Background

As described in Saelens et al. (2018)[1], studies have shown that greater accessibility to supermarkets is associated with healthier dietary intake and lower body weight, whereas greater accessibility to fast-food restaurants and convenience stores is associated with less healthy dietary intake and higher body weight. To understand the impact of food outlet accessibility on weight loss trials, the ADOPT working group recommended measuring the density of outlets within a given area around the participant's home and distance to the closest outlet.

Jones, et al. (2017)[2] (https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-016-2355-1) describes data improvement methods for commercially obtained food outlet data developed based on existing validation studies. Their process includes purchasing records from commercial business lists (InfoUSA®, now called Data Axle USA® and Dun and Bradstreet) based on store/restaurant names as well as standard industrial classification (SIC) codes, reclassifying records by store type, improving geographic accuracy of records, and de-duplicating records. This study provides a step-by-step approach to purchase and process business list data obtained from commercial vendors.

Obtaining and Cleaning the Data

In order to provide an example that illustrates tract-level density of grocery stores, fast food restaurants, and convenience stores, we followed the process outlined in Jones, et al. and obtained historical commercial business listings for the specified SIC codes and chain names from the year 2019 for the state of Colorado from InfoUSA® and Dun and Bradstreet. After receiving the data, we reviewed the lists to ensure that all of the desired SIC codes and chain names were included in the list. If a particular chain was missing and was known to have locations in the designated area, we contacted the vendor to provide the additional listings. The data we received from InfoUSA included a variety of variables including company name and address, contact information, employee size, sales volume, SIC code(s), FIPS code, Latitude, Longitude, Census Block, Census Tract and others. This data also included a match code, an indication of geocode accuracy. The Dun and Bradstreet data set included business name and address, state and county codes, employee size, sales volume, SIC code(s), DUNS number and others. The Dun and Bradstreet data did not include any geocoding information. After receiving the data, we conducted an additional batch geocoding on the two datasets, followed by iterative (manual) geocoding to improve the overall geocoding of the address data. The census tract ID of the food outlet was added to each resulting dataset based on the geocodes. Data were cleaned using the de-duplication and re-classification process described by Jones, et al. (2018)[3] (https://neighborhoodsandhealth.uic.edu/wp-content/uploads/sites/103/2018/04/Commercial-

---

[1] Saelens BE, Arteaga SS, Berrigan D, Ballard RM, Gorin AA, Powell-Wiley TM, Pratt C, Reedy J, Zenk SN. Accumulating Data to Optimally Predict Obesity Treatment (ADOPT) core measures: environmental domain. Obesity 2018; 26(suppl 2):S35–S44. doi: 10.1002/oby.22159

[2] Jones KK, Zenk SN, Tarlov E, et al. A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments. BMC Res Notes 10, 35 (2017). https://doi.org/10.1186/s13104-016-2355-1

[3] Jones K, Xiang W, Matthews SA, Zenk SN. (2018). Accessibility of businesses: Weight and Veterans' Environments Study GIS protocol, Version 1. Retrieved from Weight and Veterans' Environments Study website: https://waves.uic.edu/.

Business-Listings-Documentation.pdf). This protocol describes the process through which commercial business list data was processed and includes measure definitions, bias assessments, deduplication, and a section describing how the processed business list data were used to create national raster surfaces showing neighborhood environment measures.

To produce our datasets, we first converted the STATA code provided by Jones, et al. to SAS. The STATA code contained multiple imbedded lists of food outlet names. For ease of future modification, we moved these lists to external lookup tables (Excel) that are read by SAS code. Researchers wishing to apply this process to their study area might want to review these lists of retailer names to be sure they are applicable and appropriate for their study area. The SAS programs and lookup tables are provided.

Following this step, it was necessary to review the lists for multiple listings for the same business at each address. In each spreadsheet, duplicates in the geocoded address column were highlighted and reviewed manually. If there were multiple listings at an address we searched in google maps to identify whether multiple businesses shared that address (e.g. food court at a mall). If both listings referred to the same restaurant or store (e.g. John Smith Pizza Restaurants LLC and Dominos Pizza) where one listing is the actual storefront and the other noted the management/ownership information, we kept the listing with the restaurant name. Table 1 shows the counts per food outlet type before and after processing.

Table 1. Store and restaurant counts before and after processing, overall and by store type, 2019

	Supermarkets and grocery stores	Convenience stores	Pharmacies	Liquor stores	General merchandise stores	Limited service restaurants
Provisionally classified by SIC code	907	1,271	631	1,342	117	3,708
Reclassified by name	907	1,271	629	1,342	117	3,704
After cleaning for locational accuracy	889	1,255	622	1,337	117	3,697
After deduplication by name	871	1,005	622	1,307	NA	NA
After deduplication by address	871	1,005	621	1,307	NA	NA
Final count after cleaning and deduplication (excluding AK, HI)	871	1,005	621	1,307	117	3,697
Final count after manual cleaning and deduplication	850	949	610	1,297	104	3,568

Creating the Density Variables

To create density variables by census tract, we needed land area of the tract as well as the number of specific food outlet types per tract.

