The Pennsylvania State University

The Graduate School

ANALYZING THE IMPACT OF DISTRIBUTION CENTER LOCATION ON E-COMMERCE NETWORK DESIGN

A Thesis in

Data Analytics

by

Luke Jackson

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2022

The thesis of Luke Jackson was reviewed and approved by the following:

Robin G. Qiu
Professor of Information Science
Thesis Advisor

Raghu S. Sangwan
Professor of Software Engineering

Youakim Badr
Professor of Data Analytics

Colin J. Neill
Professor of Software and Systems Engineering
Head of the Engineering Division

ABSTRACT

Strong E-commerce demand over the past decade has resulted in market constraints on available land and labor that limit the ideal location selection of new E-commerce facilities. This research investigates the impact that Middle-Mile facility locations have on the network design process. Our research showed that Middle-Mile facility location has an outsized impact on the overall delivery speed and transportation cost of the network. A strong correlation was found between the location of Middle-Mile facilities and the population density of their surrounding Urban Areas. We measured the impact Middle-Mile (Distribution) facility locations have on the upstream First-Mile (Fulfillment) and downstream Last-Mile (Delivery) segment distances. Using these relative distances, we found an optimized design that shifted Middle-Mile facility locations 10 miles closer to their nearest population centers.

To test this design, we built a network simulation model to quantify network impact. We constructed network edges and vertices by calculating the minimum end-to-end distances from 739 E-commerce facilities to 481 Urban Areas. Population density was used to represent customer demand and weighted with network distance to give an objective measure of performance. However, the results show that the 10-mile shift approach increased distances between First-Mile and Middle-Mile more than it reduced the distances between Middle-Mile and Last-Mile segments. This resulted in a +3.5% increase in total end-to-end network distance which is cost unfavorable. Despite the poor performance of the population density approach, this research found that the simulation model provides real-world value in what-if scenario testing for network designers and decision makers. Using the simulation model, network designers can quantify the impact adding or removing facilities has on the overall network. We provide a blueprint for future work to further optimize network design using the simulation model, methodologies, and tooling developed in this research.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

FC	Fulfillment Center
DC	Distribution Center
DS	Delivery Station
UA	US Census Bureau Urban Area

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am beyond grateful for the guidance of my committee, Dr. Robin Qiu, Dr. Sangwan, and Dr. Badr, who contributed to my research questions and structure of my experiments.

I also have a profound appreciation for my wife, family, and friends who supported me throughout this process. Serving as a champion, a soundboard, and a shepherd of my wellbeing. I simply cannot thank you all enough for the unwavering support and encouragement that you've provided me. I will remember fondly how each of you took the time to help, however you could, and despite how difficult it was at times to relate to my experience.

All of my accomplishments mean nothing without each of you.

Thank you.

Author
Luke Jackson

Chapter 1

Introduction

E-commerce has changed how consumers buy and receive their goods. Specifically, the internet has made it possible to browse and buy products via multiple channels (e.g., desktop, mobile, and voice). According to a recent study, consumer sentiments towards e-commerce have reached a point where what used to be a marginal complementary activity is now in direct competition with conventional retail [1]. The research found that changes in consumer sentiment and the retail sector footprint have taken place concurrently. The sustained growth of e-commerce is shifting the broader retail footprint from commercially-accessible locations, towards transportation-accessible locations. Analysts estimate e-commerce transactions will total $5.5 trillion USD by the end of 2022 and will account for 20% of worldwide retail sales [13]. This continues a growth trend with sales up 30% ($4.3 trillion USD) since 2020 and 76% ($1.3 trillion USD) since 2014 [13].

Strong sustained demand, high commercial real estate availability, and surplus labor supply facilitated aggressive e-commerce expansion over the past decade [3]. These sustained growth cycles created a “flywheel effect” that also influenced demand momentum. Network expansion into new geographies reduced transportation costs and increased capacity that generated savings which were passed on to consumers through lower prices. Lower prices and high accessibility in turn attracted more demand [18].

Analysts now expect year-over-year e-commerce growth to continue averaging 10% through 2025 [13]. This is expected to create challenges in constrained markets [20]. Additionally, research has shown that having too many facilities within the same commutable distance radius can exhaust the local labor supply [2]. This is a big concern as labor is an essential

input for both expansion and sustaining existing operations. Similarly, commercial real estate vacancy rates are at record lows. The national average is less than 4% (96% occupied) as of Q1 2022 [20]. Thus, reduced real estate inventory will further increase competition and drive higher price per sq. ft. cost. Notably, these high recurring costs will further constrain where network expansion is viable. These constraints will motivate E-commerce companies to make more investments into location selection science and the processes that design their end-to-end networks.

E-Commerce Networks 101

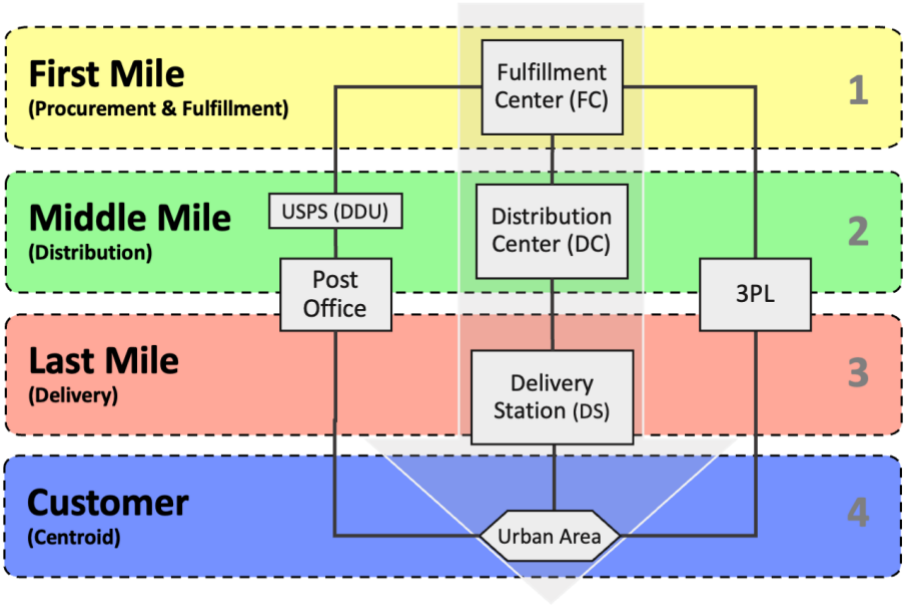


Figure 1-1: E-commerce Network Ecosystem

All supply chain networks in the E-commerce sector have a hierarchical structure. Figure 1-1 illustrates the hierarchy and sequence of facilities that goods pass through as they traverse the network. E-Commerce networks are a special form of supply chain network. The First-Mile of the

E-commerce supply chain includes the processes of stowing, retrieving, packaging, and shipping ordered goods. These processes take place in a Fulfillment Center (FC) [1]. Packaged goods are passed to the Middle-Mile" where they are consolidated by destination in a Distribution Center (DC) [2]. The last segment before the customer is the Last-Mile". Delivery Stations (DS) [3] receive sorted packages where they are assigned a route for customer delivery. For this research we've constrained the delivery locations to US Urban Areas (UA) [4]. Urbanized Areas (UAs) are legal designations from the US Census Bureau that delineate areas having a population of at least 50,000 persons. According to the 2020 cartographic files there are 481 designated Urban Areas within the contiguous 48 states [24]. This research will focus specifically on the Distribution Center Network hereon referred to as the Middle-Mile. It analyzed the specific distances for each of the E-commerce network segments illustrated in Figure 1-1. Together these segments represent the total end-to-end distance that goods travel in their journey from point-of-click to customer delivery.

To design a successful E-commerce network, it is essential to have a basic understanding of the underlying topologies that facilitate the fulfillment, distribution, and delivery of goods from point of click to package delivery. To keep the analysis focused on the simplest solution possible we've made some specific scope exclusions. First, any costs relative to constructing and operating e-commerce facilities are not considered in our model. These are well documented sectors of research with only tangential relationship to the location selection space. Additionally, 3P logistics providers are excluded. Research suggests that most major E-commerce companies are prioritizing the vertical integration of their delivery logistics, therefore the assumption is made that 1P will eventually handle 99% of end-to-end customer packages. The practice of consolidating customer demand into generalized geographical areas is a common industry practice that we will discuss in more detail in Chapter 3. Higher population areas will order more products than lower population areas. This constant allows us to infer cost assumptions on each

routes according to the population density of its Urban Area. Urban Area containers aggregate customer demand into a generalized location. The generalization of customer demand is possible due to our focus on Middle-Mile optimization. Using Urban Areas to consolidate distance measurements scope constrains the values to an appropriate resolution for leadership consumption. Urban Area level insights are sufficient to inform the initial conversation and direction. If needed, address or block level analysis can be completed to answer any lower-level impact questions. Figure 1-2 provides a national overview of the 739 E-commerce facilities and 481 Urban Areas that comprise the network used in this research.

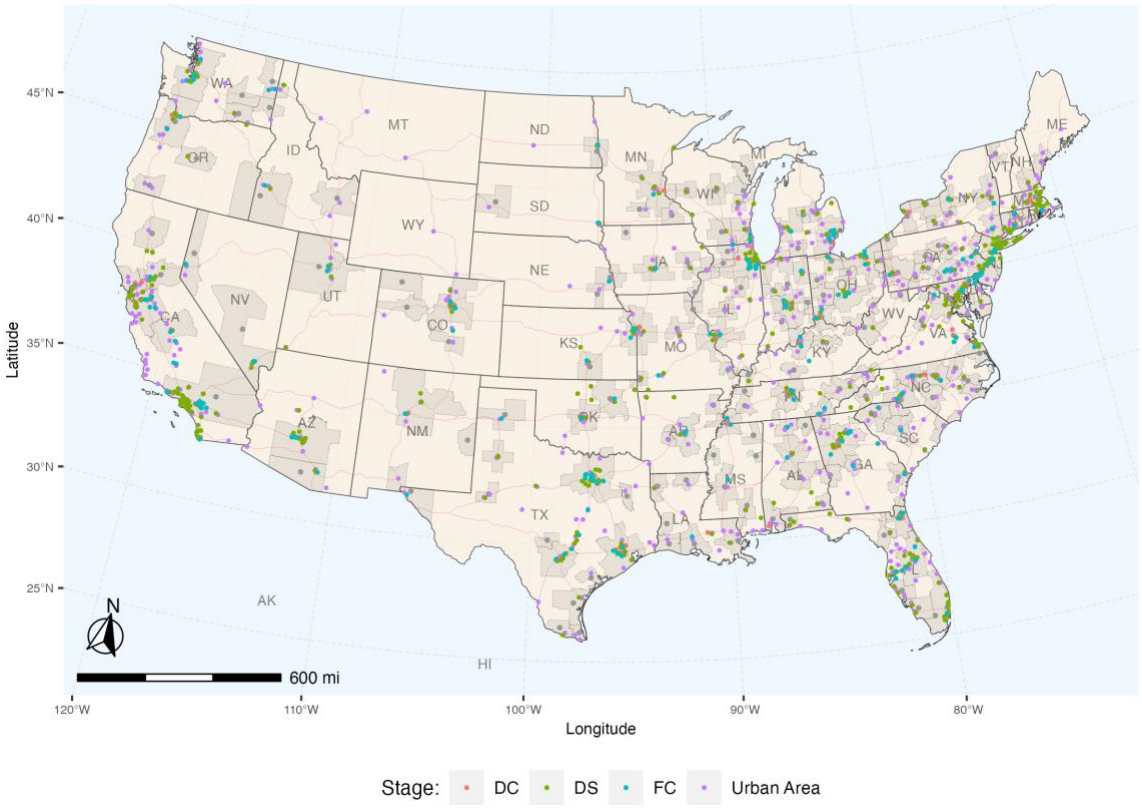


Figure 1-2: Consolidated National Map of Urban Areas and Facilities

Research Question

The principal objective of this research is to understand the impact that Middle-Mile facility locations have on the network design process and its operations. To quantify this answer, we've prepared three (3) research questions.

Research Outline

This research outline defines the steps taken during the research process to create the outcomes discussed in our results section. The questions are incremental with Question 2 building on the work in Question 1 and Question 3 connecting all of the parts into the complete answer.

Question 1 (Network Distance by Segment)

What are the upstream (First-Mile) and downstream (Last-Mile) distance impacts of the existing Middle-Mile facility locations?

To answer this question, we implemented the cleaned datasets to generate distance matrices using geocoded addresses. The distance matrices store distances pairs for all possible combinations of vertices that are then used to quantify specific distances by segment for known edges. Using a structural understanding of E-commerce Supply Chain and Nearest Neighbor analysis we construct a minimum distance network. We define a weighted metric to objectively compare distance across Urban Areas to ensure that higher populated areas are weighted higher than lower populated areas. Finally, we summarized the upstream (First-Mile) and downstream (Last-Mile) distance impacts by division.

Question 2 (Design Network around Max Population Density)

How do the upstream (First-Mile) and downstream (Last-Mile) distances change as the centroids of the Middle-Mile facilities move closer to the max population center of their local capacity radius (cluster community)?

To answer this question, we built upon the Nearest Neighbor analysis in Question 1 and aggregated nearby Urban Areas using clustering analysis. The leading clustering algorithms were tested and evaluated against a standardized cost function to determine the best fit model for our dataset. The resulting clusters were used to calculate a weighted mean centroid based on the population density of all Urban Areas in the cluster. Using this weighted centroid an initial bearing was calculated for each Distribution Center within the cluster to project a new optimized location.

Question 3 (Quantify Network Impact)

What is the impact of Middle-Mile facility locations on delivery speed and transportation cost (locally, regionally, nationally)?

To answer this final question, we integrate the results from the prior to questions to determine the holistic distance impacts for the network. Using these distances, we calculate Delivery Speed and Transportation Cost locally, regionally, and nationally. This informs the final design recommendations for the network and demonstrates the business value of the model.

Summary

In this chapter we presented the objective and scope of this research. We established a foundational knowledge of E-commerce network design and define three specific questions to guide the research. This thesis is organized into 6 chapters. In the next chapter we will review the extent of current literature, define cutting-edge, and discuss the identified research gap.

Chapter 2

Literature Review

The review of current literature identified two key constraints, the availability of 1) labor and 2) commercial land. These constraints directly limit the location of proposed new facilities. The severity of these constraints is driving a need for analytical models that can evaluate multivariate cost-benefit tradeoffs within complex multi-stage networks. To address these headwinds, geospatial analysis models have been implemented to quickly estimate the impact of facility locations without the overhead of complex Optimization Models [17]. A simplified "as the crow flies" geospatial model can be operated as a front-end field tool to quantify the potential value of a location as early in the real estate search process as possible.

All location selection decisions must consider demand uncertainty, and determine when network expansion into constrained markets makes long-term fiscal sense. The location of each E-commerce facility represents a trade-off between economies of scale, operational requirements, market areas, land use density, lead time, and land cost [1]. Research has shown that up to 60% of a company's supply chain costs are attributed to the transportation of goods between facilities in route for customer delivery [9]. Companies are incentivized to minimize transportation costs and maximize the click-to-deliver speed through increased customer demand. A delivery promise at checkout of less-than 2-days is estimated to increase checkout conversions by 11% [17].

Planning to match this "momentum" of customer demand for the end-to-end network is a difficult and dynamic challenge that increases in complexity with the number of facilities operating in the network. As the network scales to meet customer demand so does their complexity. Facilities in smaller networks can handle all product types, whereas in larger networks they must specialize based on product dimensions, conveyability, and hazard

characteristics (e.g. aerosols and batteries) [23]. Despite these complexities, the location of each facility within the network remains the most important decision in the network design process.

