Time Series Modeling of Agricultural Market Data

Jordan Turley

December 10, 2019

1 Introduction

The agriculture industry is a multi-billion dollar industry that is essential to modern life. Without the agriculture industry, it would be impossible to go to the grocery and buy meat, fruit, vegetables, or basically anything that we eat. Because of the already high and increasing demand for agricultural goods, and the volatile supply, the prices of common goods like corn, soybeans, and others are very volatile, similar to stock prices. Analyzing and predicting prices can be useful for farmers. Knowing if the price is going to go up or down in the next few days or even next month can inform the farmer to sell now or hold their supply for the future. This could allow large farms to gain an advantage over their competitors, or simply allow a small farm to make more money to support a family. In this paper, we will analyze the prices of four common crops grown and sold in the United States: corn, cotton, soybeans, and wheat. We will look at historical trends during regular economic climate and during recessions, as well as predicting future crop prices using previous crop prices of the same crop, previous prices of the other crops, and using both.

2 Data Analysis

2.1 Data Collection

Macrotrends.net has data sets of each of the four crop prices that go all the way back to the 1950s, which are well organized and do not have any missing values. Links to each of the data sets will be given at the end of the paper.

2.2 Data Exploration

2.2.1 2009 - 2019

Below in Figure 1, we see plots for the four crop prices for the last ten years; precisely, from October 23, 2009 to October 23, 2019. The last ten years were used since this will reflect how the current prices perform better than using data from the 1950s or 1960s, when the economy of the United States and of the entire world was completely different. This still gives us thousands of data points. The previous ten years also does not include any recessions, which causes crop prices to perform completely differently. We will look at recessions in the next section.

We see that, in general, the prices seem to follow each other. We see a large peak around 2012, which then decreased to a trough around 2014 to 2016, but has generally been increasing since. From these plots, it seems like we may be able to predict one price from the others fairly well. If some prices are available earlier than others, this could allow an individual to gain an advantage over others by knowing about what the future price is going to be before others know.

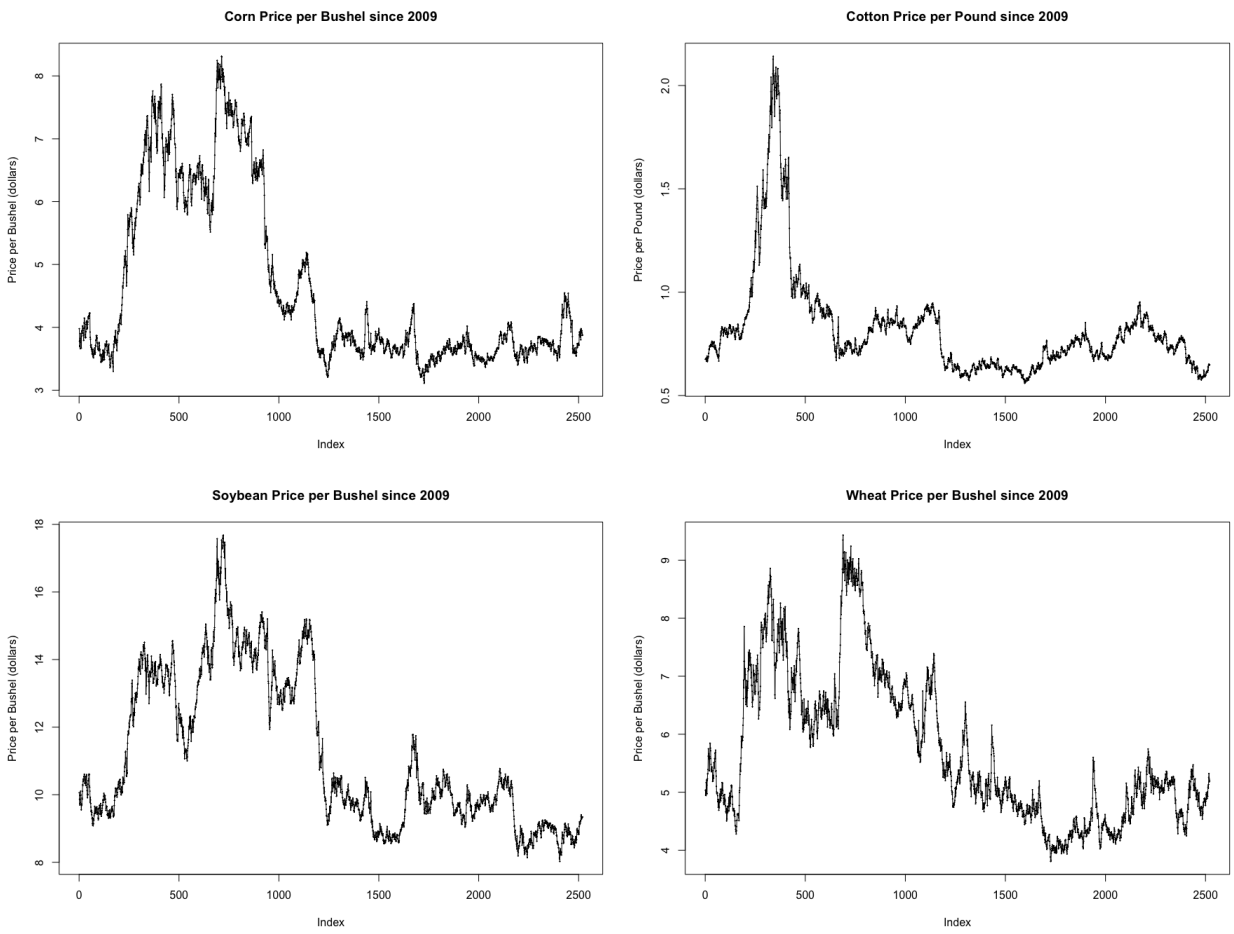

Figure 1: Prices From Last Ten Years of Corn, Cotton, Soybeans, and Wheat

2.2.2 Great Recession

Below in Figure 2, we see the plots for the four crop prices during the Great Recession, or from December 1, 2007 to June 1, 2009. Again, we see very similar general trends among the prices, which again suggests that we will be able to predict one from the others. We see somewhat surprisingly that for the first three to six months of the recession, the prices go up, but then steadily decrease, before again increasing at the end of the recession.

