# Citadel 2024 Summer Invitational

## *Desserts in Deserts: Quantifying the Hidden Costs of Processed Foods*

Giovanni M. D'Antonio[1], Abhay Srivastava[2], Ethan C. Tan[3], and Fucheng Warren Zhu[4]

[1]A.B. Statistics & Computer Science, S.M. Computer Science, Harvard University
[2]B.S. Economics, Statistics & Data Science, The Wharton School, Penn
[3]A.B. Computer Science & Statistics, S.M. Statistics, Harvard University
[4]A.B. Statistics, S.M. Statistics, Harvard University

August 2024

## Contents

# 1 Executive Summary

Our research examines the relationships between processed foods, food insecurity, health outcomes, socioeconomic factors, and race across the United States. Through exploratory data analysis (EDA) and iterative modeling, we uncovered significant associations between processed foods and food environments, racial demographics, geographic locations, and health indicators like obesity and diabetes rates.

EDA revealed broad trends, including a positive correlation between convenience store prevalence and obesity rates, and the significant predictive power of the Food Environment Index (FEI) for adult obesity. **Racial isolation correlated with decreased food environment quality for minority populations, but not for White Isolation.** We also noted substantial disparities in food access between urban and rural areas.

Our iterative modeling process began with partially linear models combining LASSO regression and random forests. This approach identified general correlations between food insecurity and negative health/socioeconomic outcomes at the national level. **State-level analysis showed stronger correlations in states with worse health outcomes.**

We then progressed to hierarchical linear models to capture nuanced state-level effects, providing a more detailed description of inter-state inequalities in food environments. Bayesian hierarchical models served as a robustness check and explored complex functional forms in our data.

Expanding to panel data, we examined trends over time using linear, hierarchical, and Bayesian models. This temporal analysis reinforced the importance of food-related variables in predicting health outcomes, even when controlling for socioeconomic and demographic factors. **We found that food insecurity, income, and racial factors (Hispanic vs. non-Hispanic) are significant predictors of obesity.** Bayesian models verified that obesity is heterogeneous across states and confirmed the robustness of our linear models.

**At the county level, our models revealed that food deserts are more prevalent in areas with higher concentrations of racial and ethnic minorities.** Low-income neighborhoods often lack access to fresh, nutritious foods and are saturated with fast-food outlets and convenience stores selling primarily processed foods.

Lastly, we examined the trends in obesity rates over time when controlling for relevant socioeconomic, lifestyle, and race related factors and found that there is a significant increase in obesity rates by year.

In conclusion, our research suggests that the impact of processed foods extends far beyond immediate health concerns, perpetuating a cycle of deepening inequality with long-term consequences for vulnerable populations. **The interplay between food environments, racial demographics, socioeconomic status, and geographic location paints a complex picture of food insecurity in America.** Our findings underscore the need for comprehensive, targeted interventions addressing not only processed foods but also underlying socioeconomic and racial disparities shaping food environments across the United States.

# 2 Introduction

The United States is the modern land of abundance. Yet, a crisis festers within its grimy underbelly: Food deserts.

Food deserts are regions with limited public transportation and few grocery stores, making their residents dependent on processed food as opposed to affordable fresh food. This is especially true for low-income households [1] and racial minorities such as Black, Hispanic, and Native American populations [2] [3], all of which are overrepresented in food deserts. [4]

It is clear, and perhaps uninspiring, to conclude that processed foods lead to general negative health outcomes. More crucially, the unequal distribution of populations in food deserts mean that there are also strong racial and socioeconomic differences in these outcomes [5], such as an outsized risk factor for Type-2 Diabetes among the aforementioned population groups [6]. Food deserts also perpetuate a vicious cycle, where affected (racial) populations perceive healthy food as *"white people food"*, leading to decreased demand for healthy food. [7]

That is to say, the true cost of food insecurity is not only the aggregate impact it has on the health of US population, but also the focused damage it has on particular communities.

Given that 5-10% of the country's population can be said to live in food deserts [8], we want to investigate not only the general country-wide costs of food insecurity, but also the disproportionate impacts in these areas, where processed foods are not a lifestyle choice, but a grim necessity.

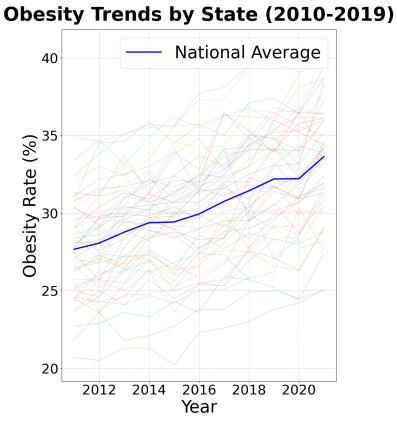



**Figure 1:** Obesity Rate trends over the years by state. Overimposition of National Average.

## 2.1 Motivation & Literature Review

Food deserts in the United States, having only been studied for two decades [1], are still a relatively young area of research. Although the causes of food deserts in the US have been well documented, ranging from the uneven expansion of large-chain supermarkets to changes in inner-city demographics [9] [10], their broader impacts are still being explored.

So far, there has been consensus in the medical community that processed foods are an major contributor to adverse downstream health problems like Diabetes and Obesity. [11] There are also a growing literature examining how vulnerable communities disproportionately suffer from the adverse consequences of processed foods. [12] [5]

We contribute to this line of questioning, hypothesizing that food deserts contribute to:

1. **Poor health**, due to the constant consumption of processed foods.
2. **Socioeconomic and racial inequality**, due to healthcare expenses and the higher acquisition cost of fresh foods in food deserts.

In this paper, we will explore whether and how food deserts and places with low food access perpetrate these outcomes, a necessary contribution to the policy conversation to promote food equity and racial equality.

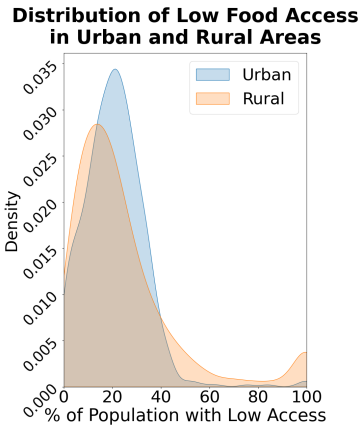



**Figure 2:** Visualizing the density of the percent of the population with low access between urban and rural counties.
Urban mean: **20.86**
Rural mean: **24.41**
Cohen's d: **-0.1937**

## 2.2 Research Questions

We aim to discuss the following questions in sequence:

1. *How are processed foods associated with public health outcomes across the United States?*

2. *To what extent do socioeconomic factors, racial demographics, and geographic disparities exacerbate the negative effects of processed food and lead to health inequities?*

3. *How can predictive modeling of health trends in vulnerable populations and geographic areas inform targeted policy interventions?*

## 2.3 Datasets & Variables of Interest

To inform our analysis, we engineered two datasets. The first dataset is a Cross-Sectional dataset from the CDC United States Diabetes Surveillance System [13]. This dataset contains data and important variables related to diabetes prevalence and associated factors for the year 2019 and just prior. The dataset includes county-level information across various counties in the United States. We then added data from the Food Environment Atlas [14]. This included the number of convenience stores, fast food restaurants, and other relevant variables. We will refer to this dataset as our **Cross-Sectional Data**.

For this dataset, we grouped our variables into:

1. **Identifiers**, such as State and County
2. **Food-Related** variables, such as the number of grocery stores or fast food restaurants per 1000 people
3. **Health-Related** variables, such as the percentage of people diagnosed with diabetes, and
4. **Socioeconomic and Race-Related** variables, such as the Isolation Index for different racial groups.

The second dataset comprises panel data from 2014 to 2022 acquired from the County Health Rankings & Roadmaps (CHR) [15]. This dataset provides county-level health data, allowing for the analysis of trends and changes across different counties. The time series nature of the data makes it suitable for examining temporal patterns and determining the impacts of health interventions over time. We collected time series data from 2014 to 2022, the years in which the Food Environment Index is available, and selected variables that were available in all 9 years. We then added control for the county population size from the US Census Bureau and USDA Economic Research Service as the population estimates available from the County Health Rankings [16] [14]. Furthermore, we evaluated the years the key variables were available, which is available in Table 10. We will refer to this as our **Panel Data**.

## 2.4 Non-Technical Summary

**Our research reveals significant disparities in health outcomes across different populations when examining factors such as race, income, and geographic location.** These findings suggest that the health problems associated with processed foods are deeply intertwined with systemic inequalities and segregation in our society.

Low-income neighborhoods often lack access to fresh, nutritious foods and are instead saturated with fast-food outlets and convenience stores selling primarily processed foods. **Our analysis confirms that "food deserts" are more prevalent in areas with higher concentrations of racial and ethnic minorities, underscoring the role of race and socioeconomic status in health disparities.**

Importantly, we found that **racial isolation correlates with decreased food environment quality for most minority populations, but not for White isolation.** This racial divide in food access contributes to the perpetuation of health inequalities.

