Food Index Forecasting

Kalyani Dacha, Ramya Cherukupalli, and Abir Sinha

1 Business Problem

Time series forecasting techniques are important to improve business process, increase efficiency, profits and reduce costs. Forecasting in any business is important because it provides an insight about the direction and the volume business is heading towards.

In this paper, we are discussing business problems of one of the leading food service providers in North America which operates in industries like education, health care, sport and entertainment, and business and government. They do not buy live animals (e.g., hogs, cattle, poultries) to make the finished product. Rather, they buy raw products and make intermediate products out of them (e.g., ground beef, steak, bacon, liquid eggs) from several food vendors. The company signs long term contracts with the food vendors either for 12 months or for 18 months in advance. Hence, it becomes critical to understand where the inflation numbers for each individual food category (U.S Bureau of Labor Statistics, BLS published non-adjusted numbers) are heading toward. Currently, they guess the average inflation number for the upcoming 12 or 18 months for each individual category, adjust their last year inflation numbers in the contract and renew it. So, to make the contract efficient and profitable for them, it is necessary to estimate the upcoming average inflation numbers effectively. We were provided with a list of 18 major food categories that contribute to 90% of their food supply business.

K. Dacha (✉) · R. Cherukupalli · A. Sinha
Deloitte Consulting, Hyderabad, India
e-mail: kalyani.dacha@gmail.com

R. Cherukupalli
e-mail: ramyac1@gmail.com

A. Sinha
e-mail: abirsinha1@gmail.com

A. K. Laha (ed.), *Applied Advanced Analytics*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-981-33-6656-5_11

Table 1 List of product categories

S. No.	Category	Sub-category
1	Beef	Ground beef
2		Beef roast
3		Beef steak
4	Pork	Bacon
5		Ham
6		Pork chops
7	Poultry	Chicken
8		Fresh and frozen chicken
9		Turkey
10	Eggs	Eggs
11	Potatoes	Potato
12		Frozen and freeze-dried prepared foods
13		Potato fries (PPI)
14	Fish and seafood	Fish and sea food
15	Dairy	Milk
16		Cheese
17		Coffee
18	Fats and oils	Fats and oils

Price index forecasting was used to forecast monthly inflation numbers for 18 product categories namely (Table 1):

2 Data Gathering

All the information used for this analysis has been downloaded using publicly available, free, external websites.

Data gathering was primarily carried out for the dependent variable (Consumer Price index). BLS website stores monthly price index/inflation numbers for over 100 food items, and updates are carried out each month. Extensive research was carried out to find out what can drive price for each product category. Instead of looking into the individual sub-categories (like ground beef, beef roast and beef steak)—research was focused on the overall categories (like what affects beef prices in this example).

We can broadly categorize key drivers across major product categories into five different levels (Fig. 1):

Few insights from the research on affecting price trends:

1. It was observed that feed expenses were high in 2012, and huge number of cows were slaughtered (Frohlich 2015) as an immediate impact. In the next