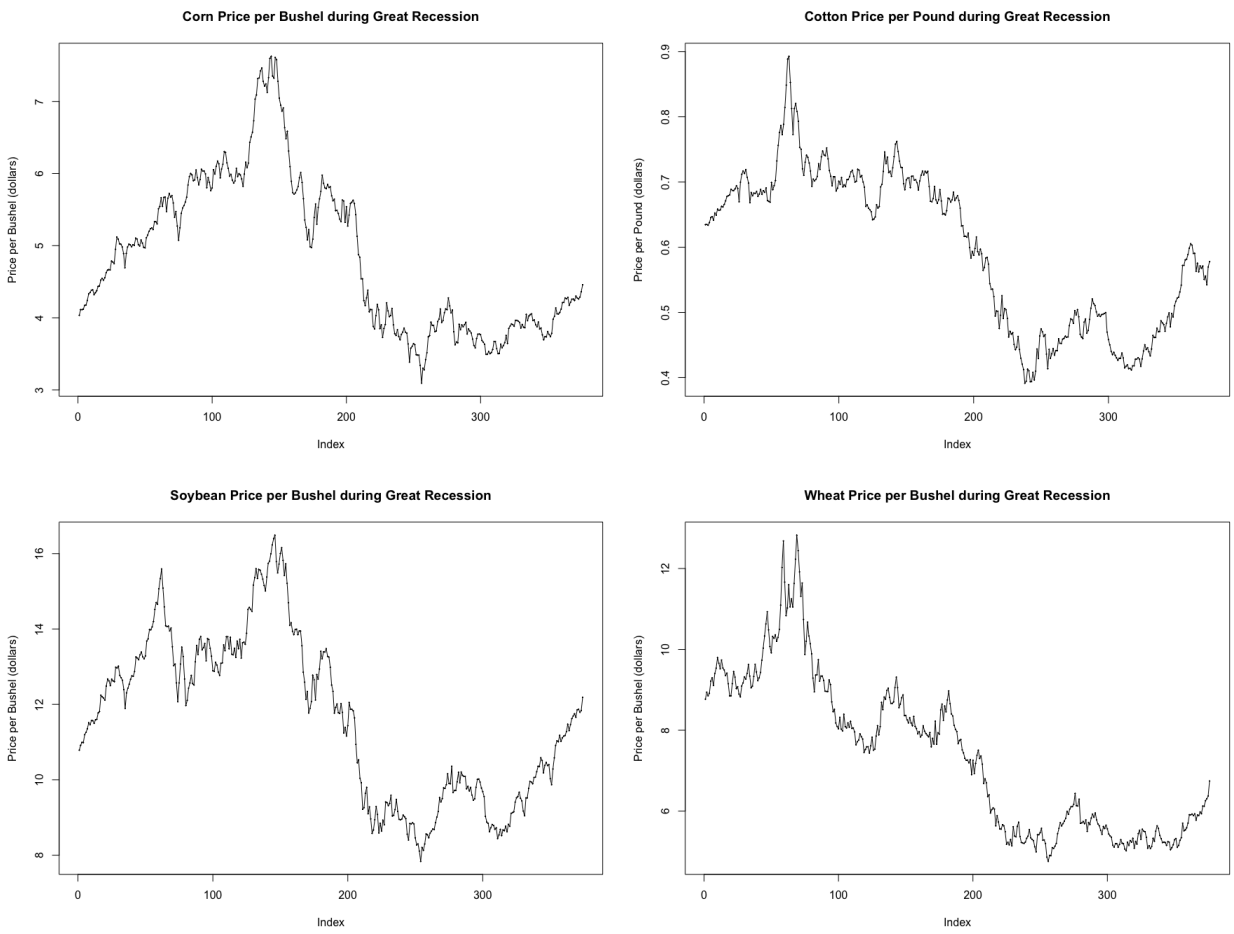

Figure 2: Prices of Corn, Cotton, Soybeans, and Wheat during Great Recession

3 Modeling

In this section we will look at how we can use different types of statistical models to predict future prices. We reserve the last month of the data to evaluate the effectiveness of the models.

3.1 Classical Time Series

First, we apply the ARIMA model to each data set. Since the data is clearly not stationary, we take the first difference, which we see below in Figure 3. We see that the first differences mostly resemble white noise, indicating that the first difference is stationary.

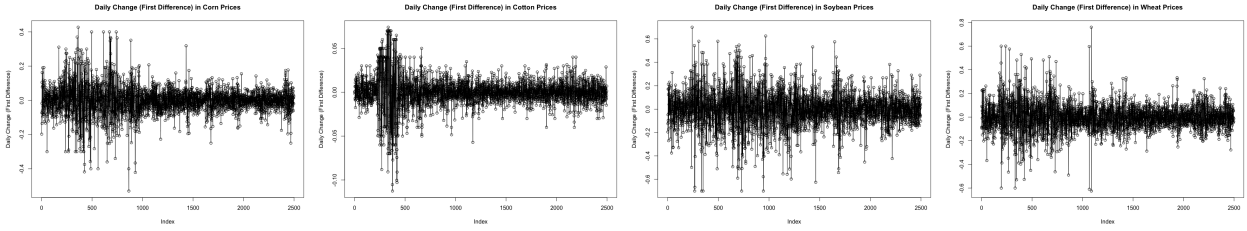

Figure 3: First Differences of Corn, Cotton, Soybean, and Wheat Prices

3.1.1 Model Selection

To select the best model, we will use three methods: Akaike information criterion (AIC), Bayesian information criterion (BIC), and the autocorrelation/partial autocorrelation plots, and we will compare how well these perform on the test set data.

Below in Figure 4 we show the AIC and BIC values, as well as the ACF and Partial ACF plots for the first differences of corn prices. The best, or lowest AIC and BIC are shown in bold. We notice a few things. First, the model suggested by BIC and the model suggested from the ACF and PACF plots is the same, which is an ARMA(1, 1) model. The model suggested by AIC is ARMA(3, 6). However, all of the AIC and BIC values are very close to each other, so it probably would not make much difference if we chose to use any of them. We repeat the same process for all four crops.

