Extra Sides: A Health and Econometrics Analysis of American Fast Food

Emily Shen
University of Pennsylvania
shenyit@seas.upenn.edu

Hasan Hussein
Harvard University
hasanhussein@college.harvard.edu

Ikenna Ogbogu
Harvard University
iogbogu@college.harvard.edu

Oliver Lin
Yale University
oliver.lin@yale.edu

Authors listed in alphabetical order of first names.

Date: August 4, 2024

Abstract

America's obesity epidemic has reached the level of a public health crisis. When addressing the issue, common-held beliefs about the causes of obesity result in data misinterpretations and misinformation surrounding key policy decisions. To understand the true causes of rising obesity rates, we analyze the evolution of obesity in America and its geographical influences on a county-by-county level. We used a Gradient Boosted model to compare the significance of various environmental factors from the geographical data and found that school lunches were a leading determinant of diabetes-related diseases. We then looked at time-series data in the U.S. for fast food stocks, obesity rates, meat manufacturing statistics, and other variables to determine and confirm the most pertinent causal relationships, which we used as the basis for forecasting. We used the ARIMA model, VAR regressions, and the LSTM model to forecast fast food stocks, unemployment levels, and obesity rates to analyze the rate at which this crisis would increase in the future. Based on our analysis, we outline policy recommendations that aim to lower the cost of fresh produce for low-income communities to address the troubling obesity rates in the U.S.

Executive Summary

The history and legacy of the United States is intricately intertwined with that of ultra-processed foods. As a global powerhouse of processed food production and consumption, the United States leans on the Big Mac and Whopper to bridge the gap between its diverse cultures.[1] The industry for fast food is just as important, with the largest Fast Food Restaurants (FFR) bringing in hundreds of billions of dollars annually. To maintain their business model, major fast food chains invest hundreds of millions of dollars on marketing campaigns to differentiate their product from a sea of competitors [2]. From TV commercials to TikTok shorts, advertisements now target every age group and denomination, giving hungry Americans an illusion of choice with a supersized drink at every turn.

Our cultural fascination with fast food has long-lasting consequences though. Despite their flashy advertising and media presence, most fast food and processed food items are void of nutritional value. To maximize broad market appeal and keep prices low, fast food items are loaded with sugars and saturated fats that have proven adverse effects on individual health. Previous studies have shown a significant relationship between processed food consumption and health concerns like diabetes and heart disease, leading many health experts to decry fast food as one of the leading causes of America's rising obesity rates.[3]

Simultaneously, America faces a formidable problem for its hospitals and doctors: obesity. Since the start of the millennium, obesity cases in the United States have increased by more than a third, costing the American healthcare system over a hundred billion dollars annually. [4] Repeated efforts have been made to analyze and understand the source of this stark increase. From so-called ”food deserts” and ”food swamps” to the shortening of lunch hours at work, it is clear that obesity is a complex issue comprised of a multitude of factors. In this report, we explore the validity of these relationships and suggest new causal factors for the growing obesity crisis.

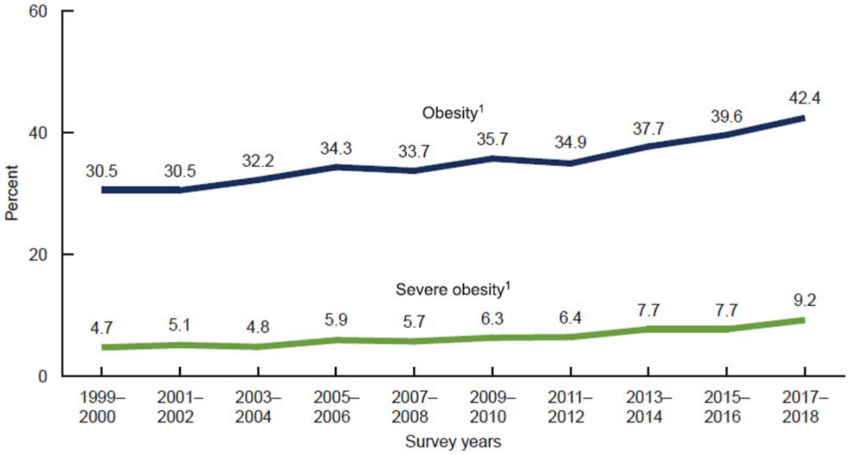


Figure. 1: Rising American Obesity Rates (NIH)

To explore the main socioeconomic factors, we collected local food access data from the Food Environment Atlas and compared it with health-related data from the Behavioral Risk Factor Surveillance System (BRFSS). Then, to confirm causal relationships for forecasting later, we analyzed the correlations between an assortment of time-series data including fast food stocks, meat production and storage, and employment data. The most salient relationships found provided the basis for our predictions, which highlight the exigency of our policy recommendation of making healthy alternatives cheaper and more accessible.

Guiding Questions

- How does geographical access to processed foods affect the prevalence of diabetes in individual neighborhoods?
- How does the economy reflect the U.S's unique trends in food consumption, and what are the associated impacts on people, economically and health-wise?

Summary of Findings

By analyzing changes in obesity rates over time, we found that fast food stocks are closely linked with American obesity, suggesting that fast food companies have managed to profit from this public health crisis. On the other hand, unemployment data suggests a causal relationship between economic growth and rising obesity, possibly as a result of rising inflation. We then traced the impact of a variety of food access indicators on diabetes rates, a significant indicator of processed food and sugar consumption. An analysis of over 3100 counties revealed that the strongest predictor of diabetes-related disease was not fast food access or lack of grocery stores, but school lunches, suggesting that processed foods in schools may be linked to childhood and adult diabetes.

By tracing the impact of a variety of food access indicators on diabetes rates, a significant indicator of processed food and sugar consumption, we found that access (distance from grocery stores) was not among the strongest predictors of diabetes-related disease. Rather socioeconomic levels, both in terms of school lunches that children in public schools consume and in welfare programs for adults, controlled people's dietary choices. While we were unable to confirm the causality of these relationships, their strong correlative effects suggest further research into the effects of means-tested welfare on participant health. Then, by analyzing obesity rates over time, we found that fast food stocks are closely tied to American obesity, suggesting that fast companies have managed to profit from this public health crisis. In our deeper correlational and causal analysis, we discovered interesting inter-causal relationships among fast food stocks, obesity rates, meat production and storage data, and unemployment rates. These relationships not only explain the past and present dynamics of the American obesity crisis but also highlight potential pathways for future mitigation efforts. Our forecasting work supports these findings. By employing VAR and LSTM models, we demonstrated that variables such as fast food stock prices, meat production, and cold storage levels are significant predictors of unemployment and obesity rates. These models performed well in capturing the trends and validating the significant predictive power of the included variables.

Overall, our findings emphasize the complex interplay between socioeconomic factors, food access, and public health. They underscore the need for targeted policy interventions that address the root causes of obesity and related diseases, rather than solely focusing on access to food.

Contents

1 Our Analytical Framework for Health

To assess the health and economic consequences of America's dominant fast-food industry, we chose to focus our analysis on obesity and type II diabetes rates. While obesity is technically defined as a body mass index (BMI) of at least 30 kg/m [6], type II diabetes is a chronic metabolic disease that occurs when the body misallocates and insufficiently produces enough insulin to offset glucose levels [7]. Although type 1 diabetes is also a pervasive disease, we analyze type II diabetes as it accounts for about 90%-95% of diabetes cases. Moreover, while type 1 diabetes is determined by genetic factors, type II diabetes largely occurs due to exceeding glucose levels, making it a preventable disease [7]. Since roughly 90% of those with type II diabetes have obesity [6], our analysis gains a specific and broad understanding of the direct health effects by looking at obesity and type II diabetes in tandem, with certain types of data for each being more suited for different analyses.

Although ultra-processed foods are bought and sold through a variety of stores and consumer markets, The health-related impact of fast food restaurants on Americans has been a prominent research area given that fast food restaurants are closely associated with preeminent causal factors of obesity and type II diabetes such as diet, calorie intake, and blood pressure [8]. The health effects associated with fast food menu items such as ultra-processed meat and sugary drinks are troubling as diabetes causes mortality and worsening states of morbidity.

After reviewing over 100,000 death certificates, the American Diabetes Association (ADA) in 2021 found that diabetes was the eighth leading cause of death in the United States [9]. Alongside the staggering health costs of this chronic health disease come its economic ramifications. Research by Emily Parker et al. (2024) shows that the medical expenses accrued by patients sum to about $19,700 annually, resulting in roughly $306.6 billion in direct medical costs on a national level in 2022 alone. Additionally, the indirect costs of higher rates of disability, presenteeism, and premature death strained the United States economy by around $95 billion in unemployment [10].

Given that worsening health reduces one's ability to be economically productive, addressing the adverse health effects of America's fast food restaurants can aid efforts in promoting economic mobility for affected communities. Our initial hypothesis treads on the common belief that fast food restaurants would best explain obesity and diabetes rates in the U.S. due to their comparatively low costs, convenience, and rampant levels of sugar and cholesterol in their food.

