APPLE PRICES ANALYSIS AND FORECASTING

TIME SERIES MODELLING

By

IFEANYI OKWUCHI

20774990

SYSTEMS DESIGN ENGINEERING, MASc

SUPERVISOR

PROFESSOR KUMARASWAMY PONNAMBALAM

Contents

INTRODUCTION AND BACKGROUND

The price of food products is largely assumed to increase over time. In this study, an analysis is done for fresh food products to see how the prices behave and then an attempt is made to forecast future prices.

The case study here is apples. In particular, we consider apple prices received by growers in the United States from 1980 till date. The data spans from January 1980 to August 2018 as this was the only period where there was consistent data for this particular product. The data is obtained from the United States Department of Agriculture Economic research Service. (USDA-ERS).

The time series being analyzed here is a monthly time series. An exploratory data analysis is done to understand the characteristics of the data after which a confirmatory data analysis is done to confirm the initial inferences.

ARMA models are fit to both the monthly time series, a forecasting experiment is done to check the model accuracy and minimum mean square error forecasts are used to predict future prices.

EXPLORATORY DATA ANALYSIS

Before any analysis is carried out, an exploratory data analysis is important to provide insight regarding the appearance of the series.

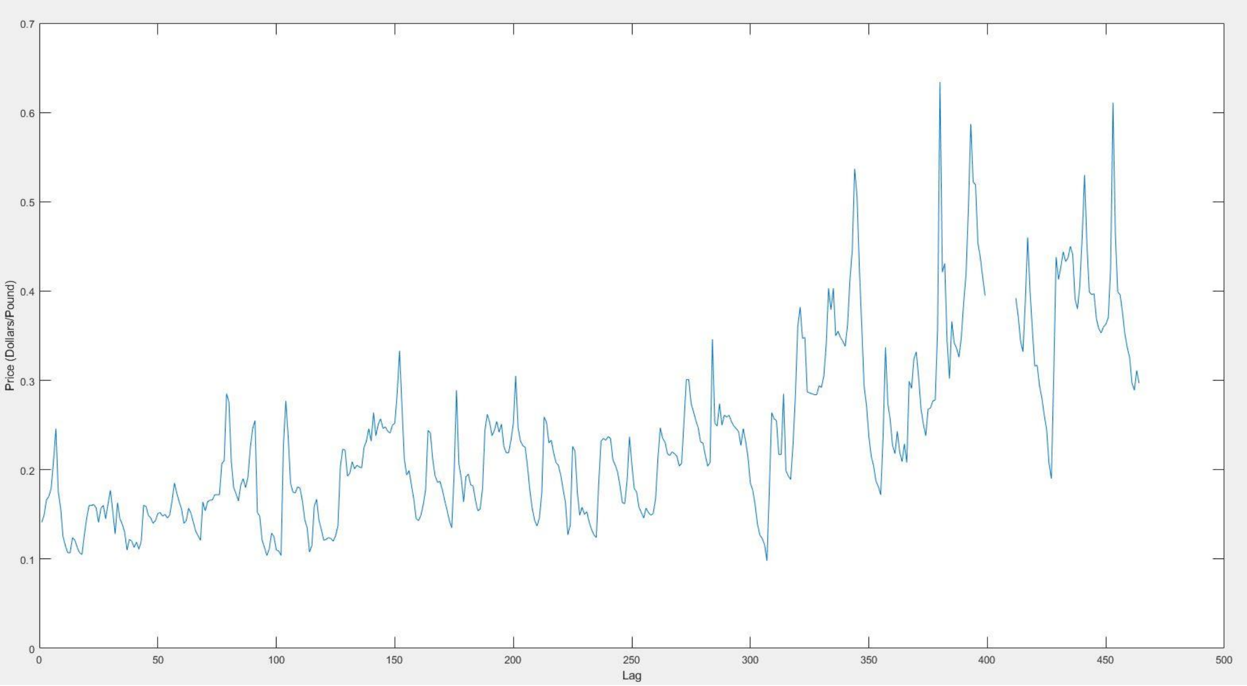


Figure 1: Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date Showing Missing Points.

The plot of the original time series data shows that the series is non-stationary over time. It also appears that the variance of the series is increasing with time hence the series can be said to be heteroscedastic. It looks like there is randomness in the data and there also seems to be a trend.

It should be noted that there are several missing data points which sum up to one year from April 2013 to March 2014.

These missing points are estimated using the Least Square Approximations as described by Fung (2006). The algorithm utilized data before and after the missing points to estimate the missing values.

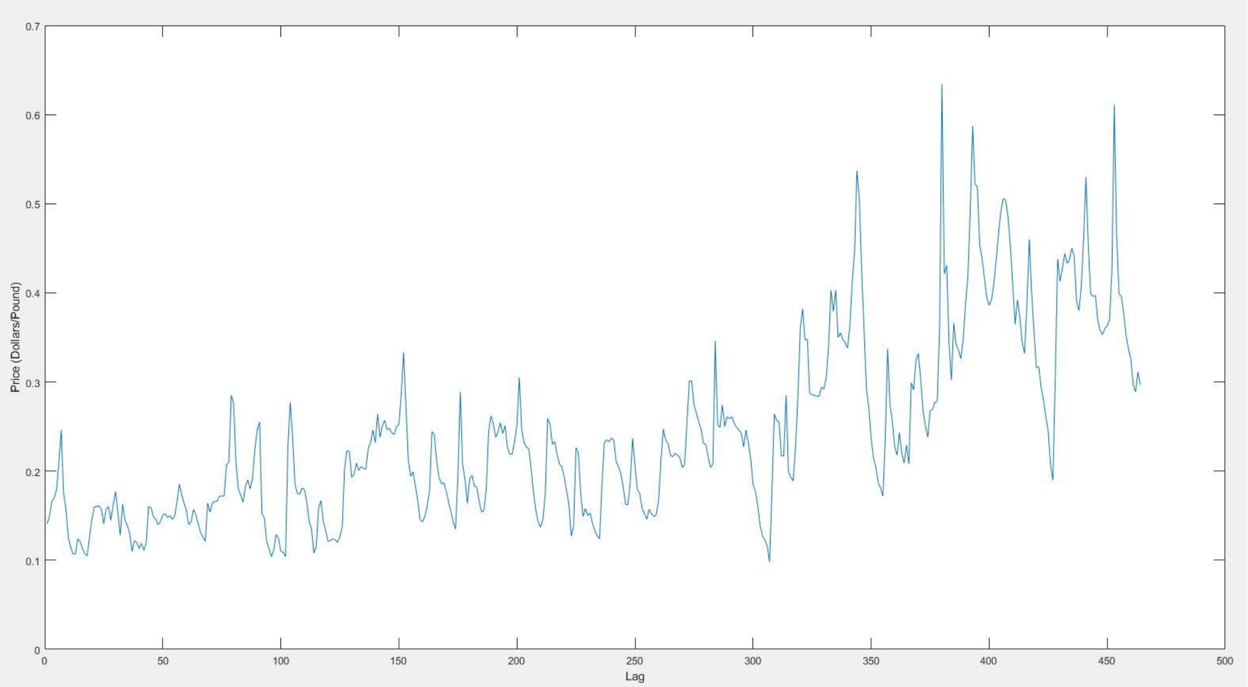


Figure 2: Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date Showing Complete Data Points.

CONFIRMATORY DATA ANALYSIS

From exploratory data analysis, we determine that the data is non-stationary. We have to confirm this information as well as fit a range of models to the data.