Current state-of-the-art models for network design show a shift in thinking from prior segment-based optimization to full end-to-end [12]. This research also found that traditional brick-and-mortar retail location selection models can be adapted to optimize E-commerce networks [19]. A notable example [15], compared the location selection strategies of Walmart and Kmart retail stores in the Greater Cincinnati Area. Their model more accurately estimated the potential revenue for each store by addressing the impact of store location in respect to all other destinations on consumer selection, which had not been considered in prior research. Their research found that Walmart unlike Kmart made a decisive decision to retrofit their older lower performing Discount stores into Super-centers and discontinue the Discount template in favor of the Super-center. This type of strategy adjustment shows an understanding of their core business model and that they are willing to make costly changes to their retail network in order to best address their customer's needs.

Another key development showed that lightweight geospatial analysis routines paired with constrained clustering models can generate fast location selection insights with no statistically significant impact to accuracy [14]. Their innovation addresses shortcomings in traditional clustering techniques that generate sub-optimal clusters, with large spatial distances between the cluster centroid and cluster points making them unusable for computing drive times. Their constrained clustering approach realized a mean computation time reduction of ~80% when run on 220,334 census block groups. Also notable in spatial clustering research is the application of density aware clustering [28]. In this research the use of a "convex hull" as a cost function for evaluating clustering performance is proposed. This research found a reduction in overall cluster diameter when compared to single and complete traditional linkage techniques.

Traditional linkage techniques measure the distances between nodes of different clusters to determine the relative efficiency of the algorithm. In the case of the convex hull the edge vertices are measured relative to the centroid of the hull. Using this approach, the minimum representation of the cluster (outer edges) is retained and the respective mean measurements of all clusters are an accurate metric of fit to the dataset. The concept of convex hull is analogous to placing a rubber band around a set of nails in a board. First, the rubber band is expanded to fit around the nails and then released to become taut around the perimeter of the outer most nails. The contracted state of the rubber band is the equivalent of the convex hull. This is also referred to as the minimum complex polygon (MCP), which defines a polygon with no angles greater than 180 degrees. Using this method, the minimum diameter of the resulting clusters can be balanced against the total assignment of all available data points to select the most optimal parameters for the model. The centroid of the clustered data can also be weighted during this process to find the population density maxima of the local cluster. Using this weighted average centroid, a target is established for additional spatial network engineering which further minimizes the end-to-end distances within the network. The contributions of these prior works confirm the importance of location selection in traditional retail and E-commerce markets.

The decision of location can be simplified into three (3) distinct impacts [2]. 1) Storage capacity, this is the total cubic feet of space available to house the inventory needed to meet customer demand. 2) Ship capacity, this is the total mechanical and labor capacity available to meet customer demand. 3) Delivery speed, this is the total time from point-of-click to package delivery. All three metrics are measured locally, regionally, and nationally to determine the overall efficiency of the network. Delivery speed is unique as it can serve as both an input and an output in the network design model. For example, defining a national delivery speed goal of 2-days or less would require the minimum distances to all customers be with a 320-mile radius [4]. This research shows facility distance to be highly correlated with both optimal E-commerce

network design and a positive customer experience. As such, it is well suited to serve as the key objective function in any location selection model.

Research Gap

Our investigation into existing literature was not able to find relevant examples of end-to-end models that address a specific link between Middle-Mile infrastructure and population density. In this research gap there is an opportunity to develop a “lightweight” clustering model that approximates the impact of adding or removing facilities during “what-if” scenario modeling. By lightweight model, the intention is to leverage a low-complexity metric for distance estimation from origin to destination. A simplified “as the crow flies” geospatial model is sufficient to quantify the potential value of a location early in the real estate search process and be accessible to field agents.

Lower accuracy is acceptable in the facility location selection space because location recommendations are not one-way doors this early in the process. Any downstream acquisition of land would still be constrained by what real estate and labor are available in the target area. Therefore, the additional computational overheads and model complexities can be foregone, providing a mechanism for low-latency two-way door decisions that add immediate business value. This lightweight approach also offers implementation advantages given that the calculations required to compute simple spatial distances can be executed on personal computers in lightweight languages such as Python, R, and Javascript. Leveraging established open-source frameworks extends the potential use cases all without the overhead of expensive licensing.

Summary

In this chapter, we curated a comprehensive assessment of existing literature and defined the current state-of-the-art approaches to the location selection problem. The key market constraints of labor and commercial land availability create an opportunity to think differently about the holistic E-commerce network. Taking note of Walmart's decisions to pivot their design template strategy and retrofit existing locations, we built on this thinking to develop our population density approach. Leveraging the alternative distance metric defined in the convex hull process we built a model that returns optimal clustering of the target dataset based on network distance. Chapter 4 will discuss in detail how these insights improved the overall performance of the final model. Further we identified a research opportunity to construct a lightweight geospatial model using "as the crow flies" distance and widely available open-source software for implementation.

Chapter 3

Datasets, Techniques, & Tools

All datasets were retrieved or compiled from publicly available sources. There are two primary sources: 1) A major E-commerce organization's list of operating buildings, and 2) the US Census Bureau. Each record in the dataset is an address for either a US Urban Area or 1 of 3 types of E-commerce Facility. All addresses are limited to the Continental US. The compiled list of 739 E-commerce Facilities includes columns for location name, operation type, and physical address. The operation type of each facility is listed in Table 3-1 with a breakdown of location count by zip code group. For zip code grouping the first digit of the zip code, *Zip1* is used to provide a simple geospatial framing of how the facilities are geographically distributed.

Table 3-1: E-commerce Facility Dataset by Zip Code

	Zip Code First Digit (Zip1)										
Op. Type	0	1	2	3	4	5	6	7	8	9	**Total**
DC	13	5	10	17	8	5	8	14	10	17	107
DS	47	44	48	55	36	18	32	47	33	77	437
FC	12	16	14	29	23	7	16	28	18	32	195
Total	72	65	72	101	67	30	56	89	61	126	**739**

Legend: DC = Distribution Center DS = Delivery Station FC = Fulfillment Center

Figure 3-1: US National Map by Zip1 Prefix

The TIGER Geodatabases and American Community Survey (ACS) data were compiled at the regional, divisional, state, and urban area levels. These containers were appended to the Urban Areas dataset to allow grouping of results relative to the respective scope of impact. A breakdown of Urban areas by Division and Region is included in Table 3-2 and visualized in Figure 3-2.

Table 3-2: US Census Bureau Regions, Divisions

Region Name	Division Name	Urban Areas (#)
Midwest	East North Central	74
Midwest	West North Central	33
Northeast	Middle Atlantic	42
Northeast	New England	22
South	East South Central	32
South	South Atlantic	103
South	West South Central	56
West	Mountain	42
West	Pacific	77

TIGER Geodatabases are cartographic shape files storing spatial information from the Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System. These files are designed for use with geographic information systems (GIS) software. The geodatabases contain national coverage (for geographic boundaries or features) or state coverage (boundaries within state). Geospatial analysis was used to convert the physical addresses into latitude and longitude coordinates. Using the resulting geocoded locations and the cartographic shape files we measured land area and "as the crow flies" distance. Joining these results with ACS 2020 survey data, we calculated population estimates, and population density for each of the 481 US Urban Areas.

Figure 3-2: US Census Bureau Regions, Divisions

Table 3-3 lists a breakdown of the complete Urban Areas dataset joined with ACS 2020 survey data for total population estimate, margin of error, and population density. ACS values are added for context on how the US population is geographically distributed across the Urban Areas.

Table 3-3: US Census Bureau Urban Area Dataset by Zip Code

	Zip Code First Digit (Zip1)										
Variable	0	1	2	3	4	5	6	7	8	9	**Total**
Count (#)	55	43	45	45	59	48	42	57	49	38	**481**
Total Pop. (Millions)	24.5	24.2	20.2	8.5	21.0	31.9	42.3	32	15.7	12.2	**232.5**
Margin of Error (Millions)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	**0.9**
Pop. per Mile (Thousands)	2.1	2.5	2.6	1.8	2.4	3.6	3.3	3.0	2.4	2.5	**26.2**

Legend: UA = US Census Bureau Urban Area = Population of 50,000 or more.

All together, these inputs comprise the base model in such that the US Urban Areas represent “customer destinations” and the E-commerce facilities represent “facility origins”. This concept of origin and destination is the main mechanism for how distance within the network is approximated. From this compiled dataset, we constructed network edges and vertices by calculating the minimum end-to-end distances from 739 E-commerce facilities to 481 Urban Areas. Next, will cover the modern techniques and tools that made the collection and spatial analysis of these datasets both possible and exceedingly efficient.

Data Cleaning

The sourcing, pre-processing, and cleaning of all datasets was completed in the R programing language using R-Studio. To ensure accuracy and consistency the Tidyverse package collection within the R programing language was utilized for its robust and intuitive set of functions that standardize data analysis routines into a simple grammar that anyone can implement. This "grammar of data manipulation", is a consistent set of verbs that help you solve the most common data manipulation challenges. For example, *mutate()* adds new variables that are functions of existing variables, *select()* picks variables based on their names, and *filter()* picks cases based on their values [25].

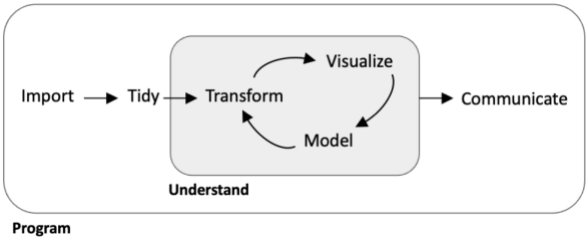


Figure 3-3: Overview of Tidy standard data process

These functions and the underlying data model uniquely facilitated the extensive pre-processing and cleaning that was required to structure the datasets into the appropriate spatial formats. A visualization of this process is included in Figure 3-3 that was adapted from the book: R for Data Science [23]. Notably the package Spatial Features (sf) were essential during preprocessing stage allowed tabular and spatial data to be managed within a single interface. A consistent feature across all Tidyverse tools is the ability to "pipe" data using an operator string "|>" to connect transformations into a single operation. This proved essential when generating the distance matrix tables required to generate the nearest neighbor links between Urban Areas and Facilities.

Pre-processing

After the datasets were imported and tidied, additional pre-processing was completed on both datasets to convert them into spatial data frames within R. The Facilities dataset was geocoded in a Geographic information system (GIS) to map the physical addresses into spherical coordinates of latitude and longitude. For the Urban Areas dataset, geocoding was not required as all US Census data is pre-geocoded. The cartographic shape files, were joined with Census survey data from the American Community Survey (ACS), to calculate the land area population estimate, and population density for each of the 481 US Urban Areas.

Summary

This chapter took a deep dive into the datasets used in this research. Data tables for Facilities and Urban Areas were provided to establish a spatial framing of how they are distributed across the nation. We introduced the Tidyverse foundational data analysis process and how this was used to ensured consistency of our unified tabular and spatial datasets. In the next chapter we will use the prepared dataset to create the data matrices needed to determine distance within our network.

Chapter 4

Constructing Network Distance by Segment

Network Topology is key concept of network design which focuses on the arrangement of facilities (nodes) and their connections (edges). An example matrix is illustrated in Table 4-1 on the right. Note the zeros along the diagonal of the matrix which represent null distance measurements for all self-to-self pairs. To construct our potential E-commerce network the pairwise distances were calculated in miles for all Facilities. The resulting matrix was symmetrical with the total number of records in the dataset. (739 x 739 = 546,121). Next, the same process was completed for all Urban Areas. The resulting matrix was also symmetrical with the total number of records in the dataset. (481 x 481 = 231,361). Lastly, the pairwise was determined for the join of Facilities and Urban Areas. The resulting matrix was a-symmetrical with the total number of records equal to the in the joined dataset. (481 x 739 = 355,459). All distance matrices calculated Haversine (or great circle) values. Haversine distance is a distance function like Euclidian or Manhattan distance used specifically for latitude and longitude coordinate pairs.

Table 4-1: Example Distance Matrix

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	0	35	18	92	93	5	70	2
B	12	0	73	85	18	24	12	70
C	43	86	0	58	49	87	7	7
D	81	43	90	0	92	82	96	31
E	91	25	34	56	0	9	59	52
F	58	54	23	61	63	0	87	57
G	79	99	98	80	52	71	0	59
H	20	46	17	1	92	44	40	0

$$hav(\theta) = hav(\varphi_2 - \varphi_1) + \cos(\varphi_1)\, cos(\varphi_2)\, hav(\lambda_2 - \lambda_1)$$

Where φ_1, φ_2 are the latitude of points 1 and 2 and λ_1, λ_2 are the longitude of points 1 and 2 [26]. Haversine can be thought of as a decimal representation of the angular distance between two

points on the surface of a sphere, in our case Earth! The Haversine function was applied during distance matrix generation and establishes the spatial glossary needed to lookup distances between origins (supply) and destinations (customer). A joint distance matrix was created for the Urban Areas -> Facilities dataset to quantify the distances between each Urban Area and its nearest facility. All distance operations utilize these three (3) baseline distance matrices to derive total segment and end-to-end distances. For symmetrical distance matrixes like the Facilities <-> Facilities and Urban Area <-> Urban Area datasets, only required the bottom triangle of the matrix to determine the forward distances between origins and destinations. Whereas in non-symmetrical cases like Urban Areas -> Facilities, all row are retained.

Nearest Neighbors

Next, to quantify the shortest segment distances in miles we find the Nearest Neighbor at each segment (hop) using the structure below (FC -> DC -> DS -> Urban Area).

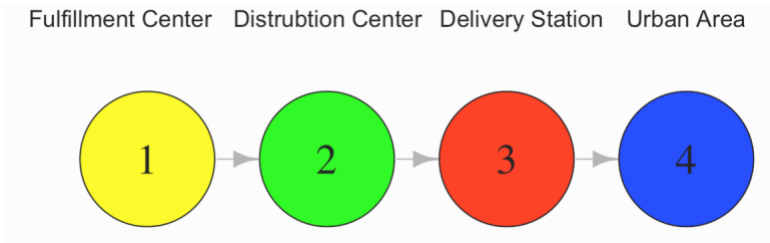


Figure 4-1: E-commerce Segments

To achieve this, the minimum end-to-end (FC -> DC -> DS -> Urban Area) distances between each segment were calculated from the generated holistic distance matrices in the prior step. The hierarchical relationship of the segments is illustrated in Figure 4-1. Using the Nearest Neighbors for each destination Urban Area, we compile the minimum and optimal network structure

between all origins and destinations. We convert Nearest Neighbors pairs into Edge (E) and Vertex (V) vectors and visualize the result for the largest component for Northern California.