2 Technical Exposition

2.1 Data Collection

Our dietary choices result from a complex web of interactions between personal preferences and necessities. Environmental factors play a major role in these relations, so we incorporated geographical data into our analysis to examine disparities based on location and socioeconomic status. We collected data from the Food Environment Atlas, a county-level databank by the US Department of Agriculture that contains over 280 access-related variables, such as physical distance to grocery stores, SNAP/WIC benefits, and fast food restaurant density. To explore the health-related consequences of fast food, we also gathered local diabetes and similar chronic disease data from the Behavioral Risk Factor Surveillance System (BRFSS). Previous studies have linked excessive consumption of processed food with these kinds of chronic diseases, allowing them to serve as important indicators for the human effects of processed foods. We confirmed these findings with the provided time-series datasets on stocks, obesity, meat production, and others (including an external dataset on employment) which we analyzed to capture correlation and causality, visualize trends across time, and forecast future impacts. [15]

2.2 Wrangling and Cleaning Process

We began by extensively cleaning and remediating all of the provided datasets. For the commodities dataset, we filled in missing values in the 'Commodity' column using forward fill to maintain continuity, and correspondingly assigning appropriate units to each commodity; we then standardized the pricing of all the commodities for consistency.

Some datasets were mislabeled, so we manually fixed nonsensical labels where we found them, like Domino's being labeled a wholesale grocery company. We also standardized relevant variables throughout our entire array of datasets to allow for meaningful comparisons and interpretation.

2.2.1 Dataset Transformations

For all the 'meats_stats' datasets, which had a dating format different to all the other datasets, we simply standardized its date column one that would allow us to properly index along years and months as needed.

Within the geographical data, we combined each dataset by county and state. The dataset was adjusted to account for response bias by controlling for age and employment. Sparsely populated counties were removed from the dataset due to low response levels, and all data was normalized per capita.

2.2.2 Quality Control

Throughout the exploration process, we ensured the data remained consistently formatted. We performed checks for stationarity using the ADF (Augmented Dickey-Fuller) test on key variables like stocks and obesity rates before conducting any further time-series-based analyses, which was crucial for the validity of our tests on causality and regressions, it confirmed our intuitions for the non-stationarity and seasonality of data like stocks and obesity. We standardized all our time series data to ensure that the results would make more sense on a graph or when forecasting and chose to aggregate our stock, commodity, meat production, and employment data by month (if it wasn't already) rather than by day to save computation time and make analysis more meaningful.

After exploring the provided dataset on nutrition, diabetes, and physical activity data for students and adults ranging from 2001 to 2022 across the U.S., we sought data that illuminated a clearer picture of the disease-related impacts associated with obesity. By combining the provided dataset with the U.S. Center for Disease Control and Prevention Center's (CDC) 2023 Chronic Disease Indicators (CDI) Release, we visualized state-level differences in mortality and morbidity rates associated with diabetes.

Each of these datasets aggregates survey responses regarding various topics with each topic containing a set of relevant questions. For instance, while the first dataset contained information on nutrition, obesity, and physical activity for high school students and adults aged 18 years or older, the latter dataset contained these topics in addition to diabetes and cancer. Most importantly, CDI Release enabled our analysis to be nuanced as the diabetes-related questions provided survey responses that differentiated between mortality and morbidity. The questions "Mortality due to diabetes reported as any listed cause of death" and "Amputation of a lower extremity attributable to diabetes," for instance, demonstrate the depth to which our analysis explored the health effects of ultra-processed foods.

To prepare this dataset, we first removed all features that either only contained null values or failed to provide additional information. For instance, we unified all categorical variables regarding race, grade level, income, and gender under consistent entry values to improve our data exploration process. Since all entry values were stored under the Data_Value column, we cleaned the dataset by filtering out rows with empty values for this feature.

When performing exploratory visualizations, we ensured that the metrics of the survey responses were standardized by aggregating all entries with the same units for the data value entry. While there were many entries utilizing crude rates as their metric, we visualized entries based on age-adjusted rates whenever possible, even if it meant leveraging a smaller data subset. Since crude rates fail to account for differing propensities for diabetes-related health symptoms amongst age groups, their results can be misleading.

2.3 Exploratory Analysis

We started by confronting some lingering questions from the existing literature. The term ”food desert” originated in Scotland in the 1990s to describe an area with low access to fresh foods through grocery stores. [5] Since then, the concept of food deserts has been pushed as a leading driver of food insecurity and malnutrition. However, new waves of research have repeatedly demonstrated that food deserts are uncorrelated with fast food consumption and its effects. We started with a wide overview by state, constructing a correlation heatmap to test this hypothesis, comparing geographic data for FFRs per capita against a host of other possible factors.

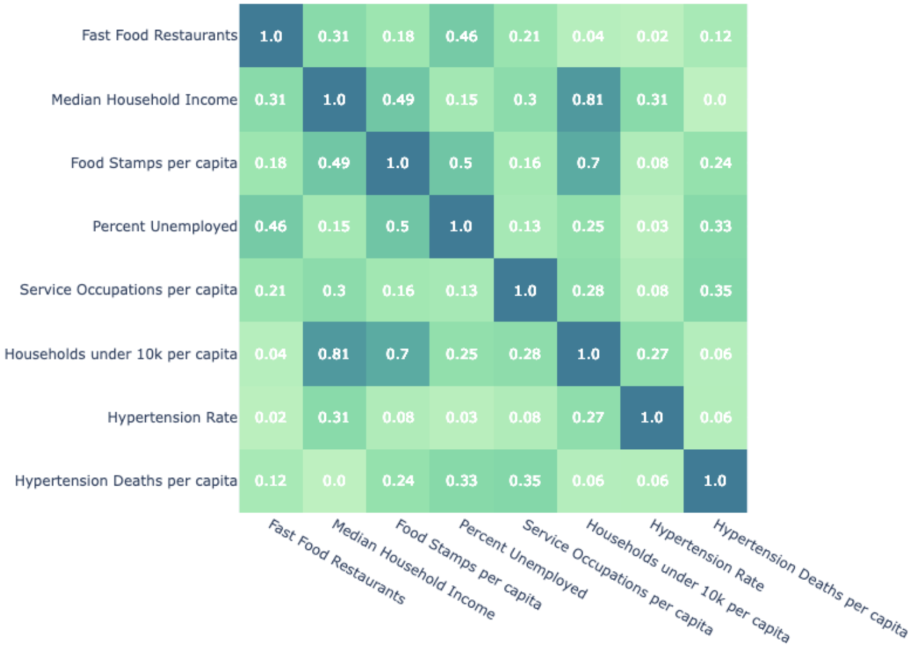


Figure. 2: Exploratory Correlation Heatmap. Each square represents the absolute value of the correlation between two variables. Fast food restaurant density (first row/column) is poorly correlated with the other values, debunking the ’food desert’ explanation for poor nutrition.

In particular, the heatmap shows a lack of correlation between FFR density and hypertension rate and mortality. As a result, we can safely reject the idea that a higher density of fast food (food swamp) is related to adverse health effects. Instead, this paper looks into overarching economic trends and finds a causal relationship between fast food stock growth and obesity rates.

2.4 Geographic Analysis

To dive into the details, we then compared individual counties. Since states are so large, there is a risk that the scale of the data will even out smaller effects. For example, the fast food industry in the Bay Area is unlikely to affect the citizens of Sacramento. Still, both of these areas would be accounted for holistically in the state of California. A county-level analysis allows for a more precise and robust understanding of local effects. Using localized data, we analyzed a variety of geographical indicators and used them to predict fast food sales and diabetes rates. We applied a gradient-boosting model to reduce the effect of outlier counties and interpret the potentially non-linear relationship of our input vector. Our results indicated that while welfare programs and general socioeconomic indicators were strong predictors of both fast food consumption and diabetes, the impact of American school food outstripped all other factors in terms of predictive power.

2.4.1 Feature Engineering

To make correlative statements, we gathered location-based data from the Food Environment Atlas (FEA). [11] This dataset included over 280 environmental factors, including low-income access, means-tested welfare programs, and grocery store prevalence. A full list of variables can be found on the USDA website.

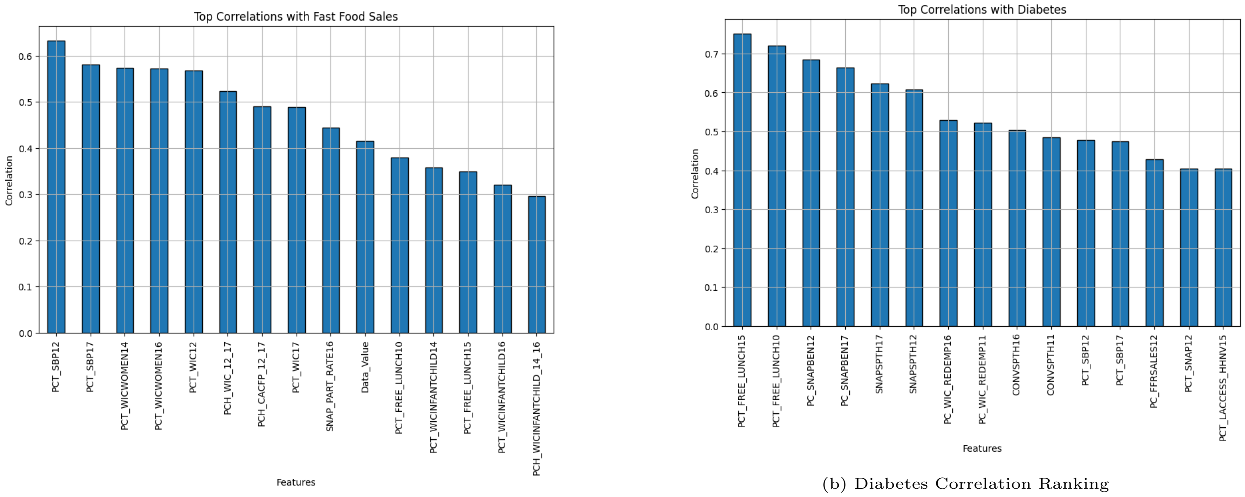
To control for the varying populations of counties, we adjusted all data to be per capita. Additionally, many of these factors are tracked periodically and may show up in the dataset for different years. Inputting multiple years of the same data would introduce endogeneity into our categorical model. To properly control the correlative power of our model, we removed duplicate factors and left only the most recent year of data.

