Predictive Modeling for Food Accessibility: Machine Learning Insights into Food Deserts and Swamps

Juan David Salazar

Advisor: Emma Hubert

Submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science in Engineering
Department of Operations Research and Financial Engineering
Princeton University

April 2024

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Juan David Salazar

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Juan David Salazar

Abstract

This thesis builds upon previous work that utilizes machine learning with census tract data to predict the percentage of healthful food retailers in a tract. Instead of using a modified Retail Food Environment Index (mRFEI) that refers to the percentage of healthful food retailers in a tract for classifications - this work focuses on taking a more granular approach based on data scraped from online food delivery websites. We process our data using the Nutrient Rich Food Index ($NRF_{9.3}$) to calculate a Food Retailer Composite Score (FRCS) for each retailer, which corresponds to its relative cost and nutritional value. Using FRCS scores and the number of retailers in the vicinity of a tract, we label each tract as either a ”Food Desert”, ”Food Swamp”, or ”Food Oasis” utilizing k-means clustering with $k = 3$. After which we apply Random Forests to predict these labels using public demographic and economic data from the US Census Bureau for each tract. Our model detects food deserts, swamps, and oasis with a prediction accuracy of 79% - showing a reasonable improvement to previous work predicting at 72% accuracy. We find that College No Degree, Poverty Rate, Property Value, and Median income are the most important features for accurate predictions. This work highlights the need for further research on the links between food quality in a region and overall health outcomes in order to better identify locations at risk to target with policy suggestions.

Acknowledgements

To my mother, Gladys Ruiz, my father, Novaro Alberto Salazar. Your love and support has always been my north star. I feel incredibly proud and grateful to be your son.

To my brother, Santiago, whose ability to bring joy to others, and live a life of service inspired my topic selection.

To my other brother, Novaro, who instilled my drive to improve in all things I do from such an early age. Without you, I would not be where I am today.

To my advisor, Emma Hubert, who guided me throughout this process, and to my graduate student advisor, Rajiv Sambharya, whose ongoing feedback helped me refine my methodology and approach.

To my friends, Victor, Paul, and Tai, with who I shared countless hours in the library working and brainstorming ideas.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Chronic Disease Epidemic

According to the Center for Disease Control and Prevention (CDC), six in ten adults in the United States have at least one chronic disease, and four in ten adults have two or more [12]. Chronic diseases are defined as illnesses that last one year or more and require ongoing medical attention. They include cancer, diabetes, heart disease, Alzheimer's disease, and many more.

The CDC cites four main risk factors contributing to the development of these illnesses: tobacco use, poor nutrition, physical inactivity, and excessive alcohol use [12]. Of these factors, three of them could be mitigated with binary decisions (i.e., start to exercise, stop smoking, stop drinking), but nutrition is shrouded in an overwhelming amount of complexity. Access to basic nutritional foods at low cost can often be out of an individual's control. Poor nutrition has been shown to negatively impact our gut microbiome, which has been shown to negatively impact mental health [5, 7]. Poor mental health is correlated with increased rates of excessive drinking and smoking [3, 25]. Therefore, nutrition can be seen as the factor with an outsized impact on chronic disease outcomes.

The National Health and Nutrition Examination Survey by the CDC shows that the number of adults classified as obese rose from 13.4 percent in 1960-1962 to 35 percent in 2022 [11]. According to the World Health Organization (WHO), obesity can be linked to increasing eating of foods “that are high in fat and sugar but low in vitamins” [23]. The CDC also finds evidence linking this epidemic to increased “consumption of food away from home; increased consumption of salty snacks, soft drinks, and pizza; and increased portion sizes.” [21, 9]. Crucially, unequal access to healthy, affordable food seems to be at the core of this issue.

1.2 Food Deserts, Swamps, and Oasis

Studies have shown that the built environment characterized by socioeconomic factors, including the presence, quantity, and accessibility of healthful food retailers, contributes to healthful food consumption [4]. Therefore, Food Deserts (FD) are a useful way to understand and model this problem. By creating a name for a place with unique features that may influence negative public health outcomes, we are better able to understand where problems exist and possibly even how to solve them.

FD have had a variety of different interpretations over the past few decades, but the definition we will be using refers to FD as geographies with limited food retailing options which are on average of very low nutritional quality, and highly convenient.

The literature on FD tends to focus more on proximity to supermarkets than the impacts of also having a large number of unhealthy food options available through fast food restaurants. However, recently many studies have started to analyze a new category - Food Swamps (FS) - to understand the difference between geographies with little food access and those with an overwhelming amount of unhealthy food access.

The issue in developed countries is over-consumption of unhealthy foods [16]. The overwhelming unhealthy food in low-income neighborhoods is a more severe problem than it is in food deserts [26]. Therefore, it is important to have a separate term - ”Food Swamp” - to describe areas where high-calorie and energy-dense foods swamp out healthy options in socially deprived areas [26].

Food Swamps (FS) are defined as areas where unhealthy food “swamps” out healthy food options, meaning although healthy food exists - people will still primarily consume from more convenient and unhealthy food retailers [26, 15, 16].

This also helps us create a differentiation between food insecurity and quality food access - as areas experiencing food insecurity are FD, whereas FS have adequate food, but it is simply mostly unhealthy food. FS residents tend to be of low socioeconomic status and tend to buy energy-dense inexpensive foods, and less fruits and vegetables, leading to an increased risk of obesity [16]. These unhealthy options tend to come from fast-food restaurants and convenience stores.

Food Oasis (FO) are areas that are not FS or FD. These areas have adequate healthy food options available compared to unhealthy options and similarly have a lower risk of obesity for their residents. They are simply locations where healthy food is available and not swamped out by unhealthy options.

This thesis aims to analyze the difference between FD, FS, and FO using primary data scraped from the internet, and filled in using random sampling on a normal distribution. This data will be processed to classify a census tract as either FD, FS, or FO based on natural groupings defined by the k-means clustering algorithm where $k = 3$. Using these natural groupings, I will label every Census tract as either a FD, FS, or FO. Then, using the Random Forests Algorithm, I will analyze how accurately these labels can be predicted using publicly accessible data from the US Census Bureau.

By analyzing the affordability and accessibility of healthy foods compared to unhealthy foods, we can identify geographic gaps in access that may correlate with statistically significant rates of chronic disease. By understanding the similarities these geographies have in terms of socioeconomic and demographic factors, but also quantity and quality of food retailers, we may be able to suggest policy actions that help provide equitable solutions aimed at reducing the rates of chronic disease.

1.3 Quantifying Nutritional Value

1.3.1 Nutrient Density vs Energy Density

The difference between what is considered high or low quality food is generally believed to be the degree of processing of the foods in question, yet this doesn't show the whole picture. Processed foods have existed for centuries - such as olive oil, cheese, and noodles which were all staples of ancient diets. Yet these foods are processed minimally, such as by purely mechanical means in the case of crushing freshly picked olives to create olive oil, or through fermenting milk in the case of cheese. Unlike these ancient staples, many modern foods can be characterized as being ultra-processed, having to undergo extreme heat and chemical treatments which involve various additives and preservatives in order to produce a food which is inexpensive, durable, and easy to mass-produce. These processes tend to denature many of the nutrients in the original ingredients, or create something edible from ingredients which humans would otherwise consider food waste. With this in mind, the NOVA food classification scheme divides foods into ultra-processed, processed, unprocessed, and culinary ingredients [13].

However, NOVA must be coupled with a more standardized and quantifiable system for food classification in order to help us better understand the quality of any given food. This has been done by analyzing different NOVA categories using the Nutrient Rich Food index $\mathrm{NRF}_{9.3}$. This NRF algorithm is represented by the sum of the percentage of the daily values of 9 nutrients to encourage (protein, fiber, vitamin A, vitamin C, vitamin E, calcium, iron, magnesium, and potassium) minus the sum of the percentage of the maximum recommended values for 3 nutrients to limit (saturated fat, added sugar, and sodium) [10].

$$NRF_{9.3} = \left(\sum_{i=1}^{9}\left(\frac{N_i}{DV_i}\right) \times 100\right) - \left(\sum_{i=10}^{13}\left(\frac{N_i}{MRV_i}\right) \times 100\right) \tag{1.1}$$

where:

- N_i: Amount of nutrient i,
- DV_i: Daily Value for nutrient i per 100kcal,
- MRV_i: Maximum Recommended Daily Value for nutrient i per 100kcal.

Using NRF$_{9.3}$ it has been observed that NOVA classified ultra-processed foods tend to be energy-dense (in kcal/g), low-cost (in $/kcal), and nutrient-poor (in NRF$_{9.3}$/kcal) when compared to unprocessed foods [14]. Importantly, the NRF$_{9.3}$ index uncovers nutritional disparities in food offerings that industry classifications overlook, providing a detailed assessment of food quality across different types of retailers. Also note that the units for each nutrient varies, but are standardized as percentages relative to daily value.

For this paper, we will define high quality food options as those which have a higher average nutrient to energy density ratio compared to the rest of the items in the dataset. This be the basis of how we will later break up grocery stores, supermarkets, and convenience stores into different quality tiers using the Food Retailer Composite Score formula defined in later sections.

The application of the NRF$_{9.3}$ index in identifying food deserts offers a more refined approach than the approach taken by [4]. It enables the identification of areas not just lacking in food availability but specifically in high-quality, nutritious options. This approach moves beyond the binary categorization of food retailers, allowing for a detailed and nuanced understanding of food accessibility. Consequently, this can inform more precise and effective interventions, targeting areas with the most significant nutritional deficits.

1.3.2 Limitations of Nutrient Rich Food index

Although NRF$_{9.3}$ is a useful baseline for quantifying food quality, recent studies highlight the link between many nutrients outside of the model which should be encouraged and limited. Many essential vitamins and minerals that are part of the FDA's nutritional guidelines - such as the family of B vitamins, copper, and manganese - are left out of the model.