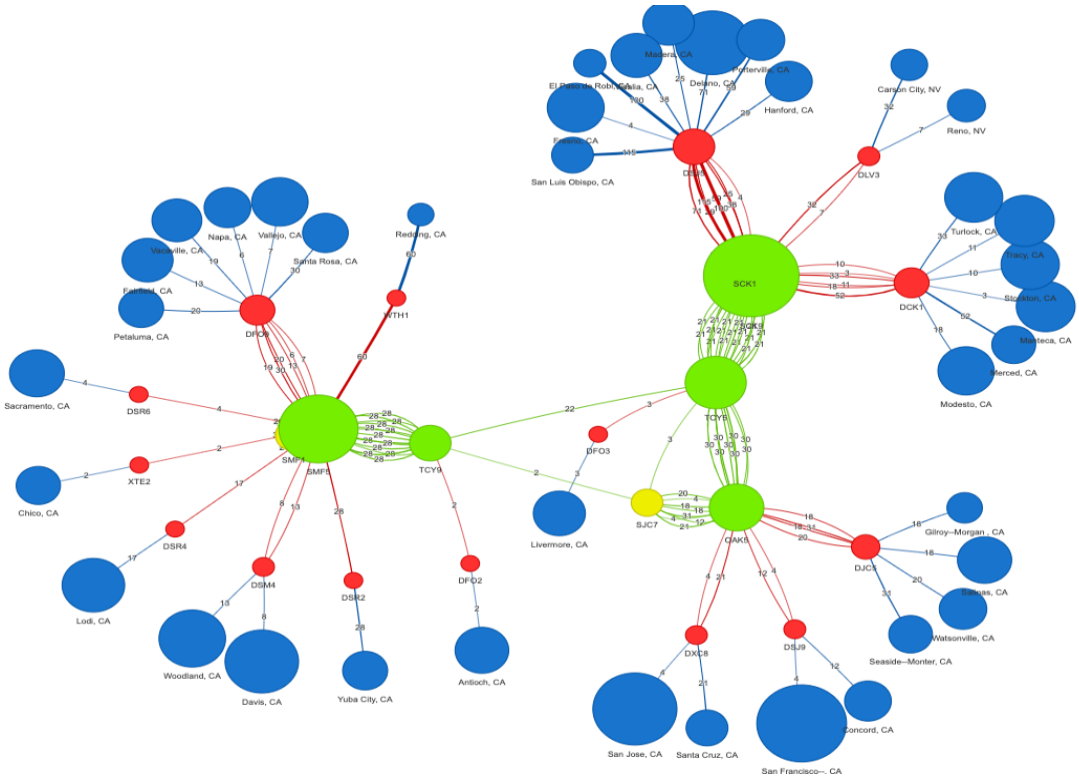


Figure 4-2: Nearest Neighbors for Northern California

Table 4-2: Northern California - Top 10: End-to-End Distance Table (Miles)

Urban Area	End-to-End	Delivery Station	Distribution Center	Fulfillment Center
San Luis Obispo, CA	230.18	115.22	114.89	0.07
El Paso de Robl, CA	215.45	100.49	114.89	0.07
Delano, CA	186.86	71.9	114.89	0.07
Redding, CA	180.06	60.32	91.6	28.14
Porterville, CA	174.87	59.9	114.89	0.07
Carson City, NV	169.9	32.33	137.5	0.07
Visalia, CA	153.8	38.84	114.89	0.07
Hanford, CA	144.81	29.85	114.89	0.07
Reno, NV	144.68	7.11	137.5	0.07
Madera, CA	140.03	25.07	114.89	0.07

Figure 4-2 illustrates the resulting network structure and the respective segment distances for the Top 10 end-to-end edges are listed in Table 4-2. From this analysis we can conclude that **San Luis Obispo, CA** has the longest end-to-end distance. In the next section we build upon these findings to assess the impact with considerations for population.

Weighted Segment Distances by Division

We use a weighted distance metric to determine which Urban Areas have the biggest opportunity for improvement relative to how many people are impacted. To calculate this while keeping our values readable we multiply the raw Segment Distance Miles calculated in the last step by Population Density (per 1,000 people) from our US Census Survey data. By grouping the weighted average segment distances, we can see how the divisions compare.

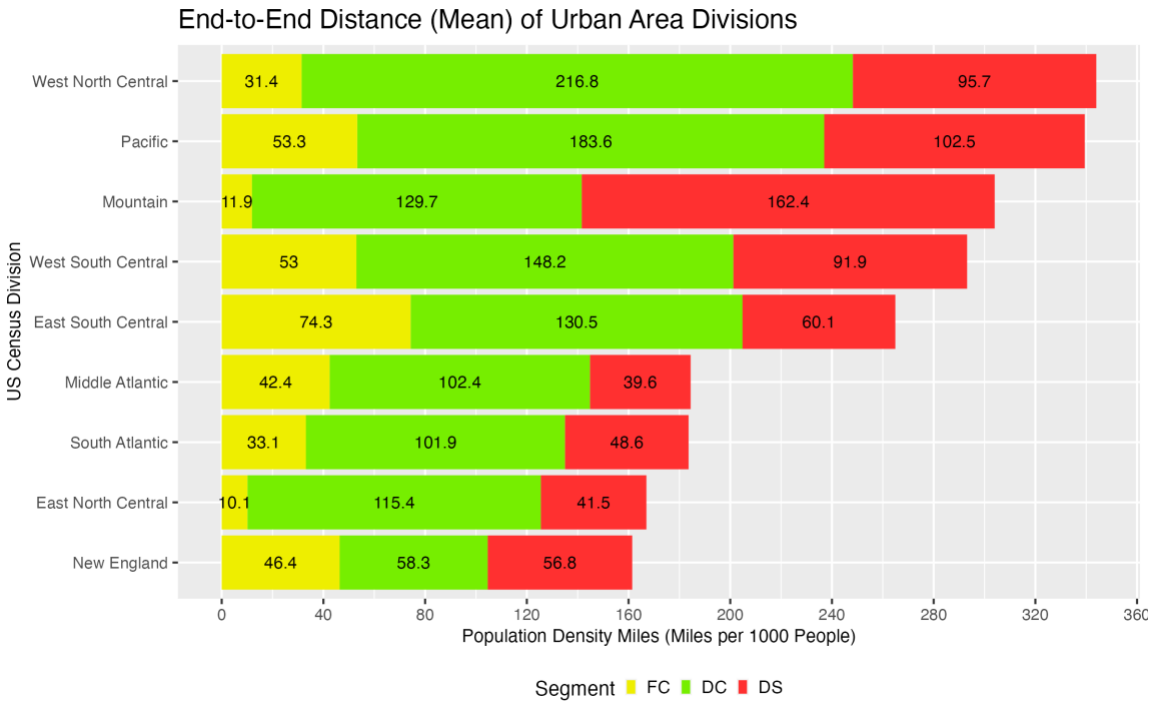


Figure 4-3: Weighted Segment Distances by Division

Figure 4-3 visualizes the aggregated segment distances by US Census division area. In this figure, we see that **West North Central** has the highest end-to-end weighted average distance within our network. Whereas, the division with the highest average DC -> FC distance is **East South Central** (74.3) and then again **West North Central** (216.8) has the highest average DS -> DC distance. We also calculated the full distances for all US Urban Areas and included it as [Appendix A].

Summary

In this chapter we implemented our cleaned dataset to create geocoded distance matrices that were leveraged to generate the distances pairs for all possible combinations of vertices. We refined this based on a structural understanding of our E-commerce network and Nearest Neighbor analysis. We defined a weighted metric to objectively compare distance across Urban Areas. Further distances to lower populated Urban areas and shorter distances to higher populated Urban areas is more optimal network than equally dividing the distance between all Urban Areas. Finally, we summarized the upstream (First-Mile) and downstream (Last-Mile) distance impacts by division. In the next chapter we will cover how these results change if we design the network with the DCs closer to the local population centers.

Chapter 5

Design Network for Max Population Density

To determine the best location for Distribution Centers, we calculated the weighted centroid of population density for each "cluster" of Urban Areas. We construct a modular clustering model with support for multiple algorithms. An illustration of the general model with input attributes is provided in Figure 5-1. There are three (3) categories of attributes considered in the mode. 1) Demographic attributes: these are attributes that quantify the population within the immediate area of a facility. Examples include; census boundaries, population density, personal consumption expenditures (PCE) (a.k.a. consumer spending patterns), unemployment rate, and available workers. 2) Geospatial attributes: these are attributes that quantify distance or quantity for features nearby a facility. Examples include; nearest highway exit ramp, nearest neighbor facility, and population within a 20-mile radius. 3) Physical attributes: these are attributes that quantify physical details of a facility. Examples include; address (latitude, longitude), land area within zip code, and building total sq. ft.

Based on our research we identified three (3) clustering algorithms to test for best fit with our Urban Areas dataset.

- Kmeans Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering and Application with Noise)

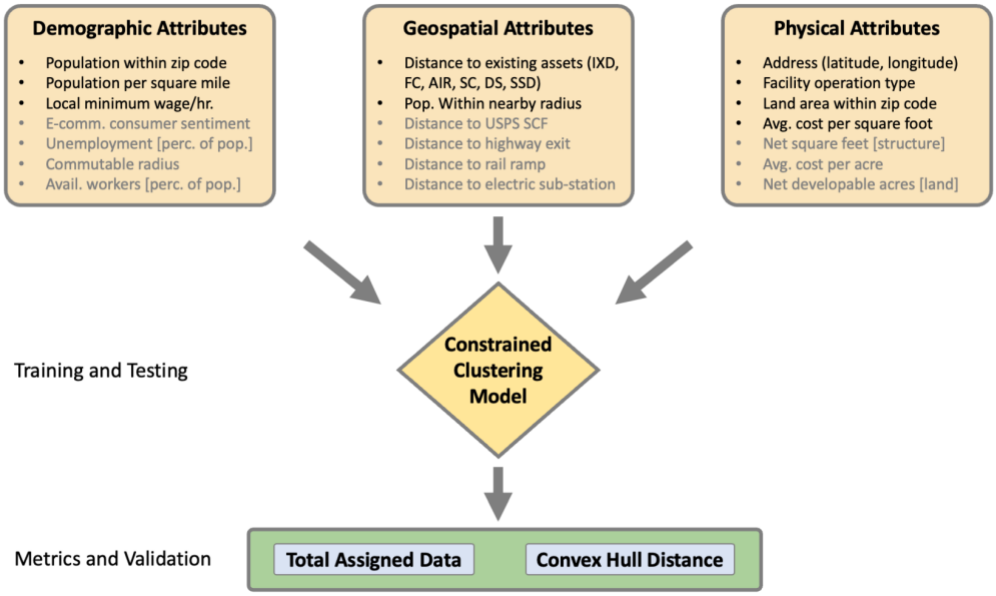


Figure 5-1: Network Design Clustering Model

Objective Cost Function

To quantify fit across the models we built an objective cost function to measure the results of each iteration of clustering. Once all nodes are assigned the result is evaluated by the cost function to determine a score. The objective of our cost function is to accurately identify areas of contiguous population. We desire clusters that are not too small and not too large. The function is comprised of two key metrics, 1) Minimum Complex Polygon (MCP) to measure cluster diameter, and 2) Percentage of Urban Areas Assigned.

Minimum Complex Polygon (MCP)

A Minimum Complex Polygon is a vector shape that doesn't have any interior angles greater than 180 degrees. The practice of applying MCP to clustering is well researched and the advantage is being able to isolate the perimeter points of each cluster. Finding the MCP allows us to measure the mean distance between the MCP vertices and the centroid of the cluster. This distance value is an indicator of fit and serves as an alternative to traditional linkage methods that measure the spacing between clusters. In this implementation we want to minimize the average vertex distance while still assigning as many Urban Areas to clusters as possible.

Percentage of Urban Areas Assigned

A counterbalance to MCP, percentage assigned metric measures the count of Urban Areas assigned to clusters for each iteration of the model. There is a natural tradeoff between density of clusters (measured in avg vertex distance) and high percentage of Urban Area assignment. Specifically in sparse areas with <3 points within 100 miles.

Clustering Model

To complete these experiments a clustering model was constructed to segment the Urban Areas into local communities. Within each local cluster the end-to-end distance impact of shifting the middle mile facilities closer to the population center can be determined. Because middle mile facilities have both upstream (FCs) and downstream (DSs) the adjustment of their location has a compounding effect and requires that a new distance matrix be generated for each set of middle mile coordinates. Specific results illustrated in the following Figures 5-2, 5-3, and 5-4.

Kmeans Clustering

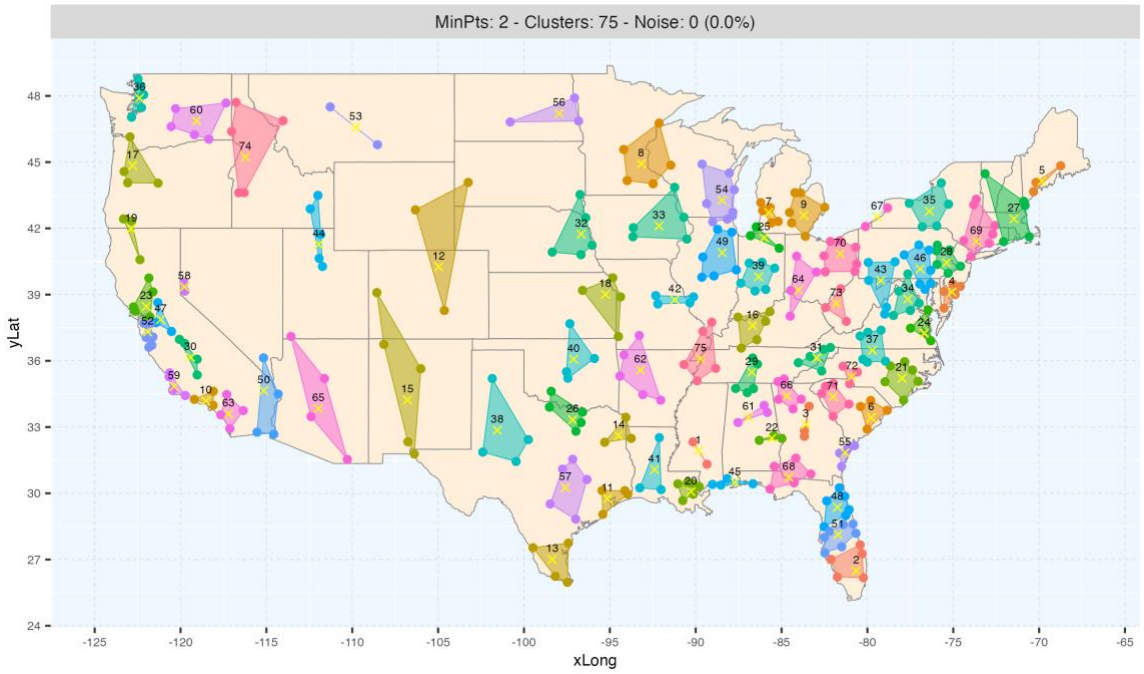


Figure 5-2: Kmeans Clustering results as Minimum Convex Polygons (MCPs)

The Kmeans clustering algorithm is designed to….

Table 5-1: Results of Kmeans Grid Scan Sorted by Avg. Dist.

rank	Clusters (k)	Clusters (#)	Avg. Vertex Dist.	Noise (%)
1	92	92	725.24	0.00%
2	64	64	729.62	0.00%
3	51	51	737.16	0.00%
4	56	56	740.51	0.00%
5	50	50	746.36	0.00%
6	90	90	749.93	0.00%
7	53	53	751.08	0.00%
8	87	87	754.7	0.00%
9	96	96	766.82	0.00%
10	66	66	767.82	0.00%

Hierarchical Clustering

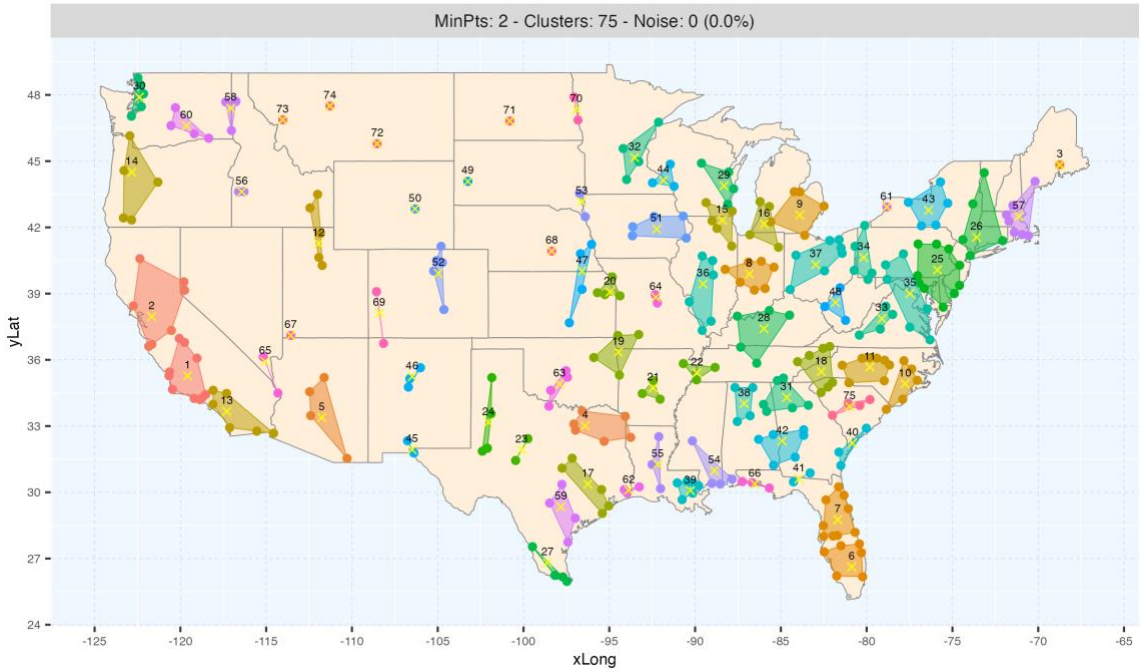


Figure 5-3: Hierarchical Clustering results as Minimum Convex Polygons (MCPs)