Our health data came from the Behavioral Risk Factor Surveillance System (BFRSS), and contains crude and age-adjusted prevalences of a variety of chronic diseases within adults. [12] Since diabetes is one of the main health concerns related with obesity and is also strongly tied to both obesity and fast food consumption, we chose to examine diabetes as a leading indicator of health impacts.

2.4.2 Exploratory Data Analysis

Our goal is to differentiate the most relevant factors of fast food consumption and diabetes. To start, we want to find strong correlative relationships between our factors and predicted values, as these types of variables will likely be good predictors as well. We ran a simple correlation test between our full list of FEA factors to find closely related variables.

For example, we plotted diabetes data against free lunch, the strongest correlation score, and discovered a generally strong upwards trend that suggests a relationship between both.



(b) Diabetes Correlation Ranking

(a) FFR Sales Correlation Ranking

Figure. 3: Bar graph of different factors encoded according to USDA standards and their correlative relationship to target variables. These correlations represent direct linear relationships, so it is possible that some strongly causal variables may not show up on this graph.

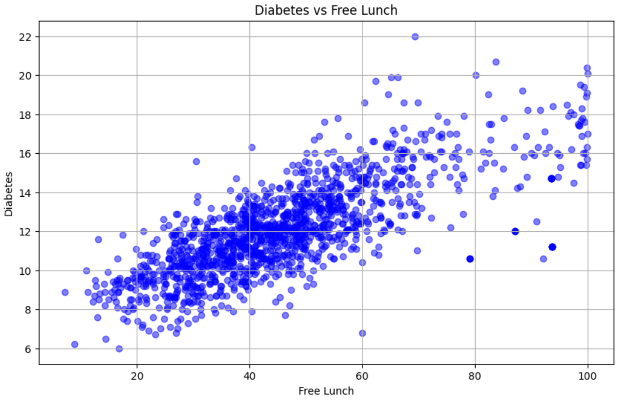


Figure. 4: Scatter plot of diabetes data. There seems to be some kind of upward relationship present between the two variables.

Most of these variables are encoded with USDA standards, which we interpreted via the FEA's associated data dictionary. The largest correlation values from this chart are likely to have some type of predictive power on their associated indicators. All units were normalized per capita or percent.

2.4.3 Assumptions and Choices

We made some simplifying assumptions about our target variables. We used fast food sales as a proxy variable for the total amount of fast food consumed. While fast food prices might differ regionally to some extent, major chains typically keep costs relatively constant to maintain a sense of familiarity. Additionally, we used diabetes as a proxy to examine the health consequences of fast food. Processed foods contain high amounts of saturated fats and sugar that have been linked to higher rates of type II diabetes. [13] Since diabetes and obesity are tightly linked, and diabetes data is both easier to access and has more robust responses, we chose to model diabetes rate per capita as our primary health indicator. [14]

We also assumed that inter-county interactions had negligible effects on in-county variables. For most day-to-day activities including grocery shopping and eating out, people typically stay near their homes and in their counties for convenience. Similarly, we discounted any sort of tourist effects, as tourism typically has a low impact on personal diets.

2.4.4 Feature Selection

While our FEA data contained over 280 variables, trying to pass all of these factors into our model would inevitably lead to significant overtraining. Instead, we chose a smaller subset of important variables to train and predict on. For our predictive models, we chose the five best-correlated factors for both indicators and then selected similar variables as auxiliary predictors. We assumed that strong linear correlations suggested predictive power, as it is extremely rare for two wholly unrelated variables to display strong correlations. A full table of the variables codes and units is provided below.

Original Name	New Name	Unit
PC_FFRSALES12	FFR Sales	$ per capita
PCT_SBP12	School Breakfast Program	Percent
PCT_WICWOMEN14	WIC Women	# per capita
SNAP_PART_RATE16	SNAP Participants	# per capita
PCT_WICINFANTCHILD14	WIC Infant/Child	Percent
FFRPTH16	FFR per capita	# per capita
PC_SNAPBEN17	SNAP Benefits	$ per capita
PCT_LACCESS_POP15	Percent Low Access to Grocery Store	Percent
Data_Value	Diabetes Rate	Percent
PCT_FREE_LUNCH15	Students on Free Lunch	Percent
PCT_SNAP17	SNAP Benefits	Percent
PC_WIC_REDEMP16	WIC Benefits	$ per capita
SNAPSPTH17	SNAP Recipients	# per capita
PCT_LACCESS_SNAP15	SNAP Recipients without Grocery Access	Percent
FFRPTH16	Fast Food Restaurants	# per capita
CONVSPTH16	Convenience Stores	# per capita
PCT_LACCESS_CHILD15	Children without Grocery Access	# per capita

Table 1: Variable Encodings and Units

2.4.5 Model Development

To predict each of these variables, we chose to implement a Gradient Boosting Machine (GBM) model. Gradient boosting is a powerful ensemble learning method that iteratively trains basic predictive models, typically decision trees, to form a stronger predictive model. Each subsequent model is built on the so-called ”pseudo-residuals” of the previous model. These pseudo-residuals represent the gradient of the loss functions of previous models, allowing each iterative model to train on previous models' weaknesses.

A GBM is defined by the sum of a series of component learners, each of which is trained on previous GBM data. The final model $F(x)$ is built as an additive model:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

where $h_m(x)$ are the base learners and γ_m are the controlled coefficients. We start by initializing a base model, F_0, using some kind of basic decision model for our dataset. Here, L represents some loss function being minimized by our basic model.

$$F_0(x) = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

We then improve the model iteratively by computing the pseudo-residuals for each observation in the training data.

$$r_{im} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right]$$

We can fit a new model $h_m(x)$ to these residuals and optimize the coefficient γ_m to minimize the loss when the new model $h_m(x)$ is added to the existing ensemble:

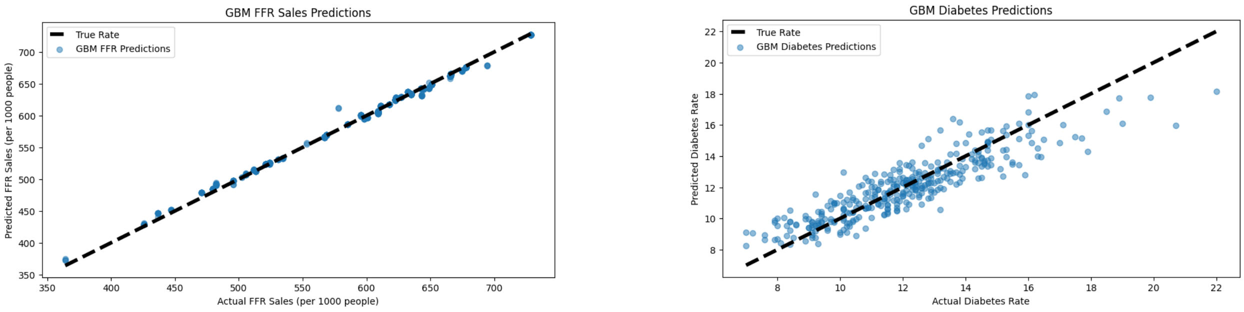
$$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

Then the model is updated according to

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

2.5 Model Performance

We trained both GBM models on over 3000 counties of available data using a randomized 4-to-1 train-test split. The below residual visualizations demonstrate the accuracy and power of our predictive models on testing data. True values are displayed along the diagonal, with variations from the diagonal signaling error. To quantify our results, we then retrieved the normalized mean square errors (MSE) of both models.



(a) FFR Sales GBM Residuals

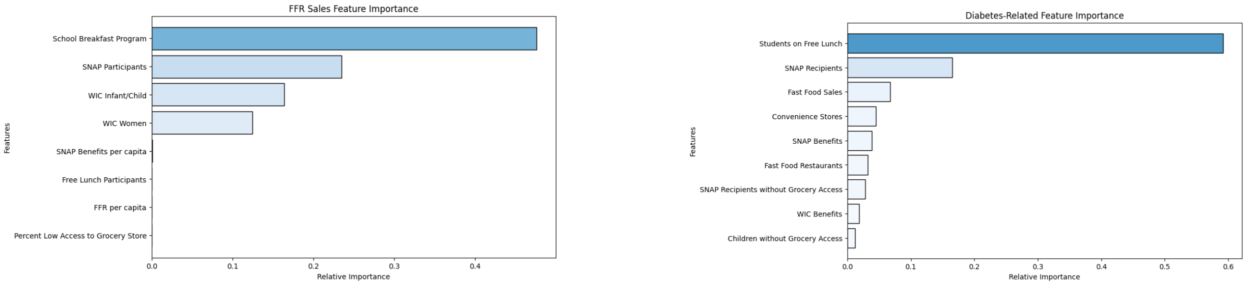
(b) Diabetes GBM Residuals

Figure. 5: Residual plots to analyze model performance. The true values lie along the main diagonal and the spread from that diagonal represents model error.

Model	Raw MSE	Normalized MSE
Fast Food Sales	26.98	0.0046
Diabetes	1.21	0.210

Table 2: Mean Squared Errors

Although the diabetes model had a significantly worse MSE, it remains within an acceptable deviation. We expected this kind of increased deviation, as there are many other potential causes of increased diabetes rates that were not captured in our model. To find the most important input components, we used the built-in `feature_importances_` attribute.



