PREDICTING FOOD DESERT CENSUS TRACTS USING MACHINE LEARNING AND CENSUS DATA

Katie Henning

khenning@umich.edu

SI 671: Data Mining

School of Information

The University of Michigan

December 13, 2021

ABSTRACT

Food deserts are geographic areas which lack convenient access to affordable, healthy food. They are defined at a census tract level which are smaller geographic spaces than state counties and have a maximum population of 8,000 residents. In addition to low access to food, those who live in food deserts are disproportionately impacted by poverty. For my project, I proposed the creation of a machine learning model to predict food deserts using the data collected from the 2015 American Community Survey-5 year estimates. I also analyzed food deserts in the United States, especially the factors which lead to their existence and the demographics of citizens which occupy these areas. From learning which features are the most important when predicting food desert status, policy choices could be better informed to prevent this scarcity from occurring as well as help already at-risk target populations.

Key words: Machine Learning, Census, American Community Survey, Poverty, Food Access

1 Introduction

The USDA defines food insecurity as "consistent lack of access (both physical and economic) to enough food for an active and healthy life"[1]. In recent decades, levels of food insecurity have been in fluctuation and for the most part have been correlated with macro-economic events. For example, the rate of food insecurity increased from 11 percent in 2007 to 14.6% in 2008 and peaked at 14.9% in 2014 due to the lasting impacts of the 2008 recession [2]. Combined with rising rates of unemployment at this time, and cost of living factors such as gas and food, a great deal of strain was put on community programs like food pantries to provide for those impacted by these economic shifts [9]. Additionally, the percent of Americans using Supplemental Nutrition Assistance Program (SNAP) funds increased from 4% to 15% in 2010 [9]. Rural areas of the United States in particular have been experiencing changes in population and employment opportunities which would bring the need, as well as the purchasing power, to retain grocery stores and supermarkets.

As younger generations leave these areas in search of higher paying jobs and the remaining generations age, there is not as strong of an economic incentive to keep local grocery stores which the community once relied on for access to food in operation [9]. These retailers are being replaced with convenience stores and the gap between need and access that this creates struggles to be filled by food pantries and assistance from family and friends. Even when members of rural communities do visit food assistance programs, they are typically given a smaller amount of food compared to visitors of urban food assistance locations due to the smaller amount of food donations available in areas that are less densely populated [9].

In 2019, due to the efforts of non-profit organizations and food banks, the level of food insecurity reached a 20 year low of 10.5% [2]. However this comparatively low percentage did not hold for long due to the hardships of securing food for children, families, and individuals during the COVID-19 pandemic. The non-profit organization "Feeding America" (which in a December 2020 Forbes report was listed as the second largest charity in the United States by revenue) projected that in 2021, 46 million people (1 in 8) including 13 million children (1 in 6) experienced hunger and food insecurity [3, 4]. A large source of this insecurity is the existence of food deserts. 39.5 million people or 12.8% of the US population lived in a food desert as of 2017 [6]. The lack of affordable, nutritious food options also factors heavily into the rise of diet-related diseases as people have fewer options to make healthy decisions based on their limited budget, transportation options, and store choices [5].

Areas are likely to be disproportionately impacted by food deserts if they have a high level of poverty, a small population, or if they are classified as "rural" [6]. Residents of food deserts have disproportionately lower levels of education and higher rates of unemployment. There are also large disparities in food insecurity based on racial makeup of communities. 21.6% of Black individuals (1 in 5) experienced food insecurity in their life compared to 12.3% of white individuals (1 in 8) [4]. Common statistical measures of food deserts include the average income of a given area, the distance to the nearest store in the area, and the transportation options most commonly available to physically access these stores whether that be a personal vehicle or public transportation [6]. There are also 4 sub-divisions of food security whose titles describe varying degrees of need and can help to more specifically classify these areas: high food security, marginal food security, low food security, and very low food security [1].

This project uses demographic and economic data collected by the United States Census Bureau and the United States Department of Agriculture to train a supervised machine learning classification model to predict food deserts on the census tract level. Machine learning tools offer insight into what factors are the most important to consider when classifying phenomenological data. Thus, this model provides statistical support for those trying to target key groups and aid in their path to achieving high food security.

