Emily Donofrio
August 3, 2025
MESA8440

Analysis of Food Desert Indicators

Data Sources for Multivariate Statistical Analysis

1. USDA Food Access Research Atlas: Food Access Research Atlas

The Food Access Research Atlas is a census tract level dataset that is comprised of various potential food access indicators and census data. The access indicators are derived from the 2019 STARS (Store Tracking and Redemption System) directory of stores authorized to accept SNAP benefits and the 2019 Trade Dimensions TDLinx directory of stores. Population data are from the 2010 Census of Population and Housing. This data contains many binary flags, some categorical data such as whether a tract is urban or rural, and the geographic state and county of the tract. Additionally, there are many continuous variables for socioeconomic and demographic data. This data can be used for many multivariate analyses. The binary flags can be used as the dependent variables in classification techniques such as logistic regression and discriminant analysis. The census data could be used for PCA, factor analysis, or clustering. Potential limitations are from potential aggregation bias. Additionally, the data is a combination of data collected from 2019 and 2010 so it does not represent a singular snapshot in time.

2. National Health and Nutrition Examination Survey (CDC): NHANES

The National Health and Nutrition Examination Survey is conducted yearly and consists of survey answers as well as information from medical examinations for a random sample of Americans. This dataset contains many data types, namely ordinal variables from the survey questions, as well as continuous variables in from the medical exams (eg: BMI). This data lends itself to MANOVA analysis, various subsets of the dataset could also be used for factor analysis or clustering. Logistic regression or discriminant analysis could also be done with this data. A limitation of this data is the complexity of the survey design which some question paths being conditional on others. Also, the survey structure changes year to year to any time related analyses may have limitations. Additionally, most data is self-reported and thus is vulnerable to bais.

3. IPUMS USA: IPUMS

IPUMS USA collects data from decennial censuses and American Community Surveys. The website allows for user driven dataset creation, with the ability to search for and select the variables you are interested in. There is a wide selection of variables to choose from spanning many different types including categorical, continuous, and ordinal. This data could be used for many multivariate techniques including factor and cluster analysis as well as logistic regression and discriminant analysis. A limitation of this data source is that the user will need to understand the collection method of each variable as well as the granularity (i.e. individual or household), and the data will likely involve a lot of preprocessing to create a single cohesive and usable dataset.

4. Gapminder: Gapminder

Gapminder has data for various metrics at the country and year level. Majority of variables are continuous (e.g. food supply, child mortality rate). Country is a nominal variable in every Gapminder dataset. The multivariate techniques that this data lends itself best to is PCA, factor analysis, and cluster analysis. A limitation of this data is that it is at the country level so it does not support individual level research questions. Additionally, Gapminder combines data from various sources and fills in the gaps where it has missing data. This needs to be taken into consideration if fully raw data is needed.

5. PSID: Panel Study of Income Dynamics (PSID) Series

The Panel Study of Income Dynamics Series is a longitudinal study following a representative sample of US individuals and families since 1968 collecting data on numerous topics including income, expenditures, family, education, and more. There are many different variable types including continuous (cost of food at home), ordinal (food security score), and categorical (received food assistance indicator) variables. Multivariate techniques that could be used with this data include MANOVA, PCA, factor analysis, and clustering. Classification methods such as logistic regression and discriminant analysis could also be used. A limitation is, similarly to NHANES and IPUMS there is a mix of family and individual level data which adds complexity to designing a researcher's desired dataset. Additionally, self-reported data is vulnerable to bias.

Dataset Selection and Data Analysis Questions

Food deserts are "geographic areas where residents' access to affordable, healthy food options (especially fresh fruits and vegetables) is restricted or nonexistent due to the absence of grocery stores within convenient traveling distance" (Food Empowerment Project).

There is no one way to define a food desert based on empirical measures. Various combinations of thresholds of distance to supermarkets, income level, and vehicle access are available as potential food desert indicators.

A group of researchers is interested in understanding what types of areas typically qualify as food deserts using these indicators. Are the trends consistent for all food desert indicators?

