Examining data sources and classification methods used in food access studies: Investigating volunteered geographic information as an adjunct to traditional data

by

Charles Maurice Hall

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

May 2017

Dedication

To MK, this wouldn’t have happened without you

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am grateful for the support and encouragement offered by my partner MK, both in preparing for and completing this program. I am grateful to the faculty at the USC Spatial Sciences Institute for all of the support and encouragement offered throughout my journey. I am grateful to my committee, John Wilson, Jennifer Swift, and Karen Kemp for their patience and support. I would like to thank USC and Adnan Choudhary for giving me the time, space, and encouragement required to complete this project. I would like to thank Alex Sosa for his assistance wrangling data. Thank you ACBC for the supportive smiles and WiFi.

List of Abbreviations

ACS	American Community Survey
API	Application Programming Interface
GIS	Geographic Information System
GISci	Geographic Information Science
HTTP	Hypertext Transfer Protocol
NAICS	North American Industry Classification Standard
NEMS	Nutrition Environment Measures Survey
OAUTH	Open Standard for Authorization
PIP	Python Package Manager
SIC	Standard Industrial Classification
SNAP	Supplemental Nutrition Assistance Program
SSI	Spatial Sciences Institute
USC	University of Southern California
USDA	United States Department of Agriculture
VGI	Volunteered Geographic Information

Abstract

This thesis performs a comparative analysis of traditional models of food access and a proposed model of food access that uses volunteered geographic information (VGI). Moreover, food businesses are often manually classified, which limits the number of businesses used for a given study. This thesis explores VGI as a potential improvement in the classification of food businesses. Field research was conducted in a subset of the selected facilities in order to determine the actual quality of the data retrieved from the experimental sources. The goal is to create a more nuanced and accurate representation of food access for a given person in a given place. Finally, data is compared for areas with different socio-economic conditions. Median income, car access, and percent minority from the 2010-2014 American Community Survey (ACS) 5-year estimates were used to define contrasting study areas. Two census tracts in Los Angeles were selected for the study area using these criteria: (1) an affluent area near La Cañada; and (2) a less affluent area in South Los Angeles. This thesis explores the quality and completeness of three data sets for census tracts with contrasting socio-economic conditions in order to identify whether or not problems exist with traditional methods and data. Furthermore, this thesis compares the data from census tracts with contrasting socio-economic conditions in order to determine whether or not the data varies based on the community served. The results of this thesis indicate that VGI does not represent a significant addition to commercial data because so few of the businesses are represented in the VGI data set. Moreover, the use of North American industry classification standard (NAICS) codes to classify businesses proved to be problematic. Specifically, numerous businesses that were classified as super markets or grocery stores were in fact smaller than convenience stores and sold fewer items. Finally, sentiment

analysis of reviews will require a larger data set and specifically trained models in order to be evaluated further.

Chapter 1 Introduction

Food deserts, food security, and food access have become popular topics of discussion in and out of academia in recent years. Cummins and Macintyre (2002) identify a 1995 document from a British policy working group as the first publication to use the term food desert. Eight years later, Walker et al. (2010) reviewed the food desert literature and identified 31 texts that had been published about food deserts. Their methods selected articles written in English, and excluded editorials, non-empirical works, works not focused on food deserts, and letters to the editor (Walker et al. 2010). Numerous books, articles, and films have been produced investigating these topics since their 2010 literature review; however, they often rely on commercial data sources to identify and describe the businesses within the study area. Though these data sets provide information about the businesses, such as size, income and number of employees, the data remains problematic because it lacks any measure of the variety and quality of goods offered. Consequently, the results of studies performed with these data can be difficult to interpret without additional work such as field evaluations being performed.

This thesis begins to explore the possibility of using data from social media in order to essentially crowd source the field evaluation portion of the data collection. Ratings and reviews can be used in place of in-person surveys of food facilities. Widener and Li (2014) collected data from the Twitter API and performed sentiment analysis on geolocated tweets in the United States in order to identify areas with healthy and unhealthy foods. This thesis investigates the overall quality and consistency of commercial data, and the utility of augmenting commercial data with volunteered geographic information (VGI) from the social media sources provided by Google and Yelp. In so doing this thesis evaluates commercial data and investigates whether or not VGI can reduce the need for field verification. The thesis uses the most current data available and is

not concerned with how access changes over time. Moreover, though this thesis addresses access to food, it does not address health outcomes. Finally, although community gardens, farmer's markets, and food trucks represent meaningful additions to the food environment, they are not discussed in this thesis. The primary focus is testing, verifying, and augmenting the data that is often used in current models.

The study compares data for two contrasting census tracts in the Los Angeles metropolitan area. Data from the 2010-2014 American Community Survey (ACS) was used to classify census tracts with regard to their median income, vehicles per occupied dwelling, and percentage of the population that is white. The initial site selection resulted in two tracts: 460700 and 224020 (Figure 1). Tract 460700 is located in La Cañada, CA, an affluent area north of downtown Los Angeles. It was chosen because the population is predominantly white with high income and multiple cars per household. Tract 224020 is located in an area of South Los Angeles, CA, and was selected because it contains a large minority population with low income and limited access to cars.

Figure 1 shows the two census tracts along with the scores calculated as follows:

$$score = \log(x_1) + \frac{x_2}{x_3} + x_4 \quad (1)$$

Where x_1 represents median income in US dollars, x_2 represents the number of cars in the census tract, x_3 represents occupied homes in the census tract, and x_4 represents the percentage of the population that is white. This method is useful for differentiating census tracts even though the units of the terms differ because it is a relative measure, not absolute. Consequently, different socio-economic situations can be identified through this methodology with a lower score representing fewer vehicles, less income and a larger minority population.

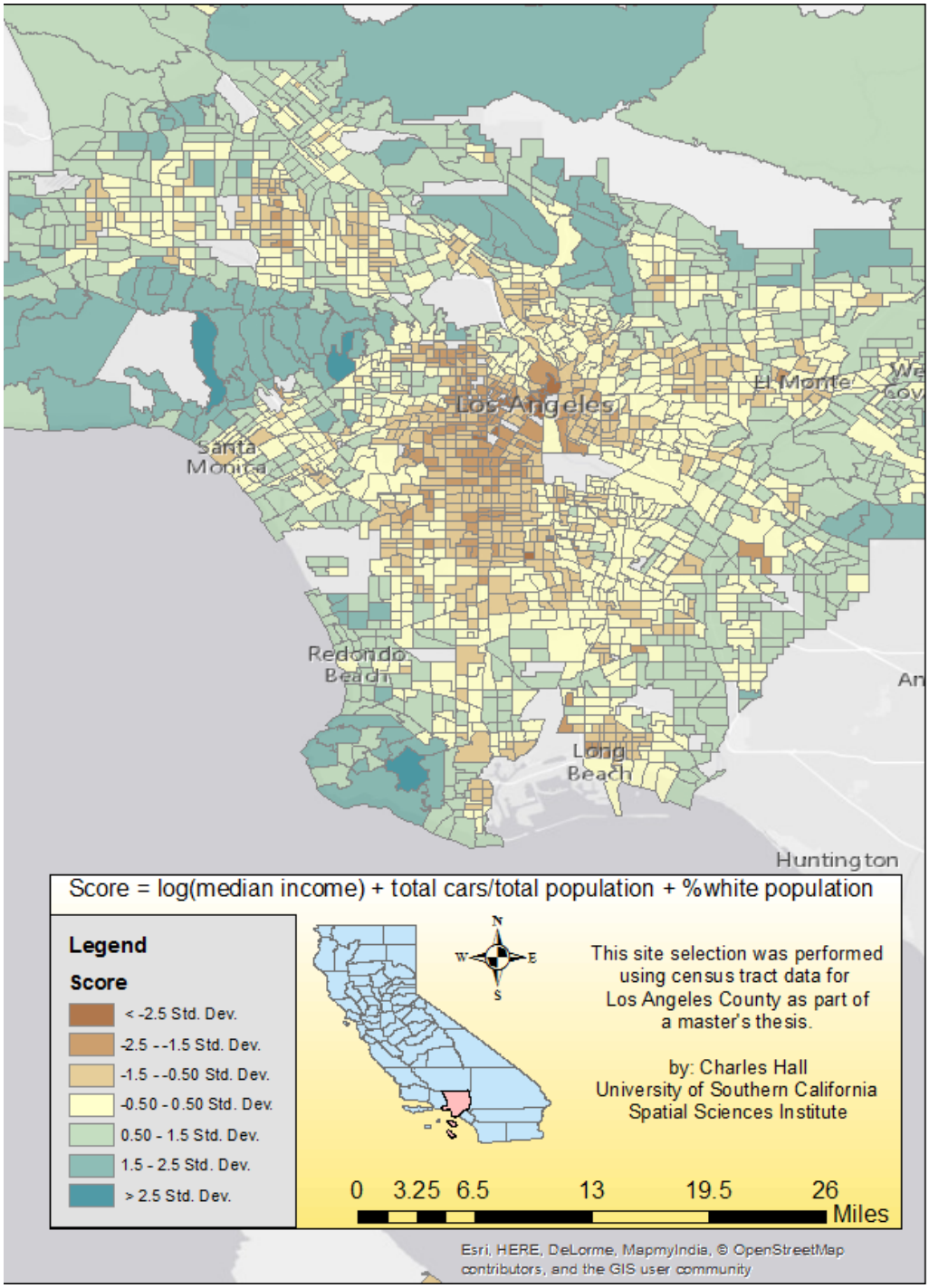


Figure 1 A map of the Los Angeles metropolitan area depicting census tracts colored by score along with a breakout map indicating the location of the Los Angeles metropolitan area within the state of California

1.1 Defining Healthy Food

A definition of healthy food is necessary in order to address the question of food security and access. The absence of a definition would erroneously show individuals who only have access to low quality food to have good access. The U.S. Department of Agriculture (USDA) publishes dietary guidelines that identify healthy food and diet choices. The USDA suggests a diet rich in fruits and vegetables that avoids processed sugars and other calorie rich, nutrient poor foods. Moreover, the guidelines suggest limiting saturated fats and sodium. This is often achieved by consuming foods that are fresh and less processed. Consequently, this thesis looked at the cost and availability of foods such as fruits, vegetables, dairy, and lean meat in each of the facilities examined. The availability and cost were used to assess the overall quality of the data that is traditionally used in food access studies.

A worksheet based on the Nutrition Environment Measures Survey (NEMS http://www.med.upenn.edu/nems/) from the University of Pennsylvania was created in order to standardize the evaluation of foods available in markets within the study areas. The USDA has published a Food Store Survey Instrument (https://www.cnpp.usda.gov/USDAFoodPlansCostofFood) that enables non-specialists to evaluate the quality of a market. The University of Pennsylvania has also published the full text of NEMS, which has also been designed to allow a non-specialist to evaluate the quality of a local market. These documents were reviewed and adapted in order to produce the final worksheet that was used in the evaluation of markets in the study area (Appendix B). The sheet includes both qualitative and quantitative measures that indicate the presence, quality, and cost of fresh food, dairy, and meat. Sections that evaluated soda, processed food, and canned foods were not used in this study because they were not relevant.

1.2 Defining Food Access

Charreire et al. (2010) reviewed 29 GIS-based food access papers and identified two major measures of food access: proximity and density. Proximity represents a measure of distance between a facility and a consumer. Euclidean and Manhattan distance are combined with buffer in order to define access. Network analysis tools can also be used to perform a proximity analysis. Density methods use tools such as cluster analysis and kernel density functions in order to visualize the number of markets within a given place. A person needs to have physical access to healthy food in a given place before they are able to make the choice to purchase and consume it. A distance of one half mile is often considered to be the maximum walking distance for a consumer, with further distances requiring a bicycle or vehicle of some type. Consequently, this study considers vehicle access as one of the variables when assessing access. The vehicle access variable was considered when selecting each of the two census tracts that are evaluated in the study. The density of facilities per local population was considered in addition to the overall quality and variety of food available in each location. This is important because food access studies often use one quarter of a mile as an acceptable walkable distance for a person without a vehicle, and density does not ensure quality.

In addition, the cost of staple items in each facility was evaluated. A person in a given place needs to be able to afford healthy food in order to choose to consume it. Prices were collected and analyzed for staple goods in each facility studied. This provided some insight into how prices change from facility to facility. Moreover, patterns could emerge that demonstrate variation from one census tract to another. It may help to address the question of how access to healthy food based on price and availability varies.

1.3 Traditional Data and VGI

Data from Esri Business Analyst, Google Places, and Yelp were aggregated and evaluated for completeness and quality in order to determine how meaningful the results of food access studies performed with these data are. Kerski and Clark (2102) enumerate five measures of accuracy that can be used when evaluating data quality: positional accuracy, attribute accuracy, logical consistency, completeness, and lineage. The data from Esri, Google, and Yelp were first evaluated using these measures as a part of this study. The data sets were compared to determine whether or not they had the same number of elements, and if not which were missing in order to address the question of completeness. The position of each element was also evaluated; however, because points are used to represent the businesses, some allowance for positional variation had to be made. This is due to the fact that individuals could choose several locations for the point: the front door or perceived center for example. Finally, the classification of the data in each data set was examined in order to assess the attribute accuracy. Data from all three sources was then compared to data collected in the field in order to determine how well the data represents the reality on the ground. This thesis suggests that an additional measure of quality that identifies how well the data represents the overall cost and quality of food in a facility is necessary in order to reap meaningful results from a food access study. It is necessary but not sufficient to know that a market is accessible. The quality of the food available in the market is also a necessary consideration.

Consequently, field work was conducted in order to evaluate the quality of the data retrieved from Esri, Google, and Yelp. A worksheet that records the availability and cost of staple food items was used to evaluate the quality of the data used. Data quality and results were compared between the two census tracts in order to determine whether or not the socio-economic

status of a census tract affects the quality of the data available for that tract. VGI, for example, might be more developed in places where greater numbers of people have access to smart phones and the internet while they are out.

1.4 Primary Research Questions

This thesis investigates the data sources used in GIS-based food access studies in order to answer several questions: (1) how well does the commercial data represent the reality of food access in the facilities that it represents; (2) does the use of VGI yield improved results; and (3) do socioeconomic factors affect either data set. Data from Esri Business Analyst was used as the commercial data source for this thesis, and selected markets were surveyed in order to determine how well the commercial data represented the reality of food access in the study area.

Data from Yelp and Google were then investigated to determine whether or not they represented a viable replacement or adjunct for commercial data. Can the results of a food access study be improved by incorporating VGI?

Finally, all three data sets were examined with respect to place in order to determine whether or not socioeconomic conditions affected the quality of the available data. Is the data from each provider consistent, or does it vary depending on place?

1.5 Motivation

Numerous studies of food access have been conducted with the aid of GIS technology. However, the data sources used in these studies are often problematic and generally require the author to visit the facilities or make assumptions about them. Traditional data sets contain limited details about the businesses that they represent, which complicates interpretation of the results. It is not sufficient to calculate the distance to the nearest market when determining

whether or not a person has access to healthy food in a given place. Additional information about the market would greatly inform the results.

Traditional food access studies often use North American Industry Classification System (NAICS) codes in order to select and classify food related businesses. NAICS codes in Esri Business Analyst include a proprietary two-digit suffix created by Dun and Bradstreet. However, there is no key available that defines the meaning of the suffixes. Esri was unaware of the existence of a key when contacted by telephone. Moreover, a Dun and Bradstreet representative was unaware of the existence of the proprietary suffix. Finally, a business librarian was also unaware of the suffix and unable to provide insight into their meaning. Manual aggregation and investigation of the codes revealed that neither the six digit NAICS code, nor the six digit code with the two digit suffix, classified food facilities into more than very general categories.

NAICS codes aggregate businesses into general categories such as market or restaurant. Consequently, food access studies often struggle with the classification of food facilities because of the use of NAICS codes. For example, it can be difficult to distinguish different types of markets and restaurants based solely on NAICS codes. Gard (2016) chose to work from the assumption that medium and large facilities represented supermarkets in urban and suburban neighborhoods, respectively. Others have resorted to manual classification, with the stipulation that the data set will only include recognizable national chains (Morganstern 2015). This paper investigates addressing the classification problem through the use of volunteered geographic information (VGI).

