Proposal: Food Accessibility at the County Level

DATA 450 Capstone

Tenzin Sherpa

2/5/23

1 Introduction

[This section should contain background and introduction to your general topic.]

2 Dataset

The project will use the Food Environment Atlas Data from 9/10/2020 found here. The dataset has been compiled by the US Department of Agriculture (USDA) Economic Research Service (ERS) for purposes of studying factors that affect food choices and the accessibility to healthy foods in communities. The dataset contains county-level information on food environment factors such as access to grocery stores/supermarkets/restaurants, local food sales, food prices, food assistance programs like SNAP (Supplemental Nutrition Assistance Program), National School Lunch Program, etc., socioeconomic characteristics, and some health/physical activities. State-level information on food taxes, and household food insecurity are also available. More extensive documentation on each variable can be found here.

The dataset has been compiled using various reports from the USDA, the Bureau of the Census, the U.S. Department of Commerce, and more. Some variable values, like those from the 2010 census, were measured directly while others were estimates calculated in the source reports. The data cover the years 2006 to 2019 as separate reports published over a decade were used to create this dataset. However, the focus of the dataset is on the years 2010 and 2015. Hence, this project will also work on analysis relevant to these years.

Some of the variables that are of interest to this project are listed below. They are listed under categories because of the large number of variables.

- FIPS: Code to uniquely identify counties

- Population with low access to stores: Living more than 1 mile from a supermarket or large grocery store if in an urban area, or more than 10 miles from a supermarket or large grocery store if in a rural area.
  - Population, low access (%) (2010, 2015)
  - Low income and low access to store (%)
  - Households, no car and low access to store (%) (2010, 2015)
  - SNAP households, low access to store (%)
  - Children, low access to store (%)
  - Seniors, low access to store (%)
  - Race group, low access (%) 2015
- Stores:
  - Grocery stores, supercenters, convenience stores, SNAP-authorized stores, specialized food stores (all in '/1000 population' values for 2011 and 2016).
- Restaurants: (2011 and 2016)
  - Fast-food restaurants, Full-service restaurants (in '/1000 population' values).
- Food Assistance: **(2012, 2017)**
  - SNAP participants (percent pop)
  - National School Lunch Program participants (percent children)
  - School Breakfast Program participants (percent children)
  - Summer Food Service Program participants (percent children)
  - Food Distribution Program on Indian Reservations (FDPIR) Sites (2012, 2015)
- State Food Insecurity (from 2012-14 and 2015-17)
  - Household food insecurity (percent, 3-year average): Households unable to provide enough food during a year for one or more household members.
- Local Foods:
  - Farmer markets /1,000 pop (2013, 2018)
  - Farms with direct sales
  - Food hubs (2018)
- Socioeconomic characteristics
  - 2010 Percentage of the county resident population that is White, Black, Hispanic, Asian, Hawaiian or Pacific Islander, 65 years or older, under 18
  - Median household income, poverty rate, and child poverty rates in 2015

3 Data Acquisition and Processing

Before analysis work, I will check if any variables of interest have missing values. Initial exploration of the dataset showed that some fields have a high number of missing values which makes such variables less useful. For example, data on the WIC (Women, Infants, and Children) program had excessive missing values across the dataset. Such variables will be discarded from any analysis. Any remaining variables should be checked to see if they have the correct data type.

Next, I will work on imputing other missing values using similar/average values for the surrounding counties. If this is not possible, I will use median values for imputation. I will also check for outliers for quantitative variables.

Addtionally, the dataset holds data in multiple spreadsheets. An important task then would be to join the variables of interest based on the FIPS number which are the unique identifiers for counties. After attaining a final dataframe with cleaned data, I will also work on recoding variables. This recoded variables will also have a lookup/list of descriptions for the new names.

4 Research Questions and Methodology

1) Which counties have the lowest accessibility to stores? Are they concentrated in certain cities/regions/states? (5 hours)
   I will investigate this question by creating a heatmap of low accessibility populations in the counties. The variables that give the percentage of the population considered to have low access will be useful for charting this. I will create separate heatmaps for the years 2010 and 2015. In the heatmaps, a darker shade of color would indicate a higher percentage of the population with low access to food, meaning these areas would have a greater problem with food accessibility.
   Additionally, I will work on how to filter for counties with low access to food. This categorization may help scale down the problem and focus on areas that need the most help.