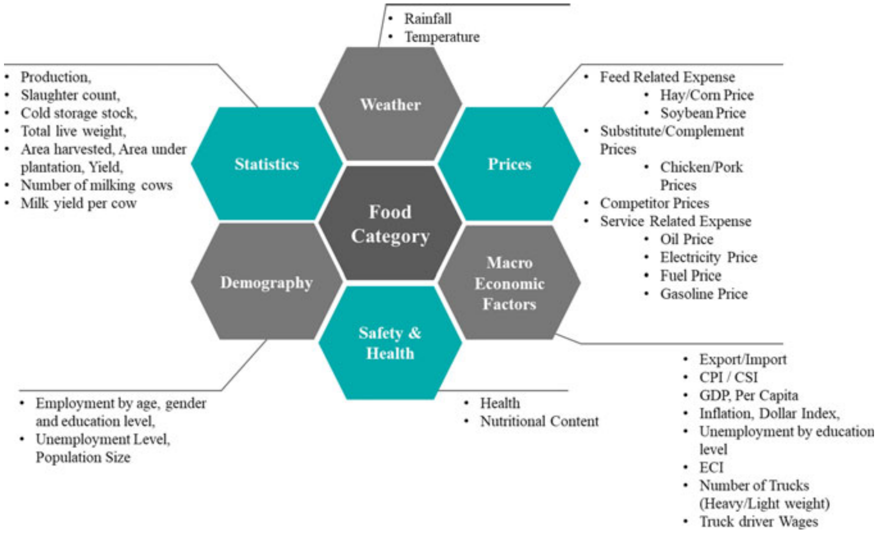

Fig. 1 Key drivers across food categories

two years, number of meat producing cows were quite less in number, and it significantly affected beef price in 2014. Hence, feed prices (hay, corn, soybean) were identified as one of the key drivers

2. Butterfat percentage: Milk (Bailey 2017; Cushnahan 2003) butterfat content lowest at peak production and highest toward the end of lactation
3. Biofuels (Rosillo-Calle et al. 2009; Parcell et al. 2018),[1] soaps, washing powders, personal care products have close interdependence to oilseed and biodiesel markets. Any changes in either US biodiesel policy or global biodiesel policy could shock oilseed prices
4. El Niño[2] is an abnormal weather pattern caused by the warming of the Pacific Ocean near the equator, off the coast of South America. In South America, there is a drastic increase in the risk of flooding on the western coast, while there is an increase in the risk of droughts on parts of the eastern coast. In eastern countries, like India and Indonesia, there is an increase in droughts. These affect the fish and seafood (OECD/FAO 2016) price index a lot.
5. The main drivers for decline of price of the commodities will be the competitive prices of substitutes (like eggs, chicken, etc.) the slowdown in demand from key markets due to sluggish economic growth and reduced production and marketing costs of aquaculture products due to lower transport and feed costs
6. Political situations in Brazil, Indonesia affects the coffee (van den Brom 2020)[3] price most as they are the leading coffee producers.

[1] An overview of the Edible Oil Markets: Crude Palm Oil vs Soybean Oil" (July 2010).

[2] Rinkesh, "What is El Niño?".

[3] Jack Purr, "What affects the price of coffee.

Data for independent variables was downloaded from various data sources for conducting multivariate time series analysis. Bureau of Labor Statistics (BLS) was one of the main data sources. Other sources include Federal Reserve Economic Data (FRED), United States Department of Agriculture (USDA), National Oceanic and Atmospheric Administration (NOAA), Data World, National Agricultural Statistics Service (NASS).

o CPI index and unemployment rate were mainly sourced from BLS,
o Import/export from USDA,
o Temperature and rainfall data from NOAA.

3 Model Development

After collecting the data for dependent and independent variables, the first task was to collate the data in a data frame so that it can be further processed for modeling. Data from different sources was collated, and time series dataset starting from January 2009 was created. The area of interest in this study was to model the Year-over-Year (Y–o-Y) inflation numbers for all the dependent variable categories. Due to the volatile nature of Y–o-Y variable, modeling was done on the actual index data for all the food categories, and then finally, the forecasted numbers were converted into Y–o-Y for further analysis.

3.1 Pre-modeling

Given this is a time series dataset, it is highly likely that lagged version of independent variables might influence the dependent variables the most. Given we are trying to forecast for future months, lagged independent variables can be directly used (given data is available), else we can also forecast directly. After the creation of lags for each of the independent variable, correlation with the dependent variables was calculated for all the 13 versions of each independent variable, i.e., original variable (Lag 0) and its 12 lagged versions. For example,

- For Ground Beef Price Index, petrol price (six months' lag) correlated the most.
- Beef slaughter count lag two variable was the highest correlated variable with Ground Beef Index with correlation −0.78.
- Lag 1 of shell egg import 1000 dozen is the highest correlated variable with Egg Index with correlation 0.34.