We start by carrying out a Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test to confirm if the data is truly non-stationary as determined earlier. This test is basically a hypothesis test where the null hypothesis says the series is stationary while the alternate hypothesis states that the series is non-stationary. Carrying out this test on MATLAB, the output ***returns 1 when the null hypothesis is rejected and 0 when we do not reject the null hypothesis***.

Result: From the KPSS Test, the returned value is 1 hence it can be confirmed that the series is non-stationary.

Also, a plot of the ACF and PACF of the series shows that the series is non-stationary. The ACF of the series shows a very slow dying off process which is consistent with non-stationary time series.

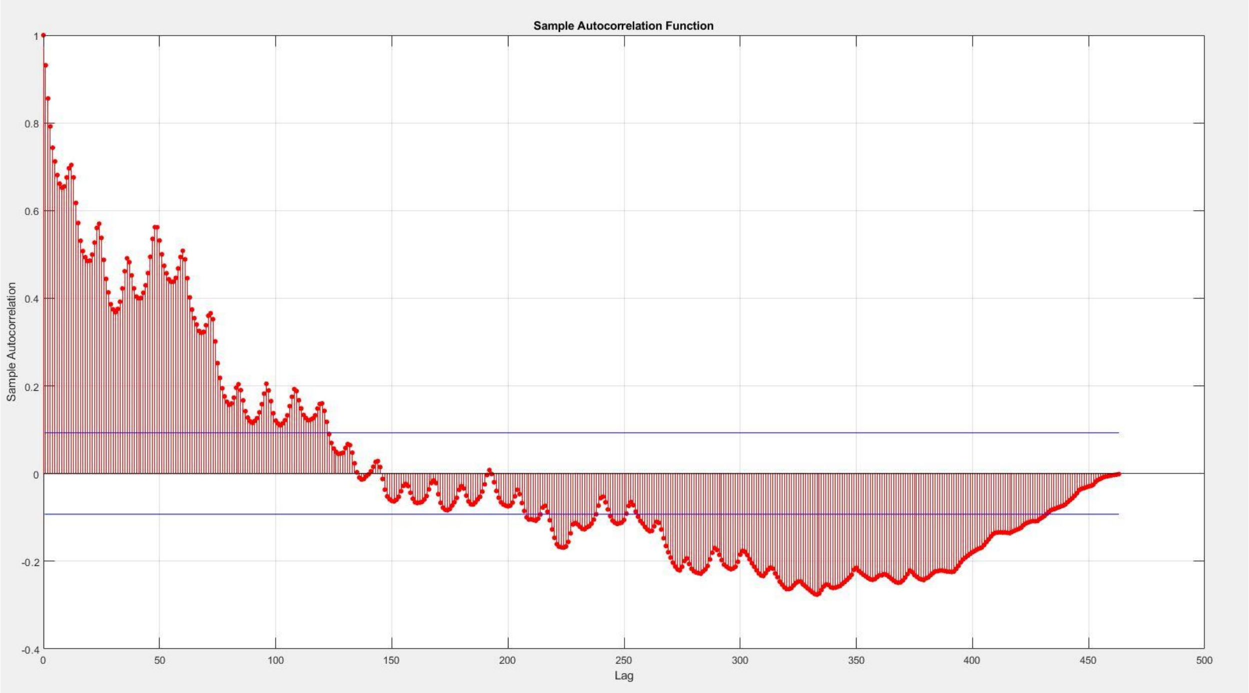


Figure 3: Sample ACF and 95% Confidence Limits for the Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date

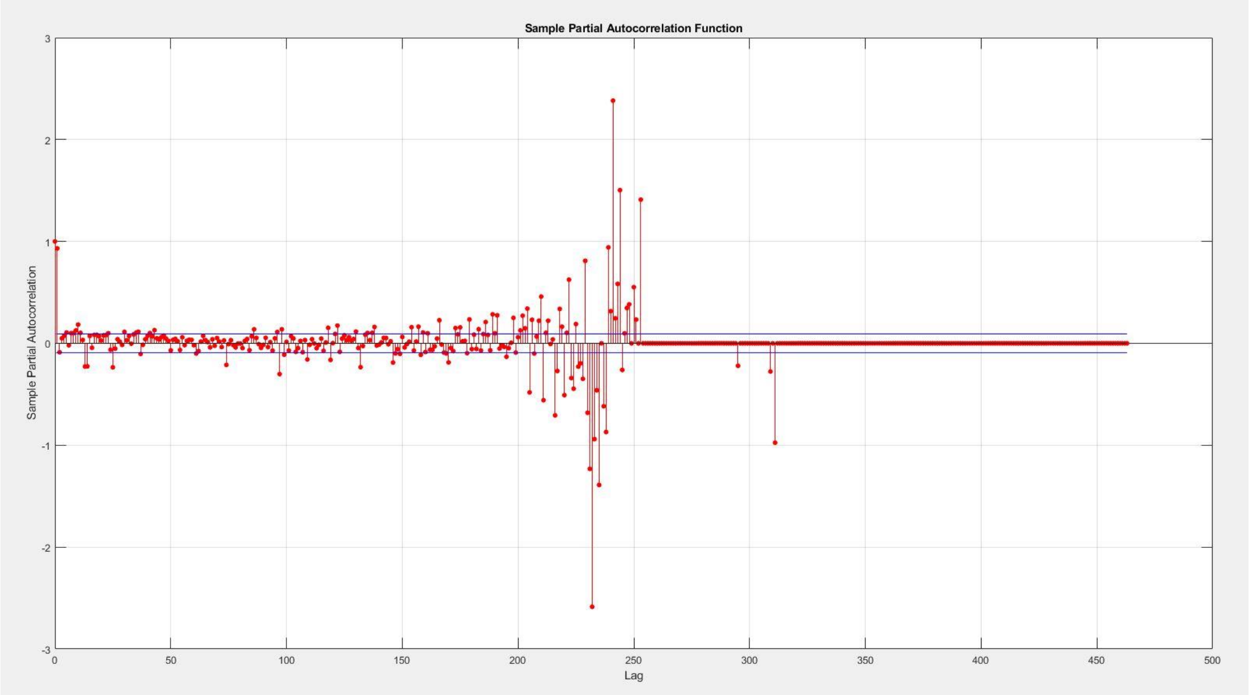


Figure 4: Sample PACF and 95% Confidence Limits for the Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date

Box-Cox Transformation and Differencing.

In modelling non-stationary time series, it is recommended to first transform the data using a Box-Cox Transformation to transform the original data which may not be normally distributed to a data set that is normally distributed. The transformation also removes heteroscedasticity from the data.

Using MATLAB, the following output was gotten from the Box-Cox transformation

Output: lambda = -0.3024, transdata is the transformed data set, Zt = 1.

The plot of the original data and that of the transformed data look very similar in shape and the transformed series also appears non-stationary. The non-stationarity is confirmed in the transformed series by the KPSS test as well since Zt =1

To remove the non-stationarity, we use differencing. We difference the already transformed data and carry out another KPSS Test.