2) What demographic is affected the most in low accessibility regions? (5 hours)
   After identifying counties with greater issues of food accessibility, I will explore the demographics of the population. To look at the overall picture, I will find the average percentage of people under different categories (race, age, income) that have low access. For example, I will use the average of % seniors with low access from different counties. A table or bar plot may be useful to communicate the findings.
   For a more detailed analysis, I will create heatmaps of the demographic category facing the lowest accessibility. This may help to understand how low accessibility affects people of different races in selected counties.

3) Does the availability of vehicles drastically change the food insecurity level? (1 hour)
For this question, I will use pie charts to see what percentage of people with low access have cars.

4) What is the distribution of stores, restaurants, local sources, and assistance programs? (4 hours)
I will illustrate this with a heatmap where low access areas will be represented by boundaries while colors represent the density of stores and other food sources there.

5) How much has food insecurity changed in states? (1 hour)
I will select counties that had some of the lowest and highest population percentages with low access to food, and create a table that records the percentage change in food security from 2010 to 2015.

6) What is the relationship between income and food access? (2 hours)
I will use a scatterplot of the median household income and low access percent population for different counties to investigate any trends.

7) How is the type of store available related to accessibility? How are the demographic & socioeconomic characteristics of the population and food access in the county connected? (6 hours)
I will use a clustering method to answer this question. Before applying a clustering algorithm, I will create sub-dataframes with only selected columns. For instance, to answer the first question, I will use the low-access population percentages along with different store types.

8) Do food assistance programs strongly benefit counties with low food access? (5 hours)
I will compare the low access percent population with the percent population receiving benefits using a scatterplot.

5 Work plan

Week 4 (2/6 - 2/12):

- Data tidying and transforming (4 hours)
- Question 1 (5 hours)

Week 5 (2/13 - 2/19):

- Question 2 (5 hours)
- Question 3 (1 hour)
- Question 5 (1 hour)
- Question 4 (2 hours)

Week 6 (2/20 - 2/26):

- Question 4 (2 hours)
- Question 7 (6 hours)

Week 7 (2/27 - 3/5):

- Question 8 (5 hours)
- Question 6 (2 hours)
- Presentation prep and practice (4 hours)

Week 8 (3/6 - 3/12): *Presentations in class on Thurs 3/9.*

- Presentation peer review (1.5 hours)

Week 9 (3/20 - 3/26):

- Poster prep (4 hours)

Week 10 (3/27 - 4/2): *Poster Draft 1 due Monday 3/27. Peer feedback due Thursday 3/30.*

- Peer feedback (2.5 hours)
- Poster revisions (2 hours)

Week 11 (4/3 - 4/9): *Poster Draft 2 due Monday 4/3. Final Poster due Saturday 4/8.*

- Poster revisions (1 hour).

Week 12 (4/10 - 4/16):

Week 13 (4/17 - 4/23): *Finish all project work*

- Draft blog post (5 hours).

Week 14 (4/24 - 4/30): *Blog post draft 1 due Monday 4/24. Peer feedback due Thursday 4/27. Blog post draft 2 due Sunday 4/30.*

- Peer feedback (2.5 hours)
- Blog post revisions (2 hours)

Week 15 (5/1 - 5/7): *Final blog post due Tuesday 5/2.*

- Final presentation prep and practice.

Final Exam Week (5/8): *Final Presentations during final exam slot, Monday May 9th 3:20-6:40pm.*

Here's an example of citing a source (see Phillips 1999, 33–35). Be sure the source information is entered in "BibTeX" form in the `references.bib` file.

6 References

[The bibliography will automatically get generated. Any sources you cite in the document will be included. Other entries in the `.bib` file will not be included.]

Phillips, T. P. 1999. "Possible Influence of the Magnetosphere on American History." *J. Oddball Res.* 98: 1000–1003.