VGI has the potential to be classified in a more granular way because of the number of people who are able to contribute to the effort. Though the risk of misclassification exists, the increased quality of classification is a worthwhile tradeoff. APIs like those provided by Yelp and

Google Places allow for the selection of facilities by geographic location, and provide significant attribute data including ratings and reviews. Unlike NAICS codes, the classification is often textual, such as restaurants -> family -> burgers. A more robust classification system would result in more robust and nuanced results.

Finally, food access studies often indicate that urban dwellers have high access scores when compared to suburban and rural populations. This conclusion, however, could be misleading because it assumes that all food facilities are of equal quality. This study intends to avoid that assumption and interrogate the data in greater detail in order to avoid this and to better understand what the results mean for a given person in a given place.

1.6 Thesis Organization

The remainder of this thesis includes a literature review, methodology, results, and conclusions. The literature review, offered in Chapter 2, provides background for this study. It will examine data and methods used in previous food access studies and provide context for the results produced by this study. The methodology described in Chapter 3, reviews the tools, data sources, and methods used to conduct this study. The Python code used to acquire data from Google and Yelp can be found in Appendix A. The results presented in Chapter 4 describe the outcomes produced by the study, and provide some descriptive statistics. Chapter 5 interrogates the results and offers conclusions about how much value VGI could potentially add to a food access study. Moreover, the results are evaluated with respect to each of the census tracts used in the study in order to determine whether or not quality differences exist. Finally, the results are evaluated with respect to the data gathered in the field in order to inform the discourse surrounding food access studies with a quality assessment of the food facilities indicated by both the business data and the VGI.

Chapter 2 Related Work

As discussed in the previous chapter, the data used in food access studies continues to be problematic. This chapter reviews methods used in food access studies that have been conducted in order to illustrate the challenges that the data presents. Various classification methods have been employed in food access studies; however, many of them rely on indirect indicators such as square footage of the facility or annual sales. Other methods employ local knowledge in order to manually classify facilities.

The intention of this chapter is to present a clear representation of recent food access studies. Particular attention is paid to the methods used to classify food facilities because classification is the main challenge presented by traditional data sets.

The remainder of this chapter offers an interrogation of recent food access studies with some discussion of the methods used by each study. A brief discussion of USDA food access methods follows because the USDA is very involved in the discussion of food access and food deserts. The chapter concludes with a brief discussion of the need for improved classification methods and makes the case for testing VGI as a potential adjunct to traditional data.

2.1 Traditional Food Access Studies

Traditional food access studies draw on a variety of data sources including business directories, focus groups, food store assessments, geographic information systems (GIS), interviews, and surveys (Walker et al. 2010). Walker et al. (2010) performed a meta-analysis of food access studies wherein they identified the data sources and common themes in studies related to food deserts. Themes include store access, income / socio-economic status, race, density, cost, and quality of available foods. Their research only produced three articles that address the quality of food available in a market, and produced numerous articles related to

access, income, and race. This thesis focused on the quality of food available, and sought to compare two distinct and different census tracts. Consequently, their work informed the site selection methodology used in this thesis. Vehicle access was used in addition to income and race because a vehicle is often necessary when shopping for groceries.

It has been noted that the price, quality, and availability of fresh foods varies in different neighborhoods (Ball et al. 2009). Ball et al. (2009) examined the variation of accessibility, availability, and price of food in neighborhoods with various socio-economic conditions. They found that access was better for those in wealthier areas, availability only slightly favored the wealthy, and price favored the poor. The data used in their study was culled from online phone books and business directories.

Moore and Roux (2006) performed a similar study focused on the Baltimore, MD area. They sourced business data using NAICS codes and compared census tracks with varied socio-economic conditions. They suggest that non-minority neighborhoods have a greater number of supermarkets, while minority neighborhoods have smaller markets. Moreover, they identified deeper patterns that emerge based on the ethnic composition of the neighborhood.

2.2 GIS Approaches to Food Access Studies

GIS-based approaches to food access studies often use NAICS codes and commercial data in order to locate and interrogate food facilities in study areas. Lee (2012) employed NETS data from Dun and Bradstreet in one such study. She classified facilities into five categories that were "primarily based on 6-digit NAICS codes, although in some cases the 8-digit SIC codes were used to refine the definition, as well as business name, trade name, employee size, and annual sales information" (Lee 2012, p. 1196). Lee's classes included: (1) supermarkets, (2) corner stores, (3) convenience stores, (4) restaurants, and (5) fast food restaurants. Though her

method sought to identify access to healthy food, it is difficult to discern quality from the classification. Consequently, her results focus on counts and densities for each of the categories.

An and Sturm (2012) employed similar methods in their study of food access in California. They sourced their data from InfoUSA; however, they too use NAICS codes when classifying facilities. An and Sturm (2012, p. 130) sampled the data and used local knowledge to identify NAICS codes that represented different business classes:

> Although there is no NAICS code for fast-food restaurants, 63 major fast-food franchises are identified with main menus containing items such as hotdogs, burgers, pizza, fried chicken, subs, or tacos under the NAICS codes 72221105-6. Convenience stores are identified as NAICS code 44512001, and small food stores (annual sales <$1 million); midsize grocery stores (annual sales $1–$5 million); and large supermarkets (annual sales>$5 million) are identified as NAICS codes 44511001-3.

The NAICS codes used in their study include the two-digit proprietary suffix as well as a third digit followed by a dash. Their methods do not explain where the additional data comes from, and a review of the InfoUSA FAQ did not yield any clues. It is possible that InfoUSA has performed some additional work to extend the classification, and could be a potential source of improved facilities data.

Shier, An, and Sturm (2012) used similar methods in their paper "Is there a robust relationship between neighborhood food environment and childhood obesity in the USA?" Their classification method uses six-digit NAICS codes and annual sales in order to classify facilities. Much like Lee (2012), their methods focus on counting facilities within their study area. In this case, Shier et al. (2012) calculated the percentage of census tracts that have at least one of each of the defined classes. Though annual sales can be used to estimate the size of a market, it is not necessarily a good indicator of the quality of goods that the market sells.

Zick et al. (2009) used Dun and Bradstreet data in their study of food environments and obesity. Their methods use standard industrial classification (SIC) codes, the predecessor to NAICS codes, to classify facilities into four categories. There are corresponding NAICS codes for each SIC code, and Esri Business Analyst data includes both codes for each facility listed. Zick et al. (2009) then identified whether or not each census block group has a single type of facility or some mix of facilities in order to define food access for a given place. Like the previously discussed studies, it is difficult to assess the quality of food available given this classification method.

Powell et al. (2007) looked at the density of chain supermarkets in different neighborhoods. They concluded that low-income neighborhoods have a lower density of chain supermarkets when compared to higher income neighborhoods. Their paper used commercial data from Dun and Bradstreet combined with data from the US Census.

2.3 USDA Food Access Estimates

The U.S. Department of Agriculture has published a report about healthy food access in the United States. The USDA combines a list of stores that accept supplemental nutrition assistance program (SNAP) vouchers with commercial data from Trade Dimensions TDLinx in order to identify retail food facilities that offer a wide range of products (Ver Ploeg et al. 2012). The USDA uses square footage and annual sales to select and classify outlets into three classes: super-center, supermarket, and grocery store. Though their method works around the NAICS selection problem, it admittedly excludes numerous smaller businesses that sell healthy food. Moreover, size and sales are not necessarily a good indicator of the quality of food sold by a given facility.

2.4 Classification and VGI

The classification methods discussed in this chapter do not provide a direct representation of the quality of the food provided by a given facility. They instead classify facilities with proxies that attempt to identify the quality of food available. Other methods classify facilities by using the local knowledge of the author; however, this method does not scale well.

Food access studies would benefit greatly from an improved classification method that represents the quality of food available more directly. Google Places and Yelp both provide access to their data sets that include classification, ratings, and reviews. The additional data could be used instead of, or in conjunction with, traditional commercial data in order to generate a more robust and nuanced measure of food access. The remainder of this thesis investigates data from Google and Yelp in order to take a step toward understanding whether or not VGI is a useful adjunct to commercial data.

2.5 Sentiment Analysis

Widener and Li (2014) performed sentiment analysis on food-related geolocated tweets in order to determine whether people had a positive or negative feeling about the food which they were tweeting. Their research aggregated tweets related to food through a keyword matching process. Widener and Li (2014) created lists of healthy and unhealthy foods that were subsequently used to filter out tweets that were not related to food. Their project collected and analyzed 148,533 geolocated tweets from the continental United States.

Widener and Li (2014, p. 192) use the Alchemy API in order to perform sentiment analysis noting that "it has been validated through earlier studies to yield more accurate sentiment classification." They used a supervised model that was trained with over 200 billion words and produced a sentiment score between -1 and 1, with 1 being strongly positive and -1

strongly negative. Widener and Li (2014) divided their data set into healthy and unhealthy foods prior to analysis in order to obtain positive and negative sentiments for both healthy and unhealthy foods. The results were subsequently visualized in a population weighted kernel density plot in order to facilitate identification of patterns. They note that their work is in agreement with previous work by Smith and Brenner (2012) who also found that urban and suburban areas had higher twitter use rates than rural areas.

The results were further interrogated in order to search for patterns in food related tweets in low income, low access areas that have traditionally been referred to as food deserts. Their results indicate that low income tracts have 0.9% more tweets about unhealthy food. The authors note that although the difference is small, it is statistically significant. The authors conclude that these results "lends credence to the supposition that there are forces driving residents of low income neighborhoods with low access to healthy food stores, like supermarkets, to maintain less healthy diets" (Widener and Li, 2014, p. 195).

Chapter 3 Methodology

The methods employed by this study required initial data aggregation, site selection, Python development, data collection, and analysis. These methods supported the collection and comparison of commercial data, VGI, and observed data in the study area. Data from the US Census and ACS were used in order to identify two census tracts that were demographically different. This thesis compares data from a wealthy white neighborhood and a less wealthy, minority neighborhood. Government data from the USDA and US Census were first collected in order to identify the study area.

3.1 US Census Data

A census tract shapefile and tabular demographic data were downloaded from the US Census website for use in ArcMap (Figure 1). The shapefile included tracts for the entire state of California, and was first pared down for ease of processing. These methods select by attribute with COUTYFPS = '037' was used to reduce the overall number of tracts that had to be processed in ArcGIS during the site selection process.

Fields that represent the white population, total population, aggregate number of cars, inhabited homes, and median household income were identified. Fields B01003e1 and B01001m1 were used to identify the total population of the tract. The object id, geoid, B001003e1, and B002002e1 were selected in the X01_Sex_and_Age table, object id and B19113e1 was selected in the X19_Income table, and finally, object id, and B25001e1 were selected from the X25_Household_Characteristics table. The selection was necessary in order to reduce the overall size of the data set, which reduced processing time.

The site selection method used is derived from site selection methods that use the raster calculator. It was necessary to retain the data in vector format because the fishnet created by raster pixels did not line up neatly with the irregular boundaries of the census tracts. Consequently, a method was adopted that produces an aggregate score for each tract, which is subsequently visualized with a choropleth map. The score is a summation of the percent white population, car access, and median income.

Some data manipulation and cleaning were necessary in order to facilitate the calculation of scores for the census tracts. Specifically, fields from several different US Census tables needed to be aggregated into a single table before the score could be calculated. Consequently, the selected data was exported from each table into a comma separated values file, which was subsequently imported into Microsoft SQL Server. The import resulted in three tables that were aggregated into a single view, joined on the geoid column. The resulting view was opened in ArcMap and saved in the original file geodatabase for subsequent processing.

Fields were added for percent white population, automobile access, median household income, log(median household income), and overall score. All of the fields were of type double. Any row that had missing data was removed because it would not be useful for the purposes of this study. Finally, the field calculator was used to populate the calculated fields. The percent white population was calculated by dividing the white population by the total population. Car access was computed by dividing the number of cars by the number of inhabited homes. The log of median household income was computed in order to be included in the final score. The value of the final score is the summation of the three previously computed fields.

The resulting tabular data was joined with the census tracts shapefile in ArcMap on the geoid field. The data for the final score was visualized by creating a choropleth map of the

census tracts in Los Angeles County. The tracts were symbolized by quantity and classified by standard deviation. The resulting map, reproduced in Figure 1, was then employed to visually identify tracts that would be useful for this study.

Census tracts with large standard deviations from the mean were visually identified. Tracts with significant deviations were more thoroughly investigated for inclusion in the study. Facilities data was included from Esri Business Analyst in order to determine whether or not a tract was a good candidate for inclusion in the study. This was necessary because tracts without facilities would not be useful for the purposes of this study.

Half-mile buffers were created around both of the census tracts (i.e. polygons) used in this study (Figures 2 and 3). Businesses within the buffer were spatially joined from each of the three data sources and visualized. The use of buffers around the tracts was employed in order to mitigate edge effects. One half mile has already been identified as the maximum consumer walking distance. Consequently, one half mile was chosen for the buffer size because it is the maximum distance that a consumer is likely to travel outside of their tract on foot. Moreover, the half-mile buffer produced a sufficient number of businesses to be compared.

Census tract 224020 shown in Figure 2, was identified as a low income area with limited car access and high minority population within the Los Angeles metropolitan area. This tract is 33% white, has a median income of $22,042, and has 0.72 cars per household.

Census tract 460700 shown in Figure 3, was selected because it has contrasting characteristics. The area is 71% white, has a median income of $177,578, and 2.6 cars per household.

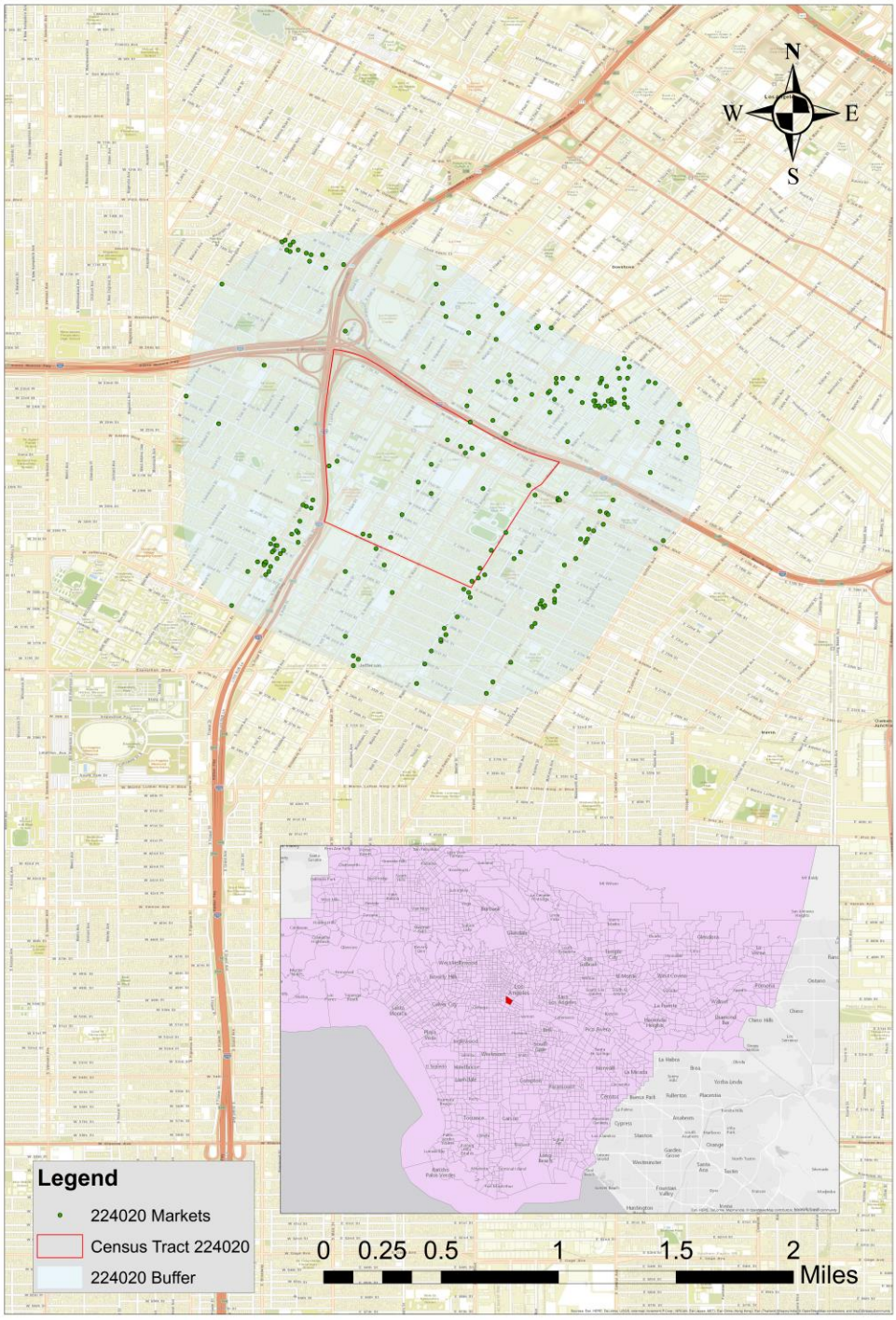


Figure 2 A map of census tract 224020 in South Los Angeles illustrating the study area, buffer, and markets surveyed in the study. The inset map illustrates the location of the census tract within the Los Angeles metropolitan area