As for essential nutrients that are not directly part of the FDA's nutritional guidelines, Linoleic Acid (LA) is one of the most relevant to note being left out of our model. This essential nutrient is found in a family of oils called Hydrogenated Oils - commonly referred to as seed oils. These oils - such as soybean, canola, sunflower, and corn oils - are ultra-processed but tend to have relatively high NRF$_{9.3}$ scores. This is primarily because they mostly contain unsaturated fats, and tend to be rich in vitamin E [19]. Yet, the unsaturated fat that they mostly contain is LA, which has been linked to increased cardiovascular risk, insulin resistance, and inflammation of the brain [18, 27]. Other oils such as olive and avocado oils also contain LA, although at much lower rates than seed oils. Natural animal fats such as butter, ghee, and tallow contain the least amount of LA of any of the cooking oils [22, 20].

Hydrogenated Oils tend to be the most inexpensive cooking oils, preferred by most restaurants and consumers alike. Yet natural animal fats and low LA cooking oils tend to be overlooked because of their price premium and low availability in most stores.

Despite these limitations, the NRF$_{9.3}$ index represents a significant advance in the analysis of food deserts. It provides a more comprehensive understanding of food quality than traditional industry classifications, offering insights for policymakers and researchers in addressing complex nutritional challenges.

Chapter 2

Research Scope

2.1 Literature Review

2.1.1 Classifying Census Tracts by Food Access

This thesis is primarily based off research titled “Predicting access to healthful food retailers with machine learning” [4]. In that paper, the authors also used Random Forests in order to classify census tracts as either a Food Desert (FD), Swamp (FS), or Oasis (FO). They trained a model using census data to predict the correct classifications and found their model had a 72 percent accuracy at predicting the correct label.

The limitations with this paper lie in how the authors arbitrarily defined FS and FD. They first introduced the following modified Food Retail Environment (mRFEI) index:

$$\text{mRFEI} = 100 \times \frac{\text{number of healthful retailers in tract}}{\begin{array}{c}(\text{number of healthful retailers in tract} \\ +\text{number of unhealthful retailers in tract})\end{array}} \tag{2.1}$$

They then used this index to classify each tract based on the following rules:

$$\text{mRFEI} = \begin{cases} 0 & \Rightarrow \text{The tract is a food desert} \\ (0, \text{median}(\text{mRFEI})] & \Rightarrow \text{The tract is a food swamp} \\ (\text{median}(\text{mRFEI}), 100] & \Rightarrow \text{The tract has good access to healthful food.} \end{cases} \tag{2.2}$$

This paper defines what is a 'healthful' food retailer using the North American Industry Classification Codes (NAICS). It states that 'Healthful food retailers include supermarkets and other grocery (except convenience) stores (NAICS 445110), warehouse clubs (NAICS 452910), and fruit and vegetable markets (NAICS 445230) within census tracts or half a mile from the tract boundary. Less healthful food retailers refer to fast food restaurants (NAICS 722211), small grocery stores, and convenience stores (NAICS 445120) within census tracts or half a mile from the tract boundary [4].'

Yet in reality, the true boundaries between what is healthy and not are not so straightforward. A Wawa may have more nutritious food than a random local convenience store where everything is packaged and processed - yet they would both be defined as NAICS 445120.

This paper aims to create a more nuanced view on the nutritional score of different stores by applying the $\text{NRF}_{9.3}$ index directly on items stocked in various stores, then weighing this nutritional score with cost in order to account for the concept of "convenience".

Then we will use a two-phased approach which involved a separate labeling and prediction procedure, in contrast to this paper which arbitrarily sets labels using mRFEI values' relation to 0, or the median of the set. We will determine labels using natural groupings on granular nutritional/cost data using k-means, then see how well we can predict these labels using public census data. My methodology aims to build upon the previous work of [4] by addressing its limitations.

2.1.2 Food Desert Research

Recent studies, such as [15], have revealed that the presence of markets in a given area is not the sole determinant of healthy food consumption. Factors like pricing and the type of stores play a pivotal role. Higher-priced stores tend to promote healthier food items, while lower-priced ones often offer unhealthier alternatives. Adding supermarkets to FD areas may not directly impact obesity rates [16]. However, they could influence overall community health, well-being, and economic vitality. This is because individuals struggling with obesity often prioritize cost and convenience over nutritional value [17].

There is a gap in quantitative research that incorporates the cost factor in the analysis of FS [4]. The literature suggests that cost is a central element in differentiating FS from FD.

The relationship between geographic distance, FD, and FS is driven by a combination of price, convenience, cultural nuances, and practical considerations [26]. Accounting for these factors in a meaningful way is an intense undertaking, but the idea is that testing whether the model works when labels are predicted using public demographic and socioeconomic data can help tie these nuances to the labels of FD/FS/FO. The concept of FS offers a clearer understanding of these dynamics [15]. In FS, the abundance of low-cost, unhealthy food options significantly impacts community health, leading to higher obesity rates.

2.1.3 Current Limitations

While numerous studies have shed light on the nature of FD and FS, a comprehensive, nuanced understanding of their intricacies and impacts on community health remains elusive. The groundbreaking work by [4] utilizes machine learning to discern between

FD, FS, and FO, yet its methodology narrowly defines 'healthful' and 'unhealthful' food retailers based on the North American Industry Classification Codes (NAICS). Such a binary classification system overlooks the complexities in food quality and nutrition that vary significantly even within the same NAICS classification.

Furthermore, much of the existing literature emphasizes the placement and mere presence of markets, overlooking the essential role of price in dictating food choices. As highlighted by studies like [15] and [16], price points at higher-end stores tend to promote healthier food options, whereas those at more affordable stores lean towards unhealthier alternatives. The implications of these pricing dynamics are significant; the cost-sensitive populations in FS are driven towards less healthy food choices, exacerbating health disparities. Despite these findings, few studies delve deeply into the cost factor, with [4] notably excluding it from its analysis.

By taking a closer look at the differences between different food retailers on more granular level, the goal is to uncover hidden natural groupings within different regions that can be useful for local policymakers to identify the proper response for a specific problem. By employing machine learning techniques to test the accuracy of predicting these classifications when primary data is not available, we may uncover a more resource-efficient route to offer actionable policy suggestions to a wide range of municipalities.

2.2 Problem Statement

As highlighted in the previous section, a current limitation in the literature is the ability to classify food retailers as healthful or not based on primary (store-level) inventory and pricing data. Most papers only account for availability, not cost of items, even though research shows that cost is an extremely important factor at play. The paper this thesis is based on also failed to account for nutritional value of individual items sold at different retailers when classifying their “healthfulness”, rather relying solely on NAICS classification codes for this assumption [4].

In this thesis, we therefore propose to use sampled inventory data from Instacart and Uber Eats, along with unsupervised learning techniques (k-means clustering) in order to classify census tracts based on the most recent findings from the literature.

Furthermore, we will compare these findings to the results from [4] in terms of how accurately can these census tract labels can be predicted by a Random Forests algorithm using publicly available US Census data.

We will then adjust the original model to see if there are any other natural groupings beyond FD, FS, and FO by applying the Elbow Method and Silhouette Scores to determine the optimal k value from our k-means implementation. We will also observe whether varying the weights placed on cost or nutritional quality, along with varying k, will impact our model's predictive accuracy.

Chapter 3

Methodology

3.1 Data Sources

The following analysis requires datasets where each data point represents a census tract, a geography with an average population of 4000, and each census tract will have a variety of features coming from different sources.

3.1.1 Geographic Data

The primary source of geographic data comes from US Census Bureau TIGER/Line® Shapefiles for the relevant geography [28]. This file provides geographic data for the latitude, longitude of the centroid as well as the geographic shape of the tract - which is crucial for using the Google Maps API. For this thesis I will focus the entire process for one region - Mercer County, NJ - which includes 84 census tracts.

Each tract includes a geometry which represents a polygon shape in terms of longitude and latitude, from which I can use Google Maps API to query food retailers within its vicinity. Google Maps API queries require me to send a Place Type along with my request. I used a specific Place Type for each category in order to get a comprehensive list of different Food Retailers in each tract, as seen in Table 3.1.

Category	Place Type
Grocery Stores, Supermarkets, and Wholesalers	`grocery_or_supermarket`
Convenience Stores	`convenience_store`
Fast Food Restaurants	`meal_takeaway`

Table 3.1: Place Types for Querying Food Retailers by Category

3.1.2 Food Retailer Data

Due to the nature of this thesis, it is unfeasible to have complete data on the inventory and cost of all items in every store within the region of investigation. Instead we have a representative subset of the complete data by scraping data from Instacart.com and Uber Eats using the Web Scraper browser extension [2].

From these sources, the following data is available for every item within the stores listed:

- Item Name
- Price

Table 3.2 contains the list of stores from my queries which were also available to scrape from either Instacart or Uber Eats. This is not a comprehensive view on all food retailers in the area, but enough to let me get an average and standard deviation Food Retailer Composite Score (FRCS) from which to randomly select values for my unknown stores. FRCS will be explained in a following section, it is a composite score we've forumalted that aims to quantify the nutritional quality and cost of a given food retailer.

Category	Observed Stores
Grocery Stores	Target, Wegmans, Shoprite, McCaffrey's, Walmart, ALDI, Stop and Shop, ACME, Costco, BJ's
Convenience Stores	7-Eleven, Wawa, CVS, Walgreens
Fast Food Restaurants	McDonald's, Wendy's, Burger King, Popeyes, Taco Bell, Dominoes

Table 3.2: Observed Food Retailer Stores by Category

Note that for Fast Food Restaurants, most items vary by price depending on the size item you order, for these I defaulted to medium sizes whenever possible. Also, many stores have different quantities or weights available for their produce by default - for the sake of simplicity I went with the default prices without normalizing for weight.

For the two wholesale stores on my list, Costco and BJ's, we will manually normalize costs based on findings discussed in [8]. This article looked at the yearly cost of purchasing at Costco compared to Stop & Shop and finds them to be \$2,194.39 and \$3,253.93 respectively. Because of this, I adjusted my data for Costco and BJ's pricing to be reduced in order to prevent the larger notional cost of items negatively impacting results.

In total, we have collected 8,943 individual food items from 20 total stores across 3 main categories.

3.1.3 Demographic and Socioeconomic Data

The US Census Bureau provides statistics on various statistics at various degrees of granularity. For this thesis I am utilizing the official US Census Bureau API, querying the `acs/ac5` dataset from the year 2022.

This is the data which we used in order to predict the census tract labels using supervised learning in the second part of this thesis. The exact features extracted from the API for each tract can be seen in Table 3.5, along with the rationale for selecting these specific features.

