AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Predicting Food Prices using Machine Learning with Data from Consumer Surveys and Search

Jisung Jo[ϕ], Jayson L. Lusk[†], Michael K. Adjemian[*], Sewon Kim[Ψ], Jinho Jung[†], Nicole O. Widmar[†]

ϕ Department of Logistics and Maritime Industry Research, Korea Maritime Institute, South Korea
† Department Agricultural Economics, Purdue University, USA
∗ Department of Agricultural & Applied Economics, University of Georgia, USA
Ψ Department of Intelligence Mechatronics, Sejong University, South Korea

Abstract

- This study explores whether unconventional consumer-oriented variables can be useful in predicting the Bureau of Labor Statistics (BLS) Food and Beverages Consumer Price Index (CPI).
- We examine the ability of an Internet search-based index related to food prices (the Google trends index) and a survey-based consumer sentiment index to predict changes in food-related BLS prices from January 2004 to July 2015.
- Based on moving window and expanding window schemes, we compare several consumer-oriented forecast models and machine learning models.
- Results show that
  - A vector autoregression (VAR) model has the best predictive performance with the moving window structure and a vector error correction model (VECM) performs best with the expanding window structure.
  - Encompassing tests reveal that our model out-predicts USDA Economic Research Service food-related CPI forecasts

Introduction

- Changes in food prices can have an important impact on household well-being.
- Coupling food price volatility with the fact that food is purchased frequently implies that consumers may be more aware of or attentive to changes in the price of food than with other items.
- **Objectives of this research**
  - Exploration of whether forward-looking variables such as Index of Consumer Sentiment (ICS) and Google Trend Index (GTI) improve the performance of Food and Beverage CPI forecast models
  - Comparison of the forecast performance of our models utilizing ICS and GTI data with the forecasts released by the USDA Economic Research Service.
    - Time Series models: Autoregressive Moving Average (ARIMA), Vector Error Correction (VAR), and Vector Error Correction with Exogenous Variables (VAR-X).
    - Machine Learning models: Support Vector Regressor (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF).
- **Contribution of this research**
  - Examination of improvement in food price forecasting with novel datasets.
  - First to adopt forward-looking variables such as survey based ICS and search based GTI for forecasting food prices.
    - Use forward looking variables other than backward looking ones
  - First to adopt machine learning estimation strategy for food price forecasting
    - Improve accuracy in forecasting food price index

Data

- Food-related CPI
  - CPI from the US Bureau of Labor Statistics.
  - The market basket of goods and services reflected in the CPI can be separated into eight categories: food and beverages, housing, apparel, transportation, medical care, recreation, education and communication, and other goods and services.
    - From 2011 to 2012, the relative importance of the food and beverage component in the CPI-U was 14.9 out of 100.
  - This research investigates the movement of the Food and Beverages CPI-U with reference base, 1982-84=100.
- Consumer Sentiment Index (ICS)
  - The University of Michigan has reported monthly ICS data since 1978, and the reference base is March 1997.
- Search-Based Index (GTI)
  - Google Trends provides a measure of the popularity of terms for which Google users have searched over time.
  - The index of Google Trends measures the number of searches conducted for a particular term, relative to the total number of searches done on Google over time.
  - In this study, we create an index based on the search term “food prices” for the USA.

Estimation Strategies

- Methodology
  - Comparison of models to examine a degree of improvement in accuracy of Food CPI.
- Time Series models
  - ARIMA(p,d,q)

$$\Delta lnFCPI_t = \theta_0 + \sum_{i=1}^{p} \emptyset_i \, \Delta lnFCPI_{t-i} + \varepsilon_t - \sum_{k=1}^{q} \rho_j \varepsilon_{t-j}$$

$\Delta lnFCPI_t$ is the first differenced Food category’s Consumer Price Index, $\Delta lnFCPI_{t-i}$ is the first differenced $i\,th$ lags of $\Delta lnFCPI_t$, and ε_t is the stochastic error term.

  - VAR(p)

$$x_t = A_0 + \sum_{i=1}^{p} A_i x_{t-i} + e_t$$

where x_t equals to $[lnFCPI_t, lnAllCPI_t, lnGTI_t, lnICS_t, lnCCS_t]^T$ where $lnFCPI_t$ a log of Food CPI, $lnAllCPI_t$ is a log of CPI with all items, $lnGTI_t$ is a log of Google Trend Index, $lnICS_t$ is a log of Index of Consumer Sentiment, and $lnCCS_t$ is a log of Consumer’s Confident Index.