AR \ MA	1	2	3	4	5	6
1	-5279.696	-5278.373	-5276.447	-5274.712	-5272.712	-5271.155
2	-5278.387	-5276.203	-5274.382	-5272.847	-5270.839	-5269.266
3	-5276.435	-5274.569	-5272.549	-5270.555	-5268.843	**-5281.509**
4	-5274.729	-5272.898	-5270.97	-5269.017	-5266.747	-5265.071
5	-5272.729	-5270.855	-5268.507	-5271.41	-5265.337	-5277.451
6	-5271.231	-5269.207	-5276.186	-5269.619	-5272.315	-5279.05

AR \ MA	1	2	3	4	5	6
1	**-5262.227**	-5255.082	-5247.333	-5239.775	-5231.952	-5224.572
2	-5255.095	-5247.089	-5239.445	-5232.087	-5224.256	-5216.86
3	-5247.32	-5239.632	-5231.789	-5223.972	-5216.438	-5223.28
4	-5239.792	-5232.138	-5224.387	-5216.611	-5208.519	-5201.019
5	-5231.969	-5224.272	-5216.101	-5213.181	-5201.286	-5207.577
6	-5224.649	-5216.801	-5217.958	-5205.567	-5202.441	-5203.353

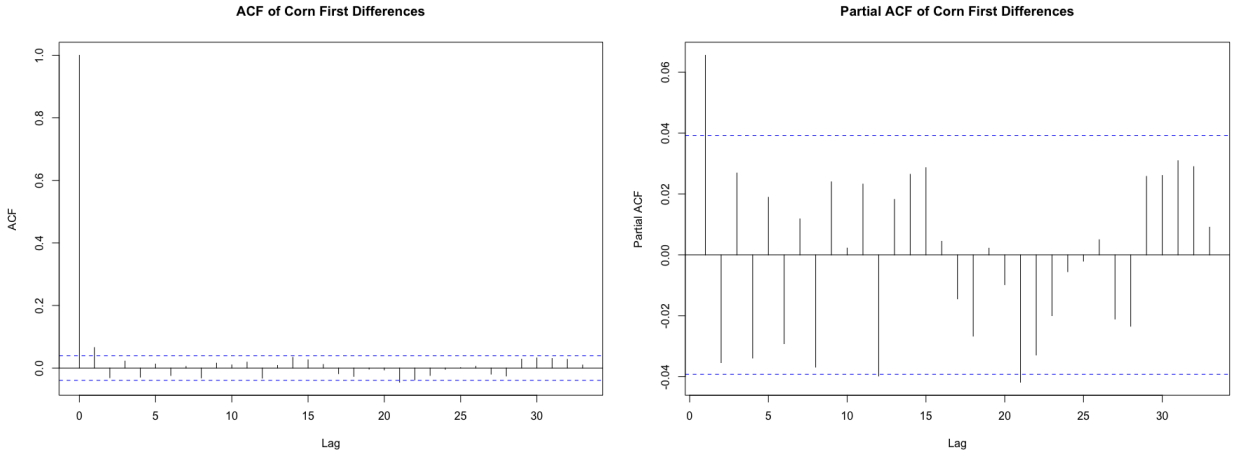

Figure 4: AIC, BIC, ACF, and PACF for First Difference Corn Prices

3.1.2 Model Performance

Below, in Figure 5, we see the actual performance of the model on the last month, which was held out from training. We plot the true values from the month, the predicted values from each model, and the 95% prediction interval as well. We see that all four look very similar for all of the models. This is somewhat expected, as we wouldn't expect drastic differences between an ARMA(1, 1) and an ARMA(5, 4) model, for example. There are small differences in the predictions and in the width of the prediction interval, but overall they are very similar.

