# Visualizing the Related Factors to Food Deserts in Urban Areas

Name: Jiatong Li, Peijia Sun, Mingzheng Wu
Github Username: lijiatong21, misaki775, Jjasperwu

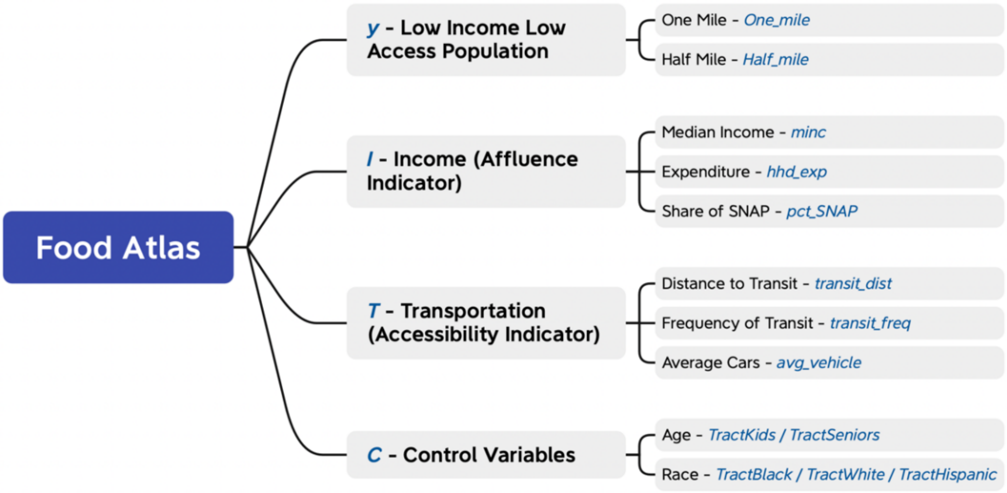
## Introduction & Research Question

Food deserts are urban areas where supermarkets are far beyond reach which plague the residents by limited access to affordable and nutritious food. These areas often emerge due to poverty and inadequate transportation. While being not synonymous to food insecurity, residing in food deserts can have detrimental health effects and reflect broader social inequalities linked to economic status. We pose the following research question: **Is the correlation between income level, transportation, and food desert well established?**

## Approach

Based on our research questions, we came up with this formula:

$$\boldsymbol{y = \beta_1 I_i + \beta_2 T_i + \gamma_1 C_i + \varepsilon}$$


Food Atlas
y - Low Income Low Access Population
One Mile - One_mile
Half Mile - Half_mile
I - Income (Affluence Indicator)
Median Income - minc
Expenditure - hhd_exp
Share of SNAP - pct_SNAP
T - Transportation (Accessibility Indicator)
Distance to Transit - transit_dist
Frequency of Transit - transit_freq
Average Cars - avg_vehicle
C - Control Variables
Age - TractKids / TractSeniors
Race - TractBlack / TractWhite / TractHispanic

We select four major cities in US: New York City, Chicago, Los Angeles, and Houston, which are the representatives of East Coast, Midwest, West Coast and South.

The definition of food desert contains information about the number and share of population in each census tract that live beyond one mile of a supermarket. The atlas also offers racial and age composition of the tracts' population. We select white, black, and Hispanic as race control, kids, and seniors as age control.

Using the median income of household, which will be complemented by the average household expenditure data, we can examine the financial affluence. Another indicator for poverty will be the share of household receiving benefit from Supplemental Nutrition Assistance Program (SNAP).

In assessing transportation convenience, our focus will be on public transportation, recognizing its crucial role for non-car owners. We'll measure proximity to the nearest transit stop, using data from policy maps, and combine this with overall transit service frequency to gauge public transit accessibility. Additionally, by examining average vehicle ownership, we can determine the community's reliance on public transit. These factors will collectively aid in identifying any transportation deficiencies within the community.

## Coding Involved

### 1. Text Processing

The initial dataset, consisting of 2020 Census Block Maps from four cities, was obtained from census.gov[1]. These maps were processed to extract the *TRACTID* for each city, essential for data integration. In the US Census Bureau's system,

[1] https://www.census.gov/geographies/reference-maps/2020/geo/2020-census-block-maps.html

a census block, the smallest geographic unit, is identified by a 15-digit GeoID. This includes a 2-digit state code, a 3-digit county code, a 6-digit tract code, and a 4-digit block code. By combining state, county, and tract codes, an 11-digit GeoID is formed for each census block in any city. We merged the data into a comprehensive dataset named '*census_no*' using the "pd.concat()" function. The *TRACTID* data was converted to integers, with zeroes appended as needed, to ensure smooth integration in the final dataset.

2. Data Wrangling

The second dataset, "Food Access Research Atlas" (referred to *main_data*), was sourced from the USDA[2] through web scraping. It provides crucial regressors and control variables, including age and race. A function named *load_4state* is used to sort variables for the four cities within this dataset.

Income variables, obtained from Policy Map as the third dataset, are processed using the same *load_4state* function. This involves sorting the cities, merging data into a long format, and selecting *TRACTID* along with its corresponding variable column. The *TRACTID* is converted into integers, with zeroes appended as needed. These variables are then merged using an inner join based on *TRACTID*, forming a dataset named *df_merge_indicator1*. Transportation variables are processed similarly, resulting in *df_merge_indicator2*. The *main_data*, *indicator_1*, *indicator_2*, and *census_no* (text processing) are combined via an inner merge to create *df_final*.

The fourth dataset, a shapefile from the Census Bureau containing geographic data for the four cities, is merged into a single long format as *df_shp*. After processing the *TRACTID* in *df_shp*, it's inner merged with *df_final* to produce the final dataset, *df_final_shp*. This dataset serves as the foundation for further plotting and OLS regression analysis.

