NCCC-134
APPLIED COMMODITY PRICE ANALYSIS, FORECASTING AND MARKET RISK MANAGEMENT

Forecasting Crop Prices using Leading Economic Indicators and Bayesian Model Selection

by

Yu Wang and Jeffrey H. Dorfman

Suggested citation format:

Wang, Y. and J. H. Dorfman. 2018. "Forecasting Crop Prices using Leading Economic Indicators and Bayesian Model Selection." Proceedings of the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. Minneapolis, MN. [http://www.farmdoc.illinois.edu/nccc134].

Forecasting Crop Prices using Leading Economic Indicators and Bayesian Model Selection

Yu Wang and Jeffrey H. Dorfman[1]

Paper presented at the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management Minneapolis, Minnesota, April 16-17, 2018.

[1] Yu Wang and Jeffrey Dorfman are a Ph.D. student and professor in the Department of Agricultural and Applied Economics at The University of Georgia.

Forecasting Crop Prices using Leading Economic Indicators and Bayesian Model Selection

Corn, wheat and soybeans are very important to the US agricultural sector as the main sources of many farmers' income. Thus, forecasting the prices of these three crops is important. When considering model specification of crop price forecasting models, this paper focuses on potential benefits from including leading economic indicators variables, both those clearly related to agriculture such as the crude oil price and interest rate and those not clearly related such as the purchasing managers index or the S&P500 stock price index. To do this, our paper tests whether leading economic indicators can be used to improve the forecasts of corn, wheat, and soybean future prices. We take a Bayesian approach to estimate the probability that a set of leading indicators belong in the forecasting model where specification uncertainty is explicitly modeled by assuming a prior distribution over a very large set of models. Model specifications considered vary by different lag lengths for leading indicators and crop prices as well as which variables are included at all. We apply this method to corn, soybean and wheat monthly spot price data from 1985 to 2016. The results show that several leading economic indicators appear to be useful for forecasting crop prices.

Key words: crop prices, forecasting, leading economic indicators, model specification uncertainty.

Introduction

Corn, wheat and soybeans constitute a major share of the American agricultural economy. The changes in prices of these three crops can affect farmers' production decisions across much of the United States. Corn and wheat with barley are the largest source of feed grains used in the livestock industry (Westcott and Hoffman, 1999). In addition, United States became the largest exporter of corn and soybeans in 1996 and has maintained a position at or near the top ever since. Thus, forecasting the prices of these major crops is important both for farmers to make optimal planting decisions and for many other participants in commodity markets who utilize price forecasts in daily business decisions and risk management actions. Given the importance of forecasting the

prices of these crops, we propose that perhaps forecasters should cast a wider net when considering the variables that belong in their forecasting models.

While in the past, such models have been confined to variables closely related to agricultural production or commodity demand, the recent positive correlation of agricultural commodity prices with a general set of financial asset prices suggests that more general variables measuring economic conditions might be worth including. Based on this motivation, we here explore the potential value of adding several leading economic indicators to agricultural price forecasting models.

In this paper we apply a Bayesian approach that formally recognizes model specification uncertainty to estimate the probability of the inclusion of five different leading economic indicators and to assess the number of lagged values that are best to include in a crop price forecasting model. We consider 4,096 potential model specifications and estimate the posterior distribution of belonging in the model for each leading indicator at three lags. We also assess how much inclusion of such variables improves our forecasting performance as measured by mean squared error.

We apply this approach to spot price data on corn, soybeans and wheat. We find non-trivial marginal probabilities (> 0.2) for the inclusion of all five leading economic indicators considered and fairly strong evidence (probabilities of > 0.4) for the inclusion of some number of lagged values for the money supply, the purchasing managers index, and the interest rate.

Methodology

There is a long history of testing model specification for the purposes of both modeling and forecasting major crop prices and acres planted. For example, time series models, both ARIMA and VAR models, were used to test different lag lengths of indicators which can affect corn price in Allen (1994). Westcott and Hoffman (1999) apply linear log regression to test how market factors and government programs affect price for corn and wheat in U.S. Westcott and Hoffman find that corn total stocks-to-use and loan-rate affect corn prices, while wheat competitor stocks-to-use, wheat feed use and loan-rate affect wheat prices. Both regressions can predict the prices for corn and wheat with high