Land Area

We obtained land area data from the Census Bureau website for Colorado counties and tracts. To produce this example:

- Go to the Census Geography TIGER Line shapefiles web page https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
- Pick the year that you want – we used 2019 to match the food outlet data
- Click "Download" using the "Web Interface"
- Download the Tract zip file for Colorado
- Download the County zip file (only available for the whole US)
- Unzip both files
- Save a copy of the DBF file from within each of the unzipped folders as an Excel file
- For the county file, delete the rows for states other than Colorado (State FIPS = 08)
- Clean up the formatting as needed
- Delete the extra columns – keep just the basic IDs, the name, and the ALAND (land area in square meters) and AWATER (water area in square meters) columns. We will just use ALAND but it is good practice to keep AWATER for reference.
- Add columns to calculate the land area in square kilometers and square miles
    - LandArea_SqKm = ALAND / 1,000,000
    - LandArea_SqMi = ALAND / 2,589,988

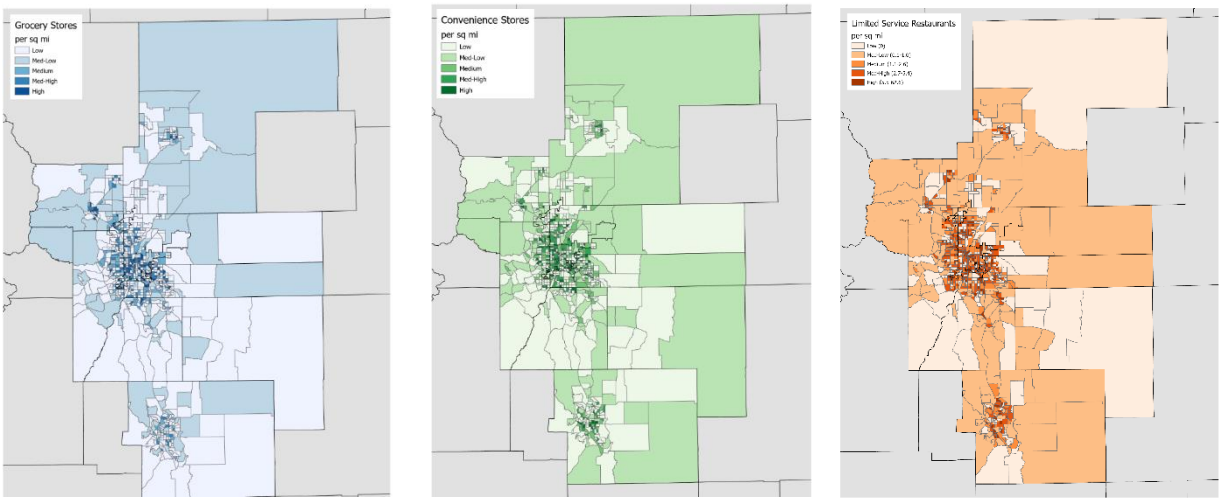
Food Outlet Types Per Tract

For each dataset (supermarketsgrocery, conveniencestores, ff_class (fast food outlets / limited service restaurants)), calculate the number per tract.

- Create a frequency table for each dataset using the primary SIC code and Tract ID variable.
- Merge the total number of food outlets per tract (the combined number of all Primary SIC codes) with the land area (square miles) per tract into one file.
- Calculate density variables by taking the total number of food outlets per tract and dividing by the land area in square miles per tract.

In Figure 1, we illustrate density of grocery stores, convenience stores, and limited service restaurants by tract for the eleven Colorado counties in which study participants reside. This illustration is not intended as an endorsement or based on a quantitative assessment of the data resources they or other companies use.

Figure 1. Density of grocery stores, convenience stores, and limited service restaurants per square mile



Calculating Other Measures

The ADOPT working group recommended measuring the density of outlets within a given area around the participant's home and distance to the closest outlet. We have shown how to calculate density of outlets by census tract. These values could be linked to census tract IDs for participant home locations. If preferred, researchers could make similar density calculations based on buffers around each specific participant's home location. Likewise, a list of participants' home locations could be compared to the list of food outlet locations to determine the distance to the nearest outlet of each type.

How to obtain data:

Data Axle USA® (Formerly Info USA®)
Jeff Jones
VP – Academic Research
Office: 402.836.1133
Mobile: 704-699-5520
Email: jeff.jones@data-axle.com
Website: https://www.dataaxleusa.com/

Dun and Bradstreet
Tara Morrow
S&MS Account Executive
Mobile: 484-723-3824
Email: morrowt@dnb.com
Website: https://www.dnb.com/