An Examination of Food Insecurity in the United States

Haley Allen, Philip Krause, Amoreena Ordonez

April 30, 2025

1 Abstract

Food insecurity in the United States has grown steadily since federal agencies began tracking it in the early 1990s and affects millions of Americans each year. Current measures in the largest publicly available dataset on the topic focus on household-level low-income indicators and distance markers from a supermarket. This study aims to use this data to assess whether machine learning models in tandem with maps and data visualizations can help examine relationships between demographics and food access, accurately predict low-access areas, and better inform public health initiatives. We also examine whether it's possible to predict census tracts nearing food insecurity classifications and utilize descriptive and inferential statistics to determine if Supplemental Nutrition Assistance Program (SNAP) benefits are reaching and benefiting households in need. By implementing k-nearest neighbors, random forest, linear regression, and logistic regression predictive models, we find it is possible to adequately predict food-insecure census tracts and recognize key factors that can be used to predict tracts nearing food-insecure thresholds. Our examination of descriptive statistics for SNAP benefit cost, usage, and distribution along with statistical testing, reveals two noteworthy conclusions. First, as SNAP benefit costs and participation rates rise, we would expect to see an inverse relationship with food insecurity levels that is nonexistent in our investigation. Second, there is a statistically significant difference between populations receiving SNAP benefits in low-income, low-access census tracts and those outside of them. We can only derive implied meaning on the efficacy of SNAP from both findings, but our research on the topic concludes that more effectual measures for analyzing the depth to which SNAP can help a community need to be developed. In addition, we conclude the geographic-based approach to food insecurity in the United States limits predictive modeling ability and may be excluding essential relationships connected to the issue.

2 Introduction

In the United States, many people face uncertainty in the ever-changing environmental, political, and economic landscape. One uncertainty affecting millions of households each year is regular access by all people in the home to enough food for a healthful life [8]. Those unable to consistently achieve this access to nutritious food are categorized as food insecure and, as such, become more susceptible to physical and mental health concerns, chronic health conditions, and higher healthcare costs [6, 3]. The depth of these risks implores government intervention, but the current measures used to assess food insecurity in the U.S. have not yet captured all relevant underlying variables and their relationship to food security.

A concept of food access that has garnered attention in recent years is the existence of food deserts; neighborhoods, often low-income, in the U.S. where access to quality food is insufficient due to the lack of local grocery stores [16]. Food deserts present a more narrow and community-based approach to analyzing the greater surrounding issue of food insecurity in this country. Identifying these geographic areas of food insecurity currently helps target public health initiatives, but a review of the literature suggests it may be leaving out vital pieces of the food insecurity puzzle. As studies expand on the concept of food deserts, researchers struggle to align on the key factors used to identify them, leaving many gaps in the current science left to explore.

The definition of a food desert varies depending on the article and does not have agreed-upon, standardized criteria [16]. Research is also deficient in understanding why food deserts exist [4], making efforts to remedy them difficult. Approaches to address issues in food desert research include (1) an economic analysis [4], (2) a critique of national-level measures of food access [15], and (3) reduced-form analyses and a structural demand model [1].

The term ”food desert” originated in the 1990s and has been the subject of several studies since then. Public health literature has argued that food deserts are important to food security because they are one cause of unhealthy eating habits in impoverished communities [1]. The U.S. government implemented federal bills and initiatives before scholars performed an economic analysis in 2011. This is not ideal, given that economics analyzes the allocation of limited resources between individuals and society, that limited resource being healthy, affordable food.

Analysis by Bitler and Haider (2011) dissects the issues of food deserts from the four basic mechanisms of an economy: relevant products, consumers, retailers, and their exchange (market). First, the relevant product in a food desert scenario is healthy food, but ”healthy food” is an incomplete definition. It is unclear which foods are considered healthy. Some definitions of food deserts only allude to fresh fruits and vegetables. Three basic food groups are left out: dairy, grains, and protein. Second, a food desert entails inaccessibility to healthy food from a fixed location, usually a place of residence, some specified distance from a food retailer. A house is a fixed location, but people are not; they take public transit, commute to work, carpool, and take their kids to school. Travel patterns change proximities to food sources, allowing other opportunities to access healthy food from other food sources. Third, a food desert lacks a food retailer, generally understood as a supermarket or large grocery store. The availability of healthy food at corner stores, farmers' markets, specialty and ethnic grocery stores, food cooperatives, and convenience stores does not qualify. Altogether, assessing what healthy food is, where it can be accessed, and how it can be accessed is incomplete [4].

Ver Ploeg, Dutko, and Breneman (2014) highlight another faultline in centering food insecurity research around food deserts: Focusing on area-based food access rather than individual-based measures. Delineating geographical areas makes for easier data aggregation and policy direction. However, not all individuals in a low-income food desert share the same income status or food access limitations. Measuring food access by areas also overlooks low-income individuals living outside designated food deserts. Low-income households are spread out, not clustered. Two problems arise: (1) a misrepresentation of food access limitations in some neighborhoods and (2) a devaluation of individual-based food access limitations in others [15]. These economic viewpoints call attention to key factors in determining food deserts, including accounting for other food sources besides supermarkets and large grocery stores, household transportation availability, individual-based food access measures, and the impact of subsidies for healthy foods.

Food deserts become a micro perspective in the larger problem of food insecurity after the Great Recession of 2008 when rates of food insecurity rise by 30% and evade significant recovery to pre-recession statistics [6]. At this point the ability to purchase enough nutritious food for a household becomes endangered as employment rates and income fell and those not living in previously defined food deserts were faced with a new food access uncertainty. During this time data on food insecurity in the U.S. become inconsistent across demographics, income levels, and geographic areas. What is found to be consistent is households with children or those composed of a non-nuclear family structure were found to suffer higher rates of food insecurity across the literature [6].

Today food insecurity is measured and reported on annually by the United States Department of Agriculture (USDA) based on responses collected in the Current Population Survey Food Security Supplement (FSS) administered by the United States Census Bureau. The survey contains 11 questions that measure a household's ability to access healthful food at any time during the year, with an extra 7 questions on the subject for households with children. The USDA classifies a food-insecure household to any survey responses indicating 3 or more circumstances of food insecurity and any 2 or more responses for households with children. In the USDA's most recent report compiled by Rabbitt et al. (2023), they announced 13.5% of households faced food insecurity at some point during the year, noting its statistically significant increase from the previous year's report [8]. Within this percentage of households in the U.S., they further delimit households into low food security and very low food security, the latter acknowledging households with one or more members dealing with active food and diet disruptions. With 11.2 million households facing low food security and 6.8 million households facing very low food security [8], it is imperative to uncover more significant identifiers for food insecurity to better inform policies and reach households at risk of nearing these levels.

Federal welfare programs and public health initiatives have been put into place to help reduce the prevalence of food insecurity, but their efficacy has been called into question by recent research [5, 7, 11, 12, 13]. As these questions and doubts are brought to the table, it's necessary to examine how they help vulnerable populations, where their efforts can be best served, and how else we may be able to help those in food-insecure households.

Historically, getting access to affordable, nutritious food has often been a difficulty for those living in poverty in the United States. Programs designed to help combat this problem at an individual or household level, however, were not implemented federally until 1964, with the signing of the Food Stamp Act [7]. A decade later, in 1974, this was made into a national program aiming to help households afford basic nutrition needs [7]. Today, this program has developed into the Supplemental Nutrition Assistance Program (SNAP). In order to be eligible for SNAP benefits, a household must meet certain income requirements, after which they are provided an Electronic Benefit Transfer (EBT) card which can be used to buy food at any authorized grocery stores or other food retail locations [7]. The implementation of food assistance programs such as SNAP account for around two-thirds of the annual budget of the USDA [12] and in 2023, SNAP helped an average of 41 million individuals a month [8] which is around 12 percent of the US population.

A large portion of recent research on food insecurity, food deserts, and the impact of SNAP highlight problems and challenges associated with the program. A common conclusion in recent studies is that SNAP is not efficient and leaves many households needing more help [5, 7, 11, 12, 13]. One cause of this problem indicated by the literature is the existence of maximum benefits while the program does not factor in differences in prices across the country. For households located in areas with high food prices, SNAP benefits do not cover necessary costs which means that they must choose cheaper, less healthy food options [5, 7]. This means that in food deserts, even the assistance of programs like SNAP may not help solve the problems faced by households. In their study, Gundersen et al. (2019) show that even outside of these specific areas, in 99% of counties in the contiguous US, the average cost of a low-income meal is still greater than the maximum SNAP benefits granted for meals [7].

Another problem associated with SNAP is the lack of impact on nutrition and healthy food choices [2, 5, 11, 12]. Solutions have been proposed such as including healthy food subsidies [2] or adding nutritional requirements [11] but there are problems associated with both. Adding healthy food subsidies would cost an additional 15% of the SNAP budget [2] and restricting what individuals can buy can be seen as discriminatory and patronizing towards those affected [11]. Some of the literature concludes that SNAP benefits can not solve the problems of food deserts on their own, but looking into the causes and trying to eliminate the existence of food deserts is the best option for helping these households [11].

2.1 Methods in Current Literature

Most research on food insecurity in the United States has relied on nationally compiled surveys and analyses of correlations between food security and one dependent variable at a time. This information and data are then used in a multitude of studies for geospatial visualizations, often broken down by census tract, income level, or ethnic and racial demographics [16]. The drawback of these research methods prevent a complete understanding of the underlying factors that can cause or worsen food insecurity [16]. Since survey data is based on self-elected participation and self-reported behaviors or feelings, there can be misrepresentations in the data [3]. Past literature has discovered statistically significant correlations between variables like transportation and the presence of chronic health conditions. However, it has yet to examine in depth how multiple metrics may interact to further exacerbate food insecurity. Implementing feature importance on all relevant circumstances to fine-tune parameters for machine learning may provide greater insight into significant relationships and the ability to identify food-insecure households accurately.