(a) FFR Sales Feature Importances

(b) Diabetes Feature Importances

Figure. 6: Feature importance plot

Feature importance is calculated by determining the total amount of Gini impurity. Gini impurity measures the homogeneity (or lack thereof) for labels within each node and is given by

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

Where p_i measures the proportion of samples belonging to the class k for a given node. We sum the feature importance across each basic decision tree to find the feature importance of the full GBM.

2.5.1 Discussion

From the feature importance plot, we can see that meals at school are surprisingly important correlative factors in both models. As our model lacks time-dependent information, it is difficult to make a causal inference on how the two variables are linked. Looking at the other high-power factors, however, we noticed that our best correlational variables are all signs of poor socioeconomic background. This suggests that while geographic access factors (i.e. food deserts) may not have a major role in determining fast food consumption and health-related consequences, limited income also limits the ability to make healthy food decisions.

The benefits of the Gradient Boosting method include the ability to determine nonlinear relationships. Training the model on pseudo-residuals allows each component decision model to make non-linear adjustments while still being easily interpretable via partial dependency plots. This allows us to minimize the mean-squared error and make more accurate comparisons.

Impurity-based feature importance can be misleading for high cardinality features, however, which restricts the number of input variables that we can include. We chose the five best linearly-correlated variables along with some hand-picked auxiliary variables in order to maximize the chance of a correlative effect, but it is also possible for non-linear causal factors to have poor correlations which might have been excluded from our model.

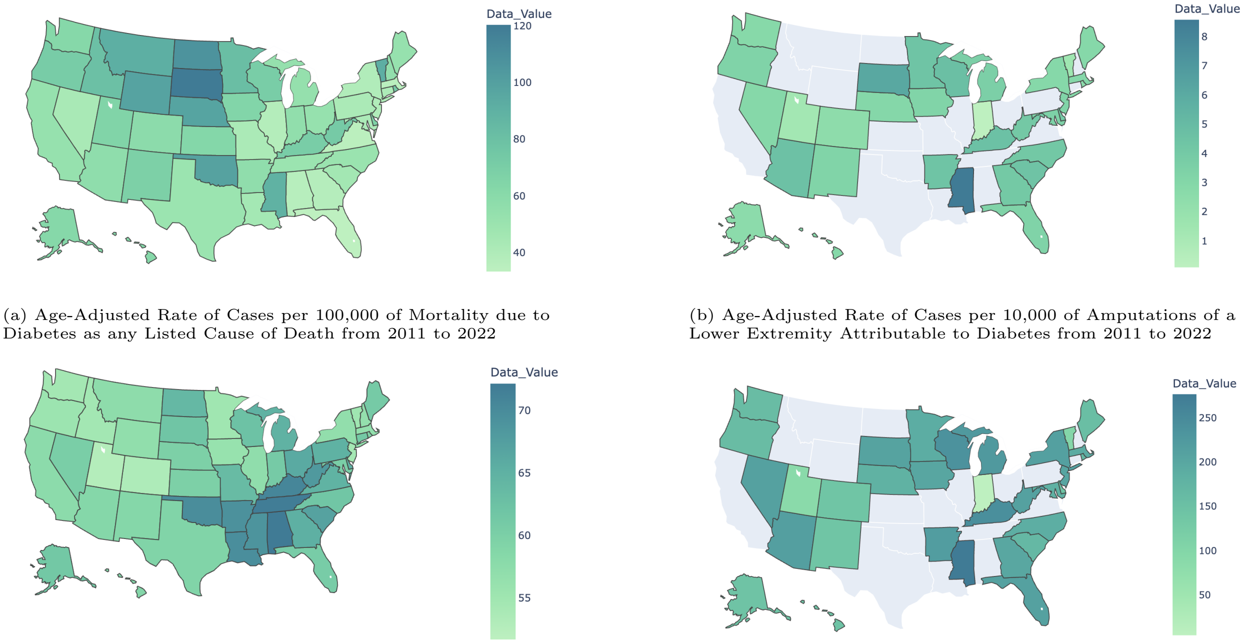
Additionally, the lack of time-series data by county prevented us from making any causal inferences on this data. In future analysis, we recommend substituting our FEA data with a proxy variable that includes time-related data in order to make causal inferences.

2.6 Case Study: Investigating Mississippi's Health-Disease Epidemic

2.6.1 Determining Mississippi

Initially, we hypothesized a strong overlap over states with high morbidity and mortality levels surrounding diabetes. Intuitively, if a region experienced excessive levels of severe medical procedures surrounding diabetes, then they would also suffer the most from the mortality effects.

However, Figure 7 highlights findings that ran contrary to our hypothesis. While the upper mid-west showed the highest mortality rates, the South had demonstrated the highest morbidity rates. Most notably, South Dakota demonstrated the highest cases per 100,000 of diabetes-related moralities while Mississippi showed the highest cases per 10,000 in diabetes-related amputation rates and percentage of the population for diabetes-related hospitalization rates.



(a) Age-Adjusted Rate of Cases per 100,000 of Mortality due to Diabetes as any Listed Cause of Death from 2011 to 2022

(b) Age-Adjusted Rate of Cases per 10,000 of Amputations of a Lower Extremity Attributable to Diabetes from 2011 to 2022

(c) Age-Adjusted Rate of High Blood Pressure Prevalence Amongst Adults 18 years or Older with Diagnosed Diabetes as a Percentage

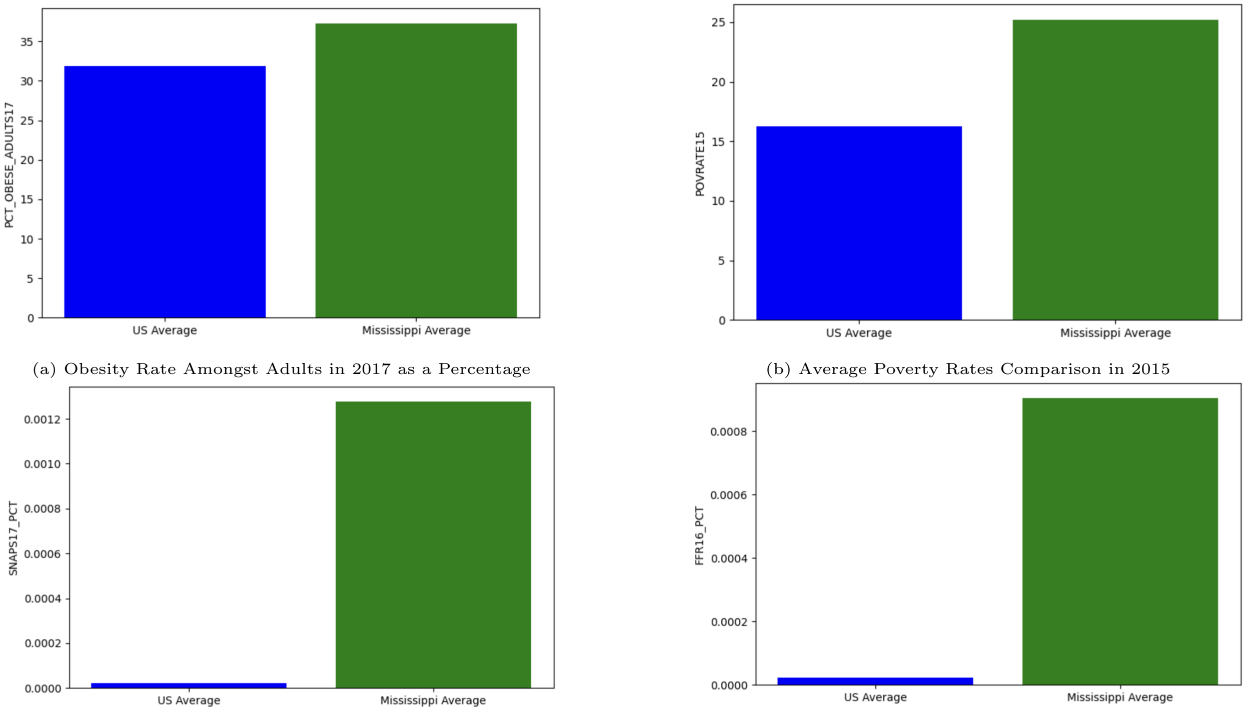
(d) Age-Adjusted Rate of Cases per 10,000 of Hospitalizations with Diabetes as a Listed Diagnosis from 2011 to 2022

Figure. 7: State-level U.S. maps outlining regional trends in diabetes-related health effects and causal factors.

2.6.2 County Level Findings

We investigated socioeconomic factors in Mississippi to better understand the demographic and associated factors with Mississippi's excessive rates of diabetes-related health effects. To do so, we grouped the county codes in Mississippi, taking the average, and then compared it to the average in the U.S. In order to understand these factors better, we divided the data by the population in Mississippi or the general population in those years to get a percentage. This approach allowed us to normalize the data and make meaningful comparisons between Mississippi and the national averages.

Upon observation, these figures indicate that Mississippi's above-average rates of diabetes are accompanied by above average rates of SNAP participation, poverty rates, and obesity relative to the U.S. average. Most notably, Mississippi showcases a stark increase in fast food restaurants per population relative to the U.S. average, suggesting a strong association between fast food prevalence, poverty, and obesity. Given that SNAP eligibility requires low levels of employment and income, SNAP participation also serves as a general proxy for income. While these graphs do not demonstrate direct correlational or causal relationships, they align with the prior finding that SNAP participation is a higher correlation to diabetes rates than fast food availability.