The Hierarchical clustering algorithm is designed to

Table 5-2: Results of Hierarchical Grid Scan Sorted by Avg. Dist.

rank	epsilon (eps)	Clusters (#)	Avg. Vertex Dist.	Noise (%)
1	100	100	636.87	0.00%
2	97	97	638.55	0.00%
3	96	96	640.48	0.00%
4	99	99	641.68	0.00%
5	95	95	642.88	0.00%
6	85	85	643.65	0.00%
7	86	86	647.11	0.00%
8	98	98	647.11	0.00%
9	94	94	648.64	0.00%
10	84	84	650.93	0.00%

DBSCAN (Density-Based Spatial Clustering and Application with Noise)

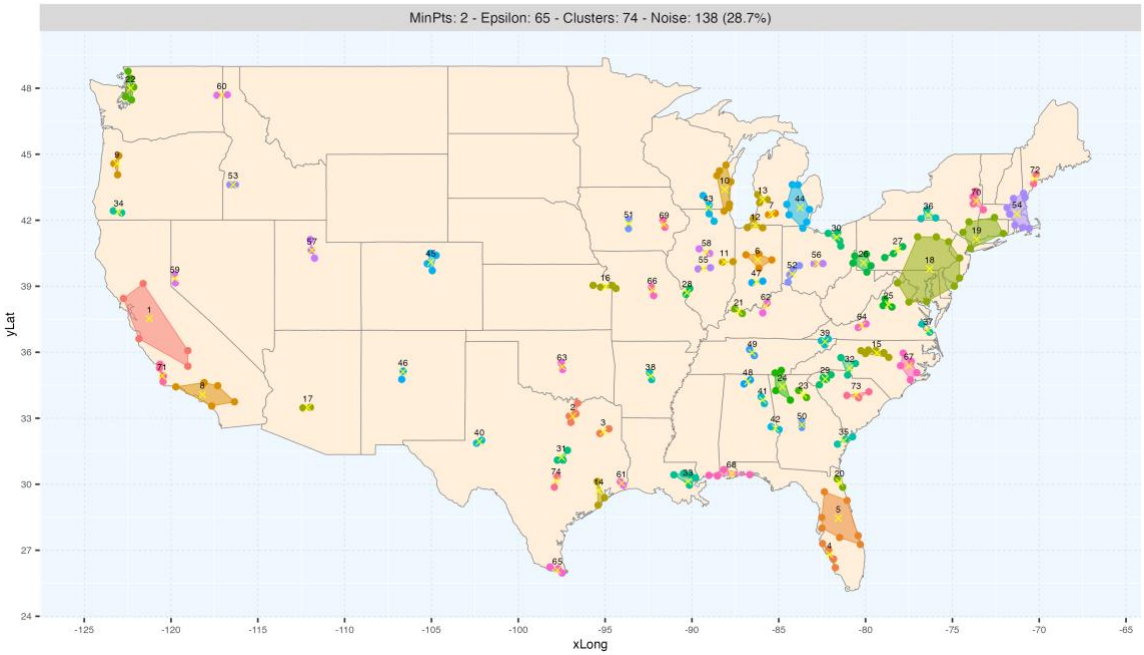


Figure 5-4: DBSCAN Clustering results as Minimum Convex Polygons (MCPs)

The DBSCAN algorithm works by identifying neighborhoods of nodes determined by counting the measuring the distances between nodes within a specific radius of a centroid point. This approach leaves the algorithm resilient to "noise" or outlier data points as they will typically fall outside of the local neighborhood radius. Two important parameters are required for the algorithm to return desirable results: *epsilon* ("eps") and *minimum points* ("MinPts") [21]. *Epsilon* defines the radius of neighborhood around a point x within which nodes will be counted. The parameter *MinPts* is the minimum number of neighbor nodes within the epsilon radius. For the extent of this research *minimum points* is defined a priori, given that our intention to a cluster our data should result in aggregation of at least 2 nodes. To determine the optimal *epsilon* value, we will build a simple grid search loop to investigate a reasonable search space and then quantify

the results based on the cost function. When interpreting the results of the cost function we also consider the number of assigned Urban Areas vs “noise” records excluded from the result.

The Vertex Centroid Cost Function measures the distance from each of the vertices to the centroid of their cluster. This provides a basic measure of clustering fit. All of the vertices for each *epsilon* are averaged for the each run of the DBSCAN algorithm and then the next value in the grid search sequence will executes. A linear approximation of this process is illustrated in Figure 5-5 on the right with exact outputs captured in Table 5-3. If you follow the linear plane downward the MCPs will get smaller, and also break into more clusters. This identified specific local clusters, and regions that are not well connected within the national network.

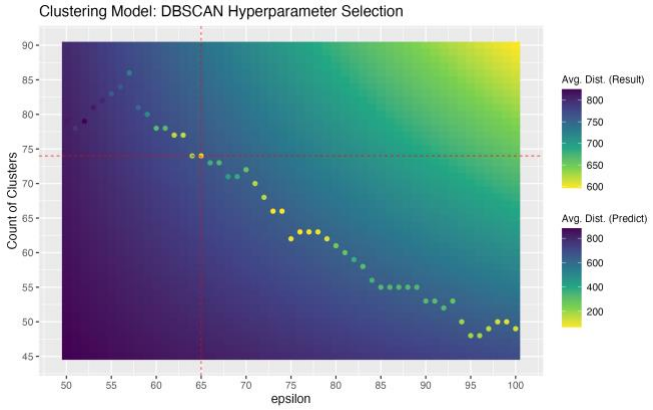


Figure 5-5: DBSCAN Hyperparameter Selection

Table 5-3: Results of Grid Scan Sorted by Avg. Dist. & Noise

Rank	Epsilon (eps)	Clusters (#)	Avg. Vertex Dist.	Noise (%)
1	73	66	595.8	22.25%
2	76	63	596.24	20.58%
3	74	66	596.49	21.62%
4	77	63	599.18	20.37%
5	78	63	601.39	19.54%
6	75	62	603.64	21.21%
7	99	50	604.55	13.51%
8	79	62	606.79	19.33%
9	65	74	608.28	28.69%
10	98	50	610.31	14.14%

Based on the cost table results of all three (3) algorithms, the DBSCAN algorithm is best suited for our Urban Area population dataset. The results are intuitive, given what we know about populations and their tendency to form local communities. There is a higher cost to building in less dense areas as the existence of critical infrastructure is lower. However, convex hull distances do still require a balancing metric such as percentage of total data records assigned.

Initial Bearing Coordinates

Using thew weighted centroids from the clustering model, an initial bearing is calculated for each Distribution Center. Figure 5-6 illustrates the new projected locations on a national map. The original latitude and longitude of each DC is shifted 10 miles to create a revised location set.

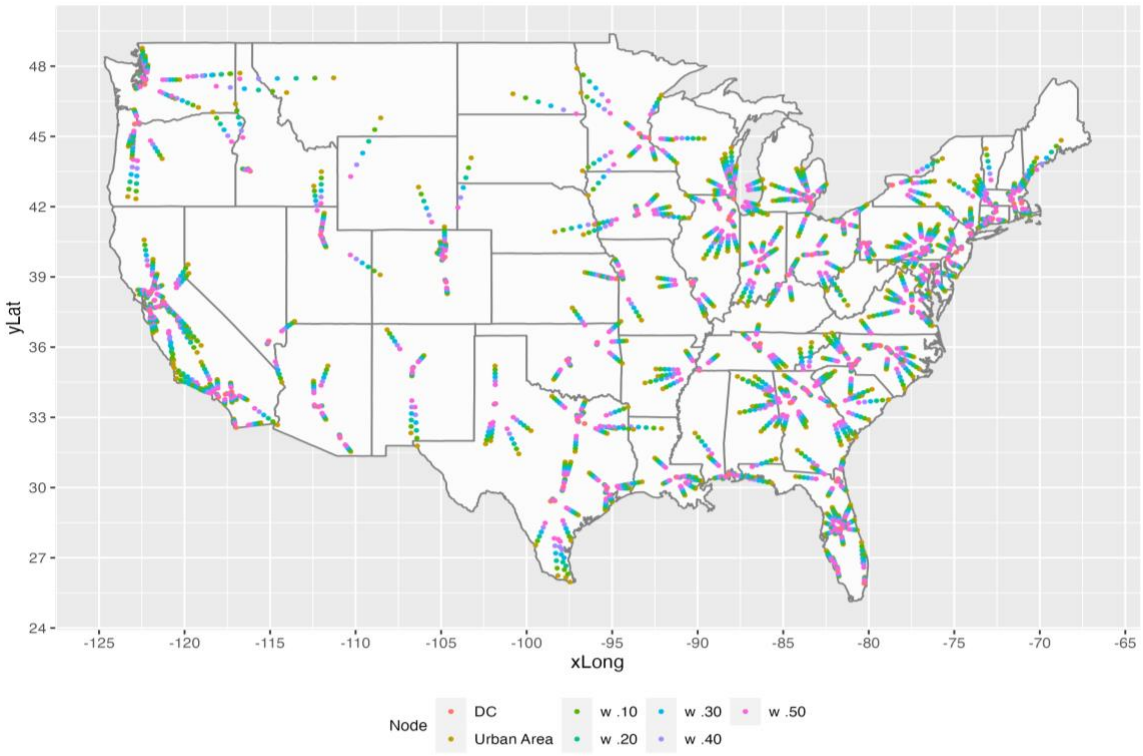


Figure 5-6: National map of Cluster model w DC centroids shifted in five steps

Weighted Segment Distances by Division (New DC Locations)

New distance matrices are created using these revised Distribution Center locations to determine the total regional impact in the same way we did for the original network.

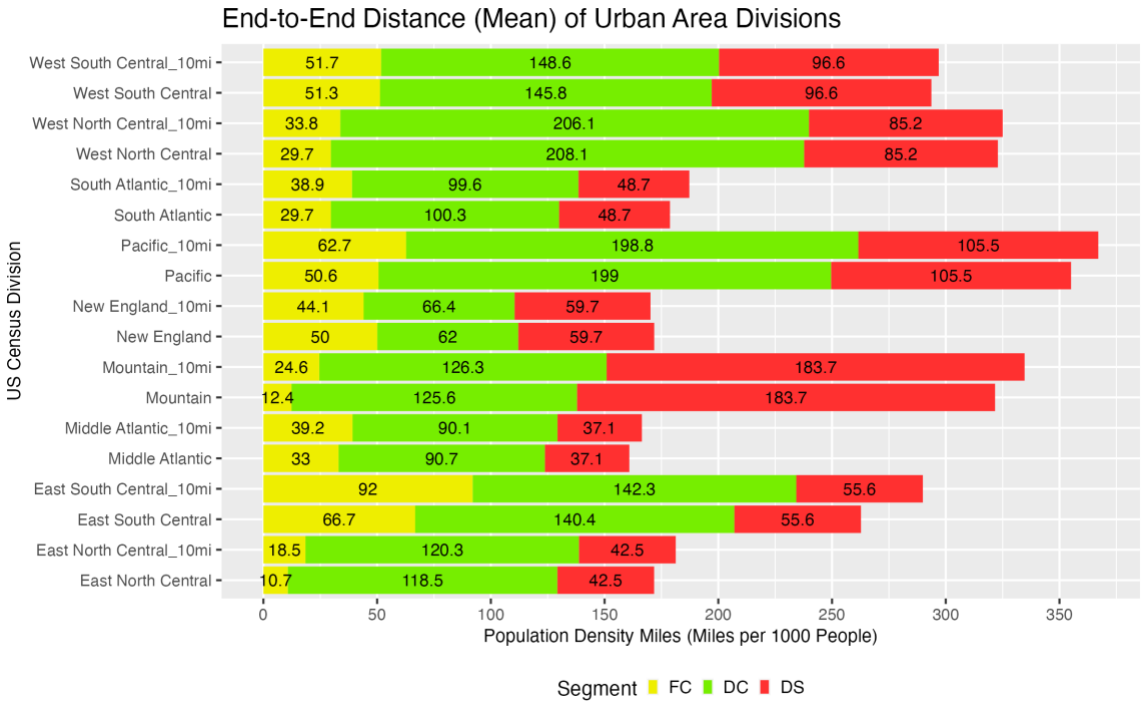


Figure 5-7: Weighted Segment Distances by Division (New -vs- Old DC Locations)

From the bar graph in Figure 5-7 we see that **Pacific** now has the highest end-to-end weighted average distance within our network. With, the **Mountain** divisions average DC -> FC distance increased 98.2%. Similarly, **West North Central** increased 13.8%. Overall, there is a 3.55% increase in total network distance by shifting the DC locations by 10 miles. The full data frame analysis was completed and provided in [Appendix B].

Summary

In this chapter we introduced the concept of maximizing Middle Mile locations by population density. We constructed a modular clustering model and tested it with 3 different algorithms. We built an objective cost function to compare results across all algorithms and selected DBSCAN as the best fitted model for our dataset. covered the approach and strategy for our analysis. Using the clustered results from DBSCAN we used geospatial analysis to determine the initial bearing from the original Distribution Center coordinates to the population maximum of the cluster. New location coordinates were created at 10 miles closer to this maximum. Finally, we connected all of the pieces and recalculated the average distances across each division to realize a 13.5% decrease in total network distance. In the next chapter we apply these distance results to Delivery Speed and Transportation Cost metrics and visualize their impact on specific example facilities.

Chapter 6

Quantifying Network Impact

Using the results from the revised Distribution Center locations we calculate the specific impact to key performance metrics Delivery Speed and Transportation Cost at the Local, Regional, and National scopes.

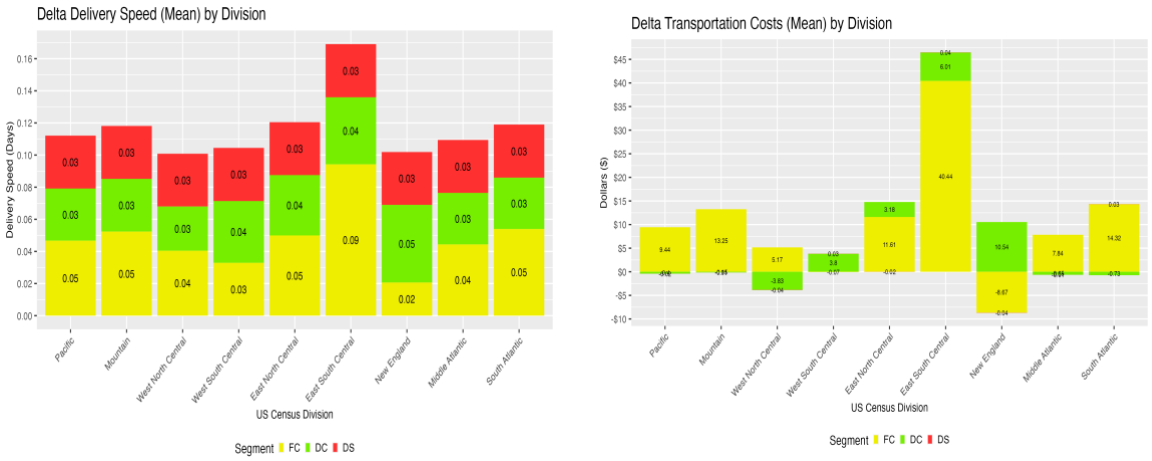


Figure 6-1: Predicted performance costs for DC simulated network

Based on the adjusted Distribution Center (DC) locations 10-miles closer to their population maxima. From the results in Figure 6-1 we can see that the impact to all most all divisions is negative with the exception of New England. Overall, the average daily impact of the simulated network to be +0.11 Days (2.64 hrs.) for Delivery Speed and +$7.39 for Transportation Cost.

