Understanding Regional and Temporal Variations in Food Prices: Insights from US Consumer Price Index and Economic Trends*

— TODO: CHANGE — A Bayesian Approach Reveals the Impact of Economic Indicators and Geographic Disparities on Price Dynamics

Tanmay Sachin Shinde

November 28, 2024

This paper focuses on building an accurate and reliable model to explore the temporal dynamics of food-at-home prices in the United States at a national level. By incorporating key economic factors such as purchase volume, food categories, and the Consumer Price Index (CPI), our Bayesian Hierarchical Model (BHM) achieves high predictive accuracy (RMSE = X) and effectively forecasts price trends. The study captures category-specific effects and temporal trends, leveraging historical data to predict future price trajectories. **Results reveal that food prices exhibit a general upward trend, with volatility concentrated in categories like dairy and fresh produce. The CPI emerges as a primary driver, strongly correlating with price increases.** Our model will provide policymakers, businesses, and researchers with a robust predictive framework to anticipate market dynamics, mitigate risks, and make data-driven decisions that enhance economic resilience and consumer well-being.

Table of contents

1 Introduction **3**

2 Data **4**

2.1 Source and Overview 4

2.2 Data Processing and Cleaning 5

*TODO - CHANGE THE LINK - Code and data are available at: https://github.com/Tanmay-Shinde/Week10Reflection.

2.3 Measurement . . . 6
2.4 Outcome variable . . . 7
2.5 Predictor variables . . . 8
2.5.1 Time . . . 8
2.5.2 Category . . . 9
2.5.3 Region . . . 9
2.5.4 Consumer Price Index (CPI) . . . 9
2.5.5 Purchase Volume . . . 10

3 Model **12**
3.1 Overview . . . 12
3.2 Model Specifications . . . 13
3.2.1 Model Justification . . . 14

4 Results **15**
4.1 Model Metrics . . . 15
4.2 Predictions . . . 16

5 Discussion **21**
5.1 First discussion point . . . 21
5.2 Second discussion point . . . 21
5.3 Third discussion point . . . 21
5.4 Weaknesses and next steps . . . 21

A Appendix **22**

B Additional data details **22**
B.1 Measurement . . . 22
B.1.1 Data Preparation . . . 22
B.1.2 Unit Values . . . 23
B.1.3 Price Indices . . . 23
B.2 EDA for Outcome Variable . . . 25
B.2.1 Regional Trends . . . 25
B.2.2 Temporal Trends . . . 25
B.3 EDA for Predictor Variables . . . 26
B.3.1 CPI . . . 26
B.4 Purchase Volume . . . 26

C Model details **28**
C.1 AR(1) Model . . . 28
C.1.1 Overview . . . 28
C.1.2 Integrating AR(1) into BHM . . . 30
C.2 Posterior predictive check . . . 30
C.3 Diagnostics . . . 30

1 Introduction

Food prices are a critical component of economic and social well-being, directly affecting food security, diet quality, and household expenditures. The Food-at-Home Monthly Area Prices (F-MAP) dataset, developed by the USDA Economic Research Service, offers detailed insights into food pricing trends across the United States. Covering the years 2012 to 2018, the dataset provides monthly price data for 90 food categories across 15 geographic areas, making it a valuable resource for understanding regional and temporal variations in food costs. This paper leverages the F-MAP data to analyze how regional disparities and time-based trends influence food-at-home prices and explores the role of economic factors such as purchase volume, food categories, and the Consumer Price Index (CPI) in shaping these patterns.

The primary focus of this analysis is to estimate how food categories, various economic factors, and time affect overall national food prices in the U.S. The estimand is the expected food price for a given food category and time period, conditional on factors such as purchase volume and CPI. By modeling these variations, we aim to uncover the drivers of price changes and predict trends in food costs.

Using a Bayesian hierarchical model, this study analyzes monthly price data from the F-MAP dataset to uncover the drivers of food price variations. The model incorporates random effects to capture categorical disparities and fixed effects to analyze the impact of economic factors, such as CPI, purchase volume, and food categories, on price trends. The analysis also highlights categories with greater price volatility and quantifies the influence of these factors on national pricing dynamics. **The findings reveal that food prices have generally increased over time, with categories like dairy, fresh produce, and meats experiencing higher price volatility.** Economic factors such as purchase volume, food categories, and CPI significantly influence these trends, with higher CPI values strongly correlating with increased prices. Furthermore, food categories with higher demand or limited supply exhibit more substantial price fluctuations, underscoring the importance of understanding market-specific dynamics.

Understanding food pricing trends is essential for addressing issues related to food affordability and access. These insights are particularly valuable for policymakers and stakeholders aiming to reduce regional disparities, promote equitable access to food, and mitigate the effects of inflation on low-income households. By analyzing the drivers of price variations, this study provides a framework for informed decision-making in food policy and economic planning.

The remainder of this paper is structured as follows: Section 2 discusses the data sources, the F-MAP dataset and its variables, and pre-processing methods. Section 3 explains the Bayesian hierarchical model and methodology used for analysis. Section 4 presents the results, followed by a discussion of the key findings and conclusion of the study, as well as the limitations of

the data in Section 5. Finally, Section A — TODO: COMPLETE WHAT THE APPENDIX INCLUDES —.

2 Data

2.1 Source and Overview

The Food-at-Home Monthly Area Prices (F-MAP) data product (U. S. Department of Agriculture 2024) is a comprehensive and detailed data product developed by the USDA Economic Research Service (ERS) that provides monthly U.S. food price data for 90 food-at-home (FAH) categories across 15 geographic areas of the United States. The dataset includes two primary price measures for each food group, geographic area, and month: (1) a mean unit value price (dollars per 100 grams) and (2) price indexes derived using advanced index formulas. These measures enable researchers to track food price trends at a granular level and compare them across geographic and temporal dimensions, while accounting for economic factors such as consumer purchasing volume, store characteristics, and inflation metrics like the Consumer Price Index (CPI). By utilizing Circana OmniMarket Core Outlets retail scanner data, the F-MAP captures detailed consumer purchasing data from over 50,000 retail stores annually, including grocery stores, supercenters, and convenience stores.

The F-MAP provides data across the following dimensions:

- Monthly, 2012–18
- 15 geographic areas
  - Nationally
  - 4 Census regions: Midwest, Northeast, South, West
  - 10 metropolitan areas: Atlanta, Boston, Chicago, Dallas, Detroit, Houston, Los Angeles, Miami, New York, and Philadelphia
- 90 ERS Food Purchase Groups (EFPGs)
  - 8 groups for grains
  - 23 groups for vegetables
  - 8 groups for fruit
  - 8 groups for dairy and plant-based milk products
  - 14 groups for meat and protein foods
  - 4 groups for prepared meals, sides, and salads
  - 25 groups for other foods

For each of these month, area, and food group combinations, the F-MAP includes the following value variables:

- Purchase_dollars_wtd: Total weighted sales in U.S. dollars (nominal, i.e., not adjusted for inflation)
- Purchase_dollars_unwtd: Total unweighted sales in U.S. dollars (nominal, i.e., not adjusted for inflation)
- Purchase_grams_wtd: Total weighted quantities in grams
- Purchase_grams_unwtd: Total unweighted quantities in grams
- Number_stores: Number of stores in geographic area
- Unit_value_mean_wtd: Weighted mean unit value per 100 grams
- Unit_value_se_wtd: Standard error of weighted mean unit value
- Unit_value_mean_unwtd: Unweighted mean unit value per 100 grams
- Price_index_GEKS: Weighted price index value, constructed using Gini-Eltetö-Köves-Szulc (GEKS) formula

The F-MAP dataset is designed to align closely with the USDA Dietary Guidelines for Americans, facilitating research into food affordability, diet quality, and food security. Unlike other datasets, F-MAP offers monthly frequency data, making it particularly valuable for tracking short-term and seasonal trends. The dataset's hierarchical structure—spanning individual food categories, metropolitan regions, and national aggregates—supports diverse research applications. For example, it can be used to model the effects of policy interventions such as soda taxes or subsidies on dietary behavior and public health outcomes.

While other datasets such as the Bureau of Labor Statistics (BLS) Consumer Price Index (CPI) and the USDA Purchase to Plate National Average Prices (PP-NAP) provide useful insights, they fall short in capturing the comprehensive geographic and categorical detail offered by F-MAP. For instance, the CPI lacks subnational comparability across regions and provides limited food category detail, while the PP-NAP is focused on prepared foods and lacks the temporal granularity needed for trend analysis. The F-MAP dataset bridges these gaps by offering a more detailed, frequent, and regionally comparable resource.