When analyzing the impacts of SNAP, the data collection and methods used in various studies are similar given the small number of wide scale data sets on the subject. Most data is taken from the Current Population Survey (CPS) [7, 12], the Survey of Income and Program Participation (SIPP) [7, 12, 13], and Nielsen reports on food prices [7, 2]. As for the methods of analysis, most of the literature uses very little predictive modeling, but rather focuses on finding correlations and patterns in existing data. A challenge that can be seen across the literature is that it can be difficult to see the exact impact of SNAP because of complexities in data. Issues like differences in policy by state and the unknown impact of private organizations and charities lead to many studies failing to find significant correlations [12]. Through our project, we aim to use inferential and descriptive statistics as well as geospatial analysis to analyze the actual impact of SNAP on food insecurity and food deserts.

2.2 Project Plan

The overall lack of application of data science and machine learning in the existing literature leaves a breadth of room for data exploration, visualizations, and the implementation of multiple models. To begin we will employ a k-nearest neighbors predictive model for its reproducibility and ability to handle the geospatial data currently available in the field. Next, random forest models will be used to identify key features in predicting food-insecure areas to inform future research and data collection. Linear and logistic Regression models will then be run to determine if more simple or complex classification models can result in superior model performance. Lastly, descriptive statistics and visualizations will largely be utilized to examine SNAP efficacy, and hypothesis testing will be conducted to report any statistically significant findings between food-insecure and non-food-insecure areas.

Together, we hope that these techniques and our analyses of them will determine if data science methodology can help to explore key relationships between demographics and food insecurity, predict food-insecure areas, and prevent this issue from impacting more communities. We also expect to confirm whether it's possible to predict census tracts nearing food-insecure thresholds and explore the efficacy of SNAP benefits on food insecurity.

3 Methods

3.1 Datasets and Pre-Processing

The primary source for our study is the USDA Food Access Research Atlas (FARA)[9]. This .xlsx file contains 72,532 observations from households across the United States with 147 columns of variables. These observations are "based on a 2019 list of supermarkets, the 2010 Decennial Census, and the 2014-18 American Community Survey (ACS)"[9], which means data points from 2020 to 2024 are not included. Variables include food access measures by census tract, urban or rural area, income level, age and ethnicity/race demographics, vehicle access, total number of households receiving SNAP benefits, and distance from a supermarket broken into 0.5, 1, 10, and 20 miles.

The National Level Annual Summary[14] is our secondary data source pulled from the USDA website. This Excel file details the average monthly benefits per person, the participation in the program, and the SNAP cost breakdown from 1969 to 2024. It serves as an insight into the cost versus benefit of the program.

A section of our exploratory data analysis findings include utilizing GIS data to map trends seen in the FARA data. In order to accomplish this, data files were taken from the US Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) system. The files contain border geometries of all census tracts in the country. These .shp files, when merged with the USDA's data, allow for geospatial mapping.

Cleaning of the FARA data as well as of the TIGER data was needed to facilitate mapping. Census tracts are encoded with an 11-digit code indicating state, county, and tract which can include leading zeroes.

In the raw data, these codes were float values, thus any leading zeroes were absent. In order to be able to merge the FARA data with the TIGER data, the code variables had to be converted to string values, then given leading zeroes if needed. Additionally, TIGER data contains all U.S. territories, so rows containing tracts located outside of the contiguous 48 states were excluded.

Exploratory data analysis revealed relationships with food insecurity across age, race, and ethnicity demographics, SNAP benefit recipients, and vehicle access within the FARA data. To perform predictive modeling, statistical testing, and examine these correlations more closely we created a new working dataframe to include all counts of these factors by tract and remove previously chosen factors for geographic visualizations. This refined dataset along with the methodology that follows allowed us to address the predictability of food deserts in the United States, the possibility of identifying areas at risk for early intervention, and understand if SNAP benefits have a significant effect on food deserts.

Once predictor variables were chosen, we identified `LILATracts_halfAnd10` as the main target variable for our methods. This variable is a binary flag for low-income and low-access census tracts measured at 0.5 miles or more from a supermarket in urban areas and 10 miles or more from a supermarket in rural areas. The low-income and low-access classifiers are defined in this context by the USDA as "Low-income: a poverty rate of 20 percent or greater, or a median family income at or below 80 percent of the statewide or metropolitan area median family income; Low-access: at least 500 persons and/or at least 33 percent of the population live more than 1 mile from a supermarket or large grocery store (10 miles, in the case of rural census tracts)" [9]. This was chosen among the three low-income and low-access (LILA) binary flag variables at different mile markers because it was the most inclusive among them and would identify the largest number of food-insecure census tracts. Once the new dataframe was created with all necessary variables, standard removal of NaN and null values was completed, leaving 71,782 observations to study.

Two subsets based on the binary classification of the target variable were created to analyze descriptive statistics and conduct statistical tests related to the SNAP: One for census tracts flagged for the target variable (`df_halfAnd10`) and a second for those not flagged for it (`df_NOThalfAnd10`). The number of independent variables were decreased from 146 to 14 and null values were dropped from both subsets. An additional column named `SNAPproportion` was inserted in the FARA dataset and in both subsets.

Before creating the linear and logistic regression models, the number of independent variables was reduced from 146 to 4 to match the feature importance from the random forest classification model. All null values in the columns of the chosen independent variables and for the target variable were dropped. The selected data were then prepared into target and predictor variables and split into training and test sets. The models were then created and trained with the training set, and predictions were made on the test set. The predictions for the linear regression model were converted from continuous to binary variables with a threshold of 0.5, where ≥ 0.5 was classified as 1 and < 0.5 was classified as 0.

3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) algorithms were applied to the FARA data to address our main question of predicting food-insecure census tracts. To ensure any household classified as low-access by the USDA was included for predictive modeling, all KNN models were run on the prepared data described above, but also included the remaining two LILA variables with all urban and rural mile markers. With the addition of binary flag variables `LILATracts_1And10` and `LILATracts_1And20`, a clustering machine learning model was chosen to most adequately handle pattern recognition processes in the data. Prior to creating multi-label models, data for the predictor variables in both the training and test sets were Z-score normalized using the `StandardScaler()` method. An initial run of KNN was performed with the default parameter of 5 k-nearest neighbors on a training set of 70% with following predictions made on a testing set of 30%. A subsequent run with k = $(\sqrt{(}n)/2)$ was performed with the same [70/30] split.

3.2.1 Data Imbalance Mitigation with Oversampling Techniques

To get a sense for the degree of data imbalance, there are 20,041 census tracts flagged for the target variable, and there are 51,741 census tracts not flagged for the target variable.

Our preference is for a model that is more sensitive to the minority class, or in our case, census tracts flagged for food insecurity. It would be in our best interest then to apply oversampling methods that will train models on those "harder to learn" instances. Adaptive Synthetic Modeling Approach for Imbalanced Learning (ADASYN) and Synthetic Minority Over-Sampling Technique (SMOTE) were chosen to conduct the resampling.

Before applying oversampling techniques, the three binary columns used in the previous multi-label KNN models [`'LILATracts_halfAnd10'`, `'LILATracts_1And10'`, and `'LILATracts_1And20'`] were recoded into one combined, categorical column named [`'LILA_All'`]. After recoding, the dataset was again divided into, what we are calling, retraining and retesting sets using a [70/30] split.

<u>ADASYN</u>:

An `ADASYN()` object was instantiated and applied to the newly split, imbalanced data with its function `fit_resample()` to create a more even resampling of the retraining set. The KNN model (k=5) was then trained with the resampled ADASYN values, and predictions were made on the retesting set.

<u>SMOTE</u>:

A `SMOTE()` object was instantiated and also applied to the newly split, imbalanced data with its function `fit_resample()` to create an equalized resampling of the retraining set. Another KNN model (k=5) was trained with the resampled SMOTE values, and predictions were made on the retesting set.

ADASYN generates synthetic data specifically for the minority class, focusing on producing more samples for minority class instances surrounded by the majority class, which tends to be more difficult to label. In contrast, SMOTE generates an equal number of synthetic samples for all minority class instances. We applied both to compare their performances in order to choose the method better suited for our research purposes. Once resampled, KNN models were then fine-tuned using GridSearchCV to discern ideal input for weights, algorithm, and n_neighbors parameters. With 'distance', 'ball-tree', and '300' as reported best input parameters respectively, KNN was run again to assess for any increased model performance.

3.3 Random Forest

Our first sub-question asks whether census tracts nearing food desert classification can be detected. To approach this question in the context of the FARA data, we utilized a random forest classification model with the target variable assigned to `LILATracts_halfAnd10`. Next, the data was separated into training and test sets with a [80/20] split, a `RandomForestClassifier()` object was instantiated, a random forest classification model was fit to the training set, and predictions were made on the test set.

To yield the most accurate and most reliable predictive power from this model, a grid search cross-validation was done on the random forest classification model. Optimal parameters chosen are as follows: `'max_depth'`: None, `'max_features'`: 'log2', `'min_samples_split'`: 5, `'n_estimators'`: 200. Using these parameters, a new random forest classification model was created and tested. Based on this random forest classification model, the top 4 most important predictive features, in descending order, include: `MedianFamilyIncome`, `PovertyRate`, `TractSNAP`, and `TractHUNV`.

Because total households receiving SNAP benefits and total households without a vehicle are two of the most important predictive variables from the random forest classification model, analyzing them to predict SNAP usage is a logical next step. Moreover, much of the current literature on food deserts indicates that these variables are integral for future research. To do this, a random forest regression model was created with the `TractSNAP` column set as the target variable. Again, the data was separated into training and test sets with a [80/20] split. Because `TractSNAP` is a continuous variable, a `RandomForestRegressor()` object was created with `'n_estimators'` = 100, a random forest regressor model was trained, and predictions were made on the test set.

