LUND UNIVERSITY
School of Economics and Management

Master's Programme in Data Analytics and Business Economics

A Comparative Evaluation of Forecasting Techniques for Public Sector Food Prices: From Statistical to Modern Methods

by

Måns Wirfelt & Vilhelm Björklund

Abstract: This paper examines the performance of traditional statistical models (SARIMA & SARIMAX) and modern machine learning models (XGBoost) for forecasting Swedish public sector food prices organized by a hierarchical index structure. The depth of the analysis is made possible by collaborating with Matilda Foodtech, who gave access to anonymized food procurement data from Swedish public organizations. A dataset spanning from 2015 to 2023, containing over a hundred food categories, has been enriched with macroeconomic indicators to serve as predictors. The results indicate that hierarchical models and modern advanced models can outperform statistical models, but that ensembles of the three yield the best forecasting performance. The study highlights the importance of adapting methods and predictors to the time and place to enhance forecasting precision. These insights are crucial for public sector stakeholders to improve budget planning and optimize procurement strategies. They can also guide future researchers in understanding underlying dynamics of food price inflation based on categorical properties of foods that constitute food price indices.

DABN01
Master's Thesis (15 credits ECTS)
May 2024
Supervisor: Joakim Westerlund
Word Count: 14768

Acknowledgements

We would like to extend our gratitude to Johnny Zackrisson and Johan Bjerning at Matilda Foodtech for making this thesis possible. Your domain knowledge, guidance on Linux, and unwavering support, regardless of the hour, have been invaluable. We would also like to thank Joakim Westerlund for supervising us and welcoming us into your office whenever questions and difficulties arose. We appreciate your guidance during our visits, which often turned what we thought were quick questions into 1.5 hours of engaging discussions.

Contents

1 Introduction **4**
- 1.1 The Relevance of Studying Food Prices 4
- 1.2 Organic and Conventional Food 5
- 1.3 Contemporary Machine Learning Techniques 5
- 1.4 This paper 6

2 Previous Research **7**
- 2.1 Food Prices 7
- 2.2 Methodology 8

3 Data **10**
- 3.1 Data Description 10
- 3.2 Data Operationalization 13
- 3.3 Data for Modelling 14

4 Methodology **15**
- 4.1 General Methodology 15
- 4.2 Statistical Approach 16
- 4.3 Hierarchical Approach 18
- 4.4 XGBoost Approach 18
- 4.5 Ensemble Methods 19

5 Empirical Analysis **20**
- 5.1 Exploratory Analysis 20
- 5.2 Rolling Average 25
- 5.3 Statistical Approach 25
- 5.4 Hierarchical Approach 27
- 5.5 XGBoost 31
- 5.6 Ensemble Methods 34
- 5.7 Model Comparison 36
- 5.8 Organic & Conventional Food 37

6 Discussion **40**
- 6.1 Insights and Challanges 40
- 6.2 Extensions 43

7 Future Research **46**

8 Conclusions **47**

Bibliography **47**

A **52**

1

Introduction

1.1 The Relevance of Studying Food Prices

After almost a decade of low inflation rates and prospering economics globally, the last four years dramatically challenged this status quo. The covid-19 pandemic, war in Europe as well as the resulting economic downturn has resulted in high inflation rates over the last few years. In the summer of 2022, the global inflation rate reached its highest since the mid 1990's (Ha et al., 2023). According to Ha et al. (2023), in their policy research paper for the World Bank, they identify oil prices as the main driving force of the inflation rate change in the last few years. However, for most people, one of the main areas where inflation is frequently present is when grocery shopping. According to Statistics Sweden (2023), food prices are one of the most significant parts of the Swedish consumer price index which is the measure of inflation domestically. Food is a limited but renewable resource that is essential, we cannot choose to not consume it. As a result there are economic incentives in forecasting the price of groceries, so that it is possible to buy when prices are low. Additionally, by accurately forecasting prices it can make good food planning more relevant and profitable, and thereby less food will be wasted as well.

This paper will be written in collaboration with Matilda Foodtech, a SaaS company focused on food procurement, meal planning and food safety within the public sector. A large portion of Swedish municipalities and other public organs are customers of Matilda Foodtech. This allows us to gain access to a large data set containing information of more than a hundred food categories over time.

By accurately forecasting the price of food in general, it will be relevant both for Matilda Foodtech and their customers as an indicator as to whether prices will increase and at what rate. Additionally, by interpreting the cost drivers of different food items it is possible to make more educated decisions on what sort of food items to prioritize in order to keep costs down. One such decision is to decide on organic or conventional food items. If it is determined that one category is more affected by a variable such as oil price (Baek and Koo, 2010), it would be possible to use that variable as an indicator of how the category's price will be affected in the future. Moreover, by learning about the seasonality and how different models capture it, it is possible to make more educated and cost efficient food planning. Similar intuition for more efficient food planning is found in papers forecasting food demand (Fattah et al., 2018).

1.2 Organic and Conventional Food

The European Corporate Sustainability Reporting Direction (CSRD) that was implemented back in early 2023 forces companies to report non-financial metrics in their annual reports. This in turn means that from this fiscal year, more companies are required to include ESG reporting in order to provide stakeholders with information regarding their environmental work (European Union, 2023). The increased focus on sustainability reporting in the European Union is regulated in the Commission's delegated regulation (2023/2772). The regulation mentions one of the EU's targets for Biodiversity, that agricultural farming should be increasingly organic and that the uptake of agro-ecological practices is significantly increased. This means that organic foods are of interest for companies conducting business within the European Union and that it is an important part of the now more comprehensive ESG reporting mandated by the union.

Matilda Foodtech's customers are mainly operating within the public domain, where the stakeholders are the taxpayers. As in accounting, the importance of reporting within the public sector is profound. Taxpayers money should be spent wisely and stakeholders gain insights through these reports and execute their control by voting. From this perspective, it is relevant to consider organic foods in this paper, since stakeholders want to maximize the utility gained from their investments, both from an environmental and financial perspectives. By gaining deeper understanding of how and why the prices of organic and conventional food items change, it is possible to both maximize the amount of organic foods and not necessarily spend more money doing so.

Lastly, there are indications that organic foods have not necessarily increased in price at the same rate as conventional foods during the last years of high inflation (KRAV, 2022). The article highlights that organic foods are less dependent on pesticides and fertilizer that are commonly imported. These items have gone up in price due to the war in Ukraine and the increased energy prices, this is because of transportation as well as the dependence of fossil fuels in fertilizer manufacturing (KRAV, 2022). This highlights that some foods might be less susceptible to the current driving forces behind inflation and why it is of interest to determine what causes what.

1.3 Contemporary Machine Learning Techniques

Swedish inflation forecasts presented by Swedish public organizations are often made using statistical models. As an example, the central bank of Sweden commonly uses a Bayesian Vector Auto-Regressive (BVAR) model to forecast consumer price index with fixed interest rate (CPIF) (Sveriges Riksbank, 2023). CPIF is the official measure for inflation in Sweden (Sveriges Riksbank, n.d.). Statistical models like these are frequently favored for their ability to elucidate the process behind forecasting, shedding light on both the methodology used and the variables influencing the forecast. In a study of the forecasting errors of inflation between 2013 and 2022, The Swedish Central bank showed lower precision in their forecasts compared to other financial institutes (Sveriges Riksbank, 2023).

There was a period during which it was believed that complex algorithms would not substantially enhance forecasting performance when contrasted with simpler statistical models. One of the main conclusions from the M3 forecasting competition in 1999 was that complex methods does not necessarily perform better than simpler ones (Koning et al., 2005). During the last decades, both academia and forecasting competitions have shown improved forecasting performance with machine learning approaches compared to statistical models. Although deep learning emerged victorious in the M4 competition in 2018, the M5 competition in 2020 revealed that tree-based gradient boosting methods, particularly LightGBM and XGBoost, wield significant forecasting power, with many winning teams incorporating them into their solutions (Januschowski et al., 2022).

1.4 This paper

Collaborating with Matilda Foodtech grants us access to invaluable data, setting our research apart from previous studies on food price forecasting. While other researchers primarily focus on predicting food price inflation, our access to high-resolution data allows for a more nuanced analysis. This granularity enables the examination and evaluation of both aggregate and disaggregate series, offering a deeper understanding of market dynamics. To illustrate the extent of the data resolution, consider the raw information it contains as the following. Each data point encompasses details regarding specific products, including brand, container size, and unique properties. For instance, a single data entry might pertain to a particular brand of milk, specifying its container size and organic certification. These individual attributes can then be aggregated to form broader categories such as "milk" or "dairy products." Furthermore, leveraging product properties allows for the differentiation between frozen, fresh, or organic goods. This becomes particularly pertinent when integrating macroeconomic variables into our models, allowing us to identify macroeconomic factors that serve as deterministic cost drivers for specific goods or categories. By considering such detailed information, our research aims to enhance the accuracy and robustness of food price forecasting models in a real-world setting, thereby offering valuable insights for both academia and industry stakeholders.

Given the shifting beliefs in forecasting methods and the unique data, this paper aims to assess models capable of enhancing forecasting accuracy beyond traditional statistical methods, leveraging variables previously identified in research as influential for food price inflation. The evaluated methods include statistical approaches, hierarchical extensions of these statistical methods, modern gradient boosting tree-based techniques, and ensembles combining all three. This multidisciplinary approach is in this paper dedicated to find the best suited forecasting model. Therefore, the emphasis will be on forecasting performance rather than inference.

Additionally, if valuable insights can be gained on which variables influence food prices, it might lead to meaningful results in terms of efficient food planning. This study will specifically investigate whether the best general food forecasting model can be effectively generalized to suit organic and conventional foods.

2

Previous Research

2.1 Food Prices

Food prices have been forecasted before. However, most papers on the topic such as the report from the Swedish National Institute of Economic Research (NIER, 2023) and Joutz (1997) utilize aggregated food price data from consumer price indices. NIER (2023) uses the Swedish consumer price index (CPI) to analyze the general price and cost-development between 2019 and 2023. They find that groceries as a category, experienced an increase in price greater than what their base model predicted on the prices in 2019 with domestic values of the predictors such as input goods and energy prices. Additionally, they modeled prices with constant levels of energy prices and concluded that energy prices only marginally contributed to the price increase in general and claim that this is due to the decline in energy prices since 2022 whilst the consumer price index still increased. NIER's (2023) paper is, however, not specifically focused on food prices but the price level in general where food constitutes a fraction of it.

Joutz (1997) summarized three forecasters' results and their primary drivers for explaining the food price inflation during the 1990s in the US. However, these different forecasting methods provide widely different predictions and Joutz (1997) claims that it is due to different assumptions on the macroeconomic variables introduced. The three forecasters had similar intuition on which drivers caused food price inflation and they determined that commodity prices at the farmgate (prices of input goods), labor costs and energy prices of the wholesalers and food processors provided meaningful results for food price forecasting.

Gilbert (2010) researched the food price inflation during the late 2000s financial crisis where the price of food commodities more than doubled from 2005 and the mid of 2008. However, he claims that the causation of this price increase is controversial and some common explanations are drought in Australia resulting in a decrease in wheat supply of 4% globally. Another factor according to Gilbert (2010) is poor grain harvest in Europe in 2007 which was offset by good harvests in Argentina, Kazakhstan and Russia. However Headey and Fan (2008) concurred that during the same time period these weather shocks did not significantly impact the prices, instead they explain this by the increase of input prices and especially fertilizer which had a lagged effect on food prices. Gilbert (2010) highlights that the rise of food prices during this period coincided with a general rise in commodity prices led by energy and metals. However, agricultural raw materials were stable during this period and followed the developments of crude oil prices which fell drastically post the Lehman Brothers crisis months where the subsequent effect was decreasing food prices as well. One theory that Gilbert (2010) put out is that oil prices might affect food prices in two ways, through increased oil prices influence agricultural production costs in nitrogen based fertilizer and transportation. Baffes (2007) determines that this cost pass-through is limited in agriculture due to it not being very energy intensive. He estimates the pass-through effect of an oil price shock on food prices to be approximately 17%. Additionally, the pass-through effect of oil on fertilizer was estimated somewhat higher at 33% and the effect on food to be 18% (Baffes, 2007).

2.2 Methodology

Much of the research presented above is mainly focused on explaining the increase in food prices and determining the drivers of increased cost. However, when focusing on the research methodology regarding forecasts of food price inflation there are different approaches used. Joutz (1997) compares forecasting performance from two different American institutes and one private consultancy firm with the baseline autoregressive integrated moving average (ARIMA) model. One institute, the Food and Agricultural Policy Research Institute at the University of Missouri (FAPRI) has created their own econometric model that is depicted as a large scale structural econometric model that combines consumer price index for food as well as economic factors, agricultural science and biological processes (Joutz, 1997). The other institute and the consulting firm also have vague descriptions of their specific model, although these models appear more complex they do not significantly outperform the baseline ARIMA model.

Toledo and Duncan's (2024) article focus on food price inflation forecasting during global crises, especially the financial crisis and the Covid-19 pandemic. They test a multitude of different model specifications and evaluate them in times of crises and pre-crisis periods. Some of the models they tried were dynamic model averaging, driftless random walk, autoregressive (AR), time-varying parameter model where each parameter evolves as a unit root process, kitchen sink approach which contains prespecified lags of the dependent independent variables, and lastly a dynamic model selection approach that uses the model with the the highest posterior probability in each forecast. These models were then also compared to baseline models such as Atkeson-Ohanian Random Walk, AR model with prespecified lag length as well as an AR model with Bayesian information criteria selected lags. They find that the dynamic model averaging performs slightly better than the rest in most cases, especially during times of crisis. However, when considering the evaluation metrics it is not a significant leap from the traditional time series models like the autoregressive ones. There are many different iterations of dynamic models that utilize multiple predictive models and combine them in different ways. The dynamic model averaging setup in Toledo and Duncan's (2024) paper is set up to average the results from different predictions. Ensemble methods like this can according to Gastinger et al. (2021) improve the forecasting accuracy, however, one additional parameter to consider and tune is the weights of each model, meaning how much it should influence the ensemble model.

Another approach used to boost performance of traditional statistical modeling methods is to consider a hierarchical modeling approach. This is when considering the time series data at a more disaggregate level. In the context of a consumer price index it implies that it is of interest to model the time series that constitutes the index individually, so food prices and prices of services etc. should be modeled separately. This can be performed using a 'top down' or 'bottom up' method where the difference is the starting point (Hyndman et al., 2011). In a bottom up model the disaggregate series are modeled and forecasted before aggregating to get the forecast at the aggregate level, top down is the opposite direction used to gain a deeper understanding of the disaggregate components. The aggregation of disaggregate series is an important factor where index weights can be used but one must be mindful of aggregation bias. Schwarzkopf et al. (1988) performed a comparison between forecasts made at the aggregate level versus using a bottom up approach. The authors state that the major downside of making individual forecasts for each item, then adding them together with a percentage distribution, is that it requires that there are no missing values in individual series and that it takes more time to compute the forecasts. They also argue for the upsides of the bottom-up approach, which can detect differences between items and is usually equally robust to a forecast modeled directly at the aggregate level even though individual series might contain more outliers. These upsides are further highlighted when NIER (2023) evaluated their PRIOR model with disaggregate level data within agriculture and groceries and were able to provide a more in-depth explanation of cost pass-through for different categories. There are various ways to improve performance of hierarchical methods. Hyndman et al. (2011) presents a way to combine the bottom up hierarchical forecasts through regression. They show good accuracy and lower variance with this method than additive bottom up specifications.