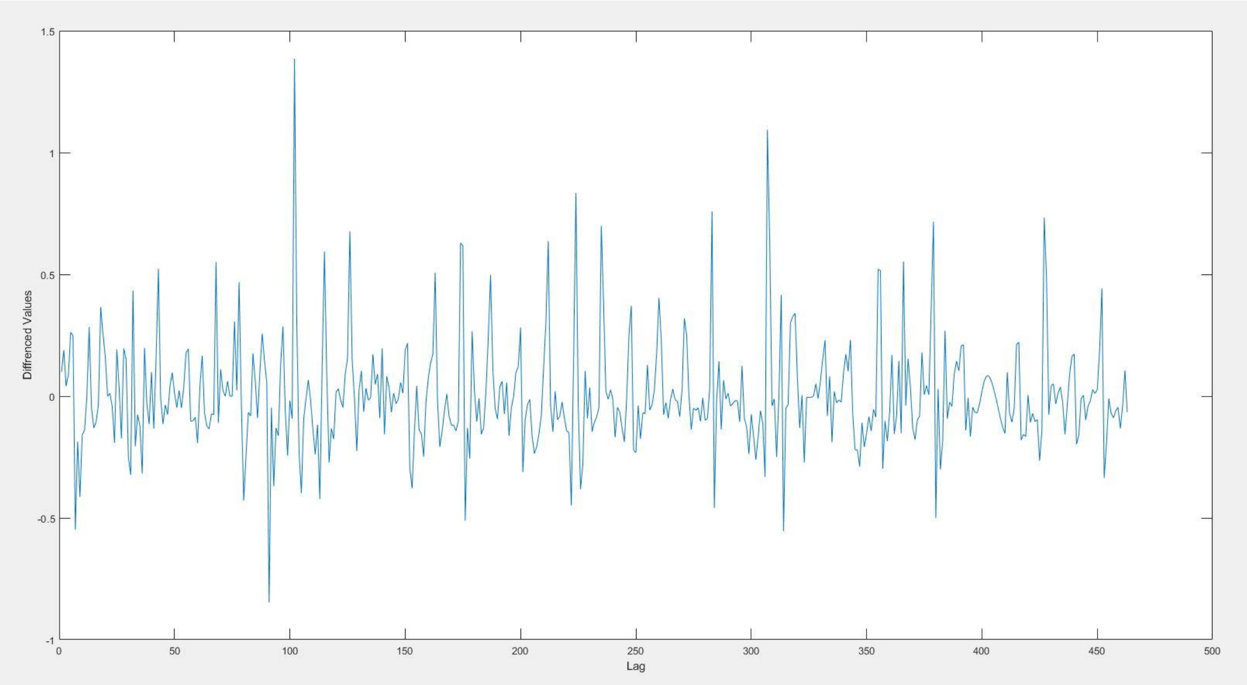


Figure 5: Box-Cox Transformation of Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date.

The value of Zd=0 implies that the series is now a stationary series after differencing was applied. The plot on figure 5 confirms the KPSS test. It shows the differenced data is stationary with an overall mean level around 0. Since the data is stationary after differencing once, there's no need for any further differencing. Note that differencing reduces the number of data points by 1.

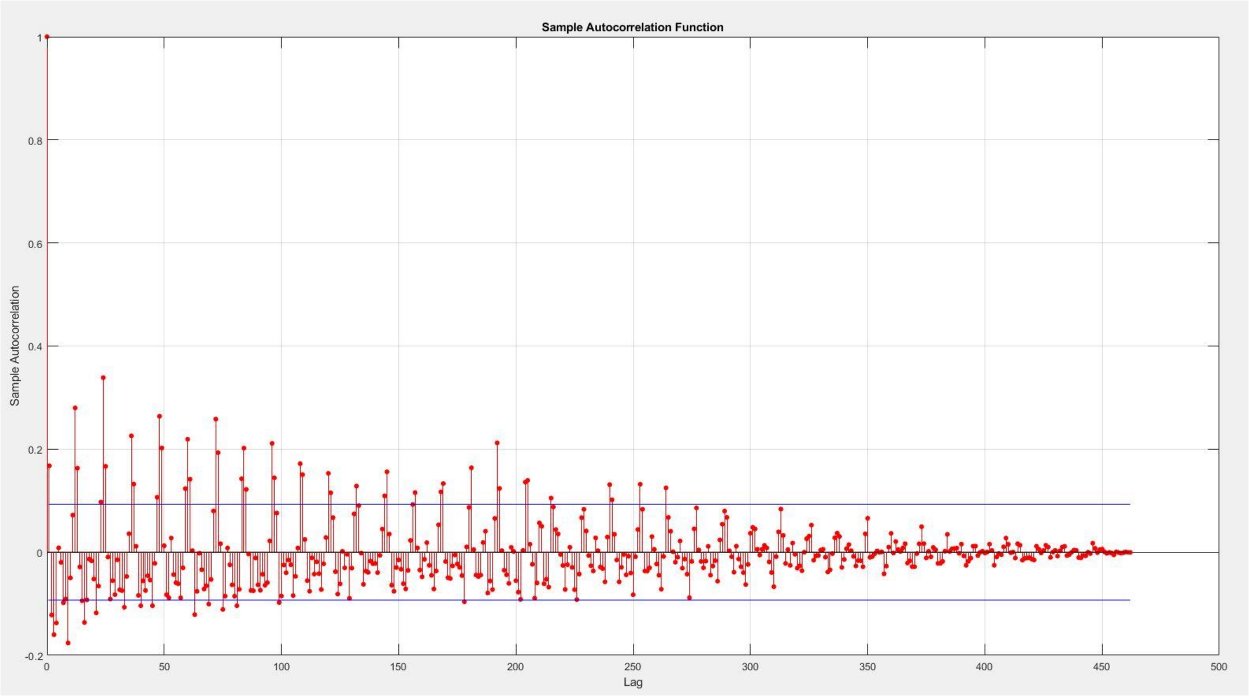


Figure 6: Sample ACF and 95% Confidence Limits for the Differenced Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date

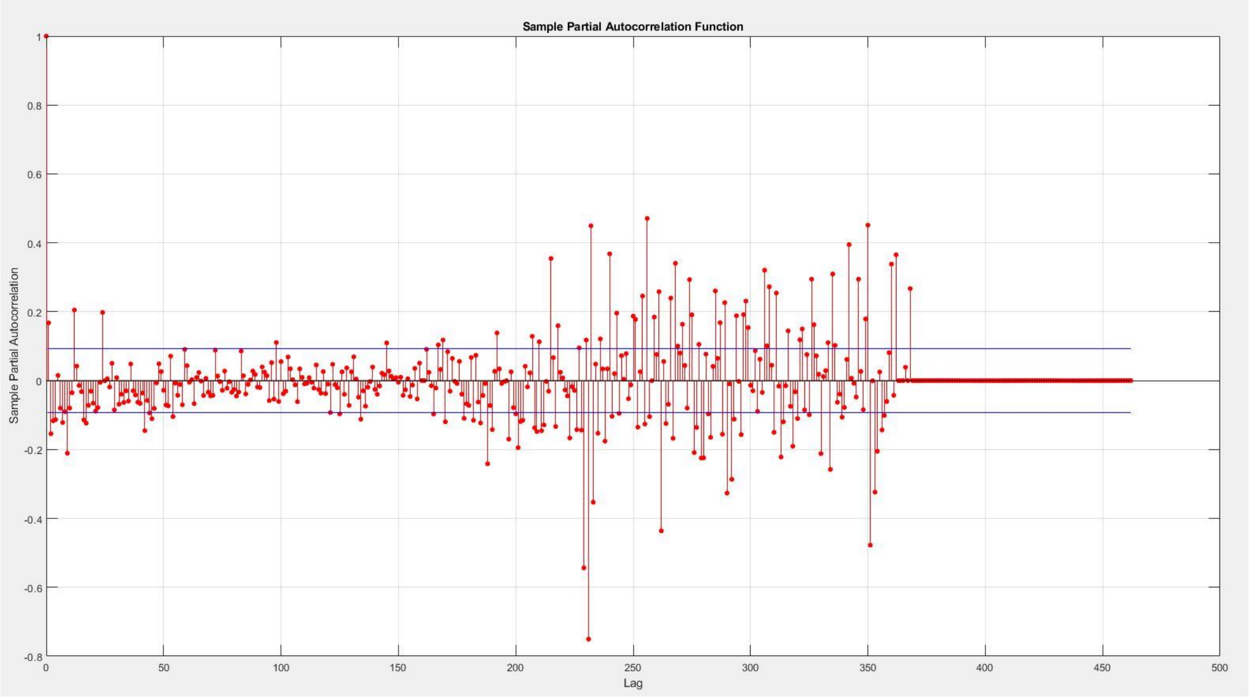


Figure 7: Sample PACF and 95% Confidence Limits for the Differenced Monthly Fresh Apple Prices Received by Growers in the US from 1980 till Date