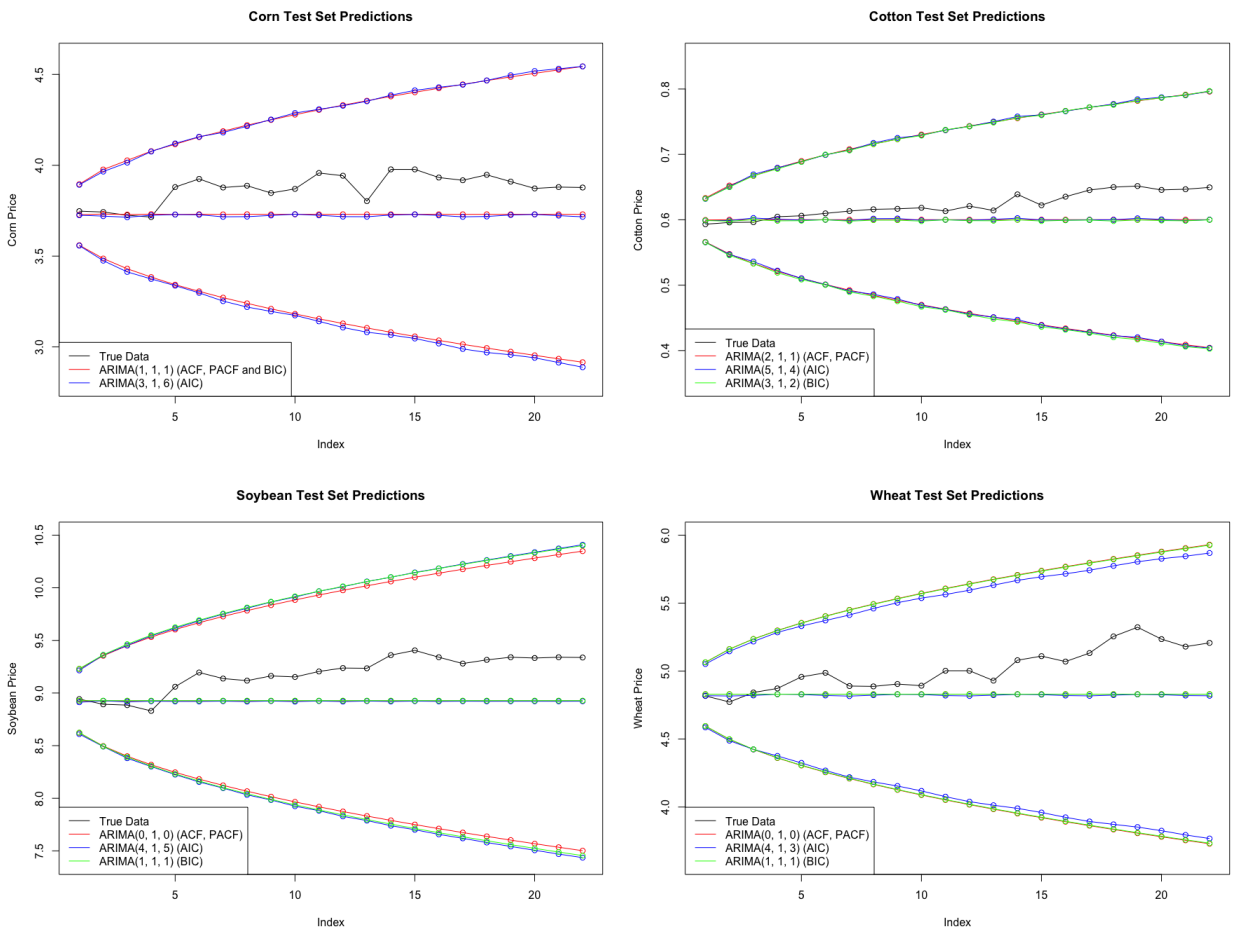

Figure 5: Predictions for Prices Using Several ARIMA Models

One interesting note was that the ACF and PACF plots mostly suggested that there was no significant autocorrelation for soybeans and wheat. This means that the data resembles white noise, or is completely random. The other two models' predictions were almost exactly the same, so there wasn't much to be gained by using an ARMA(6, 3) over an ARMA(0, 0), for example, other than small differences in the prediction intervals.

We can draw a few interesting conclusions from these models. We see that for the first few days, we are able to predict the price pretty well. The predictions for the models are very close to the true values. However, after this, we see that our models are not able to account for the random variations in the data. This is somewhat expected. As we move further into the future, it is harder to make accurate predictions.

We do see that the true value of the price is always within our prediction interval, but after 10 or more days out, the prediction interval is extremely wide to account for the random variation. This shows that this problem is tough to predict well. For the first few days, we get a pretty tight prediction interval, but as we move more days into the future, the prediction interval has to get wider and wider to account for the random variation in the price.

I was a little disappointed that the predictions were all basically just a straight line across. However, predicting a straight line does match the first few days of the test set very closely, and the prediction interval is tight enough that this would make a pretty good prediction. As we move further out, the predictions are less viable as the prediction interval gets very wide.

In a practical sense, this is still useful. Given the limitations, one could use a model like this to predict the price for the next day or even up to the next three days. It will not be effective to use this model to predict the price next year, since the confidence band will be so wide, but these models could be effective in

predicting the price for the next day. Then, once you know what the actual price is the next day, this can be added to the training data and the model can be retrained for use for the next day.

There is future work that could be done with these models. Predictions were only made on the last month of data. Since I had a significant amount of data, We could have used the first n months and then predicted for the next month, and repeated this several times, increasing n to see how the model predicts for the next month. Also, these models did not use any seasonality. By looking at the plots of the prices for the last ten years, there is no obvious seasonality. There may be some seasonality, as the prices might change due to supply going up at the end of the summer after harvest, or due to demand being higher during some parts of the year for example. However, I chose to use the basic ARIMA models rather than include seasonal terms since we were using daily data. If one wanted to predict the monthly average, for example, it may make more sense to make use of some seasonality.

3.1.3 Recession

Again, we went through the same process of model selection by using AIC, BIC, and the ACF/Partial ACF plots. Shown below in Figure 6 are the ACF and Partial ACF plots for each. We see something really interesting, in that none of the lags seem to have significant correlation. There are a few later lags in some of the plots that are significant, but this could simply be due to randomness. During the analysis of the past ten years, we did see that soybeans and wheat had no significant lags, but corn and cotton did. Here, all four have no significant lags. This tells us something very important about the prices of these crops during recessions. This tells us that all four of these crops resemble white noise, or that during the recession, the prices were more random than during regular economic climate. This is somewhat expected, since the way the economy functions and the way people act during a recession is very different than the way it functions during regular economic climate. However, it is very interesting to see numerical evidence of this in that all four crop prices during the recession resembled white noise. If you chose one of these plots and showed it to a person that didn't know any better, there would not be much reason for them to believe that this is not simply white noise.

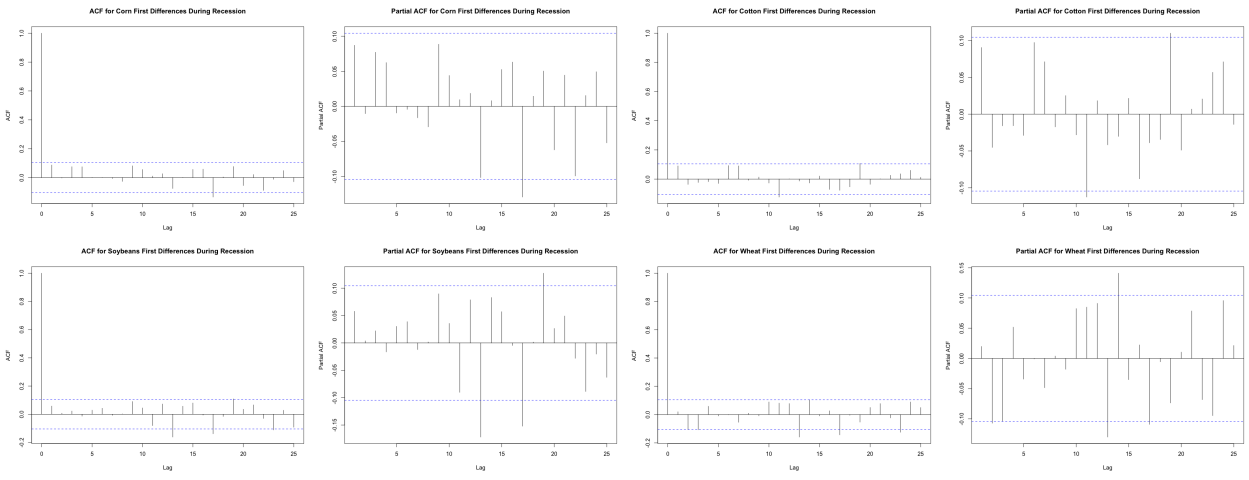

Figure 6: ACF/PACF Plots for Corn, Cotton, Soybean, and Wheat First Differences during Recession

Next, we used AIC and BIC to select models for the recession, since it is still important to attempt to predict these prices even during a recession rather than simply assuming that the prices are random, as the above suggests. The models used and predictions for the last month of the recession are given below in Figure 7.