Recent literature points towards a predictive superiority of machine learning based approaches compared to traditional statistical ones. Xu and Zhang (2023) forecasts wholesale food price index in China with an autoregressive neural network utilizing a two-layer feedforward structure. Additionally, they benchmark its performance against random walk, AR, AR general autoregressive conditional heteroskedasticity (AR-GARCH), support vector regression (SVR), regression tree (RT) and a long short-term memory recurrent neural network (LSTM-RNN). All models significantly outperforms the AR and AR-GARCH models in forecasting accuracy. This highlights a paradigm shift in time series forecasting which is supported by Januschowski et al. (2022). They discuss the superiority of gradient boosted tree based methods in M4 and M5 time series forecasting competitions. In recent years the leaderboards have been topped by gradient boosted decision tree algorithms (GBDTs) such as XGBoost. Neural networks are also represented among the leaderboards in accuracy competitions, however in second and third place. These neural networks are based primarily on DeepAR and NBEATS frameworks developed by Amazon and Facebook respectively (Januschowski et al., 2022).

3

Data

3.1 Data Description

Matilda Foodtech supplied purchasing data for five main categories: frozen goods, refrigerated goods, colonial goods, wine and spirits and nutritional goods. Within these categories there are subcategories such as dairy products, cheese and vegetables to name a few in the refrigerated goods category. For each unique combination of main category and subcategory there are monthly observations for the columns showcased in Table 3.1. This was the raw data structure as received from Matilda Foodtech. The time period for this data ranges from January 2013 until December 2023, resulting in 17306 observations.

Before receiving the data, some pre-processing steps were undertaken. Aggregation to subcategory level is the main procedure that we did not perform ourselves, it is therefore hard to control for bias at this level. For the smaller categories in particular, there is reason to believe that different products are represented in a subcategory between different periods, which might lead to less accuracy in our forecasts of these categories. Furthermore, while most public organizations feed their information into Matilda Foodtechs system on a monthly basis, some are doing it on a quarterly basis. Together with Matilda Foodtech, an active decision was made to exclude the quarterly imported data for two main reasons. Initially, fewer organizations are reporting quarterly to Matilda Foodtech, and they anticipate that this trend will continue, with very few organizations expected to do so in the coming years. Secondly, the quarterly data exerted significant influence on the behavior of our dependent variable, particularly noticeable in earlier years, resulting in distinct dips occurring every three months. These quarterly dips were diminishing over the years since more and more organizations started to report monthly. This gave many models a hard time to correctly estimate the effects of seasonality. With quarterly imported data removed, the behavior of the time series became more logical and in line with the seasonality explained by Matilda Foodtech's clients.

Table 3.1: Variable Descriptions for Matilda Foodtech Data

Variable	**Data type**	**Description**
period	timestamp	month formatted as YYYY-MM
main_category	string	the five types described above
subcategory	string	still aggregated but more detail. for example "meat", "dairy product" and "fish"
value	float	total spendings during the month
volume	float	total number of items during the month
value_organic	float	spendings on organic items during the month
volume_organic	float	number of organic items during the month
number_of_clients	integer	number of organizations that have bought items in this category during the month

Previous research highlighted the importance of macroeconomic variables in forecasting performance, feature importance and inference. These macroeconomic time series were fetched online through API's to ensure a continuous flow of contemporaneous data into the models. Global macroeconomic variables are collected from the US Federal Reserve Economic Data (FRED, 2024) API, the data itself have different sources such as the International Monetary Fund (IMF). Variables regarding Swedish macroeconomics were collected from the Swedish central bank API (Sveriges Riksbank, 2024). Lastly, the variable for the Swedish electricity price was collected from Statistics Sweden (2024) API as well. Table 3.2 lists all the macroeconomic variables, their shortened name, API and source of information.

By incorporating the use of API's, the variables are constantly up to date for as long as the API is updated online. This ensures that data is available for a great span of years and well into the future as well.

Table 3.2: Summary of macroeconomic variables, their source, API, and years covered.

Variable Code	**Variable Name**	**API**	**Source**	**Years**
POILBREUSD	Global price of Brent Crude Oil (USD)	FRED	International Monetary Fund (IMF)	1990-01-01 2024-03-01
PCU325311325311	PPI: Nitrogenous Fertilizer	FRED	US Bureau of Labor Statistics	1975-12-01 2024-03-01
PNRGINDEXM	Global Price Index - Energy	FRED	International Monetary Fund (IMF)	1992-01-01 2024-03-01
PFOODINDEXM	Global Price Index - Food	FRED	International Monetary Fund (IMF)	1992-01-01 2024-03-01
SWECPICOR-MINMEI	CPI Sweden: All Items Except Food and Energy	FRED	Organization for Economic Co-operation and Development (OECD)	1970-01-01 2023-11-01
CPGREN01-SEM657N	CPI Sweden: Energy (Fuel, Electricity and Gasoline)	FRED	Organization for Economic Co-operation and Development (OECD)	1970-02-01 2023-11-01
CSESFT02-SEM460S	Consumer Opinion Survey: Economic Future for Sweden	FRED	Organization for Economic Co-operation and Development (OECD)	1995-10-01 2024-03-01
WP5075303	PPI: Diesel Fuel	FRED	US Bureau of Labor Statistics	1986-05-01 2024-03-01
CP0111EU27-2020M086NEST	Harmonized CPI for EU: Bread and Cereals	Eurostat	Eurostat	2000-12-01 2024-03-01
PWHEAMTUSDM	Global Price of Wheat (USD)	FRED	International Monetary Fund (IMF)	1990-01-01 2024-03-01
ENRGY0EU272020-M086NEST	Harmonized CPI: Energy for European Union	Eurostat	Eurostat	2000-12-01 2024-03-01
CP0450EU27-2020M086NEST	Harmonized CPI: Electricity, Gas and Other Fuels for European Union	Eurostat	Eurostat	2000-12-01 2024-03-01
SSDManad- Elhandelpris	Electricity Prices in Sweden	Statistics Sweden	Statistics Sweden Energy Agency	2013-04-01 2024-03-01
SEKEURPMI	SEK-EUR Exchange Rate	Swedish Riksbank	Swedish Riksbank	1993-01-04 2024-04-01
SEKUSDPMI	SEK-USD Exchange Rate	Swedish Riksbank	Swedish Riksbank	1993-01-04 2024-04-01
SEKKIX92	KIX Index 92	Swedish Riksbank	Swedish Riksbank	1992-11-18 2024-04-01
SECBREPOEFF	Swedish Policy Rate (Styrränta)	Swedish Riksbank	Swedish Riksbank	1994-06-01 2024-04-01
SETB1MBENCHC	Swedish Treasury Bill Maturity 1-Month (Statsskuldsväxel)	Swedish Riksbank	Swedish Riksbank	1983-01-03 2024-04-01
EMGVB5Y	Government Bond EUR 5Y (Statsobligation EUR 5 år)	Swedish Riksbank	Swedish Riksbank	1990-01-04 2024-04-01

3.2 Data Operationalization

Unique combinations of main category and subcategory will from here on out be referred to as category. Initially, there were 163 distinct categories. However, upon visualizing the data and closely examining the information and patterns, it became clear that before 2015, there was a lack of data, leading to anomalous fluctuations in the value variables at both the category and aggregated levels. For that reason, 2013-2014 was filtered out. Following the filtering process, all categories that still contained missing values were eliminated. Categories that were either non-food or alcoholic were also removed to make sure the data was in line with the research question. 103 categories were left for the final dataset and are shown in Appendix A.1, excluding roughly a third of all categories. Regarding the total value, eliminating these categories did not have a significant impact, as the removed categories typically involved purchases in small volumes. The total value corresponding to removed categories in the period of interest between January 2015 and December 2023 is 0.8% of the total value of all categories in the original dataset in the same period. Thus the final, formatted dataset used for modeling thereby keeps 99.2% of the total value in the data handed to us by Matilda Foodtech.

To construct the aggregated price index for modeling, weights were computed for each category. These weights were derived from the category value divided by the total value over the years 2019-2021. The reason for excluding 2022-2023 (out-of-sample period discussed in Chapter 4) is to simulate a real forecasting scenario and not create biased index weights. The 36-month time frame was chosen in consultation with Matilda Foodtech to capture the general shopping habits of their clients over recent years. Furthermore, by selecting a time period that is too short, one might risk to select a biased index due to their cyclical and seasonal relevance of certain goods to the public sector. However, an overly long period could risk basing the index on the importance of categories that are no longer popular, while underestimating the significance of goods that have become more popular in recent years. The 36 month weights are thereby an effort to balance these two shortcomings.

For each category an average price variable was created. These variables are simply value divided by volume for the corresponding category and time period. The average price columns was then multiplied by the corresponding, previously calculated weights, which gave rise to yet another 103 variables, which when summed constitutes the aggregated price index. Given the trending behavior of prices, first differences were taken for each category. First differences were also applied to the price index to ensure stationarity and the ADF-test p-value for the differenced series is 0.0062 (2015-2021). It is this differenced aggregated price index that we aim to forecast in this paper.

Two additional datasets were generated, one for organic goods and one for conventional goods, employing the same methodology outlined previously. In the case of organic goods, values and volumes for organic goods were used. For conventional goods, the organic value and volume was instead removed resulting in only non-organic i.e. conventional goods being left. The organic dataset contains fewer categories due to the presence of missing values in the time period of interest for organic goods across many categories.

Some transformations of the macroeconomic variables were necessary before utilizing them in the models. The Swedish electricity market is divided into four distinct zones. These zones exhibit divergent price patterns attributable to regional disparities in electricity supply and demand. Notably, Zone 1, situated in the most northern region characterized by an electricity surplus, experiences lower prices compared to Zone 4 in the southern part of Sweden, marked by an electricity deficit (Energimarknadsbyrån, 2024). The electricity data from Statistics Sweden (2024) contains monthly average prices for non-domestic users in these four regions. The average Swedish electricity price was calculated without regards to weighing the zones differently based on energy consumption within the food industry. This newly transformed variable is from this point called 'avg_electricity_price_SWE'.

Moreover, the electricity dataset sourced from Statistics Sweden (2024) exhibited missing values for March 2021. To address this issue, a linear imputation method was employed, whereby the missing value was estimated by the arithmetic mean of the two adjacent values. This ensured that all macroeconomic variables included were consecutive time series data. The macroeconomic series were then differenced to ensure stationarity of the data. Figure 5.2 and 5.3 visualize both the original and first differenced macroeconomic data. Augmented Dickey-Fuller tests were performed on all time series between 2015-2021 (see Appendix A.2). Before differencing, only 3 variables were stationary, after differencing however, all variables were stationary at the 5% level. Starting from 2015 allowed for consistency across all variables, while ending in 2021 simulated the scenario in which these models would be trained during our selected out-of-sample period (2022-2023).

3.3 Data for Modelling

The macroeconomic dataset and three versions of the Matilda Foodtech data was then combined into three data sets. One for all Matilda Foodtech data, one only consisting of organic foods and one only consisting of conventional foods.

The three final datasets used for modelling each consist of 108 observations (time periods) and 410 variables. The difference between the datasets are the price index variables, where value and volume used to create these variables have been filtered based on underlying items in the categories are organic or not. The variables include 19 contemporary macroeconomic variables, 171 lags of macroeconomic variables, 103 category price indices, 103 first differences of category-specific indices, 12 dummy variables for months, 1 aggregated price index, and 1 first difference of the aggregated price index.

With the surging food prices that followed the recent crises, it would be interesting to incorporate some sort of proxy variable for times of crisis. However, due to the subjectivity in creating a feature like this on our own, we opted to make this into a delimitation and not include such feature in the data. Instead, the phenomenon generated by the war in Ukraine on energy prices and deficits in Ukrainian cereal production is covered by other variables. Other factors like extreme weather, war and politics are not specifically included either which is also common limitation in studies such as Ribeiro and Dos Santos Coelho's (2020). However, policy changes, subsidies and environmental factors most certainly affect the food industry and the resulting food prices.

4

Methodology

In Section 4.1 the common choices, used for all models is presented. The choices made here are to ensure comparability of results. In Section 4.2 the specifications and choices specifically applicable to the statistical models are presented and argued for. In Section 4.3 the hierarchical bottom-up approach is explained. In Section 4.4 the specifications and choices for modern machine learning approaches are presented and argued for. In the final Section, 4.5, it is presented how the best models are combined using what is commonly referred to as ensemble methods or consensus methods.

4.1 General Methodology

All models used in this study make one-step-ahead deterministic forecasts of the first differences of the monthly price index. For comparability, all models in this study use a rolling window for forecasting one-step-ahead for the 48 periods between January 2020 and December 2023. The window sizes differ between models but were all chosen through cross-validation on 2020-2021, where 12, 24, 36, 48 and 60 period rolling windows were tested. Given the available data, with the restriction of using full-year periods, 60 was the largest possible window to use. For the Rolling Average, Seasonal ARIMA (SARIMA) and SARIMA with exogenous variables (SARIMAX) models (Section 4.2) a 60 period rolling window was selected, while for XGBoost (Section 4.4) a 36 period rolling window was selected. For the hierarchical SARIMA and SARIMAX (Section 4.3), no cross-validation was performed due to extensive computational time, instead 60 was used based on the cross-validation from the non-hierarchical versions.

The sample period which will be used for modelling and evaluation is January 2015 to December 2023. January 2015 was chosen as the first observation in consensus with Matilda Foodtech by examining the data. Their business only consisted of a few major clients in the early years, but from 2015 onwards the data contains observations from a mix of clients more representative of today's clients. The product mix within categories and weights of food categories in the data is also more stable from 2015 onwards. Thereby making this a rational decision since the outlook for this paper is in the future. December 2023 was also chosen as the last observation together with Matilda Foodtech. Given that some of the macroeconomic variables are not available until April of 2014 in the API's, nine lags of these variables are used in the modelling stage. This data availability limitation is thereby used as a natural delimitation for lag lengths. However, nine lags are also a reasonable maximum lag length since groceries mainly constitute fresh goods with a limited shelf life. There is a trade-off between reducing the training data size and adding more lags of the variables.

In order to evaluate the generalizability of the models, the same in-sample and out-of-sample periods were chosen for feature selection and hyperparameter tuning for all algorithms. The year 2022 is considered to be particularly interesting since there was a sudden increase in inflation due to the war in Ukraine (Emediegwu, n.d.). The objective is to discover a model capable of being fine-tuned with pre-war data while maintaining satisfactory forecasting accuracy for the years 2022 and 2023. The years 2019-2021 i.e. 36 observations are used for hyperparameter tuning. Rolling one-step-ahead forecasts are then made between 2020 and 2023, where 2020-2021 measure "in-sample performance" which can be thought of as a measure of fitness to the sample, and 2022-2023 measure "out-of-sample performance" which instead can be viewed as the models' ability to generalize for unseen data. Results will also be presented with separate MSFEs for each year. Additionally, line graphs for each models of actual values and forecasted values, will be used to analyze which periods that are especially difficult to forecast accurately.

The initial benchmark model is a simple rolling average of 60 periods. In other words, the average of the 60 last months is the forecast for the upcoming month. Although, as the modeling phase progresses the next model, the SARIMA is the more reasonable benchmark based on previous research as discussed in section 4.2. The evaluation metric used in this study is mean squared forecasting error (MSFE). This metric is widely used in similar studies on forecasting of price differences in food such as Ribeiro and Dos Santos Coelho (2020) and Yang et al. (2017), who also compare different models to each other.

After the rolling average benchmark was established, complexity of models was added incrementally. Throughout the rest of this chapter we will present how these increasingly modern models were implemented, starting off with how to capture trends and auto-regressive aspects of the series.

4.2 Statistical Approach

The target variable as well as most of the category-level time series of the target have clear monthly seasonality, some of which is explained by the yearly cycle of Matilda Foodtech's clients buying different goods in for example December compared to August. Some is, however, explained by actual price differences of the same foods between different months due to supply and demand dynamics. As an example vegetables tend to be cheaper during summer than during winter. Considering that for instance electricity prices also have a clear seasonality, adding seasonality components can help prevent spuriousness to arise when introducing macroeconomic variables (Enders, 2015).

The first model used to try to improve on the rolling average benchmark is a SARIMA. AR and ARIMA models are often used as a benchmark in similar studies (Ribeiro and Dos Santos Coelho, 2020). In this study the seasonality effects are added to have a fair benchmark, since an ARIMA would perform poorly on the seasonal data. The hyperparameters were tuned on 2017-2021, with the grid specification as in table 4.1.