Our research indicates that the impact of processed foods extends far beyond immediate health concerns. **The true cost of processed foods lies in the vicious cycle of deepening inequality, causing health impacts not only for the current generation of vulnerable populations but potentially for many generations to come.** This cycle is reinforced by the complex interplay between food environments, racial demographics, socioeconomic status, and geographic location.

These findings highlight the urgent need for comprehensive, targeted interventions that address not only food availability but also the underlying socioeconomic and racial disparities shaping food environments across the United States.

## 2.5 Methodology

### 2.5.1 Datasets, Cleaning, & Preprocessing

**Correlation One Data**: We mainly utilized the Correlation One Data to conduct EDA and form basic hypotheses. For our specific research questions, however, we found outside sources to be more comprehensive. *The full cleaning and acquisition process can be found in Section 6.2.5.*

**Cross-Sectional Data**: We merged cross-sectional data on 2019 Diabetes rates, 2020 Obesity rates, and 19 food-related covariates with a wide range of socio-economic, racial, geographical, and healthcare-related controls. Data sources included the CDC United States Diabetes Surveillance System, Food Environment Atlas, County Health Rankings, and the 2018 census [13] [14] [15] [16]. Variables were mostly collected from 2016 to 2020, with some being multi-year averages. To merge datasets with different column names, we used the `fuzzywuzzy` string matching library. Less than 1% of rows contained missing values, which we dropped. Finally, we normalized and centered the data. This dataset allows us to examine the interplay between food, health, and various socio-economic and racial variables. *The full cleaning and acquisition process can be found in Section 6.2.1.*

**Health Trends Panel Dataset**: We collected panel data for U.S. counties from 2014 to 2022, primarily from County Health Rankings & Roadmaps (CHR). We manually indexed variables by their collection year, using the end year for multi-year averages to avoid reverse causality bias. The `fuzzywuzzy` library was used to match variable names across years. We included only variables with less than 500 missing values out of 34,000 rows. Population data was replaced with U.S. Census Bureau figures due to defects. We corrected for a change in income inequality measurement in 2017 and fixed misnamed counties. Remaining missing values (2000 out of 30 columns and 34000 rows) were proxied with the latest available data from the same county. Alaska counties and select counties from Texas and California were dropped due to missing data. The final dataset includes 3116 out of 3143 U.S. counties, providing a temporal complement to our cross-sectional data and serving as a robustness check for our analysis. *The full cleaning and acquisition process can be found in Section 6.2.2.*

### 2.5.2 Modeling

The focus of our modeling approach is making *interpretable* models, so qualitative insights can be extracted from them, allowing us to derive *robust* conclusions. In order to do so, we employed models of increasing complexity, examining whether our results were model sensitive. We build these complexity all on a linear model backbone to maintain the interpretability of our results.

We focused on exploratory modeling where results from simpler models motivates us to employ models with more sophisticated functional forms that controlled for more variables and test whether the relationships were spurious.

We further champion an iterative cycle between modeling and data collection, where the findings of our models on a previous dataset motivates us to collect more data to validate our model's findings.

To this end, we began by fitting partially linear models to our Cross-Sectional Data, identifying relationships that we explored further using hierarchical linear models and by adding interaction terms. Finally, we examined whether the associations we found are robust over time using our Panel Data.

Beyond finding relationships between food-related variables and obesity, our models also find a significant relationship between race, the degree of racial segregation, and socio-economic status on health outcomes. Further, we find that these variables significantly interact with each other. Although beyond the scope of our present study, this motivates a more nuanced examination of the relationship between food and health outcomes within the United States, where racial and socio-economic divides exacerbate, reinforce, and are enforced by processed food.

# 3 Exploratory Data Analysis

Through our EDA, we seek to explore our hypothesis that geographical, racial, and food-related factors would have significant association with health outcomes. Both the health outcomes and the results are underpinned by the consumption of processed foods.

## 3.1 General Trends

To investigate the impact of processed foods on obesity, we analyzed the correlation between convenience stores per 1,000 people and obesity rates. We hypothesize that the correlation would be positive, as convenience stores predominantly sell processed food items. Figure 4 confirms this, showing a positive correlation between the number of convenience stores per 1,000 people and the relative obesity rate per state.

We additionally utilized simple exploratory models such as a linear regression to explore the relationship between our variables.

In order to interpret our preliminary results, we would like to first define a particular key variable of interest - Food Environment Index (FEI). The FEI generally gives an index of factors that contribute to a healthy, unprocessed food environment, on a scale from 0 (worst) to 10 (best). This measure includes access to healthy foods by considering how far the population lives from a grocery store or supermarket, or other viable locations for fresh food in their communities. The Food Environment Index is available for both of our datasets, and will be used as an indicator for Food Deserts in our interpretation of our results.

Furthermore, it also includes income variables to determine the cost barriers for acquiring nutritious food. There is research which indicates that food deserts specifically are correlated with more health detriments like obesity, diabetes, etc. as supermarkets traditionally provide healthier and less processed foods compared to smaller grocery stores and convenience stores [15].

Our results are displayed in Figure 5. FEI [17] is almost as significant as physical inactivity in predicting Adult Obesity. The value of the coefficient implies as FEI increases, the Adult Obesity decreases, and vice-versa. Furthermore, limited access to healthy foods and food insecurity also have a relatively large magnitude when predicting Adult Obesity. This analysis was conducted using our Panel Dataset.

This surprisingly significant relationship, along with the high significance of variables related to race and socio-economic status indicates and supports our hypothesis of the relationship between availability of healthy food, socio-ecoonmic and racial factors, and health.

The figure also corroborates research that indicates that food deserts specifically are correlated with more health detriments like obesity, diabetes, etc. as supermarkets traditionally provide healthier and less processed foods (Higher FEI) compared to smaller grocery stores and convenience stores (Lower FEI) [18] [15]. This lended confidence in our direction of exploration and encouraged us to build more complex models to fully capture these complex relationships.

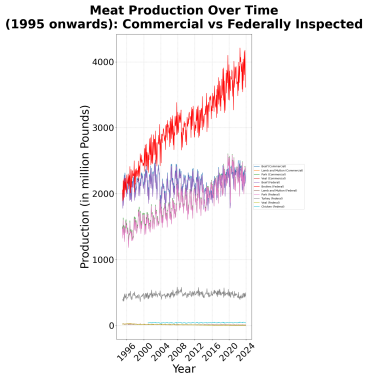



**Figure 3:** The production of meat for processed food (in red) has increased disproportionately over the last 3 decades.

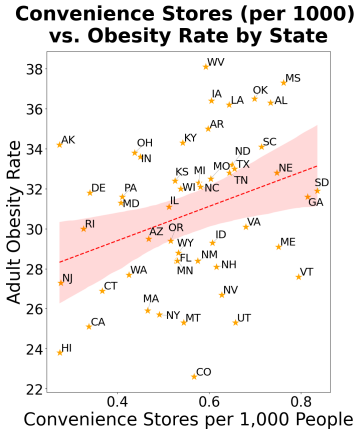



**Figure 4:** Visualizing the relationship between number of convenience stores, a popular location for processed foods, and the obesity rate by state. Source: Cross-Sectional Dataset

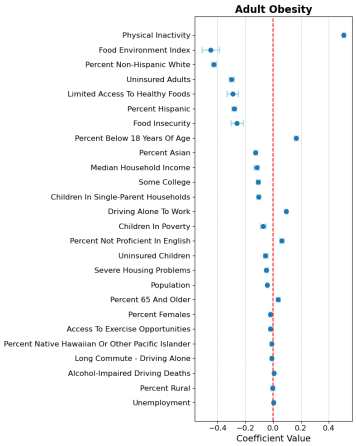



**Figure 5:** Low Food Environment Index and Access to Healthy Foods are predictive of Adult Obesity in a Linear Model, with Confidence Intervals plotted. 5-fold cross-validated $R^2$ 0.5915. All variables standardized.

## 3.2 Variable Analysis

### 3.2.1 Race-Based Factors

We put a large amount of the focus of our EDA on racial factors and how they influenced the availability of healthy food. This analysis aimed to provide information on which populations were most affected by the increase in processed food availability.

We found statistically significant differences in the mean food environment index between high and low levels of isolation for the races studied (White, Black, Hispanic). Interestingly, for all populations except White, as isolation increased, the food environment index tended to decrease. This finding provides significant insights into the racial divide and its impact on populations' access to food. We illustrate this relationship in Figure 3.2.1. The Race Isolation Index measures the extent to which minority members are exposed only to one another, and is computed as the minority-weighted average of the minority proportion in each area [16]. This analysis was conducted on our Cross-Sectional Dataset.

To further investigate the racial disparities, we conducted pairwise t-tests on each population pair (White - Black, Black - Hispanic, White - Hispanic) when highly isolated. The results, shown in Table 1, revealed that the Black-Hispanic pair was the only one that did not exhibit a significant difference.

The findings motivated us to include additional racial factors in the Panel and Cross-Sectional Data that we collected.