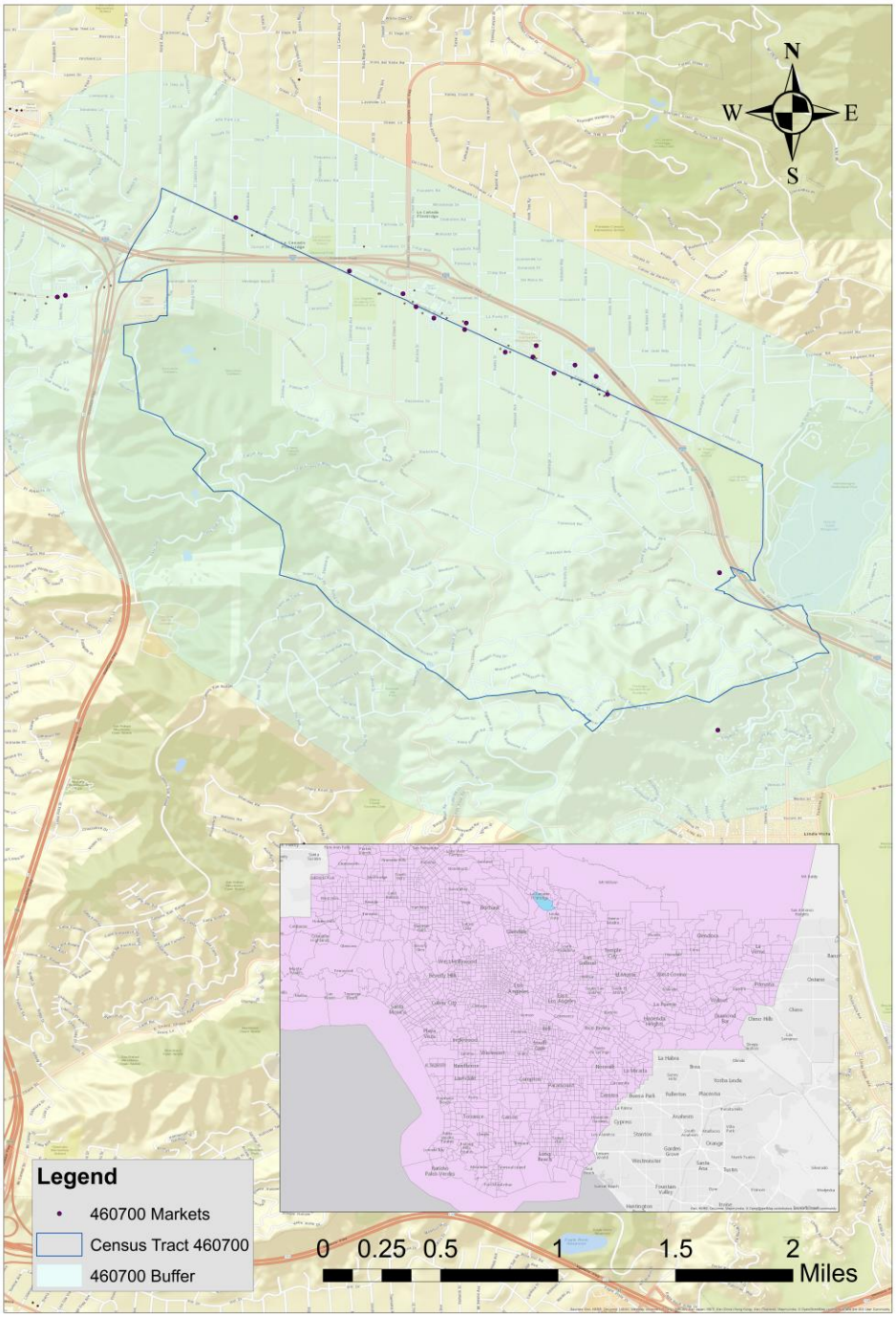


Figure 3 A map of census tract 460700 in La Cañada illustrating the study area, buffer, and markets surveyed in the study. The inset map illustrates the location of the census tract within the Los Angeles metropolitan area

3.2 Esri Business Analyst Data

Esri Business Analyst was used as a source of commercial data for this project. Data was selected by using the NAICS codes for markets, namely all NAICS codes that begin with 445. Business Analyst produced a shapefile containing point data for the businesses within the Los Angeles metropolitan area. Businesses within 1,500 m of the mean center of each tract were selected. The selection resulted in 33 grocery businesses in Census Tract 460700 and 91 grocery businesses in Census Tract 224020.

The join was performed in ArcMap with the select by location tool. Businesses that were within each Census Tract were selected, and a layer was then created from the selection. The process resulted in two layers, one for each Census Tract, which contained all of the Esri Business Analyst data for each tract.

3.3 Volunteered Geographic Information

Volunteered geographic information (VGI) was collected from both Google Places and Yelp for the two Census Tracts used in the study. Both systems use a point radius method to retrieve data from the database. They require latitude, longitude, and radius parameters when searching by location. The mean center of each tract was used with a 1,500 m radius in order to retrieve businesses that were outside of the tract but within the half-mile buffer. The businesses were then spatially joined with the buffer in ArcMap in order to produce the working data set. Python code was used to access and record the data from both APIs. The data was subsequently imported into ArcMap and point data were derived from the latitude and longitude provided in the results. The data retrieved from both systems is not projected and is in the WGS84 geographic coordinate system.

3.3.1. Computing Environment

The Anaconda package was chosen and used to install Python 2.7 onto a Windows computer because it is compatible with Windows, Mac, and Linux. Cross platform compatibility allows for the steps in the study to be repeated, and for the code to be re-used. Both APIs can be accessed with any programming language that can authenticate and make a secure http request. Python was used in this case because of the Arcpy integration with ArcGIS. Moreover, Yelp provides a Python client library that simplifies access to their API. The library is installed on the command line with pip using the following syntax: pip install yelp. The Yelp API requires OAUTH authentication, and the client library simplifies the process.

3.3.2. Google Places API

The Google Places API is accessed via a secure http request and returns paginated data. The API requires a client key to be passed with each https request as part of the query string. Python's native httplib was used to establish a connection with the Google API and collect results. A query string with all of the necessary components was assembled with concatenation operators in the Python code, and then passed to the API by httplib.

In addition to the key, the API request requires a location and a business category. The ArcGIS find mean center tool was used to identify the center of each census tract, which was then used as a location input for the API call. The categories were provided in tabular format in the API documentation. The restaurant and market categories were used for the purposes of this study.

The Google Places API returns 20 items per page along with a next page token. The Python code written for this study wrapped the API call inside a function that was recursively called with the next page token in order to aggregate the data from multiple pages into a single

data set. The API query string was concatenated and sent to the Googe API server. The http response code was checked, and if it was 200 OK, the response data was loaded into a data structure and parsed into JSON format. An iterator was then used to step through the JSON and output the latitude, longitude, business name, and classification written in CSV format. The CSV was then be opened in ArcMap and the plot XY was used to create point features that represented each business.

3.3.3. Yelp API

The Yelp API requires OAUTH1 authentication prior to being queried. Yelp provides a client library for Python that greatly simplifies the authentication process. Consequently the client library was employed in the code written for this study. The installation of the client library via pip is a prerequisite to executing any of the Yelp code provided in this thesis.

The Yelp API returns 20 businesses per function call, and uses an offset integer in order to access additional pages of data. The response data includes the total number of records returned. Consequently, an iterator is used to call the function recursively, incrementing the offset by 20 each time, until the offset is greater than the total. The results were then aggregated into CSV format so that they could be opened in ArcMap and point features could be created from the plotXY function.

3.3.4. Sentiment Analysis

Reviews were collected for each of the businesses identified during the site selection process. Sentiment analysis was performed on the reviews in order to investigate whether or not the reviews provide insight into the quality of goods available at each market. Additional census tracts were queried in order to increase the size of the data set and allow for comparison of results. Google and IBM provide REST APIs that facilitate access to their natural language

processing engines, both of which provide an interface for sentiment analysis. The Google and IBM models both return results in two parts: the sentiment and magnitude. The Google tool returns a numeric sentiment from -1 to 1 (negative to positive), while the IBM tool returns the word negative or positive. Both provide a magnitude that indicates the relative strength of the particular sentiment. This thesis limits the sentiment to positive or negative, although other sentiments do exist. The Google and Yelp APIs place strong limitations on the number of reviews returned for a given business. Google returns up to five while Yelp only provides an excerpt of a single review. The results would be more robust if a larger data set could have been used. Consequently, a scraper was built and employed in order to obtain a larger number of reviews for each business.

The scraper is a single Python function that uses httplib to execute a HTTP GET request against the Yelp servers. The Beautiful Soup library was then used to parse the returned text into XML that is easily searched. Each review has a label within the HTML paragraph tag that allows for the reviews to be easily located within the XML tree. Each page has 20 reviews, which require subsequent HTTP GET requests in order to obtain the full data set. Consequently, the scraper was executed recursively with an offset in order to collect all of the reviews. There is a review count tag at the top of the page that was used to identify the base case for the recursive function. The offset was incremented by 20 and the function was called again if the offset was less than the total number of reviews.

Sentiment analysis was performed on each review as it was collected. Functions were written to interact with both the Google and Alchemy REST APIs. The functions return the sentiment and magnitude as a comma separated list that was then concatenated with the review text and business ID information to produce the final CSV that contains all of the information for

the businesses surveyed in this study. The results were stored in a CSV file that could later be loaded into other software packages such as ArcGIS and R for further analysis. This workflow was used to get around the limitations of the Yelp API, which only provides an excerpt from a single review.

A final sentiment analysis tool based on work conducted by David Robinson was adapted in R in order to produce results that could be compared with the more opaque results returned by the online tools. The tools uses the NRC Word Association Library for R. The library was produced via an Amazon Mechanical Turk campaign that allowed people to manually assign sentiments to 14,182 words. Sentiment is evaluated for each non-trivial word in the review and then aggregated. Consequently, this tool produces a larger data set that could provide additional insight into the reviews. Moreover, all intermediate data was saved and analyzed allowing for greater insight into the words and phrases that drove the review. All of these scores were next compared and contrasted with the data collected in the field.

3.3.5. In-field evaluation of facilities

Food facilities identified in the previous steps were visited and evaluated. A worksheet, Appendix B, was employed in order to ascertain the quality and cost of food in each of the facilities within the study area. The worksheet is an abridged version of the worksheet developed by the University of Pennsylvania called the NEMS. Sections of the NEMS that address packaged foods, baked goods, and hot dogs were omitted because processed foods and baked goods do not fall within the definition of healthy that was used for this thesis.

The worksheet evaluated four categories of food in the markets for variety, price, and quality: dairy, bread, meat, and produce. These indicators provide insight into the availability of and access to healthy foods in each market. The presence of skim milk, lean meat, whole grain

bread, and fresh produce all indicate access to healthy food. Moreover, the greater variety in each of these items suggests variation in price that further facilitates access.

3.4 Data Evaluation

The data was evaluated in several ways for completeness and correctness. The Esri Business Analyst data was compared to both the VGI and field data. The Yelp and Google results were also compared to one another in order to determine where they intersected. The businesses in each data set were compared in order to identify overlapping businesses, and businesses missing from the data sets. Moreover, both census tracts were visited and the businesses were surveyed in order to confirm that they were still functioning businesses, and to determine the quantity and quality of goods available in the market.

3.4.1. Esri Business Analyst versus VGI

The Esri data and VGI were both aggregated into tabular format in CSV files, alphabetized, and examined. The comparison of the data sets was performed manually because of the small size of the VGI data set. The Esri data, being the largest data set, was used as the master data set against which the others were compared. The business names in the Yelp and Google data sets were both compared to the Esri data in order to determine whether or not they contained additional businesses that should have been included in the full data set. The data sets were merged and the duplicates were removed in order to create a working data set that was taken into the field.

3.4.2. Yelp versus Google

The tabular versions of the Yelp and Google data sets that were created in the previous steps were also compared to one another. This process was trivial given the size and lack of

intersection of both data sets. They were nonetheless compared in order to produce complete results.

3.4.3. Field Surveys

The aggregate data set was taken into the field, and all of the businesses in the data set were visited and surveyed. The first step in performing the survey was to locate the business and determine whether or not it was still operational. The location, address, and name were all verified during the survey. Businesses that were not in the data set were also noted and surveyed. The larger tract was surveyed by car over two days, and the smaller tract was surveyed on foot over three days.

Chapter 4 Results

This thesis examined two census tracts in the Los Angeles metropolitan area with very different demographic profiles. Tract 460700 is a higher income census tract located north of the city of Los Angeles, and tract 224020 is a lower income tract that encompasses the Pico Union, Fashion District, and South Los Angeles neighborhoods. A total of 96 businesses were derived from Esri Business Analyst in the two census tracts and within the half-mile buffers and 33 businesses were derived with Google and Yelp combined. Sixteen of the 33 businesses harvested form social media had at least one review. The resulting data set contained 285 reviews that were subsequently parsed into 10,538 words and analyzed with the Google, Watson, and R sentiment analysis tools.

4.1 Census Tract 224020

Census tract 224020 is located about 21.5 km north of the Los Angeles City Hall (Figure 3). The tract covers an area of 1.23 square kilometers and has a population of 2,527 with an error margin of 394 people in 1746 homes with an error margin of 65 homes. A majority of the businesses in the tract are located along Foothill Blvd, and most of the residences are located immediately north and south of this main roadway. There are a total of five major chain grocery markets in addition to a drug store and two convenience stores.

4.2 Census Tract 460700

Census tract 460700 is located about 3.3 km south of Los Angeles City Hall (Figure 2). The tract covers an area of 7.44 square kilometers and a population of 5,040 with an error margin of 373 people in 970 homes with an error margin of 60 homes. The markets in this tract are spread out in several different neighborhoods with a less obvious pattern. There are no major

chain grocery markets in the tract. The lone candidate in the data set, Fresh and Easy Neighborhood Market, has been closed. There are several markets classified with NAICS code 44511003, which indicates supermarkets and grocery stores. The majority of the businesses were small markets and convenience stores.

4.3 Esri Business Analyst Data

The Esri Business Analyst data set included 96 businesses that were located within the census tracts and half-mile buffers that comprised the study area. There are 72 businesses in tract 224020 and 24 in tract 460700. All of the major chain markets and convenience stores were included in the data set. Only two data points from the social media set were absent. One of the missing data points is now a specialty shop and was potentially filtered out when selecting by NAICS code. The other missing business is a Rite-Aid that was certainly filtered out when selecting by NAICS code.

There are significantly more businesses in the data set for tract 460700. Twenty-nine businesses are identified as supermarkets or grocery stores with NAICS code 44511003. Eight businesses are identified as convenience stores, and 11 are identified as liquor stores. Most of the businesses do not appear in the social media data set, only 17 out of 72 or about 22%. Conversely, the social media data set only included a Rite-Aid and two 99 Cents stores that were not included in the Esri Business Analyst data. Again they were likely excluded when the data was filtered by NAICS code.

There were a number of business listings in tract 224020 that were no longer in business. Several of the listings were buildings that appear to have been abandoned for some time. There were also several businesses that appear to have been misclassified as supermarkets or grocery stores. Finally, there were two large grocery stores in this tract that did not appear in the Esri data

set; however, several similarly branded markets do appear in the Esri data set for other parts of the city.