PARAMETER ESTIMATION

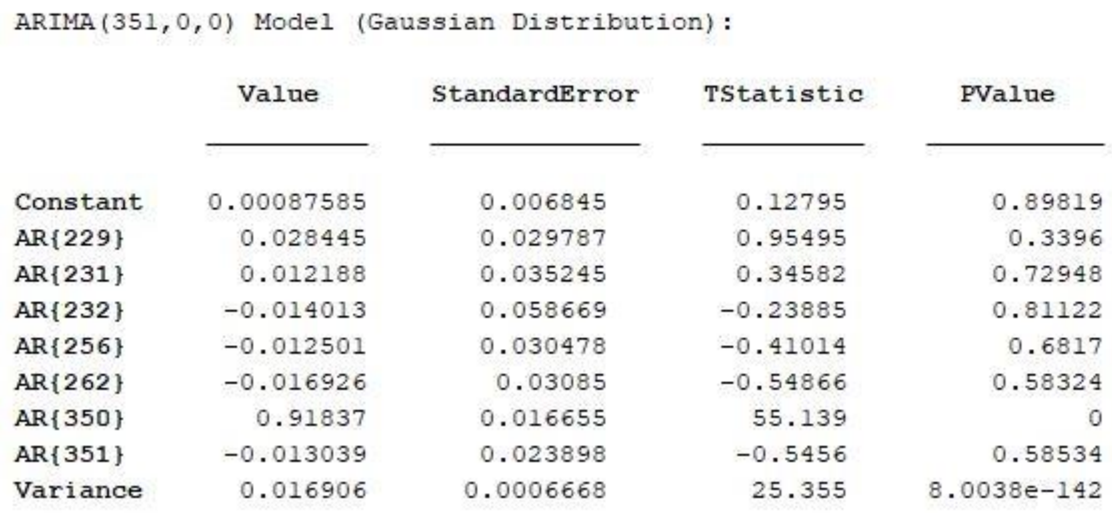
From the ACF and PACF of the differenced series, it is suggested that an AR model be fit to the data because the ACF decays gradually and the PACF cuts off after a particular lag.

However, choosing the best AR model to fit to the data is a very tricky one because the PACF exhibits significant lags 128 times. A model with this many parameters would be computationally expensive to estimate because it would be a constrained AR(362) with 128 AR parameters. Also, estimating the parameters of the model using our data was not possible as the software could not solve the problem.

To resolve this, models were selected considering as many of the most significant lags as possible so that the parameters could be estimated.

Four different AR models would be considered; a constrained AR(351) with significant values at lags 229,231,232,256,262,350,351, a constrained AR(351) with significant values at lags 1,2,4,9,12,24,42,229,231,232,256,262,350,351; a constrained AR(362) with significant values at lags 1,2,9,12,24,42,215,229,231,232,233,240,256,262,342,350,351,362 and finally a constrained AR(362) with significant values at lags
1,2,3,4,7,9,12,16,17,24,42,215,229,231,232,233,240,256,262,342,350,351,362.

For models 1, 2, 3, and 4 respectively, we obtain the following outputs from MATLAB.

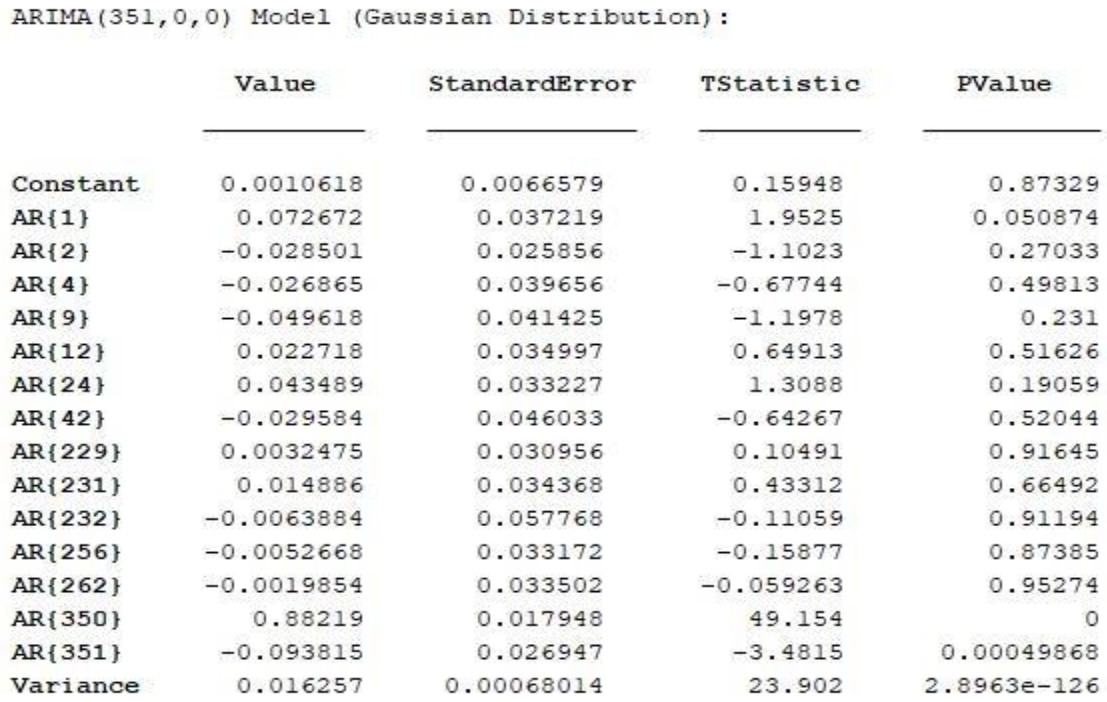


```
ARIMA(351,0,0) Model (Gaussian Distribution):

                  Value       StandardError    TStatistic      PValue
                __________    _____________    __________    __________

Constant        0.00087585       0.006845        0.12795        0.89819
AR{229}           0.028445       0.029787        0.95495         0.3396
AR{231}           0.012188       0.035245        0.34582        0.72948
AR{232}          -0.014013       0.058669       -0.23885        0.81122
AR{256}          -0.012501       0.030478       -0.41014         0.6817
AR{262}          -0.016926        0.03085       -0.54866        0.58324
AR{350}            0.91837       0.016655         55.139              0
AR{351}          -0.013039       0.023898        -0.5456        0.58534
Variance          0.016906      0.0006668         25.355    8.0038e-142
```