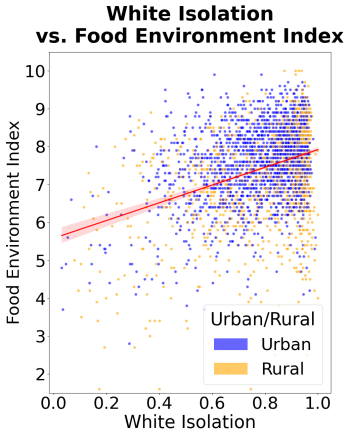
| Food Env. Index | Group 1 Mean | Group 2 Mean | p |
|---|---|---|---|
| White vs. Black | 7.51 | 6.57 | 0.000 |
| White vs. Hispanic | 7.51 | 6.77 | 0.000 |
| Black vs. Hispanic | 6.57 | 6.77 | 0.195 |

**Table 1**: Pairwise T-Tests Comparing High Isolation Areas

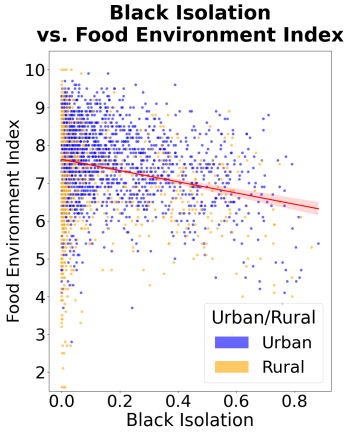
### 3.2.2 Geographic Factors

Finally, we examined the effect of geographic locations by comparing the distribution of low food access in urban and rural areas. We found a significant difference between these two populations, as illustrated in the density plot in Figure 2. While it may seem at first that urban populations have a higher percent of low food access, when taking into account the rural population's heavy tail, the mean becomes weighted towards rural, with a statistically significant difference between the two population means.

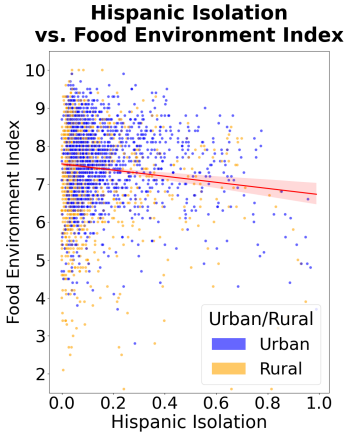
These findings highlight the complex interplay between racial, geographic, and socioeconomic factors in determining access to healthy, unprocessed food options and their potential impact on obesity rates. Furthermore, they show the detriment that processed foods have been having on specific communities. These explorations increased our confidence in exploring geographical factors as a strong variable that interacts with the relationship between processed foods and health.




**(a)** High Isolation (>0.5) Mean FEI: **7.51**
Low Isolation (≤ 0.5) Mean FEI: **6.37**




**(b)** High Isolation (>0.5) Mean FEI: **6.57**
Low Isolation (≤ 0.5) Mean FEI: **7.48**




**(c)** High Isolation (>0.5) Mean FEI: **6.77**
Low Isolation (≤ 0.5) Mean FEI: **7.45**

# 4 Modeling

As mentioned in section 2.5.2, we initially produced exploratory models that then motivated further data collection and sophisticated models. To this end, we used a 4-step approach:

1. First, we fitted a partially linear model to the Cross-Sectional Data, capturing interpretable linear relationships and existent non-linear relationships. [19] We did this on the country and state-level, and also fit a basic hierarchical linear model to state-level data where each state's parameters are drawn from a country-level distribution (thus accounting for differences between states as well as country-level effects). [20]

2. We expanded on the hierarchical structure of our dataset through Bayesian methods and using a 3-level hierarchy, where Counties exist in States, and States exist in the Country. [21] This increases the statistical power of our model as we are accounting for state and country-level confounding variables.

3. We then added interaction terms between key variables in our regression to test the robustness of the associations between food-related variables and health outcomes with a difference-in-difference style approach. [22]

4. We find that our analysis on the cross-sectional dataset gave strong indications of a relationship between food related variables and health outcomes, even when controlling for confounding variables. This motivated us to examine whether this association is *robust* over time, which led to us incorporating our 10-year Panel Dataset.

We will first describe how we dealt with variable transformation and selection, before going in detail about our approach.

### 4.0.1 Variable Transformation

We standardized all of our model response and prediction variables to have mean 0 and variance 1 to put them on the same scale. This allows for more interpretable coefficients.

When the outcomes that we are predicting are in percentages, we also attempted beta regression, a generalized linear model that predicts outcomes in $[0, 1]$. [23] However, we found that our beta regression and linear regression had similar $R^2$, with linear regression performing marginally better in most cases.

Since the ordering of variable coefficients remained nearly identical for both models, we choose to focus on linear regression because of the greater interpretability of the linear regression results.

#### 4.0.2 Evaluation Metrics

We used out-of-sample (OOS) $R^2$ and as well as MSE as ways to evaluate our models. This allows for a interpretable metric and prevents model over-fitting.

For the Bayesian models evaluated, we used the median of the Bayes $R^2$ metric which extends the traditional $R^2$ by defining $R^2$ as

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(\hat{y}) + \text{Var}(\epsilon)}$$

where:

- $\text{Var}(\hat{y})$ is the variance of the fitted values,
- $\text{Var}(\epsilon)$ is the variance of the residuals.

In a Bayesian context, this quantity is a random variable. By taking the median, we have a robust measure that has a similar interpretation to $R^2$ [24].

Because of the difficulty of evaluating Bayesian models, we use our Bayesian models mainly to explore the relationships in data with complex functional forms that other statistical methods cannot handle, but our substantive conclusions do not rely on Bayesian methodology. *Consequently, Bayesian modeling was largely utilized as a robustness check.*

#### 4.0.3 Regularization & Bayesian priors

Due to the moderate multicollinearity in our predictors as in Figure 7, in our state-level and country-level modeling for our cross-sectional analysis, we found that unregularized linear regression tends to overfit (See Table 1 Below), with LASSO regression producing better 5-fold cross-validated OOS (Out of Sample) $R^2$ values. Employing a grid search, we determined that the best regularization constant was $\alpha = 0.1$.

We employed LASSO compared to other common regularization techniques because of the ability for LASSO to set variable coefficients to 0 and has a natural interpretation of the coefficient being less than a certain parameter-specific threshold.

In the cross-sectional hierarchical modeling phase, we found that although LASSO led to less overfitting, it also led to a noticeable decrease (0.02) in 5-fold cross-validated $R^2$ with parameters determined by grid-search. Therefore, we did not employ LASSO in that context.

For our panel data, since the variables are much more carefully curated, as only important variables had wide availability on the longer time horizon, we found that the OOS $R^2$ was not noticeably different from the in sample $R^2$, and there was no need for regularization.

We chose not to use dimensionality reduction techniques like Principle Component Analysis (PCA) because the transformation in feature space that it performs decreases the interpretability of our models and the qualitative insights that we can draw from them.

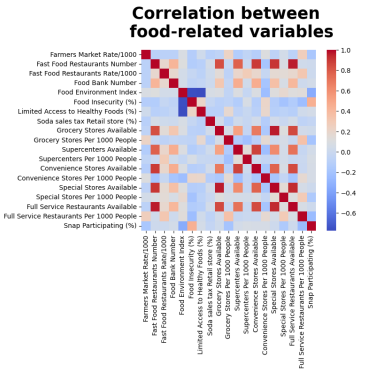



**Figure 7:** Correlation Heatmap (predictors). Note the moderate multicollinearity.

| Outcome | 5-fold Cross-Validated Mean $R^2$ | In Sample $R^2$ |
|---|---|---|
| Newly Diagnosed Diabetes Rate | 0.418591 | 0.4552656 |
| Diagnosed Diabetes (%) | 0.6275559 | 0.6521494 |
| Obesity Percent | 0.4968842 | 0.5366824 |

**Table 1:** (Unregularized) Linear Regression Model $R^2$ Values.

For the Bayesian models, we perform variable selection through using weakly regulatory Gaussian priors. Due to the ability to incorporate parameter uncertainty, over-fitting is less of a concern for Bayesian models. [20]

## 4.1 Country and State-Level Modeling

For these models, we used our food-related variables as predictors for health and socioeconomic outcomes to examine the impact of food insecurity on these areas. One model was fitted per response variable.

The model we used is partially linear:

1. We first used LASSO regression to elicit any linear relationships.
2. Then we fitted a Random Forest regression model to the residuals, identifying any nonlinear relationships.
3. We test the robustness of the relationships that we found by including relevant control variables, and still find a significant relationship between food-related predictors and response.

The partially linear regression assumes a flexible functional form as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij} + g(X_i) + \epsilon_i,$$

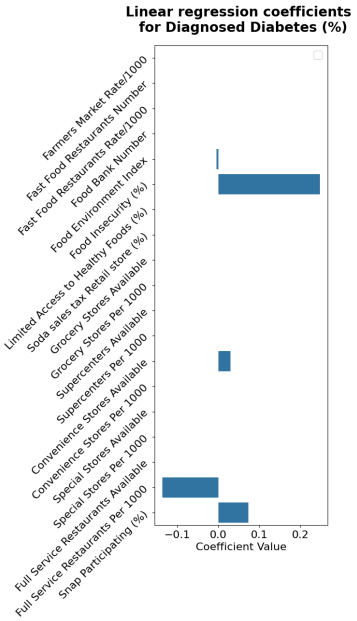
where $g(X_i)$ is a function that captures the nonlinear, nonparametric component that depends on the predictors $X_i$, allowing for a more flexible relationship between $Y_i$ and $X_i$. Here $g(X_i)$ is estimated by a random forest. The linear form allows for interpretability, whilst the random forest allows for detection of non-linear relationships.