The data is available to download in .xlsx format on the USDA Economic Research Service website. Specifically, we use the "Food-at-Home Monthly Area Prices, 2012 to 2018" dataset for our analysis.

2.2 Data Processing and Cleaning

We use the statistical programming language R (R Core Team 2023) to clean the data using various helpful packages like, readxl (Wickham and Bryan 2023), dplyr (Wickham et al. 2023), tidyr (Wickham, Henry, and Friedland 2023), lubridate (Grolemund and Wickham 2023), and arrow (Richardson et al. 2023). Further, libraries like ggplot2 (Wickham 2016), knitr (Xie 2023) were used to analyze the data and create visualizations.

The initial step involved removing unnecessary variables, retaining only those essential for analysis. This included variables such as unit_value_mean_wtd (weighted unit value), purchase_grams_wtd (weighted purchase volume), price_index_GEKS (Consumer Price Index),

EFPG (food category), metroregion_name (region), and time. Rows containing missing values in any of the key variables were removed to maintain data integrity. We also filter the data to only include the national-level data, since we want to capture price trends at a national level. The dependent variable, unitValue, was log-transformed to stabilize variance and linearize relationships, to ensure no assumptions are violations in modelling. The time variable, originally in year-month format, was converted into a continuous variable representing the number of months since January 2012. This allows temporal trends to be modeled effectively as a continuous variable. The category variable was standardized and converted into factor variables, enabling the model to treat them as categorical inputs. To improve the convergence of the Bayesian model, predictors such as cpi and purchaseVolume were normalized by centering them around zero and scaling them to have a standard deviation of one.

2.3 Measurement

The Food-at-Home Monthly Area Prices (F-MAP) dataset is built using high-frequency retail scanner data sourced from approximately 50,000–60,000 retail establishments annually. These include grocery stores, supercenters, club stores, drug stores, and convenience stores. The scanner data capture weekly sales in nominal dollars (not adjusted for inflation) and the quantities of food items sold. Weekly sales data are aggregated into monthly intervals to align with the dataset's temporal structure. If a sales week spans two months, the sales values and units are proportionately allocated based on the number of days in each month, ensuring temporal consistency.

To standardize the data, outliers in unit values are removed using the interquartile range (IQR) method, which identifies extreme values beyond 1.5 times the IQR from the 25th and 75th percentiles of the price distribution. This step eliminates inaccuracies that might arise from reporting errors or anomalous transactions. Package weights are converted into grams to ensure uniformity, using standard conversion factors (e.g., grams per ounce, fluid ounce, or pound). Prices are then expressed as unit values per 100 grams, providing a consistent measure of price across products of varying sizes.

The categorization of products into 90 detailed food categories is based on the USDA Economic Research Service (ERS) Food Purchase Groups (EFPG) system. This classification system organizes foods by their characteristics, such as ingredients, nutritional content, and convenience level. It aligns closely with the Dietary Guidelines for Americans and enables researchers to aggregate, disaggregate, or customize categories for specific research needs. These EFPG classifications are foundational to understanding price trends within and across food categories.

Retail sales data from certain retailers are reported at a broader Retailer Marketing Area (RMA) level rather than individual store locations. To ensure granularity, these RMA-level sales are disaggregated to individual stores using proportionate weighting methods, based on store-level weights developed specifically for the scanner data. These weights adjust the sales

data to reflect the population of stores nationally and regionally, ensuring that the dataset is representative of real-world purchasing behaviors. Both weighted and unweighted unit value estimates are included, enabling diverse analytical approaches.

The dataset also includes price indexes, which measure the cost of a basket of goods over time and across locations. The GEKS multilateral price index, the primary index in the F-MAP, is constructed to compare prices dynamically while accounting for product substitution and turnover. This index is based on the geometric mean of bilateral indexes (Laspeyres, Paasche, Fisher Ideal) and employs a 1-year rolling window to maintain transitivity and minimize chain drift. By capturing the cost of goods relative to a base period (2016–2018 national averages), these price indexes provide a robust measure for temporal and spatial price comparisons, making the dataset suitable for inflation and affordability analyses.

Through this rigorous process, real-world phenomena such as regional price variations, inflationary trends, and food category-specific dynamics are translated into structured data entries. The combination of high-frequency retail data, standardized unit values, and multilateral price indexes ensures that the F-MAP dataset is both comprehensive and precise, supporting its use in economic research and policy-making. A more detailed overview of the measurement process, including the data acquisition, adjustments, and methods for calculating the price index can be found in Appendix B.

2.4 Outcome variable

The response variable, unitValue, represents the weighted mean price per 100 grams of a food item within a specific food category, geographic region, and time period. In this analysis, the variable is log-transformed to create unitValue_lg, which helps stabilize variance, improve normality, and facilitate the interpretation of model coefficients in percentage terms. The log transformation allows us to understand the proportional changes in price rather than absolute changes, which is particularly useful for identifying relative price dynamics across categories and regions.

Table 1: Summary Statistics for log of Unit Value

mean	median	sd	min	max
-0.6922529	-0.7298112	0.6576483	-2.207275	1.086202

From Table 1, we can see that the transformed outcome variable, unitValue_lg, has a mean of -0.692 and a median of -0.713, indicating a near-symmetric distribution. The standard deviation of 0.687 reflects moderate variability in prices, while the range, spanning from a minimum of -2.813 to a maximum of 1.191, highlights significant disparities in food costs across categories, regions, and time periods. The log transformation normalizes the data, reducing skewness and

ensuring that extreme values in the original price variable do not disproportionately influence the analysis.

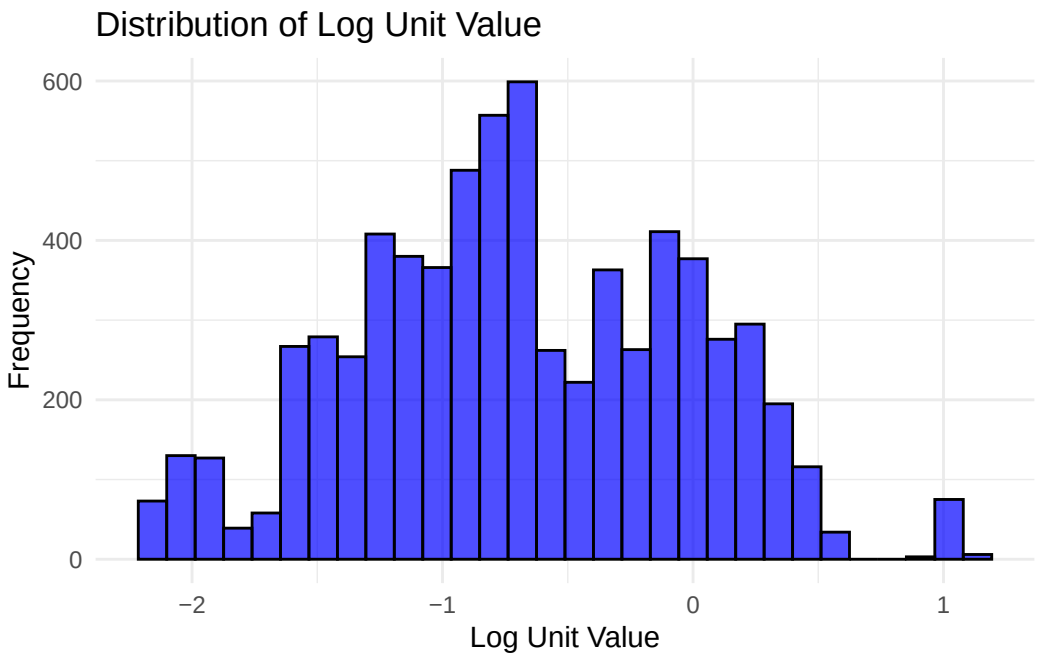

Figure 1: Distribution of Log-Transformed Food Prices (Unit Value)

Figure 1 illustrates a symmetric, unimodal distribution of unitValue_lg, centered near -0.7, with values ranging from approximately -2.8 to 1.2. This indicates that most food prices cluster around the median value, with fewer extreme high or low prices.

More details and exploratory analysis about the outcome variable can be found in Appendix B

2.5 Predictor variables

2.5.1 Time

The time variable in the analysis dataset is a continuous variable, representing the number of months since January 2012 when the data was recorded. This variable is important to help us understand temporal trends in the data and fluctuations in the unit value over time. The first month Jan-2012 is represented as 0, Feb-2012 as 1, and so on until Dec-2018 represented as 83.