The researchers decide to look at 3 commonly used food desert indicators for this analysis:

Focus solely on supermarket distance:
LA1and10 - Flag for low access tract at 1 mile for urban areas or 10 miles for rural areas

Focus on supermarket distance and income-level:
LILA1and10 - Flag for low-income and low access when considering low accessibility at 1 and 10 miles

Focus on supermarket distance and vehicle accessibility:
LAVehicle20 - Flag for tract where >= 100 of households do not have a vehicle, and beyond 1/2 mile from supermarket; or >= 500 individuals are beyond 20 miles from supermarket; or >= 33% of individuals are beyond 20 miles from supermarket

The dataset provided by the USDA Food Access Research Atlas contains census tract level food desert indicator data for 2019. The dataset also includes various tract level data points from the 2010 census. You may choose how to best utilize this data for this analysis.

You are invited to consult with the research team to answer the following questions:

a) Describe any preprocessing of data you performed and the reasoning.
b) Was there any missing data? How did you handle it? What about outliers?
c) Many of the census data variables are correlated with each other. Choose a method to mitigate this and make the data more interpretable. Explain your reasoning.
d) What assumptions did you check before this analysis? Were they passed?
e) Explain your process of deciding the final set of variables.
f) Describe your final factors extracted from the analysis. How many are there? Which variables load strongest on them? How do you interpret what each factor represents?
g) Choose a method for analyzing the relationship between these factors and the food desert indicators. Explain your reasoning. Do you plan to add additional features? Why?
h) Interpret your model results:
    a. State your null hypotheses.
    b. Present your model coefficient outputs.

c. Interpret the significance and direction of each coefficient. What does this say about the relationship between the feature and that food desert indicator?

i) Compare and contrast the results of the models. State your conclusions. What are the potential ethical considerations for choosing a food desert indicator for analyses or reports?

Food Desert Data Analysis

a) Describe any preprocessing of data you performed and the reasoning.

I first examined the provided data dictionary to determine which variables would be relevant to this analysis. I chose to keep all the continuous demographic and socioeconomic census data that was independent of any food desert low access indicators.

Many of these variables were population counts of different demographics. I made the preprocessing choice to convert these variables to the percentage of the tract population. I did this because leaving the raw population counts would make each of these variables very correlated with the population, but by normalizing by population they instead provide information about the prevalence of that specific demographic group. This aids comparison between tracts of different sizes.

Another step I took was standardizing the data to all be on the same scale. This ensures that the analyses are not influenced by varying unit sizes. This is especially important for any variance-based analysis such as a factor analysis. Without standardizing, a variable with a larger unit scale could disproportionately affect the results.

b) Was there any missing data? How did you handle it? What about outliers?

There were some missing values. Many variables had 4 missing values. This dataset contains 72531 observations, meaning that these 4 missing values constituted a 0.005% rate of missingness. All 4 missing values occurred in the same rows, so I dropped those rows, losing only 4 observations in total.

After removing these rows there were two variables with remaining missingness. PCTGQTRS has 21 missing values, and MedianFamilyIncome had 745 missing values. I considered doing a simple mean or median imputation for these values but first I wanted to check if the missingness was disproportionate. I decided to look at the poverty rate for the rows with and without MedianFamilyIncome missing values as they have a strong conceptual correlation. I found that the distribution of poverty rate for tracts with missing MedianFamilyIncome was much higher than the distribution of poverty rate for tracts without a missing MedianFamilyIncome. This told me that a simple imputation would not be a good choice to deal with missingness as the missingness was not random and simple imputation could introduce bias. I instead chose to use K Nearest Neighbors for imputation as it would impute a more realistic value based on the similar census tracts. After this, there were no more missing values.

I then investigated outliers. All variables had many outliers based solely on boxplot analyses. This is not shocking due to the very large sample size. Further analysis showed that the outliers were spread among the observations, with many rows having one outlier, rather than specific observations having many outliers. This indicates the outliers are likely due to natural variability rather than structural anomalies or data entry errors. Based on this reasoning, I decided not to remove any outliers to avoid any unnecessary data loss. A potential robustness check that could be performed as a future step of this analysis is rerunning the analysis after removing the outliers and seeing if it produces different results.

c) Many of the census data variables are correlated with each other. Choose a method to mitigate this and make the data more interpretable. Explain your reasoning.

