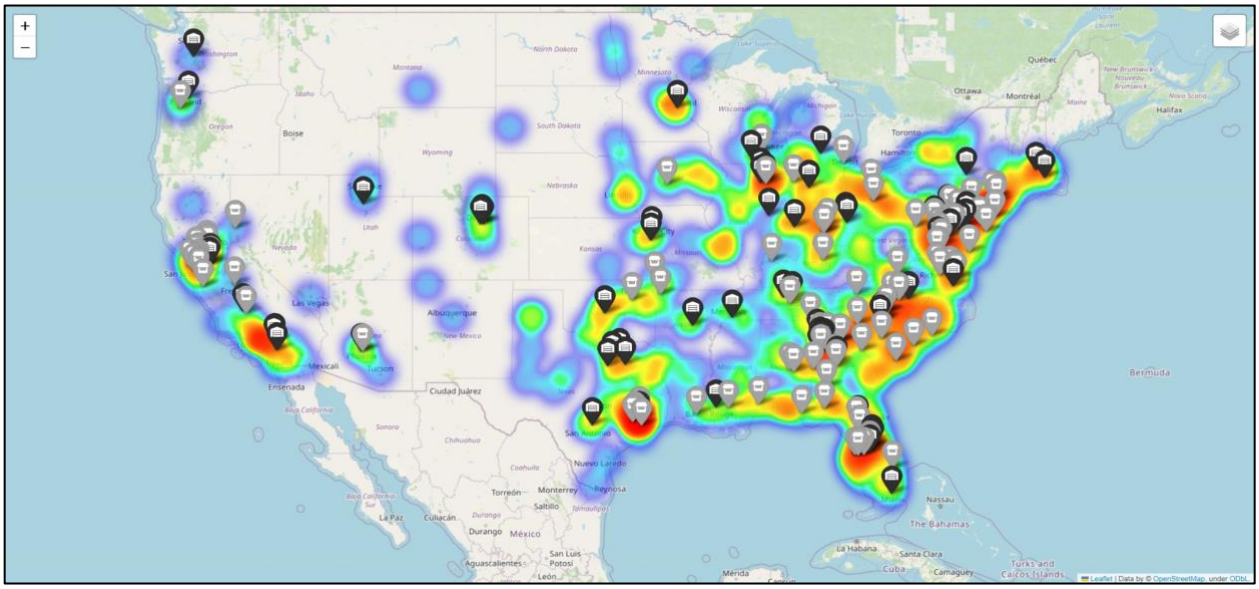

Fig. 17, *iDC Locations of Profit Maximization Model*

Similarly, we list the states having the most iDCs in the cost minimization model in Table 13. Appendix 8.2.2 provides detailed coverage results between facilities (current warehouses and selected iDCs) and clients.

Table 13. *Top 5 iDC Allocation States - Profit Model*

State Code	*No. of iDCs*
CA	15
VA	11
FL	10
GA	10
MD	5

4.4 Comparison Between Cost Minimization and Profit Maximization Models

According to the profitability data of regions, Northeast, Southwest, and Northwest are the projected most profitable regions. Through comparison between results of cost- and profit-based models, there are moves of iDC locations from southeast (e.g., Florida) to the northeast and southwest part of the country. They are justified by the iDC profitability data, which implies northeast and southwest regions have more potential of value-creation services.

In Table 14, the iDC count in the Southwest region decreases from 59 to 55 in the profit-based model, and the iDC count in the Northeast increases accordingly. However, we do not observe a significant change of the numbers. Due to the lack of traditional distribution centers in the Southeast region currently, the profits gained by adding more iDCs in the Northeast are not as profound as the cost saved by putting them in the Southeast.

Table 14. *iDC Count Comparison Between Both Models by Regions*

Region	*iDC numbers of cost minimization model*	*iDC numbers of profit maximization model*
Midwest	12	12
Northeast	7	10
Northwest	0	1
Southeast	59	55
Southwest	22	22

5. DISCUSSION

The results of our model are compelling. We can now understand that when applying revenue projections to our model, there are shifts in facility allocation to favor revenue capture. In Section 5.1 we will reflect on what our results in Chapter 4 demonstrate and any implications to Armada. We will then review some of the limitations of our model in its current state (Section 5.2). Section 5.3 discusses our recommendations to our sponsor company based on the model built, and Section 5.4 discusses our suggested approach and the next steps to be completed for the progression of value-creation offerings and making a definitive decision on iDC placement.