3.2 Data Processing

3.2.1 Associating Stores to Census Tracts

A large bulk of the work on this project has been in associating stores to census tracts. Initially I took a complex approach that followed these steps:

1. Begin with dataframe with rows being census tracts, columns being geometric polygon data.

2. Calculate max radius for each polygon in meters, using the furthest point on the polygon from its centroid.

3. Query google maps API for specific keywords within the max radius for each tract.

4. Prune stores whose latitude/longitude falls outside of the tract's geometry.

This was adequate for initial results, but I quickly realized it might not be optimal for the purpose of trying to classify a location as either a Food Desert, Oasis, or Swamp. Instead I have decided to take an approach which sets a constant hard-coded radius for my queries around each tract's centroid, and does not prune. This means I will have duplicate stores across multiple tracts, but this will better capture the concept of accessibility.

Figure 3.1 shows the distribution of max radius across the tracts in Mercer County, NJ. The main issue here is deciding a query radius that encompasses all tracts.

In small tracts with radius below 1km, it makes sense to query for stores outside of the radius, as people in that tract will reasonably be able to drive further out to find food they want - as shown by findings from [6]. This report from the CDC stated that people are driving on average over 3 miles to get food. Therefore I selected 5,000m as the hard coded value to query stores from. On the other hand, for significantly

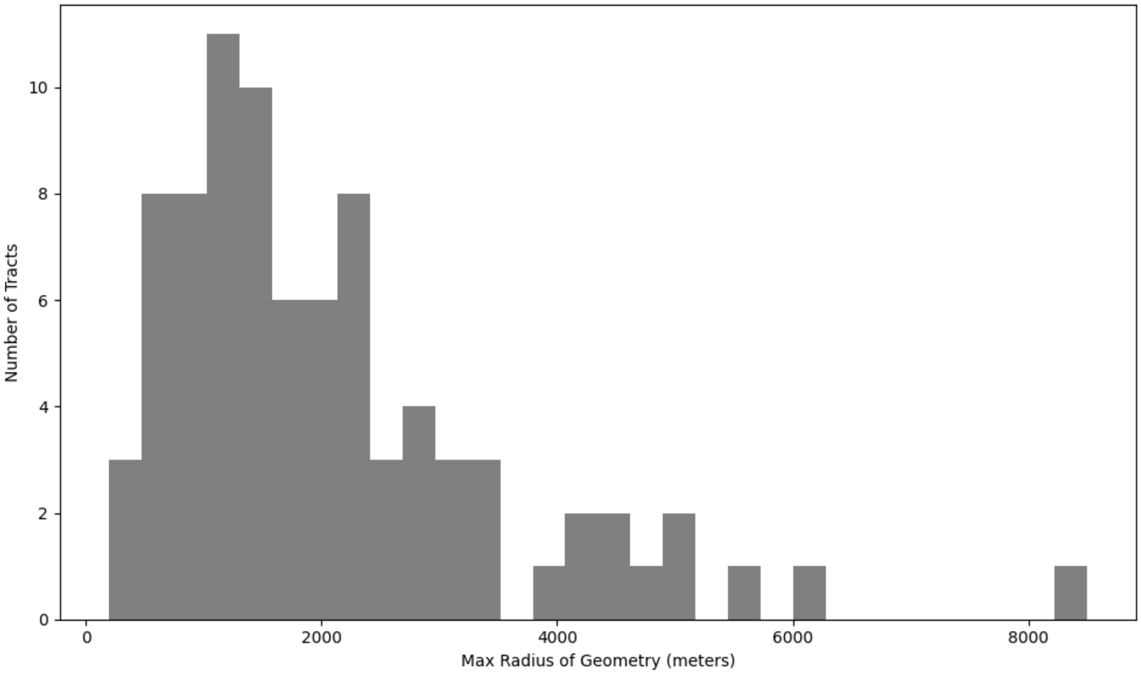


Figure 3.1: Distribution of max radius across tracts in Mercer County, NJ.

larger tracts it does not make sense to query at this constant hard coded value - as it will not encompass the entire radius of the tract itself. So in my code for tracts with a radius over 5km, I instead query using their max radius.

Note that this methodology for querying based on a hard coded radius means that many census tracts in densely populated areas essentially mesh into one entity, as neighboring tracts will have mostly the same stores within the 5km radius. Regardless, it is still a good measurement as these tracts represent one specific municipality, so regardless policy changes can still be targeted based on the tracts' classifications.

3.2.2 Using LLMs to Verify $NRF_{9.3}$ Values

An initial motivation for my methodology and thesis was to come up with a way to quantify the 'healthfulness' of a food retailer in a more precise manner than [4], who simply classified a 'healthful' retailer based on NAICS classifications.

I wrote a script using the USDA Food Data Central database API which allows me to input a string, and outputs an $NRF_{9.3}$ value. This implements formula 1.1, but has a limitation with how raw strings often output various items from the API. For example, querying `Banana` will return `Raw Banana` and `Ripe Banana` from the API, both of which have slightly different nutritional values and $NRF_{9.3}$ scores. My script takes a naive approach and simply returns the average $NRF_{9.3}$ score for all items returned for each query. This script also works with brand name items, as it will simply provide responses which match the most characters from their database to my queries.

Unfortunately, this implementation gave an inaccurate value for many items, likely leading to outliers that make no sense. I tried to prune as many of these as I could manually, but there exists many which may not be extreme outliers and therefore hard to pick out. This resulted in outputs like the ones in the Table 3.3 below.

Product	**Original NRF**	**LLM Score**	**Updated NRF**
Pringles Potato Crisps Chips	72.81	-1	-10
Lunchables Ham & Swiss Cheese with Crackers	70.14	-1	-10
NERDS Candy, Ropes, Rainbow	67.69	-1	-10
Cheetos Wheat Pretzels, Flamin' Hot Flavored	65.24	-1	-10
Takis Blue Heat Rolled Tortilla Chips Bag	65.09	-1	-10
Andy Capp's Hot Fries	61.53	-1	-10
Three Meat Burrito	58.20	0	58.20
Kellogg's Rice Krispies Treats Marshmallow Snack Bar	53.51	-1	-10
Minute Maid Aguas Frescas Mango	-20.54981301	1	10
Naked Juice, Pina Colada	-34.60001688	1	10
Juice Cocktail Blend, Orange Pineapple Flavored	53.16	0	53.16
Citrus Berry Punch Juice	52.73	0	52.73
Naked Juice, Strawberry Banana	49.67	1	49.67
Garlic Parmesan Meatballs	47.50	0	47.50
Takis Fuego Rolled Tortilla Chips Bag	46.49	-1	-10
Hard Boiled Cage Free Eggs	42.45	1	42.45
Chips Ahoy! Original Chocolate Chip Cookies	-56.13	-1	-56.13
Chips Ahoy! Chewy Chocolate Chip Cookies	-56.13	-1	-56.13
Oreo Mini Chocolate Sandwich Cookies	-57.12	-1	-57.12
Brisk Pink Lemonade Flavored Beverage	-63.45	-1	-63.45
Jack Link's Beef Stick, Smoking Hot	-66.84	-1	-66.84
Sour Patch Kids Watermelon Soft & Chewy Candy	-85.06	-1	-85.06
HARIBO Gummi Candy, Share Size	-86.96	-1	-86.96

Table 3.3: Updated $NRF_{9.3}$ Scores for Various Products Based on LLM Prompt

As you can see, it seems almost random for some items whether the script will output a large positive or negative value for $NRF_{9.3}$. The root cause of this issue is that the API being queried does not have many proprietary items in its database, and may only recognize one word from the string which it associates to a healthy item. For example `NERDS Candy, Ropes, Rainbow` gives the responses `Food: Fish, smelt, rainbow, raw` and `Food: Fish, trout, rainbow, farmed, raw` because it only recognizes the term “Rainbow” - and these two items in the responses both have high $NRF_{9.3}$ scores.

In order to attempt to clean this data, I processed the 8,943 food items and their $NRF_{9.3}$ values output from my script through the Large Language Model (LLM) Claude.ai and asked it to return a score, Ω, for each item where:

$$\Omega = \begin{cases} -1 & \Rightarrow \text{The item is potentially unhealthy} \\ 0 & \Rightarrow \text{The item's healthfulness is ambiguous} \\ 1 & \Rightarrow \text{The item is potentially healthy} \end{cases} \tag{3.1}$$

I then ran the following algorithm on my original list of 8,950 food items:

- If $\Omega = -1$ & $NRF_{9.3} > 0$, edit replace $NRF_{9.3}$ with -10,
- If $\Omega = 1$ & $NRF_{9.3} < 0$, edit replace $NRF_{9.3}$ with 10,
- else, leave the $NRF_{9.3}$ value untouched.

I used 10 as a relatively neutral benchmark that would limit the impact on my overall $NRF_{9.3}$ distribution, as the LLM is not aware of the true nutritional value of the item - but is able to improve what otherwise would have been an erroneous outlier. In total this updated 558 items to have $NRF_{9.3} = 10$ and 1,568 items to have $NRF_{9.3} = -10$. Table 3.4 presents the average $NRF_{9.3}$ values across each of the three category of food retailer present in my dataset, before and after LLM verification.

Category	$NRF9.3^{\mu}_{\textbf{before}}$	$NRF9.3^{\mu}_{\textbf{after}}$
Grocery Stores	24.08	18.48
Convenience Stores	19.45	-1.06
Fast Food	3.45	-8.52

Table 3.4: Updated $NRF_{9.3}$ Values by Category

By using a LLM to clean the data from the USDA Food Data Central API, I was able to improve the overall distribution of $NRF_{9.3}$ values in terms of ensuring that

Grocery Stores are more valuable from a nutritional perspective than Convenience Stores - which is critical for accurate clustering when I calculate FRCS scores for each store. My adjustments primarily focused on ensuring that food which is unhealthy is scored as such, rather than it being erroneously scored as healthy, and thereby severely damaging the dataset. Note that some items which were previously scored as unhealthy, but LLM scored as healthy, such as `Minute Maid Aguas Frescas Mango` (which contains a lot of sugar), will continue to exist as outliers. Regardless, the overall dataset is improved.