4.4 Social Media Data

The initial social media data set obtained by querying the Google and Yelp APIs resulted in a total of 29 businesses. There was some overlap in the results for tract 224020; however, there was no overlap for 460700. The overlap in the data set occurred for the larger markets: Sprouts, Trader Joes, Ralphs, and Gelsons. The data set was collected twice, once for the sentiment analysis portion and once for the field work, with the field work data including street address information for all of the businesses. Several rows did not include addresses and had to be sourced manually. All of the addresses that were returned corresponded with a physical place; however, as previously noted some of the locations were not markets.

The Google data set included data for businesses that were not markets on several occasions. The data included a gym, a retail store, a warehouse, and a personal residence. The Yelp data set did not include those types of businesses; however, it did include businesses that have since closed, as well as businesses in secure buildings.

The initial data set yielded very few reviews because the APIs both purposely limit the number of available reviews. A total of 285 reviews were downloaded separately in order to augment the data set used to perform sentiment analysis.

4.5 Sentiment Analysis

Sentiment analysis was performed on 61 reviews of 10 businesses in tract 224020. The tract had a mean Google sentiment of -0.1868852. Three of the 10 businesses have mean positive sentiment. Tract 460700 had 224 reviews of six businesses and a mean sentiment of 0.1571429. All of the businesses in the tract have positive sentiment with the exception of the Rite-Aid,

which has a strong negative sentiment of -0.5478261. A majority of the businesses in both tracts have mean sentiments that are very close to the median.

The IBM Watson tool delivers sentiment textually as positive or negative along with a numeric magnitude. Consequently, the magnitudes were coded as positive or negative, and the median and mean were computed for each business and for each tract overall.

Unlike the previous methods, the final analysis that was performed in R broke the reviews into words and aggregated the words and their sentiment for each tract. The process resulted in 10,538 words that were analyzed with the NRC lexicon. Words were assigned emotions in addition to positive and negative sentiments. The emotions included: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

4.6 Field Survey

Each of the markets in the Esri Business Analyst and social media data set was visited and physically surveyed. Price and availability of dairy, meat, and produce were recorded for each market. Each of the markets in tract 460700 with NAICS code 4451003 had plentiful dairy, meat, bread, and produce. Several brands of each product were readily available, along with alternatives such as organic, lactose-free, and soy. Large displays of fresh produce were visible in each market. Greens were in coolers that periodically water the produce in order to maintain freshness. Large refrigerators were stocked with numerous dairy offerings. Large coolers with packaged meat that included beef, chicken, turkey, and pork were well stocked in each of the markets. Several markets also had butcher and fish counters offering fresh custom cuts. Each market had a large bread aisle with a significant number of bread choices including whole grain, seeded, and sprouted, in addition to traditional white and wheat.

In contrast, just one of the markets in tract 224020 with NAICS code 4451003 had anything similar to that found in the previous tract. Super Farms in the Pico-Union area had a fairly large area filled with fresh produce. The market also had a dairy cooler stocked with whole and low-fat milk. Super Farms has a small butcher counter; however, there were limited prepackaged meat options. Finally, there was no bread found in the market with the exception of tortillas.

Another market in tract 224020, El Tronquito Market, is also classified with NAICS code 4451003; however, the market had a small section of a beer cooler devoted to dairy. Moreover, the market had three small plastic shopping bags of assorted produce, that was not very fresh. The bags of produce were placed on a shelf next to the cooler. There were three loaves of bread and no meat to be found.

4.6.1. Dairy

The five markets in census tract 460700 with NAICS code 4451003 are Ralphs, Trader Joes, Vons, Sprouts, and Gelso's (Table 1). Ralphs had six brands of milk including plain, organic, lactose free, and soymilk. Brands included Ralphs, Horizon, Alta Dena, Simple Truth, Broguire's, Heritage, and Lactaid. Ralphs brand was the lowest priced option for both whole and low-fat milk, while Horizon was the most costly. For both varieties, a half-gallon of Ralphs brand cost $1.99, and a half gallon of low-fat Horizon cost $4.79, a half-gallon of whole Horizon cost $5.79. The only gallons available were Ralphs brand, and cost $2.59 for low-fat and $2.89 for whole milk.

Table 1 Dairy survey results showing prices in U.S. dollars for the lowest cost choices

Market Name	Skim Quart	Skim Half Gallon	Skim Gallon	Whole Quart	Whole Half Gallon	Whole Gallon
Tract 460700						
7-Eleven 1		2.49	2.99		2.49	
Gelsons	1.79	2.79	5.49	1.79	2.79	5.49
Ralphs		1.99	2.59		1.98	2.89
Rite Aid		2.59	2.99			2.99
Sprouts		1.99	2.99	5.99	1.99	2.99
Trader Joes				1.29	1.89	2.89
Two Brother Food	This is a specialty baking supplies shop					
Vons	1.49	1.99	2.69	1.49	1.99	2.99
Whispering Gardens	This is a private residence					
Tract 224020						
99 Cents 1	0.99			0.99		
99 Cents 2	0.99		2.39	0.99		2.39
30th Street Market						3.29
7-Eleven 1	1.79	2.29		1.79	2.29	3.49
7-Eleven 2		2.49	3.49		2.49	3.49
7-Eleven 3	1.99	2.49	3.49	1.99	2.49	3.49
Adams Ranch Market	Could not locate					
Bembi's Market	Could not locate					
Black and White	Could not locate					
Cal Mart			4.84			4.84
Chayo Market				1.293.00	3.30	
Corona Market			3.80			3.80
Daily Food Market	price unknown	price unknown	3.59	price unknown	price unknown	3.59
Duran's Market						4.00
El Charrito					2.79	4.00
El Porvenir						
El Principio Market 1					2 for 3.49	3.49
El Principio Market 2					1.99	2 for 5.50
El Tronquito		3.49				5.00
Fresh and Easy	Was there but closed					
GsG II	Could not locate					
Jessie's Market						4.00
JNE	Could not locate					

Joe's Market					3.99	
KK Inc.	Could not locate					
Kola Huh	Could not locate. Main Street Market?					
La Moderna						
La Reina	Could not locate					
La Tia	Could not locate					
Main Street Market						4.00
Manuel's	Could not locate					
Mexicali Market		2.49	2.99		2.49	2.99
Numero Uno		2.49	2.79		2.49	2.79
Numero Uno		2.49	2.79		2.49	2.79
Ofelias		2.75				3.79
Pinski Portugal	This is a specialty group in a wholesale building					
Reyes Mini Mart					2.99	3.99
Rite Aid	1.79	2.59	2.99	1.79	2.59	2.99
Ron's	Could not locate					
Super Mercado San Carlos	Could not locate					
Super Farms Market		3.19	3.69			3.64
Sweet Illusion	This is a lingerie company					
Tikal Market						3.29
Trimanna Express	Could not locate this business, the building at the address is secure					
Young's Market						3.50

Trader Joes only stocked Trader Joes branded milk (Table 1). They offered standard and organic choices in quarts, half-gallons, and gallons. The quarts cost $1.29, the half gallons cost $1.89, and the gallons cost $2.89. The organic variety was only available in the half-gallon and gallon choices during my visit. The half-gallon of organic milk cost $3.99, and the gallon cost $5.99.

Vons offered eight varieties of milk: Clover, Organics, Valu, Heritage, Alta Dena, Lucerne, Fairlife, and Horizon (Table 1). The low cost option for quarts and half-gallons was Lucerne, which cost $1.49 for a quart and $1.99 for a half-gallon for both low-fat and whole

milk. The low cost gallons were Valu brand which cost $2.69 for low-fat, and $2.99 for whole milk. The highest cost choice was Horizon, which cost $7.99 per gallon.

Sprouts stocked five brands of milk at the time of my visit: Sprouts brand, Horizon, Alta Dena, Broguire's, and Organic Valley (Table 1). The market also stocked a few pints of locally produced raw milk that had not been pasteurized. Sprouts brand was the low cost choice for both varieties of milk with half-gallons priced at $1.99 and gallons priced at $2.99. Sprouts Organic and Horizon were the most costly choice, both were priced at $4.99 for half-gallons, and Sprouts Organic gallons were priced at $6.49.

Gelsons stocked a total of eight milk varieties: Rock View, Organic Valley, Fairlife, Alta Dena, Horizon, Broguire's, Saint Benoit, and Fair Life (Table 1). Rock View quarts were the low cost choice priced at $1.79, while Alta Dena was the costlier option priced at $2.29. Half-gallons were priced between $2.79 for Alta Dena, and $5.49 for Horizon. Gallons from Alta Dena were available for $5.49, and Horizon was priced at $7.99.

The 7-Eleven and Rite-Aid both stocked Alta Dena half-gallons and Swiss gallons of whole and low-fat milk. Rite-Aid also stocked Horizon half-gallons as a third choice. Both markets priced the Swiss gallons at $2.99. Half-gallons at 7-Eleven were priced at $2.49, and were 10 cents more costly at Rite-Aid, priced at $2.59. The Horizon half-gallons at Rite-Aid were priced at $5.99.

The largest market surveyed in tract 224020 was Super Farms in the Pico-Union area (Table 1). The market had a medium sized dairy case stocked with half-gallons and gallons of both Swiss and Alta Dena milk. Alta Dena half gallons were priced at $3.19 for both low-fat and whole milk. Swiss gallons were priced at $3.69 per gallon, and Alta Dena was priced at $4.89 per gallon for both low-fat and whole milk.

Two markets in the southern fashion district of downtown Los Angeles stocked very limited quantities of milk. Joe's LA Market Deli had three half-gallons of Alta Dena whole milk in the cooler, priced at $3.99. El Tranquito had two half-gallons and three gallons of Alta Dena milk priced at $3.49 and $5.00, respectively.

Two 99 Cents stores were surveyed and both stocked Golden Crème brand low-fat and whole milk. One location had quarts and gallons priced at $0.99 and $2.69, respectively. The other location only had a few quarts of each priced at $0.99. Two 7-Eleven markets were also surveyed in this tract. Both stocked Alta Dena half-gallons, and Swiss gallons, and one location also stocked Horizon half-gallons of whole milk. The Alta Dena half-gallons were priced at $2.49, the Swiss gallons were priced at $3.49, and the Horizon half-gallons were priced at $6.29.

Several small markets were surveyed in South Los Angeles: Young's Market, Cal Mart, Ofelia's, Mexicali Meat Market, Tikal Market, El Principio, Main Street Market, Daily Food, Corona Market, Chayo Market, 30th Street Market, Reyes Market, Jessie's Market, Duran's Carniceria, and El Charito Market (Table 1). Young's Market stocked a few gallons of Kirkland whole milk that were priced at $3.50. Cal Mart stocked gallons of Rock View low-fat and whole milk priced at $4.84. Both markets had limited stock that consisted of less than five gallons on a cooler shelf. Ofelia's stocked half gallons of Alta Dena low fat milk for $2.75 and gallons of Swiss whole milk for $3.79. Mexicali stock Alta Dena and Rock View half gallons and gallons of both low fat and whole milk for $2.49 and $2.99, respectively. Moreover, Mexicali stocked Lactaid, soy and almond milk. Tikal Market stocked a few gallons of Rock View whole milk for $3.29. They were sold out at the time of the survey. El Principio stocked Alta Dena half gallons and Swiss gallons for $3.49 per gallon, with the half gallons sold in pairs. El Principio 2, which is larger than the original, stocked half gallons and gallons of Alta Dena whole milk for $1.99

and $3.69, respectively. They were also running a special on gallons of Swiss whole milk, selling two gallons for $5.00. Main Street Market stocked about eight gallons of Golden Crème whole milk priced at $4.00. Corona Market sold gallons of Rock View low fat and whole milk for $3.80. Chayo Market stocked quarts, half gallons, and gallons of Rock View low fat and whole milk for $1.29 per quart, $3.00 per half gallon, and $3.30 per gallon. 30th Street Market stocked Golden Crème gallons sold for $3.29, Rock View half gallons with an unknown price, and Lactaid. Reyes Market stocked a couple of half gallons and four gallons of Alta Dena whole milk priced at $2.99 and $3.99, respectively. Daily Food stocked quarts, half gallons, and gallons of Alta Dena and gallons of Rock View low fat and whole milk. The gallons were priced at $3.59 for Rock View and $3.99 for Alta Dena. The cashier was unable to provide a price for the quarts and half gallons of Alta Dena milk. Jessie's Market and Duran's stocked a few gallons of Swiss whole milk for $4.00. El Charito stocked a few half gallons and gallons for Rock View whole milk priced at $2.79 and $4.00, respectively.

Four markets that were not in included the data set were surveyed in the interest of completeness: Rite-Aid, Numero Uno, Angel's Nutrition WIC, and WIC. Rite-Aid and WIC are located in the same strip mall as one of the 99 Cents stores. The Numero Uno and Angel's Nutrition were located between other markets on San Pedro Street. The Rite-Aid sold Alta Dena quarts and half-gallons, and Swiss gallons. The Alta Dena quarts were priced at $1.79, the half-gallons were priced at $2.59, and the gallons were priced at $2.99.

The WIC store sold Swiss, Dairy Pure, and Mother's milk. The low cost option for low fat and whole milk was Swiss priced at $2.94. The more costly choice was Dairy Pure priced at $4.28. A gallon of low fat Swiss milk was priced at $3.89, and a gallon of whole Mother's milk was priced at $3.62. The staff indicated to me that they believe that the prices on dairy items

were inflated because most of their customers used government assistance coupons that were meant to purchase the item at the listed price. Angel's Nutrition WIC Shop sold Rock View half gallons and gallons of low fat and whole milk for $2.60 per half gallon and $4.35 per gallon. They had a large, well-stocked dairy case.

Finally, Numero Uno market was a large grocery store that was conspicuously missing from the data set. I encountered several people on the street who were carrying shopping bags branded with the Numero Uno logo. There were two markets that were several blocks apart on the same street. They stocked Foremost, Swiss, Alta Dena, Horizon, and Heritage low fat and whole milk. The low cost half gallon was Swiss milk priced at $2.49 and the costliest option was Heritage at $5.39. Low cost gallons of Foremost were priced at $2.79, and Alta Dena was the high priced option at $4.49.

4.6.2. Bread

All of the major chain markets in tract 460700 had significant bread aisles with stocked shelves (Table 2). The Ralphs location had a large bread aisle containing more than six brands of bread, often with several varieties of bread per brand. Brands included Home Pride, Western Hearth, Orowheat, Van de Kamps, Natures Own, Wonder, and Sara Lee. The low cost wheat bread choice was Van de Kamps priced at $1.29, and the high price was Orowheat or Sara Lee priced at $3.99. Van de Kamps wheat bread was also priced at $1.29, and Home Pride wheat was $3.99. There were also numerous other bread offerings including sourdough, flat bread, and bagels.

Table 2 Bread survey results showing prices in U.S. dollars for the lowest cost choices

Market Name	Whole Wheat	White	Other Notes
Tract 460700			
7-Eleven 1			
Gelsons	2.49	1.99	Large variety of bread available
Ralphs	1.29	1.29	Large variety of bread available
Rite Aid	2.99	2.99	Hostess bread
Sprouts	2.59	2.59	Large variety available, limited white bread options
Trader Joes	1.99		
Two brother food			
Vons	1.19	1.29	
Whispering Gardens			
Tract 224020			
99 Cents 1		0.99	
99 Cents 2	0.99	0.99	
30th Street Market	3.49	3.49	Bimbo brand
7-Eleven 1			
7-Eleven 2			
7-Eleven 3			
Adams Ranch Market			
Bembi's Market			
Black and White			
Cal Mart			
Chayo Market	3.59	3.59	Bimbo brand
Corona Market	2 for 5.00	2 for 5.00	Bimbo brand
Daily Food Market	2 for 5.00	2 for 5.00	Bimbo brand
Duran's Market			
El Charrito			
El Porvenir			
El Principio Market 1	2.99	2.99	Bimbo brand
El Principio Market 2	2.99	2.99	Bimbo brand
El Tronquito		2.99	
Fresh and Easy			
GsG II			
Jessie's Market			
JNE			

Joe's Market			
KK Inc.			
Kola Huh			
La Moderna			
La Reina			
La Tia			
Main Street Market			
Manuel's			
Mexicali Market	2.99	2.99	
Numero Uno	1.19	1.19	Variety
Numero Uno	1.19	1.19	Variety
Ofelias		2.99	Bimbo brand
Pinski Portugal			
Reyes Mini Mart			
Rite Aid			
Ron's			
Super Mercado San Carlos			
Super Farms Market			
Sweet Illusion			
Tikal Market			
Trimanna Express			
Young's Market			

Trader Joes had a full selection of bread, all from their brand. The low and high cost choices were priced at $1.99 and $2.99, respectively, with the higher cost option being a fancier version of the same bread.