2 Related Works

While the prediction of food deserts using machine learning methods has been conducted previously, these models utilized different means and measures than the model presented herein. A 2021 publication by Modhurima DeyAmina, Syed Badruddozaa, and Jill J.McCluskey also uses census data to predict food deserts at census tract resolution by predicting a factor known as the modified Retail Food Environment Index score (mRFEI) [13]. This is a measure of the relative amount of healthful food retailers in a given census tract. They divide limited access to healthful food retailers into two classifications: food deserts and food swamps. Food deserts have no access to healthful food retailers and food swamps have very little access to healthful food retailers [13].

$$mRFEI = 100 \times \frac{number\ \ of\ healthful\ retailers\ in\ tract}{number\ \ of\ healthful\ retailers\ in\ tract + number\ of\ unhealthful\ retailers\ in\ tract} \tag{1}$$

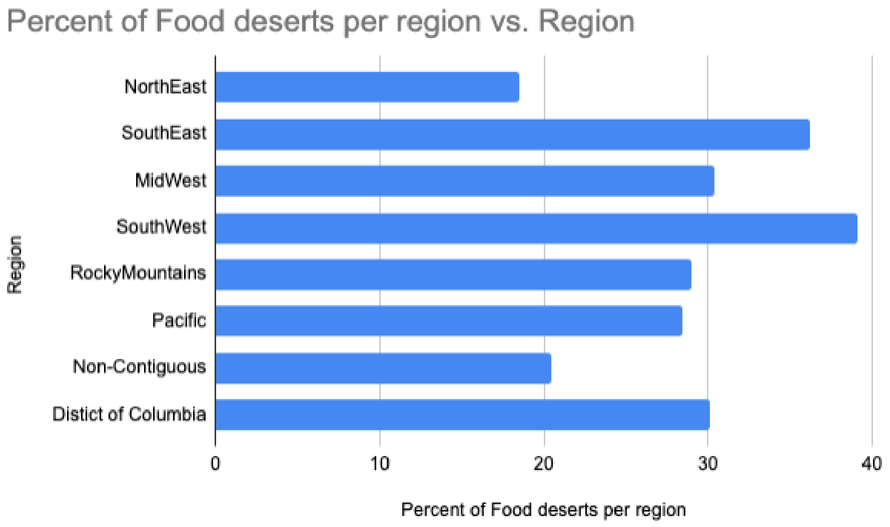
This measure was originally calculated for the purposes of the Centers for Disease Control and Prevention's Children's Food Environment State Indicator Report and is defined only by the types of food stores within a given census tract and does not represent the dominant choices of the consumers who purchase food at these locations or the availability of food at these locations [13]. The authors of this paper state that they used data from the 2010 American Community Survey and narrowed data scope to the top 50 most informative features. Their model was able to present food deserts and food swamps based on mRFEI scores with 72% accuracy [13]

An author who has published a considerable body of work regarding census data and food deserts is Michele Ver Ploeg, who was formerly the Acting Assistant Administrator of the US Department of Agriculture's Economic Research Service. She has more recently been appointed as the Director of the George Washington University Food Policy and Health Institute. Her work commonly focuses on the impacts of food deserts on the elderly population, analysis of the use of SNAP benefits, and the long term nutritional impacts of food deserts on communities. Her insights on these features provide context to the existence of food deserts and how their residents are impacted by this scarcity. Multiple papers authored and co-authored by Ver Ploeg, especially "The Impact of Food Deserts on Food Insufficiency and SNAP Participation Among the Elderly", aided in my understanding of broader applications of the policies that already exist surrounding food deserts and how the results of my statistical analysis can be further interpreted.

3 Data & Analysis

The two datasets I utilized in this project were the 2015 American Community Survey (ACS) 5-year estimates and the 2015 United States Department of Agriculture Economic Research Service Food Access Atlas. The ACS is a demographic survey conducted by the United States Census Bureau [7]. The "5-year estimate" indicates this data was collected over a period of five years or 60 months [15]. In this case, the data was collected from the beginning of 2011 to the end of 2015. Though this data might not be as current as a 1-year or a 3-year estimate, it is more accurate and therefore the most reliable [15]. Additionally, where 1- and 3-year estimates only collect data for areas that surpass a certain population threshold, the 5-year estimates collect data for all areas [15]. I specifically used two tables from this survey, "Economic Characteristics" (CP03) and "Demographic Characteristics" (CP05). This data is collected at a census tract level which is typically higher in spatial resolution than even a state county. census tracts are based on population size with a minimum of 1,200 residents in a given area, a maximum of 8,000 and an average of 4,000 [14]. census tracts have been used for over 100 years and were originally used only in major cities while block number areas covered all other areas. They became the official micro-geographic entity for the entire US in 2000. They remain relatively stable but can be split or merged based on population count as needed [14].