  - VAR-X

$$x_t = A_0 + \sum_{i=1}^{p} A_i x_{t-i} + \sum_{i=1}^{q} B_i y_{t-i} + e_t$$

where x_t is same as VAR model and y_t is a $(n\times 1)$ vector of exogenous variables.
- Machine Learning models
  - SVM: a supervised learning algorithm used for regression analysis. The goal is to find the best possible weight ‘w’ that can fit the given data points in the best possible way.
  - KNN: a supervised learning algorithm used for regression analysis. The principal of K-Nearest Neighbors Regressor is to predict the value of a target variable by finding the k-nearest neighbors of a given data point in the training dataset and using their average or median value as the predicted value.
  - RF: am ensemble learning method that combines multiple decision trees to make more accurate predictions.

Results and Discussions

- Weak exogeneity test
  - Based on the sequential reduction method of weak exogeneity, we exclude lnICS and lnCCI in the VAR model and include lnICS and lnCCI as the exogenous variable in the VAR-X model. This suggests that the search based index, lnGTI, performs better in predicting the Food CPI than the survey based index lnICS or lnCCI.
- Based on the moving window and the expanding window versions of rolling windows, we evaluate the forecasting performance of the resulting models (Table 1).
- Forecast encompassing test
  - A comparison of the forecasting performance of the Food CPI forecast model conducted by the root mean square error (RMSE) and the mean absolute percentage error (MAPE)
  - Following Fair and Shiller (1989), comparison between time series and machine learning models can be conducted.

$$lnFCPI_t = \alpha + \lambda_1 f_{1t} + \lambda_2 f_{2t} + v_t,$$

If we are able to reject $H_0: \lambda_1 = 0$, then it would indicate redundancy of f_{2t}. That is, the f_{1t} forecast encompasses the f_{2t} forecast.
  - Under the moving window scheme, we reject both null and alternative hypotheses, which means that the combined VARX forecast and RF forecast would provide a better forecast information (Table 2).
  - For the expanding window scheme, we reject the null hypothesis of $H_0: \lambda_1 = 0$ and fail to reject the alternative hypothesis of $H_1: \lambda_2 = 0$, which means that the VAR forecast encompasses the SVM forecast (Table 2).

Table 1. 1-Step Ahead Food and Beverage CPI Forecasting Comparison

Moving Window Scheme		RMSE	MAPE
Time Series Model	ARIMA	0.003	0.03897
	VAR w/ all variables	0.00296	0.036731
	VAR	0.00303	0.041033
	VAR-X	0.00281	0.037667
ML Model	SVM	0.11674	1.716776
	KNN	0.11078	1.601377
	RF	0.0894	1.306346
Expanding Window Scheme		RMSE	MAPE
Time Series Model	ARIMA	0.00303	0.039686
	VAR w/ all variables	0.00285	0.036371
	VAR	0.00303	0.040973
	VAR-X	0.00286	0.038521
ML Model	SVM	0.07568	1.10635
	KNN	0.07948	1.149681
	RF	0.08696	1.280225

Table 2. Encompassing Test by Moving Window Scheme and Expanding Window Scheme

Moving Window Scheme		
Models	t-value	Pr >t
VARX(2,1)	128.94	<.0001***
RF	-2.86	0.0058***
Expanding Window Scheme		
Models	t-value	Pr >t
VAR(4)	202.21	<.0001***
SVM	0.82	0.4142

Conclusion

- We examine whether unconventional consumer-oriented measures improve the accuracy Food and Beverages Consumer Price Index (CPI) predictions.
- The exogeneity test suggests that the consumer sentiment indicator ICS does not react to disequilibrium, and thus there is no information loss even if the ICS is excluded.
- Preliminary comparison shows that VAR and VECM are the preferred models with the moving window and expanding window scheme, respectively.
  - Thus, the models assuming GTI and CPI as endogenous variables best predicts the Food and Beverage CPI.
- The encompassing test shows that the consumer oriented VECM encompasses the information contained in the USDA ERS forecast, suggesting that accuracy could be improved by including Google search data.
- The suggested forecast model with consumer-oriented variables, which can predict food prices more accurately.