Delivery Speed

To approximate Delivery Speed, we use commercial trucking milage bands to estimate the speed impact in days from point-of-click until customer delivery. Dwell time where packages sit idle between segments is not captured. Table 6-1 below summarizes the quadratic relationship.

Table 6-1: Delivery Speed to Distance Estimate Matrix

Delivery Speed (Days)	Distance Radius (Miles)
0	75
1	150
2	250
3	350

Proposed new locations that have a positive impact to delivery speed, can be extrapolated using the delivery promise checkout conversion identified in our literature review. Using this guidance, we assume that for any segment distance less than 250 miles we will realize an 11% lift in additional potential revenue [17]. We apply this at 50% to retain a conservative estimate conversion as delivery time decreases. The total sum of this product is aggregated at the cluster level to generate a Delivery Speed impact.

Transportation Costs

Just like Delivery Speed, Transportation Costs are also a function of distance. Given the results of our simulation model we input the resulting raw mile distances to derive the total cost impact of truck freight transportation within our network. First, we establish some general assumptions about demand in order to accurately predict what the impact to transportation costs will be. We assume an estimated 10 billion packages to traverse our network annually, which is

27.5 million packages daily. US Urban Areas account for ~70% (232.5M) of total US Population (329.5M), thus we expect Urban Packages to be ~8.3% of Urban Population (~19.4M).

A per mile cost assumption of $2.90 per mile is assumed for First-Mile and Middle-Mile based on National Private Truck Council (NPTC) estimates [8]. These estimates include holistic costs to operate the needed transportation and allow for generalized inference within our model. Similarly, Last-Mile (small parcel) costs are predicted at a per mile cost of $1.46. However, within the Last-Mile scope any distances downstream of the Urban Area centroid is not captured by this model.

By establishing a generalized demand model based on the regionalized population density we are able to predict expected Transportation costs at the Urban Area level. Additional scope to expand the resolution of this functionality is added into Chapter 7. The ~8.3% of population is calculated at the Urban Area level to aggregate the total number of expected daily packages to originate from that location. This is then extrapolated using the First-Mile and Last-Mile cost assumptions described above.

Considerations for target use case are made relative to what specific leadership recommendations could be made from the results of the simulated what-if scenario. Despite the +3.5% unfavorable result of the DC population density approach, I demonstrated how the simulation model provides real-world value in What-If scenario testing and visualization for network designers and decision makers. The quantified average daily impact of the simulated network to be +0.11 Days (2.64 hrs.) for Delivery Speed and +$7.39 for Transportation Cost. These results suggest that using a lightweight simulation model early in the network design process can mitigate cost unfavorable location selection.

Local, Regional, National

Local Impact

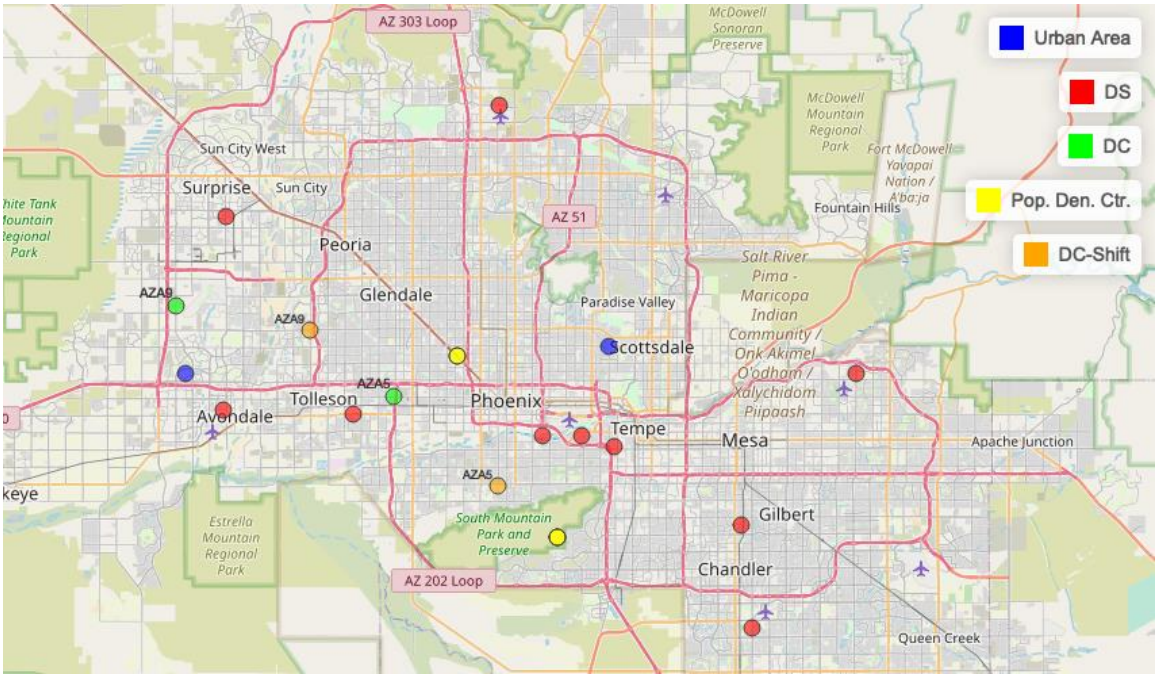


Figure 6-2: Local Results – Phoenix, AZ

In Figure 6-2 we can see that the revised location for Distribution Center AZA8 is more centrally located within the cluster. This new location reduced the Upstream (First Mile) and Downstream (Last Mile) distances within the cluster positively impacting transportation costs and delivery speed. Regional is the result of consolidated local cluster level impact.

Regional Impact

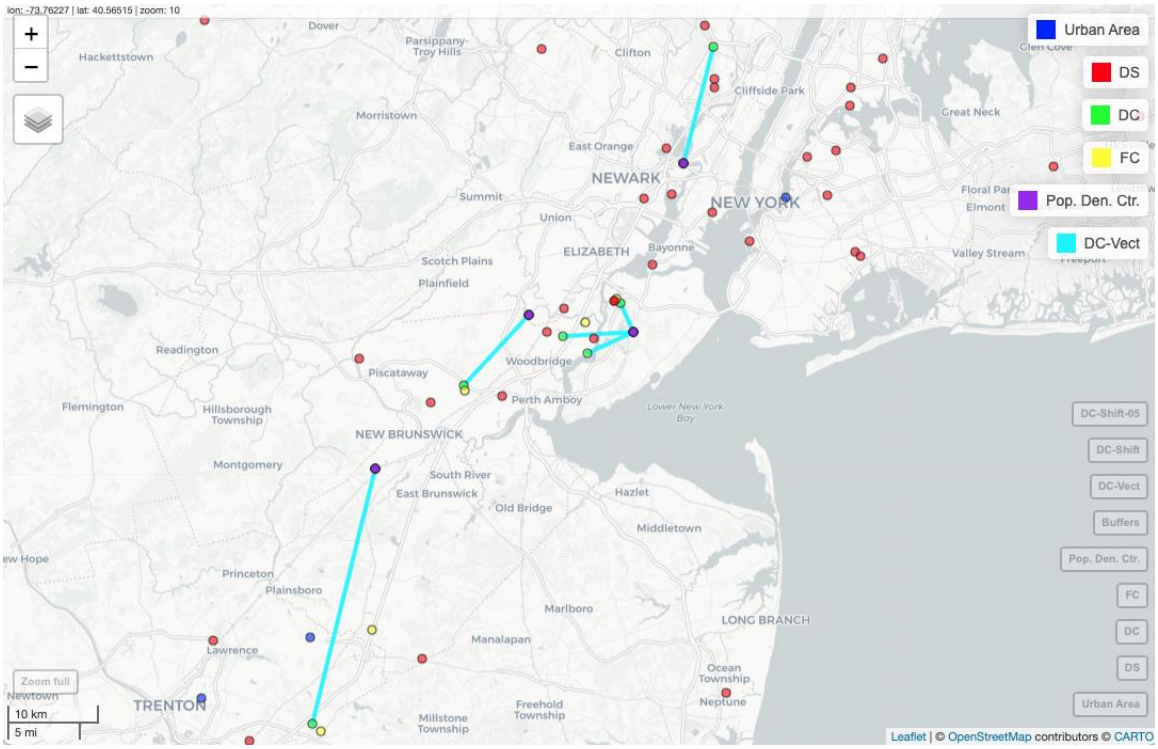


Figure 6-3: Regional Results – Pennsylvania, New Jersey, New York Tri-state area

The constructed constrained clustering model was executed on our computed distance matrix tables. Weighted means were calculated using the Urban Area population density of the cluster as the gravity to pull the centroid closer as the weights of the function increased. For Figure 6-3 we see the centralization of the shifted DCs closer to the middle of each population center. Aggregating these distances up from the local clusters will give us the regional impact. We look at this regional example to see how the DCs are pulled towards the population density maximas. In this view we can see that there is a very strong pull from New York City Urban Area with a slight counter pull from Eastern PA, NJ, and Middletown, NY Urban Areas.

National Impact

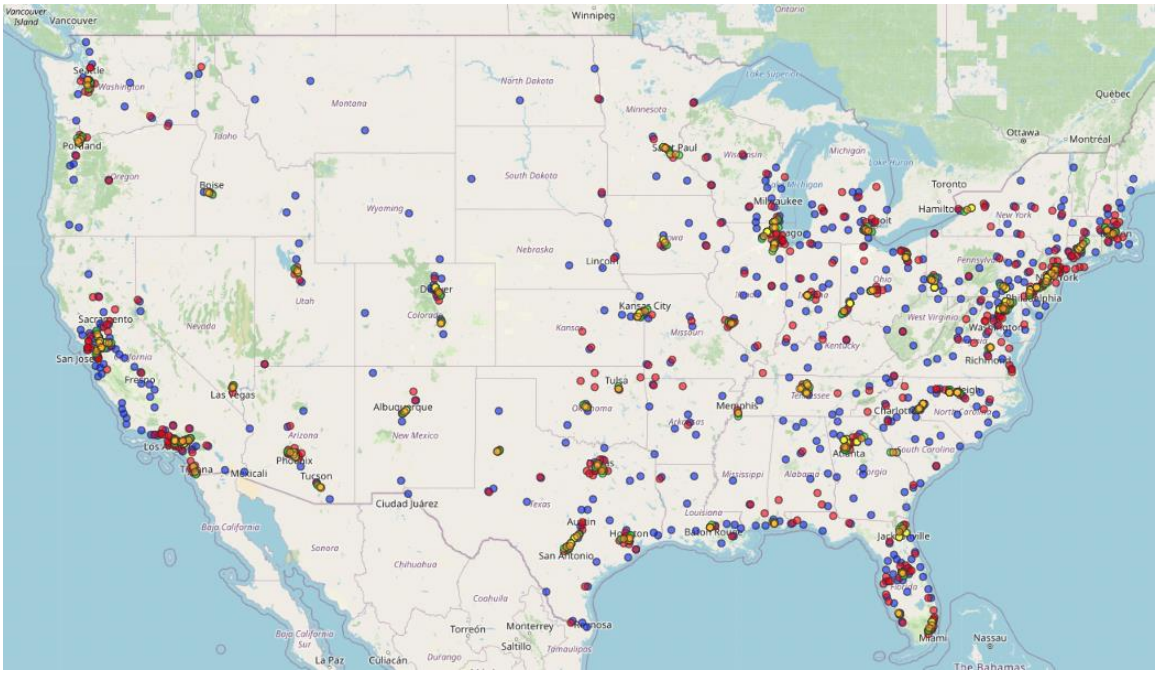


Figure 6-4: National Results – Continental United States

At the national level we can fully see the total extent of the network. In the interactive model network designers are able to toggle layers and perform additional analysis functions to quantify the impacts from the local and regional levels that comprise the final national network. Figure 6-4 visualizes the results of the shifted Distribution Centers 10 miles. Leveraging the visual results in tandem with the data table in Appendix B we can confirm that the change is not optimal relative to our baseline model. This approach increased distances between First-Mile and Middle-Mile more than it reduced the distances between Middle-Mile and Last-Mile segments. This resulted in a +3.5% increase in total end-to-end network distance which is cost unfavorable.

Chapter 7

Conclusions

Strong E-commerce demand over the past decade has resulted in market constraints on available land and labor that limit the ideal location selection of new E-commerce facilities. Finding the best alternative location requires data analysis. The significance of this research is in its ability to quantify two key performance metrics: 1) Having too many facilities within the same commutable distance radius will exhaust the local labor supply and increase operating costs. 2) Land must be suitable for development of large (1 million+ sq. ft.) facilities. The average national commercial real estate vacancy rate is less than 4% (96% occupied) as of Q1 2022. Consolidating all of the findings and results, this research demonstrated that it is possible to positively impact end-to-end network distances at the local, regional, and national levels with a supervised visualization approach.

Despite the poor performance of the holistic population density approach, we demonstrated that the simulation model provides real-world value in what-if scenario testing for network designers and decision makers. Using the simulation model, network designers can quantify the impact adding or removing facilities has on the local, regional and national network and make informed decisions based on quantified performance metrics.

Given the constrained markets and continued competitive pressures for commercial real estate in the E-commerce space any new E-commerce facilities will need to consider 2 key factors, 1) avoid cannibalizing the workforce at existing facilities, and 2) quantify an improvement e.g., Delivery Speed and Transportation Cost. Using population density in balance with the existing local facility locations network planners can derive recommendations that improve the delivery speed and transportation cost of the network using the model proposed in this research.

Limitation of the Research

A lightweight lower accuracy model is acceptable in the facility location selection space due to market constraints on available land and labor that limit the ideal location selection of new E-commerce facilities. Such a model would allow for modular expansion as new inputs are discovered and can then be integrated into the business logic. However, the current implementation is also limited by hyperparameter tuning to arrive at optimal cluster assignment. Future versions of this model should incorporate different clustering algorithms and linkage techniques to improve continuity within the local population areas, increase density of the assigned clusters, and increase quantity of total Urban Areas assigned to clusters.

A foundational assumption within the simulation model is the use of haversine distances to determine relative distance between respective node pairs. This raises a question around the realized accuracy of the haversine function given example cases where road network differs dramatically from the Euclidian context inherent in the spherical haversine distance. As the measured distance increases, the expected margin of error with using Haversine distance will decrease.

Future Work

The proposed model and methodologies discussed in this research stand to add real-world value to E-commerce organizations. Providing quantifiable location selection recommendations is an essential part of supply chain network design. By implementing the recommendations in this research, E-commerce organizations can ensure that their decision makers are setup for success!

Trade-offs between speed and cost objectives potentially require a separate field of study on their own. This research found that decision criteria for a speed optimized network will not

always be conducive to a cost optimized network. These trade-offs require additional examination in addition to considerations on environmental impact and how these could be derived into realistic cost functions. This is especially true in more sparse population areas where the demand cannot sustain the required infrastructure to realize the same speed that would be inherent in a densely population metro area.

Furthermore, considerations for traffic in metro and urban environments should be considered in future work. Current model assumptions place no cost for building in densely populated areas which are known to have high traffic impacts. This research found Traffic Studies as standard practice within the location selection process. However, traffic considerations should be more robust and could be added as a dedicated cost function during clustering.

In closing, we provide a blueprint for future work to further optimize network design using the simulation model, methodologies, and tooling developed in this research:

1. Network directed distances (road distance); available via different modes (Truck, Passenger Car) and with or without Traffic insights.
2. Traffic considerations for impact on all segments (First, Middle, and Last Mile)
3. For further optimizing clustering consider the total intersected area of interstates as an additional research avenue.
4. Adding additional metrics/datasets detailed in Figure 5-1. Extending the inputs to the clustering could be expanded by these inputs to further inform cluster generation.
5. A/B Location Evaluation (Site Selection) – Simulate prospect facility impact.
6. Predict the ideal centroids of new facility locations for each operation type.
7. Extend impact quantification to First-Mile and Last-Mile networks
8. Extend impact quantification to Sub-Operation Type (e.g., Non-conveyable, Conveyable package processing).