The features from this dataset include measures such as the total population of a tract, the number of men, the number of women, a count of residents who self-identify as one of 6 ethnicities (Black, Asian, Pacific, Hispanic, Native, white), poverty rate, child poverty rate, number of residents employed and unemployed, and more. This dataset provided a total of 37 features. Some features, such as "State" and "County," were dropped from the data used to train the model as I was more interested in the impact of the other numeric features than the names of the states and counties in a one-hot encoded format. I did however conduct some exploratory data analysis to see which geographic regions of the United States had the highest percentage of census tracts designated as food deserts. The region with the highest percentage of their tracts designated as food deserts is the Southwest. Because food deserts are partially based on income, I also dropped all income related measures from this set of features to mitigate data leakage. My goal in doing this was to look purely at the demographic data and economic-adjacent features to predict food desert census tracts.

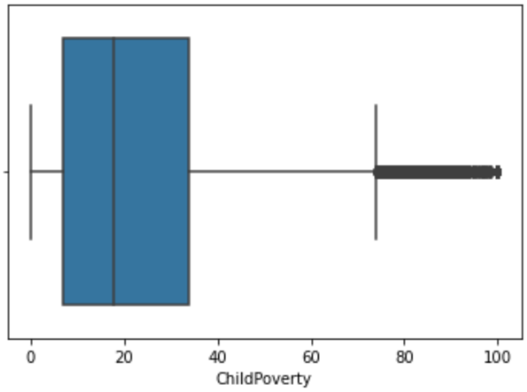


The other dataset used in this project was a Food Access Atlas which was also collected at a census tract level [8]. This Atlas dataset denoted which tracts were denoted as food deserts or not based on four variations of low-income and low-access measures. These four measures take into account the multiple different ways a census tract could be designated as a food desert. These measures are varied to allow for the different definitions of 'low-access' that could exist based on the census tract's urban/rural classification or lack of transportation. The four measures of low-access which mark a census tract as a food desert are the following:

1) "Low-income census tracts where a significant number (at least 500 people) or share (at least 33 percent) of the population is greater than ½ mile from the nearest supermarket, supercenter, or large grocery store for an urban area or greater than 10 miles for a rural area." [12]
2) "Low-income census tracts where a significant number (at least 500 people) or share (at least 33 percent) of the population is greater than 1.0 mile from the nearest supermarket, supercenter, or large grocery store for an urban area or greater than 10 miles for a rural area." [12]
3) "Low-income census tracts where a significant number (at least 500 people) or share (at least 33 percent) of the population is greater than 1.0 mile from the nearest supermarket, supercenter, or large grocery store for an urban area or greater than 20 miles for a rural area." [12]
4) "A fourth and slightly more complex measure incorporates vehicle access directly into the measure, delineating low-income tracts in which a significant number of households are located far from a supermarket and do not have access to a vehicle. This measure also includes census tracts with populations that are so remote that, even with a vehicle, driving to a supermarket may be considered a burden due to the great distance. Under this measure, a tract is considered low access if at least 100 households are more than ½ mile

> from the nearest supermarket and have no access to a vehicle; or at least 500 people or 33 percent of the population live more than 20 miles from the nearest supermarket, regardless of vehicle access." [12]

A tract is considered low-access or a food dessert if just one of these four indicators is marked in the affirmative. I used this data to create a binary label, "0" (not flagged as a food desert) or "1" (flagged as a food desert) for each tract. This was the label to be predicted by the classifiers I trained. I also kept select columns from this dataset that were not present in the ACS data such as if a tract was considered "Urban" or not, the number of housing units using SNAP benefits in each tract, the number of children in each tract (ages 0-17), and the number of seniors in each tract (ages 65+). The ACS 5-year estimates had some missing data within the features. The percent of rows with missing data was too high to simply drop these rows from the dataset and so to replace the absent values, I used the sklearn method "Simple Imputer" to fill these values with the most commonly occurring value. I chose this method of imputing because when I graphed the range of values for a few features, I observed that there was quite a wide range. I therefore did not choose another method of imputing such as 'mean' because the higher values would give a false representation of the average missing value. I mapped the features from this dataset to the ACS data by using the unique 10 digit code for each census tract. In total, there were 72,839 rows of data (one for each census tract) and 35 features.



