2011

Time Series Analysis of Stock Prices Using the Box-Jenkins Approach

Shakira Green

TIME SERIES ANALYSIS OF STOCK PRICES USING THE BOX-JENKINS APPROACH

by

SHAKIRA GREEN

(Under the Direction of Patricia Humphrey)

ABSTRACT

A time series is a sequence of data points, typically measured at uniform time intervals. Examples occur in a variety of fields ranging from economics to engineering, and methods of analyzing time series constitute an important part of Statistics. Time series analysis comprises methods for analyzing time series data in order to extract meaningful characteristics of the data and forecast future values. The Autoregressive Integrated Moving Average (ARIMA) models, or Box-Jenkins methodology, are a class of linear models that are capable of representing stationary as well as nonstationary time series. ARIMA models rely heavily on autocorrelation patterns. This paper will explore the application of the Box-Jenkins approach to stock prices, in particular sampling at different time intervals in order to determine if there is some optimal frame and if there are similarities in autocorrelation patterns of stocks within the same industry.

INDEX WORDS: Statistics, Modeling, Stationary, Stock prices, ARIMA, Box-Jenkins

TIME SERIES ANALYSIS OF STOCK PRICES USING THE BOX-JENKINS APPROACH

by

SHAKIRA GREEN

B.S. in Mathematics

A thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment

of the Requirement for the Degree

MASTERS OF SCIENCE

STATESBORO, GEORGIA

2011

TIME SERIES ANALYSIS OF STOCK PRICES USING THE BOX-JENKINS APPROACH

by

SHAKIRA GREEN

Major Professor: Dr. P. Humphrey

Committee: Dr. J. Duggins

Dr. B. Oluyede

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Humphrey for her patience, support, and knowledge. I would also like to acknowledge my husband and family for keeping me sane, Dr. Lemma for encouraging me to keep going and GSU for the opportunity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1.1 Motivation

The stock market is a general term referring to the organized trading of securities through the various physical and electronic exchanges and the over the counter market. The stock market is one of the most vital areas of a market economy, because it provides companies with access to capital by allowing investors to buy shares of ownership in a company. By buying shares of ownership, investors stand to possibly gain money by profiting from companies' future prosperity. Although there are millions to be gained by buying shares and then selling them for a profit, not all investors are successful in gaining a return on their investment, and even fewer are successful in making a lot of money. This happens because the price of stocks is constantly fluctuating and at any given moment, the price of a stock could fall below the price at which it was bought; selling the shares at this reduced price results in the investors losing money. A natural solution to not losing money would be for investors to sell their shares before they begin to diminish in value, ideally at a point when the stock's price is higher than when it was purchased by the investor.

1.2 Introduction

A statistical approach to being able to determine when this desired point in time would be is to first look at a stock's price at various times of interest, and then represent this data as a time series. Then, after an analysis of the time series is carried out, an appropriate model can be used to forecast prices and possibly help the investor determine when to buy or sell. This chapter will provide the reader with the necessary background information on time series to be able to analyze the series, but also to identify the necessary model and make forecasts. Although this paper only highlights two modeling approaches: Box-Jenkins procedures and regression models, there are other techniques that may be more applicable to different areas of interest. This paper was written with the idea in mind that these procedures are being applied to stock values, but is not limited to only this case.

1.3 Background Information

A time series is a collection of observations made sequentially and typically equally spaced in time. The special feature of time series analysis is the fact that the analysis must take into account the time order because the successive observations are usually not independent observations, whereas most other statistical theory is concerned with random samples of independent observations. Methods of analyzing time series constitute an important area of statistics. Although there are several objectives that can be satisfied by analyzing a time series, they can all be classified as descriptive, explanatory, predictive, or control [1].

Time series are often examined in hopes of discovering a historical pattern that can be exploited in the preparation of a forecast. In order to identify this pattern, it is often convenient to think of a time series as consisting of several components, and in doing so, taking a descriptive approach. The components of a time series are trend, seasonal variation, cyclic changes, and irregular factors.