### 4.1.1 Country-level Modeling (Completely Pooled)

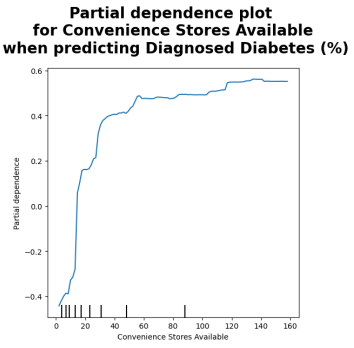
Our Cross-Sectional Dataset divides our variables by state. The most direct way to examine the general costs of food insecurity and food deserts is to disregard these divisions and treat all the datapoints as draws from the same distribution, hence 'completely pooling' them.

For each group of response variables, we see results that correlate food insecurity with negative health and socioeconomic/racial equality outcomes. We select two examples to examine here, although all the regression results can be found in our codebase.

We first examine our regression results for diabetes, a key health outcome.




**(a)** We see a positive linear relationship with food insecurity and food benefits (an indicator for low household income), and a negative one with full-service restaurants (which provide fresh food).




**(b)** We see a general negative relationship between supercenters (which provide fresh groceries) and diabetes, after an initial spike - likely because areas with zero supercenters have low to zero population.

**Figure 8:** Model results for predicting Diagnosed Diabetes (%). OOS $R^2$ (LASSO): **0.20**. OOS $R^2$ (LASSO and Random Forest): **0.50**. Note that the nonlinear estimates significantly improved model performance.

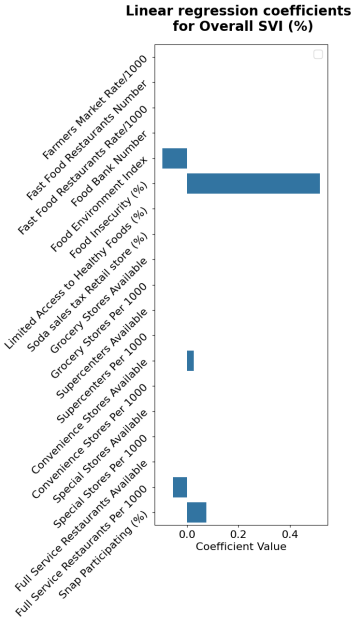
For the LASSO part of the model (Figure 8a), diabetes has a positive linear relationship with food insecurity and food benefits (indicating lower income), and a negative relationship with full-service restaurants (sources of fresh food for higher-income households). This is as expected.

Examining the most important factor in the Random Forest with a partial dependence plot (Figure 8b) reveals that convenience stores, which supply processed food, are also positively correlated with diabetes, though the relationship appears logarithmic.

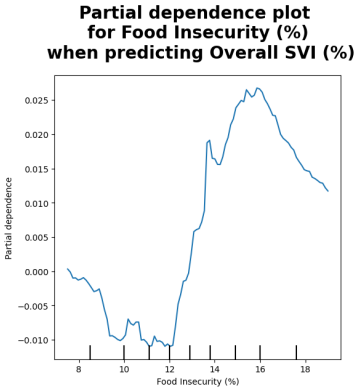
Next, we examine our results for overall Social Vulnerability Index (SVI), to elicit any general country-level relationships. The SVI refers to the index which combines several demographic and socioeconomic factors like poverty, lack of transportation, etc. that adversely impact people [13].

For the LASSO part (Figure 9a), food insecurity has a positive linear relationship with overall SVI, which is as expected.

However, for the Random Forest part (Figure 9b), we see evidence of overfitting. Food insecurity was also identified as the top factor by the Random Forest model. But when we examine the partial dependence plot, we see that the overall SVI initially decreases with food insecurity, before sharply increasing, and then decreasing again. This might suggest that different relationships between the two variables exist among different geographies or social groups, motivating a state-level analysis.




**(a)** There is a strong linear relationship between overall SVI and food insecurity.




**(b)** However, the partial dependence plot - reminiscent of Simpson's paradox - might suggest that different relationships between overall SVI and food insecurity exist among different populations.

**Figure 9:** Model results for predicting Overall SVI (%). OOS $R^2$ (LASSO): **0.52**. OOS $R^2$ (LASSO + Random Forest): **0.65**.

### 4.1.2 State-level Modeling (Unpooled)

At the state level, we noticed a trend that our models tend to have better performance when predicting health and socioeconomic outcomes for states that are worse-off in terms of those outcomes.

We formalized this intuition by scatterplotting MSE against response variable values for all states. We see, for example, in Figure 14 that when used to predict the % of people with diagnosed diabetes, food-related variables are better signals in states with higher diabetes prevalence. Similar patterns were seen for other response variables.

This implies that **lower food security is more linked to negative outcomes in states that are worse-off**, which could potentially be a manifestation of the outsized impacts of food insecurity and the proliferation of processed foods in food deserts.

However, some states have less data than others, which could also have affected model performance. Additionally, in both this and the previous section, we did not control for other non-food-related variables' effects on our target variable.

Hence, we finally move on to hierarchical modeling, which allows us to model state-level relationships without sacrificing the rest of the out-of-state data for any given state. When producing these models, we also regress on the other non-food-related variables not used as the response variable so as to control for their effects.

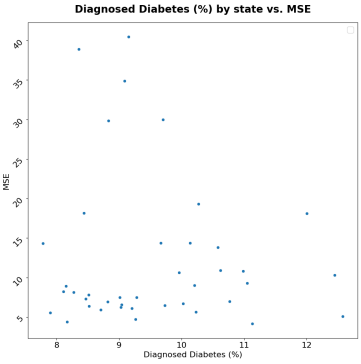



**Figure 10:** Downward trend in model MSE vs. Diagnosed Diabetes (%) by state, implying that food-related variables are better predictors of diabetes for states with higher diabetes prevalence.

## 4.2 Hierarchical modeling (Partially Pooled)

Motivated by our findings on country-wide associations and state-level differences in the previous exploratory phase, we modelled the hierarchical structure of our dataset, where counties were nested inside States inside a Country, by assuming that there are both country-wide and state-level effects by adding state-dummies into our regression. [21]

In the hierarchical modeling approach, we did not include the random forest regression on the residuals in order to improve the interpretability of our results. Further, the inclusion of dummy variables provides for our models non-linear modeling capacities, making an additional non-linear model for the residuals unessential. Indeed, we find that the hierarchical modeling approach significantly increases the $R^2$ of our models compared to the linear step on our partially linear model.

Compared to the state-level analysis, this increases the statistical power of our model as we are using more observations, whilst we are accounting for state-level confounding variables unlike the complete pooling approach. We then added interaction terms between key variables in our regression to test the robustness of the associations between food-related variables and health outcomes with a difference-in-difference style approach. [22]

By including state-fixed effects, we have the following equation for our regression:

$$Y_i = \beta_0 + \sum_{j=1}^{J} \alpha_j D_{ij} + \sum_{k=1}^{K} \beta_k X_{ik} + \epsilon_i$$

where $D_{ij}$ is a dummy variable that takes the value 1 if county $i$ is in state $j$ and 0 otherwise, and $\alpha_j$ represents the fixed effect for state $j$. Here, $J$ is the total number of states minus one (to avoid perfect multicollinearity).

By including the state-level hierarchical model, we remove the differences in state-level baseline obesity rates that affect our parameter estimates. This allows us to obtain a more accurate estimate of the relationship between our covariates **X** and our response **Y**.

Indeed, we observe that OOS $R^2$ improved when running the state-level regression to 0.65 compared to the unpooled $R^2$ of 0.63. This indicates that there are state-level effects that we did not control for, supporting the validity of our hierarchical modeling approach.

| **Diagnosed Diabetes (%)** | **out-of-sample $R^2$** | **In Sample $R^2$** |
|---|---|---|
| **Linear Regression** | 0.6275559 | 0.6521494 |
| **Fixed Effect Model** | 0.6517992 | 0.6820664 |

**Table 2:** Comparison of $R^2$ values for Linear Regression and Fixed Effect Models.

One limitation of hierarchical modeling is the absence of convenient regularization techniques. We found through a grid parameter search using LASSO implemented hierarchical models that although it reduces the difference between in sample and out-of-sample $R^2$, the out-of-sample $R^2$ remains lower than simple linear regression [25].