5.1 Reflection

The cost-based model gives us an output reflective of cost to deliver by the defined regions, while the profit-based model gives us an alternative output reflecting high revenue-generating regions. Specifically, the cost-based model suggests locating iDCs in the South (both East and West) and the Midwest regions of the United States. It does not place any iDCs in the Northwest, suggesting the cost to deliver services daily is too high. Alternatively, when running our model with the profit-maximization objective, we see a shift away from the Southeast, allocating those iDCs now to the Northeast, apart from one iDC being allocated to the Northwest. These results inform us that adding iDCs in the northern part of the United States increases profits at a higher rate than adding iDCs to the southern half.

A major factor impacting these results is that there are a lower number of traditional distribution centers in Armada's current network, located in the Southeast region. For this reason, the *profits* realized by adding one iDC to the Northeast region are not as significant as the *costs* saved by adding one iDC to the Southeast region. That said, it may be compelling for Armada to explore how opening additional traditional distribution centers in the Southeast impact the results of the

iDC location allocation model. Further stated, will an increase in traditional distribution centers in the Southeast suggest a more proportionate allocation of iDC location solutions.

5.2 Limitations

Reviewing our approach to the business problem at hand, there are some limitations that can be addressed:

1. Capacities of iDCs are not captured in our facility location models. The decision variables are locations and allocations of multiple product categories. Capacities of each selected iDC can be calculated by dividing the demand of iDC by demand-area ratio. The demand-area ratio varies between different regions and can be estimated by industry rule of thumb.
2. Fixed costs of iDCs are not incorporated into our facility location models. Because of the time constraint, relevant data are not provided or collected. It can be fixed by estimating the real estate rentals and adding the total fixed costs in the objective of the models. In addition, we can couple fixed costs and iDC capacities together to give recommendations on capacity choices of iDCs.
3. Service sensitive demand is not explored in the capstone. We primarily focus on deterministic demand. However, both logistics and value-creation services in the food supply chain are sensitive to where facilities are located. It brings uncertainty in the results of our models.

5.3 Recommendations

We recommend that Armada utilize the developed model to identify the top five to ten urban clusters to perform further analysis of the market demand, prior to deciding where to build each iDC. The current model provides 100 feasible solutions, which can be adjusted to a smaller selection, should Armada want to review fewer feasible solutions. We suggest that after internal deliberations, Armada defines the number of feasible solutions to evaluate and adjust the model

parameters to reflect that decision. The output of the current model suggests that the locations with highest profitability and lowest costs are the feasible solutions, the Northeast and Southwest regions. Knowing this, it would make sense for Armada to narrow down on a select number of feasible locations to do a market analysis of the demand for each value-creating service, in each area. There will be key next steps for Armada once the model has produced these possible iDC locations for them, namely identifying which value-creation services will best serve the selected areas. Section 5.3 details the future research needed to determine the final recommendation of value-creating services at each iDC.

5.4 Future Research

There are a few areas of future research to the iDC project that we suggest Armada, or an additional academic institution, complete to assist in finalizing their decision of where to place iDCs. Our model provides 100 optimized locations that will minimize costs and maximize profits. Armada does not want to build 100 iDCs, at least not in the immediate term. To decide the most viable solution(s) to build upon, we suggest Armada adjust the model parameters to the number of solutions they would like to reasonably evaluate. Once results are run, market research will be required to validate which value-creation services are most sought after in each of these urban regions. It will be important for future researchers to understand that while the northeast region may have a strong demand for food preparation services, leading to high profits to offer this solution, that may not be the right value-creating service to offer in the southwest. Alternatively, the southwest may strongly value reverse logistics to reduce food waste, and therefore the most feasible service to offer in this region is reverse materials handling. We suggest that for each identified location for an iDC, the contributing researchers utilize a market analysis tool like SWOT or Porter's Five Forces to understand the feasibility of implementation in each urban area. Furthermore, performing market research can help guide Armada's revenue projections, which are easily updated in our model.

6. CONCLUSION

We began this research capstone with an understanding of design of supply chain networks and the key strategic decisions they help drive in organizations. Our company sponsor, Armada, approached us with a scalable research problem of redesigning their current service network. The two components encompassed in the problem were: (1) adding new distribution centers ("iDCs") located near urban areas, and (2) deploying services that create value for Armada's client base. We sought to answer the following research questions:

- RQ1: How can Armada quantitatively identify urban clusters in their current distribution network that will be better served by an iDC than a traditional DC?
- RQ2: What kind of quantitative method should Armada use to choose iDC locations and which are the resulting locations?
- RQ3: What relationship exists between cost and distance between iDC and its location?
- RQ4: How can Armada incorporate anticipated value-creation services into the quantitative method?