Appendix A

Actual End-to-End Distance by Urban Area

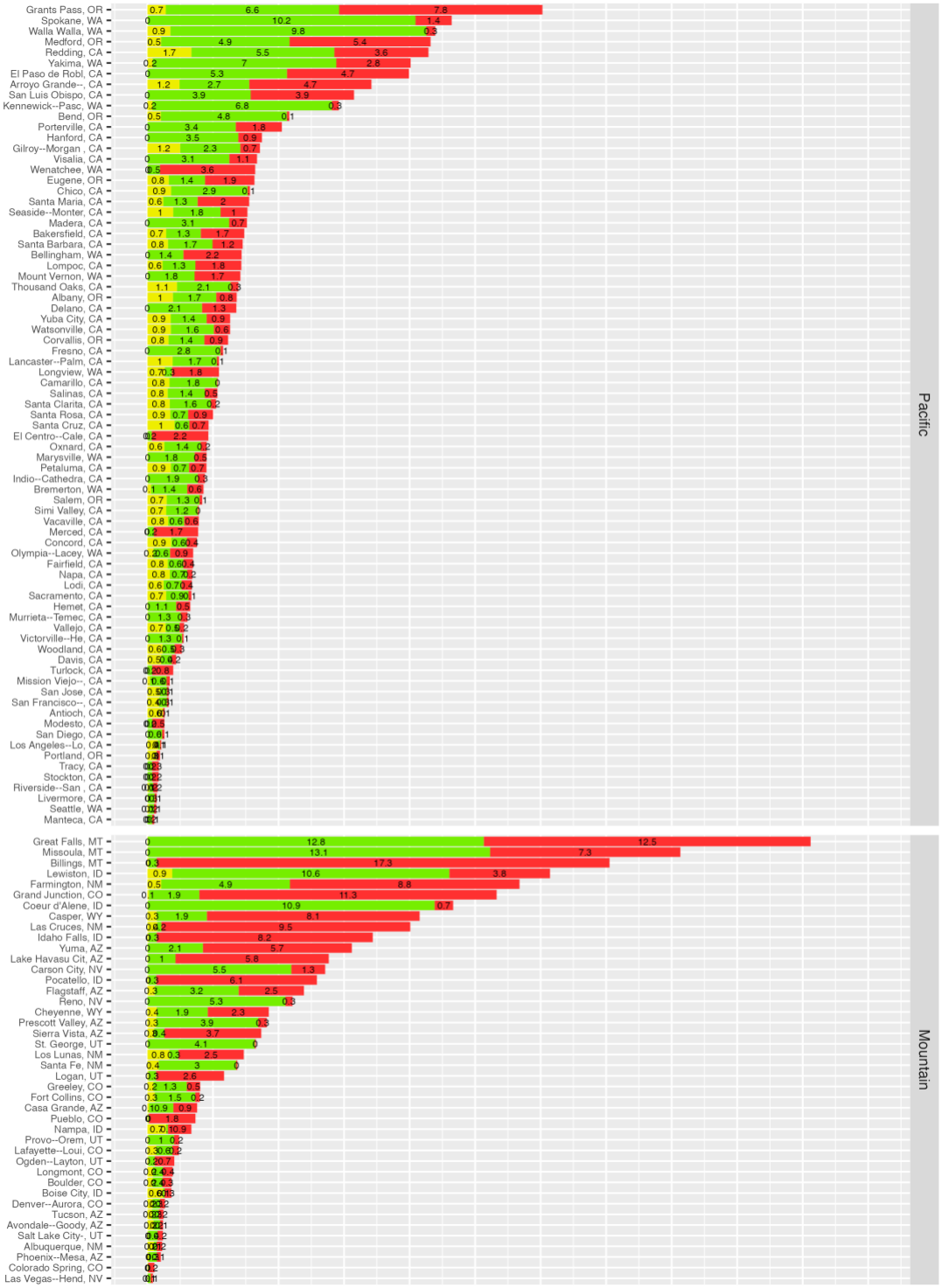
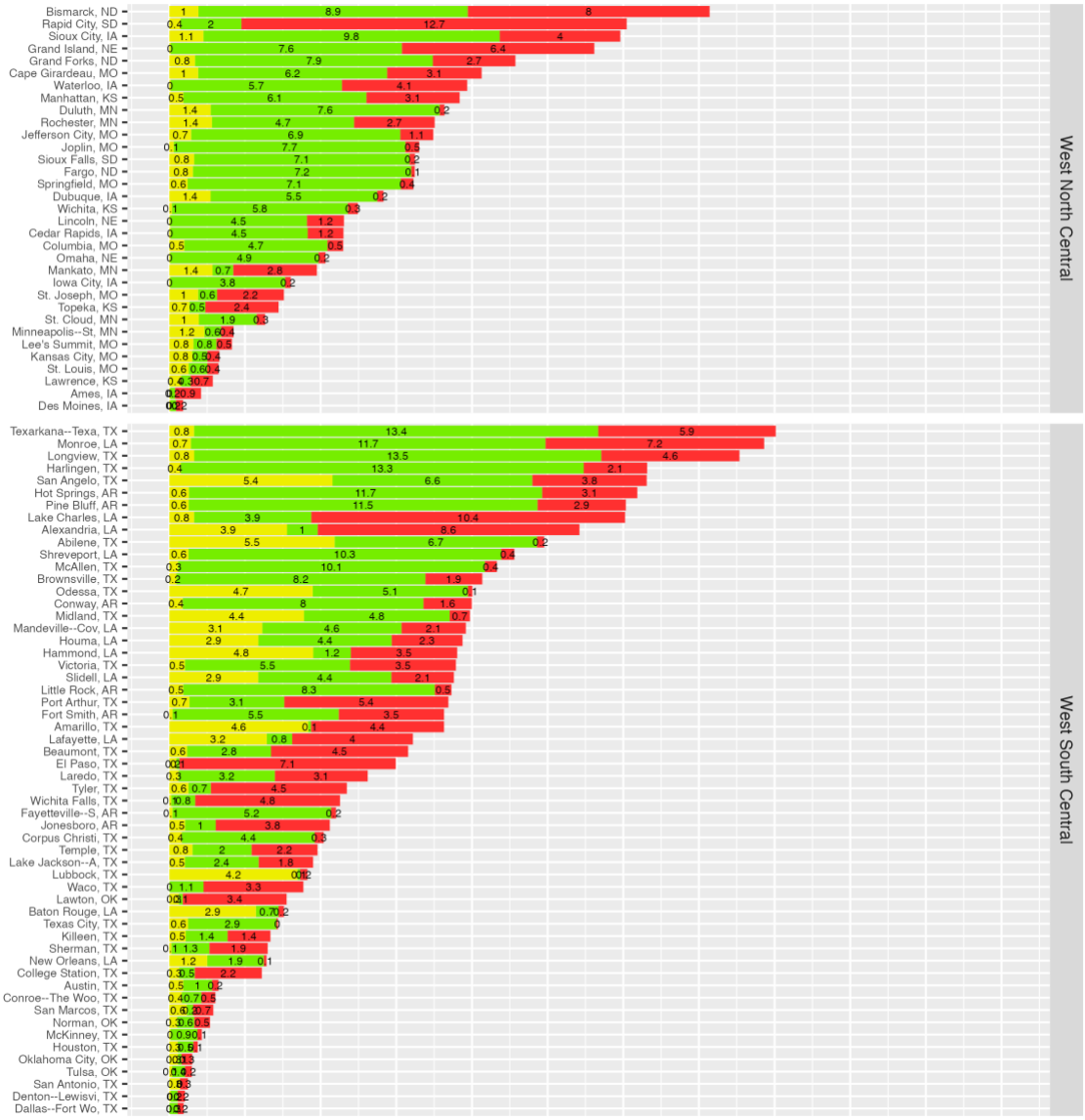
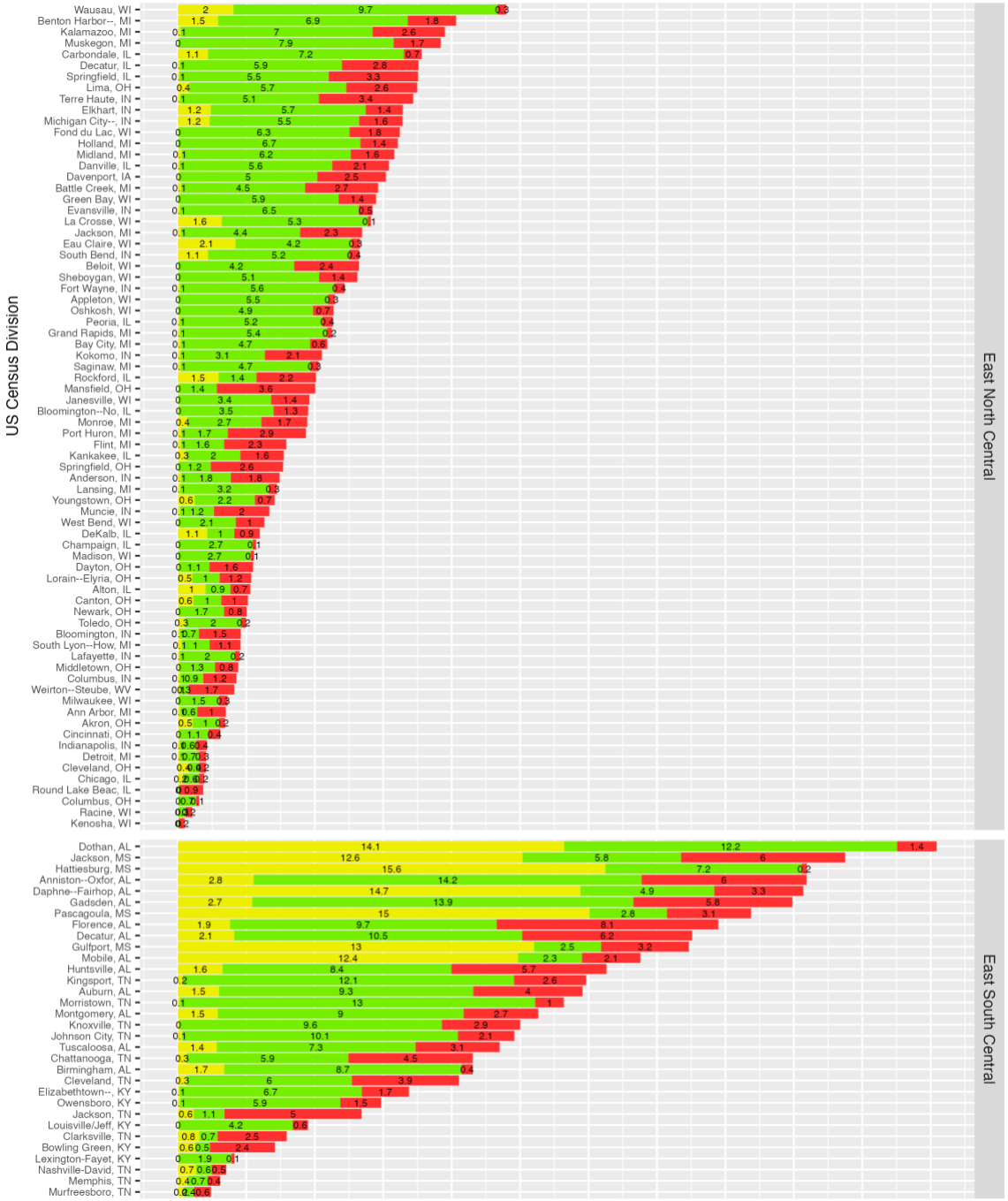
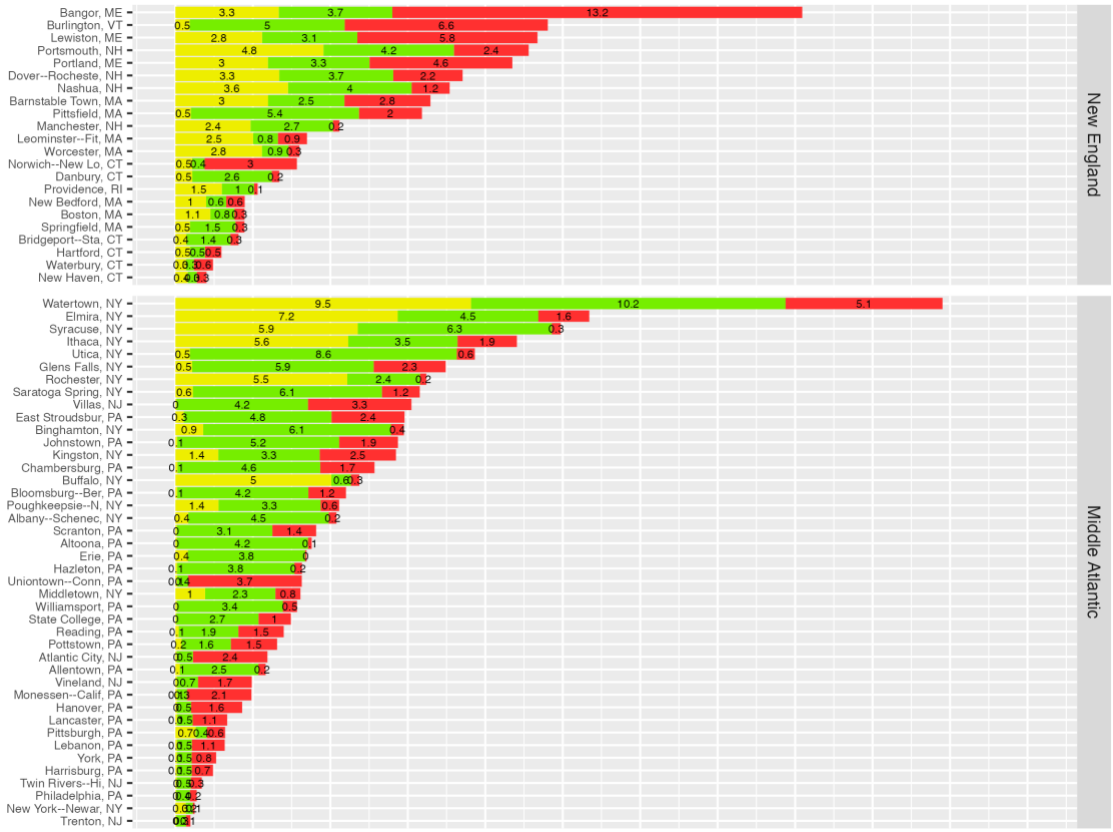
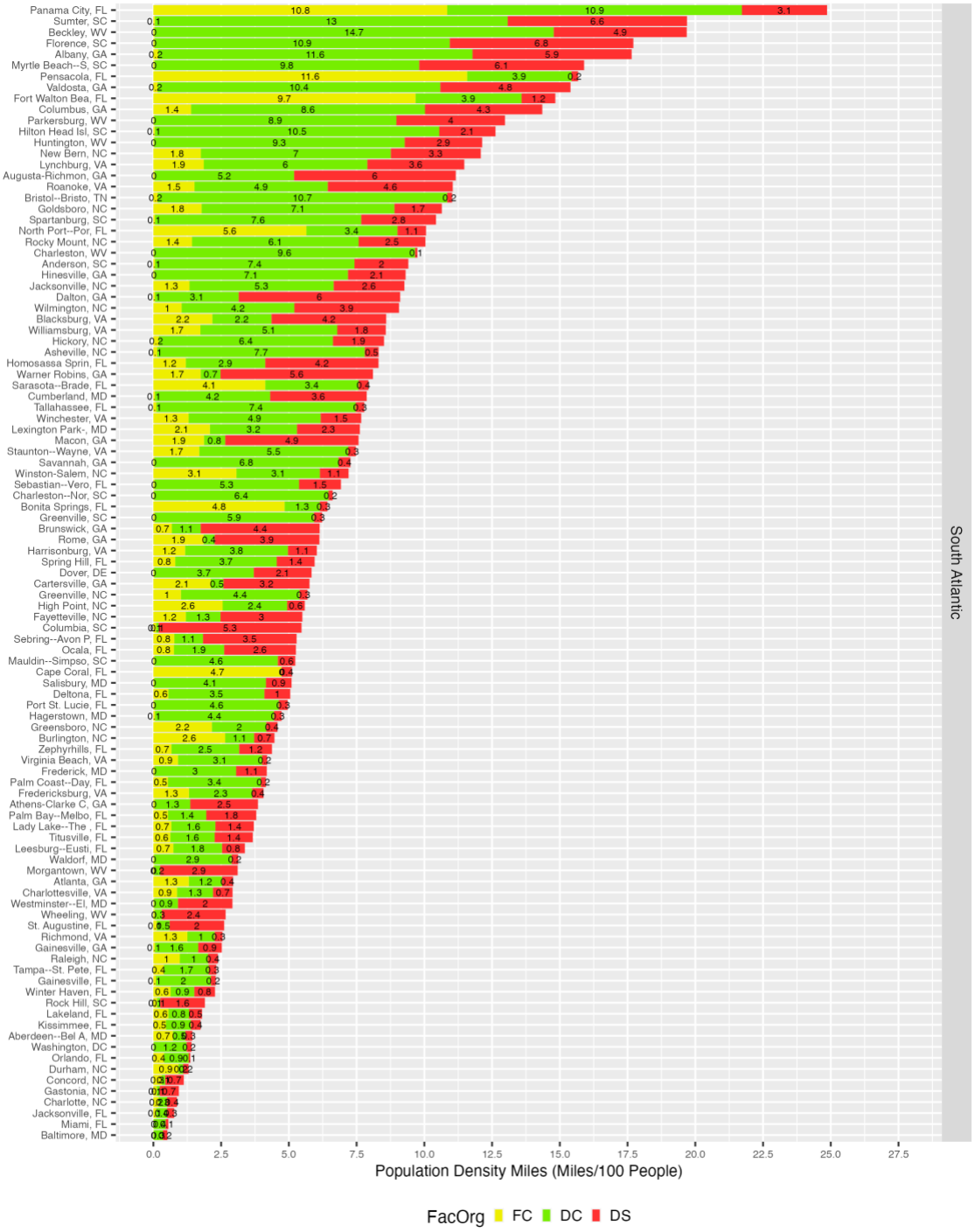


Figure Appendix A: Itemized results of segment distance for all 481 Urban Areas. The segments represent the measured distance between Customer “destination” and Fulfillment Center “source”.