Data considered for further analysis was in the period Jan 2010–May 2018, and there were no missing values within this time frame. For providing price index forecasts, five modeling techniques (univariate and multivariate) were applied. Best model was chosen from the five techniques based on the applicability, accuracy and

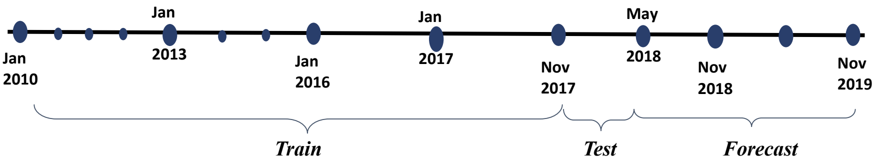

Fig. 2 Modeling timeframe

blind validation. Forward forecast inflation (average and interval range) for 12 and 18 months is provided as results.

3.2 Modeling Process

After identifying all the highly correlated variables for each of the 18 dependent variables separately, next task was to run models to identify the significant variables for each of the dependent variable category.

Train, test and forecast periods were created as given in Fig. 2 throughout the modeling process:

To define train and test period of the study, several test periods were taken into consideration like: 12 months, 9 months and 6 months, respectively. While observing the pattern of the test period (12, 9 and 6 months), it was seen that for most of the categories, distribution of the test set was completely different from that of train period, and as a result, the forecasts were going in a completely different direction. To avoid this situation, test periods were reduced to six months to better train the time series models. To compare the performance of various models, **MAAPE** was used. Below is the definition and reasoning behind using this metric.

3.3 Evaluation Metrics

In time series analysis, mean absolute percentage error (MAPE) is widely used and is calculated as below.

$$\text{MAPE} = \frac{100\%}{n} \sum \frac{|\text{Actual} - \text{Forecast}|}{|\text{Actual}|}$$

In this case study, Y–o-Y is the dependent variable, and this can take both positive and negative numbers, and using MAPE as a model evaluation criterion was found to be not applicable as errors were huge. Example is shown below:

Table 2 Comparison of MAAPE across test periods

Model #	Test period	YoY Train MAAPE (%)	YoY Test MAAPE (%)
1	12	18	83
2	9	21	40
3	6	20	**22**

- Assume for the forecast month June 2018, for the best model, the estimate is 2.50, whereas the actual number is 0.50. In this case, the MAPE would be |0.50 − 2.50|*100/|0.50| = 400% which is not reflecting the actual scenario.
- Hence, alternative model evaluation criteria MAAPE (Kim and Kim 2016) was used as it is applicable to deal with positive or negative numbers where as MAPE was not. MAAPE is calculated using the below formula:

$$\text{MAAPE} = \frac{100\%}{n} \sum \arctan\left(\frac{|\text{Actual} - \text{Forecast}|}{|\text{Actual}|}\right)$$

- For the above example, MAAPE is 132%. Another advantage is MAAPE is well defined even if Y–o-Y is zero though MAPE is not.
- Table 2 describes different test periods and the corresponding MAAPE for one of the food categories index.

The modeling techniques used in this paper is given in Table 3.

- Univariate models were run on the training set, i.e., directly considering the dependent variables as time series. Then, the fitted models were used to forecast the test period to check for the accuracy of the model. Accuracy was calculated using the six original test data points versus six forecasted data points. Once the model was finalized from the accuracy metric, model parameters were retrieved from the best model, retrain the model using the same set of parameters, but with the data of both train and test and finally forecast for the final 18 months. For example, for chicken fresh and frozen category—say ARIMA parameters $p = 2$, $d = 1$, $q = 0$ was finalized. Using the parameters, forecast for the upcoming 18 months will be developed using ARIMA parameters $p = 2$, $d = 1$, $q = 0$ on the entire train + test (95 + 6 = 101 months) data.