Trend is the long-term change in the mean level and often thought of as the underlying growth or decline component in the series. The seasonal component is concerned with the periodic fluctuation in the series within each year. Seasonal fluctuations are most often attributed to social customs or weather changes. This type of behavior where variation is annual in period is most often seen in time series for sales figures and temperature readings. Cyclic changes within a time series are similar to the seasonal component in that it is revealed by a wavelike pattern. Cyclic changes can be thought of as variation across a fixed period due to some physical cause other than seasonal effects. Cycles are normally confined to a particular fixed period and can be a behavior that takes place over a period of years.

Once the trend and cyclic variations have been accounted for, the remaining movement is attributed to irregular fluctuations and the resulting data is a series of residuals. This set of residuals is not always random, so this series of residuals is also analyzed to determine if all the cyclic variation has truly been removed.

When analyzing a time series, an essential step is to plot the observations against time and then join successive points with line segments. The line segments are not solely for aesthetics, but also to reinforce the feeling that a continuous time scale exists between the plotted points [1]. The plotted time series is used to obtain simple descriptive measures of the main properties of the series. This plot can immediately reveal features such as trend, seasonal variation, discontinuities, and outliers that may be present in the data.

As previously mentioned time series are also analyzed for the purpose of prediction or forecasting. When successive observations are dependent, future values may be predicted from past observations. If the future values of the time series can be predicted exactly then the times series is classified as deterministic. It is most often the case that these predicted values can only be partially determined from past observations.

1.4 Box-Jenkins Methodology

1.4a. Autocorrelation and Partial Autocorrelation Functions

The different Box-Jenkins models are identified by the number of autoregressive parameters (*p*), the degree of differencing (*d*), and the number of moving average parameters (*q*). Any such model can be written using the uniform notation ARIMA (*p, d, q).*

We begin the investigation of appropriate model type by looking at the autocorrelations and partial autocorrelations. The sample autocorrelation coefficient (ACF) of lag *k* is computed for the (*n*-*k*) pairs $(y_1, y_{k+1}), (y_2, y_{k+2}), (y_3, y_{k+3}), \ldots, (y_{n-k}, y_n)$

and is given by $r_k = \dfrac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$.

This quantity measures the linear relationship between the time series observations separated by a lag of *k* time units. The autocorrelation coefficient is analyzed to determine the appropriate order *p* of the model [5].

The partial autocorrelation coefficient (PACF) of lag *k*, denoted, ϕ_{kk}, is a measure of the correlation between y_t and y_{t-k} after adjusting for the presence of $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$. This adjustment is done to see if there is any correlation between y_t and y_{t-k} beyond that induced by the correlation y_t has with $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$.

One method of computing the partial autocorrelation at lag k is to perform a regression of y_t on y_{t-1} through y_{t-k}, using the resulting coefficient of the y_{t-k} term as the estimate of ϕ_{kk}. Another approach is to use an algorithm that recursively generates the partial autocorrelation coefficient from knowledge of the autocorrelation coefficients. This recursive process is most often carried out by a computer. When ϕ_{kk} is graphed for lag 1, lag 2 …lag *k* the result is the partial autocorrelation graph (PACF) of the series.

If the ACF trails off and the PACF shows spikes, then an autoregressive (AR) model with order *q* equal to the number of significant PACF spikes is considered the "best" model. If the PACF trails off and the ACF shows spikes, the moving average (MA) model with order *q* equal to the number of significant ACF spikes is the best model. If both the ACF and the PACF trail off then an autoregressive moving average (ARMA) model is used with *p* and *q* equal to one. If the data had to be differenced for it to become stationary, then the ARIMA model is used.

1.4b. ARIMA Models

Autoregressive models are used when the current level of the series is thought to depend on the recent history of the series. An autoregressive model of order *p* (AR(*p*)) or ARIMA (*p*, 0, 0)) is expressed as

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

where ϕ_i is the autoregressive parameter for y_{t-i}. In practice AR models higher than order 2 are rarely observed. The following figures show the typical ACF and PACF for stationary AR(1) and AR(2).