US Census Division
East North Central
East South Central
Wausau, WI
Benton Harbor--, MI
Kalamazoo, MI
Muskegon, MI
Carbondale, IL
Decatur, IL
Springfield, IL
Lima, OH
Terre Haute, IN
Elkhart, IN
Michigan City--, IN
Fond du Lac, WI
Holland, MI
Midland, MI
Danville, IL
Davenport, IA
Battle Creek, MI
Green Bay, WI
Evansville, IN
La Crosse, WI
Jackson, MI
Eau Claire, WI
South Bend, IN
Beloit, WI
Sheboygan, WI
Fort Wayne, IN
Appleton, WI
Oshkosh, WI
Peoria, IL
Grand Rapids, MI
Bay City, MI
Kokomo, IN
Saginaw, MI
Rockford, IL
Mansfield, OH
Janesville, WI
Bloomington--No, IL
Monroe, MI
Port Huron, MI
Flint, MI
Kankakee, IL
Springfield, OH
Anderson, IN
Lansing, MI
Youngstown, OH
Muncie, IN
West Bend, WI
DeKalb, IL
Champaign, IL
Madison, WI
Dayton, OH
Lorain--Elyria, OH
Alton, IL
Canton, OH
Newark, OH
Toledo, OH
Bloomington, IN
South Lyon--How, MI
Lafayette, IN
Middletown, OH
Columbus, IN
Weirton--Steube, WV
Milwaukee, WI
Ann Arbor, MI
Akron, OH
Cincinnati, OH
Indianapolis, IN
Detroit, MI
Cleveland, OH
Chicago, IL
Round Lake Beac, IL
Columbus, OH
Racine, WI
Kenosha, WI
Dothan, AL
Jackson, MS
Hattiesburg, MS
Anniston--Oxfor, AL
Daphne--Fairhop, AL
Gadsden, AL
Pascagoula, MS
Florence, AL
Decatur, AL
Gulfport, MS
Mobile, AL
Huntsville, AL
Kingsport, TN
Auburn, AL
Morristown, TN
Montgomery, AL
Knoxville, TN
Johnson City, TN
Tuscaloosa, AL
Chattanooga, TN
Birmingham, AL
Cleveland, TN
Elizabethtown--, KY
Owensboro, KY
Jackson, TN
Louisville/Jeff, KY
Clarksville, TN
Bowling Green, KY
Lexington-Fayet, KY
Nashville-David, TN
Memphis, TN
Murfreesboro, TN


Bangor, ME 3.3 3.7 13.2
Burlington, VT 0.5 5 6.6
Lewiston, ME 2.8 3.1 5.8
Portsmouth, NH 4.8 4.2 2.4
Portland, ME 3 3.3 4.6
Dover--Rocheste, NH 3.3 3.7 2.2
Nashua, NH 3.6 4 1.2
Barnstable Town, MA 3 2.5 2.8
Pittsfield, MA 0.5 5.4 2
Manchester, NH 2.4 2.7 0.2
Leominster--Fit, MA 2.5 0.8 0.9
Worcester, MA 2.8 0.9 0.3
Norwich--New Lo, CT 0.5 0.4 3
Danbury, CT 0.5 2.6 0.2
Providence, RI 1.5 1 0.1
New Bedford, MA 1 0.6 0.6
Boston, MA 1.1 0.8 0.3
Springfield, MA 0.5 1.5 0.3
Bridgeport--Sta, CT 0.4 1.4 0.3
Hartford, CT 0.5 0.5 0.5
Waterbury, CT 0.3 0.6
New Haven, CT 0.4 0.3
New England
Watertown, NY 9.5 10.2 5.1
Elmira, NY 7.2 4.5 1.6
Syracuse, NY 5.9 6.3 0.3
Ithaca, NY 5.6 3.5 1.9
Utica, NY 0.5 8.6 0.6
Glens Falls, NY 0.5 5.9 2.3
Rochester, NY 5.5 2.4 0.2
Saratoga Spring, NY 0.6 6.1 1.2
Villas, NJ 0 4.2 3.3
East Stroudsbur, PA 0.3 4.8 2.4
Binghamton, NY 0.9 6.1 0.4
Johnstown, PA 0.1 5.2 1.9
Kingston, NY 1.4 3.3 2.5
Chambersburg, PA 0.1 4.6 1.7
Buffalo, NY 5 0.6 0.3
Bloomsburg--Ber, PA 0.1 4.2 1.2
Poughkeepsie--N, NY 1.4 3.3 0.6
Albany--Schenec, NY 0.4 4.5 0.2
Scranton, PA 0 3.1 1.4
Altoona, PA 0 4.2 0.1
Erie, PA 0.4 3.8 0
Hazleton, PA 0.1 3.8 0.2
Uniontown--Conn, PA 0 0.4 3.7
Middletown, NY 1 2.3 0.8
Williamsport, PA 0 3.4 0.5
State College, PA 0 2.7 1
Reading, PA 0.1 1.9 1.5
Pottstown, PA 0.2 1.6 1.5
Atlantic City, NJ 0 0.5 2.4
Allentown, PA 0.1 2.5 0.2
Vineland, NJ 0 0.7 1.7
Monessen--Calif, PA 0 0.3 2.1
Hanover, PA 0 0.5 1.6
Lancaster, PA 0 0.5 1.1
Pittsburgh, PA 0.7 0.4 0.6
Lebanon, PA 0 0.5 1.1
York, PA 0 0.5 0.8
Harrisburg, PA 0 0.5 0.7
Twin Rivers--Hi, NJ 0 0.5 0.3
Philadelphia, PA 0 0.8 2
New York--Newar, NY 0 0.8 2 1
Trenton, NJ 0 0.3 1
Middle Atlantic


Panama City, FL 10.8 10.9 3.1
Sumter, SC 0.1 13 6.6
Beckley, WV 0 14.7 4.9
Florence, SC 0 10.9 6.8
Albany, GA 0.2 11.6 5.9
Myrtle Beach--S, SC 0 9.8 6.1
Pensacola, FL 11.6 3.9 0.2
Valdosta, GA 0.2 10.4 4.8
Fort Walton Bea, FL 9.7 3.9 1.2
Columbus, GA 1.4 8.6 4.3
Parkersburg, WV 0 8.9 4
Hilton Head Isl, SC 0.1 10.5 2.1
Huntington, WV 0 9.3 2.9
New Bern, NC 1.8 7 3.3
Lynchburg, VA 1.9 6 3.6
Augusta-Richmon, GA 0 5.2 6
Roanoke, VA 1.5 4.9 4.6
Bristol--Bristo, TN 0.2 10.7 0.2
Goldsboro, NC 1.8 7.1 1.7
Spartanburg, SC 0.1 7.6 2.8
North Port--Por, FL 5.6 3.4 1.1
Rocky Mount, NC 1.4 6.1 2.5
Charleston, WV 0 9.6 0.1
Anderson, SC 0.1 7.4 2
Hinesville, GA 0 7.1 2.1
Jacksonville, NC 1.3 5.3 2.6
Dalton, GA 0.1 3.1 6
Wilmington, NC 1 4.2 3.9
Blacksburg, VA 2.2 2.2 4.2
Williamsburg, VA 1.7 5.1 1.8
Hickory, NC 0.2 6.4 1.9
Asheville, NC 0.1 7.7 0.5
Homosassa Sprin, FL 1.2 2.9 4.2
Warner Robins, GA 1.7 0.7 5.6
Sarasota--Brade, FL 4.1 3.4 0.4
Cumberland, MD 0.1 4.2 3.6
Tallahassee, FL 0.1 7.4 0.3
Winchester, VA 1.3 4.9 1.5
Lexington Park-, MD 2.1 3.2 2.3
Macon, GA 1.9 0.8 4.9
Staunton--Wayne, VA 1.7 5.5 0.3
Savannah, GA 0 6.8 0.4
Winston-Salem, NC 3.1 3.1 1.1
Sebastian--Vero, FL 0 5.3 1.5
Charleston--Nor, SC 0 6.4 0.2
Bonita Springs, FL 4.8 1.3 0.3
Greenville, SC 0 5.9 0.3
Brunswick, GA 0.7 1.1 4.4
Rome, GA 1.9 0.4 3.9
Harrisonburg, VA 1.2 3.8 1.1
Spring Hill, FL 0.8 3.7 1.4
Dover, DE 0 3.7 2.1
Cartersville, GA 2.1 0.5 3.2
Greenville, NC 1 4.4 0.3
High Point, NC 2.6 2.4 0.6
Fayetteville, NC 1.2 1.3 3
Columbia, SC 0.1 5.3
Sebring--Avon P, FL 0.8 1.1 3.5
Ocala, FL 0.8 1.9 2.6
Mauldin--Simpso, SC 0 4.6 0.6
Cape Coral, FL 4.7 0.4
Salisbury, MD 0 4.1 0.9
Deltona, FL 0.6 3.5 1
Port St. Lucie, FL 0 4.6 0.3
Hagerstown, MD 0.1 4.4 0.3
Greensboro, NC 2.2 2 0.4
Burlington, NC 2.6 1.1 0.7
Zephyrhills, FL 0.7 2.5 1.2
Virginia Beach, VA 0.9 3.1 0.2
Frederick, MD 0 3 1.1
Palm Coast--Day, FL 0.5 3.4 0.2
Fredericksburg, VA 1.3 2.3 0.4
Athens-Clarke C, GA 0 1.3 2.5
Palm Bay--Melbo, FL 0.5 1.4 1.8
Lady Lake--The , FL 0.7 1.6 1.4
Titusville, FL 0.6 1.6 1.4
Leesburg--Eusti, FL 0.7 1.8 0.8
Waldorf, MD 0 2.9 0.2
Morgantown, WV 0.2 2.9
Atlanta, GA 1.3 1.2 0.4
Charlottesville, VA 0.9 1.3 0.7
Westminster--El, MD 0 0.9 2
Wheeling, WV 0.3 2.4
St. Augustine, FL 0 0.5 2
Richmond, VA 1.3 1 0.3
Gainesville, GA 0.1 1.6 0.9
Raleigh, NC 1 1 0.4
Tampa--St. Pete, FL 0.4 1.7 0.3
Gainesville, FL 0.1 2 0.2
Winter Haven, FL 0.6 0.9 0.8
Rock Hill, SC 0.1 1.6
Lakeland, FL 0.6 0.8 0.5
Kissimmee, FL 0.5 0.9 0.4
Aberdeen--Bel A, MD 0.7 0.5 0.3
Washington, DC 0 1.2 0.2
Orlando, FL 0.4 0.9 0.1
Durham, NC 0.9 0.2
Concord, NC 0.2 0.7
Gastonia, NC 0.1 0.7
Charlotte, NC 0.2 0.4
Jacksonville, FL 0.1 0.3
Miami, FL 0.1
Baltimore, MD 0.2
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0 27.5
Population Density Miles (Miles/100 People)
South Atlantic
FacOrg FC DC DS

Appendix B

Distance impact of DC simulation (DCs 10-miles)

Table Appendix B: Distance results of DC Simulation (DCs 10-miles)

Division Name	FacOrg	E2E (Orig)	E2E (10 mi.)	E2E Diff	Seg. (Orig)	Seg. (10 mi.)
East North Central	DC	171.78	181.28	-5.53%	118.4	120.8
East North Central	DS	171.78	181.28	-5.53%	42.51	42.51
East North Central	FC	171.78	181.28	-5.53%	10.72	18.49
East South Central	DC	262.72	289.93	-10.36%	140.6	142.0
East South Central	DS	262.72	289.93	-10.36%	55.65	55.65
East South Central	FC	262.72	289.93	-10.36%	66.72	91.99
Middle Atlantic	DC	160.87	166.40	-3.44%	90.75	90.13
Middle Atlantic	DS	160.87	166.40	-3.44%	37.08	37.08
Middle Atlantic	FC	160.87	166.40	-3.44%	33.04	39.19
Mountain	DC	321.68	334.66	-4.04%	125.6	126.4
Mountain	DS	321.68	334.66	-4.04%	183.5	183.5
Mountain	FC	321.68	334.66	-4.04%	12.36	24.57
New England	DC	171.82	170.19	0.95%	62.04	66.38
New England	DS	171.82	170.19	0.95%	59.73	59.73
New England	FC	171.82	170.19	0.95%	50.05	44.08
Pacific	DC	355.09	367.06	-3.37%	198.8	198.4
Pacific	DS	355.09	367.06	-3.37%	105.9	105.9
Pacific	FC	355.09	367.06	-3.37%	50.63	62.73
South Atlantic	DC	178.75	187.26	-4.76%	100.5	99.59
South Atlantic	DS	178.75	187.26	-4.76%	48.72	48.72
South Atlantic	FC	178.75	187.26	-4.76%	29.68	38.94
West North Central	DC	322.91	325.03	-0.66%	208.9	206.8
West North Central	DS	322.91	325.03	-0.66%	85.16	85.16
West North Central	FC	322.91	325.03	-0.66%	29.66	33.79
West South Central	DC	293.68	296.91	-1.10%	145.4	148.8
West South Central	DS	293.68	296.91	-1.10%	96.58	96.58
West South Central	FC	293.68	296.91	-1.10%	51.26	51.75
Total		**6,717.90**	**6,956.16**	**-3.55%**	**2,239.30**	**2,318.72**

Appendix C

Delivery Speed and Trans. Costs for simulation (DCs 10-miles)

Table Appendix C: Delivery Speed and Trans. Costs for Simulation (DCs 10-miles)