To address these research questions, we adopted a hybrid methodological approach. We first analyzed demand distribution at a zip code level to derive our demand clustering approach. We then formulated the primary facility location models; first with an objective of cost-minimization. Then, we built upon that model with the objective of maximizing total profit (revenue less costs). As a result, our hybrid approach addresses each of our key research questions and provides Armada with specified urban areas across the country and prioritizes iDC allocation based on cost reduction and profit maximization.

Contributing to the supply chain and network design industry, our primary finding relates to profit-driven network design models. Our approach incorporates both costs and revenues into a set of

models. The cost-minimization objective will locate hubs near demand dense regions, while the profit-maximizing model will use revenues to dictate location allocation. This is a key finding for overarching business strategy because it shifts the focus from driving decisions based on cost reduction, to providing freedoms to prioritize revenue generating factors in business decision making.

REFERENCES

Armada. "Mission and Vision." Accessed October 8, 2022, https://www.armada.net/about-us/mission/.

Current, J., Min, H., & Schilling, D. (1990). Multiobjective analysis of facility location decisions. European journal of operational research, 49(3), 295-307.

Karatas, M., & Dasci, A. (2020). A two-level facility location and sizing problem for maximal coverage. Computers & Industrial Engineering, 139, 106204.

Klibi, W., Martel, A., & Guitouni, A. (2010). The design of robust value-creating supply chain networks: a critical review. European Journal of Operational Research, 203(2), 283-293.

Kouvelis, P., & Rosenblatt, M. J. (2002). A mathematical programming model for global supply chain management: Conceptual approach and managerial insights. In Supply chain management: Models, applications, and research directions (pp. 245-277). Springer, Boston, MA.

Liao, K., & Guo, D. (2008). A clustering-based approach to the capacitated facility location problem 1. Transactions in GIS, 12(3), 323-339.

Mak, H. Y., & Shen, Z. J. M. (2016). Integrated modeling for location analysis. Foundations and Trends® in Technology, Information and Operations Management, 9(1–2), 1-152.

Melkote, S., & Daskin, M. S. (2001). Capacitated facility location/network design problems. European journal of operational research, 129(3), 481-495.

Melo, M. T., Nickel, S., & Saldanha-Da-Gama, F. (2009). Facility location and supply chain management–A review. European journal of operational research, 196(2), 401-412.

Mesa-Arango, R., & Ukkusuri, S. V. (2015). Demand clustering in freight logistics networks. Transportation Research Part E: Logistics and Transportation Review, 81, 36–51. https://doi.org/10.1016/j.tre.2015.06.002

Miranda, P. A., González-Ramírez, R. G., & Smith, N. R. (2011). Districting and customer clustering within supply chain planning: a review of modeling and solution approaches. IntechOpen.

Mukundan, S., & Daskin, M. S. (1991). Joint location/sizing maximum profit covering models. INFOR: Information Systems and Operational Research, 29(2), 139-152.

Ross, G. T., & Soland, R. M. (1980). A multicriteria approach to the location of public facilities. European journal of operational research, 4(5), 307-321.

Sankaran, J. K., & Raghavan, N. S. (1997). Locating and sizing plants for bottling propane in south India. Interfaces, 27(6), 1-15.

Sheffi, Y. (2013). Logistics-Intensive Clusters: Global Competitiveness and Regional Growth. In J. H. Bookbinder (Ed.), Handbook of Global Logistics (Vol. 181, pp. 463–500). Springer New York. https://doi.org/10.1007/978-1-4419-6132-7_19.

Soland, R. M. (1983). The design of multiactivity multifacility systems. European Journal of Operational Research, 12(1), 95-104.

Sridharan, R. (1995). The capacitated plant location problem. European Journal of Operational Research, 87(2), 203-213.

Wildes, Joe. Armada iDC Project Discussion. Teams Meeting, October 11, 2022.

Wu, L. Y., Zhang, X. S., & Zhang, J. L. (2006). Capacitated facility location problem with general setup cost. Computers & Operations Research, 33(5), 1226-1241.

APPENDICES

Appendix A: Data Preprocess

A1. Data combination

All three sets provided by Armada are stored by clients in different sheets of Excel files, so we first concatenate the specific data of each client into a holistic data set for every set of data (Store Sales of client 3, 8, 9, 17 into a total Store Sales). The process facilitates further analysis as we are designing networks for all clients instead of any one of them.