2.5.2 Category

It is a categorical variable representing the 90 ERS Food Purchase Groups (EFPGs). The categories include different types of grains, fruits and vegetables, dairy, meat and protein foods, prepared meals, salads, and other foods.

2.5.3 Region

The region variable represents 15 geographic regions (10 metropolitan areas and 4 Census regions and 1 National category). the region variable will help us analyze the regional and geographical trends and variations in food prices. We are only interested in the 'National' region data for this study.

For each of these time, region, and food group combinations, the F-MAP includes the following value variables:

2.5.4 Consumer Price Index (CPI)

The Consumer Price Index (CPI) variable in the dataset represents the weighted price index (GEKS index) that measures the cost of a basket of goods over time, standardized relative to a base period (2016–2018). The CPI is unitless, with values below 1 indicating prices lower than the base period average and values above 1 indicating higher prices. The cpi is normalized to improve accuracy in modelling. This variable is critical in understanding inflationary trends and regional price dynamics, as it allows comparisons of relative purchasing power and economic conditions over time and across regions.

Each value recorded under the CPI column represents the consumer price index of one particular food group out of the 90 EFPGs in a particular region for that month.

Table 2: Summary Statistics for CPI

mean	median	sd	min	max
0	-0.0558982	1	-5.500379	11.51913

Table 2 illustrates the normalization of the variable with mean as 0 and standard deviation as 1. The minimum and maximum values of -7.04 and 22.79, respectively, highlight the presence of extreme values or outliers. These statistics indicate notable regional or temporal differences in CPI values, reflecting the varying economic conditions or price trends across the observed time period and regions. Figure 2 also shows a spread of the data from -7 to 22, with most of the values clustered around 0 with a fairly normal distribution.

More details and exploratory analysis about the cpi variable can be found in Appendix B

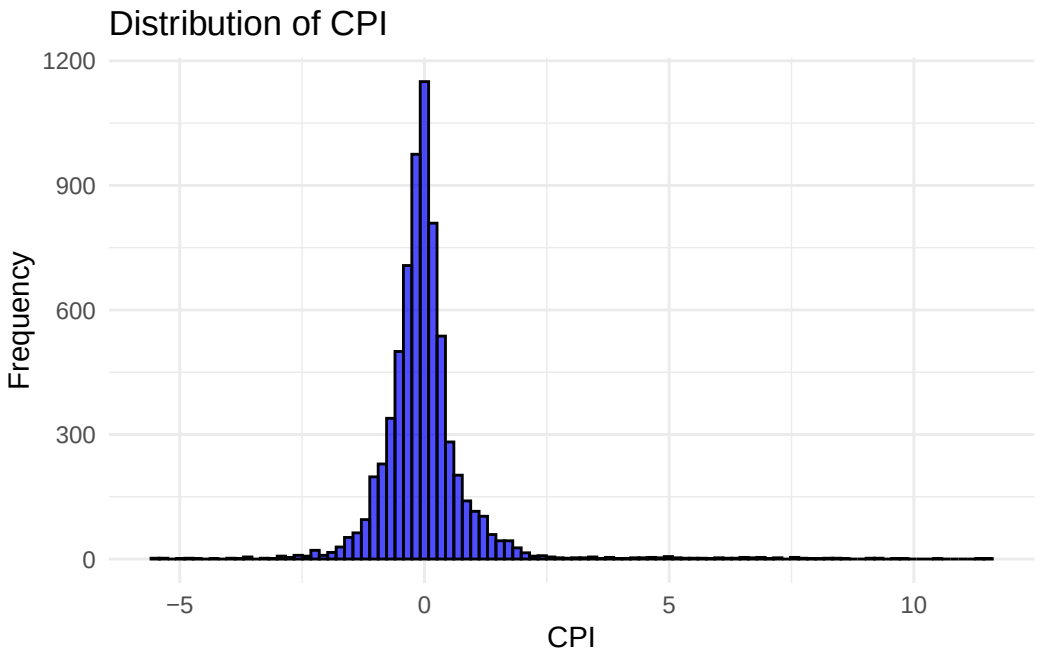

Figure 2: Distribution of GEKS Consumer Price Index (CPI)

2.5.5 Purchase Volume

The Purchase Volume variable represents the total quantity of food purchased, measured in grams, weighted by store-level survey weights to ensure representativeness across geographic areas and time. This variable is crucial in understanding consumer purchasing behavior, providing insights into the relationship between food quantity consumed and factors such as price, region, and time. It also helps in identifying demand patterns for various food categories and regions, which can influence pricing and policy decisions. High purchase volumes could indicate greater demand or accessibility, whereas lower volumes might reflect supply constraints, lower demand, or higher relative prices. This variable is key for examining the impact of economic factors like Consumer Price Index (CPI) and regional trends on consumer behavior.

Table 3: Summary Statistics for Purchase Volume

mean	median	sd	min	max
0	-0.4829133	1	-0.8398009	3.185714

Table 3 illustrates the normalization of the variable with mean as 0 and standard deviation as 1. The minimum and maximum values of -0.736 and 4.266, respectively, highlight the presence of extreme values or outliers on the right side. The median of -0.444 indicates that there might

be a slight right skew in the data. Figure 3 also shows a high spread in the data and a heavily right-skewed histogram that accurately represent the expected purchasing trends, with most purchasing volumes being at the lower end and then becoming lesser and lesser as one goes right.

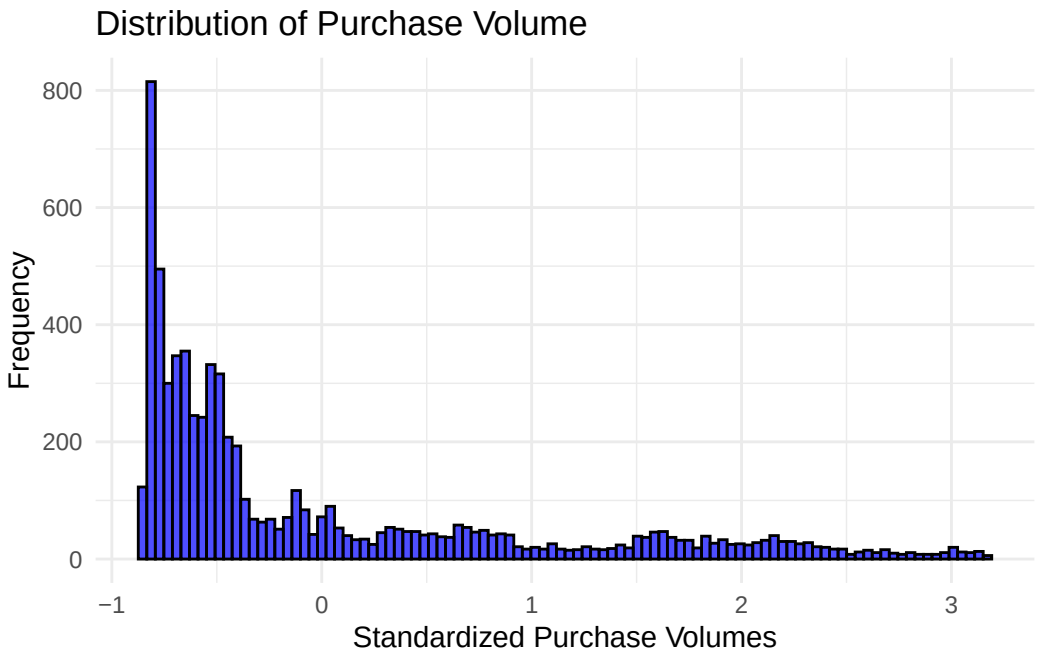

Figure 3: Distribution of Purchase Volume of Groceries

More details and exploratory analysis about the purchaseVolume variable can be found in Appendix B

3 Model

We employ a Bayesian Hierarchical Model (BHM) in this study to capture the complex relationships between food-at-home prices and their predictors, such as purchase volume, food categories, the Consumer Price Index (CPI), and time. The hierarchical structure allows the model to account for variability across different food categories while leveraging the overall data trends at the national level. By including time as a predictor, the model effectively captures temporal trends, while predictors like CPI and purchase volume elucidate economic and consumer-driven influences on prices.