There are many different but related variables in this data. I chose to use Exploratory Factor Analysis (EFA) to both address the multicollinearity and improve the interpretability of the data. EFA is a dimension reduction method that extracts latent factors from the data that consist of different linear combinations of the variables. Latent factors are unobserved variables that explain the shared variance among the observed variables. These resulting factors will be uncorrelated with each other which is useful for avoiding multicollinearity which can be obstructive in many analysis techniques. Additionally, the factors will each represent a conceptually sound feature of the data, in this case census tracts. This aids interpretation as we can get a more holistic understanding of the underlying structure of census tracts. Rather than having to interpret the relationships of each individual variable, we can utilize the latent factors leading to a simpler and more practically relevant interpretation for understanding patterns in the data.

d) What assumptions did you check before this analysis? Were they passed?

The two most important assumption tests for Exploratory Factor Analysis are Bartlett's Test of Sphericity and KMO. Bartlett's test tests for significant correlation among the variables sufficient to perform an EFA. KMO measures the Measure of Sampling Adequacy (MSA) which predicts if the data will factor well based on correlation and partial correlation.

The first step I took was to look at the correlation matrix. There were 3 variables that did not have any correlations with an absolute value greater than 0.3. These variables are TractAsian, TractNHOPI, TractAIAN. PCTGQTRS only had one correlation with an absolute value greater than 0.3. I decided to remove these variables. I then ran the assumption tests on the new subset of variables.

The Bartlett's Test chi-square was 51322 and the p-value was well below the threshold of 0.05 indicating that there is sufficient multicollinearity for an EFA. The KMO is 0.68 which is in the mediocre but acceptable range to proceed with an EFA. As I proceeded with the EFA I ended up further pruning the variables. The results of the assumption tests with the final variable selection were a Bartlett's test Chi-Square of 561568 and p-value < 0.001 and a KMO of 0.65. This KMO is still in the mediocre but acceptable to proceed range of KMOs.

e) Explain your process of deciding the final set of variables.

I iterated through many different options for performing this EFA, making decisions based on statistical and theoretical reasoning. In every iteration I used principal factor extraction as the extraction method as the goal is to understand the latent structure. The first EFA I performed was with all variables and without rotation. The resulting factors were not very interpretable and had a lot of cross-loadings. This led me to go forward with varimax rotation for all future tests as it would improve the practical interpretability of the factors which is crucial for this analysis. The resulting EFA resulted in 5 factors, although the 5th was borderline using Kaiser's rule as it has an eigenvalue of 1.001. After parallel analysis

this fifth factor loses significance. Looking at the strong factor loadings, it also lacks conceptual clarity.