Figure 8: Image of MATLAB Output of Model 1 Parameter Estimates

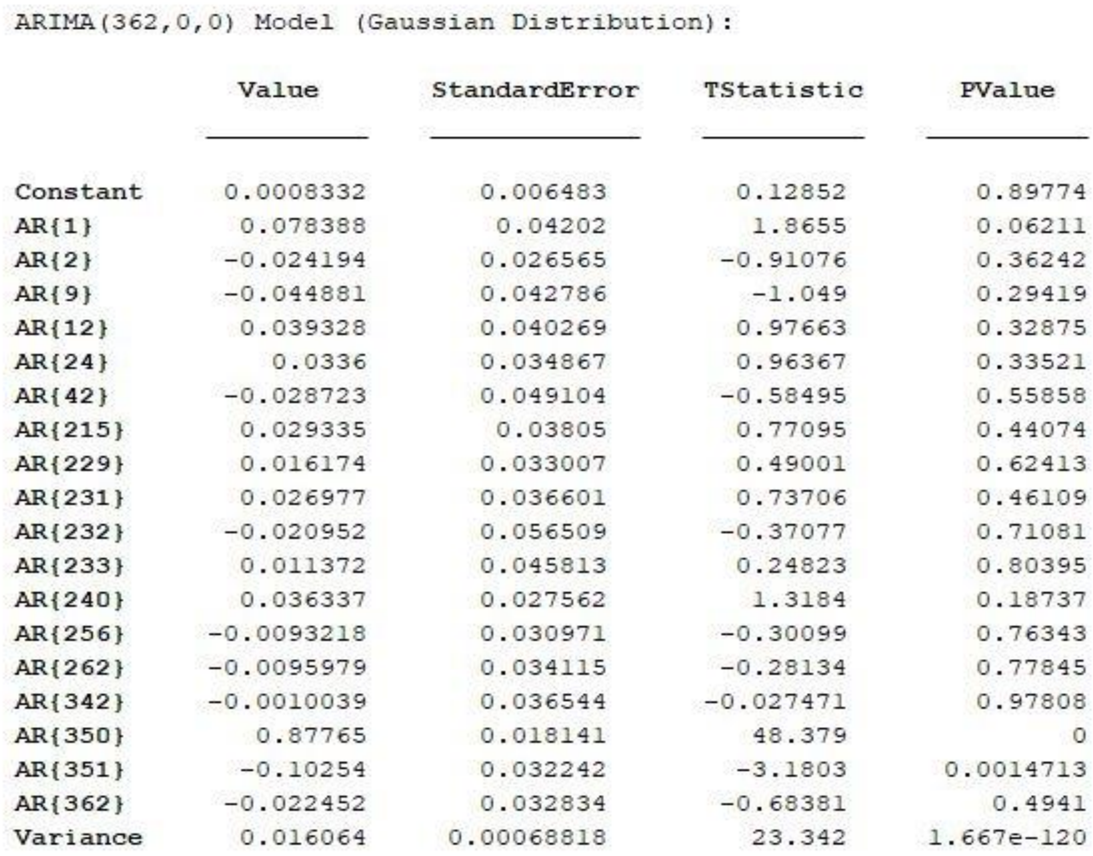


```
ARIMA(351,0,0) Model (Gaussian Distribution):

                 Value        StandardError    TStatistic      PValue   
              ___________    _____________    __________    ____________

Constant       0.0010618        0.0066579       0.15948          0.87329
AR{1}           0.072672         0.037219        1.9525         0.050874
AR{2}          -0.028501         0.025856       -1.1023          0.27033
AR{4}          -0.026865         0.039656      -0.67744          0.49813
AR{9}          -0.049618         0.041425       -1.1978            0.231
AR{12}          0.022718         0.034997       0.64913          0.51626
AR{24}          0.043489         0.033227        1.3088          0.19059
AR{42}         -0.029584         0.046033      -0.64267          0.52044
AR{229}        0.0032475         0.030956       0.10491          0.91645
AR{231}         0.014886         0.034368       0.43312          0.66492
AR{232}       -0.0063884         0.057768      -0.11059          0.91194
AR{256}       -0.0052668         0.033172      -0.15877          0.87385
AR{262}       -0.0019854         0.033502     -0.059263          0.95274
AR{350}          0.88219         0.017948        49.154                0
AR{351}        -0.093815         0.026947       -3.4815       0.00049868
Variance        0.016257       0.00068014        23.902      2.8963e-126
```

Figure 9: Image of MATLAB Output of Model 2 Parameter Estimates

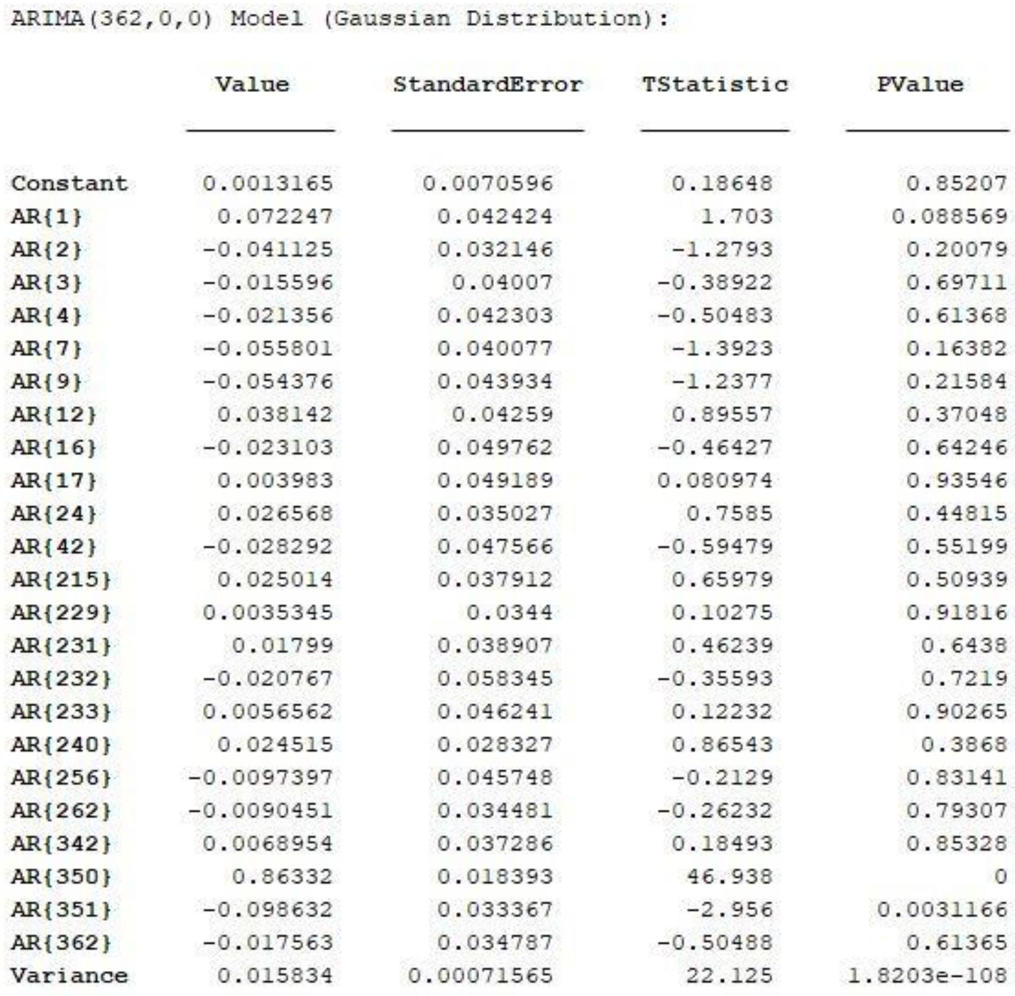


```
ARIMA(362,0,0) Model (Gaussian Distribution):

                 Value        StandardError    TStatistic      PValue
              ___________    ______________    __________    __________

Constant       0.0008332       0.006483          0.12852       0.89774
AR{1}           0.078388        0.04202           1.8655       0.06211
AR{2}          -0.024194       0.026565         -0.91076       0.36242
AR{9}          -0.044881       0.042786           -1.049       0.29419
AR{12}          0.039328       0.040269          0.97663       0.32875
AR{24}            0.0336       0.034867          0.96367       0.33521
AR{42}         -0.028723       0.049104         -0.58495       0.55858
AR{215}         0.029335        0.03805          0.77095       0.44074
AR{229}         0.016174       0.033007          0.49001       0.62413
AR{231}         0.026977       0.036601          0.73706       0.46109
AR{232}        -0.020952       0.056509         -0.37077       0.71081
AR{233}         0.011372       0.045813          0.24823       0.80395
AR{240}         0.036337       0.027562           1.3184       0.18737
AR{256}       -0.0093218       0.030971         -0.30099       0.76343
AR{262}       -0.0095979       0.034115         -0.28134       0.77845
AR{342}       -0.0010039       0.036544        -0.027471       0.97808
AR{350}          0.87765       0.018141           48.379             0
AR{351}         -0.10254       0.032242          -3.1803     0.0014713
AR{362}        -0.022452       0.032834         -0.68381        0.4941
Variance        0.016064     0.00068818           23.342    1.667e-120
```