Table 4.1: Parameter Values for Component Orders

Components	**Values tested**
AR order (p)	0, 1, 2
Integration order (d)	0, 1
MA order (q)	0, 1, 2
AR order of seasonality (P)	0, 1, 2
Integration order of seasonality (D)	0, 1
MA component of seasonality (Q)	0, 1, 2
Seasonality components (s)	0, 12

The hyperparameters were then selected by minimizing the Bayesian information criteria (BIC). Using the selected model specification, 48 models were trained using a rolling window and used to make one-step-ahead forecasts for January 2020 up until December 2023. Note that 0 is included in the grid for seasonality components, but 12 was selected (see Chapter 5, Table 5.2), which further validates the need for including monthly seasonality effects in the benchmark. Additionally, the integration order was only meaningful for a few of the category-level time series where SARIMA and SARIMAX were used hierarchically as discussed in 4.3 that needed second differences to ensure stationarity.

Considering the vast amount of lagged macroeconomic variables, a two-step feature selection procedure was used for the SARIMAX. The first step was using a simplified version of correlation analysis. Correlation is a commonly used way to perform feature selection in Machine Learning. Hall's (1999) paper explains that a good feature set should consist of variables that have high correlation with the target variable, while not being correlated with each other within the feature set. The simplified approach used in this article was to choose a correlation threshold used to keep or dismiss certain lags of certain variables. If the absolute value of the correlation between the feature and the target was greater than the threshold, the feature was kept for step two.

The second step for feature selection used was least absolute shrinkage and selection operator (LASSO). When the features selected through correlation analysis alone were used to perform out-of-sample forecasts, the model showed signs of severe overfitting. Muthukrishnan and Rohini (2016), explains how models fitted on too many variables often are hard to interpret and do not generalize well. He raises LASSO as a proven alternative for feature selection and argues that it can make models generalize better on new data. LASSO was therefore used to further filter the features with high correlation. Again, 2019-2021 was the period used to select features with LASSO. Using LASSO alone did not keep any variables in the feature set, and therefore was not considered interesting since it would be equivalent to the SARIMA model explained above.

The simplified approach used in this article was to perform in-sample cross-validation to choose a correlation threshold used to keep or dismiss certain lags of certain variables. [0.15, 0.2, 0.25, . . . , 0.4] was tested as thresholds and 0.3 gave the lowest in-sample MSFE. The same parameter grid for the SARIMA components (see table 4.1) were used for the SARIMAX and parameters were now re-selected using BIC, given the variables selected using this approach.

4.3 Hierarchical Approach

The SARIMA and SARIMAX procedures explained in 4.2 were applied directly to the aggregated first differences of the price index. However, as explained by previous research, there are potential benefits of first forecasting individual categories and then adding them back together using the same percentage share used when creating the index (Hyndman et al., 2011). In previous literature a bottom-up approach is sometimes shown to both improve model performance as well as providing valuable insights about underlying items.

The next models to be tested are therefore a hierarchical (bottom-up) SARIMA and a hierarchical (bottom-up) SARIMAX. The procedure is the same as in the aggregated case, but 103 SARIMA models and 103 SARIMAX models are now trained and tuned. This means that for each category, both the SARIMA and the SARIMAX model has its own combination of p, d, q, P, D, Q and s. This also implies that each SARIMAX model can select different macroeconomic variables depending on what features are highly correlated and kept by LASSO for that specific category. The 103 forecasts of the category-level SARIMA models are summed, which creates the forecast for the first difference of aggregated price index. The same applies to the 103 forecasts made by the SARIMAX models.

4.4 XGBoost Approach

In contrast to the traditional statistical approach above, this section covers the methodology of a machine learning based approach with gradient boosted decision trees. Namely, the extreme gradient boosted tree algorithm XGBoost. The prerequisites regarding data structure are somewhat different to the statistical approach, hence a reformatted data structure was needed. As dependent variables, both the aggregated price index as well as all the category-level indices were kept. To adapt the data for XGBoost, nine lags of differenced category price indices were created in addition to the macroeconomic variables. Nine lags were included to enable a fair comparison to the statistical modeling approach. Lastly, a new set of variables for the month was created with one-hot encoding for each month of the year to easily capture seasonality. These modifications to the data significantly increased the dimensionality, hence the importance of properly selecting features in the subsequent step.

Feature selection and feature scaling are two procedures that are different with regards to tree based models. Scaling and normalization is not necessary due to the tree structure, additionally feature selection is in a sense automatically conducted in the algorithm by the splits. Additionally, when tuning the model, regularization parameters for L1 and L2 regularization were included in the parameter grid to only include relevant features in the model. In order to train a general and well-performing model, cross validation was employed on a parameter grid. This ensures that the model is tuned to utilize the optimal values of the parameters in the grid according to a seven-fold cross validation. The parameters included in the grid were number of estimators, max depth, learning rate, lambda, alpha, gamma and subsample. The values of these parameters in the grid and the effect they have on the model is presented in table 4.2.

Table 4.2: Parameter Grid for Model Tuning of XGBoost

Hyperparameter	**Values**
Number of boosting rounds	[50, 60, 70, 75, 80, 90, 100, 125, 150]
Maximum tree depth	[2, 3, 5, 7]
Learning Rate	[0.1, 0.01, 0.001]
L2 regularization	[0, 0.1, 0.3, 0.5, 0.7]
L1 regularization	[0, 0.1, 0.2, 0.3]
Minimum loss reduction allowed for a split	[0, 0.1, 0.2, 0.3]
Subsample	[0.5, 0.7, 0.9]

These parameters were then tuned using GridSearch cross validation with seven folds on the training data set which is the 36 months prior to the first observation in the validation set. Both RandomizedSearch and GridSearch were tested. However, GridSearch did perform better although it is much more time consuming. The main difference between GridSearch and RandomizedSearch when cross validating is the methodology. When using GridSearch it systematically tests every combination of parameters in the grid. In comparison, RandomizedSearch randomly selects parameter combinations and provides us with the best one given the number of iterations it is allowed to test.

4.5 Ensemble Methods

Previous studies show that a combination of models can be used to further improve performance. Combining models in different ways is often referred to as ensemble methods (Ribeiro and Dos Santos Coelho, 2020) or consensus methods (Marmion et al., 2009). A summary of articles that show how different ensemble techniques can improve single forecasting models in price forecasting is provided by Ribeiro and Dos Santos Coelho (2020). While the ensemble methods exemplified in their paper are all showing improved performance, they tend to be complicated. In many other fields than economics, consensus methods have also shown performance boosts compared to stand-alone models. One of the simplest ways to implement consensus methods that show improved performance is taking averages of the forecasts of different models (Marmion et al., 2009). Marmion et al. (2009) among others also discuss potential improvements using weighted average ensembles, where different models are assigned different weights based on their estimated performance.

In this study an average ensemble of the best models was created using the SARIMAX, the hierarchical SARIMAX and the XGBoost models. In other words, for each of the 48 periods that were forecasted, the forecasts of the three best models were summed and divided by three. This method will also be compared to creating an ensemble of the same three models using the optimal linear combination suggested by a linear regression model applied to the 24 in-sample forecasts.

5

Empirical Analysis

This chapter will start with a exploratory analysis in Section 5.1. This will be followed by results and analysis of the different modelling approaches in Section 5.2 to 5.6. A comparison of all models performance is found in Section 5.7. Lastly, models for organic and conventional foods will be analyzed in Section 5.8.

5.1 Exploratory Analysis

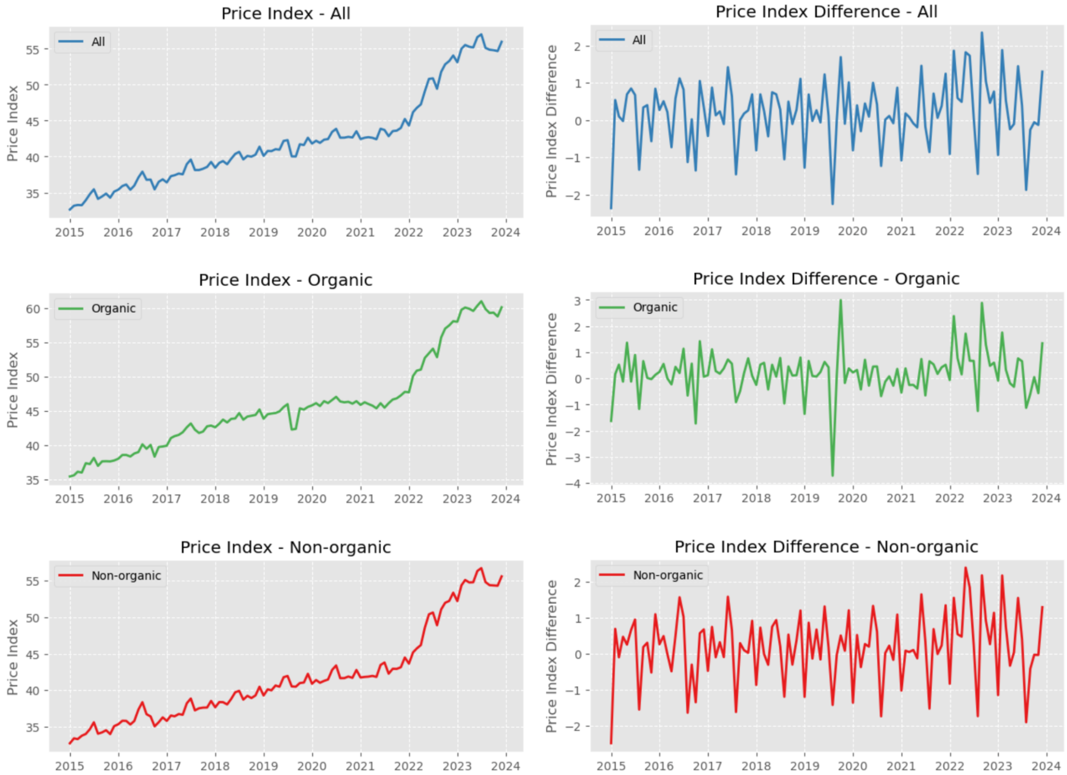


Figure 5.1: Price Indices over time for organic, conventional and the ordinary index.

Figure 5.1 depicts the price indices used in this paper. The left column shows the aggregated price indices before differencing, where the increase in food prices in 2022 is very clear. Prior to the war in Ukraine, the trend for food price inflation appears to be rather constant with a slight stagnation between 2020 and 2022. These were the years most affected by the Covid-19 pandemic which significantly impacted public sector organizations. In the right column the first differenced time series are presented. Here the series are stationary and a clear seasonality is present in the data. It appears that there is a significant downturn in food prices in January and August. Yet, whether this constitutes a genuine trend in food pricing or stems from heightened procurement by public sector entities in Sweden during these months remains to be clarified. Given that the prices reflect average costs across various grocery categories, it is plausible that the commencement of new semesters during these two months prompts bulk purchases, thereby mitigating the average unit price. However, without looking into the specific goods bought during different months it is not possible to accurately pinpoint the reason for this behavior.

Figure 5.2 and 5.3 depicts the macroeconomic variables used in the modeling, some of them are price indices, interest rates and prices in different currencies. Therefore, the scale and unit on the y-axis is different for each and every variable. However, it is still possible to interpret both the tendencies and trends in these time series. Generally, for all the series, the last years of the Covid-19 pandemic and war in Europe has affected them. Significant increases are present in all three exchange rates since the outbreak of the Russian invasion in February of 2022. The Swedish policy rate, treasury bill and government bond also increased in this time period from low and stable levels the years prior. When considering the prices of products and commodities they all significantly increased in this time period as well. It is safe to say that the last couple of years has significantly affected all of the macroeconomic variables, this in turn will hopefully provide useful information for the models to accurately model the similar pattern found in the food price index as well. In the right hand side of Figure 5.2 and 5.3 the first differenced values of all macroeconomic variables are presented. When considering the differenced series, most of them go from stable levels with constant variance to greater fluctuations in the last two to three years. The exchange rates are the only ones that do not show much difference in behavior over time. The differenced series are the input to the models.

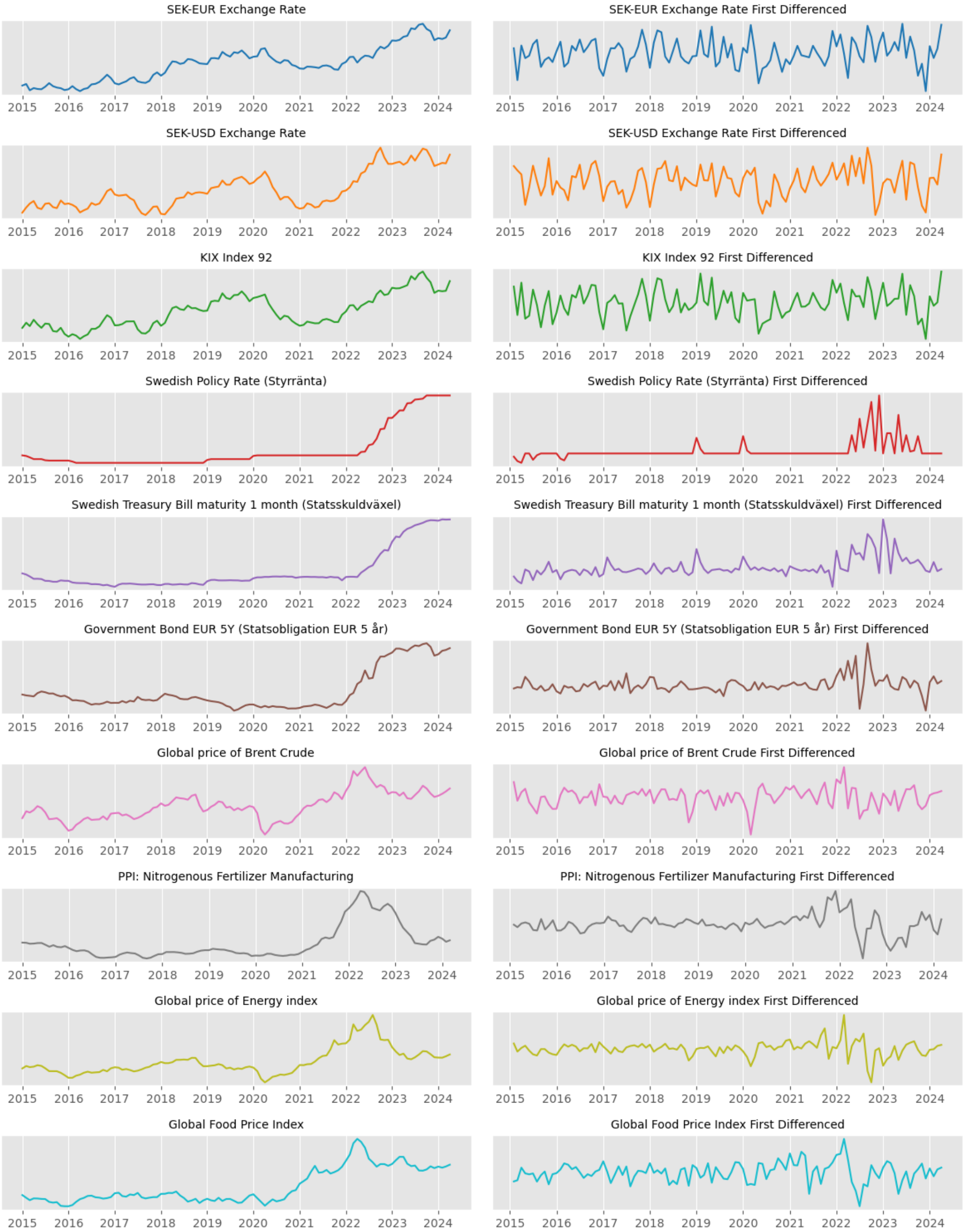


Figure 5.2: Evolution of first differences of macroeconomic variables over time.