3.2.3 Food Retailer Composite Scores

The Food Retailer Composite Score (FRCS) approximates the relative value of a store in terms of nutrition and cost by comparing a store to all other stores' average, min, and max $NRF_{9.3}$ and cost values. This means that stores are rewarded when their average $NRF_{9.3}$ value is greater than the overall set of stores' average $NRF_{9.3}$ value, and rewards it more the closer its average value is to the max overall value found in any store. Similarly, this formula penalizes stores whose average cost is higher than the average overall cost across all stores, and penalizes it more the closer it's average cost is to the max cost item across all stores whose data I have scraped. This can be seen in formula 3.2.

$$FRCS_{store} = w_1 \times \left(\frac{\overline{NRF9.3}_{store} - \overline{NRF9.3}_{all}}{P_{90}(NRF9.3) - P_{10}(NRF9.3)} \right) - w_2 \times \left(\frac{\overline{Cost}_{store} - \overline{Cost}_{all}}{P_{95}(Cost) - P_5(Cost)} \right) \tag{3.2}$$

Where:

- w_1 and w_2 are the weights for the average $NRF_{9.3}$ score and average cost, respectively.
- $w_1 + w_2 = 1$

- $\overline{NRF9.3}_{store}$ is the average $NRF_{9.3}$ score from a given store's inventory.
- $\overline{NRF9.3}_{all}$ is the average $NRF_{9.3}$ score across all stores' inventory.
- $P_{90}(NRF9.3)$ is the 90th percentile $NRF_{9.3}$ score from all items across all stores' inventory.
- $P_{10}(NRF9.3)$ is the 10th percentile $NRF_{9.3}$ score from all items across all stores' inventory.
- $\overline{Cost}_{store}$ is the average price for an item from a given store's inventory.
- $\overline{Cost}_{all}$ is the average price across all stores' inventory.
- $P_{95}(Cost)$ is the 95th percentile price from all items across all stores' inventory.
- $P_{5}(Cost)$ is the 5th percentile price from all items across all stores' inventory.

In order to account for greater sensitivity to cost, as most consumers have, we will have a large w_2 value, starting with 0.66. We also adjusted the formula to use percentiles instead of Min/Max in order to minimize the impact of outliers from the scraping and $NRF_{9,3}$ calculations process. We used 95th and 5th percentile for cost data, which had significantly fewer outliers - and 90th and 10th percentile for $NRF_{9.3}$ values which had significantly more outliers.

FRCS will be normalized so it is always in the range [-1,1]. We will then use these scores, along with the number of retailers, to cluster the data using the k-means algorithm in order to determine what is a FD, FS, or FO where ideally:

- Food Deserts have low average FRCS and few food retailers.
- Food Swamps have low average FRCS and many food retailers.
- Food Oasis have a high average FRCS and any amount of food retailers.

3.2.4 Calculating FRCS Scores

Using the methodology previously described we have census data from the US Census Bureau TIGER/Line® Shapefiles associated with the food retailers in each tract, including details such as name and category for each retailer. The next step is to go through and scrape the data from Uber Eats/ Instacart for all of the retailers from this list who are on these sites. Then, for each retailer, I have followed this process:

1. Begin with Name and Price for each item from a store,
2. Clean data to remove non-food items,
3. Calculate $NRF_{9.3}$ for each item using USDA FoodData Central database [1] API script and equation 1.1,
4. Append store data to sheet with all other stores,
5. Verify signs of $NRF_{9.3}$ values with LLM scores,
6. Calculate Average Cost and $NRF_{9.3}$ for store,
7. Calculate FRCS using formula 3.2,

At this point I have FRCS values for each retailer from Table 3.2, which allows me to calculate the mean and standard deviation for the FRCS values in each category. Using the mean, standard deviation, and assumption of normal distribution for each of the three categories - I randomly sample FRCS values for each store I queried using Google Maps API for my region which is not in the list of stores I observed. This allows me to fill in the rest of my data by providing FRCS scores for every retailer associated to every census tract.

With the missing data filled in, I am able to calculate an average FRCS for each tract, which along with the store count for each tract, is what I will be using to cluster tracts into either FD, FS, or FO.

3.3 Model Parameters

3.3.1 Unsupervised Learning: k-means Algorithm

Once the data is cleaned such that each census tract has a store count and an average FRCS value, we run a k-Means algorithm to determine natural groupings for census tracts based on their FRCS score and number of retailers in the tract. In the k-means clustering algorithm, a dataset is partitioned into k distinct clusters by minimizing the variance within each cluster. Each data point is assigned to the cluster with the nearest mean, serving as the cluster's centroid, thereby grouping data points that are more similar to each other into the same cluster while maximizing the dissimilarity between groups. The k represents the predefined number of clusters to be identified in the dataset.

We start by using $k = 3$, but will also later perform a sensitivity analysis to see if other natural groupings may fit the data better. The choice of k-means clustering was driven by its proficiency in identifying hidden patterns and natural groupings.

Clustering Process

Our implementation involved several key steps:

- **Initialization and Iteration**: The k-means algorithm commenced with the random initialization of $k = 3$ centroids. The iterative process of assigning census tracts to the nearest centroid and updating the centroids based on current assignments ensued until the centroids' positions stabilized, signaling convergence. We begin with $k = 3$ to see whether natural groupings between Food Swamps, Deserts, and Oasis are present my data.

- **Cluster Analysis**: We analyze the resultant clusters to interpret and understand the different food retail environments across the census tracts.

- **Determining the Optimal Number of Clusters**: The optimal number of clusters, k, will be established using the Elbow Method and Silhouette Scores in Chapter 4.

I used Python with `numpy`, and the `KMeans` method from the `sklearn.cluster` module in the `scikit-learn` library to implement this process.

3.3.2 Supervised Learning: Random Forests Algorithm

This thesis utilizes Random Forests for predicting the labels because it allows our results to be more directly compared to [4] which is the research we're attempting to expand upon. The Random Forest algorithm is a learning method used for classification and regression tasks, which operates by constructing multiple decision trees during training time and outputting the classification of the individual trees.

Using this algorithm and comparing the relative accuracy will serve as a benchmark for the overall effectiveness of my research. Additionally, this algorithm has inherent strengths in handling complex datasets with potential nonlinear relationships. This method is also effective at avoiding over fitting by averaging the predictions from multiple trees to improve generalizability to unseen data.

Feature Selection The selection of features aims to capture a broad range of socioeconomic and demographic characteristics that influence food accessibility and availability within census tracts. Table 3.5 outlines the features selected for the model and their descriptions. These will be used for training and testing the model. The rationale for this list was to get as close to [4] as possible in the features selected. This will allow for more comparable accuracy results from our predictive models.

Feature	**Description**
Median Income	Median household income in the past 12 months.
Poverty Rate	Percent of people living below the poverty line.
HH With SNAP	Number of households receiving SNAP benefits.
Inequality	Measures income distribution within a population.
Unemployment	Percent of unemployed individuals aged 16 and over.
Below High School	Percent of persons age 25 and over with no high school diploma.
College No Degree	Percent of adults with some college education but no degree.
Bachelors Or More	Percent of persons with a Bachelor's degree or higher.
Property Value	Median value of owner-occupied housing units.
Public Transport	Percent of workers who used public transportation to get to work.
No Vehicle	Percent of occupied housing units with no vehicle available.
Population	Total population of the area.
Black	Percent Black or African American population.
Hispanic	Percent Hispanic or Latino population.
Asian	Percent Asian population.
Native American	Percent American Indian and Alaska Native population.
Pacific Islander	Percent Native Hawaiian and Other Pacific Islander population.

Table 3.5: Descriptions of Selected Features

Implementation The dataset was split into an 70% training set and a 30% testing set. This split was chosen to provide a substantial amount of data for training the model while still reserving a significant portion for validation.

The code was written in Python and makes use of several key libraries. Pandas for data manipulation, `scikit-learn` for machine learning algorithms, and `imbalanced-learn` for addressing class imbalance. To combat the issue of data imbalances across classes, the SMOTE technique was applied from the `imblearn.over_sampling` module, enhancing the representation of minority classes in the training data. It's worth noting that I dynamically sets the `k_neighbors` parameter for SMOTE based on the smallest class size in the training data, ensuring the algorithm operates within the bounds of the available data while avoiding errors related to insufficient samples for nearest neighbor calculations.

I also used Stratified K-Fold Cross-Validation, with K set to 5, to ensure a thorough evaluation of the model's performance across different subsets of the data. Implemented using the `StratifiedKFold` method from the `sklearn.model_selection` module - this method ensures that each fold is representative of the overall dataset's class distribution, in order to provide a more accurate and reliable assessment of the model's predictive capabilities.

3.3.3 Outcome and Relevance to Project Aims

The implementation of k-means clustering aims to classify the census tracts into meaningful clusters based on their food retail characteristics. Each cluster represented a distinct grouping of tracts with similar Average FRCS and Store Counts. Based on the literature we assume $k = 3$ is an adequate representative for the natural groupings in the data, from which we will label each tract as either a FD/FS/FO. This also provides quantitative insights into the literature which discusses the presence of FD/FS/FO but scarcely defines them in a quantitative manner.

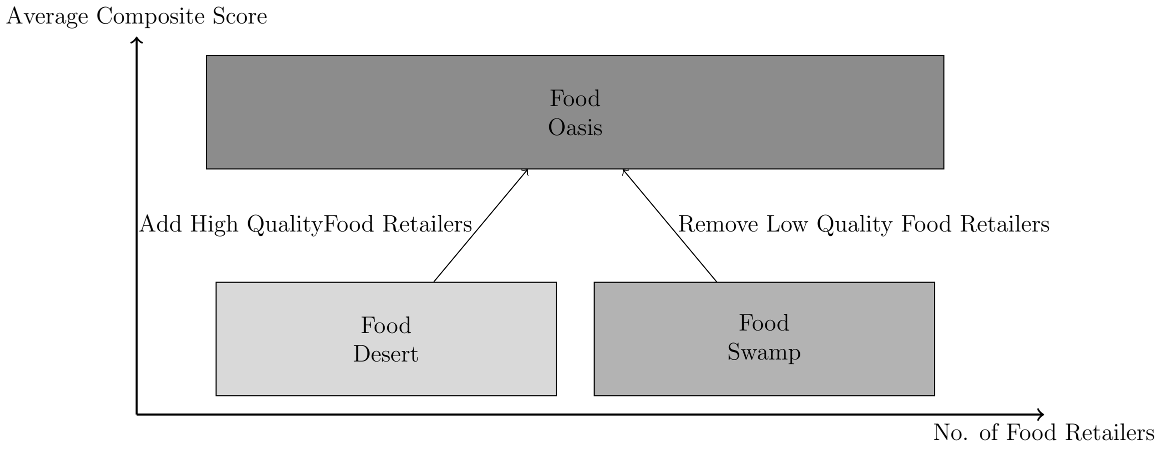


Figure 3.2: Ideal Clustering Geometry for Targeting Policy Solutions.