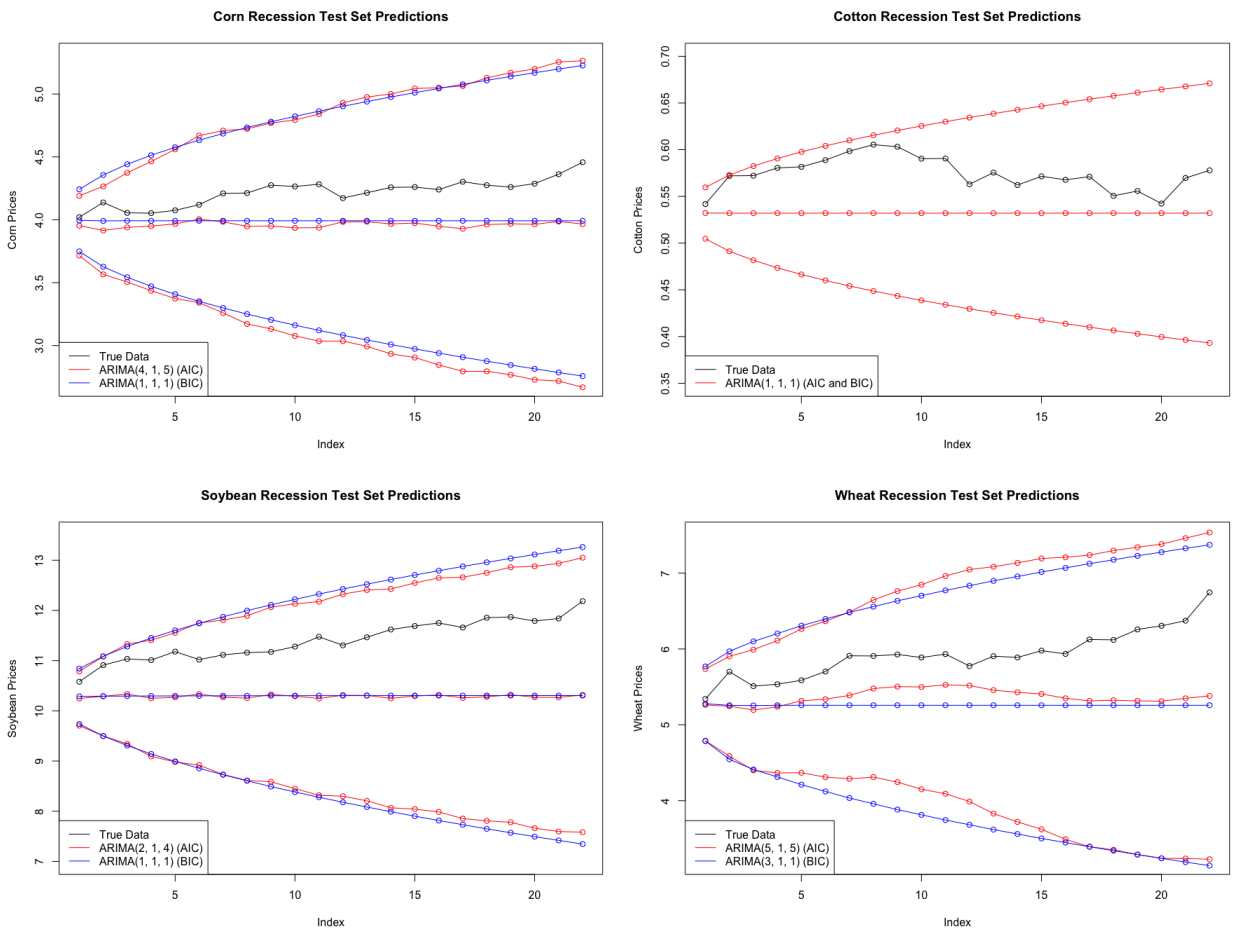

Figure 7: Predictions for Prices During Great Recession Using Several ARIMA Models

We see a few different things which are interesting. First, there is some overlap between the models found here and the models found for the regular data above, but the overlap does not seem to be significant. This suggests that the model that best represents the data during a recession is different from the model that best represents the data during regular economic climate. This is useful information, as it tells someone trying to predict crop prices that, if there is a recession, you will likely do better using a different model than the usual models.

In terms of performance, sometimes these models do well, but sometimes they do poorly. For example, in the corn predictions, we see the more complex ARIMA(4, 1, 5) model does worse than the simpler ARIMA(1, 1, 1) model, even though the simpler model basically predicts the same value all month. Around days eight to twelve, the true data goes up, but the predictions go down. However, in the wheat data, we see that the ARIMA(5, 1, 5) model actually does well following the true data as it goes up around days eight to twelve. The true data continues to increase and the predictions decrease, so as we move further out the predictions get worse, but we do pretty well up to about day fifteen.

The models's performance tells us something important about the prices during a recession, as well as the ACF and PACF plots above, which is that the prices during a recession are more volatile and random than regular times. Sometimes our model does well, like the case of wheat, but sometimes it does poorly, like the case of corn. Again, we always get the true price in our prediction interval, but the prediction interval gets very wide as we predict further into the future. A regular person could probably look at some data and say that the price of soybeans in a month will be between seven and thirteen dollars, for example.