Division Name	FacOrg	Distance Delta	Delivery Speed	Trans. Cost
East North Central	DC	1.10	0.04	$ 3.18
East North Central	DS	-0.01	0.03	$ (0.02)
East North Central	FC	4.00	0.05	$ 11.61
East South Central	DC	2.07	0.04	$ 6.01
East South Central	DS	0.03	0.03	$ 0.04
East South Central	FC	13.94	0.09	$ 40.44
Middle Atlantic	DC	-0.22	0.03	$ (0.65)
Middle Atlantic	DS	-0.01	0.03	$ (0.01)
Middle Atlantic	FC	2.70	0.04	$ 7.84
Mountain	DC	-0.05	0.03	$ (0.15)
Mountain	DS	-0.01	0.03	$ (0.01)
Mountain	FC	4.57	0.05	$ 13.25
New England	DC	3.63	0.05	$ 10.54
New England	DS	-0.02	0.03	$ (0.04)
New England	FC	-2.99	0.02	$ (8.67)
Pacific	DC	-0.14	0.03	$ (0.40)
Pacific	DS	-0.01	0.03	$ (0.02)
Pacific	FC	3.25	0.05	$ 9.44
South Atlantic	DC	-0.25	0.03	$ (0.73)
South Atlantic	DS	0.02	0.03	$ 0.03
South Atlantic	FC	4.94	0.05	$ 14.32
West North Central	DC	-1.32	0.03	$ (3.83)
West North Central	DS	-0.03	0.03	$ (0.04)
West North Central	FC	1.78	0.04	$ 5.17
West South Central	DC	1.31	0.04	$ 3.80
West South Central	DS	0.02	0.03	$ 0.03
West South Central	FC	-0.02	0.03	$ (0.07)

Appendix D

Build the full segment distance data table (R Code)

```
      ## Build the full segment distance data table


      dfDistFacUAs %>%
        # Filter to show a-symetrical matrix of UA -> Facilities
        filter(!grepl("[a-z]+", orig, ignore.case = T)) %>%
        filter(!grepl("^[0-9]", dest, ignore.case = T)) %>%
        # {.}
        # Sort by miles to get shortest paths first
        arrange(distMiles) %>%
        left_join(dfB %>% select(user_uuid, DS_min_FacOrg =
user_org), by = c("dest" = "user_uuid"), keep = T) %>%
        group_by(orig, DS_min_FacOrg) %>% #DON'T MESS WITH THIS
        # {.}
        # 1) FIRST: Summarize the dist() matrix to find closest DS
facility to each orig
        summarise(
          # nearFacUuids = paste(nearFacUuid[1:3], collapse = "|"),
          # neardistMiless = paste(floor(distMiles[1:3]), collapse =
"|"),
          DS_min_FacUuid = user_uuid[which.min(distMiles)],
          # min_FacOrg = user_org[which.min(distMiles)], #DON"T MESS
WITH THIS, use group by
          DS_min_distMiles = min(distMiles, na.rm = T),
          n_DS = n(),
          .groups = "drop"
        ) %>%
        filter(grepl("DS", DS_min_FacOrg)) %>%
        # Join Nearest Neighbor Minimum Distances for DCs
        # 2) SECOND: Join to find nearest DC facility
        left_join(dfDistFacs_min %>% select(DC_min_distMiles =
min_distMiles, DC_min_FacUuid = min_FacUuid, DC_min_FacOrg =
min_FacOrg, n_DC = n, orig),
                  by = c("DS_min_FacUuid" = "orig")) %>%
        filter(grepl("DC", DC_min_FacOrg)) %>%
        select(!DC_min_FacOrg) %>%
        # 3) THIRD: Join to find lateral DC facility
        left_join(dfDistFacs_min %>% select(DCl_min_distMiles =
min_distMiles, DCl_min_FacUuid = min_FacUuid, DCl_min_FacOrg =
min_FacOrg, n_DCl = n, orig),
```

```
                  by = c("DC_min_FacUuid" = "orig")) %>%
        filter(grepl("DC", DCl_min_FacOrg)) %>%
        select(!DCl_min_FacOrg) %>%
        # 4) FORTH: Join to find nearest FC facility
        left_join(dfDistFacs_min %>% select(FC_min_distMiles =
min_distMiles, FC_min_FacUuid = min_FacUuid, FC_min_FacOrg =
min_FacOrg, n_FC = n, orig),
                  by = c("DC_min_FacUuid" = "orig")) %>%
        filter(grepl("FC", FC_min_FacOrg)) %>%
        select(!FC_min_FacOrg) %>%
        # {.}
        # Hide match count cols for Nearest Neighbors
        select(!matches("^n")) %>%
        rowwise() %>%
        mutate(
          # Calculate the end-to-end distannces xFC.minDist +
xDC.minDist + xDS.minDist = e2eDistMiles
          e2eDistMiles = FC_min_distMiles + DC_min_distMiles +
DS_min_distMiles,
        ) %>%
        ungroup() %>%
        left_join(dfGeoStack %>% select(UA_name = name10, UA_popDens
= aPopDens10, UA_geoid = geoid) %>% st_drop_geometry(), by =
c("orig" = "UA_geoid"), keep = T) %>%
        mutate(
          UA_state = str_sub(gsub("^.*,", "", UA_name), 2, 3),
          UA_stub = str_sub(gsub(",.*$", "", UA_name), 1, 15),
          clusterId = 0
        ) %>%
        arrange(desc(e2eDistMiles)) %>%
        select(orig, e2eDistMiles, order(tidyselect::peek_vars()))
%>%
        {.} -> dfDistFacUAs_segs
      # dfDistFacUAs_segs
```

Appendix E

Grid search DBSCAN parameters (R Code)

```
      library(purrr)

      # Density-Based Spatial Clustering and Application with Noise
- analysis on a set of dissimilarities

      # Method for determining the optimal eps value
      # This is not applicable to our use case and recommends a sub
optimal value,
      dbscan::kNNdistplot(dmLocsUAsClust, k = 2) +
        abline(h = 60, lty = 2, col = "red")

      # Scrappy Grid Search, all epsilon within seq
      rm(out)
      # t <- seq(2, 3, by = 1)
      s <- seq(50, 100, by = 1)
      out <- vector("list", length(s))

      for (i in s) {
        print(i)
        out[[i]] <- dbscan::dbscan(dmLocsUAsClust, eps = i, weights
= NULL, minPts = 2)
      }

      # Current best based on visual assesment
      # out[[60]]$cluster

      # Collapse list elements into single data.frame and name
columns based on seq
      res <- bind_cols(
          # (dfGeoStack %>% pull(geoid)), # Assign rownames from
original df
          (dfMapData %>% pull(uuid)), # Assign rownames from
original df
          map(out, ~ .x$cluster)
        ) %>%
        `colnames<-`(c("geoid", s))
```

Appendix F

Weighted centroid by population density (R Code)

```
      ### Weighted xLong and yLat by Population Density

      # aPopDens10, label, label2, org, uuid, xLong, yLat, zip1,
zipcode, r150, geoid, aPopDens10_UA, xLong_UA, yLat_UA, geometry,
DC_pointBuffer, geometryPoint, geometryPoint_UA, nodeWeight,
meanW_yLat, meanW_xLong, geometryPoint_meanW
      dfMapFacBuffsUAs %>%
        distinct(uuid, org, geometryPoint_DC, geometryPoint_meanW,
.keep_all = T) %>%
        # {.}
        # colnames() |> paste(collapse = ", ") |> write_clip()
        mutate(
          targetDist1 = st_distance(geometryPoint_DC,
geometryPoint_meanW, by_element = T) %>% set_units("miles"),
          org = "DC-Shift",
          projfromGeo_DC = st_transform(st_sfc(geometryPoint_DC, crs
= st_crs(5070)), crs = st_crs(4326)),
          projfromGeo_meanW =
st_transform(st_sfc(geometryPoint_meanW, crs = st_crs(5070)), crs =
st_crs(4326)),
          projfromX = st_coordinates(projfromGeo_DC)[,1],
          projfromY = st_coordinates(projfromGeo_DC)[,2],
        ) %>%
        rowwise() %>%
        mutate(
          targetDist = st_distance(projfromGeo_DC,
projfromGeo_meanW, by_element = T) %>% set_units("miles"),
          bearingRad_meanW =
geosphere::bearing(c(st_coordinates(projfromGeo_DC)[1],

st_coordinates(projfromGeo_DC)[2]),

c(st_coordinates(projfromGeo_meanW)[1],

st_coordinates(projfromGeo_meanW)[2])) + 360,
          DC_dist10_yLat =
geosphere::destPoint(c(st_coordinates(projfromGeo_DC)[1],

st_coordinates(projfromGeo_DC)[2]),
```

```
40
41 bearingRad_meanW, 1609.34 * 8)[, 'lat'],
42           DC_dist10_xLong =
43 geosphere::destPoint(c(st_coordinates(projfromGeo_DC)[1],
44
45 st_coordinates(projfromGeo_DC)[2]),
46
47 bearingRad_meanW, 1609.34 * 8)[, 'lon'],
48         ) %>%
49         ungroup() %>%
50         mutate(
51           geometryNewPoint_DC =
52 st_transform(st_sfc(map2_dfc(DC_dist10_xLong, DC_dist10_yLat, ~
53 st_point(c(.x, .y))), crs = 4269), crs = st_crs(5070)),
54           shiftDist = st_distance(geometryPoint_DC,
55 geometryNewPoint_DC, by_element = T) %>% set_units("miles")
56         ) %>%
57         st_set_geometry("geometryNewPoint_DC") %>%
58         # {.}
59         {.} -> dfMapFacsShift
60
61
```

Appendix G

Plot bar chart for division average distance (R Code)

```
      # Summarize by Division Name

      dfFinal <- bind_rows(dfDistFacUAs_segs_lng %>% mutate(v = 0),
                           dfDistFac10_segs_lng %>% mutate(v = 10))

      dfFinal %>%
        filter(!grepl("DCl", FacOrg)) %>%
        left_join(dfGeoStack %>%
select(matches("region|division|aPopDens10|uace10")) %>%
st_drop_geometry(), by = c("orig" = "uace10")) %>%
        left_join(dfMapUAs %>% select(matches("label|uuid")) %>%
st_drop_geometry(), by = c("orig" = "uuid")) %>%
        # {.}
        mutate(
          divisionName = ifelse(v == 10,
paste0(divisionName,"_10mi"), divisionName)
        ) %>%
        group_by(divisionName, FacOrg) %>%
        summarise(
          e2eDistMiles2 = mean(e2eDistMiles) *
(mean(aPopDens10)/1000),
          distMiles2 = mean(distMiles) * (mean(aPopDens10)/1000),
          label = round(distMiles2, digits = 1),
          .groups = "drop"
        ) %>%
        # {.}
        mutate(
         FacOrg = factor(FacOrg, levels=c("FC", "DC", "DS", "DCl")),
        ) %>%
        arrange(desc(e2eDistMiles2)) %>%
      ggplot(aes(x = reorder(divisionName, divisionName), y =
distMiles2, fill = FacOrg)) +
        geom_bar(stat="identity", position = position_stack(reverse
= T)) +
        geom_text(aes(label = label), color = "black", size=3, #x =
key, e2eDistMiles2 = pct, group = FacOrg,
        position = position_stack(vjust = 0.5, reverse = T)) +
        scale_fill_manual(values = c("FC" = "yellow2", "DC" =
"chartreuse2", "DS" = "firebrick1")) +
```

```
        scale_y_continuous(
          breaks = waiver(), n.breaks = 12,
        ) +
        coord_flip() +
        labs(title = paste("End-to-End Distance (Mean) of Urban Area
Divisions"), #, format(dim(ggDim), big.mark = ",")),
             x = "US Census Division", y = "Population Density Miles
(Miles per 1000 People)", fill = "Segment") +
        theme(
            legend.title = element_text(size = 10),
            axis.text = element_text(size = 8),
            axis.title = element_text(size = 9),
            legend.key.size = unit(0.2, 'cm'),
            legend.position = "bottom"
        )

```

References

[1]. Rodrigue, J. P. (2020). The distribution network of Amazon and the footprint of freight digitalization. Journal of transport geography, 88, 102825.

[2]. Alnaggar, A. (2021). Optimization under Uncertainty for E-retail Distribution: From Suppliers to the Last Mile.

[3]. Houde, J. F., Newberry, P., & Seim, K. (2017). Economies of density in e-commerce: A study of amazon's fulfillment center network.

[4]. Lara, C. L., & Grossmann, I. E. (2016). Global optimization for a continuous location-allocation model for centralized and distributed manufacturing. Computer Aided Chemical Engineering, 38, 1009-1014.

[5]. Cranmer, S. J., Leifeld, P., McClurg, S. D., & Rolfe, M. (2017). Navigating the range of statistical tools for inferential network analysis. American Journal of Political Science, 61(1), 237-251.

[6]. Boeing, G. (2018). A multi-scale analysis of 27,000 urban street networks. Environment and Planning

[7]. Akeb, H., Moncef, B., & Durand, B. (2018). Building a collaborative solution in dense urban city settings to enhance parcel delivery: An effective crowd model in Paris. Transportation Research Part E: Logistics and Transportation Review, 119, 223-233.

[8]. Alamsyah, A. V., & Purevdorj, N. (2021). Carbon Efficient Network Design: Evaluating The Trade-Offs Between Carbon Emissions, Transportation Cost and Delivery Time For a Middle-Mile Distribution Network.

[9]. Collins, B., & Wang, H. (2019). Facility Location Optimization for Last-mile Delivery.

[10]. Greening, L. M., Dahan, M., & Erera, A. L. Middle Mile Consolidation Network Design with Fixed Origins and Destinations: Time-Constrained Rate-Based Models.

[11]. Toor, R., & Chana, I. (2021). Network Analysis as a Computational Technique and Its Benefaction for Predictive Analysis of Healthcare Data: A Systematic Review. *Archives of Computational Methods in Engineering*, *28*(3), 1689-1711.

[12]. Miller, H. J. (1996). GIS and geometric representation in facility location problems. *International Journal of Geographical Information Systems*, *10*(7), 791-816.

[13]. Global Ecommerce Forecast 2022. (2022). Retrieved 23 July 2022, from https://www.insiderintelligence.com/content/global-ecommerce-forecast-2022

[14]. Gwinn, D., Helmick, J., Kholgade Banerjee, N., & Banerjee, S. (2018, March). Comparison of Traditional and Constrained Recursive Clustering Approaches for Generating Optimal Census Block Group Clusters. In *International Conference on Geographical Information Systems Theory, Applications and Management* (pp. 28-54). Springer, Cham.

[15]. Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: A comparison of Wal-Mart and Kmart stores in Cincinnati. *Applied Geography*, *32*(2), 591-600.

[16]. Fitzgerald, M., Hansen, D. J., McIntosh, W., & Slade, B. A. (2020). Urban land: Price indices, performance, and leading indicators. *The Journal of Real Estate Finance and Economics*, *60*(3), 396-419.

[17]. Zhen, G. (2021). Optimization of e-commerce logistics distribution network based on highway service areas. *Advances in Transportation Studies*, *53*.

[18]. Schorung, M. (2021). Analysis of the spatial logics of Amazon warehouses following a multiscalar and temporal approach. For a geography of Amazon's logistics system in the United States.

[19]. Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained k-means clustering. *Microsoft Research, Redmond*, *20*(0), 0.

[20]. United States Industrial Outlook | Q2 2022. (2022). Retrieved 4 October 2022, from https://www.us.jll.com/en/trends-and-insights/research/industrial-market-statistics-trends

[21]. Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Sthda.

[22]. Bien, J., & Tibshirani, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, *106*(495), 1075-1084.

[23]. Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

[24]. 2010 Census Urban and Rural Classification and Urban Area Criteria (2022). Retrieved 5 October 2022, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html

[25]. A Grammar of Data Manipulation • dplyr (2022). Retrieved 1 July 2022, from https://dplyr.tidyverse.org/

[26]. Haversine formula – Wikipedia (2022). Retrieved 12 August 2022, from https://en.wikipedia.org/wiki/Haversine_formula

[27]. Ortiz, B., & Sinha, A. (2021). Using Image Transformations to Learn Network Structure.

[28]. Debnath, M., Tripathi, P. K., & Elmasri, R. (2015, September). K-DBSCAN: Identifying spatial clusters with differing density levels. In *2015 International workshop on data mining with industrial applications (DMIA)* (pp. 51-60). IEEE.