3.4 Descriptive and Inferential Statistics

The analysis of the efficacy of SNAP benefits began broadly and then narrowed down to more specific areas. The FARA dataset was disaggregated to focus on recipient and non-recipient SNAP census tracts inside and outside the target variable. Keeping in mind the target variable, `LILATracts_halfAnd10`, of the 71,782 census tracts within the cleaned FARA dataset, 20,041 are flagged as `LILATracts_halfAnd10`. The percentage of `LILATracts_halfAnd10` in the dataset equals 27.9%. The percentage of census tracts outside of the target variable is the complement, 72.1%. Since the FARA dataset includes information for every state in the U.S., plus the District of Columbia, it is representative of the U.S. population. This means 27.9% of the census tracts in the U.S., basically 1:4, are considered LILA.

The FARA dataset includes a variable, `POP2010`, a total population count from the 2010 census. In order to calculate the percentage of the population residing within the target variable, the population of `df_halfAnd10['Pop2010']` was divided by `fara['Pop2010']`, equaling 26.6%. Again, about 1:4 individuals live within a LILA tract.

To explore the ratio of housing units receiving SNAP benefits within each census tract, a new column was added. This new column, `SNAPproportion`, was calculated by dividing the total number housing units receiving SNAP benefits with variable `TractSNAP` by the total number of housing units with variable `OHU2010`. This way, you can appreciate the percentage of SNAP-recipient households for each census tract and within the subset dataframes.

- `SNAPproportion` within `LILATracts_halfAnd10` = 21.32%, with a median of 20.44%.
- `SNAPproportion` outside `LILATracts_halfANd10` = 9.43%, with a median of 7.4%.

If you prefer to think in whole numbers, the FARA dataset also offers a tally of the number of housing units receiving SNAP benefits within each census tract with the variable, `TractSNAP`.

- The `TractSNAP` average within `LILATracts_halfAnd10` = 322 housing units, with a median of 285.
- The `TractSNAP` average within `LILATracts_NOThalfAnd10` = 157 housing units, with a median of 114.

The investigation into SNAP benefits aims to identify a significant disparity in the numbers between census tracts within and outside the target variable. From the aformentioned descriptive statistics for both `SNAPproportion` and `TractSNAP`, there is at least a recognizable difference between the two areas. For more validating results, statistical tests are conducted with p-values compared to a significance level of 0.05. An independent samples t-test and a two-sample Welch's t-test were performed with the variable `SNAPproportion` between `LILATracts_halfAnd10` and Non-`LILATracts_halfAnd10`. A two-sample Welch's t-test was performed due to unequal variances of the variable `TractSNAP` between `LILATracts_halfAnd10` and Non-`LILATracts_halfAnd10`. In both situations, the null hypothesis claims no difference in sample means exist with either variable between the two areas.

3.5 Linear and Logistic Regression

The four most important predictive variables from the random forest classification model are as follows: `MedianFamilyIncome`, `PovertyRate`, `TractSNAP`, and `TractHUNV`. These four variables were assigned as predictor variables for both the linear and logistic regression models with `LILATracts_halfAnd10` assigned as the target variable. Next, the data was prepared into training and test sets with a [70:30] split. A `LinearRegression()` object was created and the linear regression model was fit to the training set. Predictions were made on the test set and outputted as continuous variables. These continuous values were manually converted to binary with a threshold set at ≥ 0.5, where ≥ 0.5 was classified as 1 and < 0.5 was classified as 0. Next, a `LogisticRegression()` object was created, trained, and predictions were made on the test set. Similar to the KNN method, data imbalance was adjusted for by applying the SMOTE() method to both regression models, producing a resampling of the training sets and equalizing for the minority class. Linear and logistic regression models are again instantiated and retrained with resampled SMOTE values. Finally, new SMOTE-assisted predictions were made for both models on their test sets.

4 Results

4.1 Mapping

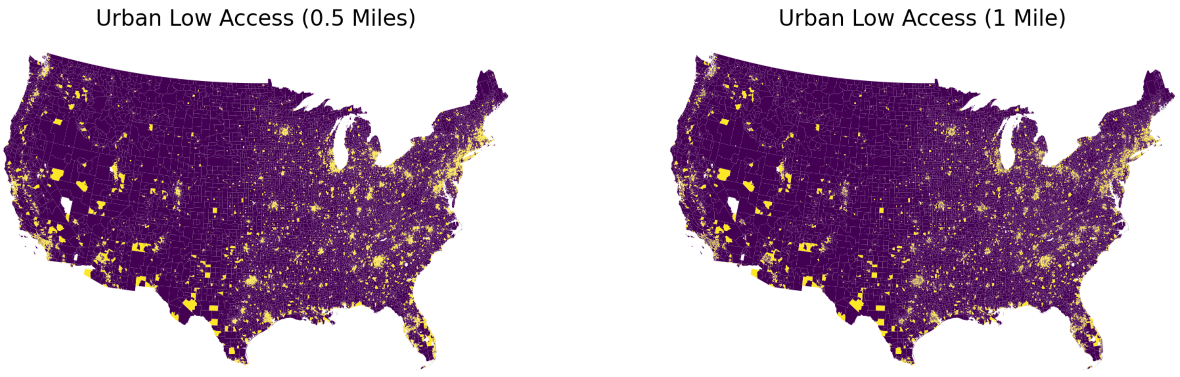


Figure 1: Urban Low Access Tracts at 0.5 and 1 Mile

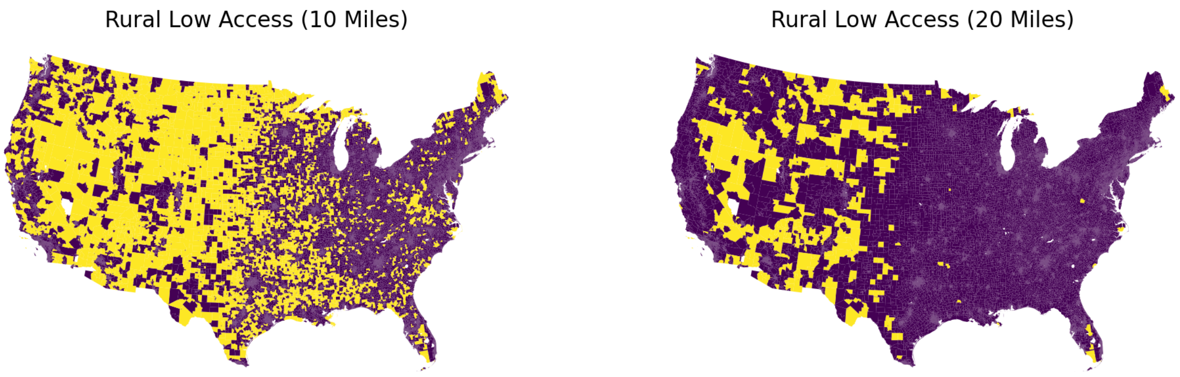


Figure 2: Rural Low Access Tracts at 10 and 20 Miles

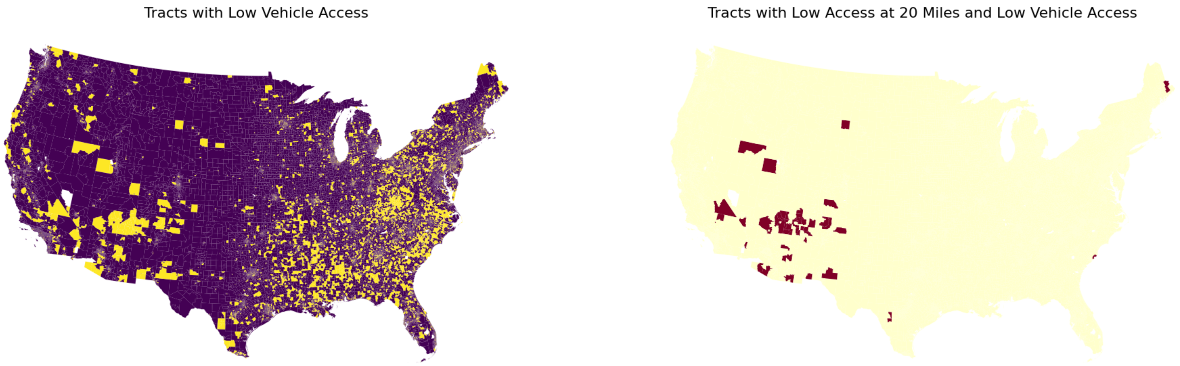

Tracts with Low Vehicle Access
Tracts with Low Access at 20 Miles and Low Vehicle Access

Figure 3: Intersection of Low Vehicle Access and Low Food Access

These nationwide maps are not capable of showing any clear relationships or meaningful patterns, especially in urban areas because the data focuses on such small areas. Nonetheless, Figures 1 and 2 give a good picture of the extent of low food access across the country. Figure 3 indicates the intersection between low food access and low vehicle access, with this clearly being the most prevalent in the southwestern portion of the country.