3.1.4 Classical Time Series Conclusions

There are many conclusions that we can draw from the classical time series modeling process. First, predicting real world data like crop prices is difficult, and the fairly simple ARIMA models we tried are not able to accurately or precicely predict the price of a crop for more than a few days in the future. However, whether it was due to luck or not, we were able to accurately predict the price for about four days into the future, and even farther into the future for some of the crops. Also, even far into the future, we capture the true price in the prediction interval of our prediction. Even though our prediction interval gets very wide as we move far into the future, it is still somewhat useful to know that we can be very confident that our prediction interval will capture the true value. We could even use a tighter interval, like a 67% prediction interval and be confident it would capture the true value or come close to capturing the true value, since it seems like our models are accounting for more variability than actually happens in the data.

We also looked at data during the Great Recession and made some interesting discoveries. When looking at the autocorrelograms and partial autocorrelograms, we saw that the recession data resembles random data, i.e. white noise, more than the regular data. We expect this during a recession since the economic climate and individual behaviors are different, but it is refreshing to see hard numeric evidence of this. In addition, we see this randomness in our predictions as well. Some models performed fairly well, while others did not. This again tells us that this is a hard problem due to the randomness of the data, and even more so during a recession. Finally, most of the models we found best using AIC and BIC were different than the models found for the regular data, which suggests that prediction during a recession should be done using different models than prediction during regular economic climate.

In the next sections, we will look at how correlated each of the crop prices is by predicting one from another, as well as using previous crop prices and current prices of other crops to predict one crop.

3.2 Predict One Crop from Another

In this section, we will predict the current price of one crop from the current price of another crop, using a simple linear regression model. The main goal is to see how correlated each of the crops is with the others, and how well we could predict one crop price from another.

In Figure 8, we see the results from several regressions of one crop on another.

Crop 1	**Crop 2**	**Coefficient**	t	p-value	R^2
Corn	Cotton	3.01	33.12	0.0	0.305
Corn	Soybeans	0.527	82.85	0.0	0.733
Corn	Wheat	0.97	98.54	0.0	0.795
Cotton	Soybeans	0.049	24.14	0.0	0.189
Cotton	Wheat	0.105	31.04	0.0	0.279
Soybeans	Wheat	1.509	81.64	0.0	0.728

Figure 8: Pairwise Regressions Between Current Price of Corn, Cotton, Soybeans, and Wheat

There are several interesting points we can draw from the above regressions. First, all of the coefficients are positive, which we expected. As one price goes up, the other prices go up as well. The value of the coefficient doesn't matter much, since the relative prices of each crop are different, but the sign does. In addition, each of the coefficients is statistically significant at any reasonable confidence level, so we can be very confident in the relationship.

When we look at the values of R^2, we see that there is a much stronger correlation between corn, soybeans, and wheat, but not between cotton, which is interesting. When we look at the raw plots of the data, we see that the general trends in the plots are the same, but the plots of corn, soybeans, and wheat are much more alike, and the plot of cotton is the odd one out. One explanation is that the market for cotton is significantly different since it is a different type of good than the other three. Cotton is used in making fabric, where corn, soybeans, and wheat are all used for food. The general trends are the same, but the three food goods are much more correlated than the non-food good.

3.3 Prediction using Current Crop Prices

In this section, we will use the same time series concepts we used in the previous section, but rather than only using the previous prices of the crop, we will also use previous and current prices of other crops. I must admit that I do not know the logistics of the markets that these goods are sold on, but one could imagine a setting where the prices of crop A, B, and C are somehow available before the price of crop D, so one could use the current and historic prices of A, B, and C, along with the historic prices of crop D to predict the current price of crop D. If the prices of A, B, C, and D are correlated, this could potentially give a very accurate prediction of the current price of D before it is publicly available.

We will use linear regression to build an autoregressive model of the current price of the crop on the prices of the crop from the last six days, as well as the other three crops' current and previous prices from the last six days. Instead of calculating AIC and BIC for model selection, I chose to use six days for all four models so that we can see which crops and which lags are the most significant. The results are shown below in Figure 9.

There are several things to notice. First is that we get a great R^2 value, which is mostly due to the fact that we have much more data here than before, when using classical time series methods. Now we have access to other current crop prices, and they do a very good job of predicting the current price of the crop in question. In almost every model, at least one of the other crop prices is extremely significant and plays a large role in predicting the current price of the given crop. Cotton seems to be the one exception, but this is likely because it is a different type of crop and follows different trends, as we discussed before. The most signfiicant coefficients for cotton are the previous cotton prices, along with a few seemingly random ones, like corn for lags 5 and 6, soybeans lags 4 and 6, and wheat lag three. In the other models, the most significant coefficients are the lags of the current crop, as well as at least one or both of the current prices of the other crops, ignoring cotton.

While these results may seem intuitive or obvious, this still tells us a lot and allows us to draw several conclusions. First, we again have statistically significant evidence that cotton performs differently than the other three crops. There is definitely future work to be done investigating why this is. Next, it is unclear if this would be feasible in the real world, but if it is, one could most definitely exploit the current prices of other crops to predict the current price of a given crop. In practice, this would almost be like a form of arbitrage. If an individual could learn the current price of one crop before the current price of the other crop is known, one could use this information, along with the previous prices of the crops, to predict the crop price accurately. If crop prices work like stock prices, then this likely would not work unless there is a delay in timing for some prices for some reason, but if prices are set and released at a given time, and some are released before others, then this could be used to ones advantage.