The reason why we want to differentiate between FD and FS is that they have inherently different, mutually exclusive policy solutions, as visualized in Figure 3.2. A FD can not be fixed by removing low quality food retailers - as they are lacking high quality food retailers by definition. On the other hand a FS is defined as too much low quality, low cost food 'swamping' out the healthy food - so adding high quality food retailers fails to solve the problem.

It is possible that a different number of clusters results in a more optimal solution for k-means, or the geometry of clusters does not match this hypothetical Figure. If this is the case, I will not label each cluster as either FD/FO/FS based on the geometry in Figure 3.2 but rather based on the socioeconomic and demographic factors associated with each cluster. Alternate clustering geometry would also suggest changes to optimal policy solutions.

Later on I will label the tracts based on those optimal groupings and try to come up with natural naming conventions for the new groupings within the context of FD/FS literature if k is not 3. I will also touch on what this means for policy solutions.

Chapter 4

Results

4.1 Census Tract Classifications

4.1.1 Processed Primary Data

Table 4.1 displays a summary of the processed data for each category. We observed about 30 percent of the data, while the remainder was sampled based on a normal distribution assumption. We set $w_1 = 0.66$, and $w_2 = 0.34$ to get these results.

Category	**Observed Data %**	$NRF_{9.3}^{\mu}$	$Cost^{\mu}$	$FRCS^{\mu}$	$FRCS^{\sigma}$
Grocery Stores	28.44%	18.48	5.64	0.0158	0.0294
Convenience Stores	43.58%	-1.06	5.19	-0.0999	0.0690
Fast Food	20.74%	3.45	-8.52	-0.2173	0.0637

Table 4.1: Summary of Store Data by Category

With about 30% of the overall data manually scraped from Instacart and Uber Eats, we have a representative sample for the wider food retail environment. This represents a significantly more granular view on access to healthy food than [4]. In Table 4.2 we can see a more detailed breakdown of the data from Table 4.1. Note that Appearances are not exactly store counts because my methodology double counts stores for neighboring census tracts. For example if a store is within 5km of 10 tracts, that counts as 10 data points for that specific store. From the table can see that Fast Food has a much lower nutritional quality compared to the Convenience Stores, which have a much lower nutritional quality compared to Grocery Stores - as expected.

Category	Store	Items	Appearances	$NRF^{\mu}_{9.3}$	$Cost^{\mu}$	FRCS
Grocery Store	McCaffreys	524	16	20.28	6.70	-0.0065
Grocery Store	ACME	717	22	16.88	5.49	0.0100
Grocery Store	ALDI	663	75	19.28	3.71	0.0830
Grocery Store	BJs	590	12	18.89	6.20	0.0003
Grocery Store	Costco	471	12	25.15	6.60	0.0290
Grocery Store	ShopRite	1008	85	12.19	5.20	-0.0121
Grocery Store	Stop & Shop	718	12	20.30	6.50	0.0002
Grocery Store	Target	810	16	13.78	5.43	-0.0090
Grocery Store	Walmart	1136	58	15.38	4.20	0.0415
Grocery Store	Wegmans	1377	8	22.72	6.32	0.0220
Convenience Store	Wawa	56	103	9.81	4.76	-0.0137
Convenience Store	CVS	225	201	-7.26	6.33	-0.1779
Convenience Store	7/11	122	232	-8.82	4.31	-0.1232
Convenience Store	Walgreens	207	57	2.03	5.36	-0.0847
Fast Food	Domino's	47	103	-7.40	10.98	-0.3286
Fast Food	Taco Bell	94	43	-5.23	7.27	-0.1946
Fast Food	McDonalds	98	75	-11.17	6.47	-0.2082
Fast Food	Wendys	60	75	-6.96	6.19	-0.1713
Fast Food	Burger King	41	73	-11.85	5.56	-0.1835

Table 4.2: Updated Observed Data NRF, Cost, and FRCS Outputs with Instances

By taking a closer look at the data, Figure 4.1 shows us how Grocery Stores tend to have similar FRCS scores, with ALDI being the clear outlier in terms of high quality food at a low cost. On the other hand, Fast Food and Convenience Stores all score extremely low - this can be attributed to their extremely high item costs. Notably, Wawa scores the best among Convenience Stores by far, while Walgreens trails behind. Another factor to consider is the large amount of data gathered from Grocery Stores compared to the other categories, as well as the potential erroneous outputs that our $NRF_{9.3}$ calculations may still result in for some values. Generally, though, the plot seems accurate to what one would expect.

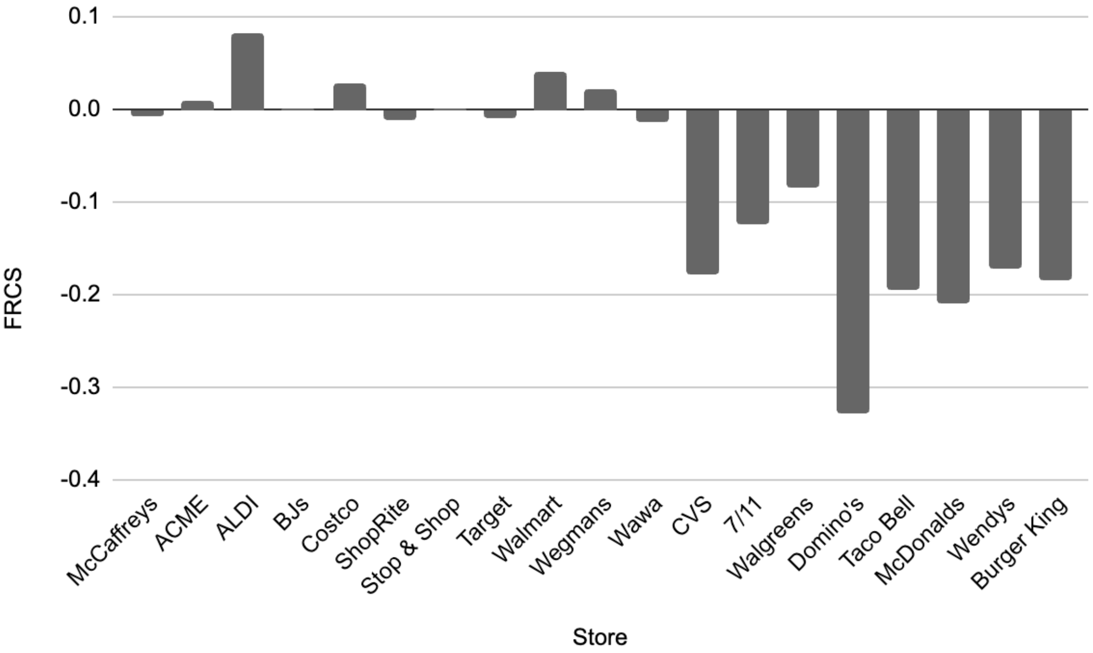


Figure 4.1: FRCS From Observed Stores

4.1.2 Clustering Results

The clustering results for $k = 3$, $\Theta = 0.66$ can be seen in Figure 4.2a which clusters the stores into three groups, but visually it's clear that the clusters don't match the hypothesized ones previously referenced in Figure 3.2. Instead, there is a linear relationship between Store Count and Average FRCS across census tracts, with a density of tracts that have 60 stores representing the Trenton area.

K-means does not have any way of determining which cluster corresponds to either Food Desert, Swamp, or Oasis. Based on our definitions, the clustering features we selected - FRCS and Store Count - are meant to cluster Food Deserts as the lowest of both features, and Food Oasis with the largest values of both features. We will base our labeling off this assumption, even though if we analyze the demographic & socioeconomic factors of each cluster in Table 4.3, it conflicts with the notion that tracts with better economic conditions are likely to be Food Oasis, as these tracts have price insensitive consumers who tend to buy healthier food [17].