3.1 Overview

The outcome variable is **unitValue_lg**, which represents the log of weighted mean price per 100 grams of a food item within a specific food category and time period. he log-transformation stabilizes variance and helps handle skewness in the price distribution, ensuring that the Gaussian family used in the model is appropriate.

cpi, **purchaseVolume**, and **time** are treated as fixed effects predictors. CPI models the overall inflationary trends and their influence on food prices. It is a continuous variable that captures macroeconomic factors affecting food price levels. Purchase Volume represents the total weighted quantity of food purchased. It acts as a demand-side predictor, reflecting consumer purchasing behavior's effect on price trends. Time is continuous variable representing the time elapsed since January 2012. It models the temporal trend in food prices, allowing the capture of long-term changes over time.

category captures variability in price trends across different food categories. By converting this variable to a factor, we allow each food category to have its own baseline price level while maintaining the overall hierarchical structure of the model.

To train the model, we use data from the year 2012 to 2017, and use the 2018 data as a benchmark to test the predictive accuracy of the model.

3.2 Model Specifications

Let y_{ijt} denote the log-transformed unit value (unitValue_lg) for category i and time t. Then, the model can be written as:

$$y_{it} = \beta_0 + \beta_i^{category} + \beta_1.CPI_{it} + \beta_2.purchaseVolume_{it} + \beta_3.time + \epsilon_{it}$$

Where:

- β_0: Global intercept (average log unit value across all regions and categories).
- $\beta_j^{category}$: Random effect for category i, capturing category-specific deviations.
- β_1: Coefficient for the CPI predictor.
- β_2: Coefficient for the purchaseVolume predictor.
- β_3: Coefficient for the time variable.
- $\epsilon_{ijt} \sim N(0, \sigma^2)$: Residual Error

We use the following heirarchical priors for our model:

- $\beta_0 \sim N(0, 10^2)$: Weakly informative prior for the global intercept.
- $\beta_j^{category} \sim N(0, \sigma^2_{category})$: Category-level random effects.
- $\beta_1, \beta_2, \beta_3 \sim N(0, 10^2)$: Weakly informative priors for the predictors
- $\sigma^2_{category} \sim HalfCauchy(0, 5)$: Priors for variance components.

We run the model in R (R Core Team 2023) using the `brms` package of Bürkner (2023).

3.2.1 Model Justification

The prediction of long-term trends in the outcome variable, log-transformed unit value (unitValue_lg), requires a model that can effectively handle complex data structures and temporal dynamics. We use a Bayesian Hierarchical Model (BHM) since the dataset has a natural hierarchical structure with food categories as grouping factors. BHM can model the variability within categories while borrowing strength across these groups through partial pooling. This approach improves predictive accuracy, by allowing group members to have some influence over one another. The data spans multiple years and involves time-dependent variations influenced by economic factors, purchase volumes, and Consumer Price Index (CPI).

We facilitate out-of-sample testing by dividing the dataset into two distinct, non-overlapping temporal segments: a training set (2012–2017) and a testing set (2018). The training set is used to build and calibrate the Bayesian Hierarchical Model (BHM), leveraging historical patterns in food prices, consumer purchasing volume, and other predictors like the Consumer Price Index (CPI). The testing set, comprising data exclusively from 2018, remains untouched during the model training phase, serving as unseen data for evaluating the model's performance.

Underlying Assumptions:

1. **Hierarchical Structure Validity**: The assumption that categories introduce random effects is reasonable given the heterogeneity in geographic and product-specific dynamics. The assumption is based on the nested nature of the data, where observations are grouped by region and food categories. This reflects real-world dynamics, as food prices are influenced by category-specific factors (e.g., perishability, demand patterns). By introducing random effects for regions and categories, the model captures these unobserved heterogeneities effectively. For example, the average price trend in one region may systematically differ from others, and this variation is modeled through the random intercepts. Without this hierarchical framework, the model would risk oversimplifying the underlying relationships, potentially leading to biased estimates and less reliable predictions.

2. **Normality of Residuals**: The BHM assumes that the residuals (differences between observed and predicted values) are normally distributed. This assumption underpins the use of Gaussian likelihoods in Bayesian modeling and ensures that the posterior distributions are well-defined and interpretable. Normality of residuals implies that the model captures all systematic patterns in the data, leaving random noise as the primary unexplained variation. Posterior predictive checks are used to validate this assumption, comparing the observed data distribution with the model-predicted posterior distributions. If deviations from normality are detected (e.g., skewness or heavy tails), adjustments such as alternative likelihoods (e.g., Student-t distributions) can be incorporated. The normality assumption is reasonable here because the dataset represents aggregated measures (e.g., regional or category-specific prices), which tend to exhibit normal-like properties due to the Central Limit Theorem.

Limitations:

1. **Limited Explanatory Variables**: The model incorporates CPI and purchase volume as key predictors, which are useful but may not fully capture the multifaceted nature of food price dynamics. External factors such as supply chain disruptions, climatic events, or geopolitical influences on agriculture could significantly impact prices yet remain unaccounted for in the model. The exclusion of such variables may lead to omitted variable bias, where the model attributes unexplained variability to random effects rather than systematic external factors. Expanding the set of predictors to include weather patterns, fuel prices, or global trade indices could enhance the model's comprehensiveness and accuracy, albeit at the cost of increased complexity.
2. **Computational Complexity**: Bayesian hierarchical models, especially when paired with ARIMA components, require substantial computational resources due to their iterative sampling-based inference methods, such as Markov Chain Monte Carlo (MCMC). While modern software like brms in R or PyStan in Python facilitates these computations, running the model on large datasets with multiple hierarchical levels can still be time-intensive and hardware-dependent. This limitation might restrict the ability to conduct extensive sensitivity analyses or real-time updates. Additionally, ensuring model convergence requires careful tuning and monitoring, which demands expertise and can prolong the modeling process. Despite these challenges, the benefits of Bayesian methods—such as flexibility and probabilistic interpretation—often justify the computational overhead.

4 Results

4.1 Model Metrics

The Bayesian model, presented in Table 4, explores the relationships between the unit value of a food item within a specific food category and predictors such as cpi, consumer purchase volumes, and time.

The intercept estimate of -0.745 represents the baseline log-transformed unit value (unitValue_lg) when all predictors—Consumer Price Index (CPI), purchase volume, and time—are at their reference or mean levels. A negative intercept suggests that the baseline food-at-home prices, in log scale, are below zero, implying prices below 1 on the original scale.

The CPI predictor has an estimate of 0.044, indicating that a one-unit increase in CPI leads to a 0.044 increase in the log-unit value, holding all other variables constant. This positive relationship aligns with expectations, as rising inflation levels typically correlate with increasing food prices.

The purchase volume variable exhibits a negative relationship with the response, with an estimate of -0.170. This means that for every one-unit increase in normalized purchase volume,

the log-unit value decreases by 0.170, holding other variables constant. This result is consistent with economic principles, reflecting economies of scale where bulk purchases reduce per-unit costs.

The time variable shows a small positive effect on log-unit value, with an estimate of 0.001. This result suggests a gradual increase in food prices over time.

Table 4: Coefficients of Predictor Variables

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
Intercept	-0.745	0.073	-0.890	-0.583	1.234
cpi	0.044	0.001	0.043	0.046	1.002
purchaseVolume	-0.170	0.005	-0.179	-0.161	1.010
time	0.001	0.000	0.001	0.001	1.000

Across all predictors—CPI, purchase volume, and time—the standard errors are small, indicating high precision in the parameter estimates. This precision is particularly evident in the case of CPI and time, where the standard errors are close to zero, signifying that the data provides strong evidence for the relationship between these variables and the log-unit value.

Table 5 provides a clear assessment of the model's performance. With 5944 observations, the model uses a substantial dataset, enhancing the reliability of its estimates. The R^2 value of 0.744 shows that 74.4% of the variability in the log-transformed unit value is explained by the predictors, indicating a strong fit and relevance of variables like CPI, purchase volume, and time. An AIC of -16810.9 reflects the model's effectiveness in balancing fit and complexity. The high log-likelihood (Log.Lik.) value of 8405.5 further supports the model's ability to explain the data well. The Root Mean Square Error (RMSE) of 0.058 indicates high predictive accuracy, showing minimal difference between observed and predicted values. These metrics confirm that the model is robust and reliable for capturing trends and making future predictions.

4.2 Predictions

In this section, we use the model to make some predictions, to visualize and assess the accuracy of the model. As mentioned before, we the data from Jan 2018 to Dec 2018 serves as our test dataset. We will look at the model's predictions for the unit values of 5 food categories and compare them against the actual recorded values.

Figure 4 visualizes the comparison between the actual and predicted log-transformed unit values (unitValue_lg) for whole fruit prices during the test period (January to December 2018).