3. Analysis

Firstly, separate *df_final* with different *y* (one_mile, half_mile), organize the columns into a sequence of *y*, *Income* variables, *Transportation* variables, and *Control* variables to create *df_ols_one* and *df_ols_half*, ensuring only rows with complete data are retained. Subsequently, perform regression analysis using *statsmodels.api* on all combinations of *aff_var* and *acc_var* for each city. This will yield a total of 72 regression results.

4. Plotting

We created six bar charts. The x-axis represents the four cities, and the y-axis represents the variables, which are divided into two groups based on *Income* and *Transportation*. This visual representation helps intuitively understand the impact of each variable on different cities, facilitating the comprehension of OLS results.

Based on the results of regressions, we only generate plots for regression combinations where the p-values of the two variables are simultaneously significant using *matplotlib.pyplot*. This process results in the creation of 8 scatter plots, each enhanced with its corresponding regression line.

5. Shiny

The interface features four tabs at the top: Maps, Variable & Model Lookup, Visualization, and Dataset. In the Food Desert Mapping section, the left side allows selection of six variables, measurement distance, and city, subsequently generating two maps. One map displays the distribution of the selected variable in the chosen city, and the other illustrates the population in food deserts within that city, offering a comparative view. The Model Lookup tab contains our formula and various variables. In the Visualization tab, users can generate maps showing the distribution of food desert populations at different scales for each city, providing an intuitive understanding. The final tab presents our ultimate data wrangling result, the *df_final* dataset.

- Weaknesses / Difficulties

1. Limited Sample Size

Although the consolidated final data frame contains over 3000 rows, only 1366 of these rows have complete data in *df_ols_hf*, and 316 rows in *df_ols_one*. This limitation in the dataset size restricts the scope of our analysis, preventing

[2] https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/

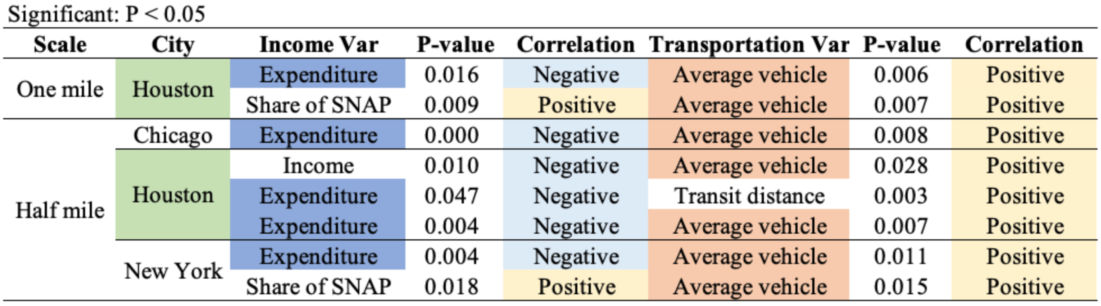
us from achieving comprehensive results.

2. Incomplete Factor Listing

Despite identifying six factors that influence the dependent variable y, there are many other potential factors that have not been exhausted. This incomplete enumeration of influences leads to a somewhat one-sided analysis. We must acknowledge the possibility of omitted variable bias, which could impact the validity of our conclusions.

- Result and Future Use

1. Result



Significant: P < 0.05

| Scale | City | Income Var | P-value | Correlation | Transportation Var | P-value | Correlation |
|---|---|---|---|---|---|---|---|
| One mile | Houston | Expenditure | 0.016 | Negative | Average vehicle | 0.006 | Positive |
| | | Share of SNAP | 0.009 | Positive | Average vehicle | 0.007 | Positive |
| Half mile | Chicago | Expenditure | 0.000 | Negative | Average vehicle | 0.008 | Positive |
| | Houston | Income | 0.010 | Negative | Average vehicle | 0.028 | Positive |
| | | Expenditure | 0.047 | Negative | Transit distance | 0.003 | Positive |
| | | Expenditure | 0.004 | Negative | Average vehicle | 0.007 | Positive |
| | New York | Expenditure | 0.004 | Negative | Average vehicle | 0.011 | Positive |
| | | Share of SNAP | 0.018 | Positive | Average vehicle | 0.015 | Positive |

1. One Mile regression:

In Houston, controlling for age and race, it is evident that lower expenditure increases the likelihood of falling into a food desert. Furthermore, there is a positive correlation between average vehicle ownership, share of SNAP benefits, and the prevalence of food deserts.

2. Half mile regression:

Controlling for race and age, it is observed that lower expenditure is associated with a higher risk of food deserts. This positive relationship between average vehicle ownership and food deserts is consistent across Chicago, Houston, and New York. Specifically in Houston, there is an inverse relationship between income and food deserts, and a direct positive correlation between distance to amenities and food deserts. In New York, there is a positive correlation between the share of SNAP benefits and the prevalence of food deserts.

3. Future Use

1. Expenditure and Food Desert correlation:

Given the observed relationship between lower expenditure and the increased prevalence of food deserts, especially in the one-mile scale analysis, further investigation into how and why expenditure impacts food desert formation would be insightful.

2. Vehicle ownership and Access to amenities

The positive correlation between average vehicle ownership and food deserts suggests that areas with lower vehicle ownership may lack adequate access to essential amenities, including grocery stores. Future research could explore transportation policies, urban planning, and the availability of public transit options to improve access in these areas.

3. Targeted geographic studies

The distinct patterns observed in Houston, Chicago, and New York indicate that food desert dynamics can vary significantly by location. Future studies could focus on these cities to understand local factors contributing to food deserts, potentially leading to tailored policy interventions.