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.009743   0.007444  -1.309 0.190702
cornTrain$y_minus_1    1.048185   0.020128  52.077  < 2e-16 ***
cornTrain$y_minus_2   -0.169861   0.029015  -5.854 5.43e-09 ***
cornTrain$y_minus_3    0.117589   0.029223   4.024 5.90e-05 ***
cornTrain$y_minus_4   -0.089084   0.028112  -3.169 0.001549 **
cornTrain$y_minus_5    0.070728   0.026569   2.662 0.007818 **
cornTrain$y_minus_6    0.015133   0.018362   0.824 0.409949
cottonTrain$value      0.121960   0.074315   1.641 0.100898
cottonTrain$y_minus_1 -0.057884   0.115533  -0.501 0.616406
cottonTrain$y_minus_2 -0.114690   0.116663  -0.983 0.325660
cottonTrain$y_minus_3  0.208256   0.117383   1.774 0.076161 .
cottonTrain$y_minus_4  0.152572   0.117488   1.299 0.194196
cottonTrain$y_minus_5 -0.377119   0.117073  -3.221 0.001293 **
cottonTrain$y_minus_6  0.078619   0.076223   1.031 0.302441
soybeansTrain$value    0.239579   0.008844  27.089  < 2e-16 ***
soybeansTrain$y_minus_1 -0.248389  0.013381 -18.562  < 2e-16 ***
soybeansTrain$y_minus_2  0.002530  0.014281   0.177 0.859376
soybeansTrain$y_minus_3  0.012393  0.014278   0.868 0.385486
soybeansTrain$y_minus_4  0.012433  0.014260   0.872 0.383357
soybeansTrain$y_minus_5 -0.006379  0.014257  -0.447 0.654582
soybeansTrain$y_minus_6 -0.013844  0.010012  -1.383 0.166847
wheatTrain$value       0.079184   0.010890   7.271 4.76e-13 ***
wheatTrain$y_minus_1  -0.084521   0.015281  -5.531 3.52e-08 ***
wheatTrain$y_minus_2   0.268026   0.015698  17.074  < 2e-16 ***
wheatTrain$y_minus_3  -0.246842   0.017015 -14.507  < 2e-16 ***
wheatTrain$y_minus_4   0.039021   0.017579   2.220 0.026525 *
wheatTrain$y_minus_5  -0.063748   0.017528  -3.637 0.000282 ***
wheatTrain$y_minus_6   0.018156   0.012597   1.441 0.149638
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06163 on 2464 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.998,     Adjusted R-squared:  0.998
F-statistic: 4.647e+04 on 27 and 2464 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.0002877  0.0020175  -0.143  0.88662
cornTrain$value        0.0089527  0.0054552   1.641  0.10090
cornTrain$y_minus_1   -0.0130675  0.0078994  -1.654  0.09821 .
cornTrain$y_minus_2   -0.0024431  0.0079154  -0.309  0.75761
cornTrain$y_minus_3    0.0074562  0.0079420   0.939  0.34791
cornTrain$y_minus_4   -0.0001638  0.0076320  -0.021  0.98288
cornTrain$y_minus_5    0.0196079  0.0071981   2.724  0.00649 **
cornTrain$y_minus_6   -0.0215291  0.0049567  -4.343 1.46e-05 ***
cottonTrain$y_minus_1  1.1925197  0.0200692  59.421  < 2e-16 ***
cottonTrain$y_minus_2 -0.1765647  0.0314138  -5.621 2.12e-08 ***
cottonTrain$y_minus_3 -0.0473945  0.0318093  -1.490  0.13636
cottonTrain$y_minus_4  0.0364978  0.0318343   1.146  0.25170
cottonTrain$y_minus_5  0.0052198  0.0317860   0.164  0.86957
cottonTrain$y_minus_6 -0.0138011  0.0206542  -0.668  0.50407
soybeansTrain$value    0.0006736  0.0027298   0.247  0.80512
soybeansTrain$y_minus_1 -0.0002510  0.0038707  -0.065  0.94831
soybeansTrain$y_minus_2 -0.0001790  0.0038694  -0.046  0.96311
soybeansTrain$y_minus_3  0.0041052  0.0038681   1.061  0.28867
soybeansTrain$y_minus_4 -0.0076105  0.0038612  -1.971  0.04883 *
soybeansTrain$y_minus_5 -0.0027159  0.0038624  -0.703  0.48202
soybeansTrain$y_minus_6  0.0058545  0.0027110   2.160  0.03091 *
wheatTrain$value       -0.0047537  0.0029805  -1.595  0.11086
wheatTrain$y_minus_1    0.0073056  0.0041633   1.755  0.07943 .
wheatTrain$y_minus_2   -0.0080519  0.0044947  -1.791  0.07335 .
wheatTrain$y_minus_3    0.0121641  0.0047967   2.536  0.01128 *
wheatTrain$y_minus_4   -0.0022667  0.0047673  -0.475  0.63450
wheatTrain$y_minus_5   -0.0091466  0.0047582  -1.922  0.05469 .
wheatTrain$y_minus_6    0.0065098  0.0034120   1.908  0.05652 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0167 on 2464 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9957,    Adjusted R-squared:  0.9957
F-statistic: 2.129e+04 on 27 and 2464 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.026581   0.014879   1.786  0.07415 .
cornTrain$value        0.957809   0.035358  27.089  < 2e-16 ***
cornTrain$y_minus_1   -1.006171   0.054694 -18.397  < 2e-16 ***
cornTrain$y_minus_2    0.075164   0.058396   1.287  0.19817
cornTrain$y_minus_3    0.069695   0.058605   1.189  0.23446
cornTrain$y_minus_4    0.021109   0.056322   0.375  0.70784
cornTrain$y_minus_5   -0.054873   0.053189  -1.032  0.30234
cornTrain$y_minus_6   -0.050464   0.036705  -1.375  0.16931
cottonTrain$value      0.036684   0.148669   0.247  0.80512
cottonTrain$y_minus_1 -0.233157   0.230969  -1.009  0.31285
cottonTrain$y_minus_2 -0.026879   0.233310  -0.115  0.90829
cottonTrain$y_minus_3  0.136039   0.234838   0.579  0.56245
cottonTrain$y_minus_4  0.460497   0.234811   1.961  0.04998 *
cottonTrain$y_minus_5 -0.113993   0.234566  -0.486  0.62703
cottonTrain$y_minus_6 -0.271209   0.152341  -1.780  0.07515 .
soybeansTrain$y_minus_1  1.005587   0.020139  49.933  < 2e-16 ***
soybeansTrain$y_minus_2  0.007881   0.028555   0.276  0.78258
soybeansTrain$y_minus_3 -0.037930   0.028542  -1.329  0.18400
soybeansTrain$y_minus_4 -0.005226   0.028518  -0.183  0.85461
soybeansTrain$y_minus_5 -0.002872   0.028507  -0.101  0.91977
soybeansTrain$y_minus_6  0.026874   0.020018   1.342  0.17956
wheatTrain$value       0.040628   0.021992   1.847  0.06481 .
wheatTrain$y_minus_1   0.012471   0.030743   0.406  0.68503
wheatTrain$y_minus_2   0.026971   0.033188   0.813  0.41648
wheatTrain$y_minus_3  -0.091935   0.035396  -2.597  0.00945 **
wheatTrain$y_minus_4  -0.023924   0.035180  -0.680  0.49654
wheatTrain$y_minus_5   0.035916   0.035134   1.022  0.30675
wheatTrain$y_minus_6  -0.001934   0.025198  -0.077  0.93883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1232 on 2464 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.997,     Adjusted R-squared:  0.997
F-statistic: 3.07e+04 on 27 and 2464 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.017859   0.013625   1.311 0.190056