Table 5: Explanatory model of Unit Value

```
\begin{table}
\centering
\caption{\label{tab:tbl-modelresults2}Model Metrics}
\centering
\begin{tabular}[t]{l|r}
\hline
Metric & Value\\
\hline
Num.Obs. & 5944.000\\
\hline
R² & 0.744\\
\hline
AIC & -16810.900\\
\hline
Log.Lik. & 8405.500\\
\hline
RMSE & 0.058\\
\hline
\end{tabular}
\end{table}
```

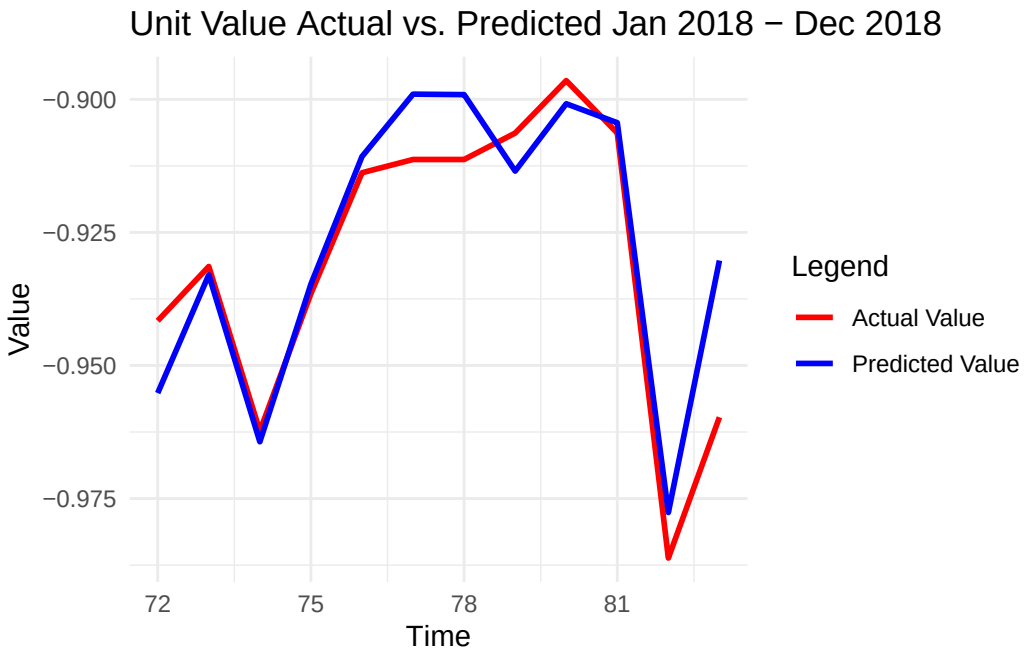

Figure 4: Prediction for Whole Fruit Prices

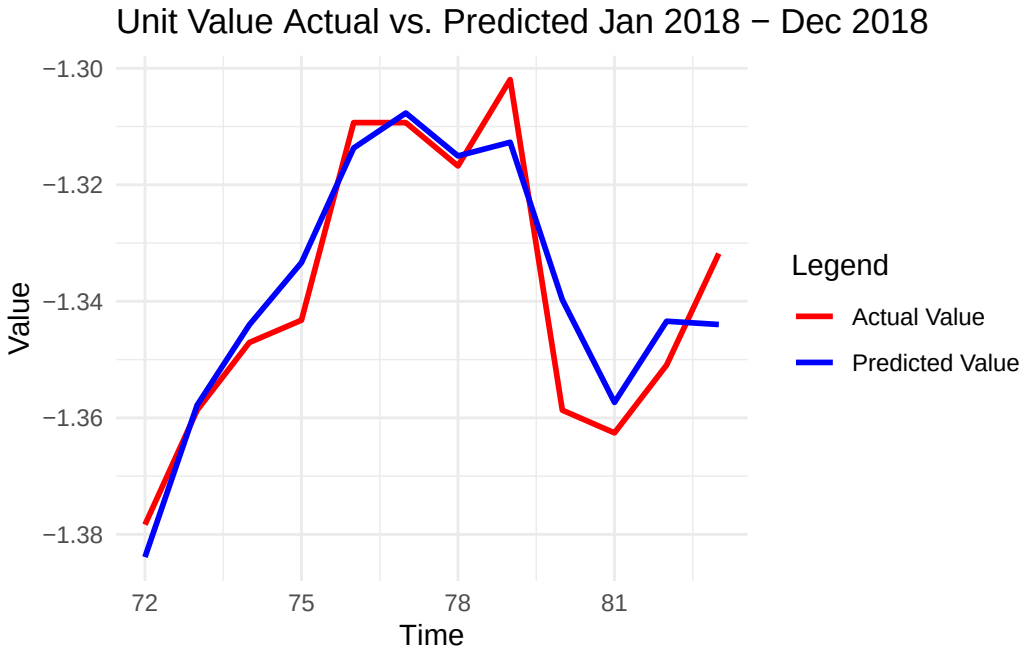

Figure 5: Prediction for Canned Tomatoes Prices

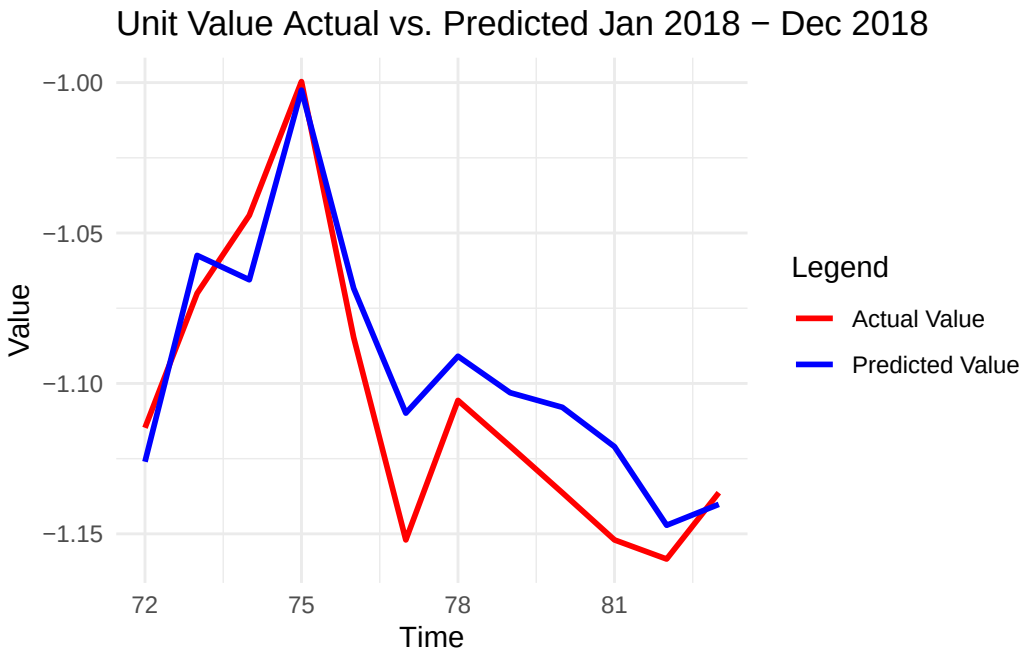

Figure 6: Prediction for Eggs and Egg Substitutes Prices

Figure 5 visualizes the comparison between the actual and predicted log-transformed unit values (unitValue_lg) for tomatoes (canned) prices during the test period (January to December 2018).

Figure 6 visualizes the comparison between the actual and predicted log-transformed unit values (unitValue_lg) for eggs and egg substitutes prices during the test period (January to December 2018).

Figure 7 visualizes the comparison between the actual and predicted log-transformed unit values (unitValue_lg) for canned starchy vegetable prices during the test period (January to December 2018).

Figure 8 visualizes the comparison between the actual and predicted log-transformed unit values (unitValue_lg) for bacon, sausage, and other lunch meat prices during the test period (January to December 2018).

The predicted values for all categories closely follow the actual trend, demonstrating the model's ability to capture temporal and category-specific variations effectively. While minor deviations are visible at certain points, the overall alignment underscores the predictive accuracy of the Bayesian hierarchical model. These results validate the model's reliability in forecasting price trends, providing valuable insights for economic analysis and policy-making.