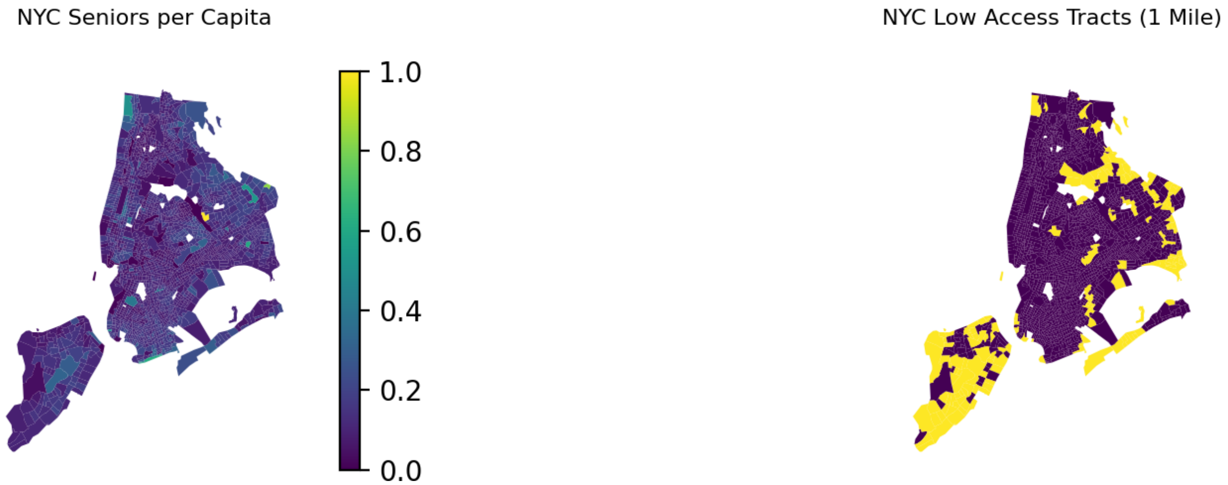


Figure 4: Seniors and Low Access Tracts in New York City

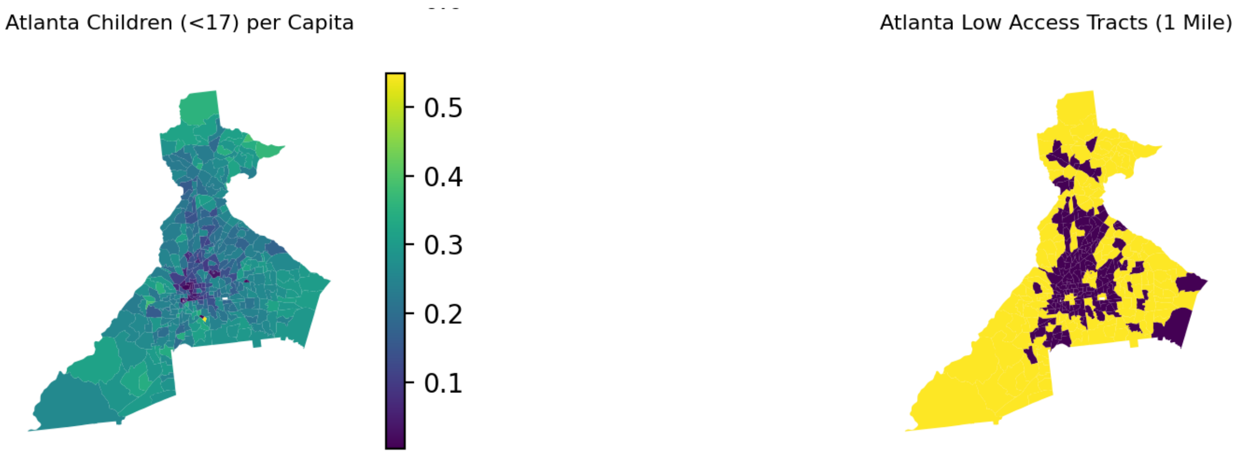

Atlanta Children (<17) per Capita
0.5
0.4
0.3
0.2
0.1
Atlanta Low Access Tracts (1 Mile)

Figure 5: Children and Low Access Tracts in Atlanta

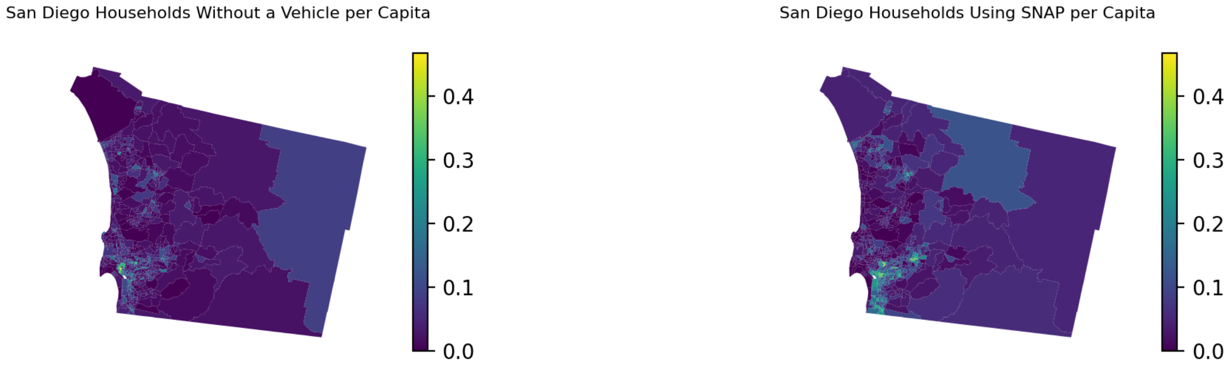

San Diego Households Without a Vehicle per Capita
0.4
0.3
0.2
0.1
0.0
San Diego Households Using SNAP per Capita
0.4
0.3
0.2
0.1
0.0

Figure 6: Vehicle Access and SNAP Participation in San Diego

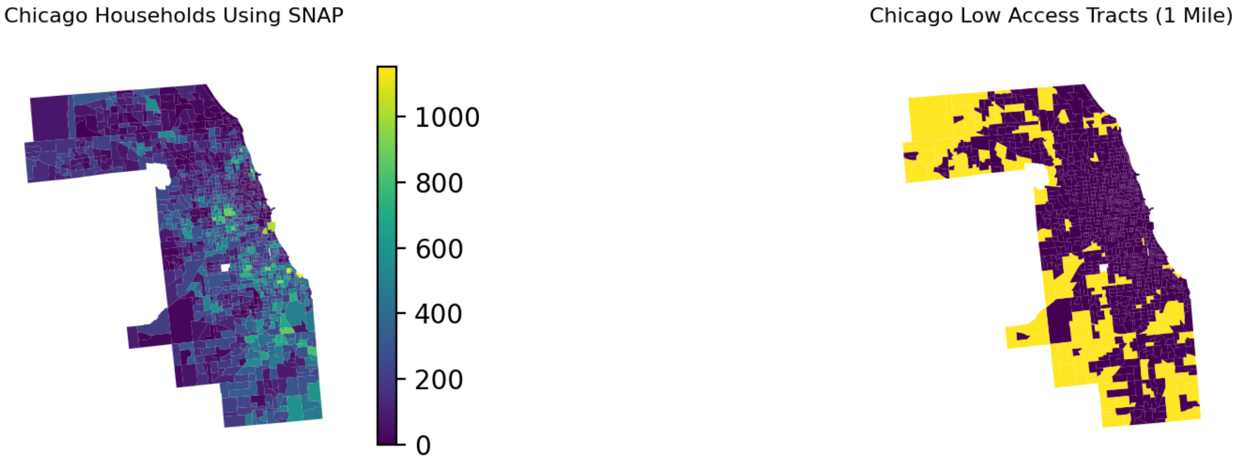
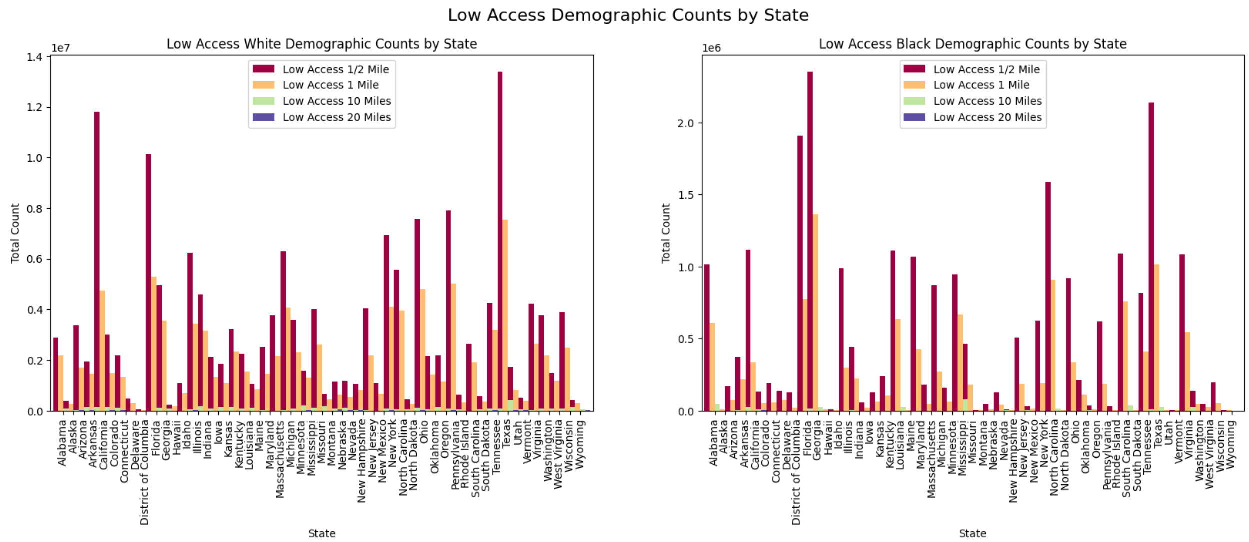

Chicago Households Using SNAP
1000
800
600
400
200
0
Chicago Low Access Tracts (1 Mile)