accuracy. Baba and Narain (2011) propose a data-based algorithm to select a subset of leading indicators from a large set. They develop a simple linear relationship between the forecasted price and leading indicators. Baba and Narain (2011) evaluate the accuracy of their forecasting model based on both the Quadratic Probability Score (QPS) and the Log Probability Score (LPS). They also apply the Clements and Harvey Test to pick the best model and test whether the best model encompasses other models. Baba and Narain's (2011) algorithm chooses the Ratio of Index of Help-Wanted Advertising to Number of Persons (LHELX), Month's Supply at Current Sales Ratio (HNR) and Housing Starts, Private Including Farm (HUSTS1) as the best variables to predict the price in the short run. In long run, the algorithm chooses four totally different variables as the best predictors of prices. Chiaie, Ferrara, and Giannone (2017) forecast commodity prices using a dynamic factor models. The authors decompose the price series into a global component, blocks of commodities and an idiosyncratic component. Chiaie, Ferrara, and Giannone (2017) develop a modified version of information criterion (IC) to select the best factors to model commodity prices. The authors find that non-fuel and food and beverage factors are the most important factors to forecast changes in commodity prices. Clearly, the literature has not reached consensus on the correct model specification for crop price forecasting.

Given this uncertainty of model choice, Dorfman and Sanders (2004) present a Bayesian approach to a generalized hedge ratio estimation under model uncertainty. We follow a similar approach, but for the forecasting of corn, soybean, and wheat prices. We consider a range of exogenous variable types to augment standard lagged prices and more commonly used variables within our set of models.

Bayesian Model Construction

We define the set of potential models as $M = \{J_i, i = 1, \dots J\}$. For each possible model specification, we develop the general model form as:

$$y = X_i\beta_i + \varepsilon_i, \text{ for } i = 1,2, \dots J. \quad (1)$$

In equation (1), i is the model index, denoting the ith model, y is the vector of dependent variables which does not vary when the model specification changes, X_i is the matrix containing the independent variables (including different lags of some

variables) and lagged dependent variables, and ε_i is the error term vector in the ith model. The number of independent variables and the lengths of their lags differ over models, but the dependent variable (in both form and number of observations) stays the same.

To perform Bayesian estimation of these models, we need prior distributions for all the parameters. Following Dorfman and Sanders (2004), the prior distribution of β_i is defined as:

$$p(\beta_i) \sim N(\mu_{0i}, \sigma_i^2 \Sigma_{0i}),\ i=1,2,\ldots J \quad (2)$$

where N denotes the multivariate normal distribution, μ_{0i} is the prior mean of β_i and $\sigma_i^2 \Sigma_{0i}$ is the prior variance matrix of β_i in ith model. Using a standard Bayesian framework, the inverse of σ_i^2 has a prior distribution of:

$$p(\sigma_i^{-2}) \sim \mathrm{G}(\gamma_{0i}^{-2}, D_{0i}),\ i=1,2,\ldots J \quad (3)$$

where G denotes the gamma distribution, γ_{0i}^{-2} is the prior mean for σ_i^{-2}, and D_{0i} is the prior degrees of freedom for the inverse error variance. A higher value of D_{0i} indicates that the prior distribution is more informative.

Given the distributions above and assuming the error term is normally distributed, the likelihood function of each model follows a standard form. For model *i*, the likelihood function can be written as:

$$L_i(y|\beta_i, \sigma_i^2, X_i) = (2\pi\sigma_i^2)^{-n/2} \exp\{-\frac{(y-X_i\beta_i)'(y-X_i\beta_i))}{2\sigma_i^2}\},\ i=1, 2,\ldots J \quad (4)$$

Combining the likelihood function above with the prior distributions in equations (2) and (3), we derive the joint posterior distribution as:

$$p(\beta_i, \sigma_i^2 | y, X_i) \sim NIG(\mu_{pi}, \Sigma_{pi}, D_{pi}, \gamma_{pi}^2),\ \ i=1,2,\ldots J \quad (5)$$

In equation (5), NIG stands for joint normal inverse-gamma distribution. The parameters of this posterior distribution are:

$$\mu_{pi} = \Sigma_{pi}({\Sigma_{0i}}^{-1}\mu_{0i} + (X_i{}'X_i)\hat{\beta}_i), \quad (6)$$

$$\Sigma_{pi} = ({\Sigma_{0i}}^{-1} + X_i{}'X_i)^{-1}, \quad (7)$$

$$D_{pi} = D_{0i} + n_i, \quad (8)$$

$$\gamma_{pi}^2 = {D_{pi}}^{-1}[(n_i - k_i)\gamma_i^2 + \left(\hat{\beta}_i - \mu_{0i}\right)'\Sigma_{pi}\left(\hat{\beta}_i - \mu_{0i}\right) + D_{0i}\gamma_{0i}^2]. \quad (9)$$

In the above equations, $\hat{\beta}$ is the coefficients from equation (1) if estimated by OLS, γ_i^2 is the sum of squared errors from the same regression estimates of equation (1), and n_i and k_i are the number of rows and columns in ith independent matrix, X_i.