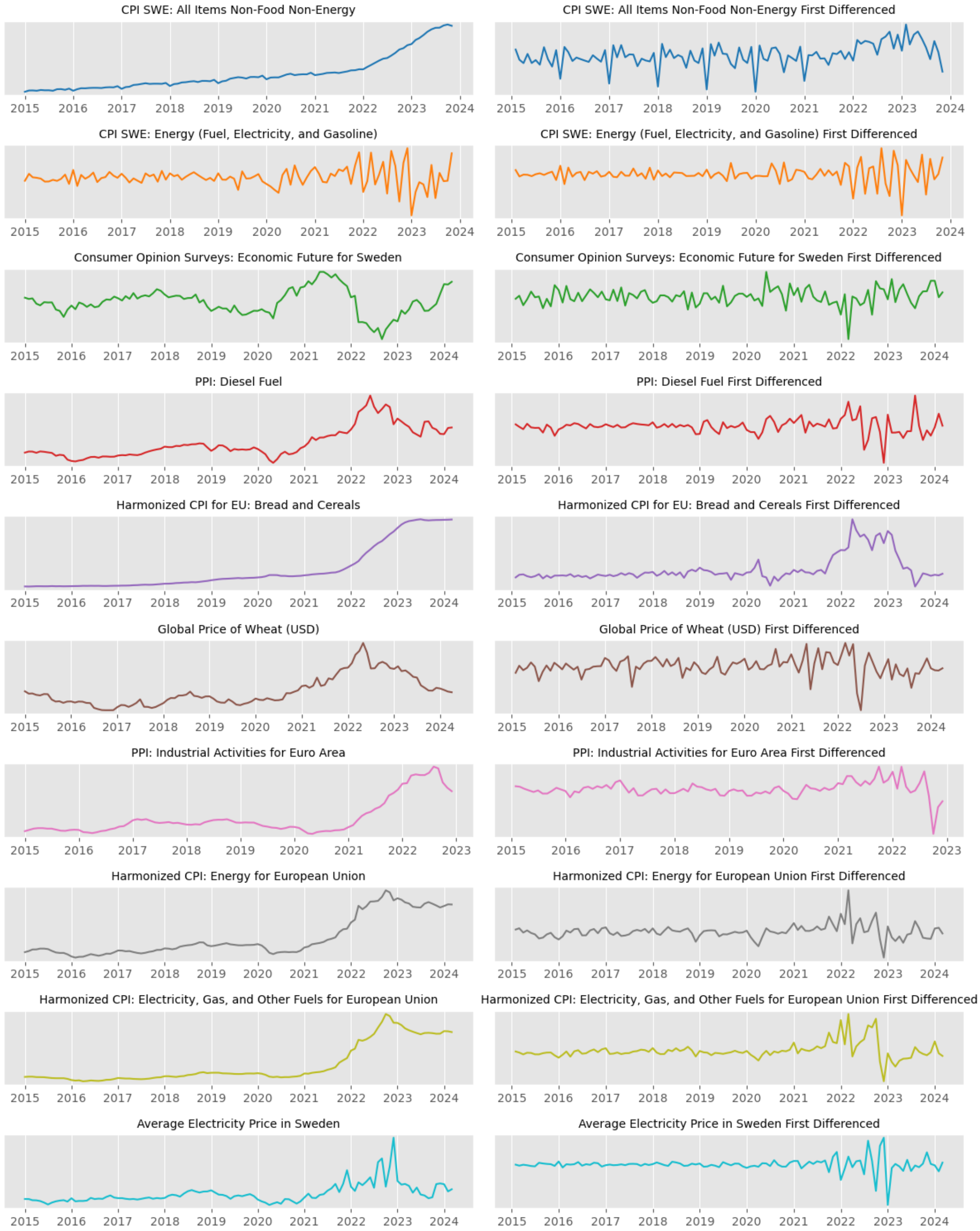


Figure 5.3: Evolution of first differences of macroeconomic variables over time.

One way to assess the relevance of macroeconomic variables and their lags is to consider correlation matrices. The correlation matrix in figure 5.4 presents the correlation between the price index and all nine lags of the macroeconomic variables between January 2019 and December 2021, the same period as the models are trained and tuned on. Significant correlation is present in the second lag of the exchange rates of Swedish krona against euro, dollar and the combined exchange rate of 27 trade partners. The Swedish policy rate and treasury bill is most correlated at lag 5. However, the opposite relationship exists between these two and the price index at lag 7. Other interesting patterns visible in the figure is that the global price of wheat is most correlated at the first lag and then it is very close to zero. The price of Brent crude oil is however not very correlated with the food price index during this time period, even though previous research by Baffes (2007) suggests that an oil price shock will in part influence the food prices as well. Interestingly, the strongest correlation in figure 5.4 is not present in the first lag. Insinuating that in this period, forecasting with longer horion than one month ahead is viable.

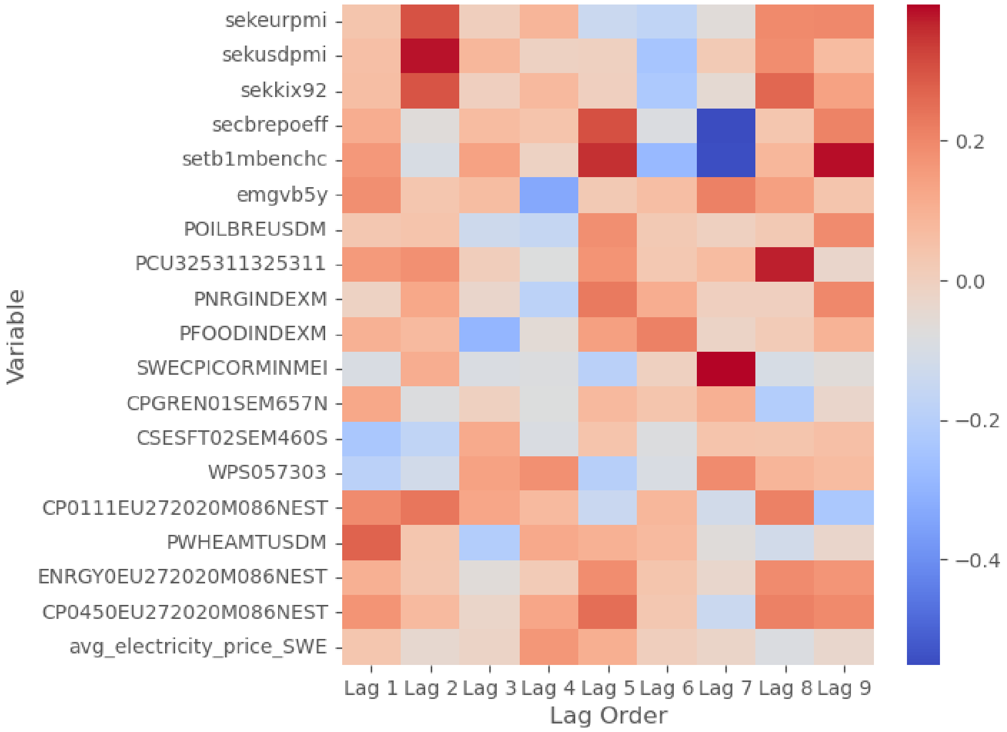


Figure 5.4: Matrix with correlation for all lags of the macroeconomic variables with the independent variable 'First Difference of Price Index' between January 2019 and December 2021.

One important aspect to highlight is the difference in the selection of macroeconomic variables and lags. The statistical models SARIMA and SARIMAX selects its macro variables based on the period 2019-2021 as shown in figure 5.4. XGboost on the other hand re-evaluates the variable selection for each forecast and hence allowing for a more dynamic model due to the rolling window forecasting approach. The statistical models can be configured to do this as well, but it increases the computational complexity significantly and thus not implemented in this paper.

5.2 Rolling Average

The rolling average model is the benchmark model in this paper. It is a simple way to benchmark and mostly used as a proof of concept (Conol, 2020). The length of the rolling average window is found to be 60 months, i.e. five years, through cross validation. The forecasting performance of this benchmarking model is depicted in figure 5.5. This figure shows both the in-sample and out-of-sample forecast. In Table 5.1, the evaluation metric MSFE that will be compared with the other models is presented.

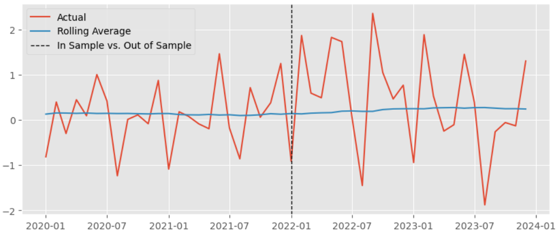


Figure 5.5: Rolling Average forecast

As one might expect, the rolling average is forecasting a slight positive value indicating a fairly constant price inflation. As exemplified in Figure 5.1, starting from 2022 the food price inflation increased from this fairly steady level. Hence, the benchmark should easily be beaten by other models that can capture the dynamics of food prices.

Table 5.1: Mean Squared Forecasting Errors for rolling window forecast (36 months) both in and out-of-sample

Model	**MSFE: 2020-2021**	**MSFE: 2022-2023**
Rolling Average	0.4467	1.2572

The MSFE in Table 5.1 is lower in sample (2020-2021) than it is out-of-sample (2022-2023). This is all due to the sudden change in food prices which the rolling average will have difficulties capturing due to its 60 month window. Conversely, if the food price inflation stabilizes around a lower level again in the future, it will take time for it to adapt to this more steady state.

5.3 Statistical Approach

The tuned and selected hyperparameters by BIC for the SARIMA and SARIMAX models are presented in Table 5.2. These are very similar for the two models with the main difference being that the SARIMAX does not include any AR components and instead uses the seasonal MA-component. The parameter p, d and q represents the ARIMA parameters where p is the autoregressive component, d is the differences and q is the MA component. In the seasonal models, P, D and Q are the seasonal autoregressive, differences and MA components. In effect, the SARIMAX model is therefore an MA model with seasonal effects and exogenous variables whilst the SARIMA includes the AR component and no MA component in the seasonal effect.

Given the complexity of the situation, it is best to simplify by referring to them as SARIMA and SARIMAX moving forward, as these are the frameworks in use.

Table 5.2: SARIMA Hyperparameters

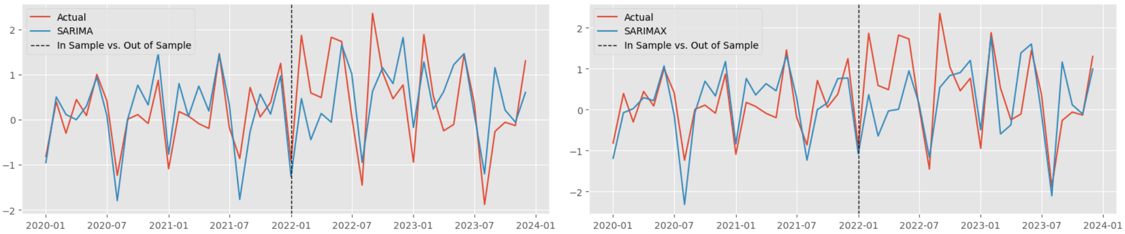
Hyperparameter	p	d	q	P	D	Q	s
SARIMA	1	0	2	2	1	0	12
SARIMAX	0	0	2	2	1	1	12

The performance of the two models in Table 5.3 are very similar. Although, the SARIMAX which includes macroeconomic variables is slightly better performing both in and out-of-sample. When comparing this with the rolling average forecast, the statistical modeling approach is superior.

Table 5.3: Mean Squared Forecasting Errors for SARIMA and SARIMAX

Model	**MSFE: 2020-2021**	**MSFE: 2022-2023**
SARIMA	0.2321	0.7829
SARIMAX	0.2209	0.7469
Rolling Average	0.4467	1.2572

This is also evident in figure 5.6 where both models follows the dynamics of the actual values. Nevertheless, it is evident that both models encounter difficulties in the year 2022, where they underestimate the increase in food prices. While they accurately capture the direction of the fluctuations, they fall short in accurately representing the magnitude of the change. This is unsurprising, as the model lacks significant knowledge of the war in Europe and its potential impact on prices in the future.



(a) SARIMA

(b) SARIMAX

Figure 5.6: Comparison of non-hierarchical SARIMA and SARIMAX forecasts

The addition of macroeconomic variables improves the forecasting performance compared to the SARIMA model. The improvement in MSFE is small but both models perform significantly better than the rolling average benchmark both in- and out-of-sample.

An interesting aspect of the SARIMAX model is that it is possible to identify which macroeconomic variables it selected and the lag order of these, which is found in Table 5.4. The global food price index three months ago, Swedish treasury bill with one month maturity five months ago and the government bond are selected to name a few. Notably, most of these are Swedish macroeconomic variables except global food price index and harmonized consumer price index for bread and cereals in the European Union.

Table 5.4: Selection of macroeconomic variables for SARIMAX

Feature	**Lag**
PFOODINDEXM	3
setb1mbenchc	5
emgvb5y	4
setb1mbenchc	6
secbrepoeff	7
CP0111EU272020M086NEST	8

5.4 Hierarchical Approach

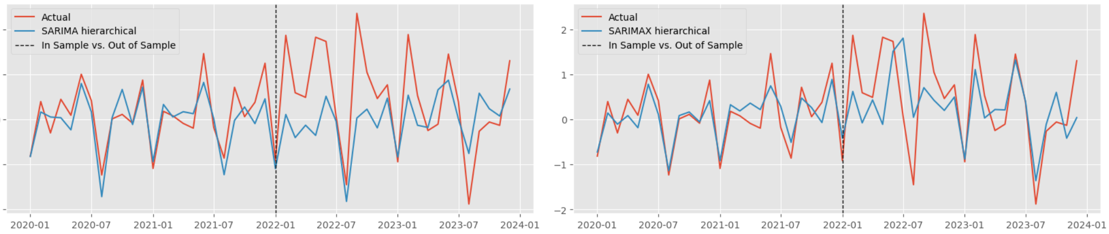
As described in section 4.3, the same modelling approach as for the SARIMA and SARIMAX is used for the hierarchical version. In the hierarchical approach however, 103 separate models on disaggregate data are tuned and trained before forecasts are added together to create the index. The parameter specifications construct a vast matrix that will not be presented in this paper. All disaggregate models are however tuned according to the same principle as before. While the disaggregate models may not strictly adhere to the full SARIMA or SARIMAX specifications, they do operate within the framework and have the flexibility to utilize all parameters in their specifications. Hence, the overarching modeling approach is SARIMA and SARIMAX.

In table 5.5 the performance of the hierarchical approach is shown to be more reliable on incorporating macroeconomic variables than the non-hierarchical approach. Out-of-sample the hierarchical SARIMAX is equivalent to the statistical models while the hierarchical SARIMA is clearly worse. The in-sample performance is better for both hierarchical models compared to the non-hierarchical SARIMA.

Table 5.5: Mean Squared Forecasting Errors for hierarchical SARIMA and hierarchical SARIMAX

Model	**MSFE: 2020-2021**	**MSFE: 2022-2023**
Hierarchical SARIMA	0.1397	0.9687
Hierarchical SARIMAX	0.1000	0.7544
Rolling Average	0.4467	1.2572
SARIMA	0.2321	0.7829

Considering figure 5.7, it is clear that the dynamics of 2022 are poorly captured by the models. On the other hand, the hierarchical SARIMAX is quite accurate in 2023 compared to the other models. This indicates that the effects of 2022 are not as present in 2023 since the model is performing similarly to the in-sample period. Thereby, it is possible that food price inflation is settling down to pre-war levels. Conversely, it is also possible that the model is slowly adapting and learning the new reality since the war started. However, whatever the true effect is, the forecasting performance is better in 2023.