Figure 7: SNAP Households and Low Access Tracts in Chicago

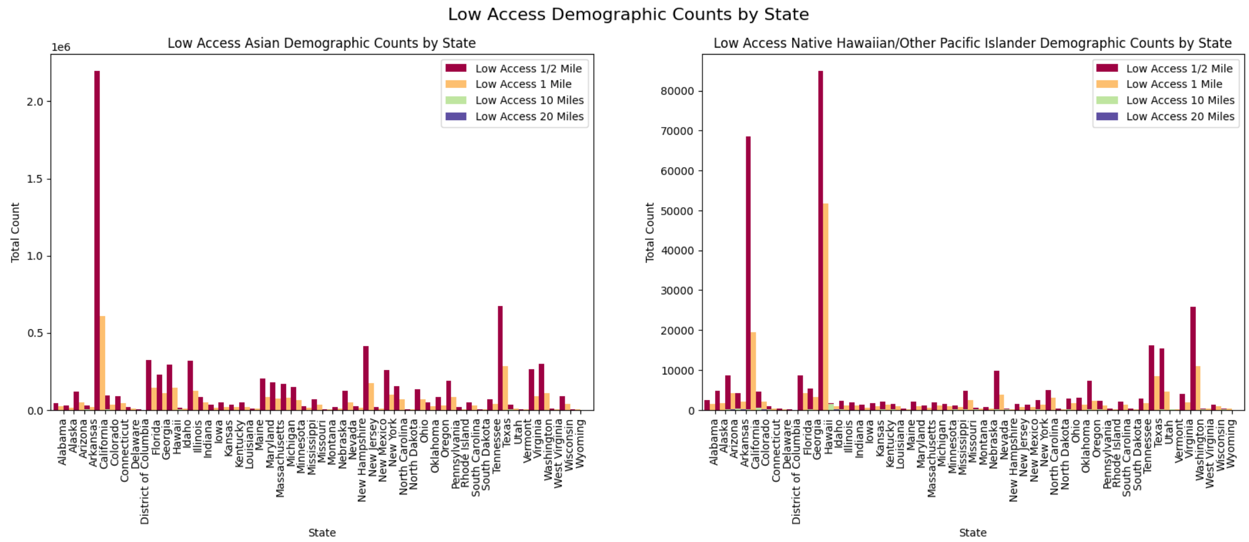
These county-wide maps containing major cities give more clear pictures of possible relationships in the data. Figure 4 visually indicates some correlation between senior populations and low access tracts. Figure 5 shows a stronger visual relationship between child populations and low access tracts. A slight visual relationship between vehicle access and SNAP usage can be seen in Figure 6. Finally, Figure 7 appears to show an inverse relationship between SNAP usage and low access.

4.2 Food Access Demographics


Low Access Demographic Counts by State

Low Access White Demographic Counts by State

Low Access Black Demographic Counts by State


Low Access Demographic Counts by State

Low Access Asian Demographic Counts by State

Low Access Native Hawaiian/Other Pacific Islander Demographic Counts by State

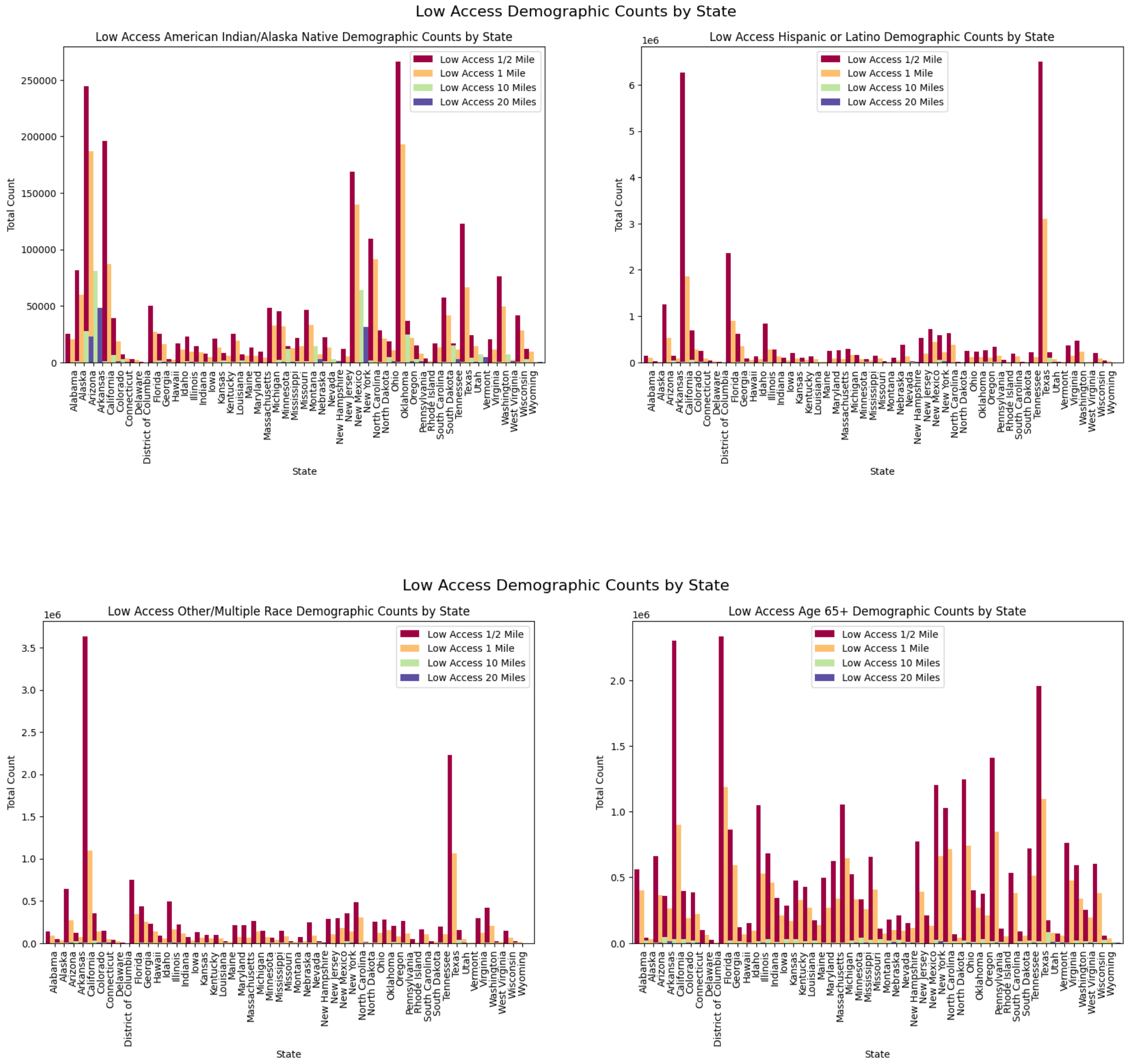
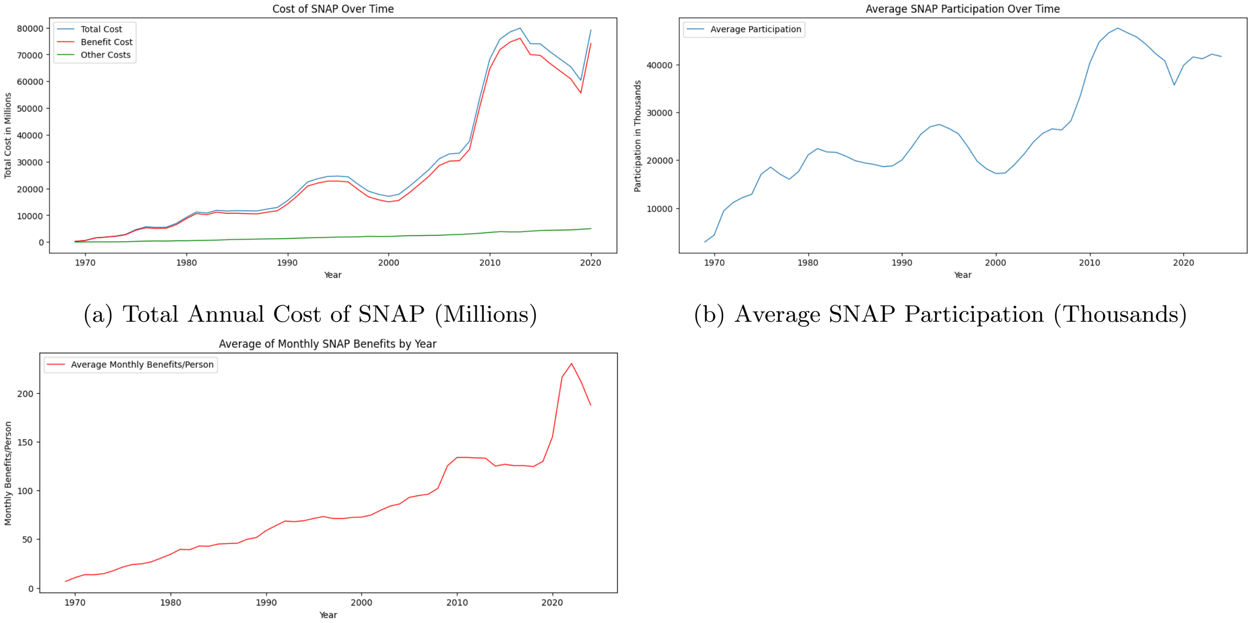


Figure 8: Demographic Counts of Food Access Across the United States

The various graphs in Figure 8 show many of the demographic trends seen in this data. Overall, the majority of households categorized as low access by the USDA are represented by those located 0.5 miles or more away from a supermarket. Places and populations with a large number of households more than 20 miles away from a supermarket are American Indian/Alaska Natives in Alaska, Arizona, Montana, and New Mexico. Additionally, Arkansas consistently has high amounts of low access populations at every demographic except Black. Seniors aged 65+ have the next largest population of low access households with mile markers above 0.5 and 1 across all states. It is important to note that some states simply have larger populations of specific demographics than others which explain spikes in states like Hawaii for Native Hawaiians and California and Texas for multiple demographics. The broad results of this analysis are that millions of people are affected by low access to food and all populations are vulnerable in their own ways.