Vons stocked at least seven brands of bread including: Nature's Own, Sara Lee, Natures Harvest, Bimbo, Orowheat, Signature, and Dave's. Signature white and wheat breads were the low cost option priced at $1.19 and $1.29, respectively. Roman Meal what was the costly choice priced at $3.99, and Wonder was the costly white bread choice at $2.99. Vons also stocked numerous other varieties of bread such as sourdough, seeded, flat, bagels, and dinner rolls.

Sprouts has a bread section near the store entrance with a variety of brands and types. Brands stocked include Sara Lee, Dave's, Apple Valley, Nature's Own, and Orowheat. Nature's Own was the low cost wheat bread choice priced at $2.59, and Dave's was the costly choice priced at $5.59. There was only one white bread choice, Nature's Own, which was priced at $2.59.

Gelsons had a well-stocked bread aisle that included brands such as Nature's Harvest, Roman Meal, Rudi's, Orowheat, and Home Pride. The low cost choice was Nature's Harvest priced at $2.49, and the high cost choice was Roman Meal at $4.49. Other specialty wheat varieties such as Rudi's were priced even higher, at $5.79; however, it was excluded because it was more like a specialty bread than a traditional bread. There were several other specialty choices such as Ezekial bread, which was also priced above $5.00 per loaf.

The 7-Eleven did not stock any bread; however, the Rite-Aid did stock a few loaves of Hostess white and wheat bread priced at $2.99 per loaf. The bread was shelved with other Hostess products that are sugary desserts.

Thirteen of the markets in tract 224020 sold bread: El Tranquito, both 99 Cents locations, Daily Food, Corona Market, El Principio 1 and 2, Ofelia's, Mexicali, Chayo market, 30th Sreet Market, and Numero Uno (Table 2). El Tranquito had three loaves of white bread on a shelf priced at $2.99. Both 99 Cents locations sold white and wheat Golden Baked bread for $0.99. Both locations had a significant amount of bread available for sale. El Principio 1 and 2, Ofelia's, and Mexicali sold Bimbo white and wheat bread for $2.99, Daily Food and Corona Market sold Bimbo white and wheat loaves for $5.00. 30th Street Market stocked Bimbo white and wheat for $3.49, and Chayo stocked the same for $3.59. Angel's Nutrition stocked Nature's Harvest white and wheat for $3.65 per loaf.

Numero Uno had a large bread aisle stocked with Roman Meal, Home Pride, Orowheat, Bimbo, Springfield, and Sara Lee. Springfield was the low cost option for both white and wheat priced at $1.29 per loaf. Roman Meal was the costly wheat option priced at $3.79, and Bimbo was the costly white choice at $3.49.

4.6.3. Meat

All of the major chain markets in tract 740600 had extensive meat sections containing various cuts of meat packed in foam meat trays and wrapped in plastic (Table 3). All of them sold ground beef and turkey, and most also sold ground chicken and pork. Ralphs sold ground beef with 27%, 20%, and 10% fat, with the 20% priced at $4.99 per pound and the 10% priced at $6.99 per pound. Ground turkey was priced at $2.49 per pound and ground chicken at $4.99 per pound. The market also offered items such as ground pork, bison, and organic options. Moreover, there were multiple varieties of ground beef available to choose from.

Trader Joes offered ground beef with 10% and 15% fat, with the 15% lean being an organic option priced at $7.49 per pound. The 10% option was not organic and was priced at $6.49 per pound. The store also sold ground turkey for $2.99 per pound. They did not, however, sell ground chicken.

Vons had the clearest labels on their beef with respect to fat content and price per pound. They listed the options as 20% and 10% fat with a clear price per pound on the label. The 20% fat was priced at $4.49 and 10% was priced at $5.99. The market also sold ground turkey for $7.49 per pound, and ground chicken for $4.99 per pound.

Table 3 Meat survey results showing prices in U.S. dollars for the lowest cost choices

Market Name	Beef	Prices	Turkey	Prices	Chicken	Prices
Tract460700						
7-Eleven 1						
Gelsons	90% / 80%	8.99 / 7.99	Dark / Breast	5.99 / 7.99	ground	10.99
Ralphs	90% / 80%	6.99 / 4.99	ground	2.99	ground	4.99
Rite Aid						
Sprouts	15%	4.99	ground	7.49	ground	2.99
Trader Joes	90% / 85%	6.49 / 7.49	ground	2.99		
Two brother food						
Vons	90% / 80%	5.99 / 4.49	ground	7.49	ground	4.99
Whispering Gardens						
Tract 224020						
99 Cents 1						
99 Cents 2						
30th Street Market						
7-Eleven 1						
7-Eleven 2						
7-Eleven 3						
Adams Ranch Market						
Bembi's Market						
Black and White						
Cal Mart						
Chayo Market						
Corona Market	Unknown	5.50 / lb				
Daily Food Market	Unknown	4.69 / lb			ground	2.99 / lb
Duran's Market	Unknown	3.99 / lb				
El Charrito						
El Porvenir						
El Principio Market 1	Made to order	Various				
El Principio Market 2	Unknown	4.19 / lb				
El Tronquito						
Fresh and Easy						
GsG II						
Jessie's Market						
JNE						

Joe's Market						
KK Inc.						
Kola Huh						
La Moderna						
La Reina						
La Tia						
Main Street Market						
Manuel's						
Mexicali Market	Unknown	4.89				
Numero Uno	85% / 70%	4.49 / 3.69				
Numero Uno	85% / 70%	4.49 / 3.69				
Ofelias						
Pinski Portugal						
Reyes Mini Mart						
Rite Aid						
Ron's						
Super Mercado San Carlos						
Super Farms Market	Lean	4.99				
Sweet Illusion						
Tikal Market						
Trimanna Express						
Young's Market						

Sprouts only had 15% fat ground beef at the time of my visit, which was priced at $4.99 per pound. The market sold ground turkey for $7.49 per pound, and ground chicken for $2.99 per pound. There was also a small butcher counter that could potentially accommodate special requests.

Gelsons sold ground beef with 15% and 9% fat for $7.99 and $8.99, respectively. They sold two varieties of ground turkey, dark meat and breast, for $5.99 and $7.99, respectively. Ground chicken was available for $10.99 per pound. There was also a butcher counter with a butcher available to accommodate custom requests and offer guidance with respect to the meat products.

Several markets in tract 224020 offer non-processed meat, including Super Farms, Daily Food, Duran's, Corona Market, El Principio 1 and 2, Mexicali, and Numer Uno (Table 3). Super Farms had a small butcher counter that offered a variety of meat. The ground beef was simply labeled lean, with two varieties available for $4.79 and $4.99. There was nobody at the counter available to answer questions.

Daily Food stocked lean ground beef, with an unknown fat quantity, and ground chicken for $4.69 and $2.99, respectively. Duran's was a small carniceria with a butcher counter that sold ground beef of unknown fat content for $3.99 per pound. Corona market had a small butcher case that contained ground beef of unknown fat content priced at $5.50 per pound.

El Principio 1 and 2 both had small butcher counters stocked with various types of meet including beef, chicken, and sausage. El Principio 1 ground meat to order, and the price and quality consequently varied. El Principio 2 had ground beef of unknown fat content in the butcher case priced at $4.19 per pound. Mexicali was a larger space with a butcher counter that provided a single variety of ground beef with unknown fat content for $4.89 per pound.

Numero Uno had a butcher counter as well as a large walk up meat cooler that was well stocked. They had packages marked 70/30 and 85/15 fat content as well as ground beef labeled lean priced at $3.69 and extra lean priced at $4.49 per pound.

4.6.4. Produce

All of the major chain markets in tract 740600 had large produce displays that were well stocked and clearly labeled (Table 4). The produce was well organized and of good quality. The variety of the produce and the price was clearly marked. Green vegetables were kept in coolers that regularly watered the produce in order to maintain freshness.

Table 4 Produce survey results showing prices in U.S. dollars for the lowest cost choices

Market Name	Apples $	Bananas $	Oranges $	Carrots $	Tomatoes $	Broccoli $
7-Eleven 1	0.99 each		0.99 each			
Gelsons	2.49 / lb	0.29 / lb	2.29 / lb	0.99 / lb	2.49 / lb	1.99 / lb
Ralphs	0.99 / lb	0.69 / lb	1.29 / lb	0.99 / lb	0.99 / lb	1.99 / lb
Rite Aid						
Sprouts	0.98 / lb	0.69 / lb	0.88 / lb	0.99 / lb	0.98 /lb	1.49 / lb
Trader Joes	0.69 each	0.19 each	0.79 each	1.49 / lb	0.29 each	1.79 / bunch
Two brother food						
Vons	1.49 / lb	0.69 / lb	3.99 / bag		1.69 / lb	1.29 / lb
Whispering Gardens						
99 Cents 1		0.49 / lb	0.99 / bag		0.99 / lb	
99 Cents 2		0.49 / lb	0.99 / bag	0.99 / bag	0.99 / lb	
30th Street Market						
7-Eleven 1	0.99 each	0.50 each				
7-Eleven 2	0.89 each	0.89 each	0.89 each			
7-Eleven 3	0.69 each	2 for 1.00	0.69 each			
Adams Ranch Market						
Bembi's Market						
Black and White						
Cal Mart						
Chayo Market						
Corona Market		0.69 / lb				
Daily Food Market		0.69 / lb	0.59 / lb		0.99 / lb	
Duran's Market					0.99 / lb	
El Charrito						
El Porvenir						
El Principio Market 1	0.89 / lb		0.59 / lb	0.79 / lb	0.79 / lb	1.29 / lb
El Principio Market 2	unknown		unknown		unknown	
El Tronquito		1.00 each	1.00 each			
Fresh and Easy						
GsG II						
Jessie's Market						
JNE						
Joe's Market		1.00 each				

KK Inc.						
Kola Huh						
La Moderna						
La Reina						
La Tia						
Main Street Market						
Manuel's						
Mexicali Market		0.79 / lb				
Numero Uno	0.69 / lb	0.59 / lb	1.00 / 2 lbs	0.59 / lb	0.99 / lb	0.99 / lb
Numero Uno	0.69 / lb	0.59 / lb	1.00 / 2 lbs	0.59 / lb	0.99 / lb	0.99 / lb
Ofelias						
Pinski Portugal						
Reyes Mini Mart	0.50 each		0.50 each		0.50 each	
Rite Aid						
Ron's						
Super Mercado San Carlos						
Super Farms Market	0.99 / lb	0.50 / lb		2.99 / lb	0.89 / lb	
Sweet Illusion						
Tikal Market						
Trimanna Express						
Young's Market		0.25 each	0.50 each			

Ralphs stocked red delicious, golden delicious, fuji, granny smith, and gala apples for between $0.99 and $1.29 per pound. Bananas, priced at $0.69 per pound, were green to yellow in color and medium size. Oranges with good color and minimal bruising were priced at $1.29 per pound. Fresh carrots were kept cool and priced at $0.99 per pound. Fresh, red, Roma tomatoes were priced at $0.99 per pound. Fresh broccoli that was kept cool and moist was priced at $1.99 per pound.

Trader Joes stocked gala, sweet tango, envy, honey crisp, and fuji apples for between $0.69 and $1.29 per apple. The apples were clean and bright looking with minimal blemishes. Bananas that were green to yellow in color were sold for $0.19 each. Large navel oranges that were of good quality were sold for $0.79 each. Bagged baby carrots that were kept cool and

clean sold for $1.49 per pound. Red Roma tomatoes were priced at $0.29 each. Bagged fresh broccoli was priced at $1.79 per bunch.

Vons stocked envy, red delicious, gold delicious, fuji, gala, kiku, honey crisp, and jazz apples. The fuji apples were the lowest price at $1.49 per pound, and the honey crisp was the costliest at $3.49 per pound. Green to yellow colored, medium sized bananas were priced at $0.69 per pound. Bagged oranges with some discoloration and bruising were priced at $3.99 per bag. Bagged baby carrots were kept cool and sold for $7.45 per bag. Red Roma tomatoes with some discoloration and blemishes were sold for $1.69 per pound. Fresh broccoli that had been recently sprayed was sold for $1.29 per pound.

Sprouts stocked jazz, red delicious, golden delicious granny smith, gala Fuji, tango, Braeburn, honey crisp, and pink lady apples. Gala apples were the low priced option at $0.98 per pound. All of the apples were clean and appeared to be unblemished and of good quality. Medium sized bananas with green to yellow color were priced at $0.69 per pound. Fresh and unblemished navel oranges were priced at $0.88 per pound. Good quality carrots were priced at $0.99 per pound. Tomatoes that were not yet fully ripe were priced at $0.98 per pound. Broccoli that was fresh and green was priced at $1.49 per pound. Sprouts also had a separate section that contained organic fruits and vegetables.

Gelsons stocked Fuji, honey crisp, envy, pink lady, red delicious, golden delicious, and granny smith apples. The lowest cost was $2.49 per pound for several varieties of apple. The apples were well organized, bright colored, and appeared to be blemish and bruise free, medium sized bananas with green to yellow color were $0.29 per pound, with an organic option priced at $0.99 per pound. Navel oranges with some discoloration visible were priced at $2.29 per pound. Fresh orange carrots were priced at $0.99 per pound. Red Roma tomatoes with good color and no

visible blemishes were priced at $2.49 per pound. Several other tomato options such as heirloom, on the vine, and beefsteak were also available. Fresh broccoli with good color was priced at $1.99 per pound.

The 7-Eleven had a few red apples with slight blemishes for $0.99 each. The oranges were of low quality with obvious discoloration, and were priced at $0.99 each.

Many of the markets in tract 224020 had a limited selection of produce, usually composed of some combination of bananas, apples, and oranges for sale (Table 4). Much of it was of moderate quality and limited quantities. The three exceptions to this were the Super Farms, Numero Uno and WIC stores, all of which had a moderate to large selection of fresh produce (Table 4).

Several of these markets only sold a single item on the survey list. Joe's Market had a few medium bananas that were yellow to brown in color. Mexicali market had fresh medium bananas that were green to yellow in color for $0.79 per pound. Corona market had medium bananas that were yellow to brown in color for $0.69 per pound. Duran's had a small box of tomatoes that were on sale for $0.99 per pound.

Some of them sold small quantities of two or three items from the survey. Chayo Market sold medium sized yellow bananas for $0.69 per pound, and tomatoes for $0.89 per pound. 30th Street Market sold brown bananas for $0.65 per pound and fresh tomatoes in a cooler for $1.00 per pound. Young's Market had medium yellow bananas for $0.25 each, and oranges with limited blemishes for $0.50 each. El Tranquito had three small grocery bags of produce that included bananas for $1.00 each, oranges for $1.00 each, lettuce for $1.50, and onion and cilantro for $0.75 each. Reyes Market had a small box with apples, oranges and tomatoes in the cooler for $0.50 each. Daily Food had a small produce area that included medium sized green to

brown bananas for $0.69 per pound, moderate quality oranges for $0.59 per pound, and Roma tomatoes for $0.99 per pound. El Principio 1 and 2 stocked apples for $0.89 per pound, oranges for $0.59 per pound, carrots for $0.79 per bag, tomatoes for $0.79 per pound, and broccoli for $1.29 per pound

Both of the 99 Cents stores had a produce section; however, the location on Washington was significantly larger and better stocked. The Pico location had medium sized yellow bananas for $0.49 per pound. Bagged oranges with limited blemishes or bruising were priced at $0.99 per bag. Tomatoes that were green to red in color were priced at $0.99 per pound. All of the produce was in bins in a corner of the store. The Washington location had a larger produce section that was also not climate controlled. Medium sized green to yellow bananas were priced at $0.49 each. Oranges with bruises were priced at $0.99 per pound. Bagged baby carrots that appeared to be fresh were priced at $0.99 per bag. Tomatoes that were green to red in color were priced at $0.99 per pound.