(a) SARIMA

(b) SARIMAX

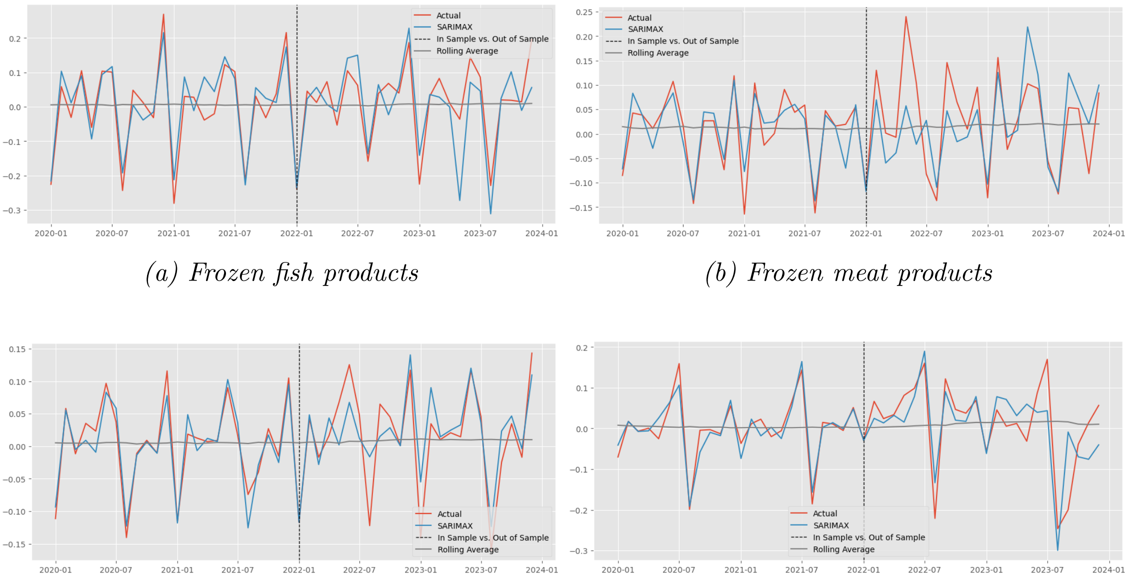
Figure 5.7: Comparison of hierarchical SARIMA and SARIMAX forecasts

The hierarchical modeling approach enables the analysis of results from models focused on specific categories. The specifications and results of the four best models, compared to rolling average, are presented in table 5.6. These categories are frozen fish products, frozen meat products, fresh meat products and fresh dairy products. The MSFE is less than half of the corresponding rolling average across all four categories. Moreover, these categories hold significant weight in the aggregated price index, with fresh dairy products having the largest share among them. An intriguing and intuitive observation from the table is that the variables chosen for fresh meat products, which are predominantly Swedish (approximately 99% during the in-sample period), consist solely of Swedish macroeconomic variables. Conversely, for frozen fish products, where only about 40% are Swedish, predominantly global and European macroeconomic variables are selected.

Table 5.6: Statistical performance and component orders by category

Product	**MSFE SARIMAX (OOS)**	**MSFE Rolling avg (OOS)**	**Features**	**SARIMA Orders**
Frozen fish products	0.0057	0.0135	setb1mbenchc, CP0111EU272020M086NEST, POILBREUSDM, WPS057303	(0, 0, 1), (1, 0, 0, 12)
Frozen meat products	0.0017	0.0065	WPS057303	(0, 0, 1), (1, 0, 0, 12)
Fresh meat products	0.0045	0.0113	setb1mbenchc, emgvb5y, sekeurpmi	(1, 0, 0), (1, 0, 1, 12)
Fresh dairy products	0.0025	0.0087	setb1mbenchc, CP0450EU272020M086NEST, secbrepoeff, PWHEAMTUSDM, sekkix92, ENRGY0EU272020M086NEST, CP0450EU272020M086NEST	(0, 0, 0), (0, 1, 0, 12)

In figure 5.8, the four best performing submodels from the hierarchical SARIMAX are visualized. In these four models, the SARIMAX predictions follow the actual values fairly closely. It is also possible to distinguish differences in the behaviour of the prices between the categories. As an illustration, Matilda Foodtech elucidated that Swedish Christmas ham, a component of fresh meat products, could elucidate the price surges during December months. This is due to its relatively high cost and the fact that it is primarily purchased in December, aligning with traditional consumption patterns. Clear seasonal spikes and dips like these are generally well-captured by the models. Issues arise when price fluctuations deviate from these established patterns, resulting in significant price changes occurring in months that were not previously affected by such fluctuations. This is a sign that the effect of macroeconomic variables on price changes is not constant, and their relative importance could change over time. The solid fit of the fresh meat product model in-sample and in 2023, contrasted with its poor performance in 2022, suggests that during periods of economic instability, additional variables may influence outcomes compared to more stable economic periods. Another possible explanation is that the models mostly captures seasonality, and given the large amount of features to choose from, some features are simply selected by chance and should not actually be part of the feature set.



(a) Frozen fish products

(b) Frozen meat products

(c) Fresh meat products

(d) Fresh dairy products

Figure 5.8: Four best models in the hierarchical SARIMAX model

Table 5.7 displays the frequency with which the twelve most selected macroeconomic variables are employed in the category-level models. The global wheat price is most commonly used, followed by the Swedish treasury bill and US diesel prices. Additional insights can be gleaned from less frequently chosen features like the SEK-USD exchange rate. This feature is predominantly selected for categories typically produced in the US and South America, such as colonial dairy products and colonial coffee. Conversely, the SEK-EUR exchange rate is identified in categories like colonial pasta, illustrating its varied usage across different product categories.

Table 5.7: Most common features all categories.

Rank	Feature	Count
1	PWHEAMTUSDM	56
2	setb1mbenchc	46
3	WPS057303	43
4	avg_electricity_price_SWE	37
5	sekkix92	31
6	PFOODINDEXM	30
7	CSESFT02SEM460S	28
8	CP0450EU272020M086NEST	25
9	PCU325311325311	21
10	CP0111EU272020M086NEST	18
11	PNRGINDEXM	17
12	POILBREUSDM	16

This can further be broken down by main category as in table 5.8, which can provide insights about differences between the groups. An intriguing example is that Swedish electricity prices are ranked as third, fourth, seventh, and are not included for fresh, frozen, colonial, and nutrition, respectively. The share of Swedish goods in these main categories in the sample period is 73%, 52%, 41% and 0.6%. Another example is that global price index for energy, which is the eleventh most selected overall, is the fifth most selected for frozen. Note that it is not possible to make inference-like conclusions from this analysis, one can only tell that feature selections often seem intuitive.

Table 5.8: Most common features by main category.

Rank	Frozen
1	WPS057303
2	setb1mbenchc
3	PWHEAMTUSDM
4	avg_electricity_price_SWE
5	PNRGINDEXM
6	CSESFT02SEM460S
7	PCU325311325311

Rank	Fresh
1	PWHEAMTUSDM
2	setb1mbenchc
3	avg_electricity_price_SWE
4	sekkix92
5	PFOODINDEXM
6	WPS057303
7	CP0450EU272020M086NEST

Rank	Colonial
1	PWHEAMTUSDM
2	setb1mbenchc
3	CSESFT02SEM460S
4	sekkix92
5	PFOODINDEXM
6	WPS057303
7	avg_electricity_price_SWE

Rank	Nutrition
1	WPS057303
2	PWHEAMTUSDM
3	POILBREUSDM
4	NA
5	NA
6	NA
7	NA

5.5 XGBoost

The seven-fold cross-validation when tuning the model provided the model specification of parameters shown in table 5.9. Surprisingly, it favored relatively large trees with a max depth of seven, compared to stumps which is what James et al. (2023) describes as the most efficient specification.

Table 5.9: Hyperparameter specification from cross validation and tuning.

Hyperparameter	**Value**
Number of boosting rounds	150
Max tree depth	7
Learning Rate	0.1
L2 regularization	0
L1 regularization	0
Minimum loss reduction allowed for a split	0
Subsample	0.7

Interestingly, it does not pick any regularisation parameters. Manually adding some regularization can be beneficial if the tuned model does not generalize to new unseen data very well. Figure 5.9 shows the performance of the model. Generally, the model follows the patterns and fluctuation in the actual data very well, especially in sample. However, it does not fully capture the amplitude of these fluctuations even though it appears to go in the same direction as the actual observed values. Although, the sample ends right before the war in Ukraine and the increasing food price inflation as visible in figure 5.1. By considering the MSFE in table 5.10, it is clear that the error is considerably lower in sample than out-of-sample. It is however still a well performing model with slightly lower MSFE than the statistical models.

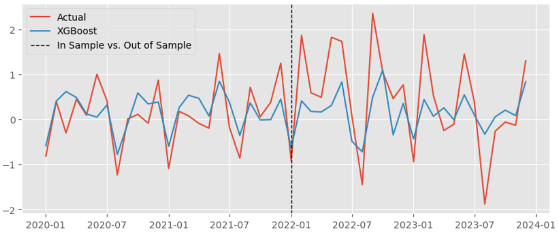


Figure 5.9: XGBoost forecast in- and out-of-sample

Table 5.10: Mean Squared Forecasting Errors for XGBoost and benchmarking models.

Model	**MSFE: 2020-2021**	**MSFE: 2022-2023**
XGBoost	0.2232	0.7180
Rolling Average	0.4467	1.2572
SARIMA	0.2321	0.7829

One advantage by using XGBoost is that one can easily extract the most important features. Feature importance in this case is the built in function in the XGBoost algorithm, rating the importance of each feature by how much impurity it reduced in the tree by splitting on the feature in a tree (Marsh, 2023). The average for all trees are then used to create the feature importance figures in this paper (Marsh, 2023). Figure 5.10 depicts the 25 most important features from the 36 month long training period between 2019 and 2021. In comparison, figure 5.11 shows the most important features from the last forecast, i.e. the forecast for December 2023. Notice that the macroeconomic variables are significantly more important in the latter period.

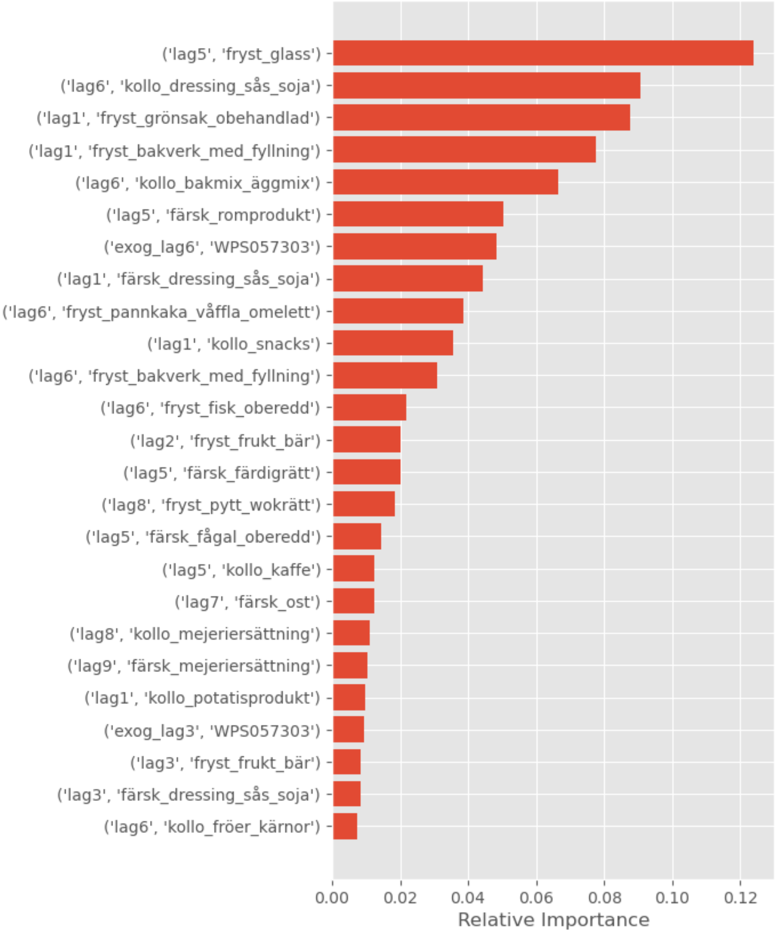


Figure 5.10: The 25 most important features from the training period 2019-2021

In figure 5.10, the relatively most important feature is the fifth lag of frozen ice cream. Whether this is true or a seasonal effect not captured by the model is hard to differentiate between. However, ice cream is most likely a very seasonally dependent good and thereby it is also likely that it is the latter. The six most important features are disaggregate index variables with varying lags, meaning that the autoregressive endogenous features can to a larger degree explain the future behavior of food price inflation in the training period 2019-2021 compared to figure 5.11 which used the 36 months prior to the last forecast of December 2023. Macroeconomic features that are present among the 25 most important features in the training set are the diesel price six months ago and the diesel price three months ago.

In comparison, the most important macroeconomic features in the 36 months prior to the last forecast (2020-2023) are very different. These are the European Energy Price Index in EU 6 months ago, diesel price six months ago, Swedish consumer price index for energy one month ago, five year European government bond one month ago, global wheat price one month ago and many more. This highlights that the macroeconomic features plays a bigger role during the time period with

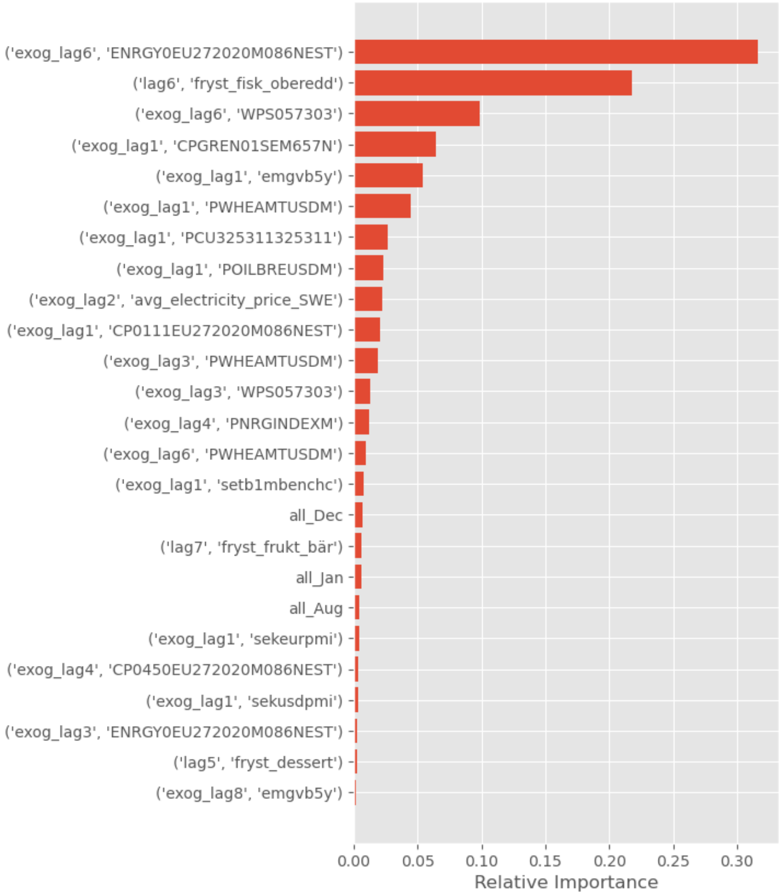


Figure 5.11: The 25 most important features from the 36 months prior to the last forecast in December 2023.

higher food price inflation. The effect of an oil price shock is only visible in times where shocks are present, thereby the difference in features validates Baffes's (2007) results that the effect of an oil price shock spills over on food prices. Interestingly, the first lag of oil prices is most important meaning that a change in oil price effects food prices one month later. This is rather fast and thus not likely to be caused by fertilizer prices and instead more likely increased transportation costs. In contrary, the diesel fuel price is most important with six lags which means that it is not as closely related to transportation. Instead, the time frame is better suited to be the result of farming and agricultural machines running on diesel fuel. However, the diesel price in the model is from the US so even though oil related goods are sold on a global market and thus prices are interlinked, it does not account for subsidies and stimulus grants currently present in the EU (European Commission, 2024). Another interesting feature is the sixth lag of frozen unprepared fish. As shown in Table 5.8, frozen goods are from the statistical approach dependent on electricity prices and energy. It would make sense since fishing as an agricultural activity is dependent on fuels and frozen goods are linked with electricity prices to stay frozen.