4.3 SNAP Trends



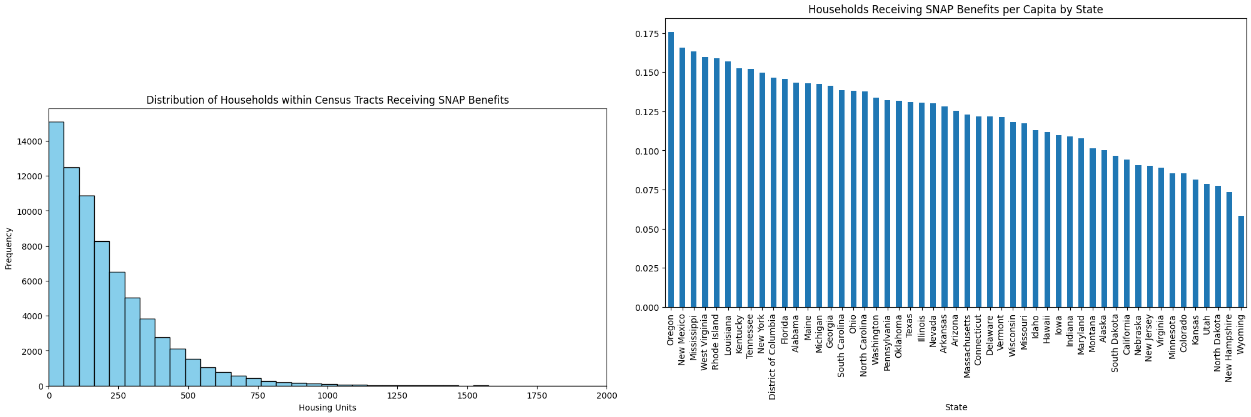
(a) Total Annual Cost of SNAP (Millions)

(b) Average SNAP Participation (Thousands)

(c) Average Monthly SNAP Stipend Amount (Dollars)

Figure 9: SNAP Data Exploration

All three line graphs seen in Figure 9 show a general trend of increasing costs and benefits over time, with larger spikes seen around 2008 and 2020. Additionally, Figure 9a highlights a lack of change over time in ”other costs” which include project funding, education, employment, and training which all support the SNAP program.



(a) SNAP Household Totals Histogram

(b) SNAP Benefit Recipients per Capita by State

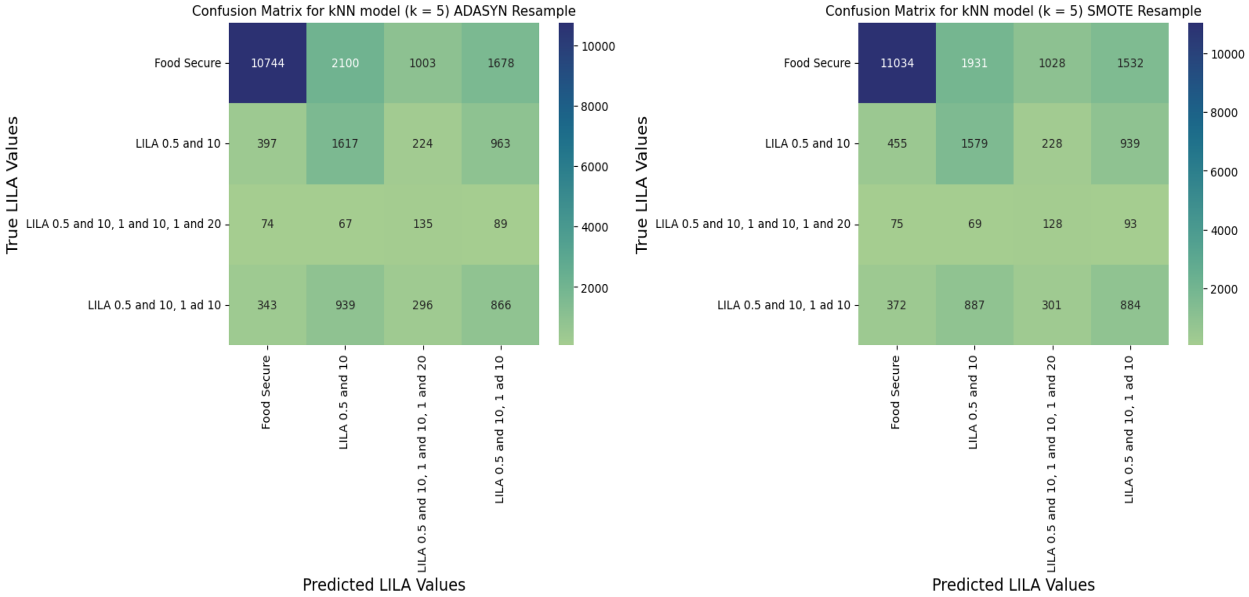
Figure 10: General SNAP Trends

Figure 10a shows a right-skewed distribution of SNAP recipients, indicating that few census tracts have many households receiving stipends. This draws the average to the right of the median, making it a misleading indicator of central tendency. Figure 10b shows the states with the highest per capita SNAP usage. Oregon, New Mexico, Mississippi, and West Virginia are the states with the highest per capita usage.

4.4 K-Nearest Neighbors (KNN)

A total of six KNN models were run on the FARA data, adjusting for parameter tuning and resampling methods. The initial KNN models using three binary target variables for all LILA mile markers included the default k value input of five and k = $(\sqrt{(n)}/2)$ parameters. Reported accuracy for these models were 75.54% and 76.17%, respectively.

The next two KNN models were run after the FARA data was resampled using ADASYN and SMOTE to add synthetic samples to the minority classifications in the target prediction variables. This adjustment for the skewed distribution resulted in lower, but presumably more true scores of model performance. The ADASYN KNN model resulted in 60.44% accuracy and the SMOTE KNN model resulted in 61.74% accuracy. Figure 11 shows the confusion matrices for each resample method.

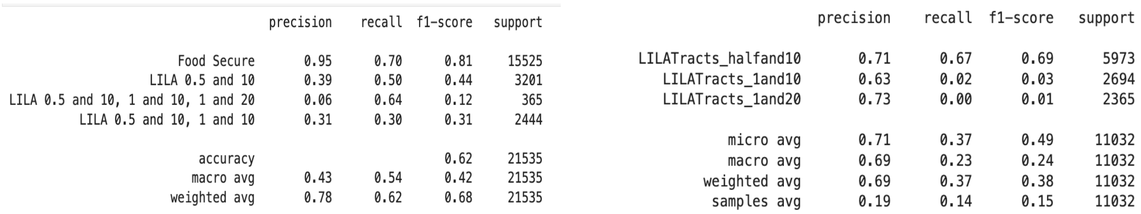


(a) ADASYN Confusion Matrix

(b) SMOTE Confusion Matrix

Figure 11: Confusion Matrix Heat Maps for Resampled Data

With decreased accuracy levels in mind, KNN models that were fine-tuned using GridSearchCV to discern ideal input for weights, algorithm, and n_neighbors parameters were run with and without SMOTE applied. Scores for each model remained close to their counterparts with a GridSearch SMOTE KNN accuracy level of 62.29% and a GridSearch KNN without SMOTE accuracy level of 75.80%. Figure 12 displays the classification reports for each model run.



```
                                    precision    recall  f1-score   support

                        Food Secure      0.95      0.70      0.81     15525
                      LILA 0.5 and 10      0.39      0.50      0.44      3201
LILA 0.5 and 10, 1 and 10, 1 and 20      0.06      0.64      0.12       365
            LILA 0.5 and 10, 1 and 10      0.31      0.30      0.31      2444

                           accuracy                          0.62     21535
                          macro avg      0.43      0.54      0.42     21535
                       weighted avg      0.78      0.62      0.68     21535
```

(a) KNN GridSearch with SMOTE Resample

```
                       precision    recall  f1-score   support

LILATracts_halfand10       0.71      0.67      0.69      5973
   LILATracts_1and10       0.63      0.02      0.03      2694
   LILATracts_1and20       0.73      0.00      0.01      2365

           micro avg       0.71      0.37      0.49     11032
           macro avg       0.69      0.23      0.24     11032
        weighted avg       0.69      0.37      0.38     11032
         samples avg       0.19      0.14      0.15     11032
```

(b) KNN GridSearch without SMOTE Resample

Figure 12: Classification Report Tables After GridSearch

4.5 Random Forest

The random forest classification model targeting LILA tracts, using data resampled with SMOTE, and hyperparameterized through a grid search cross validation yielded an accuracy of 85.54%. The model's Receiver Operating Characteristic (ROC) curve yielded an Area Under the Curve (AUC) of 0.93, as seen in Figure 13.

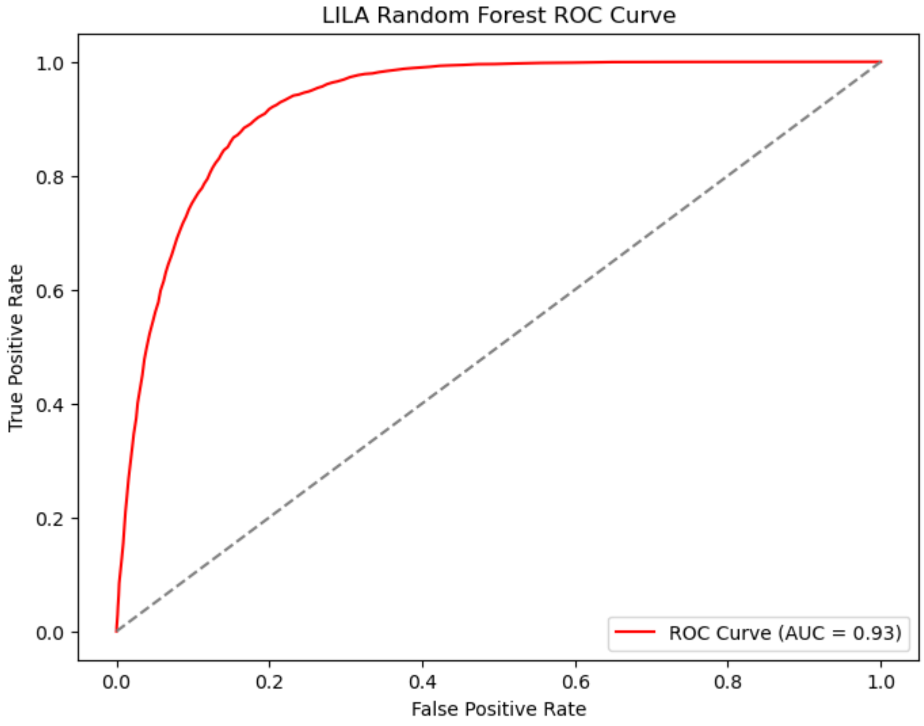
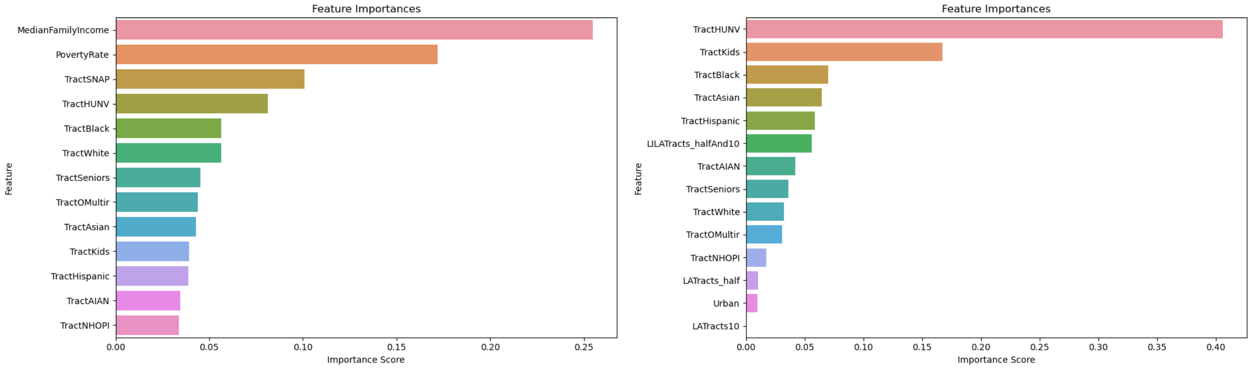