All three of the 7-Eleven locations in this tract had red and green apples; however, they were all priced differently. The San Pedro Street location priced them at $0.99 each, Figueroa priced them at $0.89 each, and the Adams location at $0.69 each. All three locations also sold medium sized bananas that were green to yellow in color, again at different prices. San Pedro Street sold them for $0.50 per banana, Figueroa for $0.89 each, and Adams for $0.69 each. Finally, both Figueroa and Adams sold oranges, $0.89 on Figueroa and $0.69 on Adams.

The WIC location on Washington was notable because they had a moderate amount of produce in bins and coolers that was very fresh, clean, and unblemished. The staff noted that they received multiple shipments of fresh produce each week, and customers would often come to them for fresh produce. They sold fresh red apples for $2.75 per pound, green bananas for

$2.25 per pound, oranges for $3.50 per pound, carrots for $1.00 per pound, and tomatoes for $2.50 per pound.

The Super Farms was immediately notable because of the size and selection compared to other markets that I had visited in the tract (Table 4). There was a medium sized produce area with displays of fruits and vegetables that were clearly marked for variety and price. Greens were kept in a cooler and sprayed regularly to maintain freshness. They stocked Fuji, green, gala, red, and dorada apples. The Fuji apples were priced at $0.79 per pound and the dorados were $1.29 per pound. They sold medium sized green to yellow bananas for $0.50 per pound. Carrots were kept in the cooler and sold for $2.99 per pound. Fresh red tomatoes were sold for $0.89 per pound.

Similar to Super Farms, Numero Uno was a surprising find given that it was not present in either the Esri or social media data set. The Jefferson location was the larger of the two, with a large produce area with displays of fruits and vegetables, that the staff was regularly maintaining while I was there. They stocked gala, red delicious, gold delicious, granny smith, and Fuji apples. The red delicious and granny smith were priced at $0.69 per pound. The Fuji and golden delicious were priced at $1.29 per pound. The apples were clean and appeared to be bruise and blemish free. Medium sized green to yellow bananas were priced at $0.59 per pound. Oranges with some discoloration visible were priced at $1.00 for two pounds. Bagged carrots that were kept in a vegetable cooler were priced at $0.59 per bag. Fresh red tomatoes were priced at $0.99 per pound. Fresh broccoli that had been recently sprayed was priced at $0.99 per pound.

The results summarized in Tables 1-4 and accompanying commentary might have been able to be used to evaluate the oft-cited assumption that larger stores means better access to fresh dairy, fruit, vegetables, and meat. This was not possible because all of the stores in Census Tract

224020 were labeled as class A (0-2,500 ft2) in Esri's Business Analyst and the relatively small numbers of stores in Census Tract 460700 were spread across Class A (3 stores), Class B (3 stores), and Class D (1 store). The relative paucity and richness of choice in Census Tract 224020 and 460700, respectively point to the difficulty of drawing sweeping conclusions across very different geographies and the total numbers of stores in Classes B, C, and D were not sufficient to support such analysis in this case.

Chapter 5 Discussion and Conclusions

The social media data from both Google and Yelp produced inconsistent results, with all of the major businesses represented in both the social media data set and the Esri Business Analyst data set for tract 740600; however, the social media and Esri data sets only had two businesses in common in tract 224020. Moreover both data sets contained entries that were misclassified, closed, or absent. Finally, the APIs offered incomplete access to the social media data necessitating the use of other means.

5.1 Commercial Data

How well the commercial data represents reality is one of the research questions explored in this thesis. The results of the field survey suggest that the Esri Business Analyst data produces inconsistent results. The data set included businesses in both tracts categorized with NAICS code 44511003, super market or grocery store. Most of the businesses in the data set were classified with square footage code A or B (i.e. 1 – 2,499 square feet and 2,500 – 9,999 square feet, respectively), with the Ralphs in tract 740600 being the lone business classified in the D 40,000+ square footage class.

Both Sprouts and Gelsons in tract 740600 are classified as 44511003 with square footage class A. Field surveys indicate that both markets are well stocked with fresh foods. It is illustrative to compare these markets with businesses such as Tikal Market in tract 224020 because of the contrasting survey results. Tikal Market, also categorized with NAICS 44511003 and square footage class A, only stocked a few gallons of milk and Ezekial sprouted tortillas. Tikal Market is not an outlier in the data set, numerous markets in tract 224020 are classified as supermarkets or grocery stores; however, they stock very little fresh food. None of the markets in tract 224020, with the exception of Numero Uno and Super Farms, stocked all of the items

enumerated on the survey, which is an abbreviated version of the complete University of Pennsylvania survey. This suggests that analysis performed with this data set that excludes a field survey component could produce misleading results.

5.2 Social Media Data

The social media data from both Google and Yelp proved to be inconsistent and incomplete, with a majority of businesses in tract 224020 missing from the data set. Moreover, the Google data set was very sparse and excluded more businesses than the Yelp data set in both tracts. The Yelp data set was more complete in tract 460700: it included all of the major chain markets as well as several convenience stores. Furthermore, the Yelp data set for tract 460700 had significantly more reviews for fewer businesses.

5.2.1. Google Places

The Google Places API returned basic information about businesses in the data set; however, it did not enhance the data that is available in the commercial data set. Moreover, the data set contained a limited and incomplete set of businesses with limited reviews. The Google classification system is not more granular than the NAICS classification. Consequently, it is not possible to achieve better classification by replacing or augmenting the commercial data with Google data. Finally, the absence of so many businesses in the Google data set makes it difficult to justify the use of Google Places data as an adjunct to commercial data.

5.2.2. Yelp

Though more complete than the Google Places data, the Yelp data is also problematic in that it excluded some businesses and provided limited access to the review text. Only two of the businesses from the commercial data set for tract 224020 were included in the Yelp data set,

which is not sufficient to augment the commercial data set in a meaningful way. Furthermore, the API provides an excerpt from a single review, which does not allow for meaningful sentiment analysis to be performed. This thesis worked around the limitation by using a scraper; however, the use of a scraper would not be recommended as a best practice in food access studies.

5.2.3. Demographics

The results of this thesis suggest that there is a difference in participation in sites such as Google Places and Yelp that is based on community and neighborhood characteristics. There were significantly more reviews for fewer businesses in tract 740600, while a larger number of businesses in tract 224020 had fewer or lacked reviews all together. The data set for tract 224020 contained reviews for 10 businesses, many of them near the University of Southern California and the Los Angeles Convention Center. The cluster of businesses in the southeast quadrant of the tract were notably absent from the data set. In contrast, every major market in tract 740600 was included in the data set and had reviews. This suggests that the people in tract 740600, a wealthier tract with fewer minorities, are more engaged with social media. Moreover, the reviews for tract 740600 tended to be more positive, suggesting that people in the tract are more willing to engage social media for positive reasons as well as negative ones.

5.2.4. Sentiment Analysis

Six of the nine businesses in tract 224020 have an overall negative sentiment computed by the Google engine, while all but one have a positive sentiment in tract 740600. The IBM Watson sentiment was more challenging to interpret given the textual nature of the sentiment. Consequently, a ratio of the sum of positive over negative sentiment was produced in order to interpret the results. Tract 740600 exhibited results similar to those generated by Google, with

Rite-Aid being the sole business to provoke a net negative sentiment. Tract 224020, however, produced different results with seven businesses receiving more positive than negative reviews. This suggests that differences in the Google and IBM sentiment analysis methods can and do produce different and contradictory results.

The sentiment analysis performed in R, which parsed each review into words and assigned a per word sentiment, produced some descriptive statistics that suggest that reviews in tract 224020 are far more negative than those in 740600. The analysis suggests large increases in the use of words with negative sentiment such as anger, sadness, and disgust in the 224020 census tract as indicated by Figure 4 below.

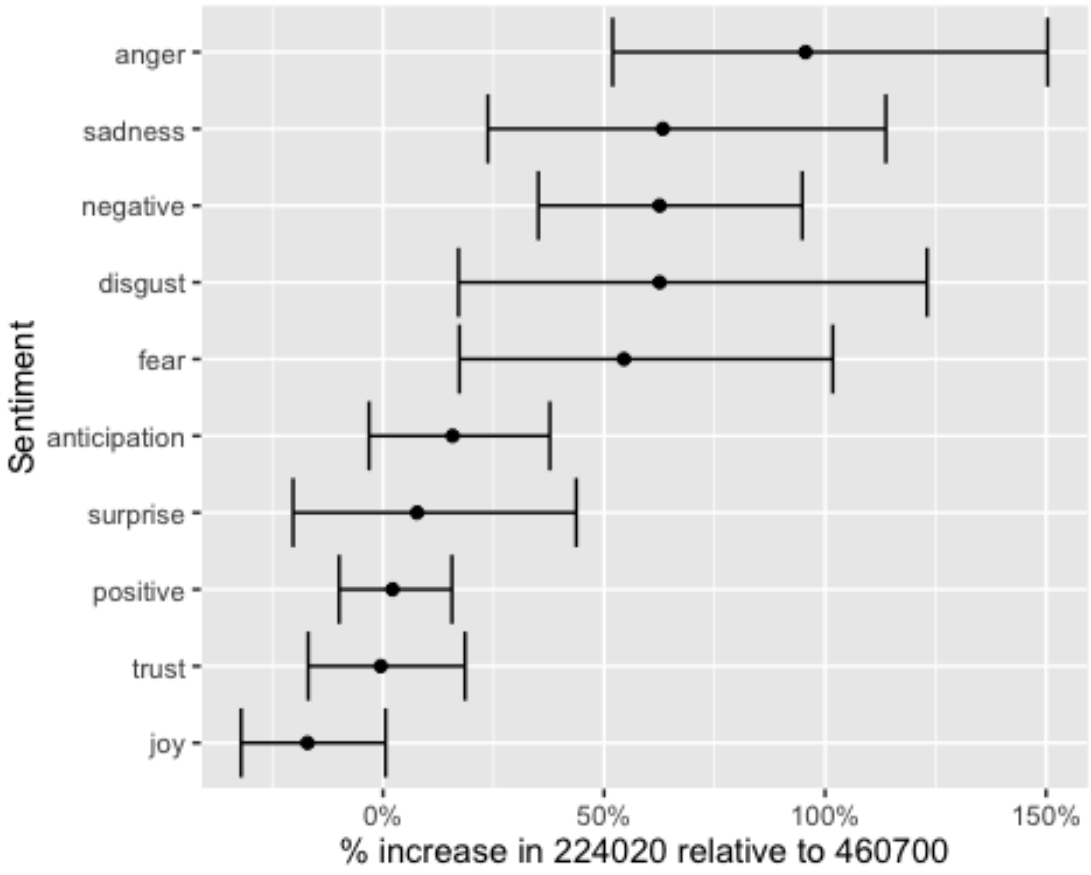


Figure 4 Graph comparing the sentiment analysis in R for the two census tracts

The results of the initial sentiment analysis in R motivated further analysis that considered 10 additional census tracts in the Los Angeles metropolitan area. The column statistics were used in ArcMap to determine the minimum and maximum scores for tracts in the Los Angeles metropolitan area. Tracts were then selected by attribute by increasing or decreasing the criteria and moving them away from the minimum or maximum until five additional tracts were identified for each extreme. Eight of the 10 tracts had social media data associated with them; however, only one of the high scoring tracts had businesses with reviews associated with it. Consequently, it was difficult to extend the sentiment analysis to a broader study area. Moreover, Yelp became aware of the scraper being used and instituted a recaptcha process meant to prevent this type of program from harvesting reviews. This further complicated the expansion of sentiment analysis to the larger area. This is discussed further under future work below.

5.3 In-field market surveys

The in-field market surveys revealed large variation in the markets that are classified with NAICS code 44511003 (supermarkets and grocery stores). There were notable differences in availability, choice, and price of fresh food in the markets that were surveyed.

5.3.1. Availability

Fresh food was available in both census tracts; however, it was far more abundant in census tract 460700. Many of the small markets in census tract 224020 had limited dairy, bread, meat, and produce options. Many of them had very small sections with very few items of moderate quality. The notable deviations from this trend were Numero Uno and Super Farms, which were in opposite corners of census tract 224020. Census tract 460700, in contrast, had numerous markets offering fresh food throughout the tract. Markets including Gelsons, Ralphs, Trader Joes, and Vons had an abundance of fresh food including produce, bread, dairy, and meat.

The aforementioned markets in census tract 460700 also had abundant parking further improving the availability of fresh foods in the tract.

5.3.2. Choice

The smaller square footage of the markets in tract 224020 was a limiting factor in the number of choices available to consumers. The smaller markets in this census tract contained markedly fewer choices than those in census tract 460700. The fresh produce areas were often a small corner of the store or a small section in a refrigerator. The fresh food in the markets of census tract 460700 was large and well curated, with displays of fresh fruits and vegetables that were automatically watered at regular intervals. The larger square footage of these markets allowed for large quantities of fresh food to be displayed, including multiple varieties of fruits, meats, breads, and dairy. This large selection included alternative choices such as organic and local products, choices that were notably absent in many of the markets in census tract 224020.

5.3.3. Price

The limited choices in census tract 224020 often came at higher price points. There were often low priced options in census tract 460700 in addition to the more expensive options, both of which were absent in census tract 224020. In fact, people shopping in census tract 224020 were often paying higher prices for the same or inferior products.

5.4 Limitations

There are several limitations that affect this thesis: the (1) relatively small data set; (2) limited access to reviews; (3) imperfect sentiment analysis models; (4) abbreviated surveys; and (5) the incomplete data. Additional cities and census tracts should be investigated in order to determine whether or not the results in this thesis with regard to classification and data quality

can be generalized. It would be interesting and useful to identify how many of the businesses in the commercial data set are misclassified or misleading.

Moreover, it would be interesting and instructive to perform a more thorough survey of markets in a larger data set. The results presented in this thesis suggest that there are large differences in markets that are classified in the same way. Assumptions and conclusions made in studies may need to be altered if the patterns identified in this thesis hold; however, the data set and surveys are not large enough to support any change in methods.

The sentiment analysis models used in this study are not specifically trained to identify healthy food options. A specifically trained model could potentially improve the results produced by this study by identifying healthy choices instead of the more general sentiment.

Finally, this thesis makes use of a limited number of surveys and social media data points. It would benefit from having a larger, more complete data set. This could potentially be achieved by using the Yelp data set that has been released for research. That data set was not used for this thesis because of the challenge it presented with respect to ground truthing.

5.5 Suggestions for Future Work

Future work should focus on larger review data sets that span all four store square footage classes, supervised sentiment models that are specifically trained for healthy food, and additional field work. Future work in this area should consider a much larger data set because the results in this thesis suggest that there are problems with both the commercial and social media data sets. These problems include missing, closed, and misclassified businesses that should be further investigated. This is a large task that will require a substantial investment in field surveys in order to identify additional problems in the data set. The reward would be substantial and it

could offer the opportunity to validate or repudiate often-held assumptions such as the belief that larger stores mean more choice and greater access to healthy foods.

Additional sentiment analysis can be performed with a supervised model that is specifically trained to identify healthy food. This is important because traditional sentiment could assign a positive sentiment to unhealthy choices such as fried or sweet foods. The results presented in this thesis do not clearly suggest that sentiment analysis of reviews can be a useful adjunct to the commercial data set because of missing data and variation in the number and quality of reviews.

5.6 Conclusions

The commercial and social media data sets remain problematic. Consequently, studies performed using those data sets can potentially be misleading. Additional work is required to validate and update the commercial data set. Moreover, there are large problems with the completeness of the social media data set. This thesis has demonstrated that inconsistencies in the data could generate misleading results in food access studies.

REFERENCES

An, Ruopeng, and Roland Sturm. 2012. "School and residential neighborhood food environment and diet among California youth." *American Journal of Preventive Medicine 42 (2):* 129-35.

Ball, Kylie, Anna Timperio, and David Crawford. 2009. "Neighborhood socioeconomic inequalities in food access and affordability." *Health and Place 15:* 578 - 585.

Charreire, Hélène, Romain Casey, Paul Salze, Chantal Simon, Basile Chaix, Arnaud Banos, Dominique Badariotti, Christiane Weber, Jean-Michel Oppert. 2010. "Measuring the food environment using geographical information systems: a methodological review." *Public Health and Nutrition* 13 (11): 1773–1785.