We then employed Bayesian fixed effect modeling with weakly regulatory priors to examine whether the relationships change as a robustness check [20]. We find little difference in the fixed-effects regression and Bayesian estimates. We further found that the Bayes $R^2$ is within 0.01 range from the fixed effects regression. Due to the difference in interpretation between the 2 $R^2$ measures, and the similarity in the results of the two models, we do not determine which model coefficients are more accurate, but take their convergence as a positive indication for the robustness of our results that these 2 different modeling approaches arrive at the same relationships. Below shows statistically significant results that we found:

| Variable | Estimate | Std Error |
|---|---|---|
| Physical Inactivity Percent | 0.5189 | 0.0189 |
| Isolation Index Non Hispanic White | 0.2275 | 0.0874 |
| Children In Poverty | -0.2257 | 0.0613 |
| Poverty Percent All Ages | 0.1777 | 0.0642 |
| Single Parent Households Percent | 0.1094 | 0.0449 |
| Multi Unit Structures Percent | 0.0897 | 0.0379 |
| Snap Participating Percent | -0.0885 | 0.0403 |
| Isolation Index Non Hispanic Native | 0.0787 | 0.0218 |
| Isolation Index Non Hispanic Asian | 0.0737 | 0.0361 |
| Severe Housing Cost Burden Percent | -0.0698 | 0.0283 |
| Severe Housing Problems | 0.0603 | 0.0264 |
| Unemployment | 0.0573 | 0.0253 |
| Percent Females | 0.0515 | 0.0211 |
| Driving Alone To Work | 0.0464 | 0.0230 |
| Poverty Estimate All Ages | 0.0449 | 0.0222 |
| Food Bank Number | 0.0333 | 0.0165 |

**Table 3:** Estimates and Standard Errors for Hierarchical State-Level Fixed Effect Modeling for Predicting Percentage of County Population that is Obese. Significant race and socio-economic effects were found. Food-related policy interventions were found to be effective.

We see that segregation and socio-economic status are significant indicators of high obesity rates, with Food Banks and SNAP both having a significant contribution in reducing obesity. This confirms our hypothesis in the contribution of race and food related variables in our analysis. Policy interventions also has an effect on reducing obesity rates even when controlling for relevant confounding variables.

At last, in order to examine the interaction between selected covariates and geographical location, we include interaction terms between the selected covariates and states. It has a difference-in-difference interpretation that leads to qualitative insights about policy implications. [22]

$$Y_i = \beta_0 + \sum_{j=1}^{J} \alpha_j D_{ij} + \sum_{k=1}^{K} \beta_k X_{ik} + \sum_{j=1}^{J} \sum_{k=1}^{K} \gamma_{jk} (D_{ij} \cdot X_{ik}) + \epsilon_i$$

Because of the amount of interaction terms that is added, interaction terms are only added for Food Environment Index and race-related Isolation Indices.

The added $\gamma_{jk}$ coefficient for the interaction term of represents how the difference in obesity a difference in 1 unit of FEI will predict, given that all the other socio-economic-racial-geographic control factors remain the same, *and that the county is in the same state $j$ as another county*. Thus, it measures the *difference* in obesity outcome of a county in the same state

with another county in the same state with a *different* FEI value.

Given this interpretation, the $\gamma_{jk}$ coefficients being significantly away from 0 would indicate that there are certain policies that some states are implementing that other states are not, which helps disadvantaged communities cope with their lack of food.

We implemented this functional form by using default weakly regulatory Bayesian priors, as standard linear regression techniques does not support the large number of variables that this approach adds. We see that there are indications that $\gamma_{jk}$ is different from 0. However, we acknowledge that we have wide credible intervals (Bayesian confidence intervals) because of our small sample size. Notibly, we see that in the interaction terms between state dummies and FEI, the most significant coefficients are for Alabama and Tennessee. These were states that we know have a high concentration of food deserts, which gives us confidence in collecting a larger panel dataset that would have enough statistical power to detect these relationships.

## 4.3 Panel Data Modeling

In our panel data, we exclusively use percentage of adults that are obese as our outcome variable, as this was the most available response variable in the dataset. While we initially intended to apply time series forecasting methods such as Vector Autoregression (VAR) and deep learning models like Long Short-Term Memory and Transformers, the limited temporal resolution of our dataset made this infeasible, as observations are only on a yearly basis. We had insufficient data to train and validate a VAR model as well as deep learning models, let alone capture seasonality or other fine-grained trends.

We found that more traditional statistical methods for panel data performed better for our dataset, and our approach is similar with the approach we took for cross-sectional data, but the addition of panel data enables us to model the effects of individual year on obesity. This increases both the **robustness** of our results by controlling for confounding variables along the time axis, and also makes our results more **interpretable**, as the year effects we found has the natural interpretation of the general trend of obesity in United States counties, *after controlling for relevant poverty, socio-economic, lifestyle, and race-related factors*.

We found that our results are consistent with our findings in the cross-sectional data, and that in our panel dataset, the coefficients of food-related variables are magnified. After performing a sensitivity analysis on our cross-sectional dataset, we found that this is due to the lack of racial segregation indicator (isolation index) in our panel data. When we removed isolation index in a pooled linear regression, we find that the coefficients for FEI went from $-0.0683$ to $-0.0961$, and for Limited Access to Health Foods, from $-0.0268$ to $-0.0425$, with the same standard errors. *This confirms our hypothesis that racial segregation and food deserts are inextricably linked in their relationship to health.* Further, we found that there were concerning trends in the year effects on obesity. Although out of the scope of the present study, this supports the importance of examining policy interventions that could slow down or reverse the present trend of rising obesity rates.

#### 4.3.1 Linear & Hierarchical Models

In our panel data, we found that unlike the cross-sectional dataset, linear regression and hierarchical linear regression with state fixed effects. Further, although we have much fewer variables, our in-sample and out-of-sample $R^2$ are similar to the Cross-Sectional dataset. This gives promising indication that we have captured important variables related to the temporal dimension and have not lost major predictive variables that had limited temporal availability. However, we acknowledge that we are predicting percentage obese in the total population in the cross-sectional data, which may lead to a difference in the interpretation of $R^2$ metric. Because our panel data had a larger sample size, we were

| Model | In Sample $R^2$ | Cross-Validated $R^2$ |
|---|---|---|
| Linear Regression (LR) | 0.593 | 0.592 |
| LR + Year + State | 0.702 | 0.700 |
| LR + Y + S + FE + SI | 0.712 | 0.703 |

**Table 4:** Comparison of $R^2$ values for different models. We observe little over-fitting. Note: FE + SI = Fixed Effects and Selected Intervention Terms. We only added interaction terms between Food Environment Index and State Dummies.

able to include interaction terms between Food Environment Index and State Dummies. We did not include more interaction terms to avoid model over-fit, which has begun to happen Table 5. We found that whilst our more sophisticated modeling approaches improved upon the $R^2$ significantly, the ordering in the importance of the coefficients remained the same. Therefore we include the coefficients for the model with all the interaction terms, as it had a marginally higher cross-validated $R^2$.

We found that after controlling for lifestyle, food-related variables were the most significant, along with income and race-related variables. Strikingly, the Food Environment Index is almost as significant as Physical Inactivity in affecting obesity *after controlling for relevant confounding variables*.

**Table Name:** Coefficients and Confidence Int for Interaction Dummies

| Term | Estimate | Lower | Upper |
|---|---|---|---|
| Physical Inactivity | 0.355031 | 0.343260 | 0.366872 |
| Food Environment Index | -0.354640 | -0.446698 | -0.262787 |
| Percent Non Hispanic White | -0.285615 | -0.305277 | -0.266046 |
| Food Insecurity | -0.242270 | -0.304249 | -0.180235 |
| Limited Access To Healthy Foods | -0.238277 | -0.296331 | -0.180324 |
| Median Household Income | -0.197294 | -0.240305 | -0.154664 |
| Percent Hispanic | -0.187654 | -0.208942 | -0.166404 |

**Table 5:** The most important variables in predicting percentage of adults obese in US counties. It is striking that the Food Environment Index is as significant as Physical Inactivity in predicting Obesity Rates. There is a natural drop after the these first 7 with the next most important variable having a coefficient of $-0.11$.

At last, we see that with the increased sample size in panel data, we could see significant state-level heterogeneity. 20 out of 50 of the interaction terms 'coefficient were statistically significant. We found much more significant coefficient estimates than our exploratory Bayesian hierarchical modeling on the cross-sectional dataset, where the highest absolute estimate of our parameter was 0.14 for Alabama due to the larger sample size and higher statistical power of the panel dataset.

We further see that there is correlation between state obesity rates and the Difference in Difference (DiD) coefficient estimates of our interaction terms. This implies that in states with lower baseline obesity rate, the

| Term | Estimate | Lower CI | Upper CI |
|---|---|---|---|
| Maine | -0.309426 | -0.454145 | -0.186075 |
| Florida | -0.209377 | -0.260307 | -0.159540 |
| Alabama | -0.132654 | -0.170589 | -0.094139 |
| New Jersey | -0.116294 | -0.199016 | -0.031175 |
| Maryland | -0.106998 | -0.175540 | -0.037400 |
| Illinois | 0.109832 | 0.068995 | 0.153825 |
| Oklahoma | 0.110996 | 0.064462 | 0.156677 |

**Table 6:** Table of estimates with their corresponding lower and upper confidence intervals.

difference in obesity rate between counties with higher food environment index and lower food environment index is greater. However, the correlation is weak and more data collection and modeling efforts should be done to disentangle this complex effect.