Table 3 Modeling techniques used

Method	Technique
Univariate	ARIMA (Box and Jenkins 1970)
Univariate	Holt-Winters (Chatfield and Yar 1988)
Univariate	Exponential smoothing (Broze and Mélard 1990)
Multivariate	Regression (Ramcharan 2006)
Multivariate	ARIMAX (YuanZheng and Yajing 2007)

- For multivariate time series models, independent variables were forecasted first to get the final forecast numbers for the dependent variables. For each of the dependent categories, regression models were run using independent variables in the training period. Once the model variables were finalized, next step was to forecast for the test period. This was done in a two step processes. First was to forecast for the independent variable, and once this was complete, next step was to get the forecast numbers for the dependent variable from the regression equation. Forecast of the independent variables was done using the three univariate time series methods (listed in Table 3), and selected method was the one for which the test MAAPE was the least. Once all the forecast numbers were handy for all the independent variables, regression equation was used to get the final forecast numbers for the dependent variables.
- For ARIMAX model, same set of independent variables and their corresponding forecasted numbers were used. To select the optimal parameters of the ARIMAX model, grid search was carried out.

4 Results

Out of the 18 models built for the study, we have achieved greater accuracy (<35% MAAPE) for 50% of the models, and error range for the rest of the models was between 40 and 60%.

Among the 18 food categories, majority were stable (Fig. 3) with the overall Y–o-Y range between ±5%, few categories like eggs (Fig. 4) are a volatile category with the Y–o-Y range between ±40%, however, our models were robust, and we achieved great accuracy in such scenarios too.

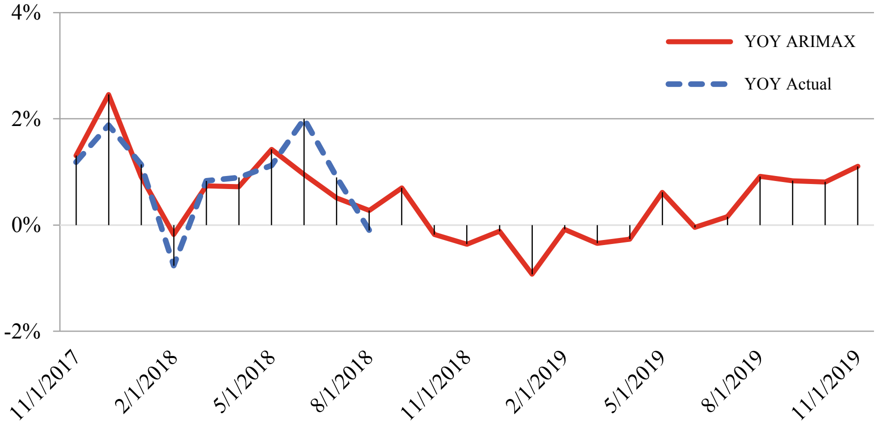

Fig. 3 Figure showing Y–o-Y plots comparison between Actual (Historical Cheese values shown as blue dotted line) versus ARIMAX (forecasted values based on the model selected based on performance as red line)

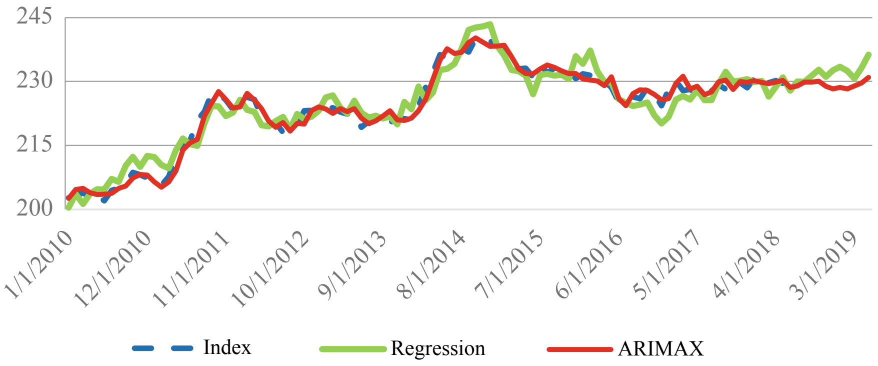

Fig. 4 Price index comparison between Actual (Historical Cheese values shown as blue dotted line) versus ARIMAX (forecast values based on ARIMAX model as red line) versus Regression (forecast values based on regression model as green

Below is the summary report of 18 models built with details on final model selected (based on performance); MAAPE metrics for train and test periods are provided in Table 4. ARIMAX had been predicting better than rest of the models in most categories. Beef category predictions were 70–80% accurate, and these were the main categories. Of all categories, dairy and poultry are having greater accuracy. Except for pork and potatoes category, rest all category forecasts were in the range of 60–80%. Pork and potatoes categories were highly volatile, and the greatest accuracy achieved in these cases is about 50%.