Cummins, Steven, Ellen Flint, and Stephen A Matthews. 2014. "New neighborhood grocery store increased awareness of food access but did not alter dietary habits or obesity." *Health Affairs 33 (2):* 283-91.

Cummins, Steven, Sally Macintyre. 2002. "“Food deserts”—evidence and assumption in health policy making." (BMJ) 325: 436 - 438.

Gard, Julienne. n.d. *A spatial accessibility analysis of the Los Angeles foodscape (PhD diss, University of Southern California, 2015).*

Kerski, Joseph J., Jill Clark. 2012. *The GIS Guide to Public Domain Data.* Redlands, CA: Esei Press.

Lee, Helen. 2012. "The role of local food availability in explaining obesity risk among young school-aged children." *Social Science and Medicine 74 (8):* 1193-203.

Moore, Latetia V and Ana V. Diex Roux. 2006. "Associations of Neighborhood Characteristics With the Location and Type of Food Stores." *Research and Practice 96 (2):* 325 - 331.

Morganstern, Seth. 2015. *Disparities in Food Access: An Empirical Analysis of Neighborhoods in the Atlanta Metropolitan Statistical Area (MS thesis. University of Southern California, 2015).*

Powell, Lisa M, Sandy Slater, Donka Mirtcheva, Yanjun Bao, and Frank J. Chaloupka. 2007. "Food store availability and neighborhood characteristics in the United States." *Preventive Medicine 44* 189 - 195.

Shier, Victoria, Rong An, and Robert Sturm. 2012. "Is there a robust relationship between neighbourhood food environment and childhood obesity in the USA?" *Public Health 126 (9):* 723-30.

Smith, Aaron and Joanna Brenner. 2012. "Twittter Use 2012." *Pew Internet & American Life Project.* May 31. Accessed November 2016. http://dougleschan.com/the-recruitment-guru/wp-content/uploads/2014/01/Twitter-Use-2012-Pew-Internet-American-Life-Project.pdf.

Ver Ploeg, Michael, Vince Breneman, Paula Dutko, Ryan Williams, Samantha Snyder, Chris Dicken, and Phil Kaufman. 2012. "Access to Affordable and Nutritious Food." *United States Department of Agriculture.* November. Accessed November 2016. https://www.ers.usda.gov/webdocs/publications/err143/33845_err143.pdf.

Walker, Renee E., Christopher R. Keane, Jessica G. Burke. 2010. *Disparities and access to healthy food in the United States: A review of food deserts literature.* Vol. 16. Health and Place.

Widener, Michael J, Wenwen Li. 2014. "Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US." *Applied Geography* 54: 189-197.

Zick, Cathleen D., Ken R. Smith, Jessie X. Fan, Barbara B. Brown, Ikuho Yamada, and Lori Kowaleski-Jones. 2009. "Running to the store? the relationship between neighborhood environments and the risk of obesity." *Social Science & Medicine 69 (10):* 1493-500.

Appendix A: Code Used to Retrieve Yelp and Google Data

```
#-------------------------------------------------------------------
----------
# Name:        Google Places API Nearby Request
# Purpose:      USC SSI Master's Thesis Project
#                   Food access in Los Angeles
# Author:      CMH
#
# Created:     06/03/2016
# Copyright:   (c) CMH 2016
# Licence:     Attribution-NonCommercial-ShareAlike 4.0 International
#-------------------------------------------------------------------
----------

# Import all of the libraries
import httplib
import urllib
import json
import pprint

#Yelp Python Client Support
from yelp.client import Client
from yelp.oauth1_authenticator import Oauth1Authenticator

auth = Oauth1Authenticator(
    consumer_key='NLz9tZJ465VfGE1ZF4wSvw',
    consumer_secret='hPmTpTwFLNskvTNO-lu4HUNskMo',
    token='HLBajJ7AO9-EVkz-eRiaD0DDfFz4kE1q',
    token_secret='RLEhQPqvobQ9umvCKiCN_wJxFHc'
)

yelpClient = Client(auth)


def yelpit( lat, lng, types, offset ):
    #This funciton pulls business data from the Yelp API

    params = {
        'term': 'food',
        'radius_filter': '1500',
        'offset': offset,
        'sort': '0',
        'category_filter': types
    }
    response =  yelpClient.search_by_coordinates(lat,lng,**params)
```

```
    #print response.total
    for key in response.businesses:
        print key.location.coordinate.latitude, ",",
        print key.location.coordinate.longitude, ",",
        print key.name, ",",
        print key.categories[0][1], ",",
        print key.rating, ",",
        print key.review_count

    offset += 20
    #print offset, "\n"
    if offset < response.total:
        yelpit(lat, lng, types, offset)

def deets(bus_id):
    conn = httplib.HTTPSConnection("maps.googleapis.com")
    APIkey = "AIzaSyB864Llir0-1NrFaQ1yr3TIzG9fB09IP7c"
    reqstring = "/maps/api/place/details/json?placeid=" + bus_id + "&k
ey=" + APIkey
    #print reqstring
    conn.request("GET", reqstring)
    response = conn.getresponse()
    #print response.status, response.reason
    if response.status == 200:

        # Get and print the actual data
        data = response.read()

        # parse the json into a more useful data structure
        parsed_json = json.loads(data)
        #pp = pprint.PrettyPrinter(indent=4)
        #pp.pprint(parsed_json)
        if "rating" in data:
            print parsed_json['result']['rating'], ",",
        else:
            print "0,",
        if "user_ratings_total" in data:
            print parsed_json['result']['user_ratings_total']
        else:
            print "0"

    conn.close()

def googit(lat, lng, types, next_page_token = None):
    conn = httplib.HTTPSConnection("maps.googleapis.com")
```

```
    #headers = {"":""}
    #Lat Long for center of census tract
    #lat = "34.0465960"
    #lng = "-118.2515835"

    radius = "1500"
    #types = "restaurant"
    #types = "grocery_or_supermarket"
    APIkey = "AIzaSyB864Llir0-1NrFaQ1yr3TIzG9fB09IP7c"

    if next_page_token is None:
        conn.request("GET", "/maps/api/place/nearbysearch/json?locatio
n=" + lat + "," + lng + "&radius=" + radius + "&types=" + types + "&ke
y=" + APIkey)
    else:
        conn.request("GET", "/maps/api/place/nearbysearch/json?locatio
n=" + lat + "," + lng + "&radius=" + radius + "&key=" + APIkey + "&pag
etoken=" + next_page_token)

    # Get the response and print the response information eg. 200 OK o
r 404 Not Found
    response = conn.getresponse()
    #print response.status, response.reason

    if response.status == 200:

        # Get and print the actual data
        data = response.read()

        # parse the json into a more useful data structure
        parsed_json = json.loads(data)

        # Load the pretty printer so that we can better see the struct
ure of the data
        #pp = pprint.PrettyPrinter(indent=4)
        #pp.pprint(parsed_json)
        #pp.pprint(parsed_json['pagination'])
        #pp.pprint(parsed_json['meta'])
        #pp.pprint(parsed_json['results'])
        #pp.pprint(data)
        #json.dumps( parsed_json, sort_keys=True, indent=4, separators
=(',', ': ') )

        items = parsed_json['results']
        for item in items:
            #theLine = ""
```

```
            try:

                #print item['place_id']
                #pp.pprint(item)
                print item['geometry']['location']['lat'], ",",
                #theLine = theLine + item['location']['latitude'] + ",
"
                print item['geometry']['location']['lng'], ",",
                #theLine = theLine + item['location']['longitude'] + "
,"
                print item['name'], ",",
                print item['types'][0], ",",
                deets(item['place_id'])
                #print ""
                #theLine = theLine + item['link'] + ","
                #print item['images']['standard_resolution']['url']
                #theLine = item['location']['latitude']
                #theLine = theLine + item['location']['latitude'] + ",
" #+ item['location']['longitude'] + "," + item['link'] + "," + item['
images']['standard_resolution']['url'] + "\n"
                #print theLine
                #print (",")
                #print "\n"
                #f.write(theLine)
            except TypeError:
                print ",type error"
                pass
            except KeyError:
                print ",key error"
                pass
        # Close the connection
        #f.close()
        conn.close()

        if "next_page_token" in data:
            #print "recurse"
            #print parsed_json['next_page_token']
            googit(lat,lng,types,parsed_json['next_page_token'])

print "lat, long, name, type, rating, review_count"
#googit("34.1929284","-118.1988009","grocery_or_supermarket");
#yelpit("34.1929284","-118.1988009","grocery,convenience",0);

googit("34.0304827","-118.2686569","grocery_or_supermarket");
#yelpit("34.0304827","-118.2686569","grocery,convenience",0);
```

Appendix B: Food Outlet Survey Sheet

Store ID: _____

Store Name: _____

Store Location: _____

- ☐ Grocery Store
- ☐ Convenience Store

Date: _____

Comments: _____

_____

Dairy

- Skim Milk
  - Brands available: _____
  - Lowest price quart w/ brand: _____
  - Highest price quart w/ brand: _____
  - Lowest price half gallon w/ brand: _____
  - Highest price half gallon w/ brand: _____
  - Lowest price gallon w/ brand: _____
  - Highest price gallon w/ brand: _____
- Whole Milk
  - Brands available: _____
  - Lowest price quart w/ brand: _____
  - Highest price quart w/ brand: _____
  - Lowest price half gallon w/ brand: _____
  - Highest price half gallon w/ brand: _____
  - Lowest price gallon w/ brand: _____
  - Highest price gallon w/ brand: _____

Comments: _____

Bread

- Whole wheat brands available: _____
  - Low price loaf w/ brand: _____
  - High price load w/ brand: _____
- White brands available: _____
  - Low price load w/ brand: _____
  - High price load w/ brand: _____

Comments: _____

Meat

- Ground Beef fat percentages available: _____
  - Price per pound for 90% lean: _____
  - Price per pound for 80% lean: _____
- Ground turkey price per pound: _____
- Ground chicken price per pound: _____

Comments: _____

Produce

- Apple varieties: _____
  - Lowest price apple per pound _____
  - Quality of apples: _____
- Banana price per pound: _____
- Banana quality / color: _____
- Orange price per pound: _____
- Orange quality: _____
- Carrots cost per pound: _____
- Carrots quality: _____
- Tomatoes cost per pound: _____
- Tomatoes quality: _____
- Broccoli cost per pound: _____
- Broccoli quality: _____

Appendix C: Python Scraper

```
#-------------------------------------------------------------------------------
# Name:        module2
# Purpose:
#
# Author:      CMH
#
# Created:     22/08/2016
# Copyright:   (c) CMH 2016
# Licence:     Attribution-NonCommercial-ShareAlike 4.0 International
#-------------------------------------------------------------------------------

# Main libraries
import urllib
import urllib2
import httplib
import json

# Google Cloud API for Sentiment Analysis
import argparse
from googleapiclient import discovery
import httplib2
import json
from oauth2client.client import GoogleCredentials

def get_watson(review_text):
    '''Get Sentiment analysis from IBM Watson '''
    WatsonKey = 'caa1116164b2d90fab5e8a6af28b1447f3b40a87'
    conn = httplib.HTTPSConnection("gateway-a.watsonplatform.net")
    reqstring = "/calls/text/TextGetTextSentiment?apikey=" + WatsonKey +
"&outputMode=json&text=" + urllib.quote(review_text)
    #print reqstring
    conn.request("GET", reqstring)
    response = conn.getresponse()
    #print response.status, response.reason
    if response.status == 200:

        # Get and print the actual data
        data = response.read()

        # parse the json into a more useful data structure
        parsed_json = json.loads(data)
        #pp = pprint.PrettyPrinter(indent=4)
        #pp.pprint(parsed_json)
        #print parsed_json['docSentiment']['score'] + ':' + parsed_json['docSentiment']['type']
        if "score" in data:
```

```
        return parsed_json['docSentiment']['type'] + '|' + parsed_json['docSentiment']['score']
    else:
        return "null|null"

def get_sentiment(review_text):
  '''Run a sentiment analysis request on review text'''

  http = httplib2.Http()

  credentials =
GoogleCredentials.get_application_default().create_scoped(['https://www.googleapis.com/auth/c
loud-platform'])
  http=httplib2.Http()
  credentials.authorize(http)

  DISCOVERY_URL = "https://language.googleapis.com/$discovery/rest?version=v1beta1"
  service = discovery.build('language', 'v1beta1', http=http,
discoveryServiceUrl=DISCOVERY_URL)
  service_request = service.documents().analyzeSentiment(
  body={
   'document': {
     'type': 'PLAIN_TEXT',
     'content': review_text,
   }
  })
  response = service_request.execute()
  polarity = response['documentSentiment']['polarity']
  magnitude = response['documentSentiment']['magnitude']
  #print('Sentiment: polarity of %s with magnitude of %s' % (polarity, magnitude))
  #print polarity * magnitude
  return str(polarity) + '|' + str(magnitude)

def scraper(bus_id, tract, group, offset = 0):
  url = "https://www.yelp.com/biz/" + bus_id
  #print url
  filename =
'C:/Users/CMH/Desktop/USC_GIST/SSCI594b/ThesisDocs/Thesis/addtl_social_reviews.csv'
  #filename = 'C:/Users/CMH/Desktop/'
  target = open(filename, 'w')
  offset = int(offset)
  if(offset != 0):
    parts = url.split('?')
    url = parts[0] + "?start=" + str(offset)
    #print 'url: ' , url
  #Query the website and return the html to the variable 'page'
  page = urllib2.urlopen(url)
```