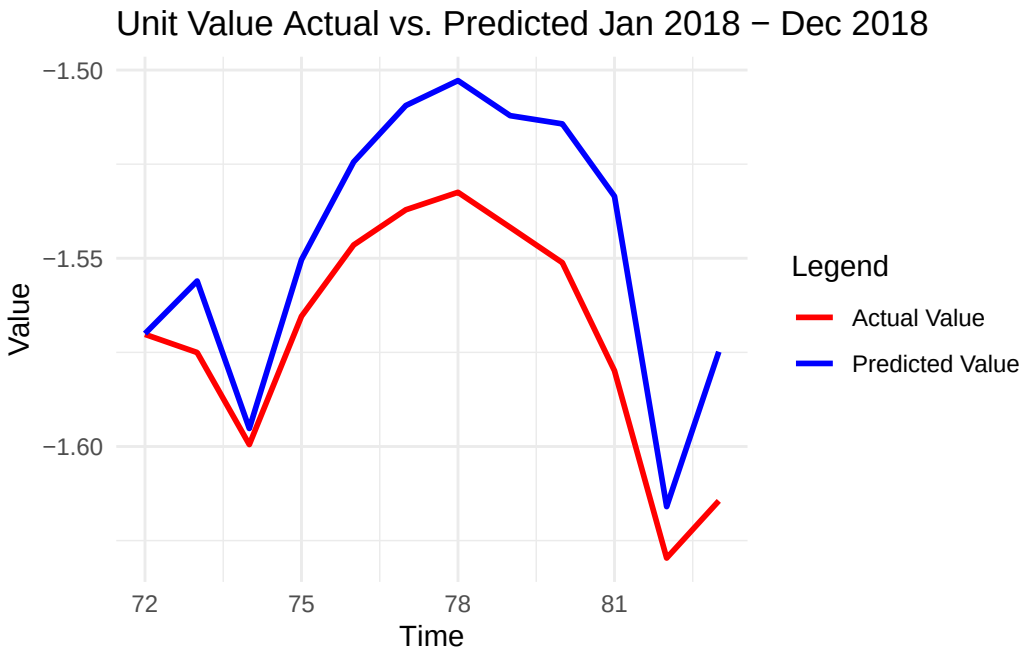

Figure 7: Prediction for Canned Starchy Vegetables Prices

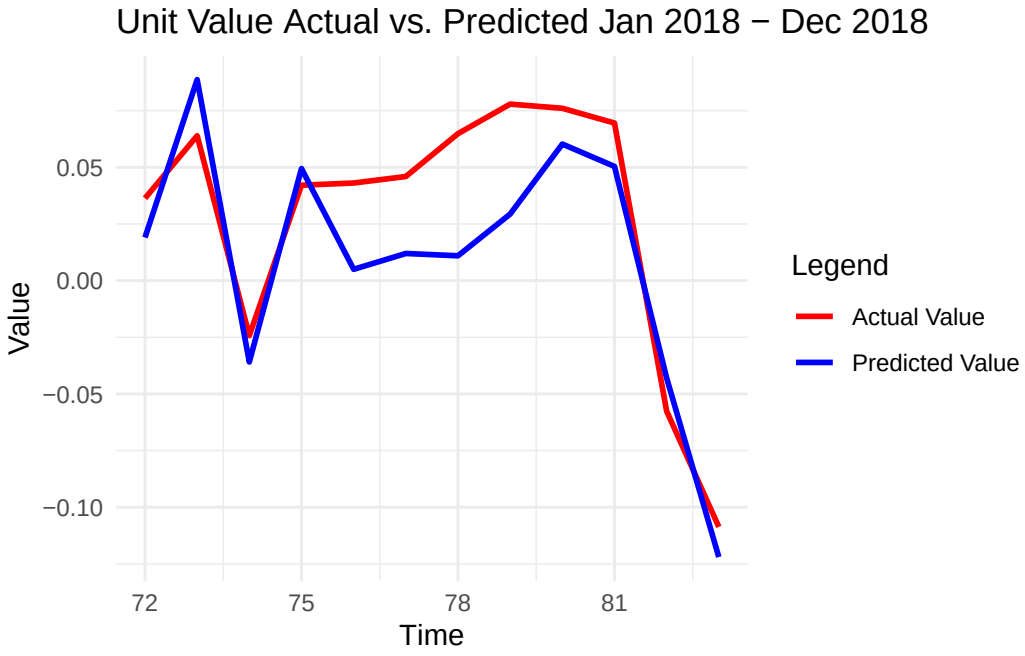

Figure 8: Prediction for Bacon, Sausage, and Lunch Meat Prices

Plots for all 90 categories are available in the "other/plots/" directory. The code to generate all plots is available in "scripts/05_model_data.R"

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

B Additional data details

B.1 Measurement

This section describes the methods for constructing the price measures in the F-MAP.

B.1.1 Data Preparation

The process to prepare the datasets for creating the F-MAP data product are as follows:

1. The retail scanner data report sales on a weekly basis. Weekly sales are grouped into the respective months that the sales occurred. In cases where the week straddles 2 months, sales units and values are allocated proportionately based on the number of days in each month.
2. Unit value outliers are eliminated using the interquartile range (IQR) method. The IQR is the difference between the 25th and 75th percentiles of the price distribution, in this case across all unit values by store and week for each item. A unit value is considered an outlier if the value is below the 25th percentile minus 1.5 multiplied by the IQR or above the 75th percentile plus 1.5 multiplied by the IQR.
3. The weights of each package are converted into grams to calculate unit values on a per 100-gram basis:
    a. Convert from ounces: gram weight = 28.35 x ounces per package
    b. Convert from pounds: gram weight = 28.35 x 16 x pounds per package
    c. Convert from fluid ounces: gram weight = 29.57 x fluid ounces per package
4. Individual food items sold in the scanner data (about 600,000 per year) are identified and categorized into 90 food groups, based on the EFPG classification system.
5. Retailer Marketing Area (RMA) sales data are disaggregated to individual stores. In the retail scanner data, most retailers release data by individual store location. However, some retailers only release data by RMA, a grouping of stores in a retailer-defined geographical area. To disaggregate the RMA sales data to individual stores, the RMA sales data are proportioned to individual stores based on the store sales values in the store-level weight files developed for the retail scanner data. For more information about the store weights, see the Using Proprietary Data page.

6. Store-level survey weights are applied to each store. Stores in the retail scanner data are not a representative sample of stores, and store-level weights adjust the sales data to be representative of the population of stores nationally and for each geographic area in the F-MAP. In the F-MAP datasets, unit values are provided as both weighted and unweighted estimates, and the price indexes were calculated using weighted data.

B.1.2 Unit Values

Sales (in U.S. dollars) and quantity (in grams) are summed over each month, EFPG, and geographic area. Mean unit values in the F-MAP are calculated by dividing the food group sales by the food group quantity and are standardized to the price per 100 grams. This process is completed as follows: (1) calculate the total purchase values in dollars and in grams for each EFPG in a given month and geographic area, weighted by the store weight for that year of data (note, weight is 1 for unweighted estimates); and (2) divide the total (weighted or unweighted) purchase dollars by the total (weighted or unweighted) grams to get the unit price.

The F-MAP also includes standard errors for the weighted unit values. Standard errors are a measure of the precision of survey estimates and can be used to construct confidence intervals for an estimate. Confidence intervals represent a range of values that are likely to include the actual population mean. Standard errors of the weighted unit values are calculated by re-estimating the weighted unit values 200 times, using replicate weights, and then using the general formula for standard errors. These calculations are described in more detail in the Development of the Food-at-Home Monthly Area Prices Data report.

B.1.3 Price Indices

Price indexes are a unitless measure of the cost of a basket of goods and are used to measure price changes over time. A price index converts many item-level price comparisons into a single value that quantifies the overall price of the basket at a time and location relative to a base period. The base period for the F-MAP is the national average for each EFPG from 2016 through 2018. Index values lower than 1 indicate prices lower than the national average from 2016 through 2018, while index values higher than 1 indicate prices higher than the national average from 2016 through 2018.

The primary F-MAP price index is constructed using a weighted GEKS index formula (named for contributors Gini, Eltetö, Köves, and Szulc). GEKS is a multilateral price index specifically designed to compare prices over time and space. A GEKS index can also be extended for future years without revising the index numbers that have already been published. A GEKS price index is available for all years of the F-MAP (2012–18). A set of supplemental indexes is also available for 2016–18 as a research series, which includes the bilateral Laspeyres, Paasche, Törnqvist, Fisher Ideal indexes and the multilateral Caves-Christensen-Diewert (CCD) index.

Multilateral price indexes are transitive, which means that any month-area pairing (or entity) can be compared directly with another pairing or through a third pairing, and the ratio between any two pairings is independent of the choice of base period. Transitive indexes are advantageous if the mix of goods being measured is dynamic; that is, if the basket of goods changes due to product turnover. Indexes based on scanner data are dynamic because the indexes include all goods sold in stores, which may change in each time period, rather than a sample of goods selected through a survey. Indexes that are transitive also allow spatial comparisons, regardless of the choice of area used for the base.