XGBoost is the well-rounded stand-alone model that overall suits the data best in this paper, with consistent forecasts both in and out-of-sample and good adaptation to changes such as the war in Ukraine. The advantage of using a tree based method is its flexibility, ease of use and interpretation ability. As long as the time series data frame is correctly formatted with new columns for each autoregressive lag, then it is able to cope with large amounts of data and many additional variable. Another benefit is that combining both numerical and categorical variables in the algorithm is straight-forward with little to none preprocessing required. One drawback is that the algorithm isn't specifically designed for time series data, resulting in that one must create shifted value columns to allow for autoregressive components.

Another benefit of XGBoost is that it can handle non-stationary data as well. Resulting in relaxed settings were more variables can be included and thus setting the stage for a well-performing model. In a volatile and uncertain environment like the food market in 2022, XGBoost's flexibility and adaptability may have allowed it to generate more accurate and robust predictions than the statistical methods.

So, why did it perform significantly better during the troubled year of 2022 and thus being the best performing stand-alone model in this paper? The algorithm can handle complex non-linear relationships, making it flexible and adaptable than the SARIMAX. Additionally, if the seasonality aspect changed or was disrupted, the algorithms flexibility can capture this more easily due to its adaptability. From Figure 5.2 and 5.3, there were some clear disruptions in the series in this time period. Such as the spiking energy prices, wheat prices and exchange rates that might temporarily violate the stationarity assumptions of other models rendering XGBoost the most viable approach in terms of crises.

5.6 Ensemble Methods

There are different ways to combine forecasting models as alluded to previously, one way is to average different model, another is to through a regression of the forecasts in sample construct a linear combination of the forecasting models. One upside of this is that if a model accurately captures one aspect of the series that another model misses, then combining them can improve the performance. However, the interpretation is then not as straight-forward as a stand-alone model.

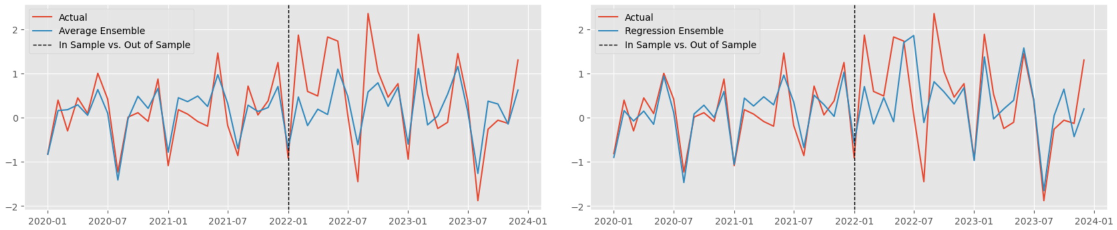
The Average Ensemble model is constructed by the arithmetic mean of the forecasts for each period from the the SARIMAX, hierarchical SARIMAX and XGBoost models above. Although averaging the best performing models is primarily chosen as a linear combination for its simplicity, it is still proven to improve performance in previous research (Marmion et al., 2009).

However, if validation of the linear combination of the same three models is performed through linearly regressing the in-sample forecasts to find the ideal linear combination. This should definitely improve the in-sample forecasts, but does not necessarily generalize to new unseen data very well.

The results for the ensemble models are presented in table 5.11 confirms this. In-sample the regression ensemble is, as expected, better than all other models presented in this study. Out-of-sample it is however not better than any of the three models indicating that the regression has overfitted the ensemble to the in-sample data. As mentioned for other models, this could also be due to the rapidly increasing food price inflation where other macroeconomic variables are important starting 2022.

Table 5.11: Mean Squared Forecasting Errors for Average Ensemble, both in and out-of-sample

Model	**MSFE: 2020-2021**	**MSFE: 2022-2023**
Average Ensemble	0.1104	0.5621
Regression Ensemble	0.0824	0.7552
Rolling Average	0.4467	1.2572
SARIMA	0.2321	0.7829



(a) Average ensemble

(b) Regression ensemble

Figure 5.12: Comparison of Average Ensemble and Regression Ensemble in and out-of-sample.

Considering the evaluation metrics, the ensemble method performs better both in and out of sample than all other models above. It does follow the fluctuations very nicely, but just as with the previous models it does struggle with the amplitude of the fluctuations. However, it appears that 2022 is the worst year and that it captures the movement nicely in 2023 again.

The difference in out-of-sample MSFE for regression ensemble model is approximately 0.2 compared to the average model, resulting in a performance not significantly better or worse than a stand-alone model.

When summarizing the results in this paper, the average ensemble model is the clear winner of this study in terms of out-of-sample forecasting performance. Once all single models are trained and forecasts are made it is also relatively simple to create both the average ensemble and the linear ensemble. It does however require a lot of computation to create each single model. Therefore, in terms of effort, the ensemble methods are the worst. When considering the comprehensibility of which variables drive price changes, ensemble methods exhibit evident drawbacks. This is because it must be deciphered using the linear combination and all the input models.

5.7 Model Comparison

Table 5.12 shows the performance of each model for each year. The SARIMA model displays a significant improvement in forecasting performance in comparison with the benchmark rolling average. It is therefore reasonable to, from here on out consider the SARIMA as the benchmark model, which is also more in line with previous studies. In 2020 and 2021 the hierarchical SARIMAX outperforms the other models. However, this is in sample where the model is trained. The worst performing models during these two years are XGBoost and the SARIMA model. In 2022, the best single model is XGBoost although it is just slightly beaten by the average ensemble model. When considering 2023 there are some interesting results, the average ensemble model is the best performer but more importantly, the hierarchical SARIMAX is not far behind and they are both significantly better than XGBoost. It is fair to say that 2022 was a special year, not only for food price inflation. One way to look at these results is that it is clear that some models cope with anomalies better than other and when the global crises settles somewhat, the hierarchical SARIMAX provide us with accurate forecasts once more.

Table 5.12: MSFE for all models for each year separately. 2020-2021 is in sample and 2022-2023 is out-of-sample.

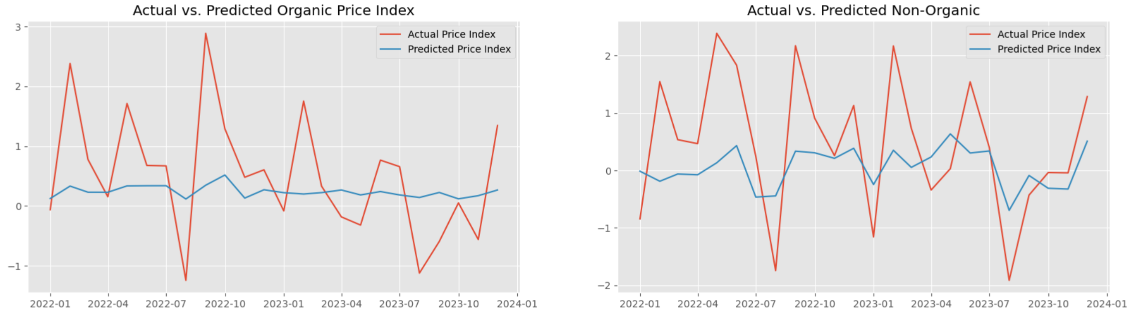
Model	MSFE			
	2020	2021	2022	2023
Rolling average	0.3789	0.5144	1.4926	1.0218
SARIMA	0.1513	0.3129	1.0136	0.5522
SARIMA hierarchical	0.0932	0.1862	1.4531	0.4844
SARIMAX	0.2162	0.2256	0.9796	0.5142
SARIMAX hierarchical	0.0570	0.1430	1.2098	0.2990
XGBoost	0.2227	0.2236	0.8897	0.5463
Ensemble - average	0.0707	0.1501	0.8577	0.2665
Ensemble - linear regression	0.0442	0.1238	1.1650	0.2588

Among the ordinary models, the hierarchical SARIMAX outperforms the other models in three out of four periods. However, the combined effect out-of-sample results in better performance by the XGBoost model. The color coding in table 5.12 highlights the best performing single and ensemble model. Consequently, the best performing stand-alone model is the XGBoost, which will be the foundation for analysing organic and conventional data sets.

SARIMAX based models exhibits improved forecasting performances compared to their SARIMA counterparts. The SARIMAX model performs better than the equivalent SARIMA model for both out-of-sample years, which is also true, and more substantial, for the hierarchical approach. While this analysis is not focused on the inference of whether a certain variable actually influences the price or whether it just functions as an indicator, it is safe to say that macroeconomic variables can be useful for food price forecasting.

5.8 Organic & Conventional Food

This section can be considered as a test of generalizability of the best performing single model (XGBoost) to see whether it is a valid procedure for these subcategories. Additionally, by considering the feature importances between organic and conventional foods, see whether the impact of macroeconomic variables are different for the two categories.



(a) Organic *(b) Conventional*

Figure 5.13: Comparison of XGBoost forecasts between organic and conventional foods out of sample.

The dataset only containing organic food items was used and XGBoost tuned and validated a new model with seven-fold cross validation, just as in the general model case above. The forecasting performance of that model is illustrated in Figure 5.13a. The poor prediction in Figure 5.13a is also reflected in the evaluation metric in Table 5.13. It is slightly better than the rolling average benchmark, but it is clear that it does not fit the movements in the data very well. Instead, it hovers just above zero, resulting in a behavior close to the benchmark rolling average of forecasting a steady inflation rate.

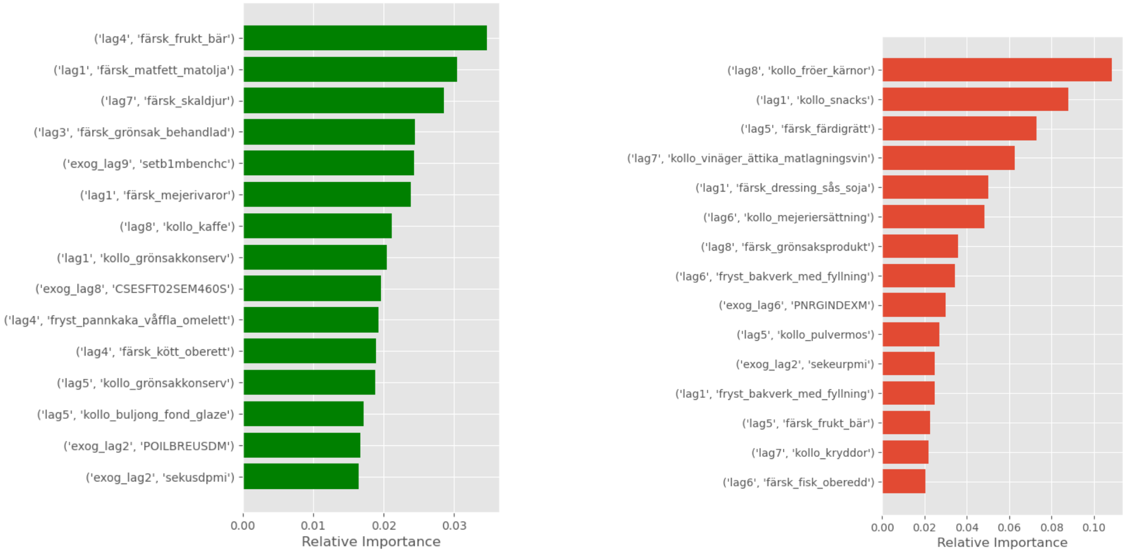
The same procedure as for organic foods was conducted on the conventional food dataset. Figure 5.13b illustrates the forecasts in comparison to the actual values. It is clear that it does not fully capture the patterns of food price inflation of conventional foods. It does however, follow the actual values better than for organic goods even though it does stay close to zero throughout the forecasting period.

It seems like the conventional model is rather restrictive in its forecasts. The amplitude of the fluctuations is not a big as the actual values. Potentially, this could be rectified by multiplication to get an even better performing model. However, since it is trained on a 36 month rolling period the period prior to increased food price inflation it is reasonable that it does not fully capture this. By linearly transforming the model forecast one might loose generalizability on new unseen data if overfitting on the validation period.

Model	Organic	Conventional
XGBoost	0.9685	1.1265
Rolling Average	1.1067	1.4278

Table 5.13: MSFE for Organic and Conventional foods using XGBoost and rolling average.

Figure 5.14 show the feature importance of the organic and conventional models. There are not any features that truly stand out compared to the full price index forecast in figure 5.10. It could be assumed that fossil fuels would be less significant for organic products, given the reduced dependency resulting from not using fertilizers. Although, there are some macroeconomic features that are relatively important such as the Swedish treasury bill nine months ago and the Brent crude oil price two months ago. There are not any outstanding differences in which features that impact organic and conventional foods in Figure 5.14. Despite assumptions about potential differences, the data fails to demonstrate them. Although, since these plots are only showing which features were important between 2019 and 2021 some things could have changed. Especially since the energy prices has gone up.



(a) Organic Foods

(b) Conventional Foods

Figure 5.14: Comparison of feature importance between organic and conventional food price indices

The forecasting performance of organic and conventional foods is noticeably inferior to that of the ordinary price index, showing only marginal improvement compared to their respective rolling average benchmarks. Whether this is the effect of data loss when using subsets or due to incorrect specification is difficult to determine. However, it should be noted that the parameter grid used for tuning the models was the exact same as for the ordinary price index. For the ordinary price index, the grid was evolved by multiple iterations of tuning the price index model and by adding more parameter values close to the selected values. This was not conducted for these subsets, rendering the model tuning less thorough than for the ordinary price index.

A big portion of the macroeconomic variables are energy related. Baffes (2007) found the pass through effect of oil on food to be 18% and 33% on fertilizer. Since fertilizer is only used on conventional foods, the assumption before analyzing was that we would see a difference in feature importance of fossil fuels on conventional foods compared to organic. However, there was no clear difference in Figure 5.14a and 5.14b on the top feature importances between organic and conventional foods. Conversely, Baffes (2007) researched the influence of oil shocks in the US. The food market is not entirely global, some produce are sold locally and thus it might be that his results are not directly transferable to the Swedish and European market. Whether the EU and Sweden are less dependent on fertilizer, fuels and oil than the US is outside the scope of this paper, but still relevant for correctly modeling the series.

What is intriguing is that both models exhibit similarly poor performance, even though one might assume that conventional foods would not be far off the ordinary price index. It would make more sense if the organic forecast was worse, since the majority of the macroeconomic variables revolve around energy and fuels that theoretically suit conventional foods better. Otherwise, it could be that other predictors are more useful for organic foods such as labor rates assuming it is more labor intensive by not using pesticides and chemicals to the same extent.

In summary, the results were not as distinct as one might think in advance. Whether it was due to model specification, feature selection or data loss is hard to say. Potentially, the time period chosen (22-23) was more complex and unforgiving for this test compared to a more stable period. Nevertheless, it is still relevant to predict these series individually. Especially since KRAV's (2022) article claimed that the price inflation grew differently for organic and conventional foods. For public sector organizations where environmentally friendly procurement is relevant for stakeholder, it would be very useful to forecast these series accurately.

6
Discussion

Section 6.1 will focus primarily on the validity of this study to what extent conclusion can be made. In Section 6.2, some practical applications and trustworthiness of forecast will be discussed.

6.1 Insights and Challanges

Having model evaluation of forecasting performance as the main goal of this paper quickly increased the complexity of the models. The primary issues include numerous lags, an abundance of features, and the multiplicative interaction between the two. The methodology of this study does not facilitate the discovery of significant evidence for particular food price drivers. While inference is not within the scope of this study, the chosen features can offer insights to guide future research. As suggested by Joutz's (1997), prices of input goods have been shown useful for forecasting in this study. There are logical suggestions from the feature selection procedures both in the hierarchical SARIMAX case and in the XGBoost case. A few such interesting things are that the global price of wheat is the most common variable selected in the feature selection process for the hierarchical SARIMAX, especially for colonial goods wheat is commonly an input into goods such as pasta and bakeries. Other interesting patterns are noticable at the main category level, particularly intriguing and intuitive is that electricity prices are more commonly selected for frozen and fresh goods than for colonial goods. The diesel price is also one of the most commonly chosen variables in the hierarchical SARIMAX. This variable is also found to be important by XGBoost on both occasions shown in the result.