cornTrain$value        0.265283   0.036484   7.271 4.76e-13 ***
cornTrain$y_minus_1   -0.217436   0.053215  -4.086 4.53e-05 ***
cornTrain$y_minus_2    0.059590   0.053461   1.115 0.265118
cornTrain$y_minus_3   -0.096274   0.053628  -1.795 0.072741 .
cornTrain$y_minus_4   -0.014756   0.051558  -0.286 0.774749
cornTrain$y_minus_5   -0.032209   0.048697  -0.661 0.508402
cornTrain$y_minus_6    0.039790   0.033604   1.184 0.236487
cottonTrain$value     -0.216953   0.136026  -1.595 0.110856
cottonTrain$y_minus_1  0.341631   0.211365   1.616 0.106156
cottonTrain$y_minus_2  1.100928   0.212422   5.183 2.36e-07 ***
cottonTrain$y_minus_3 -1.237107   0.213540  -5.793 7.78e-09 ***
cottonTrain$y_minus_4 -0.314266   0.215025  -1.462 0.143998
cottonTrain$y_minus_5 -0.062879   0.214733  -0.293 0.769679
cottonTrain$y_minus_6  0.392545   0.139321   2.818 0.004878 **
soybeansTrain$value    0.034046   0.018429   1.847 0.064805 .
soybeansTrain$y_minus_1 -0.062801  0.026118  -2.404 0.016269 *
soybeansTrain$y_minus_2  0.005247  0.026140   0.201 0.840933
soybeansTrain$y_minus_3  0.026783  0.026132   1.025 0.305509
soybeansTrain$y_minus_4  0.004973  0.026105   0.190 0.848939
soybeansTrain$y_minus_5  0.013658  0.026094   0.523 0.600724
soybeansTrain$y_minus_6 -0.017850  0.018328  -0.974 0.330213
wheatTrain$y_minus_1   0.976906   0.020118  48.558  < 2e-16 ***
wheatTrain$y_minus_2  -0.115607   0.030295  -3.816 0.000139 ***
wheatTrain$y_minus_3   0.135300   0.032332   4.185 2.96e-05 ***
wheatTrain$y_minus_4  -0.042832   0.032196  -1.330 0.183534
wheatTrain$y_minus_5   0.078944   0.032130   2.457 0.014077 *
wheatTrain$y_minus_6  -0.047614   0.023047  -2.066 0.038938 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1128 on 2464 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9922,    Adjusted R-squared:  0.9921
F-statistic: 1.165e+04 on 27 and 2464 DF,  p-value: < 2.2e-16
```

Figure 9: Autoregression Results for Crop Price Prediction Using All Data. From top left to bottom right, Corn, Cotton, Soybeans, Wheat.

4 Conclusion

The first thing to acknowledge in this study is that predicting real world data like this is hard. If this was an easy problem, then it would not be one worth studying, and the world simply doesn't work that way. There is randomness that we will never be able to account for, such as the weather causing farmers to have a good or bad harvest, or in human decision making when an individual decides to buy or sell when they otherwise might not have. Data like this is similar to stock market data, and being able to accurately predict real world data like this is an open problem.

Although this is a hard problem, we see that even with relatively simple models, we are able to make somewhat intelligent predictions. When we made predictions on data held out from training, we were able to make very accurate predictions for three or four days into the future. In addition, as we predicted further into the future, our predictions were not wildly inaccurate, and we always captured the true value of the stock within our prediction interval. Being able to perform analysis like this using simple models is useful. One could attempt to build a very complex model using a neural network or other sophisticated models, but there is value in applying a simple model and seeing what happens. In this case, we didn't discover anything groundbreaking, but we were able to make predictions that could be useful in the real world to an individual trying to decide if they should sell their crop now or wait a few days before selling.

In addition, we saw statistically significant evidence that crop prices perform differently during a recession than in a regular economic climate. During regular times, we saw significance of lags as we looked at ACF and PACF plots, which suggested models to use, but in all four crops during the Great Recession, we did not see any compelling significance. This means that our data resembled white noise more during a recession than regularly. One would expect this since recessions causes corporations and individuals to act differently and more randomly, but we see sound statistical evidence of this.

Finally, we see that one can use the current and previous price of crops to predict the current price of another crop. The real world use of this may be limited, but if a use could be found in some market where it would apply, an individual could greatly benefit from this.

There is significant future work that could be done for a study like this. We saw that cotton acted differently than the other three crops in several cases. This could be investigated, and cotton could be compared to other crops that have similar uses, like hemp for example, since the uses of cotton are inherently different than the uses of corn, soybeans, and wheat. Next, seasonality could be applied to the models. There was no obvious seasonality in the plots, but this could be investigated further. Adding a useful seasonality term could significantly increase the accuracy of our models. Also, more sophisticated methods could be applied to this data to try to make more accurate predictions. Models like Hidden Markov Models or recurrent neural networks are commonly applied to time series data and may work better than the models we tried in this study. Finally, these same methods could be applied to other time series data, like other agricultural markets, e.g. livestock or other crops, the stock market, cryptocurrency market, or basically any market where goods are bought and sold. It would be interesting to see how the markets differ between each other, and see where accurate predictions can be made versus predictions on data that is seemingly random.

5 Data Sets

https://www.macrotrends.net/2532/corn-prices-historical-chart-data

https://www.macrotrends.net/2533/cotton-prices-historical-chart-data

https://www.macrotrends.net/2531/soybean-prices-historical-chart-data

https://www.macrotrends.net/2534/wheat-prices-historical-chart-data