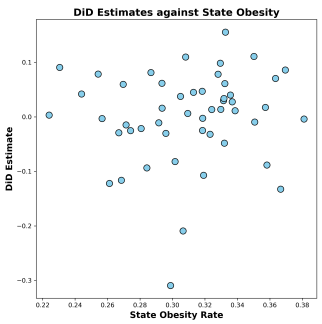



**Figure 11:** There are weak to moderate associations between the difference in difference interaction term estimate between State Indicator and Obesity with State Obesity Rates, implying that in states with lower baseline obesity rate, the difference in obesity rate between counties with higher FEI and lower FEI is greater. State-level policies should therefore be considered.

At last, we note that our modeling approach also produces qualitative insights on the *trends of obesity in the United States*. We see that after controlling for all factors, the fixed year effect on county-level obesity rates increases steadily. Further, the significance in the fixed year effects means that there are significant uncaptured effects on the *residuals* of a simple linear model without year. This increases the need for further examination of the relationship between Food Environment Index, Race, Socio-economic Status, and Obesity.

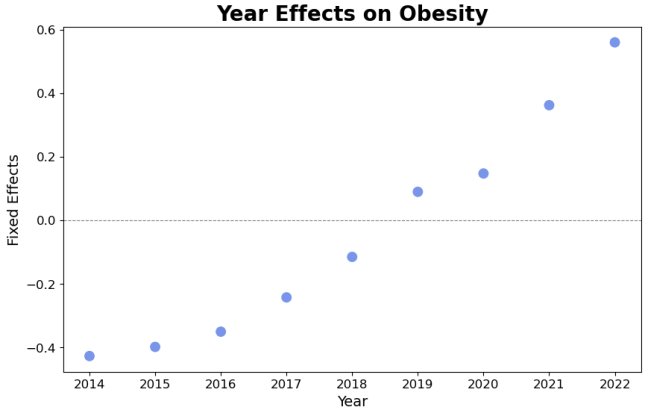



**Figure 12:** Controlling for relevant socio-economic, race, health, and lifestyle related factors, we still see a significant increase in year effect on obesity.

### 4.3.2 Bayesian Models

At last, we use Bayesian Modeling to verify the robustness of the results we got from our hierarchical linear regression. We examine the target variable in isolation by creating a hierarchical model where state-level observations are drawn from distributions parameterized by national-level hyper-parameters. We used Normal and Half-Normal priors as is standard in this case.

We see that the Monte-Carlo Markov Chains used to draw samples from our Bayesian model converge well (evidenced by the 'fuzziness' of the trace plots on the right). More importantly, consider the plot for `state_prior_means` (Left, 2nd From Bottom). This plot visualizes the distributions of the average obesity prevalence in each state, one bell curve per state. We see that different states have vastly different distributions for average obesity prevalence, verifying that this variable is indeed heterogenous across states.

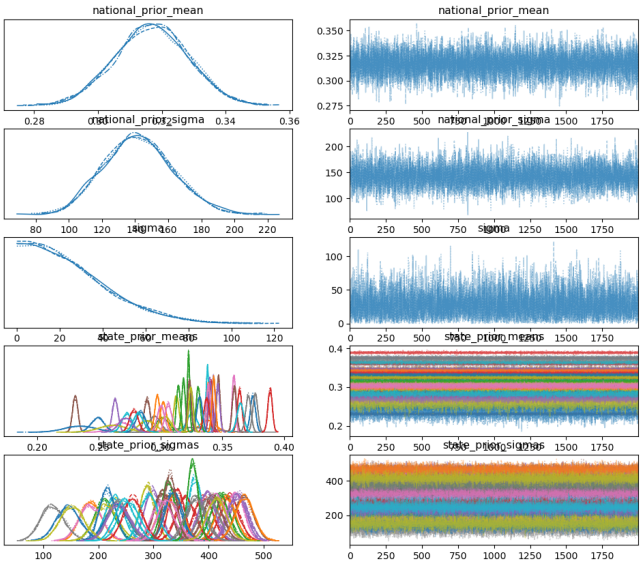



**Figure 13:** Markov Chain Monte Carlo trace plots for estimating the distributions in our hierarchical model.

Secondly, to examine predictive robustness, we also fitted a Bayesian linear regression counterpart to the models described in the previous section. It had a Bayes $R^2$ of 0.7503988, and had coefficient estimates similar to the hierarchical model, with the same ordering of the important coefficients. Due to the similarity of the results, we do not include the results table below, but note that this is another positive indication for the robustness of our findings.

## 4.4 Discussion

We find that lack of access to unprocessed, healthy foods and mitigating the effects of processed foods is a multifaceted issue that needs multiple perspectives to resolve. Our analysis of the Food Environment Index (FEI) shows that as FEI increases, there is a large increase in health detriments such as Diabetes and Obesity, with disproportionate impact on areas with low income and high degree of segregation.

One crucial policy recommendation is for the government to provide subsidies to supermarkets and super-centers to open in areas where there is largely low food access. Our findings heavily indicate that additional supermarkets are outlets for healthy, unprocessed foods like fresh fruit and vegetables, and they correlate to less diabetes. However, these policies should be on a state-government level, as our analysis revealed state-level patterns are much more discernible. In addition, there are several differences across states, as shown by Figure 14.

Expanding and enhancing the Supplemental Nutrition Assistance Program (SNAP) is another crucial strategy. Our analysis found that SNAP participation was more likely and correlated with lack of food access. This indicates that the SNAP programs are targeting the correct populations. It has been found that SNAP reduces the overall prevalence of food insecurity by as much as 30%, with even more substantial reductions in populations who are the most vulnerable [26].

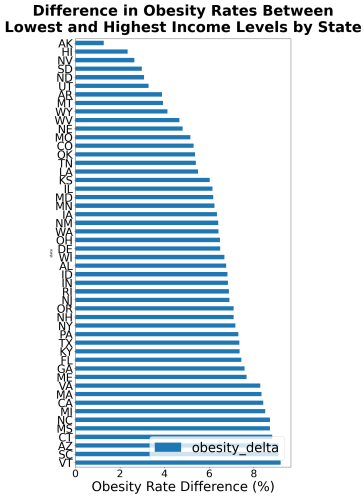



**Figure 14:** Displays the differences in Obesity Rate by State, encouraging a state-based policy.

# 5 Conclusion

Processed and unhealthy foods have been rampaging throughout America, and they have been particularly negatively affecting minority and rural communities. It is paramount to recognize the intricate connection between malnutrition, poverty, geography, and race in America. Addressing food access issues which lead to the significantly increased consumption of processed foods extends far beyond merely providing healthy food options; it requires us to rethink and stop the vicious cycle that perpetuates both poverty and poor nutrition, especially in the most vulnerable populations.

Our analysis supports that a comprehensive state-based system would be the most effective in mitigating and stopping food insecurity, as state-level impacts are much more intuitive and predictable than considering the entire United States at once. Current programs such as SNAP and WIC (Women, Infants, and Children Nutrition) are broad steps in the right direction, but they are merely the beginning of a more targeted and robust effort.

The results of our analysis compel to to confront a profound question: **Are processed foods truly the root of the issue, or are they a dark consequence of more deeper, systemic issues?** Our findings suggest that the prevalence of processed foods and the health consequences associated with them may be a symptom of entrenched food inequality. **This realization challenges us to shift our perspective from simply vilifying processed foods to addressing the fundamental societal structures that create and perpetuate food deserts, limit access to fresh produce, and trap communities in cycles of poor nutrition [27].** It forces us to consider the connection between race, class, and geography when shaping our food landscapes.

Moving forward, it is important for us to take a comprehensive approach to food justice, urban planning, and social equity to make sure that access to unprocessed, nutritious food is not just determined by one's zip code, race, or income level, but rather recognized and protected as a fundamental human right.