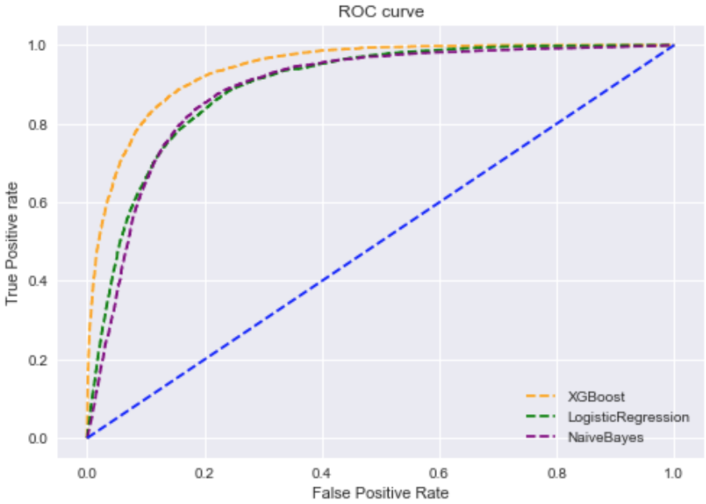
Distribution of Child Poverty rates across all tracts

With the missing data imputed, the set of features finalized and the label or 'y' values created, I then began training classifier models to predict the previously described census-tract-resolution binary food desert label. I trained these models using feature selection (using the top 10 features as designated by a ChiSquare test) as well as on the entire set of features to determine if this would have an effect on accuracy. The models trained were XGBoost, Logistic Regression, Naive Bayes, and a baseline Dummy Classifier.

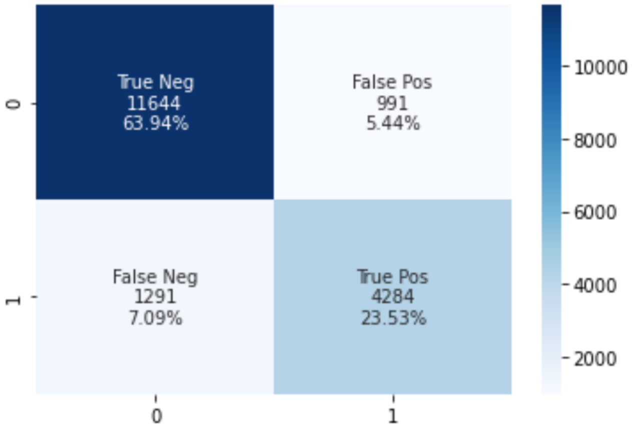
4 Findings

The result of the ChiSquare test showed that of all total features, the 10 which were selected as the most informative were 'TotalPop', 'White', 'Black', 'Citizen' (number of citizens), 'Poverty', 'ChildPoverty', 'Employed' (number of people employed), 'PovertyRate', 'Seniors', and 'Snap'. When trained with just these 10 features, the accuracy and the Area Under the (ROC)

Curve (AUC) scores were lower than the accuracy and AUC scores of the models which were trained using the entire dataset. The model with the highest accuracy score and the highest AUC score was XGBoost with an accuracy of 87.5% and an AUC score of 94.1% percent. The mean score of a 5-fold cross validation on this model was 87.4%. A confusion matrix was also generated for this top performing model and the precision, recall, and F1 scores were calculated in order to better understand the model's various tradeoffs. Those scores were all fairly comparable.