Our research interest is to see if new, seemingly-unrelated economic variables can help forecast crop prices, so we will focus not on the coefficient estimates in these different models, but in the probability of model inclusion for the different independent variables and the best number of lags to include. To do this, we focus on the posterior probabilities of variable inclusion, not of individual models. These probabilities are based on sums of posterior model probabilities across all models that include a given variable.

Computation of posterior model probabilities goes as follows. Begin with prior model probabilities,

$$p(M_i) = m_i \text{ and } \sum_{i=1}^{J} m_i = 1 \quad (10)$$

Given the large number of models estimated here (4,096) and our lack of insight on the likely inclusion of many of the variables, it is natural to use an uninformative prior over the models, that is, $m_i = \frac{1}{J}, \ \forall i$. By integrating out the parameters, we can derive the marginal likelihood function for each model,

$$p(y_i|M_i) = \theta_i[\Sigma_{pi}/\Sigma_{0i}]^{1/2}(D_{pi}\gamma_{pi}^2)^{-D_{pi}/2} \quad (11)$$

where $\theta_i = \frac{\Gamma(D_{pi}/2)(D_{p0}/2\gamma_{0i}^2)^{-/D_{pi}/2}}{\Gamma(D_{pi}/2)\pi^{n/2}}$. In θ_i, $\Gamma(.)$ is the gamma function. Based on the marginal likelihood function, we can derive the marginal posterior probability for each model:

$$p(M_i|y_i) \propto m_i\left[\left|\Sigma_{pi}\right|/|\Sigma_{0i}|\right]^{1/2}\left(D_{pi}\gamma_{0i}^2\right)^{-D_{pi}/2} = m_i\, p(y_i|M_i),\ i=1,2,\ldots J \qquad (12)$$

To ensure the sum of all the marginal posterior probabilities for each model is equal to one, we need to normalize these posterior probabilities. The simplest way to normalize them is to divide every probability by the sum of them all. Thus, we arrive at posterior model probabilities given by

$$\rho_i = \frac{m_i\, p(y_i|M_i)}{\sum_{i=1}^{J} m_i\, p(y_i|M_i)},\ i=1,2,\ldots J \qquad (13)$$

Summing subsets of these posterior model probabilities for all models that contain a particular variable or a specific number of lags of a variable provides us with the marginal posterior probability of that variable's inclusion in the model, or inclusion with a specific number of lags, respectively. These marginal posterior probabilities are the main empirical results of our paper.

Data

For the dependent variables, we use U.S. spot prices for corn, soybeans, and wheat, with the values collected from the USDA's website. We consider five U.S. leading economic indicators as independent variables: the money supply (M2), the purchasing managers' index (PMI), the interest rate (IR), crude oil prices (CL) and the S&P 500 index (SP 500). The independent and dependent variables are monthly data from January 1985 to December 2016, resulting in 384 observations. The lags of independent and dependent variables are included in the data matrix X. Given the data are monthly, we set the maximum number of lagged values considered for model inclusion to 3. Thus, there are four possible specifications for each independent variables and for the dependent variables: no lags in models, one lag, two lags, or three lagged values included in the model. When two or three lags are included, the shorter lags are also included (e.g., if X_{t-2} is in the model, so is X_{t-1}). This ensures that there is no "hole" in lag structure in the model.

Empirical Results

Models and priors

For each of the three crop prices to be forecast, we construct models with all possible combinations of lagged dependent variables and the five considered leading economic indicators. We also include dummy variables for months and years to allow for trends and seasonality. Considering that there are 4 possible specifications in lag structure for each variable and there are 6 variables including dependent variable in models constructed for one commodity, there are 4^6 = 4096 potential model specifications for each of the three commodity prices. Since the dummy variables are always included, the smallest model contains 48 regressors and the largest model contains 66 regressors (48 plus 3 lags of the commodity prices, 3 lags of M2, 3 lags of PMI, 3 lags of IR, 3 lags of CL, and 3 lags of SP500). Thus, the prior weight for each model is $m_i = \frac{1}{4096}, \forall i.$

Through equations (2) to (9), we already build up Bayesian theoretical framework for deriving posterior distributions for the parameters, while equations (10) through (13) show how to derive posterior probabilities for each model and for variable inclusion. We should note that the marginal prior distribution of parameters is fixed when the model specification changes. This means that the prior for one regressor will not change if another regressor is or is not included in matrix X.