As additional years of price data become available beyond the base period, the GEKS index can be updated using a rolling window, or the time period over which the index is calculated. In standard multilateral indexes, as new data become available beyond the initial base period, the index numbers for existing entities must be recalculated because the multilateral index compares product prices in an entity with prices in all other entities. A rolling-window GEKS index compares product prices in a new entity with prices of entities within a rolling window. The F-MAP GEKS uses a 1-year rolling window, which allows maintaining published indexes without revising historical numbers.

Bilateral indexes with a fixed base period can become less representative of the cost of food as the indexes move further away from the base, due to the effects of product turnover, as products are discontinued and new products are introduced. Although bilateral price indexes can be updated using chained indexes, which capture product substitution, chained indexes are subject to drifting. Chain drift is a phenomenon in which the price index drifts lower even as item-level prices return to their base levels. Multilateral indexes are fully transitive and free of chain drift. While chain drift is possible in rolling-window multilateral indexes, if a wide window length is chosen, the rolling-window index will be largely free of drift despite not being fully transitive. The 1-year rolling window used in the F-MAP GEKS has been found to be sufficient to remove chain drift caused by high-frequency data and seasonal variation in variety.

The GEKS index builds upon bilateral indexes as elements. The multilateral GEKS index is the geometric mean of all possible Fisher Ideal index month-area pairings. The Fisher Ideal index is, in turn, based on the geometric mean of the Laspeyres and Paasche indexes. Therefore, constructing a GEKS index requires first calculating Laspeyres, Paasche, and Fisher Ideal indexes. We avoid going into deeper details of the above indices as the math gets extremely complicated. For more information about the construction of the F-MAP GEKS, the rolling-window GEKS, and the five supplemental indexes (i.e., Laspeyres, Paasche, Törnqvist, Fisher Ideal, and Caves-Christensen-Diewert (CCD)), you can refer to ERS's report on the F-MAP methods. (Sweitzer et al. 2024)

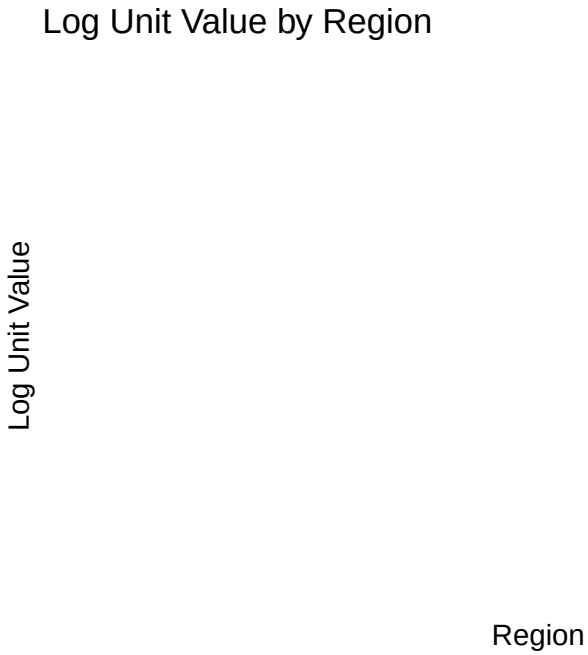

Figure 9: Regional Variations in Log-Transformed Food Prices

B.2 EDA for Outcome Variable

B.2.1 Regional Trends

Figure 9 reveals slight regional variations in log unit values. The Northeast and West regions show marginally higher median prices compared to the Midwest and South, reflecting higher food costs often associated with coastal or urban areas. The Midwest exhibits the largest variability, including a notable outlier on the high end, suggesting either an expensive food category or a regional anomaly. Despite these differences, the overall consistency in median prices across regions highlights a broadly stable pricing structure with minor geographic deviations. These trends emphasize the need to investigate the economic and geographic factors contributing to these regional disparities.

B.2.2 Temporal Trends

Figure 10 illustrates the log-transformed unit value of food prices from 2012 to 2018, revealing a general upward trend over time. Initially, from 2012 to 2014, there is only a modest increase in the log-transformed values. However, in 2014, prices increase sharply followed by a period of 3 years characterized by heightened volatility, with pronounced peaks and troughs, and

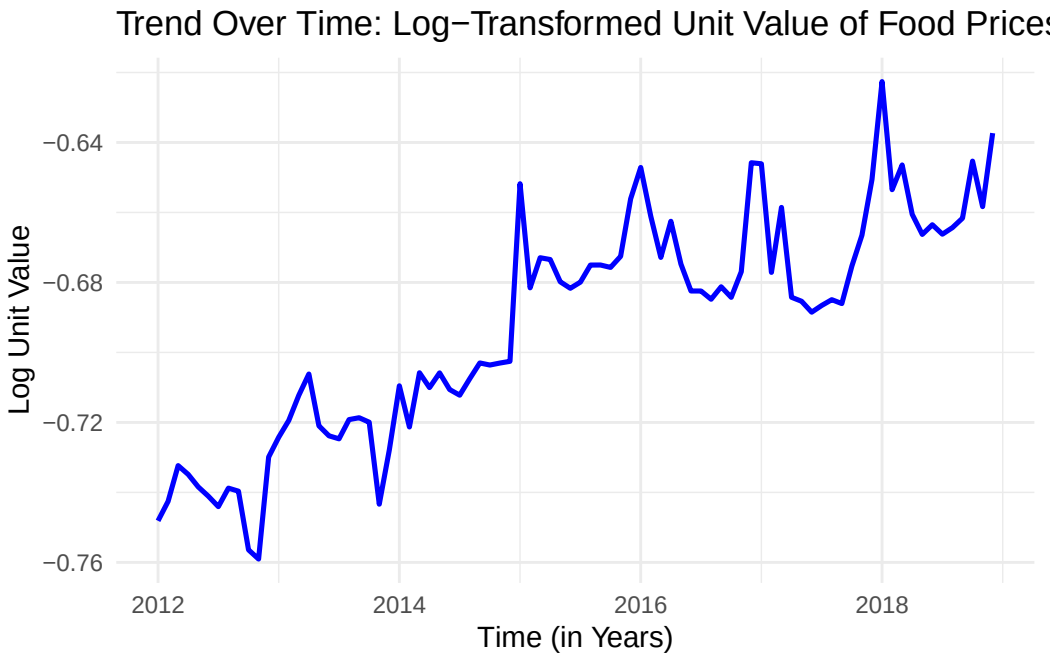

Figure 10: Temporal Variations in Log-Transformed Food Prices

another increase in the prices in 2018. Thus, the overall trajectory for the log-transformed unit value of food prices from 2012 to 2018 remains upwards.

B.3 EDA for Predictor Variables

B.3.1 CPI

Figure 11 reveals noticeable patterns and variations across Census regions. Each region exhibits a centralized cluster of values around the median, with relatively narrow interquartile ranges, indicating that the bulk of CPI values are similar within each region. However, all regions display significant outliers. From the boxplot, we can clearly see that the overall prices in the northeast and midwest regions are comparatively higher than prices in the south and the west census regions, with the west having some of the lowest cpi values out of all the regions.

Figure 12 also illustrates the general upward trend in the CPI values from 2012 to 2018.