ROC Curve for all classifiers

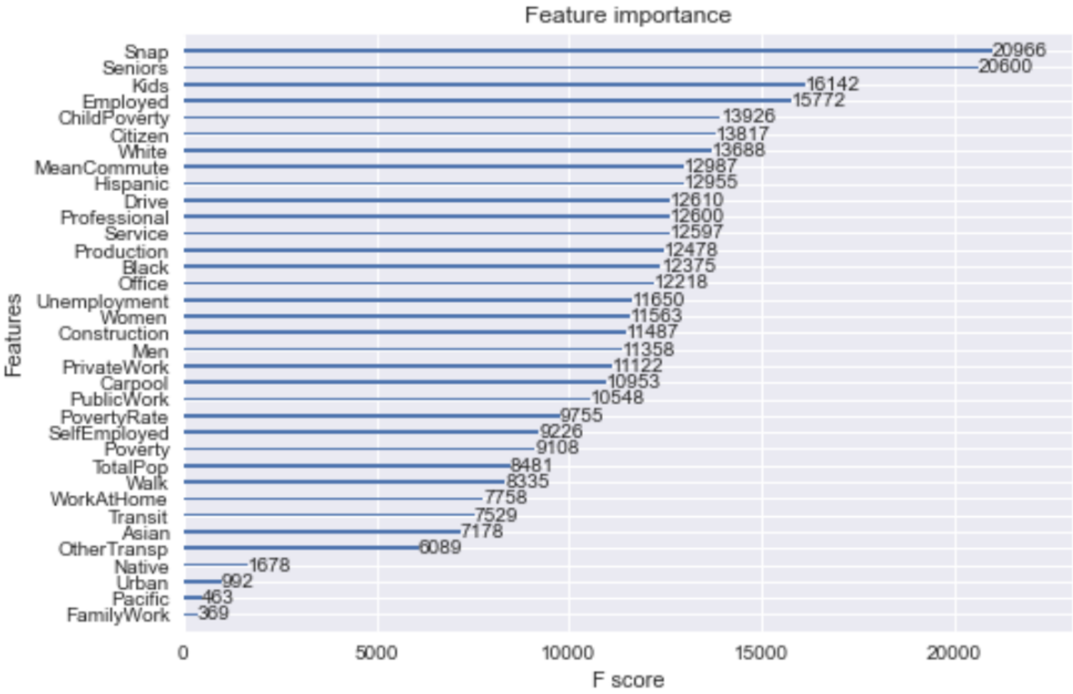


Confusion Matrix for XGB model

Precision	Recall	F1-Score
0.812	0.768	0.790

Feature importance analysis was also conducted to determine how important a given feature actually was in aiding the model to make its classifications. The top two features (housing units receiving SNAP benefits and the number of Seniors residing in a census tract) clearly had a large impact on the result of the classifications. The least useful features were the percentage of residents participating in unpaid family work, the amount of residents who self-identified as Native Hawaiaan or Pacific Islander, and surprisingly, if the tract was flagged as being an Urban tract. Since this was a binary classification, 1 represented an Urban tract and 0 represented that the tract is not Urban and so might be considered Rural instead. It is interesting that this feature carried such little weight in the designation of food desert status because the research from the USDA suggests that rural census tracts are more likely to be classified as food deserts.

Feature Importance for XGB model



Sensitivity analysis was performed by changing key hyperparameters for the XGBoost model to determine how dependent the model's performance was on their tuning. To test this, I trained three additional XGBoost models varying the parameters 'n_estimators' and 'max_depth' both separately and together (changing each one independently and changing both simultaneously). With these changes, the model's accuracy score based on comparing a sample of predictions with the actual label classification changed only 3% at maximum. This helped me to check that the robustness of the model rested in the features rather than small changes in parameter tuning.

5 Discussion

The results of feature importance analysis showing that the number of seniors that live within a given tract have a large role in determining the tract's status as a food desert stands to reason since the elderly are more susceptible to limited transportation if they are no longer able to drive, live on fixed incomes such as a pension, and are unlikely to leave an area even if it is facing food insecurity due to long-term connections within their local city or neighborhoods [10]. The elderly population is also less likely to be enrolled in receiving SNAP benefits as only 35.1% of eligible elderly adults are enrolled compared to 75.6% of SNAP-eligible non-elederly adults [10]. Of all of these factors, access to transportation is the most inhibiting for the elderly to achieve food security as "elderly food desert residents that do not own a vehicle were 12 percentage points more likely to report food insufficiency than otherwise similar food desert residents who owned a vehicle" [10]. This suggests that a key goal to strive for in decreasing the number of census tracts classified as food deserts with a high percentage of elderly residents is to improve the quality of transportation so this demographic has the opportunity to choose from more diverse food retailers.

SNAP benefits being a large indicator of food desert status also has contextual support because though the purpose of SNAP benefits is to increase access to food that is nutritionally dense, the ability to use SNAP benefits heavily depends on the existence of food retailers nearby where these benefits can be redeemed. Even when SNAP benefits are increased, other factors such as transportation serve as an impediment to residents of food deserts being able to benefit from this program [11]. Because of this, there is reason to believe that the best way to address this gap in food access is to provide economic incentivization to good retailers to remain or open new locations in these food-insecure locations [11].