It is also important to highlight that the impact of these macroeconomic variables seem to be changing over time. By comparing figure 6.1, presenting a correlation matrix between the macroeconomic variables and food price index between January 2022 and December 2023, to figure 5.4 (2019-2021), we suggest that the influence of macroeconomic variables has shifted. Most significantly the correlation between the global wheat price at one lag demonstrates a lot more correlation. Additionally the Swedish average electricity prices are more important and lastly the price of diesel and price of oil at lag 8 are also much more correlated to the price differences.

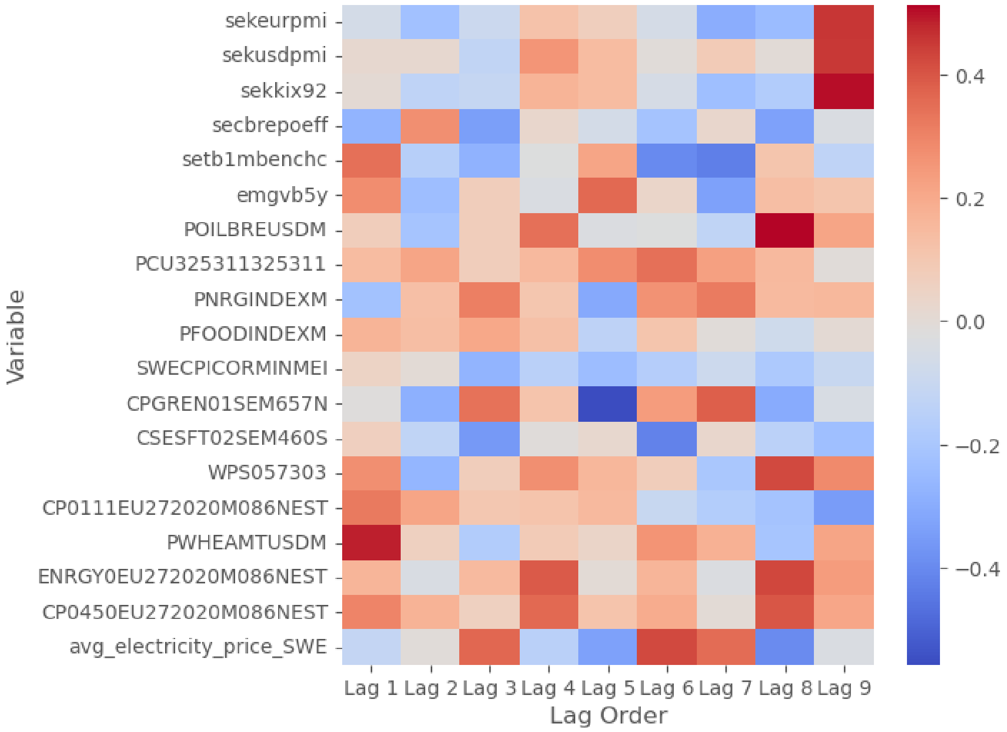


Figure 6.1: Matrix of macroeconomic features and their nine lags correlation with the price index between January 2022 and December 2023.

Methodological Insights

The study has shown that complex and modern models which include macroeconomic variables can outperform models that do not in terms of forecasting, which confirm Januschowski et al.'s (2022), Ribeiro and Dos Santos Coelho's (2020) and many other contemporary studies. The exact reason for why some models are better than others can however not be concluded by this study. The main reason is the differences in features used as input for the models. However, the algorithms work a bit differently and might capture different aspects of the relation. Most importantly, the XGBoost has access to all lagged variables, including lags of specific food categories, meaning that another variable can explain the difference in inflation than what is available for the SARIMAX. Furthermore, even though models at times select the same variables, they often choose different lags. While both XGBoost and SARIMAX models includes wheat prices, the most common lags selected by SARIMAX (lag 6), is only the third most important lag of wheat according to XGBoost. As mentioned in Chapter 4, a major difference between the SARIMAX models and the XGBoost is that the latter automatically select new variables for each training round, while the SARIMAX models use the variables selected based on 2019-2021. Same goes for price of diesel fuel where SARIMAX most commonly chooses lag 4, while XGBoost chooses lag 3 and lag 6.

In addition to that correlation between the dependent and independent variables changes over time, some macrovariables become non-stationary after 2021. P-values from ADF-tests including 2022-2023 are found in A.3. Thereby, not only should for example the first lag of wheat prices be included for out-of-sample forecasts, some variables should be completely dropped or differenced once more to not introduce spuriousness (Enders, 2015). Most importantly, the category-level SARIMAX models commonly selected the Swedish one-month treasury bill which is not stationary when including 2022, which could be part of the reason why hierarchical SARIMAX is performing much worse in 2022 than in other years. To make this problem even worse, the differenced aggregated price index is also non-stationary when including 2022. When excluding 2022 and looking at 2015-2021 and 2023 the series is again stationary. This is furthermore a potential explanation of why XGBoost is the best performing model in 2022, as it does not assume stationarity for either the dependent or independent variables.

Alternative Methods

When writing a thesis, there are many methodological considerations to me made. One of them is regarding feature importance since it can be calculated in multiple ways. An alternative approach to the one used is to use SHapley Additive exPlanations (SHAP) feature importances which are derived from game theory (Marsh, 2023). SHAP-values are the contribution of each variable but corrected to not be distorted by feature scaling and thus more accurate and consistent than the built-in feature importance in XGBoost according to Marsh (2023). SHAP was introduced by Lundberg and Lee (2017) to better suit the human intuition and allow for more complex models to be easily interpreted. In hindsight, this method would allow for more accurate and consistent feature importance's from the XGBoost models as well as providing more insights about the adaptivity of the features.

There are also considerations regarding the inclusion of macroeconomic variables. In this paper, these variables are included as after differencing and pre-processing. However, an interesting approach that was considered was to use principal components and use these as the input instead. According to Stock and Watson (2002), the benefit is that a large number of predictors easily can be incorporated without increasing the computational complexity and thus generate better forecasts. Moreover, Stock and Watson (2002) shows that using principal components are asymptotically efficient as variables and times series length increases. A drawback is however, that it does not allow for models that benefit from the presence of heteroskedastic and serially correlated uniquenesses (Stock and Watson, 2002). Due to the limited number of macroeconomic variables, this paper does not perform principal component analysis. Although, it would be interesting to use many more macroeconomic variables from a larger API source and perform principal component analysis in terms of organic and conventional foods. Since the forecasting performance was limited, adding more variables through principal components could potentially better explain the difference between organic and conventional foods.

When discussing macroeconomic variables, feature engineering becomes a factor to consider. For instance, considering variables like the global wheat price reported in US dollars, it would make sense to create a new variable combining exchange rates with the reported currency to more accurately reflect the impact in Sweden. However, this insight came later and was consequently omitted due to time constraints.

Aggregation Bias

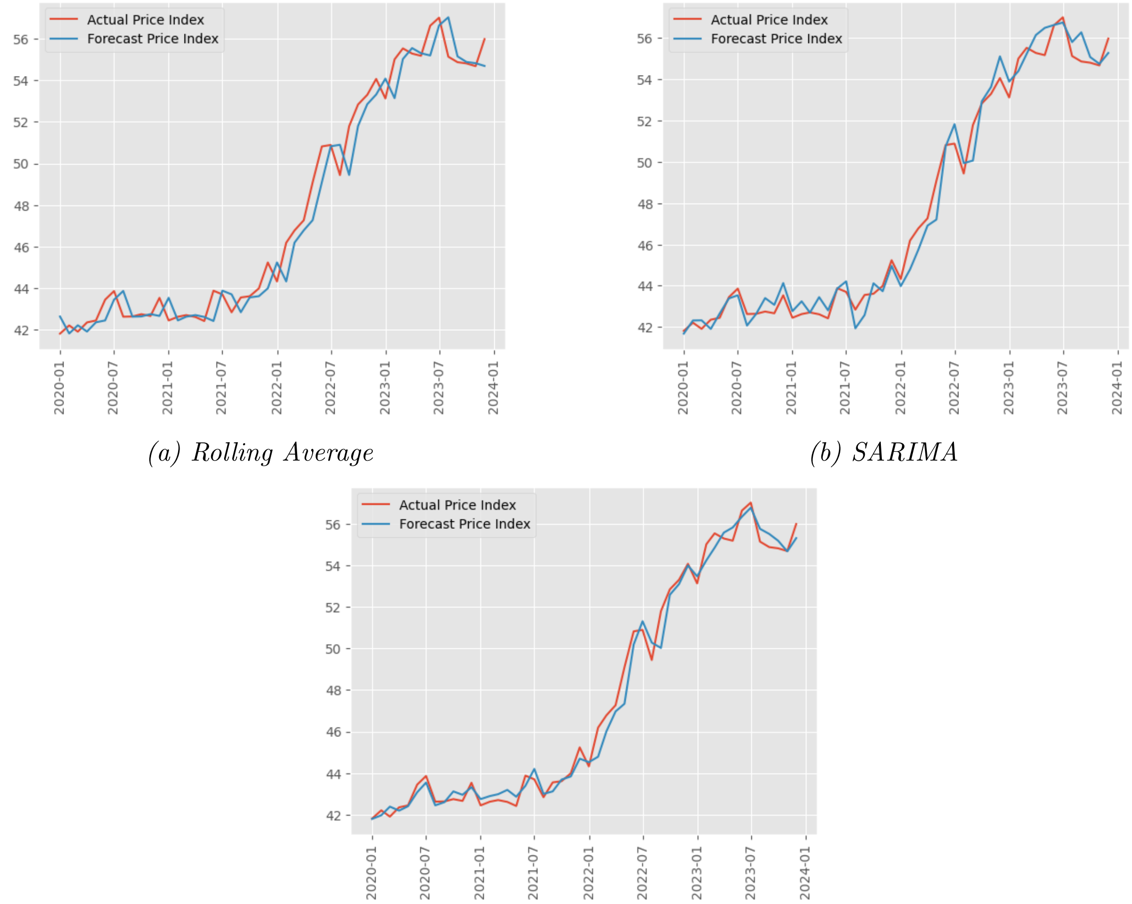
By aggregating the data into a price index, one should be cautious about aggregation bias. It occurs when aggregating data to a higher level if the disaggregate data has different scales, units or if things change in the data that are not accounted for (Luloff et al., 1980). In our data, aggregation is based on value and volume. The value is always reported in SEK, however, as the data aggregation was not conducted by us it is not possible to determine whether volume is always measured in weight (kg) or if there are any discrepancies. This can induce aggregation bias since the price index is constructed with the volume for each category. Another source of potential aggregation bias is the increasing customer base for Matilda Foodtech. Even though it is mainly public sector organizations, it is possible that different organizations and municipalities have different purchasing patterns and purchase different goods which might affect the aggregation accuracy. To somewhat account for this, only customers reporting on a monthly basis are included in this paper. It is still no safeguarding against aggregation bias since the trend is that more and more customers switch to monthly from quarterly reporting during the years analyzed. Aggregation bias can result in the ecological fallacy, where one might assume that what is true on aggregate level also is true for the disaggregates (Piantadosi et al., 1988). In this paper, the bottom up hierarchical approach is used in the statistical models and thus no conclusions are drawn in the opposite direction.

Computational Complexity and Ease of Use

In comparison with statistical time series methods, these are generally constructed to use lagged values within the same column, making it simple to adjust the input data without changing the data frame. In a setting where the data is fairly well known, some assumptions of how seasonality affects the series can make traditional statistical approaches both quicker and less computationally intensive than machine learning based methods. Here, good results are the product of intensive tuning and validation, instead of knowledge and experience from similar data. When considering Januschowski et al.'s (2022) results from forecasting competitions, the best results are found with machine learning methods. Although, if there is a time constraint for model selection, an experienced time series economist could potentially get better results in less time with statistical methods by having more experience and therefore specify the model better. However, domain knowledge is also beneficial with a machine learning approach where the addition of relevant predictors and feature engineering is fundamental here as well. Another modeling approach are pre-trained algorithms like Januschowski et al. (2022) mention about Amazon and Facebook's frameworks. A promising method using a pre-trained model is Googles TimesFM, which is a trained on billions of time series observation online (Upadhyay, 2024). However, this is a brand new and unproven method within this domain.

6.2 Extensions

This section is primarily included to enhance understanding and usability for Matilda Foodtech and public sector stakeholders. To obtain the forecast of the next month in terms of the undifferenced price index, it is necessary to add the forecast being made to the current month's price index. As explained in section 3.2, the indices are created using weights representing the most bought categories between 2019-2021 in terms of value. Furthermore the prices themselves are derived by dividing value by volume. Therefore the values on the y-axis in figure 6.2 represents the average price of one unit in the "average shopping bag between 2019-2021" of the public sector.



(a) Rolling Average

(b) SARIMA

(c) Ensemble

Figure 6.2: Visualizing the undifferenced time series forecasts for the different models.

The average ensemble model in Figure 6.2c is notably "smoothing" out the errors made by the benchmark SARIMA model in Figure 6.2b, both in- and out-of-sample. While it still has a hard time to follow the sudden inflation in 2022, it limits the big misses made by the SARIMA. Most clearly at some peaks such as July 2022 and December 2022, as well as some dips such as in January 2023 and September 2023. Furthermore, it can also be concluded by looking at the plot of the rolling average in Figure 6.2a, that it is useless in any real world situation for Matilda Foodtech.

Another way to analyze the performance and usability of our models is to look at how often they predict the correct sign, i.e. if prices increase, the model should forecast a price increase and the other way around. In terms of sign, the overall accuracy of the average ensemble model out-of-sample is 75%. However, the accuracy on changes smaller than 1% is 44%, while accuracy on changes larger than 1% is 93%. Since small price changes likely affect the clients of Matilda Foodtech less than large changes, it is more important that it correctly forecasts the sign of changes greater than 1%. When changes are greater than 1.5%, which they are half of the time during 2022 and 2023, the model forecasts the correct sign 100% of the time (12 times).

This analysis can be reversed, if the model forecasts a price increase, the prices should actually increase in order for the model to be trustworthy. Again, the overall performance of the average ensemble is 75%, and again bigger changes are more important, and the model has a higher precision on big changes. When the model forecasts a change smaller than 1%, it is correct 50% of the time, but when it forecasts changes greater than 1% it is correct 100% of the time (12 times).

For in-sample forecasts the average ensemble is correct about the sign 79% of the time, which can be compared to the rolling average that is correct 63% of the time. The hierarchical SARIMAX which intuitively would suffer from the most over-fitting is correct in-sample 79% of the time in-sample, 67% of the time out-of-sample when changes are smaller than 1%, and 80% of the time when changes are greater than 1%. The XGBoost model is only correct about the sign-in-sample 63% of the time, making it the only one not beating the in-sample performance of the rolling average. Out-of-sample the XGBoost however has 33% accuracy for changes smaller than 1%, but 100% accuracy for changes greater than 1%.

Again, this analysis suggests that the ensemble approach keeps the strengths of the strongest in-sample models for 2023. At the same time it reduces the errors of such models in 2022 when it instead seems to use the strengths of the "under-fitted" XGBoost. The ensemble model therefore seems to be very well balanced in regards to both MSFE and sign forecasting. The weakness of the ensemble is its low explainability and that it requires more training and computer resources than any single model.

7

Future Research

The results shown in this study might not generalize to other food price indices. Partly because of the delimitation of using Swedish public sector data, and partly because the manner in how the price index was created in this study does not exactly replicate other indices. Therefore, it is imperative to test these methods on additional data before making any substantive claims about their effectiveness in forecasting for example Swedish consumer food price index or Swedish CPIF.