# 6 Appendices

## 6.1 Appendix 1: Variable descriptions for Cross-Sectional and Panel dataset

| **Food Related Variables** |
|---|
| Enrolled in Free or Reduced Lunch (%) |
| Farmers Market Rate |
| Fast Food Restaurants Rate/1000 |
| Food Bank Number |
| Food Environment Index |
| Food Insecurity (%) |
| Limited Access to Healthy Foods (%) |
| Soda sales tax Retail store (%) |
| Grocery Stores Available |
| Grocery Stores Per People |
| Supercenters Available |
| Supercenters Per People |
| Convenience Stores Available |
| Convenience Stores Per People |
| Special Stores Available |
| Special Stores Per People |
| Full Service Restaurants Available |
| Full Service Restaurants Per People |
| Snap Participating (%) |

**Table 7:** Predictors (food-related variables)

| Non-Food Related Variables |
|---|
| Primary Care Physicians Rate |
| No Health Insurance (%) |
| Isolation Index Hispanic |
| Isolation Index Non Hispanic American Indian Alaska Native |
| Isolation Index Non Hispanic Asian |
| Isolation Index Non Hispanic Black |
| Isolation Index Non Hispanic Native Hawaiian Other Pacific Islander |
| Isolation Index Non Hispanic White |
| Rent of household income Proportion |
| Severe Housing Cost Burden (%) |
| Vacant housing units Proportion |
| Access to Exercise Opportunities (%) |
| Overall SVI (%) |
| Overall Socioeconomic Status (%) |
| Below Poverty (%) |
| Unemployed (%) |
| Income Vulnerability (%) |
| No High School Diploma (%) |
| Overall Household Composition Disability (%) |
| Aged 65 or Older (%) |
| Aged 17 or Younger (%) |
| Civilian with a Disability (%) |
| Single Parent Households (%) |
| Overall Minority Status Language (%) |
| Minority (%) |
| Speaks English Less than Well (%) |
| Overall Housing Type Transportation (%) |
| Multi Unit Structures (%) |
| Mobile Homes (%) |
| Crowding (%) |
| No Vehicle (%) |
| Group Quarters (%) |
| Commute >= 60 minutes (%) |
| Public Transportation (%) |
| Urban Indicator |
| Overall Socioeconomic Status |
| Population |
| Poverty Estimate All Ages |
| Poverty (%) All Ages |
| Median Household Income |

**Table 8:** Predictors (non-food-related variables)

| Health-Related Outcomes |
|---|
| Diagnosed Diabetes (%) |
| Newly Diagnosed Diabetes Rate |
| Obesity (%) |
| Physical Inactivity (%) |

**Table 9:** Response Variables (health-related outcomes)

| Panel Data Variable (%) | Available Years | Breaks |
|---|---|---|
| Adult Obesity | 2007–2021 | None |
| Uninsured Adults | 2005–2021 | None |
| Unemployment | 2008–2022 | None |
| Children in Poverty | 2007–2022 | None |
| Some College | 2005–2022 | None |
| Children in Single-Parent Households | 2005–2022 | None |
| Diabetes Prevalence | 2008–2021 | None |
| Physical Inactivity | 2008–2021 | None |
| Median Household Income | 2008–2022 | None |
| Driving Alone to Work | 2005–2022 | None |
| % Below 18 Years of Age | 2009–2022 | None |
| % 65 and Older | 2009–2022 | None |
| % Asian | 2009–2022 | None |
| % Native Hawaiian/Other Pacific Islander | 2009–2022 | None |
| % Hispanic | 2009–2022 | None |
| % Not Proficient in English | 2009–2022 | None |
| % Females | 2009–2022 | None |
| % Rural | 2000–2010 | None* |
| Uninsured | 2005–2021 | None |
| Limited Access to Healthy Foods | 2006–2019 | None |
| Uninsured Children | 2008–2021 | None |
| % Non-Hispanic White | 2009–2022 | None |
| Food Environment Index | 2010–2021 | None |
| Access to Exercise Opportunities | 2010–2023 | None |
| Alcohol-Impaired Driving Deaths | 2008–2021 | None |
| Severe Housing Problems | 2006–2020 | None |
| Long Commute - Driving Alone | 2008–2022 | None |
| Food Insecurity | 2011–2021 | None |
| Population | 2009–2022 | None |

**Table 10:** Variable Availability Analysis

## 6.2 Appendix 2: Full Cleaning Procedure

### 6.2.1 Cross-Sectional Dataset

After formulating our hypothesis, we merged a cross-sectional data with 2019 Diabetes rates and 2020 Obesity rates and 19 food related covariates like "Food Environment Index" (accessibility of healthy foods), "Food Insecurity" (availability of food in general), and "Fast Food Restaurants (Rate/1000)". It also includes a wide range of socio-economic, racial segregation, age, geographical, educational, housing, transportation, and healthcare-related controls in order to examine the interplay between processed food, health, and socio-economic and racial covariates.

After downloading and merging every variable near the 2019 range from the CDC United States Diabetes Surveillance System, we fetched data from the Food Environment Atlas and merged variables containing information on grocery stores, super-centers, convenience stores, restaurants, and SNAP participation. We also fetched population data for each of the counties, and added in income data from the 2018 census. Furthermore, we added variables missing from this dataset which were present in our Panel Data, which were originally retrieved from County Health Rankings. We predict Obesity Outcomes from 2020 and Diabetes outcomes from 2019 as this was the latest year at which there were comprehensive data available on all food-related variables. Our variables were mostly

collected from the year 2016 to 2020, with some being multi-year averages. We could not collect covariates contemporaneous with health-related outcomes that we wish to study because many of the covariates are not collected yearly, and some had high missingness in certain years. We note that this temporal difference would bias the relationship of our variables with health outcomes towards 0, and have attempted to find as many contemporaneous variables as possible. In order to merge these datasets, which often have different column names due to being compiled from different sources, we utilized a string matching library, `fuzzywuzzy`, and created a final fully merged data-frame. We found that there were less than 1% of rows with NaNs in our dataset, so we dropped the missing rows.

#### 6.2.2 Panel Dataset

In order to examine examine the relationship between processed food and health across time, we collected an panel data of each county in the United States from 2014 to 2022. It contains less controls than the cross-sectional data as there are limited availability of many of these controls over the longer time horizon, but complements the cross-sectional data with an added temporal component. The use of both datasets in tandem act as a robustness check on our final results.

The Panel Dataset was largely acquired through the County Health Rankings & Roadmaps (CHR) yearly data [15]. In order to acquire all variables of interest, each year's documentation from 2014 to 2022 was searched. Since many of the variables used in CHR were from previous years due to limited availability, we manually indexed each variable from each year's data by the the year it was collected and sorted this information into a dictionary to map each variable to their corresponding year. If the variable is a multi-year average, we chose the end year for the variable. This process ensures that we have the most accurate collection of the health related variables of each year, and choosing the upper bound ensures that the relationships we find between our covariates and outcomes would not be biased by reverse causality in time. At last, we used `fuzzywuzzy` to match dictionary names throughout different years and combine them to create a final dataframe.

We then examined the variable availability and determined that more than half of the variables were not present before 2014, and therefore curated our Panel Dataset from 2014 to 2022. We included only variables with less than 500 missing values in the whole time-span out of $34,000$ rows in order to not introduce substantial bias into our analysis. Afterwards, we performed outlier checks through histogram plots and found that the population counts were defective. Therefore we replaced the County Health Rankings population data with population data from the US Census Bureau. We also found that the way income inequality was measured changed on 2017, and differences by a factor of 100, and corrected for this change. During the quality assessment, we further found that many counties were misnamed (with "County" represented as "Count" or "Co"), and replaced them with their correct names. After these changes, we were left with 2000 missing values in total in a dataset of 30 columns and 34000 rows. Due to the sparsity of missing values, we proxy the missing values with the latest available values from the same county in a previous year. After this, we still observe that a significant number

of county-level measurements are missing from Alaska in all years, and dropped them from our analysis. There were also select counties from Texas and California that were dropped. We keep in mind the potential for this to introduce bias. However, we believe that since we only excluded 27 counties from our analysis, In total, our final dataset contain 3116 counties out of the 3143 counties in the United States.

#### 6.2.3 County Health Rankings & Roadmaps Analytic Data, 2012-2022

The variables were renamed so they are consistent with each other across the years. "Binge Drinking" was changed to "Excessive Drinking". Some variables had punctuation and other special symbols which was not consistent with their naming throughout the years, this was also renamed through a string match. Furthermore, as previously mentioned, the string library `fuzzywuzzy` was used to match up columns. This was then sanity checked, as we had the library match columns and then printed asking us whether these two columns were actually the same. Depending on the answer, we renamed the column or we added it as a unique column to our list of columns. We then merged everything by year and column names and ended with one large dataset which contained variables that spanned a range of 2012 to 2022. We then further compared some of the columns to the actual year dataset, in order to check whether the data was correct. Finally, we sanity checked the data itself by plotting and describing various distributions of the variables, ie. making sure that population was not over a certain value for the counties.

We fetched out data from County Health Rankings (CHR) for the CHR CSV Analytic Data for each year [15]. Then, for each year, we used the CHR Analytic Data Documentation to map the variables to the years for which they were collected, as many variables from different previous years were used in a year's CHR Analytic data due to limited availability.

#### 6.2.4 County Health Rankings

* % Rural uses 2010 data for all years after 2010.

**Notes:**

- "Year" is not included as it's the dataset year, not a variable.
- "State" and "Name" are likely identifier columns, not time-series data.
- Some variables have changed names or been slightly modified over the years, but are grouped together (e.g., "Uninsured" and "Uninsured Adults").
- All variables show continuous availability without breaks within their respective ranges.

#### 6.2.5 Correlation 1 Datasets

**File 1: all_commodities.csv**

1. Read the CSV file and display basic information.
2. Check for missing values, duplicate rows, and display summary statistics.

3. Convert "Date-Time" to datetime and sort data by "Date-Time". Change format to yyyy-mm.
4. Drop the "Unit" column and rows with NaN values.
5. Remove the row with commodity "Corn".
6. Rename "Value" column to "Value (in Cents per Pound)" and convert to float.
7. Remove leading/trailing whitespace from string columns.
8. Save the cleaned data to a new CSV file.