B.4 Purchase Volume

From Figure 13, we can see that across all regions, the mean purchase volume is relatively consistent, suggesting similar central tendencies in the amount of food purchased. However,

Figure 11: Regional Variations in CPI

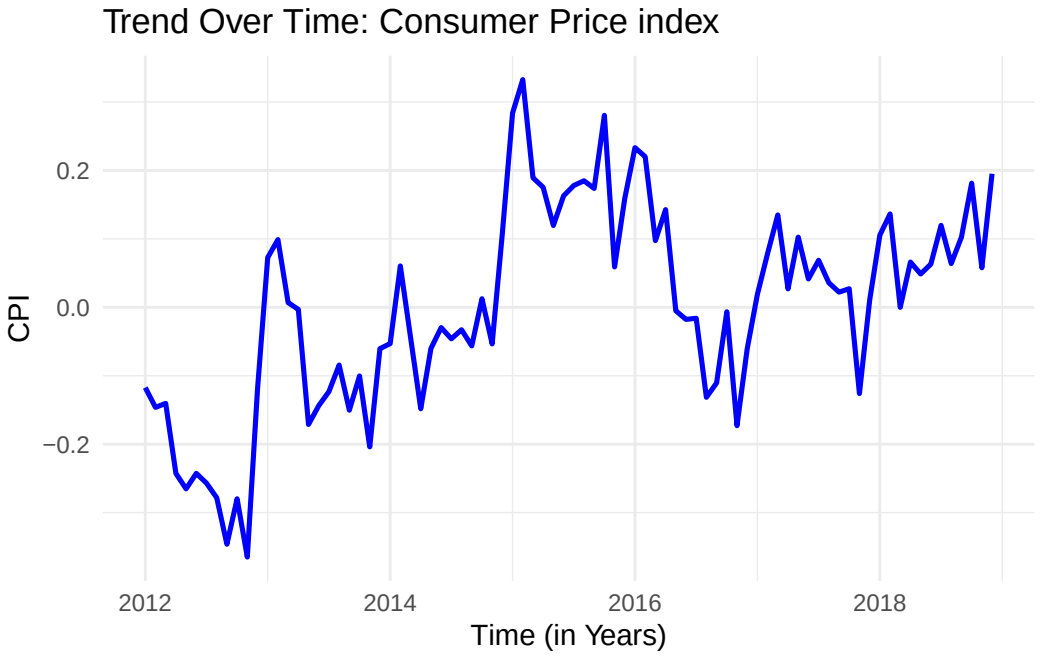

Figure 12: Temporal Variations in the Consumer price Index (CPI)

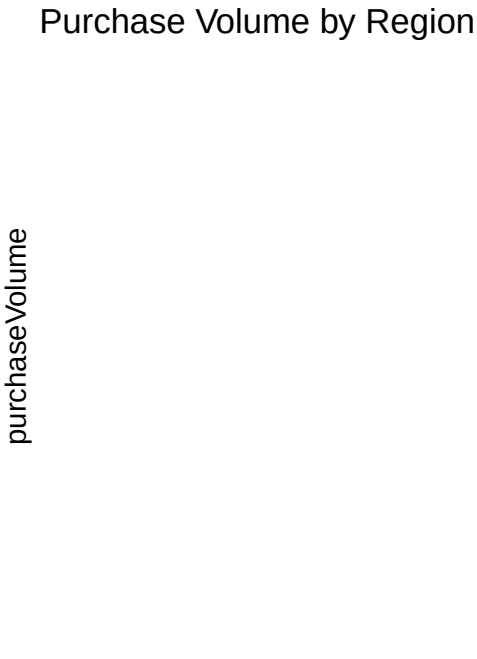

Figure 13: Regional Variations in Purchase Volume

the South exhibits a slightly higher mean and a wider interquartile range, indicating greater variability in purchase volume compared to other regions - which might be due to the lower cpi values in the South as revealed in Figure 11 above.

Figure 14 shows notable fluctuations from 2012 to 2018, suggesting periodic variability in consumer purchasing behavior. Peaks are observed around specific intervals, likely reflecting seasonal trends (peaks usually occur during the year-end i.e. the holiday season) or external economic factors influencing consumer demand. Over time, the periodic nature of these fluctuations appears consistent, with no clear upward or downward trend, indicating that the purchase volume stabilizes around recurring patterns.

C Model details

C.1 AR(1) Model

C.1.1 Overview

An Autoregressive (AR) Model of order 1, denoted as AR(1), is a type of time series model where the current value of the time series is linearly dependent on its immediate past value and a random noise term. Mathematically, an AR(1) model can be expressed as:

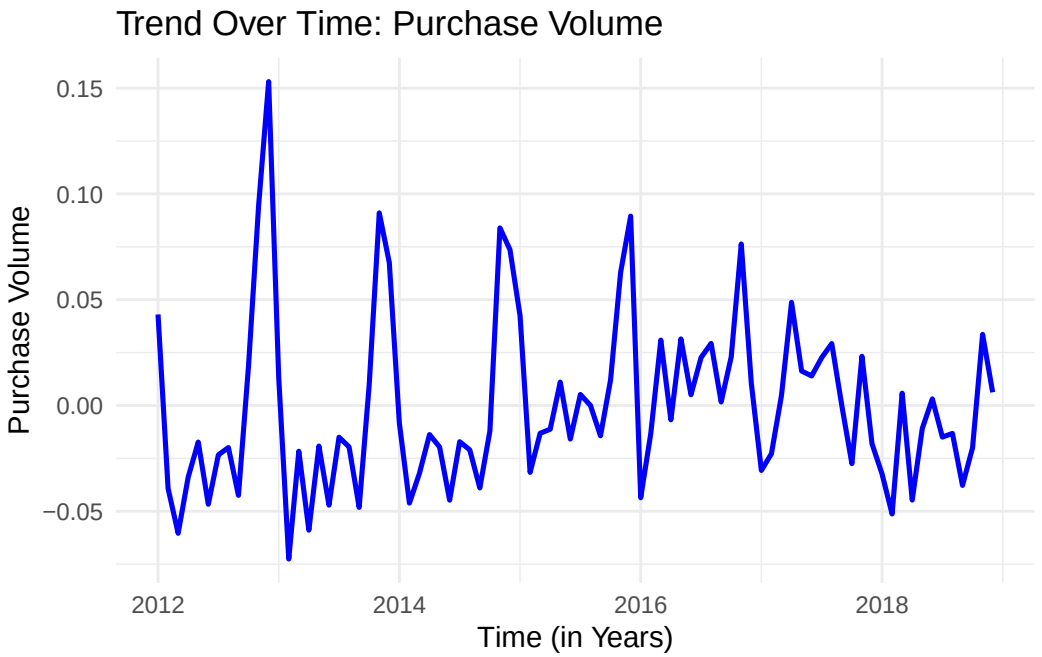

Figure 14: Temporal Variations in the Purchase Volume of Groceries

$$y_t = \phi.y_{t-1} + \epsilon_t$$

Where:

- y_t: Value of the variable at time t.
- ϕ: The AR(1) coefficient (a scalar between -1 and 1) that captures the strength and direction of temporal autocorrelation.
- y_{t-1}: Value of the variable at time $t-1$ (previous time step).
- ϵ_t: Random error or noise term at time t assumed to be independently and identically distributed ($\epsilon_{ijt} \sim N(0, \sigma^2)$).

The AR(1) model is suitable for time series that exhibit correlation between consecutive observations, making it useful for capturing short-term temporal dependencies.

Role of the ϕ Term:

The ϕ term captures temporal autocorrelation, which is the relationship between observations at consecutive time steps. Here's what it represents:

1. Positive Temporal Correlation ($\phi > 0$): If $\phi > 0$, then the current value y_t tends to be similar to y_{t-1}
2. Negative Temporal Correlation ($\phi < 0$): If $\phi < 0$, then the current value y_t tends to move in the opposite direction of y_{t-1}
3. No Temporal Correlation ($\phi = 0$): If $\phi = 0$, then there is no dependence between y_t and y_{t-1}

C.1.2 Integrating AR(1) into BHM

In a Bayesian Hierarchical Model (BHM), the ARIMA components (Autoregressive Integrated Moving Average) are implemented as part of the time-series modeling within the hierarchical framework. Here's how the ARIMA(1,0,0) model (which is just AR(1)) might be integrated:

C.2 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

C.3 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

Please include an Appendix where you focus on an aspect of surveys, sampling or observational data, related to your paper. This should be an in-depth exploration, akin to the "idealized methodology/survey/pollster methodology" sections of Paper 2. Some aspect of this is likely covered in the Measurement sub-section of your Data section, but this Appendix would be much more detailed, and might include aspects like simulation, links to the literature, explorations and comparisons, among other aspects.

References

Bürkner, Paul-Christian. 2023. *Brms: Bayesian Regression Models Using 'Stan'*. https://cran.r-project.org/package=brms.

Grolemund, Garrett, and Hadley Wickham. 2023. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Dewey Dunnington, Ian Cook, Nic Crane, Antoine Pitrou, and Uwe Korn. 2023. *Arrow: Integration to 'Apache Arrow'*. https://CRAN.R-project.org/package=arrow.

Sweitzer, Megan, Anne Byrne, Elina T. Page, Andrea Carlson, Linda Kantor, Mary K. Muth, Shawn A. Karns, and Chen Zhen. 2024. "Development of the Food-at-Home Monthly Area Prices Data." TB-1965. U.S. Department of Agriculture, Economic Research Service. https://www.ers.usda.gov/publications/pub-details/?pubid=108813.

U. S. Department of Agriculture, Economic Research Service. 2024. "Food-at-Home Monthly Area Prices [Data Product]." U.S. Department of Agriculture.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org/.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Lionel Henry, and Evan Friedland. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.