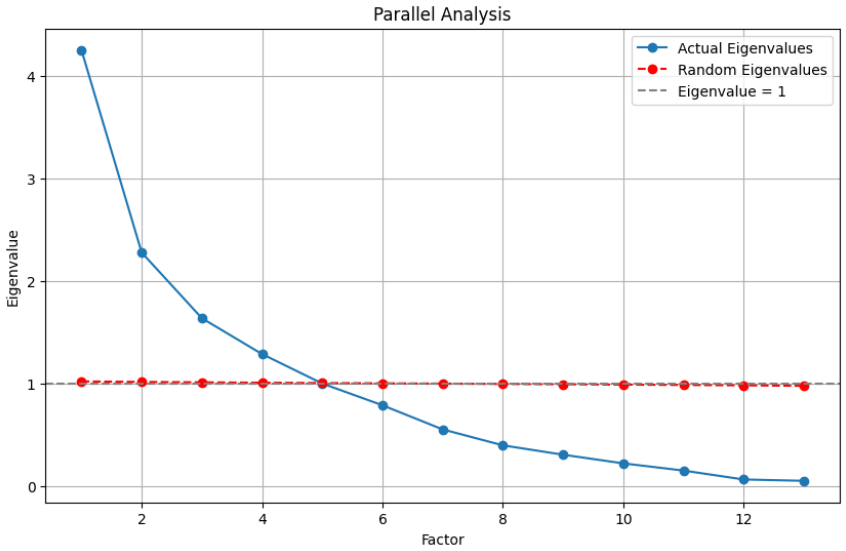

Chart I. Parallel Analysis of Initial EFA with Varimax Rotation

I looked at the communalities and factor loadings of the different variables to examine whether further variable pruning may improve the EFA. TractKids, TractSeniors, and TractHUNV (households without vehicles) were three variables that had the lowest communalities, loaded weakly across multiple factors, and contributed primarily to the conceptually vague fifth factor. For example, TractSeniors had the lowest communality (0.57), while TractKids and TractHUNV also had high factor score complexity**,** suggesting they lacked clear alignment with any single factor. Conceptually, these 3 variables don't necessarily aid the interpretation of the latent factors they load on. I chose to remove them and see how the EFA changes. The updated EFA has 4 factors with eigenvalues greater than 1 and they each have a clear theory-based meaning.

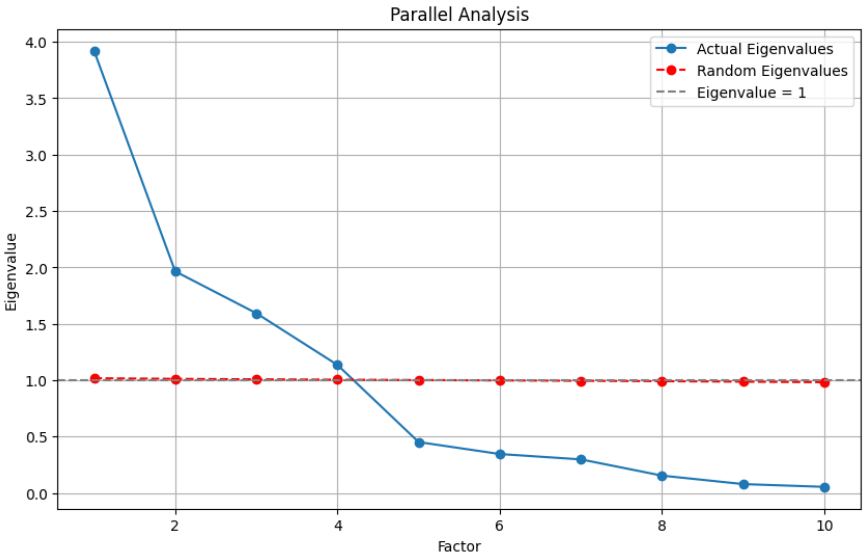

Chart II. Parallel Analysis of EFA with Varimax Rotation after Variable Pruning

f) Describe your final factors extracted from the analysis. How many are there? Which variables load strongest on them? How do you interpret what each factor represents?

There are 4 factors in my final EFA. Cumulatively they explain 86% of the variance.

a. Socioeconomic Need:
   i. Strong positive loadings: TractSNAP, TractLOWI, PovertyRate.
   ii. Strong negative loadings: MedianFamilyIncome
   iii. This factor measures economic hardship and reliance on public assistance.
b. Tract Size:
   i. Strong positive loadings: Pop2010, OHU2010
   ii. This factor captures the overall population and housing density of a tract. This functions as a measure of geographic and demographic scale.
c. Ethnic Diversity:
   i. Strong positive loadings: TractOMultir, TractHispanic
   ii. Mediocre/Weak negative loadings: TractWhite, TractBlack
   iii. This factor represents prevalence of ethnic diversity, specifically non-majority white or black populations. It is primarily driven by Hispanic and multiracial populations.
d. Racial Composition
   i. Strong positive loading: TractBlack
   ii. Strong negative loading: TractWhite
   iii. This factor represents the racial composition balance between Black and White residents in the tracts, distinguishing tracts with higher proportions of Black residents from those with higher proportions of White residents.

g) Choose a method for analyzing the relationship between these factors and the food desert indicators. Explain your reasoning. Do you plan to add additional features? Why?