(a) Obesity Rate Amongst Adults in 2017 as a Percentage

(b) Average Poverty Rates Comparison in 2015

(c) Average SNAP Participation as a Percentage

(d) Fast Food Restaurants (FFR) per Population

Figure. 8: U.S. vs Mississippi Comparisons along various socioeconomic factors

2.7 Time-Series Analysis

Having shown the geographic distribution of the health-based impacts on people, we then wanted to move on to analyzing these across time, and comparing them to the growth of fast food companies and the market at large, to find correlations that wouldn't be found in an ACS dataset, for example; to confirm causality between several different factors to provide meaningful real-world insight; and to forecast for the future.

2.7.1 Feature Engineering

To analyze the relationship between all the time-series data, several features were engineered from the datasets. Notable mentions include the fast food stock prices, for which we calculate the returns as the percentage of change in closing prices; the annual obesity rates, which we averaged for the entire nation yearly from the state-level data, and aggregating commodities together based on the least common sampling rate for whichever group of datasets we'd be working with. Since the effects of variables are rarely immediate, we also want to implement some kind of time 'lag' to measure the effects of past events.

2.7.2 Exploratory Data Analysis

When first exploring our data, we looked into common beliefs about fast food in the U.S. to inform our guiding questions. After graphing fast food stocks and obesity rates across the years, we found that there indeed was a relationship between the growth of these companies and the growth of food-caused illness in the U.S., and that both stocks and obesity rates were non-stationary according to the *The Augmented Dickey-Fuller (ADF) test* stationary test.

The ADF test is a statistical test used to determine whether a time series is stationary or has a unit root, indicating non-stationarity. The ADF test is based on the following regression equation:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \ldots + \delta_p \Delta y_{t-p} + \epsilon_t \tag{1}$$

where Δy_t is the first difference of the series, α is a constant, βt is a time trend, γy_{t-1} is the lagged level of the series, and $\delta_i \Delta y_{t-i}$ represents lagged differences. The null hypothesis of the ADF test is that the series has a unit root (i.e., it is non-stationary), which is tested by checking if $\gamma = 0$. If the null hypothesis is rejected, the series is considered stationary. For context, the average growth rate for fast food stocks was around 20.51% per year, and that for obesity rates was around 9.79%.

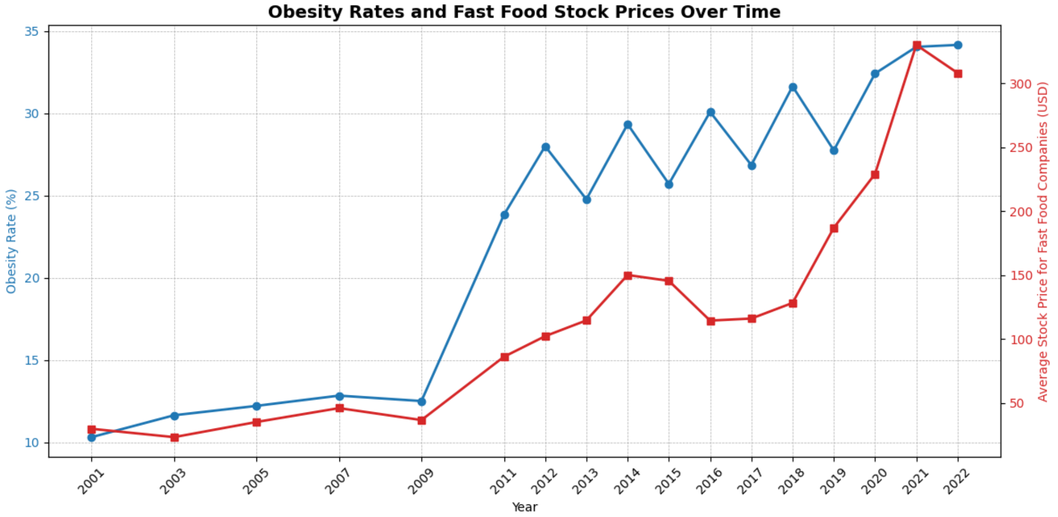


Figure. 9: A plot of the annual average value of fast food stocks and obesity rates; graphed yearly due to low resolution of obesity data

In our analysis, we aimed to explore the pairwise relationships between all variables that we considered relevant to economic impact and fast food consumption. Cross-correlation, which measures the similarity between a time series and a lagged version of another time series, was employed to uncover these relationships. For this, we utilized two main methods that would allow us to systematiclly and thoroughly identify any significant relationships and describe them accurately: cross-correlation, and Granger causality tests.

Cross-correlation allows us to determine the extent to which one time series can predict another by shifting one of the series in time. Formally, for two time series X_t and Y_t, the cross-correlation function $r_{xy}(k)$ at lag (or shift in values by time steps) k is given by:

$$r_{xy}(k) = \frac{\sum_t (X_t - \bar{X})(Y_{t+k} - \bar{Y})}{\sqrt{\sum_t (X_t - \bar{X})^2 \sum_t (Y_{t+k} - \bar{Y})^2}}$$

where $\bar{X}$ and $\bar{Y}$ are the mean values of X_t and Y_t, respectively. Immediately, one can see that $r_{xy}(0)$ is just the regular correlation coefficient between X_t and Y_t

To advance our understanding beyond correlation and establish potential causality, we employed *Granger Causality tests.* The Granger causality test, for which ADF stationarity is a prerequisite (stationarity would provide better results) examines if past values of one time series can predict future values of another time series. For two time series X_t and Y_t, X_t is said to Granger-cause Y_t if Y_t can be better predicted using past values of X_t rather than using only past values of Y_t.

The mathematical formulation involves estimating the following regression models:

$$Y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i Y_{t-i} + \epsilon_t$$

$$Y_t = \beta_0 + \sum_{i=1}^{n} \beta_i Y_{t-i} + \sum_{j=1}^{m} \gamma_j X_{t-j} + u_t$$

where:

- X_t Y_t is the value of the time series X,Y at time t
- α_0 and β_0 are the intercept terms.
- α_i and β_i are the coefficients for the past values of Y (i.e., Y_{t-i}).
- γ_j are the coefficients for the past values of X (i.e., X_{t-j})
- ϵ_t and u_t are the error terms at time t

In the first regression model, only the past values of Y are used to predict the current value of Y. In the second regression model, both the past values of Y and the past values of X are used to predict the current value of Y.

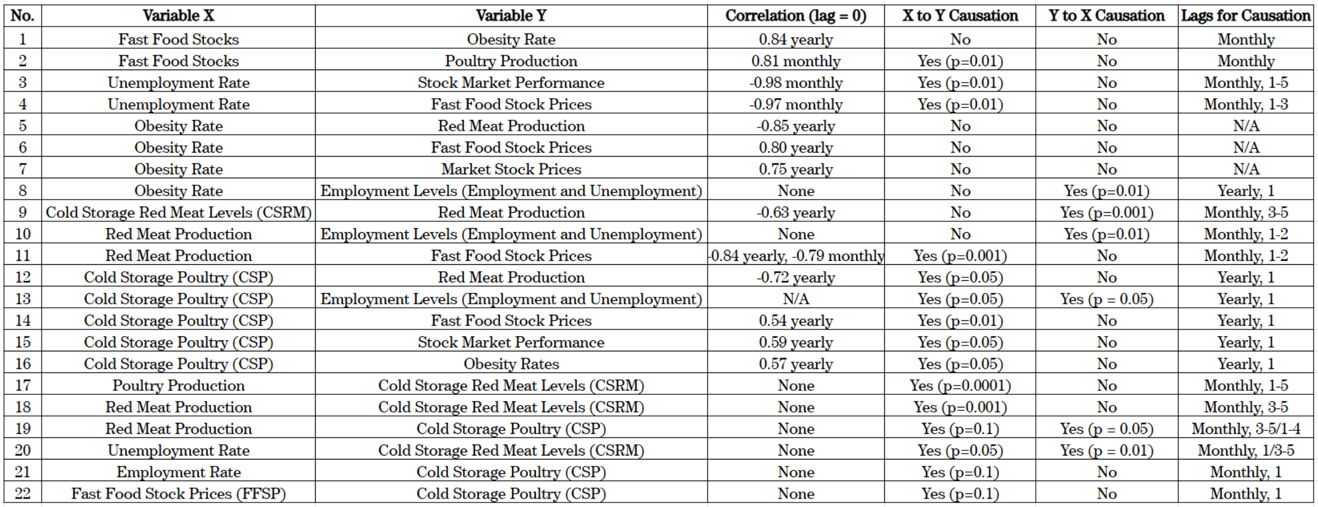
The null hypothesis (H_0) of the Granger causality test states that the coefficients $\gamma_j = 0$ for all j, which means that X_t does not 'Granger-cause' Y_t. If the null hypothesis is rejected, then the past values of X_t provide statistically significant information about future values of Y_t.

To support or reject the null hypothesis, we look at the p-values associated with the γ_j coefficients in the second regression model. A p-value below a certain significance level (generally 0.05, though many of the causal relationships we found were statistically significant at the 0.01 level or even lower), indicates that the past values of X_t significantly contribute to the prediction of Y_t. We can then write that X_t Granger-causes Y_t.

Granger causality tests were run twice for each pair of variables we examined; it is entirely possible (and in fact occured many times) that some time series X would Granger-cause another time series Y, but not the other way around.