A2. Missing values in iDC-Warehouse Sales

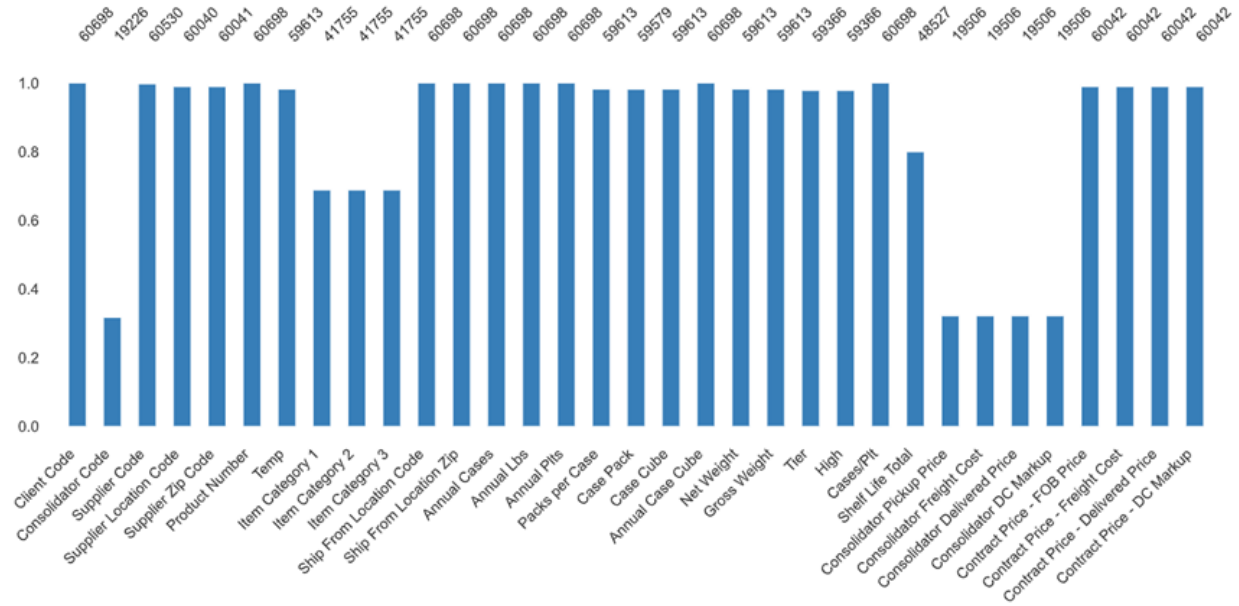

Fig. A2-1, *Missing Values in iDC-Warehouse Sales*

We focus on columns that will be used in later analysis, while leaving alone columns that are auxiliary even though they are heavily missed.

Fixed by:

1. Fill “Temp” with “DRY”, “Consolidator Pickup Price” with 0, “Consolidator Freight Cost” with 0, “Consolidator Delivered Price” with 0, “Consolidator DC Markup” with 0, “Contract Price – FOB Price” with 0, “Contract Price – Freight Cost” with 0, “Contract Price – Delivered Price” with 0, “Contract Price – DC Markup” with 0.

A3. Missing values in Store Sales

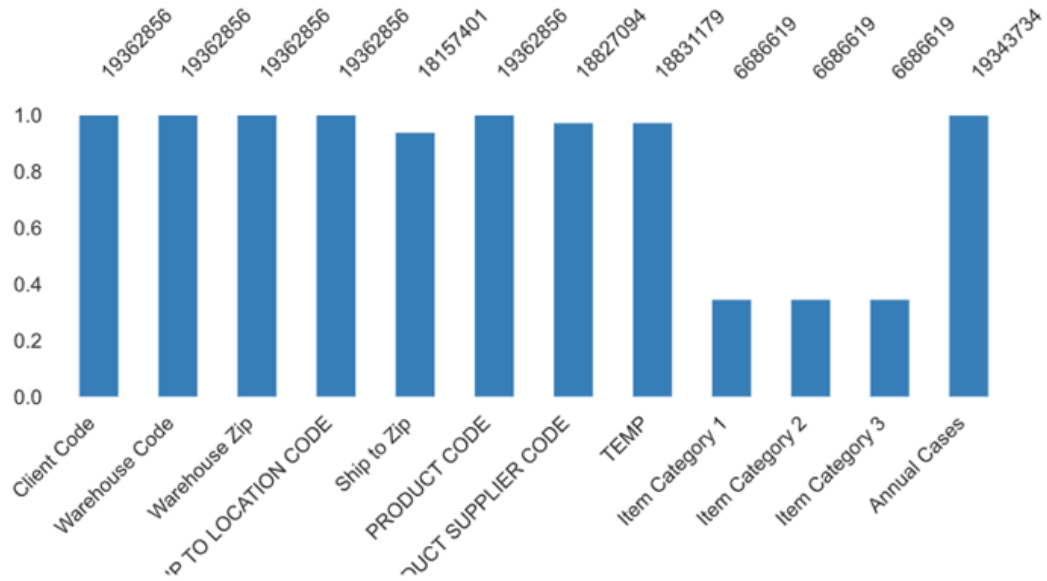

Fig. A3-1, *Missing Values in Store Sales*

Fixed by:

1. Delete rows with missing “Ship to Zip” or “Annual Cases”;
2. Fill “TEMP” with “DRY”.

A4. Missing values in Items and Current Pricing

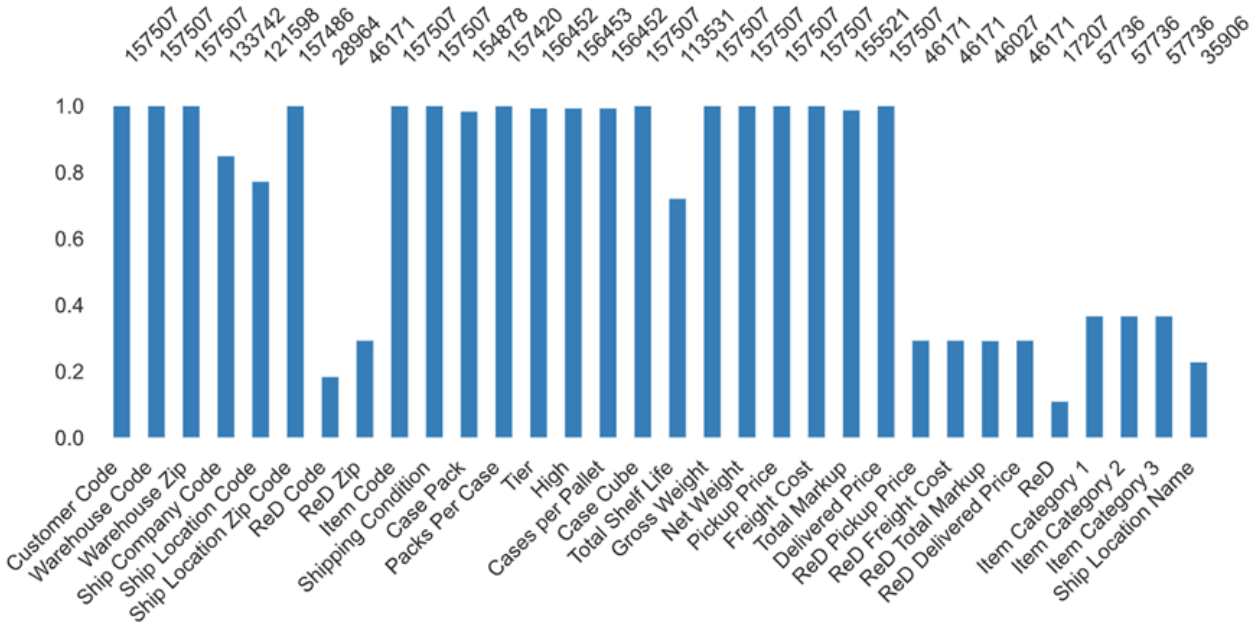

Fig. A4-1, *Missing Values in Items and Current Pricing*

Fix by:

1. Delete rows with missing “Ship Location Zip Code”;
2. Fill “Pickup Price”, “Freight Cost”, “Total Markup”, “Delivered Price”, “ReD Pickup Price”, “ReD Freight Cost”, “ReD Total Markup”, and “ReD Delivered Price” with 0.

A5. Duplications

Fixed by:

1. Delete duplicates in all data files.

A6. Consistency of zip code

Fixed by:

1. Add zero(s) at the beginning of zip codes which have less than 5 digits;
2. Only keep first part of zip codes whose forms have pattern of "*****-****";
3. Delete rows with other irrelevant zip codes.

Appendix B: iDC Location Output

B1. Cost minimization model result
https://docs.google.com/spreadsheets/d/1jEZJk136dGWJe3Bv07C6Kd5lZaDd_86j/edit?usp=sharing&ouid=100135311333344325180&rtpof=true&sd=true

B2. Profit maximization model result
https://docs.google.com/spreadsheets/d/1ubUSFy7wIdb3MZygu7tFM5lRPU8TJEuv/edit?usp=share_link&ouid=100135311333344325180&rtpof=true&sd=true