Figure 10: Image of MATLAB Output of Model 3 Parameter Estimates



```
ARIMA(362,0,0) Model (Gaussian Distribution):

                  Value        StandardError    TStatistic      PValue
               ___________    _____________    ___________    ___________

Constant       0.0013165        0.0070596        0.18648        0.85207
AR{1}           0.072247         0.042424          1.703       0.088569
AR{2}          -0.041125         0.032146        -1.2793        0.20079
AR{3}          -0.015596          0.04007       -0.38922        0.69711
AR{4}          -0.021356         0.042303       -0.50483        0.61368
AR{7}          -0.055801         0.040077        -1.3923        0.16382
AR{9}          -0.054376         0.043934        -1.2377        0.21584
AR{12}          0.038142          0.04259        0.89557        0.37048
AR{16}         -0.023103         0.049762       -0.46427        0.64246
AR{17}          0.003983         0.049189       0.080974        0.93546
AR{24}          0.026568         0.035027         0.7585        0.44815
AR{42}         -0.028292         0.047566       -0.59479        0.55199
AR{215}         0.025014         0.037912        0.65979        0.50939
AR{229}        0.0035345           0.0344        0.10275        0.91816
AR{231}          0.01799         0.038907        0.46239         0.6438
AR{232}        -0.020767         0.058345       -0.35593         0.7219
AR{233}        0.0056562         0.046241        0.12232        0.90265
AR{240}         0.024515         0.028327        0.86543         0.3868
AR{256}       -0.0097397         0.045748        -0.2129        0.83141
AR{262}       -0.0090451         0.034481       -0.26232        0.79307
AR{342}        0.0068954         0.037286        0.18493        0.85328
AR{350}          0.86332         0.018393         46.938              0
AR{351}        -0.098632         0.033367         -2.956      0.0031166
AR{362}        -0.017563         0.034787       -0.50488        0.61365
Variance        0.015834       0.00071565         22.125    1.8203e-108
```

Figure 11: Image of MATLAB Output of Model 4 Parameter Estimates

MODEL SELECTION

The best model from the lot is selected using Akaike's Information Criterion. The model with the least AIC value is the best model to fit to the data. The AIC values are evaluated from MATLAB and the output is given as;

Table 1: AIC Values of the Various Models

Model Type	AIC Value
Model 1	-557.1330
Model 2	-561.2666
Model 3	-558.8082
Model 4	-555.4798

From the results of the AIC, Model 2 has the least AIC value and it will be selected.

DIAGNOSTIC TEST

The next stage of the model development process is the diagnostic test. After a model has been selected, we check to see if the residuals are white, if it is homoscedastic and if it follows a normal distribution. Here, we try to make sure that the ACF of the residuals are not significant. In essence, were trying to ensure the residuals are white noise without any correlation.

To carry out the diagnostic test, we first simulate data using the model we built. After simulating the data, we then subtract the values of the simulated data from the actual differenced data to get the residuals. This is calculated by the MATLAB infer function. The next step is to plot the autocorrelation function of the residuals to see if it is effectively white noise.

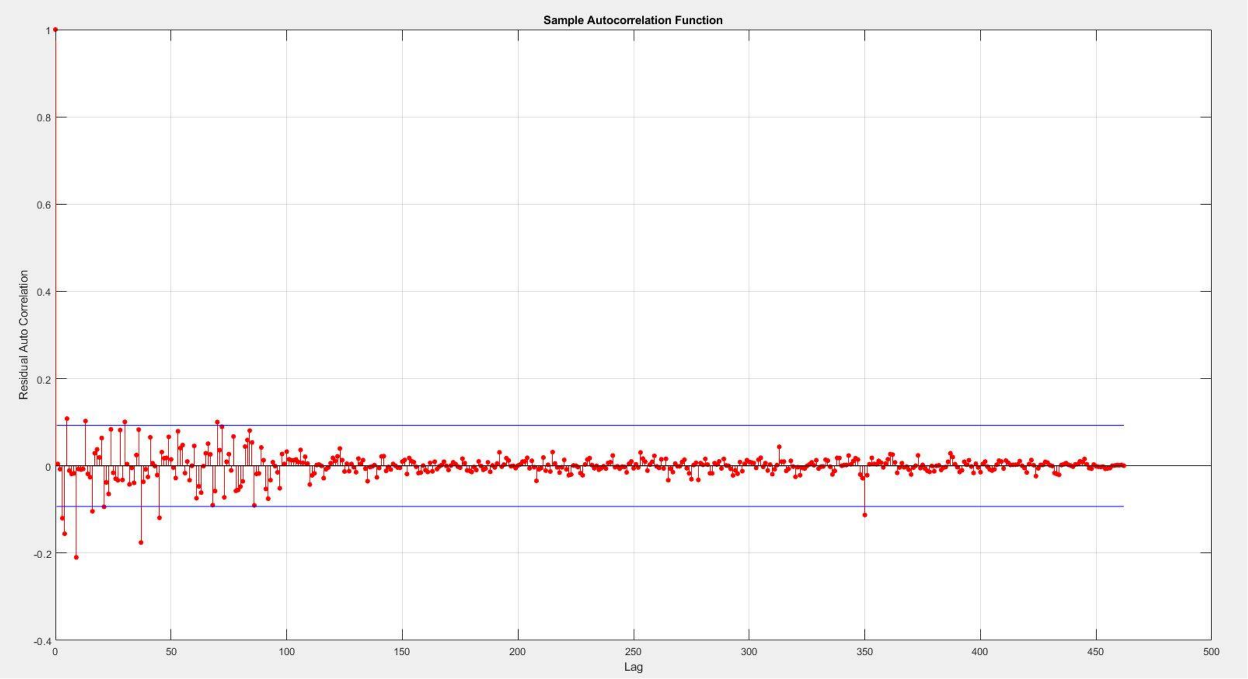


Figure 12: RACF and 95% Confidence Limits for the Selected, Constrained AR(351) Model

The residuals mostly fall within the 95% confidence interval with about 9 points appearing slightly significant. The residuals may or may not be assumed to be white noise but confirm this, a Ljung-Box test should be carried out.