2.7.3 Patterns Noticed

From the methods described above, we noticed several patterns in our data that warranted more exploration. We applied our causality model to our variables pairwise and compared the p-values. The table below summarizes the correlational and causal relationships that we found most pertinent to our study:



No.	Variable X	Variable Y	Correlation (lag = 0)	X to Y Causation	Y to X Causation	Lags for Causation
1	Fast Food Stocks	Obesity Rate	0.84 yearly	No	No	Monthly
2	Fast Food Stocks	Poultry Production	0.81 monthly	Yes (p=0.01)	No	Monthly
3	Unemployment Rate	Stock Market Performance	-0.98 monthly	Yes (p=0.01)	No	Monthly, 1-5
4	Unemployment Rate	Fast Food Stock Prices	-0.97 monthly	Yes (p=0.01)	No	Monthly, 1-3
5	Obesity Rate	Red Meat Production	-0.85 yearly	No	No	N/A
6	Obesity Rate	Fast Food Stock Prices	0.80 yearly	No	No	N/A
7	Obesity Rate	Market Stock Prices	0.75 yearly	No	No	N/A
8	Obesity Rate	Employment Levels (Employment and Unemployment)	None	No	Yes (p=0.01)	Yearly, 1
9	Cold Storage Red Meat Levels (CSRM)	Red Meat Production	-0.63 yearly	No	Yes (p=0.001)	Monthly, 3-5
10	Red Meat Production	Employment Levels (Employment and Unemployment)	None	No	Yes (p=0.01)	Monthly, 1-2
11	Red Meat Production	Fast Food Stock Prices	-0.84 yearly, -0.79 monthly	Yes (p=0.001)	No	Monthly, 1-2
12	Cold Storage Poultry (CSP)	Red Meat Production	-0.72 yearly	Yes (p=0.05)	No	Yearly, 1
13	Cold Storage Poultry (CSP)	Employment Levels (Employment and Unemployment)	N/A	Yes (p=0.05)	Yes (p = 0.05)	Yearly, 1
14	Cold Storage Poultry (CSP)	Fast Food Stock Prices	0.54 yearly	Yes (p=0.01)	No	Yearly, 1
15	Cold Storage Poultry (CSP)	Stock Market Performance	0.59 yearly	Yes (p=0.05)	No	Yearly, 1
16	Cold Storage Poultry (CSP)	Obesity Rates	0.57 yearly	Yes (p=0.05)	No	Yearly, 1
17	Poultry Production	Cold Storage Red Meat Levels (CSRM)	None	Yes (p=0.0001)	No	Monthly, 1-5
18	Red Meat Production	Cold Storage Red Meat Levels (CSRM)	None	Yes (p=0.001)	No	Monthly, 3-5
19	Red Meat Production	Cold Storage Poultry (CSP)	None	Yes (p=0.1)	Yes (p = 0.05)	Monthly, 3-5/1-4
20	Unemployment Rate	Cold Storage Red Meat Levels (CSRM)	None	Yes (p=0.05)	Yes (p = 0.01)	Monthly, 1/3-5
21	Employment Rate	Cold Storage Poultry (CSP)	None	Yes (p=0.1)	No	Monthly, 1
22	Fast Food Stock Prices (FFSP)	Cold Storage Poultry (CSP)	None	Yes (p=0.1)	No	Monthly, 1

Figure. 10: Correlation Coefficient Table of Various Causal Factors

2.7.4 Discussion of Relationships

1. **Fast Food and Obesity**
   - There is a strong positive correlation (0.84) between fast food stocks and obesity rates, but no direct causal relationship either way.
   - **Reasoning**: High-calorie, fat, and sugar content in fast food contribute to obesity when consumed in excess. Increased availability and affordability lead to higher consumption, especially in populations lacking access to healthier options. The parallel but not predictive relationship indicates that there are other external factors which might be influencing the trends for both of these variables

2. **Fast Food and Chicken Production**
   - Strong correlation between fast food stock prices and poultry production (0.81).
   - **Reasoning**:
     - **Poultry Production**: Poultry is the number one staple used by the fast food industry in many items like fried chicken and nuggets. Naturally, increased demand in the fast food sector drives poultry production.
     - **Causal Relationship**: Good fast food stock performance leads to higher meat production due to anticipated sales. Conversely, increased meat production signals higher supply, potentially lowering costs and boosting profitability.

3. **Unemployment Rate**
   - Strong negative correlation between unemployment rate and both stock market performance (-0.98) and fast food stock prices (-0.97).
   - **Reasoning**:
     - **Stock Market and Fast Food Stock Prices**: Higher employment strengthens the economy, leading to increased consumer spending, including on fast food.
     - Economic theories like the Philips Curve suggest that low unemployment leads to higher inflation, spurring economic growth.
     - **Lag Effects**: The labor market is a leading economic indicator for market performance, reflecting changes within 1-5 months.

4. **Cold Storage and Meat Production**
   - Strong negative correlation between cold storage red meat levels and red meat production (-0.63).
   - **Reasoning**:
     - **Red Meat**: Higher storage levels indicate a surplus, leading to reduced production to manage inventory and price stabilization.

  - **Causative Factors**: Fast food demand influences storage levels as companies stock up in anticipation of higher sales. Economic conditions impact storage capacities and strategies.
  - **Poultry Storage**: High correlation with stock prices and labor market can be linked to how the poultry industry adjusts to demand signals. Higher storage would ensure stable supply amid fluctuating demand.

5. **Employment and Meat Production**
   - Meat production (both red and poultry) Granger-causes employment levels, indicating production needs a steady workforce.
   - **Reasoning**:
     - Lower unemployment implies robust labor availability, facilitating higher production.
     - Increased meat production requires more labor, particularly in the fast food sector, affecting employment levels.
6. **Economic Conditions and Public Health**
   - Unemployment rates influence obesity rates.
   - **Reasoning**:
     - Economic downturns (higher unemployment) can increase stress and reduce physical activity, contributing to obesity.
     - Better economic conditions provide access to healthier food options and lifestyles and reduces obesity rates.

2.7.5 Assumption and Choices

The analysis is based on the assumption that time series data requires accounting for lags, one-way causality, and non-stationarity. We used the Augmented Dickey-Fuller (ADF) test to confirm the non-stationarity of the data, explicitly earlier for fast food stocks prices and obesity rates, and implicitly with all of the Granger causality tests.

Fast food stock prices were used as a proxy for fast food consumption, and obesity rates were chosen as a direct measure of health outcomes. Meat production and cold storage levels were included to their influence on food stocks and consumption patterns/obesity. Commodity prices were tested but no significant results were found in terms their potential causality on food prices and consumption patterns, though their correlations with fast food stock prices were really strong for monthly lags even up to two years.

2.7.6 Feature Selection

The features selected for this study include fast food stock prices, obesity rates, unemployment rates, and meat production and cold storage levels. Fast food stock prices were chosen due to their low volatility compared to the general stock market, especially during periods of economic disruption such as COVID-19. Despite efforts, we couldn't find better free dataset alternatives for obesity and unemployment data.

2.7.7 Model Performance and Limitations

Model performance for cross-correlation and Granger causality tests were evaluated based on predictive accuracy and the statistical significance of the results; i.e. we made sure that for each pair of variables, their correlational and causal relationships made sense and only gave them weight if they were statisticaly significant. Limitations included potential data quality issues, such as missing values and the granularity of temporal data. The lack of more granular and high-quality free data for obesity and unemployment rates posed constraints on the analysis. Additionally, Granger tests model predictability rather than true causality, so there may be some intervening factors that we are unaware of. Without some controlled experimental data, it is difficult to establish direct causal relationships.

2.7.8 Motivations and Meanings behind Statistical Tests and Visualizations

The motivations behind using statistical tests like Granger causality and visualizations were to identify and quantify relationships between variables over time. The ADF test was used to check for non-stationarity, ensuring reliable Granger causality analysis. Granger causality tests helped determine whether one time series could predict another, providing insights into the dynamics between food culture, health outcomes, and economic impacts and allowing us to formulate effective policy recommendations. Visualizations were used to illustrate and communicate these relationships clearly, highlighting the stability of fast food stocks and their potential resilience during economic disruptions like the COVID-19 pandemic.

3 Forecasting and Policy Recommendation

3.1 Forecasting

From our time series and geographical analysis, we noticed that factors such as meat production, cold storage levels, and fast food stock prices are strongly correlated with obesity and unemployment. To extract further information, we attempted to create models for predicting unemployment and obesity rates based on these factors.

3.1.1 VAR (Vector Autoregression) Analysis

Method

For forecasting the unemployment rate, we employed a Vector Autoregression (VAR) model. This method is appropriate for multivariate time series data and allows us to capture the linear interdependencies among multiple time series. We included variables such as fast food stock prices, red meat production, poultry production, and cold storage levels for both red meat and poultry. These variables were all highlighted to have a strong correlation with unemployment, as highlighted in section 2.7.3. We also used ARIMA (Autoregressive Integrated Moving Average) models to check the variables for the VAR analysis, but the detailed ARIMA results are provided in the appendix for further reference.

We first resampled the data to a monthly frequency and aligned the time series to ensure consistency across variables. The data was split into training and testing sets, excluding the COVID-19 spike period to avoid distortion. We used the Akaike Information Criterion (AIC) to determine the optimal number of lags for the VAR model. The model was then trained and used to forecast the unemployment rate for the next 12 months.

Results

Our VAR model performed reasonably well, capturing the trends in unemployment rates as shown in the graph below. We further validated our model using the Granger Causality Test, which confirmed the significant predictive power of the included variables on the unemployment rate, with all p-values at or below 0.05. Notably, the p-value for the impact of cold storage poultry on the unemployment rate was exceptionally low (p=0.0010), indicating a very strong predictive relationship, while the p-value for fast food stock prices was 0.0363.

These results indicate a statistically significant relationship, demonstrating that fast food stock prices, red meat production, and cold storage levels Granger-cause changes in unemployment rates, thereby confirming that the VAR model was a good choice for this analysis.