I am opting to use Binary Logistic Regression to analyze the relationship between tract features and food desert indicators. I will build 3 separate logistic regression models, one with each of the chosen food desert indicators as the dependent variable. I chose logistic regression for a few reasons. First, the goal of this analysis is to interpret the relationships, and logistic regression provides a readable and easily interpretable output in terms of log-odds. Additionally, logistic regression is more robust to assumption violation than a method such as Linear Discriminant Analysis. The data I am using is observational, not from a designed experiment, meaning it likely would violate the LDA assumptions.

I plan to add one additional variable, a categorical indicator of whether a tract is urban or rural. Urban tracts are coded as 1, rural tracts as 0. This feature has strong conceptual relevance but is not necessarily captured in the factors from the EFA.

After adding urban status I rechecked the multicollinearity assumption by looking at the VIF. The VIF for all features was between 1.1 and 2.5 which are well below the threshold of 5, indicating that there is not problematic multicollinearity.

h) Interpret your model results:
   a. State your null hypotheses.
   b. Present your model coefficient outputs.
   c. Interpret the significance and direction of each coefficient. What does this say about the relationship between the feature and that food desert indicator?

a) Null Hypotheses: The null hypotheses are identical for all 3 models but the food desert indicator is interchanged.

H_0:$\beta_{SocioeconomicNeed}=0$

After controlling for all other predictors, there is no relationship between socioeconomic need and whether a census tract is a food desert (LA1and10, LILA1and10, or LAVehicle20).

H_0:$\beta_{TractSize}=0$

After controlling for all other predictors, there is no relationship between census tract size and whether a census tract is a food desert (LA1and10, LILA1and10, or LAVehicle20).

H_0:$\beta_{EthnicDiversity}=0$

After controlling for all other predictors, there is no relationship between ethnic diversity and whether a census tract is a food desert (LA1and10, LILA1and10, or LAVehicle20).

H_0:$\beta_{RacialComposition}=0$

After controlling for all other predictors, there is no relationship between racial composition and whether a census tract is a food desert (LA1and10, LILA1and10, or LAVehicle20).

H_0:$\beta_{Urban}=0$

After controlling for all other predictors, there is no relationship between whether a census tract is urban or rural and whether it is a food desert (LA1and10, LILA1and10, or LAVehicle20).

b) Model Outputs:

DEPENDENT VARIABLE	SOCIOECONOMIC NEED	TRACT SIZE	ETHNIC DIVERSITY	RACIAL COMPOSITION	URBAN
LA1AND10	-0.232 ***	0.360 ***	-0.359 ***	-0.097 ***	0.071 **
LILA1AND10	0.531 ***	-0.006	-0.195 ***	-0.015	0.907 ***
LAVEHICLE20	0.438 ***	0.015	-0.339 ***	0.125 ***	0.328 ***

Significance indicators:

- *** p < 0.001
- ** p < 0.01
- * p < 0.05

c) Interpretations

Model 1: LA1and10

- All coefficients are statistically significant at the 0.05 level. This means we can reject the null hypotheses that there is no relationship between socioeconomic need, tract size**,** ethnic diversity, racial composition, and urban status, and the dependent variable LA1and10. We therefore accept the alternative hypotheses, indicating that each of these predictors is significantly associated with the likelihood that a tract is classified as a food desert under the LA1and10 definition.
- Socioeconomic need has a coefficient of -0.232. This means that as socioeconomic need increases, the likelihood of being a food desert under the LA1and10 definition decreases.
- Tract size has a coefficient of 0.360. This means that as tract size increases, so does the likelihood of being a food desert under the LA1and10 definition.
- Ethnic diversity has a coefficient of -0.359. This means that as Hispanic and multi-racial ethnic diversity increases, the likelihood of being a food desert under the LA1and10 definition decreases.
- Racial composition has a coefficient of -0.097. This means that as the proportion of Black residents in a census tract compared to White residents increases, the likelihood of being a food desert under the LA1and10 definition decreases.
- Urban status has a coefficient of 0.071. This means that an urban census tract is more likely to be classified as a food desert under the LA1and10 definition than a rural census tract. Specifically, the odds of an urban census tract being classified as a food desert are 7.3% higher than a rural tract.