Figure 13: ROC Curve for LILA Random Forest Classification

The random forest regression model targeting tract SNAP usage, again using the resampled data and hyper-parameterized, yielded a root mean squared error of 110.12 and an R^2 value of 0.71.

Figure 14 shows the plotted importance scores for the variables used in both random forest models. In the LILA random forest classification, the variables with the most predictive importance were median family income, poverty rate, households using SNAP, and households with low vehicle access. In the SNAP random forest regression, the variables with the most predictive importance were households with low vehicle access and child population.



(a) LILA Random Forest Features

(b) SNAP Random Forest Features

Figure 14: Random Forest Feature Importances

4.6 Descriptive and Inferential Statistics

The percentage of SNAP-recipient housing units within the target variable's census tracts is 21.32%. Ostensibly, the remaining 78.68% of housing units within the target variable's census tracts do not receive SNAP benefits. Within the subsetted target variable dataframe, `df_halfAnd10`, there are 20041 observations, each representing a census tract. Each census tract instance contains a count of SNAP-recipient households under the variable `TractSNAP`. To summarize, 21.32% of the 20041 census tracts equals 6,456,948 SNAP-recipient housing units. On the flip side of the same coin, the percentage of SNAP-recipient housing units outside the target variable's census tracts, `LILATracts_NOThalfAnd10`, is 9.43%. Yes, this proportion is smaller, but here is a breakdown of the two percentages:

- 21.32% of the 20041 target variable census tracts = 6,456,948 SNAP-recipient households out of a total of 30,283,453 households.
- 9.43% of the 51741 non-target variable census tracts = 8,126,731 SNAP-recipient households out of a total of 86,146,616 households.

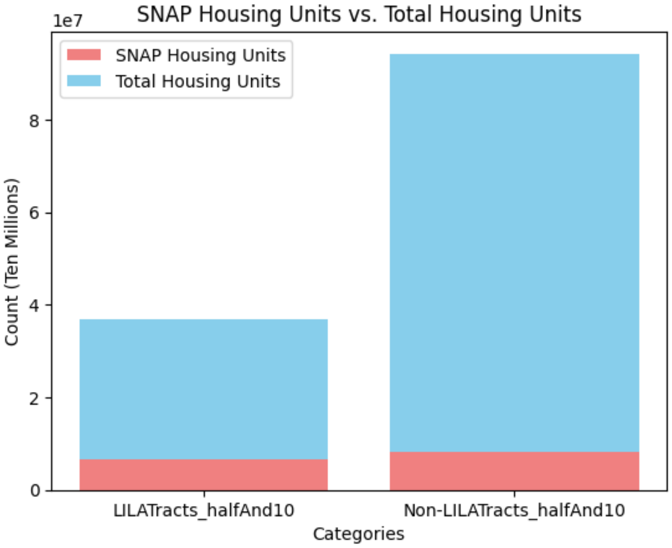
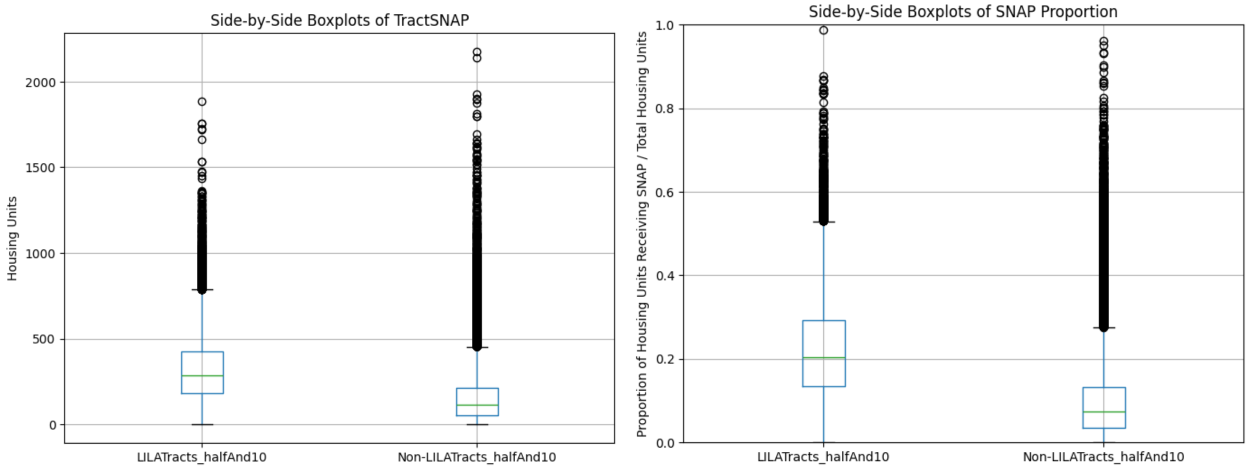


Figure 15: Comparison of SNAP Proportion and Number of SNAP-Recipient Housing Units

The higher SNAP proportion for LILATracts_halfAnd10 census tracts can be appreciated visually from the stacked bar chart seen in Figure 15. However, comparing the red-shaded ”SNAP Housing Units” between the two categories shows that non-LILATracts_halfAnd10 census tracts have a higher number of SNAP-recipient households. From the estimations calculated above, this number is almost 2 million.



(a) TractSNAP Distribution

(b) SNAPproportion Distribution

Figure 16: Distributions of LILATracts_halfAnd10 and Non-LILATracts_halfAnd10

4.6.1 TractSNAP Statistical Test

A two-sample Welch’s t-test was performed due to unequal variances of the variable TractSNAP between LILATracts_halfAnd10 and Non-LILATracts_halfAnd10.

Null hypothesis: There is no difference in the number of housing units receiving SNAP benefits between LILATracts_halfAnd10 and Non-LILATracts_halfAnd10. This statistical test returned a t-statistic of 105.04, and a p-value of 0.0; the p-value was so small the datatype, float64, did not allow enough decimal digits for the leading zeros.

Result - Reject the null hypothesis: There is a significant difference between the sample means.

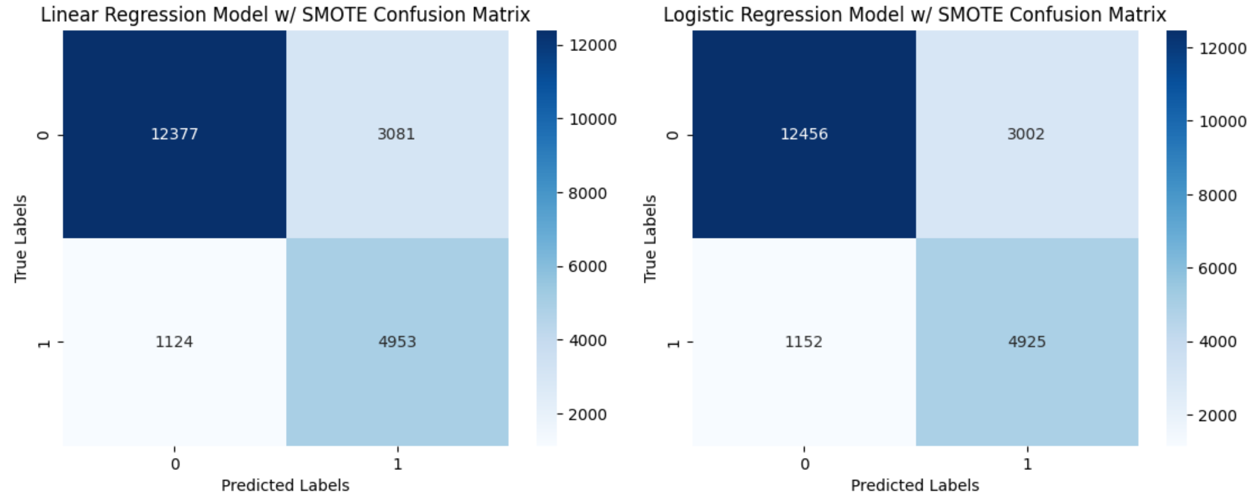
4.6.2 SNAPproportion Statistical Tests

An independent samples t-test and a two-sample Welch’s t-test were performed with the variable SNAPproportion between LILATracts_halfAnd10 and Non-LILATracts_halfAnd10.