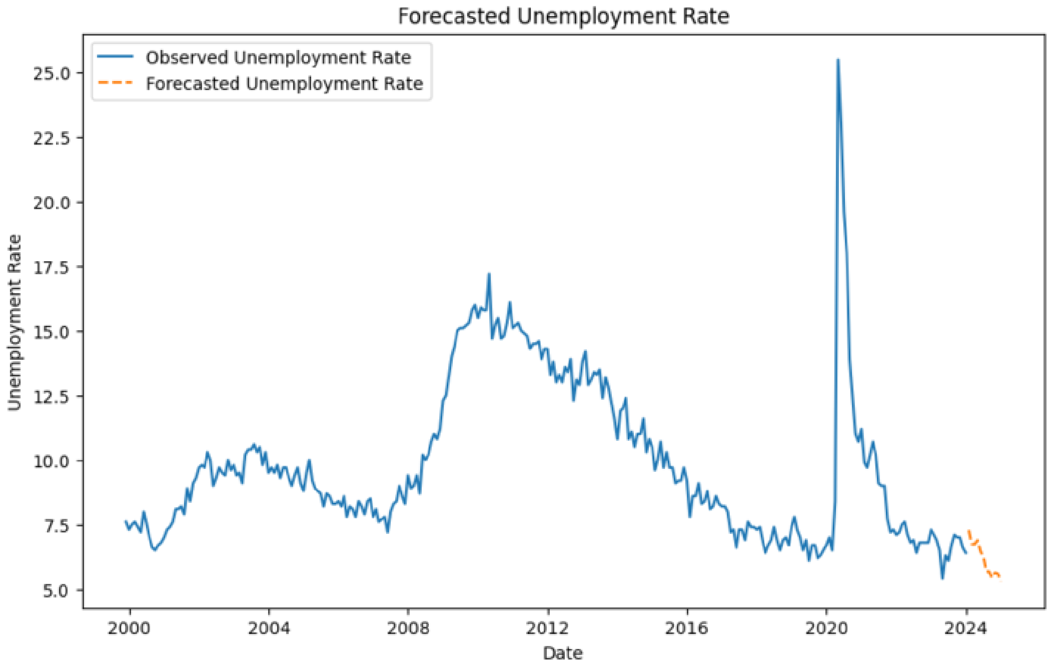


Figure. 11: VAR prediction of unemployment levels using fast food stock prices, red meat production, poultry production, and cold storage levels for both red meat and poultry

3.1.2 LSTM (Long Short-Term Memory) Analysis

Method

Given the observed correlation between unemployment and obesity, we extended our analysis to forecast obesity rates, which was one of the main themes of our exploration. For this task, we utilized a Long Short-Term Memory (LSTM) neural network model, enhanced with Gated Recurrent Units (GRU) for improved performance. LSTM and GRU are particularly well-suited for time series forecasting because they can capture long-term dependencies and patterns in the data. By using these advanced neural network architectures, we aimed to achieve a more accurate and robust prediction of obesity rates, leveraging the temporal relationships inherent in the data.

We opted not to use the Vector Autoregression (VAR) model for forecasting obesity rates because VAR models are typically better suited for capturing linear relationships among multiple time series. In contrast, the relationship between obesity rates and other influencing factors, such as unemployment, fast food stock prices, and meat production, is likely to be more complex and nonlinear. LSTM and GRU networks, with their ability to learn and model complex, non-linear relationships, provide a more flexible and powerful approach for forecasting in such scenarios.

To prepare the data, we normalized it using the MinMaxScaler, a common preprocessing step for LSTM models to ensure optimal performance. We created a dataset with a time lag sequence length of 24 months to allow the model to learn from past data points. We used a 4-to-1 train-test split to train our model and confirm its results.

Our LSTM model architecture included two GRU layers, each followed by a dropout layer to prevent overfitting. The model was compiled with the Adam optimizer and mean squared error as the loss function. We trained the model for 200 epochs with a batch size of 32.

Results

The LSTM model achieved a good fit with the observed data, as indicated by an R-squared value of 0.84, a Train-RMSE of 9.02, and a Test-RMSE of 7.7. The predicted obesity rates closely followed the actual observed rates, with adjustments applied to reduce fluctuations and improve accuracy. We aimed to evaluate its performance on existing data to see how well it could capture known trends. As shown in the graph, the model performs well on the existing data, validating our approach and demonstrating the potential of using LSTM models for this type of time series analysis for future forecasting.

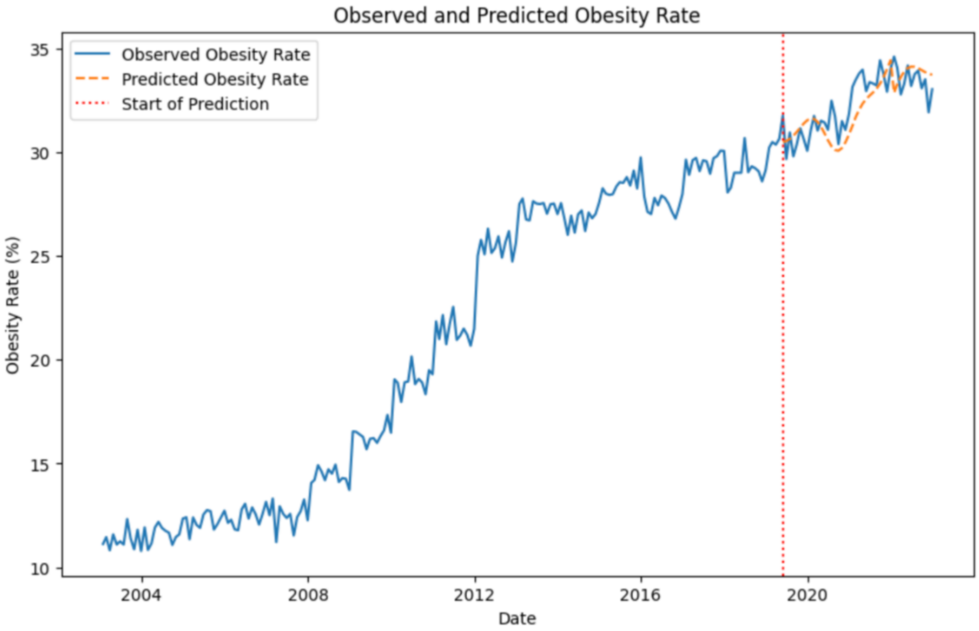


Figure. 12: caption here

3.1.3 Discussion

Our results suggest that both the VAR and LSTM models are effective in capturing the trends in unemployment and obesity rates. The strong predictive performance of these models underscores the importance of the included variables, which were selected based on their high correlation with the target variables.

In particular, the incorporation of meat production, cold storage levels, and fast food stock prices as predictors significantly enhanced the model's ability to forecast unemployment and obesity rates. These findings further highlight and reinforce the interconnected nature of food production, economic factors, and public health outcomes.

Considering the almost monotonically increasing nature of the fast food industry, and their practical resistance to economic crises due to them just being the cheapest food option across any scenario, we can safely say that fast food stock prices will very likely continue to increase as the years go on, and with them obesity rates and all the associated impacts on human life, from its quality to the amount spent on healthcare costs by individuals and the government.

Future work could focus on incorporating additional features such as economic indicators, geopolitical events, and market sentiment to further improve the models' predictive capabilities, especially in the context of unexpected market events. We could also incorporate our findings from county data to understand if specific areas in the US, such as Mississippi, could be more heavily impacted. Overall, this approach demonstrates the potential of advanced time series forecasting methods in policy planning and decision-making, providing valuable insights for addressing public health and economic challenges.

3.2 Policy Recommendation

Many current policies for addressing food insecurity target so-called ”food deserts,” or areas that lack access to grocery stores. New research has begun disputing this claim, however, arguing instead that lower-income communities aren’t able to afford healthy groceries, even when placed next to full-service stores. [16] Our findings confirm the second hypothesis, suggesting that America’s obesity-related health problems are primarily a result of poor economic access to healthy alternatives. When prices rise due to inflation, the easiest way to reduce costs is to purchase cheaper processed foods. In under-resourced areas, financial restrictions limit the amount of fresh produce that people can purchase. To rein in spending on processed foods, we recommend increasing access to low-cost nutritional options rather than focusing on expanding grocery store locations.

While the U.S. Supplemental Nutrition Assistance Program (SNAP) increases the purchasing power of low-income communities for produce in grocery stores, it’s important to note that there are likely drawbacks to SNAP’s approach to produce affordability. Historically, SNAP has been responsible for a massive reduction in the poverty level by focusing on supplementing families caloric intake. Families are given a discretionary dietary budget set by the US Department of Agriculture and calculated to provide a healthy diet on a limited budget. New studies have uncovered that SNAP households spend a larger proportion of their budget on processed sweetened foods than their non-SNAP counterparts. [17] Additionally, research has found that consumers report being more likely to purchase healthier options if the prices were lower. [18] As such, we suggest two potential policies to reduce food insecurity and obesity rates.

1. **Restrict SNAP Options.** Conservative leaders including Senator Marco Rubio have already begun pushing for SNAP restrictions to limit the discretionary spending of SNAP funds. Evidence pointing towards SNAP participants making choices based on personal preference suggests that restricting options could cut back on unhealthy options. [17] By limiting the ability to purchase processed foods in the first place, we could restrict hedonic spending and make SNAP benefits more nutritious.