Model 2: LILA1and10

- Socioeconomic need, ethnic diversity, and urban status have coefficients that are statistically significant at the 0.05 level. This means we can reject the null hypotheses that there is no relationship between socioeconomic need, ethnic diversity, and urban status and the dependent variable LILA1and10. We therefore accept the alternative hypothesis for these predictors, indicating that socioeconomic need, ethnic diversity, and urban status are significantly associated with the likelihood that a is classified as a food desert under LILA1and10.
- Tract size and racial composition have coefficients that are NOT significant at the 0.05 level. This means that we cannot reject the null hypotheses and conclude that there is no

evidence of a relationship between tract size and racial composition, and the dependent variable LILA1and10.

- Socioeconomic need has a coefficient of 0.531. This means that as socioeconomic need increases, so does the likelihood of being a food desert under the LILA1and10 definition.
- Ethnic diversity has a coefficient of -0.195. This means that as Hispanic and multi-racial ethnic diversity increases, the likelihood of being a food desert under the LILA1and10 definition decreases.
- Urban status has a coefficient of 0.907. This means that an urban census tract is more likely to be classified as a food desert under the LILA1and10 definition than a rural census tract. Specifically, the odds of an urban census tract being classified as a food desert are approximately 148% higher than a rural tract.

Model 3: LAVehicle20

- Socioeconomic need, ethnic diversity, racial composition, and urban status have coefficients that are statistically significant at the 0.05 level. This means we can reject the null hypotheses that there is no relationship between socioeconomic need, ethnic diversity, racial composition, and urban status and the dependent variable LAVehicle20 We therefore accept the alternative hypothesis for these predictors, indicating that socioeconomic need, ethnic diversity, racial composition, and urban status are significantly associated with the likelihood that a is classified as a food desert under LAVehicle20.
- Tract size has a coefficient that is NOT significant at the 0.05 level. This means that we cannot reject the null hypothesis and conclude that there is no evidence of a relationship between tract size and the dependent variable LAVehicle20.
- Socioeconomic need has a coefficient of 0.438. This means that as socioeconomic need increases, so does the likelihood of being a food desert under the LAVehicle20 definition.
- Ethnic diversity has a coefficient of -0.339. This means that as Hispanic and multi-racial ethnic diversity increases, the likelihood of being a food desert under the LAVehicle20 definition decreases.
- Racial composition has a coefficient of 0.125. This means that as the proportion of Black residents in a census tract compared to White residents increases, the likelihood of being a food desert under the LAVehicle20 definition increases as well.
- Urban status has a coefficient of 0.328. This means that an urban census tract is more likely to be classified as a food desert under the LAVehicle20 definition than a rural census tract. Specifically, the odds of an urban census tract being classified as a food desert are approximately 39% higher than a rural tract.

Each of these models have low predictive performance with pseudo-R-squared values of 0.047, 0.056, and 0.050 respectively and poor accuracy metrics. This is not unexpected as the purpose of these models is explanation rather than prediction and the data is observational. However, these poor performance metrics suggest that there are additional unmeasured variables, such as geography or public policy, that may influence a tract's likelihood to be classified as a food desert under any of these definitions.

i) Compare and contrast the results of the models. State your conclusions. What are the potential ethical considerations for choosing a food desert indicator for analyses or reports?

Analysis of the three logistic regression models shows that there are some similarities but also many differences in the types of areas that qualify as food deserts under these differently defined food desert indicators.

The most consistent relationship found across the three definitions between whether a census tract is urban or rural and how likely it is to be a food desert. Within all 3 definitions, it is significantly positive, meaning that urban tracts are more likely to be food deserts. This aligns with common understanding of food deserts. Notably, the increased odds for LA1and10 is small at 7% whereas it is larger in LAVehicle20 at 39% and very large in LILIA1and10. Even when the relationship is in the same direction for each definition of food desert, the magnitude varied drastically.

Socioeconomic need is significant in all 3 models, indicating that is has a relationship with all 3 indicators. However, the relationship is strongly negative for LA1and10 while it is strongly positive for LILA1and10 and LAVehicle20. LA1and10, which is purely distance based, is less likely to classify tracts with high poverty and reliance on public assistance as food deserts. This is counterintuitive as the phrase food desert is typically meant to encompass places where people cannot access fresh produce, and economic well-being can facilitate access. LILA1and10 and LAVehicle20 show the expected relationship of higher socioeconomic need corresponding with a greater likelihood of being a food desert. This is because these definitions account for other elements that impact access such as income and vehicle access.