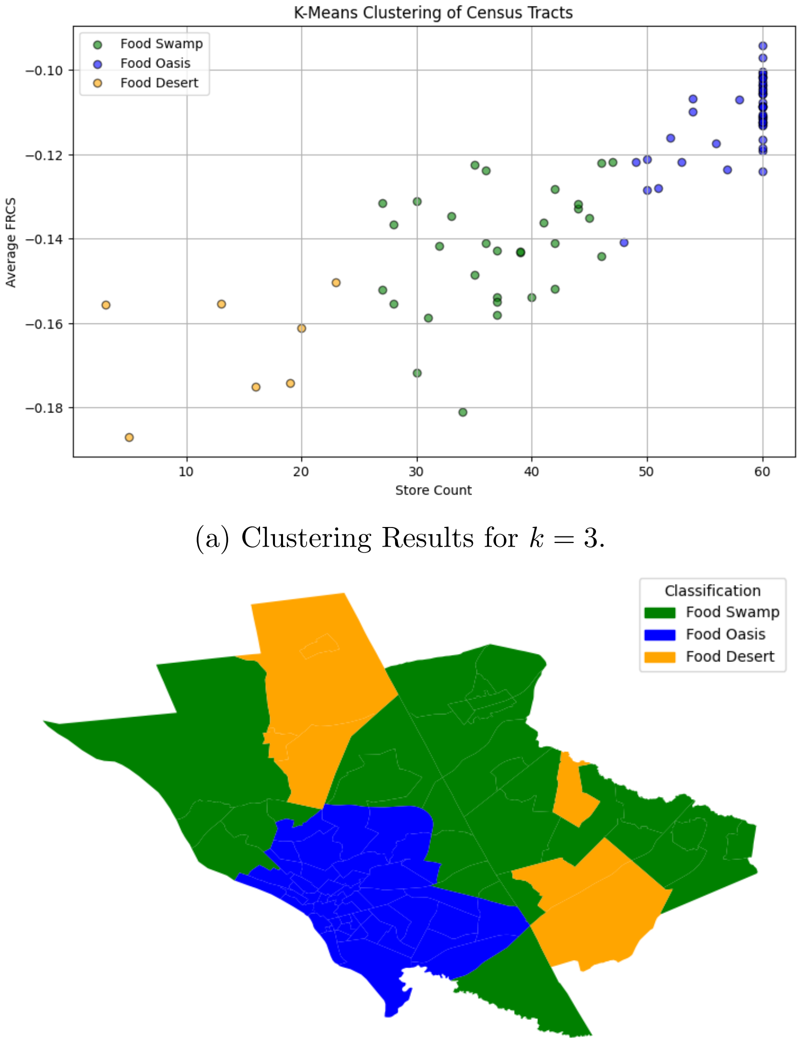
Feature	Food Desert	Food Oasis	Food Swamp	All
College No Degree	50% (6%)	16% (10%)	45% (13%)	29% (18%)
Poverty Rate	2% (1%)	18% (13%)	6% (6%)	12% (12%)
Property Value	$615,886 ($116,457)	$185,145 ($94,111)	$593,194 ($366,186)	$371,629 ($309,413)
Median Income	$194,810 ($41,090)	$69,079 ($28,707)	$143,965 ($48,356)	$107,193 ($57,719)
Bachelors Or More	23% (5%)	11% (6%)	22% (4%)	16% (8%)
No Vehicle	0% (0%)	1% (1%)	1% (1%)	1% (1%)
Unemployment	1% (0%)	4% (3%)	3% (1%)	3% (2%)
Hispanic	4% (2%)	27% (22%)	10% (9%)	19% (19%)
Black	6% (5%)	37% (28%)	7% (6%)	24% (26%)
Asian	22% (20%)	4% (3%)	21% (18%)	12% (15%)
HH With SNAP	1,295 (474)	1,647 (603)	1,818 (717)	1,681 (644)
Native American	0% (0.01%)	0% (0.01%)	0% (0.01%)	0% (0.01%)
Population	3712 (1453)	4435 (1742)	4959 (1865)	4568 (1774)
Inequality	0.40 (0.04)	0.45 (0.09)	0.42 (0.07)	0.43 (0.08)
Below High School	0% (0%)	2% (2%)	0% (0%)	1% (2%)
Public Transport	3% (3%)	2% (2%)	3% (2%)	3% (2%)
Pacific Islander	0%	0%	0%	0%

Table 4.3: Summary Statistics and Labels of Clusters, Mean (Standard Deviation)

The census tracts which all have 60 Store Count represent the smaller census tracts in the more densely populated Trenton area. This is visualized in Figure 4.2b which displays the map of FD/FS/FO across Mercer County NJ.

In Table 4.3 we can see that Food Oasis align with significantly worse economic factors, such as lower Property Value, Median Income, and higher Poverty Rates. Interesting, Food Deserts have better economic factors compared to Food Swamps. We can also see that Food Deserts have the least amount of demographic diversity, while Food Oasis have the most. All of which conflicts with our preconceived notions of what should be classified as a FD or FO respectively.

In Figure 4.2b we can see that the majority of the county is classified as a Food Swamp, which makes sense as in most of the county we mainly have fast food retailers littered among minimal amounts of healthy food retailers. The Trenton area seems to be a Food Oasis based on the FRCS scores and number of stores in the area - which is the opposite of what economic and demographic factors suggest.



(a) Clustering Results for $k = 3$.

(b) Labeled Census Tracts in Mercer County, NJ.

Figure 4.2: Clustering Results

4.2 Predicting Labels with Public Census Data

The decision to utilize the Random Forests algorithm for predicting whether a census tract is a Food Desert (FD), Food Swamp (FS), or Food Oasis (FO) was primarily based on providing a direct comparison to the results from [4].

4.2.1 Predictive Model Performance

The model demonstrated an overall accuracy of 79%, indicating a substantial ability to predict the correct food environment classification based on the census data. Interestingly, the paper that inspired this thesis reported a 72% accuracy [4]. Note that I took an alternative approach to my prediction methodology, but my slightly unreasonable labeling won't impact the algorithm's predictive power. If the clustering methodology is refined, the model's capacity to make predictions that impact policy solutions would increase.

Meaning that labels themselves would be correlated to the potential for averse public health outcomes due to the food retail environment - and more applicable for targeting policy solutions - independently of whether we're able to predict those labels in a more accurate manner or not.

Regardless, the performance for my model can be seen in Table 4.4 below. These values were predicted using Random Forests, with SMOTE applied for addressing imbalances in my dataset, and Stratified K-Fold Cross-Validation, with $K = 5$ to ensure that each fold used in model training and validation has a representative distribution of all classes. The 79% accuracy represents the average across all 5 folds.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Food Desert	0.50	0.29	0.36	7
Food Swamp	0.69	0.77	0.73	46
Food Oasis	0.89	0.87	0.88	31
Accuracy		0.79		
Macro Avg	0.69	0.64	0.66	84
Weighted Avg	0.78	0.79	0.78	84

Table 4.4: Classification Report for All Folds

While the model is highly precise and reliable in identifying Food Oasis, it struggles with Food Desert predictions, and is moderately useful for Food Swamps.

4.2.2 Feature Importance

Table 4.5 provides insight into which factors impacted the classifications from my model the most.

Feature	**Importance**
College No Degree	13.49%
Poverty Rate	13.40%
Property Value	12.79%
Median Income	12.64%
Bachelors Or More	7.15%
No Vehicle	6.64%
Unemployment	5.59%
Hispanic	4.88%
Black	4.05%
Asian	3.35%
HH With SNAP	3.25%
Native American	2.99%
Population	2.97%
Inequality	2.61%
Below High School	2.10%
Public Transport	1.65%
Pacific Islander	0.47%

Table 4.5: Average Feature Importance across folds in Random Forest Model

College No Degree, Poverty Rate, Property Value, and Median Income were the most important predictors. This finding aligns with the prevailing understanding that economic constraints significantly impact access to healthy food options. It also highlights the large impact of educational levels.

4.3 Sensitivity Analysis

In this section we will see how the previous results are impacted by varying the key parameters of my model, in order to determine if we can further optimize our results by adjusting key parameters.

4.3.1 Varying w_1 & w_2 in FRCS Calculations

Recall the formula for FRCS:

$$FRCS_{store} = w_1 \times \left(\frac{\overline{NRF9.3}_{store} - \overline{NRF9.3}_{all}}{P_{90}(NRF9.3) - P_{10}(NRF9.3)} \right) - w_2 \times \left(\frac{\overline{Cost}_{store} - \overline{Cost}_{all}}{P_{95}(Cost) - P_5(Cost)} \right) \tag{4.1}$$

Where $w_1 + w_2 = 1$.

For my initial results I assumed $w_1 = 0.66$ and $w_2 = 0.34$. This is because the key factor that influences health outcomes, and therefore a key factor for determining classifications, is the nutritional content of food. On the other hand, research has shown that cost plays a significant role in the phenomena observed that even when high nutrition food is available, lower cost unhealthy food will be preferred. In order to account for this I wanted FRCS to reward higher nutritional quality of food in a store, but still punish high cost. Since cost has not been taken into account for any of the previous quantitative research in the FD literature, it seemed like an appropriate assumption to give high nutrition twice the weight of high cost as a benchmark.

In order to begin to understand the impact of varying the weight parameters we can analyze Figure 4.3. This is a plot that shows the average FRCS scores of a given category as a function of Θ where:

$$\Theta = \frac{w_1}{w_1 + w_2}, \quad \Theta \in [0, 1] \tag{4.2}$$

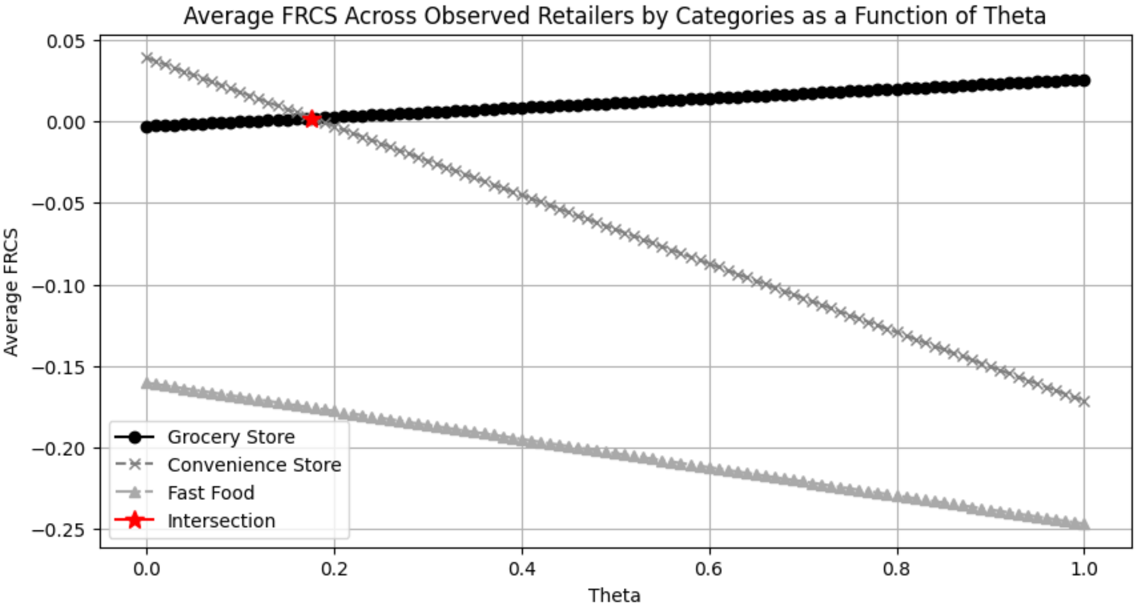


Figure 4.3: $FRCS^{\mu}$ Across Observed Retailers by Categories as a Function of Θ.

There are a few interesting observations from this plot:

- Grocery and Convenience Stores' $FRCS^{\mu}$ values intersect at $\Theta = 0.18$, where $FRCS^{\mu} = 0$,
- Both $FRCS^{\mu}$ for Convenience Stores ($FRCS^{\mu}_{GS}$) and $FRCS^{\mu}$ for Fast Food Stores ($FRCS^{\mu}_{CS}$) **decrease** as a function of Θ,
- $FRCS^{\mu}$ for Grocery Stores **increases** as a function of Θ.