Null hypothesis: There is no difference in the proportion of SNAP-recipient households between LILATracts_halfAnd10 and Non-LILATracts_halfAnd10 census tracts. The independent samples t-test returned a t-statistic of 135.77 and a p-value of 0.0. The two-sample Welch’s t-test returned a t-statistic of 122.79 and a p-value of 0.0; again, the p-values were so small the datatype, float64, did not allow enough decimal digits for the leading zeros.

Result - Reject the null hypothesis: There is a significant difference between the sample means.

4.7 Linear and Logistic Regression



(a) Linear regression model accuracy = 80.47%.

(b) Logistic regression model accuracy = 80.71%.

Figure 17: Confusion Matrices for Linear and Logistic Regression Models Using SMOTE

Linear Regression Model w/ SMOTE:

- Precision = 61.65%
- Recall AKA Sensitivity = 81.50%
- F-score = 70.20%
- Specificity = 80.07%

Logistic Regression Model w/ SMOTE:

- Precision = 62.13%
- Recall AKA Sensitivity = 81.04%
- F-score = 70.34%
- Specificity = 80.58%

After applying the SMOTE() object, the linear and logistic regression models exhibit very similar performance measures on all accounts. If we had to choose between the two, the logistic regression model could be a better fit because of its higher precision and F-score.

5 Discussion

5.1 Visualizing and Predicting LILA Tracts

Looking at Figures 1 and 2, it is clear to see that low access, as defined by this dataset, is present in census tracts all across the country. The areas of low access at 20 miles cover much of the western United States which can be initially alarming, but it is important to understand that these are sparsely populated rural areas and if a household has access to a vehicle, then there is not a great problem. This is why Figure 3 can be helpful in finding areas of great need. Households that do not have a supermarket within 20 miles and also do not have access to a vehicle are likely those who are in need of the most aide and this map highlights where those problem areas exist. Government intervention to provide incentives for development in these areas will be crucial as the introduction of a full-service supermarket into a food desert has been found to improve food security as well as nutrition [10].

In our statewide mapping, we see more specific geographical trends arise in this data. The patterns seen in Figures 4 and 5 indicate a correlation between age demographics and low access, with higher levels of both seniors and children being associated with a classification of low access. Figure 6 indicates a relationship between vehicle access and SNAP usage which is explored further through our predictive modeling. It is important to note that the patterns observed in these major cities can not necessarily be generalized to all urban areas, but nonetheless they give a good foundation on which to build predictive models.

While many of our maps show benefits to using geographical data to analyze this problem, Figure 7 highlights a weakness that exists. In this map of the Chicago area, there seems to be a visually strong inverse relationship between SNAP usage and low access. Taking a closer look, the areas of higher SNAP usage are those in the city and specifically those in poorer areas of the city. The tracts classified as not low-access are in many of the same areas, but this is likely due to the fact that the inner city is more densely populated with people and with supermarkets. Simply looking for geographical patterns can lead to misleading conclusions and there clearly need to be more factors considered. Some of the current literature advocates against these geography-based classifications and support looking more into the needs of individual households [15] and after analysis of our mapping, we find strength in this argument.

Using the foundations of relationships in FARA established through mapping, our KNN, random forest, and linear and logistic models help show the efficacy of predicting tracts that are at a food insecure classification. Our KNN model, while not yielding the best accuracy score out of our different models even after accounting for imbalanced data and hyperparamter tuning (62.29%), helped establish possible issues with using predictive models on this data. We looked at all defined distances of LILA in this model and even while resampling the data to diminish the minority effects, Figure 11 shows that the largest number of accurate predictions were in the food secure category. This is not ideal considering the goal of our main research question is to accurately identify food-insecure census tracts. Ultimately, while KNN was chosen for its ease of use and reproducibility, maximum accuracy rates seem to have been achieved for KNN models and remain unsatisfactory for reliably predicting food-insecure areas across the United States.

We see much higher accuracy scores in our random forest (85.54%), linear regression (80.47%), and logistic regression (80.71%) models. All three of these models establish that it is possible to predict LILA tracts. Utilizing hyper-parameterization techniques such as grid search cross validations ensures that these models are obtaining the highest accuracy scores possible on the data. Additionally, all of these models use the SMOTE object to artificially resample the data adjusting to the majority class, further ensuring that the accuracy scores realistically represent the patterns in the data.

The random forest model predicting LILA tracts, in addition to the random forest model predicting SNAP usage, further shows an interesting pattern seen in Figure 14. Much of recent research into these problems has been done through mapping of food deserts, analysis of food environments across geographic areas, and survey responses, all through the lens of racial, ethnic, and income demographics [16]. Both of these random forest models show the strong predictive power of the variable of low vehicle access. These results support Ver Ploeg et al.'s argument that more individualized factors should be examined when looking at food insecurity in the United States [15].

5.2 The Effects of SNAP on Household Food Insecurity

The statistical tests confirm that the proportion and total number of SNAP-recipient households are significantly different within the target variable census tracts versus outside of them. However, this does not corroborate the SNAP program's efficacy, only its larger presence in the target variable areas. SNAP benefits reaching LILA areas is a small victory, but it is only one part of a bigger story.

The side-by-side comparison from Figure 15 suggests two things: SNAP-recipient households are not limited to LILA census tracts, and the number of SNAP-recipient households is greater outside target variable census tracts. These findings align with the Ver Ploeg, Dutko, and Breneman (2015) study, that low-income households are dispersed and not clustered into specific geographical areas [15] and underscores the limitations of areas-based food-access measures.

Because more SNAP-recipient households are outside of labeled LILA areas, other factors contribute to the need for SNAP benefits beyond low-income demographics and low-access measures. Again, this reinforces another sentiment from the Ver Ploeg, Dutko, and Breneman (2015) article: areas-based food access measures overlook low-income individuals living outside flagged LILA tracts [15].

If we speak in terms of geographical regions and distances to grocery stores similar to former food desert studies, the descriptive statistics reveal a wide gap of missed opportunities for the SNAP program, leave room for improvement in identifying other circumstances that lead to applying for SNAP benefits, and call for more informative data collection and outreach investment.

According to the annual cost breakdown of SNAP in Figure 9, the lion's share of federal funding for SNAP covers the benefits cost. Therefore, an increase in federal funding for other initiatives such as research, data collection, and program evaluation and modernization to support the Supplemental Nutrition Assistance Program appears unlikely.

6 Conclusions

Our work confirms that machine-learning techniques can be effective at predicting food-insecure census tracts to target current and developing areas in need of public assistance programs. They have also helped identify a key factor in the food insecurity equation: households without vehicle access. However, more data collection is required to continue and further food insecurity research with predictive modeling and data science methods. Additionally, identifying underlying contributing factors like individual-based food access measures, price variation among states, and time commitment will be necessary to combat food insecurity in the United States and improve upon the current inefficiencies of SNAP.

References

[1] H. Allcott, R. Diamond, J.-P. Dubé, et al. *The geography of poverty and nutrition: Food deserts and food choices across the United States.* Number w24094. National Bureau of Economic Research Cambridge, MA, USA:, 2017.

[2] H. Allcott, R. Diamond, J.-P. Dubé, J. Handbury, I. Rahkovsky, and M. Schnell. Food deserts and the causes of nutritional inequality. *The Quarterly Journal of Economics*, 134(4):1793–1844, 2019.

[3] J. Berrett-Abebe and S. C. Reed. Exploring the relationship between food insecurity, chronic health conditions, and serious mental illness in the united states: Implications for social work. *Health & social work*, 49(3):147–156, 2024.

[4] M. Bitler and S. J. Haider. An economic view of food deserts in the united states. *Journal of Policy Analysis and Management*, 30(1):153–176, 2011.

[5] S. Carlson, J. Llobrera, and B. Keith-Jennings. More adequate snap benefits would help millions of participants better afford food. *Center on budget and policy priorities*, 2019.

[6] . Z. J. Gundersen, C. Food insecurity research in the united states: Where we have been and where we need to go. *Applied Economic Perspectives and Policy*, 40(1):119 – 135, 2018.

[7] C. Gundersen, E. Waxman, and A. S. Crumbaugh. An examination of the adequacy of supplemental nutrition assistance program (snap) benefit levels: impacts on food insecurity. *Agricultural and Resource Economics Review*, 48(3):433–447, 2019.

[8] M. P. Rabbitt, M. Reed-Jones, L. J. Hales, and M. P. Burke. Household food security in the united states in 2023 (report no. err-337). *U.S. Department of Agriculture, Economic Research Service*, 2024.

[9] A. Rhone. Food access research atlas, 2018.

[10] A. S. Richardson, M. Ghosh-Dastidar, R. Beckman, K. R. Flórez, A. DeSantis, R. L. Collins, and T. Dubowitz. Can the introduction of a full-service supermarket in a food desert improve residents' economic status and health? *Annals of epidemiology*, 27(12):771–776, 2017.

[11] J. D. Shenkin and M. F. Jacobson. Using the food stamp program and other methods to promote healthy diets for low-income consumers, 2010.

[12] T. A. Smith and C. A. Gregory. Food insecurity in the united states: measurement, economic modeling, and food assistance effectiveness. *Annual Review of Resource Economics*, 15(1):279–303, 2023.

[13] C. A. Swann. Household history, snap participation, and food insecurity. *Food Policy*, 73:1–9, 2017.

[14] E. R. S. USDA. Snap data tables, 2019.

[15] M. Ver Ploeg, P. Dutko, and V. Breneman. Measuring food access and food deserts for policy purposes. *Applied Economic Perspectives and Policy*, 37(2):205–225, 2015.

[16] R. E. Walker, C. R. Keane, and J. G. Burke. Disparities and access to healthy food in the united states: A review of food deserts literature. *Health & place*, 16(5):876–884, 2010.

7 Code Appendix

Project Code: `https://github.com/hiamoreena/An-Examination-of-Food-Insecurity-in-the-United-States.git`