Tract size only has a significant relationship with LA1and10, which is more likely to classify more highly populated tracts as food deserts. Tract size does not have a significant relationship with LILA1and10 and LAVehicle20 indicating that these definitions may do a better job at normalizing for population.

Ethnic diversity had a negative relationship with food desert classification for all 3 definitions. This indicates that Hispanic and multiracial communities are less likely to be food deserts. This could possibly be due to the geographic locations that Hispanic communities tend to concentrate.

Racial composition tells a different story for each of the three definitions of food deserts. For LA1and10, it is negative and significant, indicating that census tracts with a higher proportion of White residents compared to Black residents are more likely to be food deserts. This is opposite for LAVehicle20, where it is significant and positive indicating that census tracts with a higher proportion of Black residents compared to White residents are more likely to be food deserts. For LILA1and10, racial composition is not significant, indicating there is no relationship. This is a very interesting demonstration of how the varying definitions are not interchangeable but can tell different stories.

These findings suggest that future analyses could benefit from including interaction terms**,** particularly between the EFA-derived factors and urban status, to explore how these relationships differ in urban versus rural settings.

The results of this analysis show that the different methods of defining a food desert are capturing different types of areas. Beyond the structural makeup, they also categorize different proportions of the country as food deserts. LA1and10 is the loosest definition,

categorizing about 38% of census tracts as food deserts. LAVehicle20 has a tighter definition, only classifying 21% of tracts as food deserts. LILA1and10 is the most restrictive, classifying on 13% of tracts as food deserts.

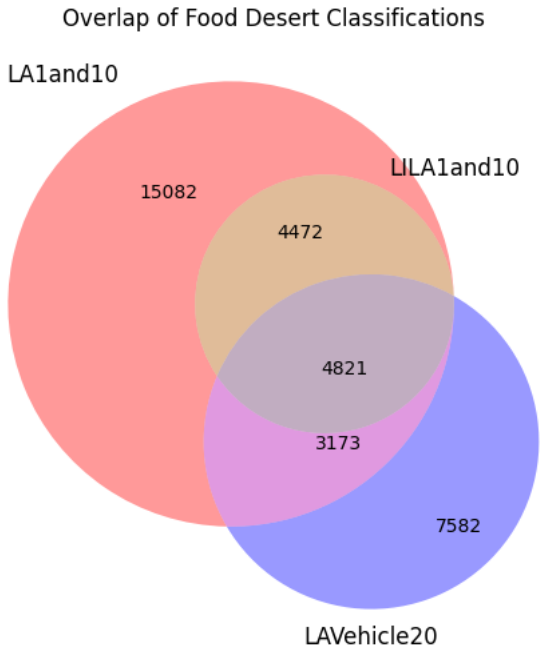

Chart III. Venn Diagram of Food Desert Indicators

There are important ethical implications of choosing a food desert definition to consider as highlighted in this analysis. Many studies use one of these definitions to represent food deserts. This analysis shows that food deserts are not a one definition fits all topic. Analyses can vary greatly based on which definition is chosen. The research team needs to be very intentional in the definition that they choose so that it best fits the assumptions of a food desert for their research questions. Transparency is also very important. Research about food deserts should explicitly call out how they are defining a food desert and how it may differ from definitions used in other research. When appropriate, researchers could also use multiple food desert indicators for their analyses. However, they must be careful not to inject bias by presenting only the definitions that show the analysis results they are hoping for.

Finally, researchers should consider how their chosen definition may over or underrepresent certain communities and clearly communicate these limitations. Food desert research influences policy and funding decisions. The definition of food desert chosen can thus have real-world impacts.

References:

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Access Research Atlas, https://www.ers.usda.gov/data-products/food-access-research-atlas/

Food Empowerment Project. (n.d.). *Food deserts*. Food Is Power. https://foodispower.org/access-health/food-deserts/