Based on Dorfman and Sanders' work, we choose a prior variance large enough so that the effect of the prior means on the posterior distribution is minimal. As the only difference in our three crop model sets is the own commodity price lags, we use the same vector of prior means and prior variance for the corn, soybean and wheat models. The prior means are set to:

$$\mu_0 = [0,0,0,0,0,0.9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1, \dots 1]' \quad (14)$$

Since we are not sure whether these leading indicators are included in the model, we assume their prior means at all lags are zero. We expect the once lagged own crop price to have a positive coefficient and since there is likely a near unit root in lagged prices, we

assign 0.9 as prior means of prices at time $t-1$. The ones at the end of the prior mean vector are for the monthly and year dummies.

For the prior variance matrix Σ_{0i}, we select a diagonal matrix with ones on the main diagonal except for the elements with nonzero prior means. These elements are assigned prior variances of 0.01. We set the prior mean of the inverse error variance, γ_{0i}^2, to 1 and the prior degrees of freedom parameters, D_{pi}, to 20.

Under the assumptions above, we can compute the marginal posterior distribution of the regression model parameters β_i according to equation (2). The posterior model weights can also be computed based on equations (12) and (13).

Posterior results

Since there are too many posterior model weights to present (given our 4,096 potential models), we focus on the marginal probability of model specification features (variable inclusion) in Table 1. Such marginal posterior probabilities are the sum of individual model posterior probabilities of all models with that specification feature. For example, the posterior probability of only one lag of corn price in the corn price forecasting model is 0.1169, which is the sum of the posterior probabilities of all models which model contain the specific regressor, p_{t-1} and no additional lags of corn price.

The posterior probabilities in Table 1 should be interpreted as support for the model specification containing that specific feature. Based on the third row of the corn price model in Table 1, we conclude that 61.34% of the posterior support is placed on models which include two lags of corn price as regressors compared to alternative lag lengths (0, 1, or 3).

Thus, the results in Table 1 for the corn price forecasting models suggest forecasters should include two or three lags of corn price. In addition, posterior support is quite evenly split on the inclusion of M2, PMI, and interest rates, with the posterior support for exclusion of those leading indicators all near 50%. The exercise is more definitive with regard to crude oil prices and the S&P500, with roughly 2/3 and 3/4 of the posterior support for those variables favoring exclusion.

For the soybean price model, we again find strong posterior support for two lags of the soybean (own crop) price, a slight tilt toward inclusion of M2 and PMI, a virtual tie on inclusion/exclusion of the interest rate, and strong evidence in favor of excluding lagged values of crude oil prices and the S&P500 index value.

For wheat, two lags of the own price again receive the highest posterior support, the only leading indicator to get over 50% posterior support for inclusion is M2, and the results lean pretty heavily toward leaving out interest rates, crude oil prices, and the S&P 500 index.

In addition, we calculated the mean squared errors (MSE) for all the different model specifications. Table 2 shows the resulting MSE values for four possible model specifications: the minimal model with only year and monthly dummy variables, the model with the dummies plus all three lagged own prices, the full model including three lags of all considered variables, and the model specification with the highest posterior probability out of all 4,096 possibilities. Because the marginal likelihood values used to compute posterior model probabilities integrate over all possible coefficient values that receive posterior support, the point forecasts generated from the posterior means of the parameters need not generate the smallest forecast MSE when the posterior model probability is the highest. In fact, we find such a result with the full model producing better forecasts than the most likely model for all three crops.

While the posterior probabilities of inclusion were not especially high for the leading economic indicators, including them does slightly increase the forecasting performance (reduces the MSE) compared to models with only the lagged own crop price included for all three commodity prices tested. Thus, while the posterior model probabilities were rather evenly balanced or slightly against using the leading indicators in a forecasting model, the actual forecasting performance suggests that the idea has some merit and deserves further scrutiny.

Conclusions

We explore variable and model selection problems under a Bayesian framework when model specification is uncertain with respect to crop price forecasting for three major US crops: corn, soybeans, and wheat. Under an assumption of a normal-inverse gamma

likelihood function, analytical inference can be accomplished to both produce price forecasts and to estimate the posterior support for particular model features.
Such an approach can help practitioners to choose suitable lag lengths and variables to include in their forecasting models.