The Ljung-Box test is a test statistic which states that If h=1, we reject the null hypothesis that there is no autocorrelation and if h=0, we do not reject the null hypothesis. This implies that when h=1, there is a correlation and when h=0, there's no correlation and the residuals are white.

Output:

H = 1.

The zero value implies that the residuals of the series are not white noise. The next step given this discovery should be to go backwards and select a different model with probably more parameters but doing so in this case implies that the parameter estimates would not be solvable. One way to then check the accuracy of the selected model is to perform a forecasting experiment.

FORECASTING EXPERIMENT

A forecasting experiment would be carried out using the all of the data less the last data point to predict the last data point. The forecast value would then be compared to the actual value. The forecasted value is the differenced value in the transformed domain hence it must be converted back to the original domain. The forecasting, differencing removal and inverse transformation are done in MATLAB.

After running the forecast on MATLAB, the output is given as;

$$\hat{Y}_{464} = 0.3119$$

The actual value at that last data point is 0.2970.

Comparing both values, the percentage error is 4.8%. It can be assumed that the forecast is not significantly different from the actual value.

FORECASTING

We can now use our model to forecast the prices of apples in future months. Here, minimum mean square error (MMSE) forecasts would be used. A forecast for a 10 year period will be calculated using one step ahead forecasting.

Recall that the model was built on the differenced, transformed series. As a result, we must first convert the forecast to the original transformed version by eliminating the differencing after which we apply an inverse Box-Cox transformation.

Table 2: Forecast of Apple Prices for the Next 10 Months

Steps Ahead (Months)	Price Forecast (dollars/pound)
1	0.2978
2	0.2986
3	0.2994
4	0.3002
5	0.3010
6	0.3018
7	0.3026
8	0.3034
9	0.3042
10	0.3050

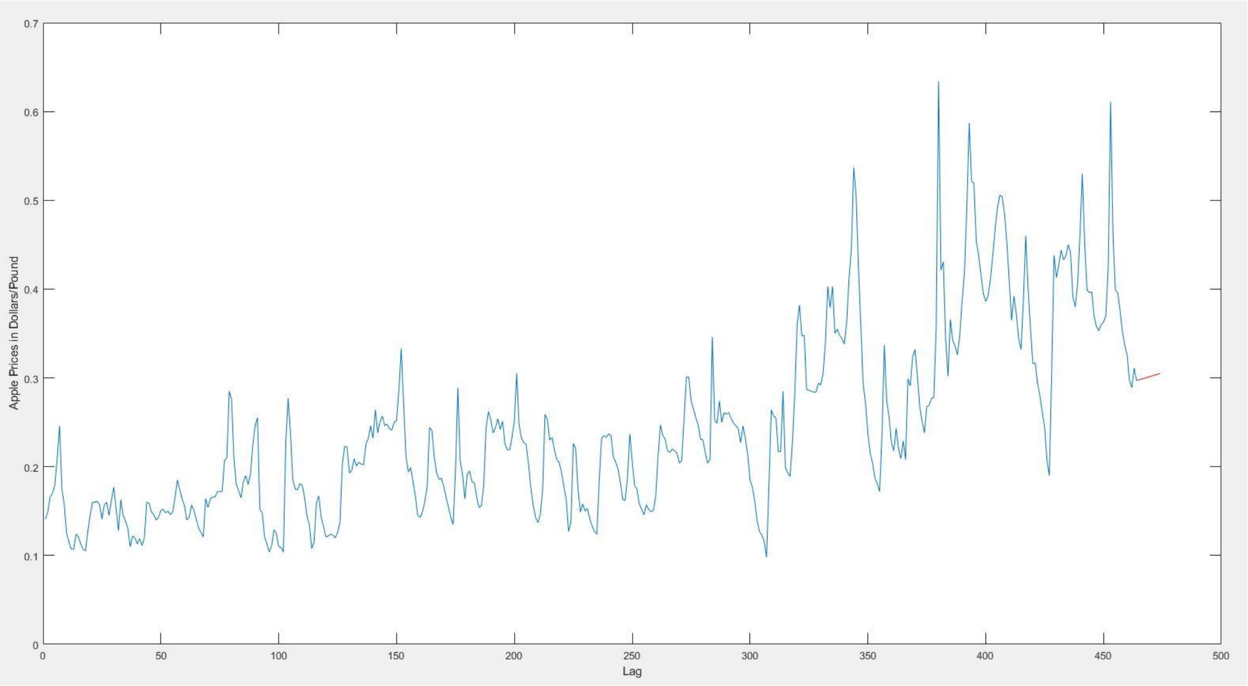
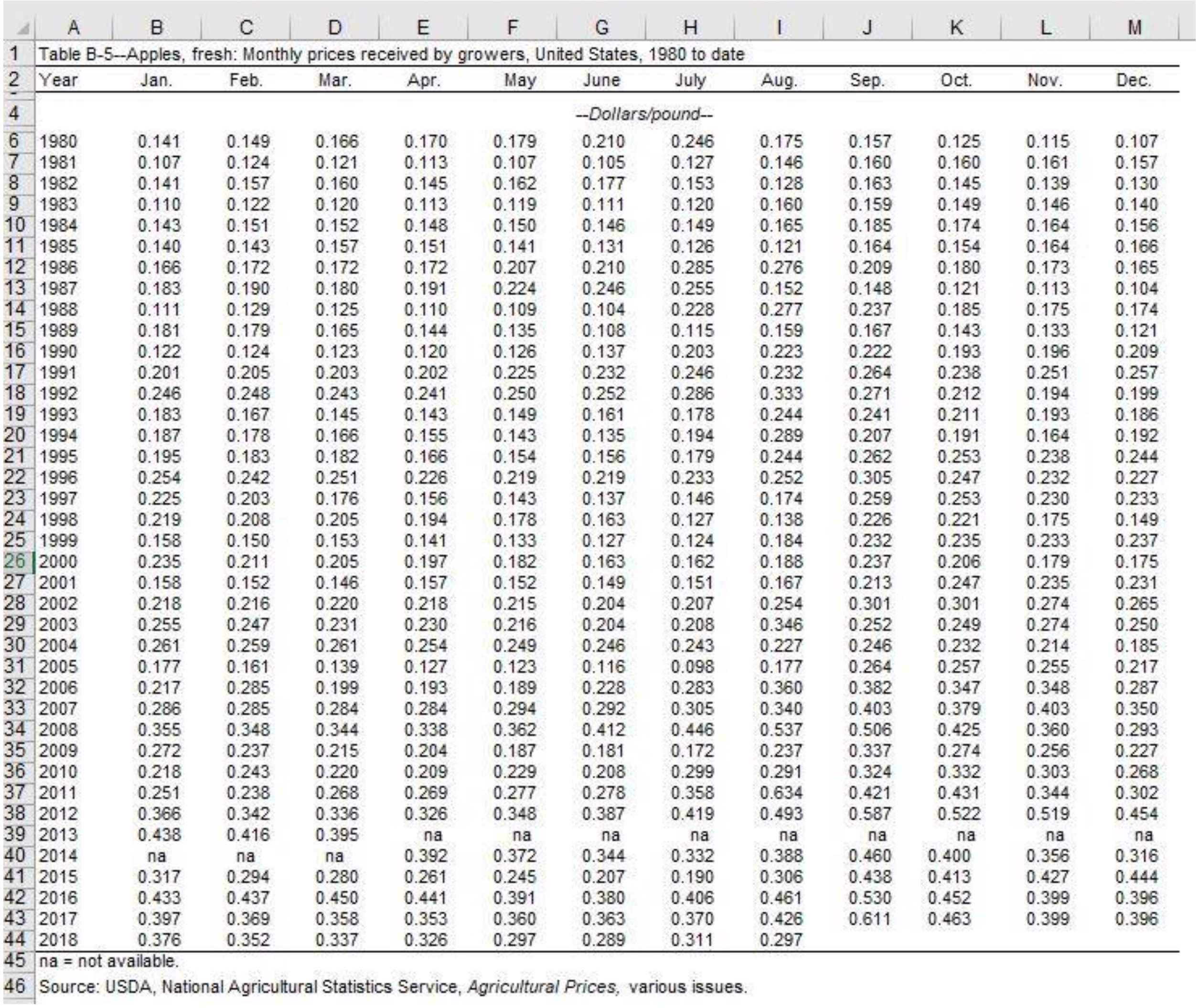