2. **Increase SNAP Purchasing Power.** A contradictory stance to the first policy argues that such restrictions would be difficult to implement and ultimately reduce the ability to provide adequate nutrition, arguing that a healthy diet currently exists outside of a standard SNAP budget. A study in the Journal of the Academy of Nutrition and Dietetics argues that economic barriers prevent SNAP participants from affording a healthy diet and pushes them towards calorically dense processed foods. [20] Instead, an increase in SNAP funding could make healthy options more accessible and attractive.

Our proposed policies are backed by social science research that investigated longitudinal data on children’s participation in SNAP. As per Lorenzo Almada and Rusty Tchernis’ findings, additional SNAP benefits reduce the probability of being obese as an adult [19]. By restructuring an existing program that focuses on affordability for low-income communities as opposed to incentivizing private companies like Target to create more affordable grocery stores, we provide immediately actionable solutions to some of the hardest-hitting issues in the U.S.

Appendix

Other ancillary stuff we looked into:

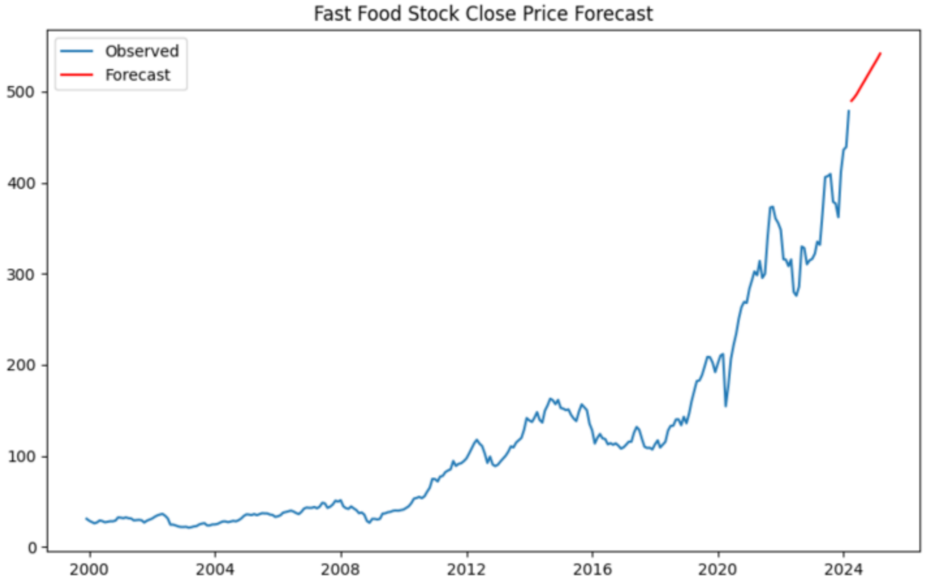


Figure. 13: ARIMA prediction for fast food stock levels

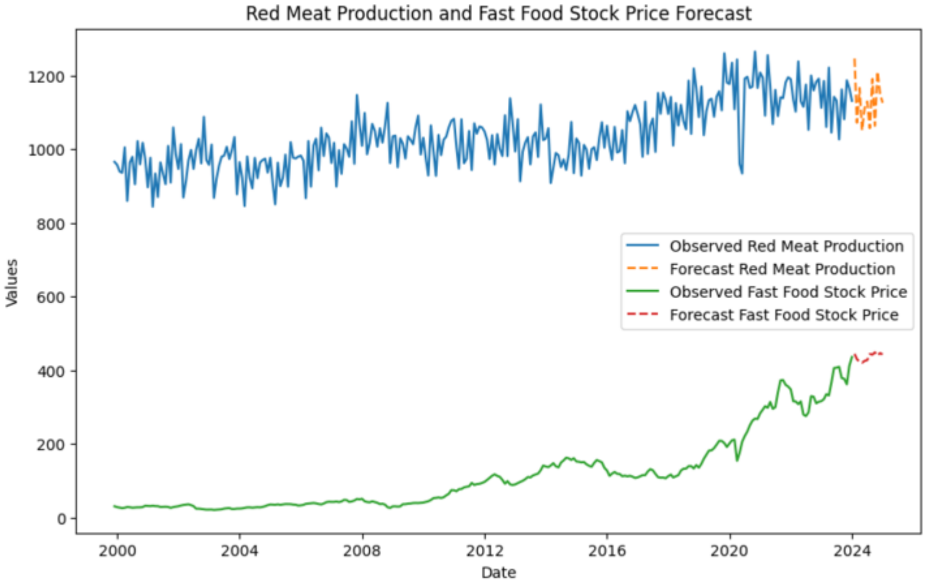


Figure. 14: ARIMA prediction for fast food stock levels and poultry production

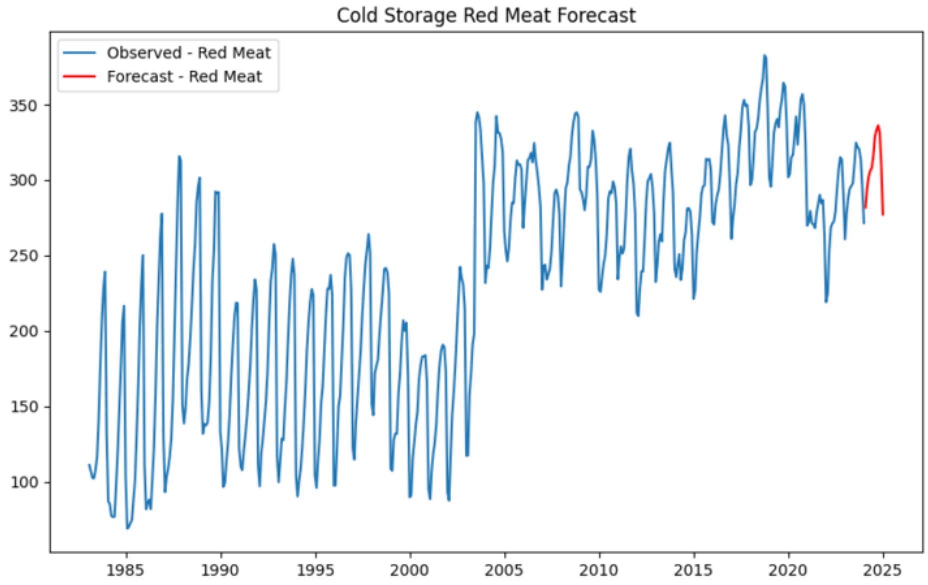


Figure. 15: Seasonal ARIMA prediction for cold storage red meat levels

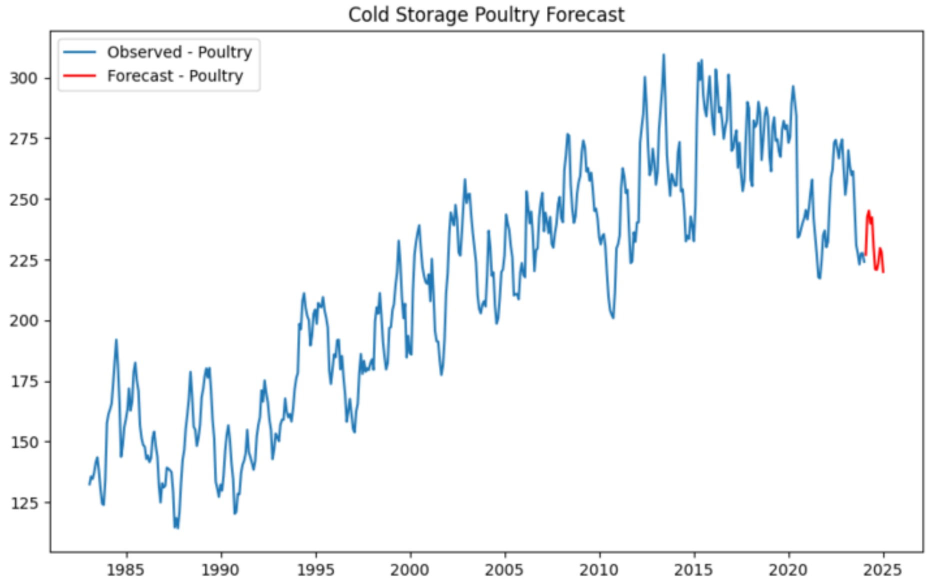


Figure. 16: Seasonal ARIMA prediction for cold poultry storage levels

References

[1] https://worldpopulationreview.com/country-rankings/fast-food-consumption-by-country

[2] https://www.statista.com/statistics/286541/mcdonald-s-advertising-spending-worldwide/

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6146358/

[4] https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity

[5] https://www.cdc.gov/pcd/issues/2009/jul/08_0163.htm#: :text=The%20term%20%E2%80%9Cfood%

[6] https://asmbs.org/resources/type-2-diabetes-and-metabolic-surgery-fact-sheet/

[7] https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4772793/

[9] https://diabetes.org/about-diabetes/statistics/about-diabetes#: :text=Deaths,a%20total%20of%20399%2C401%20certificates.

[10] https://pubmed.ncbi.nlm.nih.gov/37909353/

[11] https://www.ers.usda.gov/data-products/food-environment-atlas/

[12] https://www.cdc.gov/brfss/index.html

[13] https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2757497

[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6220876/

[15] https://catalog.data.gov/dataset/employment-unemployment-and-labor-force-data

[16] https://academic.oup.com/qje/article-abstract/134/4/1793/5492274?redirectedFrom=fulltext&login=false

[17] https://www.supermarketnews.com/news/why-do-snap-households-purchase-more-unhealthy-food

[18] https://time.com/4242944/making-healthier-foods-cheaper-could-save-millions-of-lives/

[19] https://www.nber.org/papers/w22681

[20] https://pubmed.ncbi.nlm.nih.gov/23260725/