In the application here, we particularly focus on whether forecasts of these three commodity prices can be improved by the inclusion of five possible leading economic indicators. The motivation for widening the set of possible included variables is the recent positive correlation of agricultural commodity prices with a general set of financial asset prices, suggesting that measures of general economic conditions might help forecast agricultural commodity prices. Based on this motivation, we explored the potential value of adding five leading economic indicators to agricultural price forecasting models: the money supply, the purchasing managers' index, interest rates, crude oil prices, and the S&P 500 stock price index.

Because we consider up to three lags of each economic indicator plus three lags of the own price, our model candidate set includes 4,096 different models. After deriving posterior model probabilities for all models considered, we summarize support for different variables by focusing on the posterior probability of specific model features (e.g., two lagged interest rates), rather than on each of the 4,096 specific models. Based on these results, we find that for all three crops, two lags of their own price are the most supported model specification. Posterior feature probabilities offer reasonable support for the inclusion of M2, PMI, and interest rates with some number of lags and less support for inclusion of any lags of either crude oil prices or the S&P500 index. However, in contrast to the model feature posterior probabilities, the MSE values of our forecasting models show slight improvement when lagged values of the five leading economic indicators are included. Thus, we think this initial look at widening the set of considered crop forecasting models was successful and encourage future research to more carefully investigate variables that might improve our crop forecasting models.

References

Allen, P. G. (1994) Economic forecasting in agriculture. International Journal of Forecasting 10 (1994) 81-35.

Bada, C., and T. Kisinbay. (2011). Predicting Recessions: A New Approach For Identifying Leading Indicators and Forecast Combinations, Working Paper, IMF.

Chiaie, S., Ferrara, L. and D. Giannone. (2017). Common Factors of Commodity Prices, Working Paper, European Central Bank.

Chipman, Hugh, Edward I. George and Robert E. McCulloch (2001). The Practical Implementation of Bayesian Model Selection. Monograph Series (2001) Volume 38,100-178.

Dorfman, J. H., and Sanders, D. R. (2004). Generalized Hedge Ratio Estimation With An Unknown Model, NCR-134 Conference on Applied Commodity Price Analysis, St. Louis, MO.

Franke, J., W. Härdle, and C. Hafner. (2008). Statistics of Financial Markets: An Introduction (Third ed.). Springer.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2014). Bayesian Data Analysis (Third Edition), 141-149.

Westcott, Paul C., and Linwood A. Hoffman (1999). Price Determination for Corn and Wheat: The Role of Market Factors and Government Programs. U.S. Department of Agriculture. Technical Bulletin No. 1878.

Table 1: Model Feature Posterior Probabilities

Feature	Corn Model Post. Prob.	Soybean Model Post. Prob.	Wheat Model Post. Prob.
no price lag	0.0000	0.0000	0.0000
p_{t-1}	0.1169	0.0004	0.1004
p_{t-2}	0.6134	0.8182	0.6405
p_{t-3}	0.2697	0.1814	0.2591
no M2 lag	0.4483	0.4568	0.4452
$\mathrm{M2}_{t-1}$	0.4039	0.4074	0.4052
$\mathrm{M2}_{t-2}$	0.0848	0.0801	0.0862
$\mathrm{M2}_{t-3}$	0.0630	0.0557	0.0634
no PMI lag	0.5060	0.4873	0.5045
PMI_{t-1}	0.4095	0.4119	0.4092
PMI_{t-2}	0.0653	0.0763	0.0666
PMI_{t-3}	0.0192	0.0245	0.0197
no interest rate lag	0.5841	0.5047	0.6269
IR_{t-1}	0.3164	0.3673	0.2835
IR_{t-2}	0.0784	0.0903	0.0692
IR_{t-3}	0.0211	0.0377	0.0204
no crude oil price lag	0.6698	0.7438	0.6674
CL_{t-1}	0.2603	0.2110	0.2582
CL_{t-2}	0.0584	0.0365	0.0618
CL_{t-3}	0.0115	0.0087	0.0126
no SP 500 lag	0.7778	0.7425	0.7780
$\mathrm{SP\ 500}_{t-1}$	0.1919	0.2089	0.1917
$\mathrm{SP\ 500}_{t-2}$	0.0265	0.0267	0.0265
$\mathrm{SP\ 500}_{t-3}$	0.0038	0.0219	0.0038

Table 2: MSE for various model specifications

	corn model	soybean model	wheat model
Only dummy variables included	0.108646	0.240617	0.108646
Dummies + own price lags	0.024355	0.056410	0.025281
Most Likely Model	0.024248	0.055633	0.024870
All variables included	0.022209	0.052035	0.023085