Figure 13: Time Series Plot of the Data Showing the Forecasts

REFERENCES

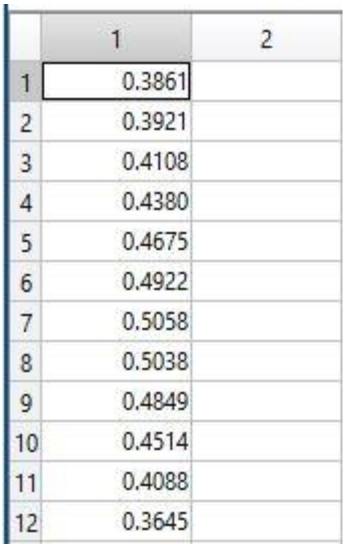
1. D.S Fung (2006). Methods for the Estimation of Missing Values in Time series. Retrieved from https://ro.ecu.edu.au/theses/63

APPENDIX



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Table B-5--Apples, fresh: Monthly prices received by growers, United States, 1980 to date												
2	Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
4							*--Dollars/pound--*						
6	1980	0.141	0.149	0.166	0.170	0.179	0.210	0.246	0.175	0.157	0.125	0.115	0.107
7	1981	0.107	0.124	0.121	0.113	0.107	0.105	0.127	0.146	0.160	0.160	0.161	0.157
8	1982	0.141	0.157	0.160	0.145	0.162	0.177	0.153	0.128	0.163	0.145	0.139	0.130
9	1983	0.110	0.122	0.120	0.113	0.119	0.111	0.120	0.160	0.159	0.149	0.146	0.140
10	1984	0.143	0.151	0.152	0.148	0.150	0.146	0.149	0.165	0.185	0.174	0.164	0.156
11	1985	0.140	0.143	0.157	0.151	0.141	0.131	0.126	0.121	0.164	0.154	0.164	0.166
12	1986	0.166	0.172	0.172	0.172	0.207	0.210	0.285	0.276	0.209	0.180	0.173	0.165
13	1987	0.183	0.190	0.180	0.191	0.224	0.246	0.255	0.152	0.148	0.121	0.113	0.104
14	1988	0.111	0.129	0.125	0.110	0.109	0.104	0.228	0.277	0.237	0.185	0.175	0.174
15	1989	0.181	0.179	0.165	0.144	0.135	0.108	0.115	0.159	0.167	0.143	0.133	0.121
16	1990	0.122	0.124	0.123	0.120	0.126	0.137	0.203	0.223	0.222	0.193	0.196	0.209
17	1991	0.201	0.205	0.203	0.202	0.225	0.232	0.246	0.232	0.264	0.238	0.251	0.257
18	1992	0.246	0.248	0.243	0.241	0.250	0.252	0.286	0.333	0.271	0.212	0.194	0.199
19	1993	0.183	0.167	0.145	0.143	0.149	0.161	0.178	0.244	0.241	0.211	0.193	0.186
20	1994	0.187	0.178	0.166	0.155	0.143	0.135	0.194	0.289	0.207	0.191	0.164	0.192
21	1995	0.195	0.183	0.182	0.166	0.154	0.156	0.179	0.244	0.262	0.253	0.238	0.244
22	1996	0.254	0.242	0.251	0.226	0.219	0.219	0.233	0.252	0.305	0.247	0.232	0.227
23	1997	0.225	0.203	0.176	0.156	0.143	0.137	0.146	0.174	0.259	0.253	0.230	0.233
24	1998	0.219	0.208	0.205	0.194	0.178	0.163	0.127	0.138	0.226	0.221	0.175	0.149
25	1999	0.158	0.150	0.153	0.141	0.133	0.127	0.124	0.184	0.232	0.235	0.233	0.237
26	2000	0.235	0.211	0.205	0.197	0.182	0.163	0.162	0.188	0.237	0.206	0.179	0.175
27	2001	0.158	0.152	0.146	0.157	0.152	0.149	0.151	0.167	0.213	0.247	0.235	0.231
28	2002	0.218	0.216	0.220	0.218	0.215	0.204	0.207	0.254	0.301	0.301	0.274	0.265
29	2003	0.255	0.247	0.231	0.230	0.216	0.204	0.208	0.346	0.252	0.249	0.274	0.250
30	2004	0.261	0.259	0.261	0.254	0.249	0.246	0.243	0.227	0.246	0.232	0.214	0.185
31	2005	0.177	0.161	0.139	0.127	0.123	0.116	0.098	0.177	0.264	0.257	0.255	0.217
32	2006	0.217	0.285	0.199	0.193	0.189	0.228	0.283	0.360	0.382	0.347	0.348	0.287
33	2007	0.286	0.285	0.284	0.284	0.294	0.292	0.305	0.340	0.403	0.379	0.403	0.350
34	2008	0.355	0.348	0.344	0.338	0.362	0.412	0.446	0.537	0.506	0.425	0.360	0.293
35	2009	0.272	0.237	0.215	0.204	0.187	0.181	0.172	0.237	0.337	0.274	0.256	0.227
36	2010	0.218	0.243	0.220	0.209	0.229	0.208	0.299	0.291	0.324	0.332	0.303	0.268
37	2011	0.251	0.238	0.268	0.269	0.277	0.278	0.358	0.634	0.421	0.431	0.344	0.302
38	2012	0.366	0.342	0.336	0.326	0.348	0.387	0.419	0.493	0.587	0.522	0.519	0.454
39	2013	0.438	0.416	0.395	na	na	na	na	na	na	na	na	na
40	2014	na	na	na	0.392	0.372	0.344	0.332	0.388	0.460	0.400	0.356	0.316
41	2015	0.317	0.294	0.280	0.261	0.245	0.207	0.190	0.306	0.438	0.413	0.427	0.444
42	2016	0.433	0.437	0.450	0.441	0.391	0.380	0.406	0.461	0.530	0.452	0.399	0.396
43	2017	0.397	0.369	0.358	0.353	0.360	0.363	0.370	0.426	0.611	0.463	0.399	0.396
44	2018	0.376	0.352	0.337	0.326	0.297	0.289	0.311	0.297				
45	na = not available.												
46	Source: USDA, National Agricultural Statistics Service, *Agricultural Prices,* various issues.												

Figure 1A: Monthly Prices of Fresh Apples Received by Growers in the United States from 1980 to Date



	1	2
1	0.3861	
2	0.3921	
3	0.4108	
4	0.4380	
5	0.4675	
6	0.4922	
7	0.5058	
8	0.5038	
9	0.4849	
10	0.4514	
11	0.4088	
12	0.3645	

Figure 2A: Least Square Approximations of Missing Data from the Apple Prices Series