No conclusions about specific macroeconomic variables' influence on food prices can be made from this study. Instead, given the results, selected features are to be seen as guidance or suggestions for future researches wanting to do more inference-focused studies on drivers of food price changes. There are signs of interesting relationships between lags of macroeconomic variables and food prices, both for specific food categories, as well as the aggregated index.

Another area for further research is trying to improve the models used in this study. One concrete suggestion is to engineer features by for example multiplying global wheat prices with the SEK-USD exchange rate. Avoidance of multicollinearity issues for the statistical models was left for LASSO to handle in this study, it could be beneficial to more carefully remove some variables manually using correlation matrices and prior beliefs before leaving the feature set for LASSO to deal with.

As mentioned in previous research, there are many modern forecasting methods such as SVMs and NNs. XGBoost was chosen for this study since it often is one of the best algorithms in forecasting competitions, but other modern algorithms could potentially be better for this data. As time passes and more data becomes available, it might also be possible to improve the SARIMA and SARIMAX models simply by increasing the window size, as the selected window size of 60 periods was the maximum. Especially in the case where macroeconomic variables are used it could be less prone to overfitting.

Matilda Foodtech provided us with both main category and subcategory. A lot is left to explore in terms of hierarchical models. The bottom-up SARIMAX approach that is the best performing single model both in-sample and in 2023, only forecasts unique main category-subcategory combinations, and then immediately aggregates them to the price index. Oftentimes, hierarchical models have more layers and it would be interesting to also forecast main categories, before aggregating all the way up to the price index. Furthermore, a next step could be to combine the category forecasts in the optimal way as suggested by Hyndman et al. (2011).

8

Conclusions

This paper has shown that ensemble methods that include modern machine learning models can improve forecasting performance as previously proven by Ribeiro and Dos Santos Coelho (2020) amongst others. More specifically, this paper has demonstrated that ensemble models improve performance in real-world scenarios where data is not always collected explicitly for researchers or forecasting organizations. Compared to previous research, we however propose ensembles not purely made up of machine learning models. Instead, we have shown that when combining statistical methods with machine learning methods, ensembles can use strengths of the two by capturing both linear and non-linear relationships between macroeconomic variables and food prices. Strengths and weaknesses of hierarchical methods discussed in Schwarzkopf et al. (1988) are also validated in this paper. We see that the best performing category-level models are the ones corresponding to the largest categories, which have the most data and least amount of outliers. The smaller categories are in general much harder to forecast and their models seem to perform poorly.

Although the average ensemble is difficult to interpret, it is possible to explain the behavior within the underlying categories once all the individual models are constructed. Different organizations and researchers such as NIER (2023), Joutz (1997) and Headey and Fan (2008) has shown that prices of input goods affect food prices, which also seems to be the case in this paper. However, different studies point towards the importance of different input goods. This paper also highlights the importance of input goods. However, due to the complexity of food—considering factors like country of origin, the variety of staple foods used in products, and the storage methods of finished goods—different inputs such as electricity prices, wheat prices, diesel prices, and exchange rates have varying levels of significance depending on the time, location, and product involved. This leads to the conclusion that there are no definitive factors that should be consistently used to explain food price inflation at the aggregated level. Instead it depends on hundreds of variables at the lower level of the hierarchy whose significance varies over time.

Bibliography

J. Baek and W. W. Koo. Analyzing factors affecting us food price inflation. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 58(3):303–320, 2010.

J. Baffes. Oil spills on other commodities. *Resources Policy*, 32(3):126–134, 2007. doi: 10.1016/j.resourpol.2007.07.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0301420707000542.

C. Conol. Benchmarking methods for deep learning-based time series forecast, 2020. URL https://medium.com/analytics-vidhya/benchmarking-methods-for-deep-learning-based-time-series-forecast-ec45f78b61e2.

L. Emediegwu. Update: How is the war in ukraine affecting global food prices? *Economics Observatory*, n.d. URL https://www.economicsobservatory.com/update-how-is-the-war-ukraine-affecting-global-food-prices.

W. Enders. *Applied econometric time series*. Wiley, fourth edition edition, 2015. ISBN 978-1-118-80856-6.

Energimarknadsbyrån. Elområden, 2024. URL http://www.energimarknadsbyran.se/el/elmarknaden/elomraden/. [Accessed 29 April 2024].

European Commission. CAP at a glance - european commission, 2024. URL https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-glance_en.

European Union. Commission delegated regulation (eu) 2023/2772 of 31 july 2023 supplementing directive 2013/34/eu of the european parliament and of the council as regards sustainability reporting standards, 2023. URL http://data.europa.eu/eli/reg_del/2023/2772/oj/eng. [Accessed 2 April 2024].

J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab. Forecasting of demand using arima model. *International Journal of Engineering Business Management*, 10:1847979018808673, 2018.

FRED. Federal reserve economic data — fred, 2024. URL https://fred.stlouisfed.org/. [Accessed 26 April 2024].

J. Gastinger, S. Nicolas, D. Stepic, M. Schmidt, and A. Schulke. A study on ensemble learning for time series forecasting and the need for meta-learning. In *2021*

International Joint Conference on Neural Networks (IJCNN), page 1–8, Shenzhen, China, 18 July 2021. IEEE. doi: 10.1109/IJCNN52387.2021.9533378. URL https://ieeexplore.ieee.org/document/9533378/.

C. L. Gilbert. How to understand high food prices. *Journal of Agricultural Economics*, 61(2):398–425, 2010. doi: 10.1111/j.1477-9552.2010.00248.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1477-9552.2010.00248.x.

J. Ha, M. A. Kose, F. Ohnsorge, and H. Yilmazkuday. What explains global inflation. *World Bank Group*, 2023.

M. A. Hall. Correlation-based feature selection for machine learning. Technical report, University of Waikato, 1999. URL https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf.

D. Headey and S. Fan. Anatomy of a crisis: The causes and consequences of surging food prices. *Agricultural Economics*, 39(s1):375–391, 2008. doi: 10.1111/j.1574-0862.2008.00345.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1574-0862.2008.00345.x.

R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589, 2011. doi: 10.1016/j.csda.2011.03.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947311000971.

G. James, D. Witten, T. Hastie, R. Tibshirani, and J. E. Taylor. *An Introduction to Statistical Learning: With Applications in Python*. Springer, 2023.

T. Januschowski, Y. Wang, K. Torkkola, T. Erkkilä, H. Hasson, and J. Gasthaus. Forecasting with trees. *International Journal of Forecasting*, 38(4):1473–1481, 2022. doi: 10.1016/j.ijforecast.2021.06.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207021001679.

F. L. Joutz. Forecasting cpi food prices: An assessment. *American Journal of Agricultural Economics*, 79(5):1681–1685, 1997. doi: 10.2307/1244403. URL https://onlinelibrary.wiley.com/doi/10.2307/1244403.

A. J. Koning, P. H. Franses, M. Hibon, and H. Stekler. The m3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21(3):397–409, 2005. ISSN 01692070. doi: 10.1016/j.ijforecast.2004.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207004000810.

KRAV. Låt oss prata om priset på ekomat, 2022. URL https://www.krav.se/aktuellt/lat-oss-prata-om-priset-pa-ekomat/. [Accessed 2 April 2024].

A. Luloff, P. Greenwood, et al. *Definitions of community: an illustration of aggregation bias, Station Bulletin, no. 516*. New Hampshire Agricultural Experiment Station; Hanover, NH, 1980.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 2017-December, page 4766 – 4775, 2017. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044542379&partnerID=40&md5=78f2bc16fd361e274004b0f78b3ef44a. Cited by: 10158.

M. Marmion, M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, 15:59–69, 2009. doi: 10.1111/j.1472-4642.2008.00491.x.

E. K. Marsh. XGBoost feature importance, 2023. URL https://medium.com/@emilykmarsh/xgboost-feature-importance-233ee27c33a4.

R. Muthukrishnan and R. Rohini. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, page 18–20, Coimbatore, India, October 2016. IEEE. doi: 10.1109/ICACA.2016.7887916. URL http://ieeexplore.ieee.org/document/7887916/.

NIER. Pris och kostnadsutvecklingen 2019-2023: Analys med en prismodell. Technical report, NIER, 2023. URL https://www.konj.se/download/18.311e072818c6242ef9e1f12b/1702630331133/2023-12-18%20Pris-och-kostnadsutvecklingen-2019%E2%80%932023.pdf. [Accessed 29 February 2024].

S. Piantadosi, D. P. Byar, and S. B. Green. The ecological fallacy. *American journal of epidemiology*, 127(5):893–904, 1988.

M. H. D. M. Ribeiro and L. Dos Santos Coelho. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied Soft Computing*, 86:105837, 2020. doi: 10.1016/j.asoc.2019.105837. URL https://linkinghub.elsevier.com/retrieve/pii/S1568494619306180.

A. B. Schwarzkopf, R. J. Tersine, and J. S. Morris. Top-down versus bottom-up forecasting strategies. *International Journal of Production Research*, 26(11):1833, 1988. doi: 10.1080/00207548808947995. URL https://eds-p-ebscohost-com.ludwig.lub.lu.se/eds/pdfviewer/pdfviewer?vid=1&sid=74fa9a97-1eaf-4119-9fec-b2ad6cbd5484%40redis.

Statistics Sweden. Historisk ökning av matpriserna senaste året, 2023. URL https://www.scb.se/pressmeddelande/historisk-okning-av-matpriserna-senaste-aret/. [Accessed 29 February 2024].

Statistics Sweden. Elhandelspriser på elenergi (exkl. skatt och nätavgift) efter avtalstyp, elområde och kundkategori. månad 2013m04 - 2024m03, 2024. URL http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__EN__EN0301__EN0301A/SSDManadElhandelpris/. [Accessed 26 April 2024].

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):

1167–1179, 2002. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214502388618960. URL http://www.tandfonline.com/doi/abs/10.1198/016214502388618960.

Sveriges Riksbank. Riksbanksstudie: Utvärdering av riksbankens prognoser. *Riksbanksstudie*, 2023. URL https://www.riksbank.se/globalassets/media/rapporter/riksbanksstudie/svenska/2023/riksbanksstudie-utvardering-av-riksbankens-prognoser.pdf.

Sveriges Riksbank. Home, the riksbanks' api portal, 2024. URL https://developer.api.riksbank.se/. [Accessed 26 April 2024].

Sveriges Riksbank. Hur mäts inflation?, n.d. URL https://www.riksbank.se/sv/penningpolitik/inflationsmalet/hur-mats-inflation/.

P. Toledo and R. Duncan. Forecasting food price inflation during global crises. *Journal of Forecasting*, n/a(n/a), 2024. doi: https://doi.org/10.1002/for.3061. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3061.

V. Upadhyay. TimesFM: How google's pre-trained model can revolutionize time-series forecasting, 2024. URL https://vivekupadhyay1.medium.com/timesfm-how-googles-pre-trained-model-can-revolutionize-time-series-forecasting-

X. Xu and Y. Zhang. Wholesale food price index forecasts with the neural network. *International Journal of Computational Intelligence and Applications*, 22(04):2350024, 2023. doi: 10.1142/S1469026823500244. URL https://www.worldscientific.com/doi/10.1142/S1469026823500244.

K. Yang, F. Tian, L. Chen, and S. Li. Realized volatility forecast of agricultural futures using the har models with bagging and combination approaches. *International Review of Economics & Finance*, 49:276–291, 2017. doi: 10.1016/j.iref.2016.12.011. URL https://linkinghub.elsevier.com/retrieve/pii/S1059056016301927.

Appendix A

Table A.1: Categories included in index

Category
Frozen pastry with filling
Frozen dessert
Frozen fine bread
Frozen unprepared fish
Frozen fish products
Frozen fruit and berries
Frozen ready meals
Frozen unprepared poultry
Frozen poultry products
Frozen ice cream
Frozen untreated vegetables
Frozen vegetable dish
Frozen coffee
Frozen pastry dough
Frozen pastry bakery
Frozen spices
Frozen meat unprepared
Frozen meat products
Frozen bake-off bread
Frozen ready-made bread
Frozen pancake, waffle, omelette
Frozen pasta dish
Frozen potato product
Frozen puree
Frozen hash and stir-fry
Frozen pie crusts
Frozen seafood
Frozen timbale
Fresh baking accessories
Fresh dessert
Fresh dressing, sauce, soy
Fresh unprepared fish
Fresh canned fish

Continued on next page

Table A.1 – *Continued from previous page*

Category
Fresh fruit and berries
Fresh ready meals
Fresh unprepared poultry
Fresh poultry products
Fresh treated vegetables
Fresh untreated vegetables
Fresh peeled vegetables
Fresh vegetable products
Fresh juice, nectar, drink
Fresh spices
Fresh unprepared meat
Fresh meat products
Fresh mayonnaise and products
Fresh mayonnaise salad
Fresh ready-made bread
Fresh cooking fats and oils
Fresh dairy substitutes
Fresh dairy products
Fresh cheese
Fresh potatoes
Fresh potato products
Fresh caviar products
Fresh shellfish
Fresh main course sauces
Fresh eggs
Fresh egg products
Colonial baking mix
Colonial baking accessories
Colonial legumes
Colonial baby food
Colonial stock, fond, glaze
Colonial dessert
Colonial dressing, sauce, soy
Colonial canned fish
Colonial breakfast cereals
Colonial fruit and berries
Colonial fruit and berries canned
Colonial seeds and kernels
Colonial ice cream accessories
Colonial grains
Colonial canned vegetables
Colonial vegetable dish
Colonial dry bread and baked goods
Colonial juice, nectar, drink

Continued on next page

Table A.1 – *Continued from previous page*

Category
Colonial coffee
Colonial pastry bakery
Colonial carbonated drink
Colonial confectionery
Colonial spices
Colonial ready-made bread
Colonial cooking fats and oils
Colonial dairy substitutes
Colonial dairy products
Colonial flour
Colonial pasta product
Colonial potato product
Colonial mashed potato powder
Colonial thickener, starch
Colonial rice
Colonial salt
Colonial mustard, tomato products
Colonial snacks
Colonial sugar, sweeteners
Colonial soup, stew, powder
Colonial jam, marmalade, jelly
Colonial sauce for main dishes
Colonial tea, chocolate drink
Colonial vinegar, cooking wine
Nutritional supplement
Nutritional products

Variable	**p-value**
sekeurpmi	0.0026
sekusdpmi	0.0
sekkix92	0.0154
secbrepoeff	0.0
setb1mbenchc	0.0
emgvb5y	0.0
POILBREUSDM	0.0
PCU325311325311	0.0
PNRGINDEXM	0.0
PFOODINDEXM	0.0
SWECPICORMINMEI	0.0017
CPGREN01SEM657N	0.0
CSESFT02SEM460S	0.0
WPS057303	0.0
CP0111EU272020M086NEST	0.0004
PWHEAMTUSDM	0.0
ENRGY0EU272020M086NEST	0.0
CP0450EU272020M086NEST	0.0
avg_electricity_price_SWE	0.0

Table A.2: P-values from ADF-test of first differenced macroeconomic variables 2015-2021

Variable	**p-value**
sekeurpmi	0.0
sekusdpmi	0.0
sekkix92	0.0
secbrepoeff	0.2635
setb1mbenchc	0.2489
emgvb5y	0.0084
POILBREUSDM	0.0
PCU325311325311	0.0079
PNRGINDEXM	0.0086
PFOODINDEXM	0.0
SWECPICORMINMEI	0.383
CPGREN01SEM657N	0.0001
CSESFT02SEM460S	0.0
WPS057303	0.0042
CP0111EU272020M086NEST	0.1413
PWHEAMTUSDM	0.0
ENRGY0EU272020M086NEST	0.0404
CP0450EU272020M086NEST	0.0002
avg_electricity_price_SWE	0.0

Table A.3: P-values from ADF-test of first differenced macroeconomic variables 2015-2023