$\Theta = 0.18$ indicates the weight at which the trade off between cost and nutritional quality is at equilibrium between Grocery and Convenience Stores - leading to indifference in our how our model values these types of stores appearing in different tracts.

For all $\Theta < 0.18$ FRCS scores indicate more value is given to Convenience Stores over Grocery Stores, while $\Theta > 0.18$ indicates the opposite. As you move to a Θ further from this value, the difference in FRCS scores between the two categories increases.

It is expected to see Convenience and Fast Food Stores both decrease as a function of Θ, because one would assume that as you put a greater emphasis on nutrition - the less-nutritionally-rich retailers would get lower scores.

Similarly, as you increase your emphasis on nutrition, the more nutritionally rich Grocery Stores increase their value to the model.

The fact that the average values for Fast Food restaurants does not intersect with the other lines indicates that there is no value at which the trade off between cost and nutritional quality is at equilibrium with any other category. For all Θ values, Fast Food stores are inferior in terms of cost and nutrition. This discrepancy can be attributed to the fact that my data was biased toward higher cost Fast Food. Locations like Domino's had extremely low FRCS scores because their items, on average, cost a lot more than other stores individual items - simply because they are selling larger amounts of food per item.

The extreme values of this figure are further explained in Table 4.6 below:

Θ	$FRCS^{\mu}_{\mathbf{GS}}$	$FRCS^{\mu}_{\mathbf{CS}}$	**Description**
0	-0.0031	0.0390	Only punish high cost
0.18	0.0020	0.0020	Equal FRCS between GS and CS
0.66	0.0158	-0.0999	Original Θ for experiment
1	0.0256	-0.1714	Only reward high $NRF_{9.3}$

Table 4.6: Noteworthy Average FRCS for Grocery and Convenience Stores.

In order to get well defined clusters that correspond to FD, FS, and FO we want to have a significant difference between these categories' FRCS values while still accounting for cost, therefore 0.66 is an adequate value for Θ.

4.3.2 Determinining Optimal Number of Clusters

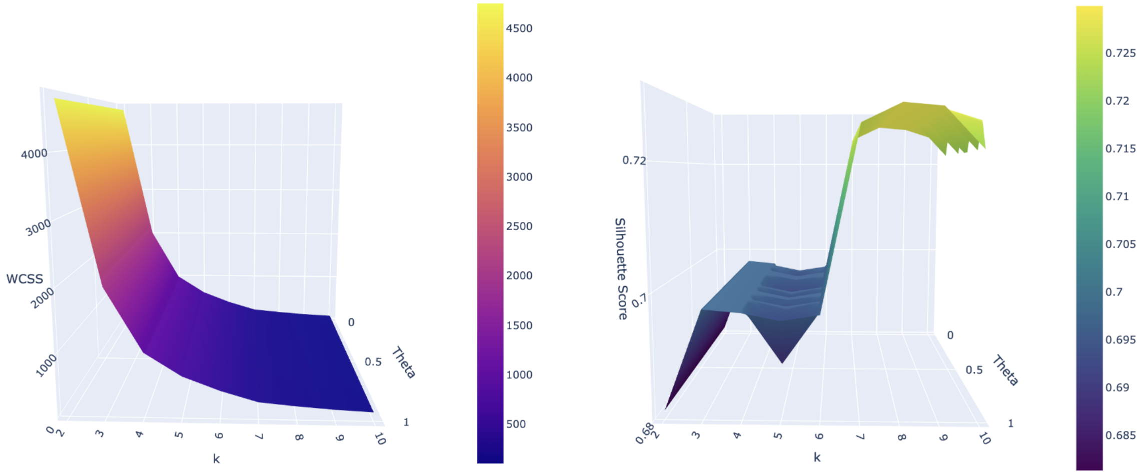
To determine the optimal number of clusters, k, two methods were utilized: the Elbow Method and the Silhouette Score.

The Elbow Method identifies an inflection point in the Within-Cluster-Sum-of-Squares (WCSS) plots [24]. The 'elbow' refers to finding the value at which the rate of change for WCSS decreases by the largest amount. WCSS is defined as the sum of the squared distance between each member of the cluster and its centroid. Ideally, one should select a number of clusters so that adding another cluster does not significantly improve the total WCSS. Beyond this point, the meaningfulness of the individual clusters decreases because the sum of the squared deviations only changes slightly. Figure 4.4a shows the resulting graph for WCSS values as we vary k and Θ. I calculated WCSS values using the `kmeans.inertia_` value output from the `sklearn.cluster` module of the `scikit-learn` library.

The Silhouette Score measures the consistency within clusters [24]. It ranges from -1 to +1, where a high value indicates that the data points are are well matched to their own cluster and distinct from neighboring clusters. It calculates the silhouette score for each data point by measuring how similar the point is to its own cluster (cohesion) compared to other clusters (separation). Figure 4.4b shows the resulting graph for the Silhouette Score when we vary Θ and k. I used the `silhouette_score` function from the `sklearn.metrics` module of the `scikit-learn` library to calculate the values. Like The Elbow Method, Silhouette Scores are not the only determinant of an optimal k, and they are best used in combination with one another.

Instantly it can be observed that in both datasets, Theta Θ has negligible impact on the optimal k.

In Figure 4.4b it can be observed that the Silhouette Scores suggests an optimal value at $k = 8$. This means that when $k = 8$ we have the most distinct clusters. However, in Figure 4.4a when observing WCSS across Θ and k there is a pronounced



(a) WCSS across Θ and k

(b) Silhouette Scores across Θ and k

Figure 4.4: Determining Optimal k With WCSS and Silhouette Scores

'elbow' at both $k = 3$ and $k = 4$. This suggests that beyond this point, the meaningfulness of adding more clusters diminishes.

Going back to Figure 4.4b we can also observe another local maxima at $k = 3$, indicating $k = 3$ is a better choice than $k = 4$, while $k > 4$ are all poor choices. Therefore, considering both metrics together, $k = 3$ is chosen as the optimal number of clusters.

Although the Silhouette Scores suggest that more clusters could be meaningful, the Elbow Method clearly indicates that increasing the number of clusters beyond three yields negligible improvement in WCSS, implying that three clusters are sufficient to capture the significant structure of the data. This result is consistent with the underlying assumption that beyond Food Oasis, Food Swamps are important to differentiate from Food Deserts - validating our choice of $k = 3$. In order to further validate our parameter selection of $\Theta = 0.66$ and $k = 3$, we will need to investigate whether another set of parameters allows us to optimize our Random Forest model's predictive power.

4.3.3 Checking Parameter Impact on Prediction Accuracy

In the previous section we determined that the optimal k for k-means is likely 3 for my data, regardless of our Θ parameter. This indicates that the breakup between Food Deserts, Swamps, and Oasis is likely to be valuable in terms of providing the theoretical policy suggestions depicted in Figure 3.2 across three categories.

Now we want to determine whether there exists a relationship between k, Θ (Theta), and Accuracy of our Random Forests model. We utilized a modified version of the script used to generate the Elbow Method and Silhouette Score graphs to calculate the predictive accuracy of our model for all k, Θ pairs. This was done using the same methodology we used for ($\Theta = 0.66$, $k = 3$) with Random Forests, SMOTE, and Stratified K-Fold Cross-Validation.

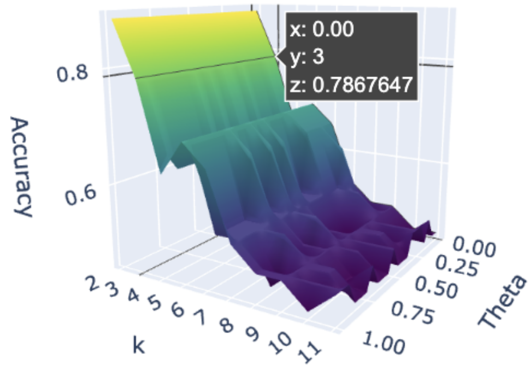


Figure 4.5: Accuracy across Θ and k

In Figure 4.5 it can be observed that when $k = 2$ predictions are approaching 96% accuracy regardless of Θ, and at $k = 3$ all values for model accuracy are indeed 79%. We also see a slight jump up at $k = 5$ where accuracy hits 69%. For all values when $k > 3$, Accuracy is below 79%. Therefore, the only way to increase the predictive accuracy is to use $k = 2$, which we showed in the previous section is sub-optimal when you take into account both the Elbow Method and Silhouette Scores.

Chapter 5

Discussion

5.1 Summary of Findings

By scraping primary data for the cost and nutritional quality of every item in a representative subset of stores across various categories, we were able to achieve results that are comparable in predictive power to [4] with far fewer assumptions on the nature of healthy food. We were able to achieve a prediction accuracy of 79% on our dataset of 84 census tracts in Mercer County, NJ. The dataset included about 3,000 appearances of stores, where about 1,000 appearances come from stores which we scraped data for. In total we scraped 8,943 food items from 20 stores in 3 categories. Predictions were made on labels determined through k-means clustering with $k = 3$, and predicted using various demographic variables from each tract.

My initial experiment yielded promising results, identifying Food Oasis with high precision. However, the model's performance in differentiating Food Deserts from Food Swamps points to the general challenge of capturing the complex factors contributing to food accessibility using a limited set of features and data points. Although our predictive accuracy was high, our clusters did not match the geometry we expected with respect to FD/FS/FO relationships to Store Counts and FRCS scores.

5.2 Research Limitations

5.2.1 Limitations in Calculating $NRF_{9.3}$

The anecdote used to motivate this work is that most people who have been to the stores will agree that the nutritional quality you find at Wawa is on average significantly higher than 7/11. Looking at Table 4.2 we can see this validated by the data, where Wawa has $NRF_{9.3}^{\mu}$ of 29.34 while 7/11 is at 7.90.

Regardless of whether these high level benchmarks make sense for how $NRF_{9.3}$ is calculated across different stores, we had to make a lot of adjustments to our $NRF_{9.3}$ scores which diverged from the initial goal of calculating the scores strictly based on the nutritional facts of each item. For items which did not have appropriate nutritional facts available in the Food Data Central API, we used LLM prompts to pick out possible outliers and adjust their values to something more reasonable. This improved the overall quality of our dataset, but it is still far from perfect.