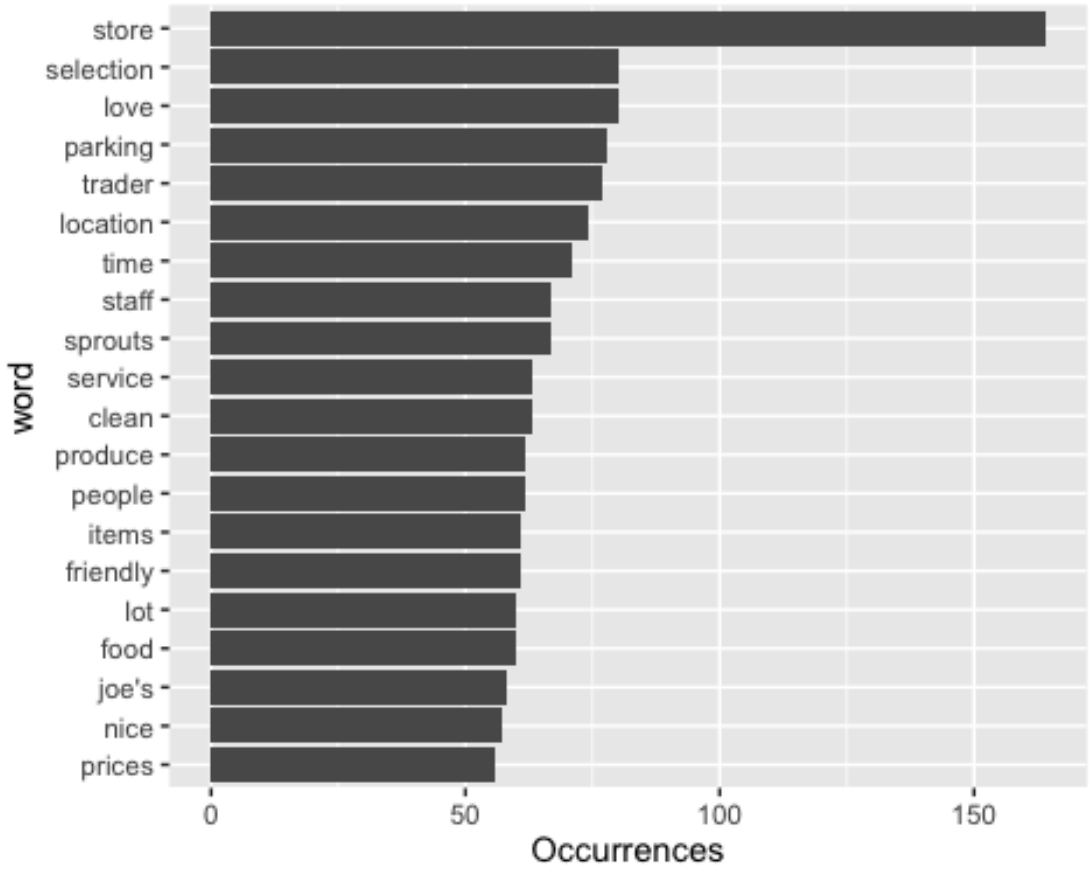
```
    #import the Beautiful soup functions to parse the data returned from the website
    from bs4 import BeautifulSoup

    #Parse the html in the 'page' variable, and store it in Beautiful Soup format
    soup = BeautifulSoup(page, "html.parser")

    next = 0

    total = soup.findAll("span",  {"itemprop" : "reviewCount"})
    for t in total:
        next = unicode(t.string)
        next = int(next)
        #print next

    all_p = soup.findAll("p", {"itemprop" : "description"})
    #print all_p
    for p in all_p:
        r_text = str(p)
        r_text.replace('<p itemprop=description lang=en>',")
        r_text.replace('<br>',")
        r_text.replace('</br>',")
        r_text.replace('</p>',")
        #watson = get_watson(r_text)
        #google = get_sentiment(r_text)
        the_line = bus_id + '|Yelp|' + r_text +'|' + '0|0' + '|' + '0|0' + '|' + tract + '|' + group
        #the_line = bus_id + '|Yelp|' + r_text +'|' + str(watson) + '|' + str(google) + '|' + tract + '|' +
group
        target.write(the_line)
        target.write('\n')
        print the_line
    target.close()

    #Recursively call the function to move on to the next page
    offset += 20
    if(offset < next):
        #print 'offset: ' , offset
        scraper(bus_id,tract,group,offset)


#specify the url
#business = "https://www.yelp.com/biz/7-eleven-los-angeles-50"
#business = "7-eleven-los-angeles-50"
#scraper(business)
#scraper("gelsons-market-la-cañada-flintridge-3")
#scraper("sprouts-farmers-market-la-canada-flintridge-2")
```

```
#scraper("trader-joes-la-canada")
#scraper("ralphs-la-canada-flintridge")
#scraper("7-eleven-la-canada-flintridge-2")
#scraper("rite-aid-la-canada-flintridge")
#scraper("7-eleven-la-canada-flintridge-3")

#scraper("joes-la-market-and-deli-los-angeles")
#scraper("trimana-express-los-angeles-2")
#scraper("cal-mart-beer-and-wine-food-store-los-angeles")
#scraper("7-eleven-los-angeles-122")
#scraper("7-eleven-los-angeles-50")
#scraper("rite-aid-los-angeles-50")
#scraper("el-tronquito-mkt-los-angeles")
#scraper("bembis-market-los-angeles")
#scraper("7-eleven-los-angeles-142",1,1)
#scraper("super-farms-market-los-angeles")
#filename =
'C:/Users/CMH/Desktop/USC_GIST/SSCI594b/ThesisDocs/Thesis/social_reviews.csv'
#filename = 'C:/Users/CMH/Desktop/addtlReviews.csv'
filename = 'E:\Program Files\RStudio\AdditionalBus.csv'
#bus_list = open(filename, 'w')

import csv
with open(filename) as csvfile:
    reader = csv.DictReader(csvfile)
    for row in reader:
        #print row
        id = row['id']
        id = id.strip()
        tract = row['tract']
        group = row['group']
        scraper(id,tract,group)
        #print id
        #print(row['id'].strip(), row['tract'], row['group'])
```

Appendix D: R Code

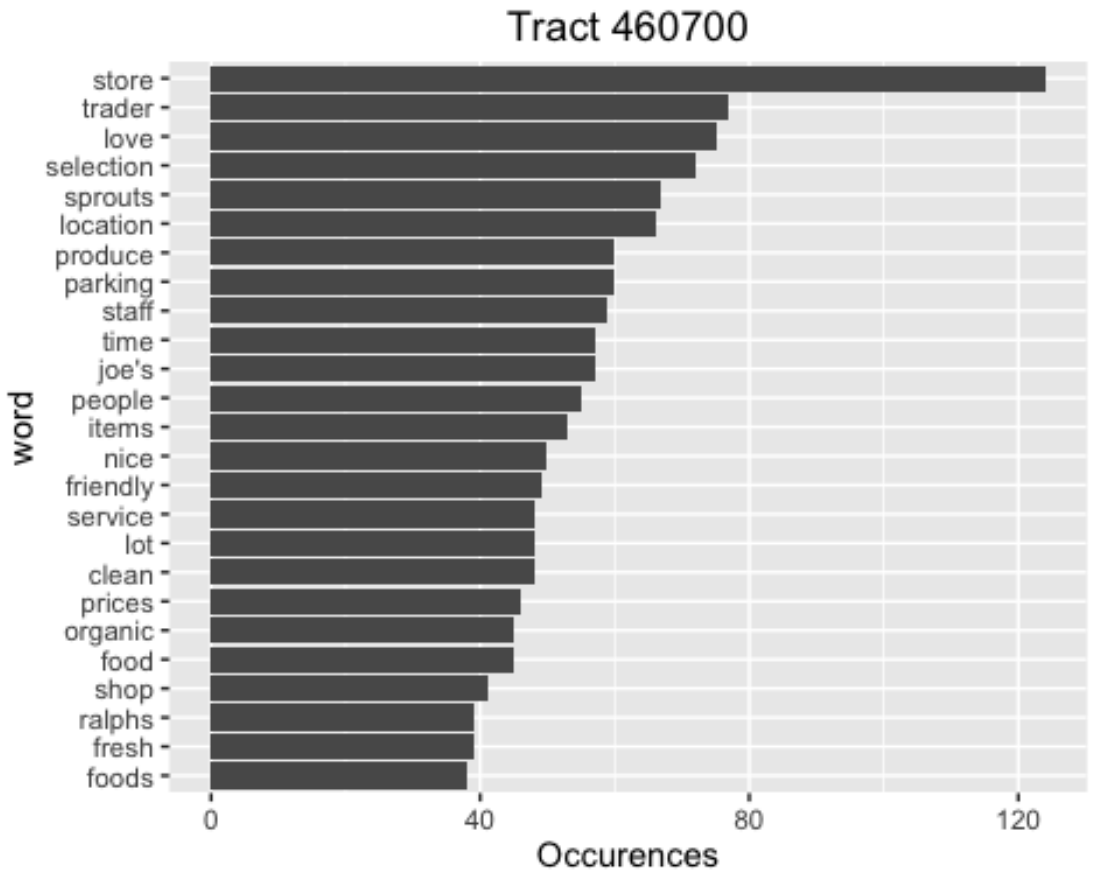
```
library(tidytext)

## Warning: package 'tidytext' was built under R version 3.2.5

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.2.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.2.5

library(stringr)
library(ggplot2)

load("Desktop/all_reviews.rda")

reviews$review_text <- gsub("<br>","", unlist(reviews$review_text))
reviews$review_text <- gsub("</br>","",unlist(reviews$review_text))
reviews$review_text <- gsub("</p>","",unlist(reviews$review_text))
reviews$review_text <- gsub("<p itemprop=\"description\" lang=\"en\">"
,"",unlist(reviews$review_text))
reviews$tract <- as.factor(reviews$tract)

reg <- "([^A-Za-z\\d#@']|'(?![A-Za-z\\d#@]))"
review_words <- reviews %>%
  unnest_tokens(word, review_text, token = "regex", pattern = reg) %>%
  filter(!word %in% stop_words$word,
         str_detect(word, "[a-z]"))

review_words %>%
  count(word, sort = TRUE) %>%
  head(20) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_bar(stat = "identity") +
```

```
  ylab("Occurrences") +
  coord_flip()
```

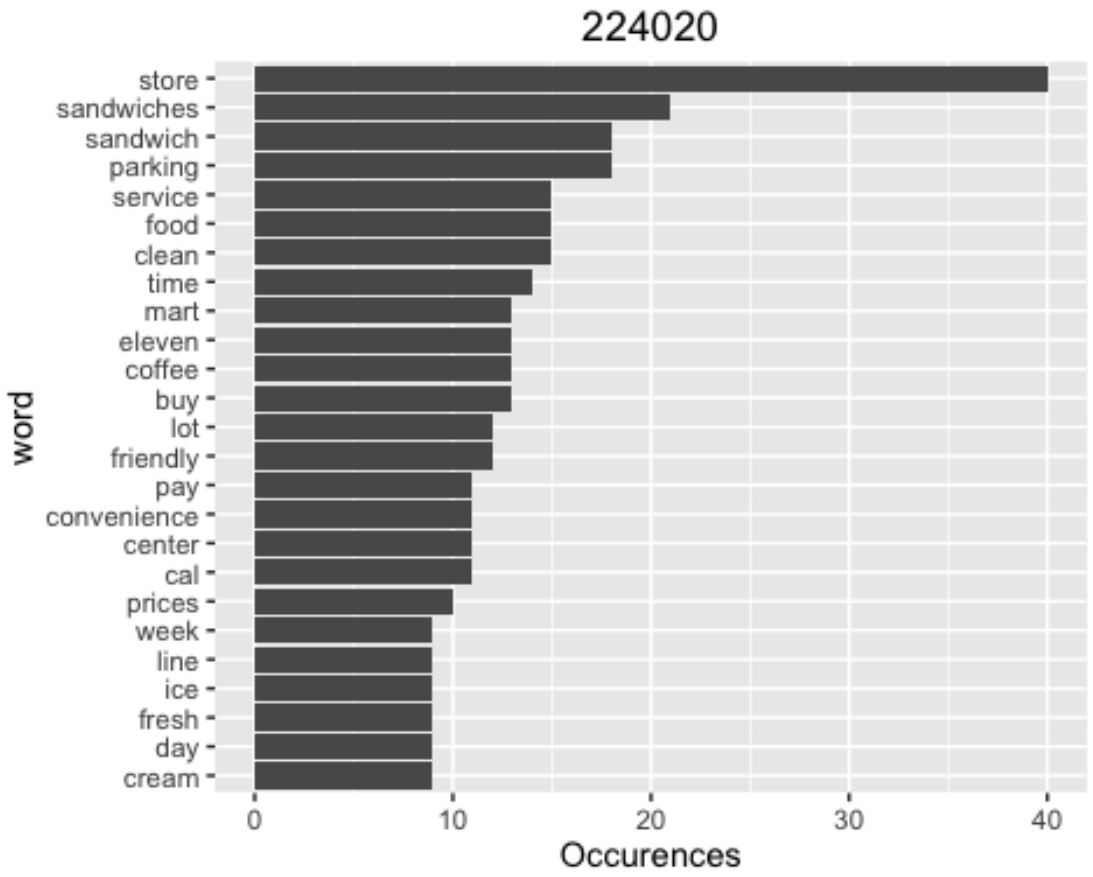


The reviews were subsetted by tract in order to obtain word counts of the top 20 words per tract which were then visualized.

```
library(tidytext)
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)
LC <- subset(review_words, tract=='460700', select = bus_id:word)
LC$tract <- as.factor(LC$tract)
LC %>%
  count(word, sort = TRUE) %>%
  head(25) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_bar(stat = "identity") +
  ylab("Occurences") +
  coord_flip() +
  ggtitle("Tract 460700")
```




```
LA <- subset(review_words, tract=='224020', select = bus_id:word)
LA$tract <- as.factor(LA$tract)
LA %>%
  count(word, sort = TRUE) %>%
  head(25) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_bar(stat = "identity") +
  ylab("Occurences") +
  coord_flip() +
  ggtitle("224020")
```



Sentiment analysis with the NRC lexicon was subsequently performed on each tract. Per tract sentiment was determined by generating a data frame with columns for the tract, sentiment, total number of words, and words that evoke the given sentiment.

```
library(tidytext)
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)

nrc <- sentiments %>%
  filter(lexicon == "nrc") %>%
  dplyr::select(word, sentiment)

tracts <- review_words %>%
  group_by(tract) %>%
  mutate(total_words = n()) %>%
  ungroup() %>%
  distinct(bus_id, tract, total_words)

tracts$tract <- as.factor(tracts$tract)
```

```
by_tract_sentiment <- review_words %>%
  inner_join(nrc, by = "word") %>%
  count(sentiment, bus_id) %>%
  ungroup() %>%
  complete(sentiment, bus_id, fill = list(n = 0)) %>%
  inner_join(tracts) %>%
  group_by(tract, sentiment, total_words) %>%
  summarize(words = sum(n)) %>%
  ungroup()
```

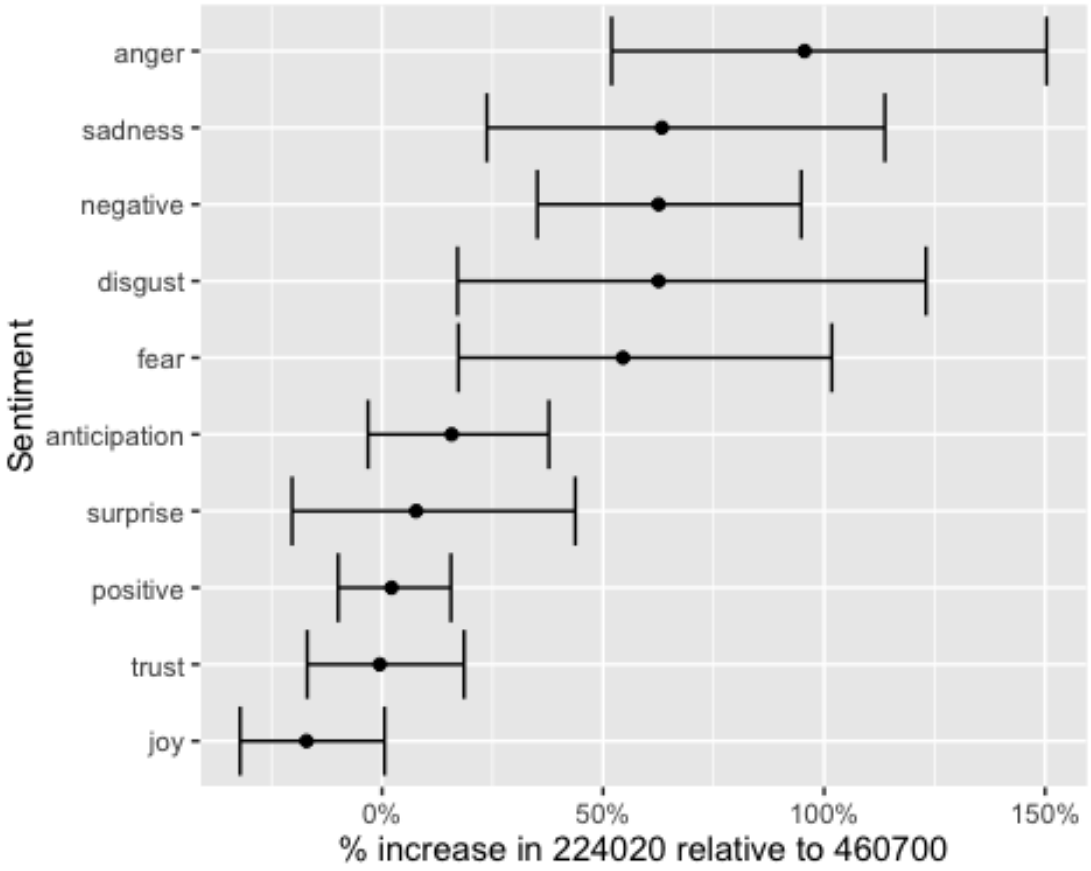
```
## Joining, by = "bus_id"
```

Finally, the difference between the tract sentiments was visualized.

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.2.5
```

```
sentiment_differences <- by_tract_sentiment %>%
  group_by(sentiment) %>%
  do(tidy(poisson.test(.$words, .$total_words)))

library(scales)

sentiment_differences %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, estimate)) %>%
  mutate_each(funs(. - 1), estimate, conf.low, conf.high) %>%
  ggplot(aes(estimate, sentiment)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  scale_x_continuous(labels = percent_format()) +
  labs(x = "% increase in 224020 relative to 460700",
       y = "Sentiment")
```


anger
sadness
negative
disgust
fear
anticipation
surprise
positive
trust
joy
Sentiment
0%
50%
100%
150%
% increase in 224020 relative to 460700