**File 2: all_stock_and_etfs.csv**

1. Read the CSV file and display basic information.
2. Check for missing values, duplicate rows, and display summary statistics.
3. Check for any negative values in numeric columns.
4. Convert "Date-Time" to datetime and sort data by "Date-Time" and "Ticker_-Symbol".
5. Remove extreme outliers from numeric columns using the IQR method.
6. Ensure "Volume" is non-negative.
7. Save the cleaned data to a new CSV file.

**File 3: Meat_Stats_Cold_Storage.csv**

1. Drop rows with NaN values.
2. Convert "Date" column to datetime format and drop "Year" and "Month" columns.
3. Rename "Animal" column to "Object" and update "Other Chicken" value.
4. Drop the "Unit" column and rename "Weight" column to "Weight (in mln Pounds)".
5. Sort data by "Date" and "Object".
6. Save the cleaned data to a new CSV file.

**File 4: Meat_Stats_Meat_Production.csv**

1. Drop rows with NaN values and update "Other Chicken" value.
2. Convert "Date" column to datetime format and drop "Unit" column.
3. Rename "Production" column to "Production (in mln Pounds)".
4. Map "Commercial or Federally Inspected" column to numeric values.
5. Adjust "Production (in mln Pounds)" column to float.
6. Drop "Year" and "Month" columns.
7. Sort data by "Date" and "Animal".
8. Save the cleaned data to a new CSV file.

**File 5: Meat_Stats_Slaughter_Counts.csv**

1. Drop rows with NaN values and update "Other chickens" and "Beef Cows" values.
2. Convert "Date" column to datetime format and drop "Unit" column.
3. Rename "Count" column to "Count (in 1k Heads)".
4. Map "Commercial_Or_Federally_Inspected" column to numeric values.
5. Adjust "Count (in 1k Heads)" column to float.
6. Drop "Year" and "Month" columns.
7. Sort data by "Date" and "Animal".
8. Save the cleaned data to a new CSV file.

**File 6: CPI Percent Changes.csv**

1. Read the CPI data and melt the dataframe to convert years to a single column.
2. Convert "Year" to datetime and pivot the dataframe to get items as columns and years as rows.
3. Sort by "Year" and convert "Year" back to string format.

4. Save the transformed data to a new CSV file.

**File 7: PPI Percent Changes.csv**

1. Read the PPI data and melt the dataframe to convert years to a single column.
2. Convert "Year" to datetime and pivot the dataframe to get items as columns and years as rows.
3. Sort by "Year" and convert "Year" back to string format.
4. Save the transformed data to a new CSV file.

**File 8: statecpi_beta.csv**

1. Convert "year" to datetime and create a "date" column combining year and quarter.
2. Sort by "state" and "date".
3. Calculate year-over-year percent changes for "pi", "pi_nt", and "pi_t".
4. Drop rows with NaN percent changes and reset index.
5. Map state names to their initials and select only the necessary columns.
6. Convert "year" to string and save the transformed data to a new CSV file.

# Bibliography


[1] Roni A Neff et al. 'Food systems and public health disparities'. In: *Journal of hunger & environmental nutrition* 4.3-4 (2009), pp. 282–314 (cited on page 3).

[2] Valarie Blue Bird Jernigan et al. 'Food Insecurity among American Indians and Alaska Natives: A National Profile using the Current Population Survey–Food Security Supplement'. In: *Journal of Hunger Environmental Nutrition* 12.1 (2017), pp. 1–10. DOI: 10.1080/19320248.2016.1227750 (cited on page 3).

[3] Kristen Cooksey Stowers et al. 'Racial Differences in Perceived Food Swamp and Food Desert Exposure and Disparities in Self-Reported Dietary Habits'. In: *International Journal of Environmental Research and Public Health* 17.19 (2020), p. 7255. DOI: 10.3390/ijerph17197255 (cited on page 3).

[4] United States Department of Agriculture, Economic Research Service. *Access to Affordable and Nutritious Food: Measuring and Understanding Food Deserts and Their Consequences*. Report to Congress. United States Department of Agriculture, June 2009 (cited on page 3).

[5] Jennifer M. Poti et al. 'Highly Processed and Ready-to-Eat Packaged Food and Beverage Purchases Differ by Race/Ethnicity among US Households'. In: *The Journal of Nutrition* 146.9 (2016), pp. 1722–1730. DOI: 10.3945/jn.116.230441 (cited on page 3).

[6] American Diabetes Association. *Food Insecurity and Diabetes*. Accessed: 2024-08-01. 2024. URL: https://diabetes.org/food-nutrition/food-insecurity-diabetes (cited on page 3).

[7] Mengyao Zhang and Ghosh Debarchana. 'Spatial Supermarket Redlining and Neighborhood Vulnerability: A Case Study of Hartford, Connecticut'. In: *Transactions in GIS* 20.1 (2016), pp. 79–100. DOI: 10.1111/tgis.12142 (cited on page 3).

[8] Allison E Karpyn et al. 'The changing landscape of food deserts'. In: *UNSCN nutrition* 44 (2019), p. 46 (cited on page 3).

[9] Michele Ver Ploeg, Paula Dutko, and Vince Breneman. 'Measuring food access and food deserts for policy purposes'. In: *Applied Economic Perspectives and Policy* 37.2 (2015), pp. 205–225 (cited on page 3).

[10] Renee E. Walker, Christopher R. Keane, and Jessica G. Burke. 'Disparities and access to healthy food in the United States: A review of food deserts literature'. In: *Health Place* 16.5 (2010), pp. 876–884. DOI: https://doi.org/10.1016/j.healthplace.2010.04.013 (cited on page 3).

[11] Muneerh I Almarshad et al. 'Relationship between ultra-processed food consumption and risk of diabetes mellitus: a mini-review'. In: *Nutrients* 14.12 (2022), p. 2366 (cited on page 3).

[12] Kelly M Bower et al. 'The Intersection of Neighborhood Racial Segregation, Poverty, and Urbanicity and its Impact on Food Store Availability in the United States'. In: *Preventive Medicine* 58 (2014), pp. 33–39. DOI: 10.1016/j.ypmed.2013.10.010 (cited on page 3).

[13] Centers for Disease Control and Prevention. *U.S. Diabetes Surveillance System*. https://www.cdc.gov/diabetes/php/data-research/data-statistics/index.html. 2024 (cited on pages 4, 5, 12).

[14] Economic Research Service. *Food Environment Atlas*. https://www.ers.usda.gov/data-products/food-environment-atlas/ (cited on pages 4, 5).

[15] University of Wisconsin Population Health Institute. *County Health Rankings & Roadmaps*. Used to acquire datasets. 2024. URL: https://www.countyhealthrankings.org/health-data (cited on pages 4, 5, 7, 23, 24).

[16] U.S. Census Bureau. *Census.gov*. https://www.census.gov/en.html. Population and income data were retrieved from this source. 2024 (cited on pages 4, 5, 8).

[17] *Food Environment Index*. Accessed: 2024-08-04. 2024. URL: https://www.countyhealthrankings.org/health-data/health-factors/health-behaviors/diet-and-exercise/food-environment-index?year=2024 (cited on page 7).

[18] UTSA Today. 'UTSA researchers: Those with inadequate access to food likely to suffer from obesity'. In: *UTSA Today* (May 2024) (cited on page 7).

[19] Victor Chernozhukov et al. *Applied Causal Inference Powered by ML and AI*. Accessed: 2024-08-03. 2023 (cited on page 9).

[20] Andrew Gelman et al. *Bayesian Data Analysis*. Third Edition. Chapman and Hall/CRC, 2013 (cited on pages 9, 11, 14).

[21] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Accessed: 2024-08-03. New York, NY, USA: Cambridge University Press, 2007 (cited on pages 9, 13).

[22] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Accessed: 2024-08-03. Princeton, NJ, USA: Princeton University Press, 2009 (cited on pages 9, 13, 14).

[23] Silvia L.P. Ferrari and Francisco Cribari-Neto. 'Beta Regression for Modelling Rates and Proportions'. In: *Journal of Applied Statistics* 31.7 (2004), pp. 799–815 (cited on page 9).

[24] Andrew Gelman et al. 'R-squared for Bayesian regression models'. In: (Nov. 2018). Unpublished manuscript (cited on page 10).

[25] Juerg Schelldorfer. *Immlasso: Linear mixed-effects models with Lasso*. https://rdrr.io/cran/Immlasso/. R package version 0.1-2. 2014 (cited on page 13).

[26] Steven Carlson and Brynne Keith-Jennings. *SNAP Is Linked with Improved Nutritional Outcomes and Lower Health Care Costs*. Policy Futures. Jan. 2018. URL: https://www.cbpp.org/research/snap-is-linked-with-improved-nutritional-outcomes-and-lower-health-care-costs (cited on page 18).

[27] Faareha Siddiqui et al. 'The Intertwined Relationship Between Malnutrition and Poverty'. In: *Frontiers in Public Health* 8 (2020), p. 453. DOI: 10.3389/fpubh.2020.00453 (cited on page 19).