5.2.2 Limitations of Clustering to Determine Labels

The Food Retailer Composite Score (FRCS) is central to this thesis, as it is the mathematical formula that allows us to quantify nutritional quality and cost in order to give a store a single score that will help classify the census tract based on geometry depicted by Figure 3.2. Food Desert literature indicates that the concept of convenience, which is primarily characterized by cost, has a significant impact on consumption patterns and therefore health outcomes.

What we observed instead in Figure 4.2a is that there is actually a linear relationship between clusters with respect to Average FRCS and Store Count.

As a reminder of our definitions:

- FD are geographies with limited food retailing options which are on average of very low nutritional quality, and highly convenient,

- FS are defined as areas where unhealthy food “swamps” out healthy food options, meaning although healthy food exists - people will still primarily consume from more convenient and unhealthy food retailers,

- FO are simply locations where healthy food is available and not swamped out by unhealthy options.

We expect that tracts with low FRCS and low Store Counts correspond to Food Deserts, while tracts with high FRCS and high Store Counts correspond to Food Oasis. We also expect tracts classified as Food Deserts to include populations that are in more dire economic conditions, whereas Food Oasis are expected to have populations that are thriving economically [17].

What we see from our clustering results is the opposite of this - although Food Deserts have low Store Counts and Average FRCS scores, they have the highest Median Income, lowest Poverty Rate, and highest Property Value of all the clusters. This suggests that there is a more effective manner of clustering the tracts based on FRCS or a similar metric in order to ensure FD match our definition, but also our economic expectations.

At its core FRCS is an imperfect score because it aims to find a balance between what people do vs what people perceive as best for their health outcomes. When looking at this with the goal of classifying geographies in order to target positive public health outcomes - it is extremely difficult to quantify the true balance between these two factors. People act based on convenience, which is largely driven by cost sensitivity - but people want to avoid unhealthy food in general. If Food Swamps exist then even when healthy options exist, the scoring metric should be low.

We adjusted the FRCS formula to use percentiles in order to remove the noise from outliers, and that provided more clean results but failed to impact the core dilemma of not matching the ideal geometry for clusters. With Average FRCS as a scoring metric, where tracts include stores from a 5km radius, we can see that no tracts have low Average FRCS scores with high Store Counts. This suggests that the tracts in the Trenton area multiplied the impact on the Average FRCS score of multiple neighboring tracts by counting the same handful of higher quality food retailers multiple times, which artificially increases the Average FRCS. Another important factor is the role cost plays in normalizing FRCS scores across stores which have vastly different nutritional profiles - even with $\Theta = 0.66$ there are still cases where a stores marginally lower cost increases its FRCS score above another store with significantly higher $NRF_{9.3}$ values.

FRCS may not be an optimal metric for capturing the underlying core issues with respect to public health outcomes, but it invites us to understand how to score different food retailers based on the quality and cost of food available there. This process highlighted how using a metric like FRCS is more accurate at the store-level than simply using NAICS classification, as in [4].

5.3 Areas for Future Study

When expanding on this work, there should be a larger emphasis on linking health outcomes to the way we determine true classifications. The classifications they generate should have a high alignment with the types of public health outcomes in each census tract. This would involve a greater investigation into which kinds of health outcomes are more prevalent in locations that can be characterized as Food Swamps or Food Deserts. This will also require accessing additional datasets from public health sources.

The approach of scraping data from representative samples for each food retailer category proved to be valuable, especially when coupled with LLMs for performing sanity checks. This could be expanded to include additional sanity checks, along with a more rigorous validation of the impacts LLMs have on the overall dataset.

The general relationship between Θ and average FRCS scores by category matches the relationships we would expect - in particular giving significantly greater scores to Grocery Stores as Θ increases. Yet how we use FRCS to cluster can be improved to better match our expected clustering results.

If this can be done, classifications should be more accurate in terms of the goals they're trying to achieve. If labels are validated in terms of public health outcomes, then the rest of the procedure can be followed similarly.

Clustering algorithms do not automatically place natural language labels on clusters. The clusters from my results, though, were not arbitrary. Instead they followed the definitions set in the introduction and methodology of this thesis, and in the end that resulted in labels which disagreed with the prevailing findings of the rest of the literature in terms of socioeconomic and demographic makeups of areas considered FD, FS, and FO. This faces us with the choice of selecting the labels as per the original definitions, or based on the demographic and socioeconomic expectations for each label. Ultimately we defined the clustering problem under specific definitions, so those are the ones we went with. In future work, a key point should be to ensure that clustering results match **both** the definitions of the model **as well as** the expected results in terms of demographic and socioeconomic factors explained in the wider literature. This would represent an ideal result that is applicable in the real-world.

Along with this, future work may also include Random forests, with SMOTE technique for imbalances, and Stratified K-Fold Cross-Validation as appropriate choices for creating a predictive model. It should also compare predictive results across multiple models to re-validate these findings for the new data. Finally, future work can expand the procedure outlined here, from scraping to predictions, by including a wider range of geographies on which to perform clustering and label predictions.

Chapter 6

Conclusion

This study builds upon previous work employing machine learning in the food retail environment. Instead of using an arbitrary index of healthful food retailers - we used primary data on the cost and nutritional quality of specific food items from a subset of food retailers. This is the the first study in the FD literature to incorporate data scraped from public sources for model training, the first to incorporate pricing data, as well as the first to utilize LLMs for automating the process of performing sanity checks on large amounts of nutritional scores for individual items.

Our data was processed through a Food Retailer Composite Score (FRCS) given to each retailer, which is based on the weight placed on nutritional quality vs cost. Results suggest FD are characterized by high income, highly educated populations, whereas FO include lower income, more diverse populations. FS are between, but much more similar to FD. Although our clusters failed to capture classifications as expected, we have shown that the optimal number of clusters for our dataset is in fact 3. This indicates that a third type of food retail environment is likely to exist.

By adjusting our features for clustering such that labels align with both our definitions, as well as the literature's findings, our model could be improved to offer valuable policy suggestions aimed at improving public health outcomes.

Appendix A

Code & Data

The code and dataset for this thesis can be found in the following public GitHub repository: https://github.com/jsalazar-1207/ORFE-Senior-Thesis

Bibliography

[1] Fooddata central. https://fdc.nal.usda.gov/.

[2] Web scraper - free web scraping. https://www.webscraper.io/, 2023. Accessed: 2023-04-12.

[3] Alcohol and mental health. *Mental Health Foundation*, n.d.

[4] M. D. Amin, S. Badruddoza, and J. J. McCluskey. Predicting access to healthful food retailers with machine learning. *Food Policy*, 99:101985, Feb 2021.

[5] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi. The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems. *Annals of Gastroenterology*, 28(2):203–209, 2015.

[6] Centers for Disease Control and Prevention. Beyond neighborhood food environments: Distance traveled to food establishments in 5 us cities, 2009–2011. https://www.cdc.gov/pcd/issues/2015/15$_0065.htm, 2015. [Online; accessed10 - February - 2024]$.

[7] T. Christensen. Anxiety is linked with smoking – but how is still hazy. *Current Psychiatry Reports*, 2019.

[8] CNET. Ever wonder how much you'd save shopping at costco for a year? we did the math. https://www.cnet.com/home/kitchen-and-household/how-much-can-i-save-shopping-at-costco/, 2023.

[9] S. Daily. Frequent consumption of meals prepared away from home linked to increased obesity. 2021.

[10] A. Drewnowski. The nutrient rich foods index helps to identify healthy, affordable foods. *Am J Clin Nutr*, 91(4):1095S–1101S, 2010.

[11] C. for Disease Control and Prevention. Adult obesity facts, 2021.

[12] C. for Disease Control and Prevention. Chronic diseases in america, 2022.

[13] M. Gibney. Ultra-processed foods: Definitions and policy issues. *Current Developments in Nutrition*, 3(2):nzy077, 2018.

[14] S. Gupta, T. Hawk, A. Aggarwal, and A. Drewnowski. Characterizing ultra-processed foods by energy density, nutrient density, and cost. *Frontiers in Nutrition*, 6:70, 2019.

[15] H. Jin. *Examining the geography of food deserts and food swamps in Austin, Texas.* PhD thesis, Texas State University, San Marcos, Texas, 2019.

[16] H. Jin and Y. Lu. Evaluating consumer nutrition environment in food deserts and food swamps. *Int J Environ Res Public Health*, 18(5):2675, 2021.

[17] A. E. Karpyn, D. Riser, T. Tracy, R. Wang, and Y. Shen. The changing landscape of food deserts. *PubMed*, 2020.

[18] F. Marangoni, C. Agostoni, C. Borghi, A. L. Catapano, H. Cena, A. Ghiselli, C. La Vecchia, G. Lercker, E. Manzato, A. Pirillo, G. Riccardi, P. Risé, F. Visioli, and A. Poli. Dietary linoleic acid and human health: Focus on cardiovascular and cardiometabolic effects. *Atherosclerosis*, 293:90–98, 2020.

[19] U. MyPlate. All about oils, 2021.

[20] News-Medical.net. Oils rich in linoleic acid, 2021.

[21] U. D. of Agriculture. Food away from home, 2022.

[22] J. of Cardiovascular Nursing. The fatty acid composition of vegetable oils and their potential health implications. 2019.

[23] W. H. Organization. Healthy diet, 2022.

[24] Rocketloop. Machine learning: Clustering in python. https://rocketloop.de/en/blog/machine-learning-clustering-in-python/, 2023.

[25] R. K. Singh, H.-W. Chang, D. Yan, K. M. Lee, D. Ucmak, K. Wong, M. Abrouk, B. Farahnik, M. Nakamura, T. H. Zhu, T. Bhutani, and W. Liao. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, 15(1):73, 2017.

[26] D. B. Stein. Food deserts' and 'food swamps' in hillsborough county, florida: Unequal access to supermarkets and fast-food restaurants. M.a. thesis, University of South Florida, 2011.

[27] A. Taha. Linoleic acid–good or bad for the brain? *npj Science of Food*, 4:1, 2020.

[28] U.S. Census Bureau. Census tracts shapefiles. https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2022layergroup=Census+Tracts, 2022.