6 Conclusion

While the factors that contribute to the existence of food deserts can be confounding, the root cause of their existence is by no means a mystery. Those who live in census tracts with lower incomes are not able to afford healthy nutritious food such as fresh produce which is more expensive than less-nutritious, processed foods. Because grocery stores and supermarkets are subject to the demands of the economic market, it does not serve their interests to continue to operate a location that does not draw in a large amount of profit. Yet, the people who live in these lower-income census tracts where there is not a large amount of diversity in food retailers still deserve access to foods that are necessary to fuel a healthy lifestyle. Subsidizing the shoppers, while a wonderful goal that is giving residents a direct increase in their food budget, is not the best option as other factors can reduce the amount of positive impact programs like SNAP can ultimately achieve. One direct improvement that can be made that addresses physical inequity is that instead, government aid opportunities should bring retailers more directly into these conversations and determine what it would take for their stores to open in these locations. In the mean-time, transportation issues to food retail locations should be targeted. Access to healthful food choices in lower-income communities would help children receive education about the importance of nutrition that could impact them for the rest of their lives.

Moving forward, this project would benefit from access to more current data from the American Community Survey of 2020. Unfortunately, the pandemic delayed plans for the release of this data until Nov 30, 2021 , or that would have been the data utilized instead. Adding features to this dataset which label tracts as being subject to specific food and health policies would also add another level of depth to the scope of this project. This would provide another source of analysis for policy makers to consider when addressing the problem of food deserts.

References

[1] Feeding America.(n.d.).What is Food Insecurity? Hunger and Health: FeedingAmerica. Retrieved from https://hungerandhealth.feedingamerica.org/understand-food-insecurity/.

[2] United States Department of Agricul-ture.(n.d.).Food Security and NutritionAssistance. Economic Research Service. Retrieved from https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/food-security-and-nutrition-assistance/.

[3]Barrett,W.P.(2020 ,December 11). America'sTopCharities.forbes.com. Retrieved November 2, 2021, from https://www.forbes.com/lists/top-charities/6dd172a15f50.

[4] Feeding America.(2021, March).The Impact of the Coronavirus on Food Insecurity in 2020 amp; 2021. feedingamerica.org. Retrieved November 2, 2021, from https://www.feedingamerica.org/sites/default/files/2021-03/National

[5] Weatherspoon, D. D., Dutko, P., amp;Ver Ploeg, S. (2012). An Evaluation of Food Deserts in America. Choices, 27(3), 1–4.

[6]Annie E. Casey Foundation. (2021,February 13).Food Desert in the United States [web log].Retrieved November 2, 2021, from https://www.aecf.org/blog/exploring-americas-food-deserts.

[7] US Census Demographic Data (n.d.).Retrieved November 2, 2021 from https://www.kaggle.com/muonneutrino/us-census-demographic-data.

[8] USDA Economic Research Services(2015). 2015 Food Access Research Atlas Data. Retrieved from https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/

[9] Whitley, Sarah (2013). Changing Times in Rural America: Food Assistance and Food Insecurity in Food Deserts. https://www.tandfonline.com/doi/full/10.1080/10522158.2012.736080?casa_token=-CUIYPQLYJwAAAAA%3AFHm_NV6RNV8DoeZpT45hX_LmRffY0k_RqJDEFOUwBfjYHCewdpHZ0kteRweShS5-d06u0wojx1j5RjQ

[10] Fitzpatrick, Katie, Greenhalgh-Stanley, Van Ploeg, Michele (2015) The Impact of Food Deserts on Food Insufficiency and SNAP Participation among the Elderly. In *American Journal of Agricultural Economics. I Volume:98 Issue: 1.* 19-40.

[11] Andres, Margaret, Bhatta, Rhea, and Van Ploeg, Michele (2012). An Alternative to Developing Stores in Food Deserts: Can Changes in SNAP Benefits Make a Difference? In *Applied Economic Perspectives and Policies. Volume:35, Issue: 1.* 150-170.

[12] USDA Economic Research Services(2015). 2015 Food Access Research Atlas Documentation. Retrieved from https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/

[13] DeyAmin, Modhurima, Badruddoza,Syed, and McCluskey, Jill J. (2021). Predicting access to healthful food retailers with machine learning. In *Food Policy, Volume: 99.*

[14] U.S. Census Bureau (n.d.) Census Tracts: Geographic Products Branch. Retrieved from https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf.

[15] U.S. Census Bureau (n.d.) When to Use 1-year or 5-year Estimates. Retrieved from https://www.census.gov/programs-surveys/acs/guidance/estimates.html.