5 Conclusion

In this study, we have tried to address food price forecasting using various univariate and multivariate techniques at monthly level. More than 60 explanatory variables were tested for each category based on extensive research for forecasting consumer price indexes of 18 food categories. The forecasting performance of the model is measured using MAAPE, and accuracy achieved for most of the models is <15%. This price-forecasting model is useful in capturing economic demand-pull factors such as food use, substitute prices, feed prices, weather, macro-economic factors and income in the food price changes. All the data used for the analysis is using publicly available, external data. The approach used here is robust—as we were able to capture trend for highly volatile category and stable category likewise and obtain satisfactory performance.

Table 4 Detailed results for all the categories

Model #	Category	Sub-category	Final model	Train MAAPE (%)	Test MAAPE (%)
1	Beef	Ground beef	ARIMAX	20.32	22.31
2		Beef roast (round)	ARIMAX	25.19	34.29
3		Beef steak	ARIMA	27.66	46.24
4	Pork	Pork bacon	ARIMAX	34.22	58.42
5		Pork ham	ARIMAX	29.45	53.04
6		Pork chops	ARIMAX	33.16	46.78
7	Poultry	Chicken	ARIMAX	28.02	28.32
8		Fresh and frozen chicken parts	ARIMA	36.66	25.98
9		Turkey	ARIMAX	40.42	36.34
10	Eggs	Eggs	ARIMAX	36.72	35.37
11	Potatoes	Potatoes	ARIMAX	37.55	111.84
12		Frozen and freeze-dried prepared foods	ARIMAX	23.28	59.62
13		Potato fries [PPI]*	ARIMAX	46.49	16.23
14	Fish and seafood	Fish and sea food	ARIMAX	28.02	34.28
15	Dairy	Milk	ARIMAX	27.09	25.20
16	Coffee	Coffee	ETS	27.68	31.74
17	Dairy	Cheese	ARIMAX	26.77	28.65
18	Fats and oils	Fats and oils	ARIMA	33.19	43.89

References

Bailey, H. (2017, September). Dairy risk-management education: Factors that affect U.S farm-gate milk prices.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day Inc.

Broze, L., & Mélard, G. (1990). Exponential smoothing: Estimation by maximum likelihood. *Journal of Forecasting, 9,* 445–455.

Chatfield, C., & Yar, M. (1988). Holt-winters forecasting: Some practical issues. *Journal of the Royal Statistical Society. Series D (The Statistician), 37*, 129–140.

Cushnahan, A. (2003, April). Factors influencing milk butterfat concentration.

Frohlich, T. C. (2015, April). States killing the most animals for food. https://www.usatoday.com/story/money/business/2015/04/15/247-wall-st-states-killing-animals/25807125/.

Kim, S., & Kim, H. (2016, September). A new metric of absolute percentage error for intermittent demand forecasts.

OECD/FAO. (2016). Fish and seafood. In *OECD-FAO Agricultural Outlook 2016–2025*.

Parcell, J., Kojima, Y., Roach, A., & Cain W. (2018, January). Global edible vegetable oil market trends.

Ramcharan, R. (2006). Regressions: Why are economists obessessed with them? Accessed 2011–12–03.

Rosillo-Calle, F., Pelkmans, l., & Walter, A. (2009, June). A global overview of vegetable oils, with reference to biodiesel.

van den Brom, J. (2020). Coffee.

YuanZheng, W., & Yajing, X. (2007). Application of multi-variate stable time series model ARIMAX. *Statistics and Operation, 9B*, 132–134