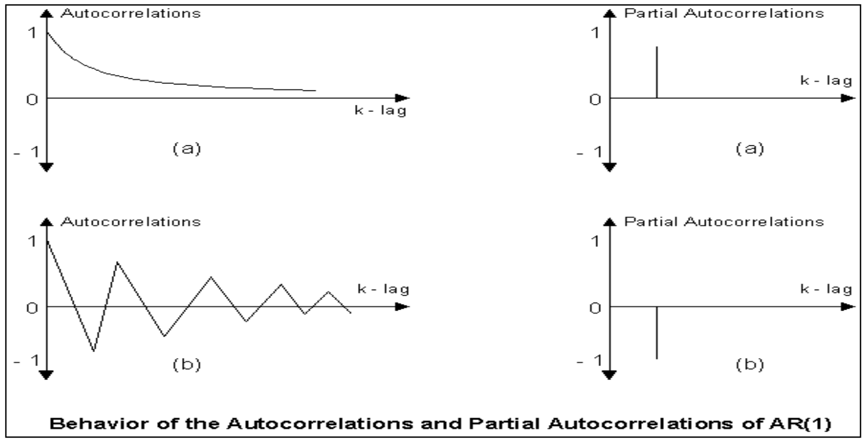

Figure 1.1: Typical autocorrelation and partial autocorrelation functions for stationary AR(1). After lag 1 the ACF dies off to 0 as an exponential function or damped sine wave while the PACF=0

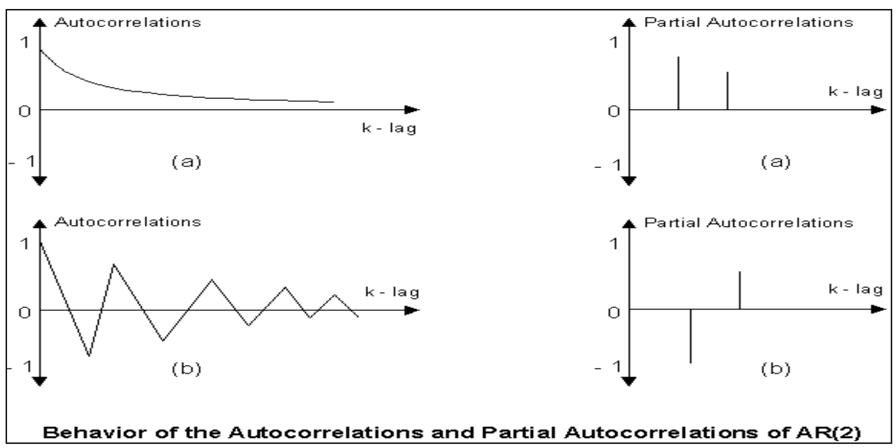

Figure 1.2: Typical autocorrelation and partial autocorrelation functions for stationary AR(2). After lag 2 the ACF dies off to 0 as an exponential function or damped sine wave while the PACF=0

Moving average models are based on the fact that y_t may not be most influenced by past values, but more so by recent shocks or random errors to the series. That is, the current value of a series may be best explained by looking at the most recent *q* error terms. The moving average model of order *q*, (MA(*q*)) or ARIMA (0, 0, *q*), is expressed as,

$$Y_t = \theta_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

where θ_i is the moving average parameter for ε_{t-i}. We assume $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ are uncorrelated with one another. The MA model is appropriate when it is believed that the effects of the random component, ε_t, may be felt for a few periods beyond the current one[2]. It is also worth noting that there is an invertibility imposed on MA models. This restriction forces the sum of the moving average parameters to be less than one. This condition causes decreasing weights to be given to the past values of the series and more importantly guarantees that there is a unique moving average model of order q, MA(*q*), for a given autocorrelation factor [2]. The following figures show the typical ACF and PACF for stationary MA(1) and MA(2).

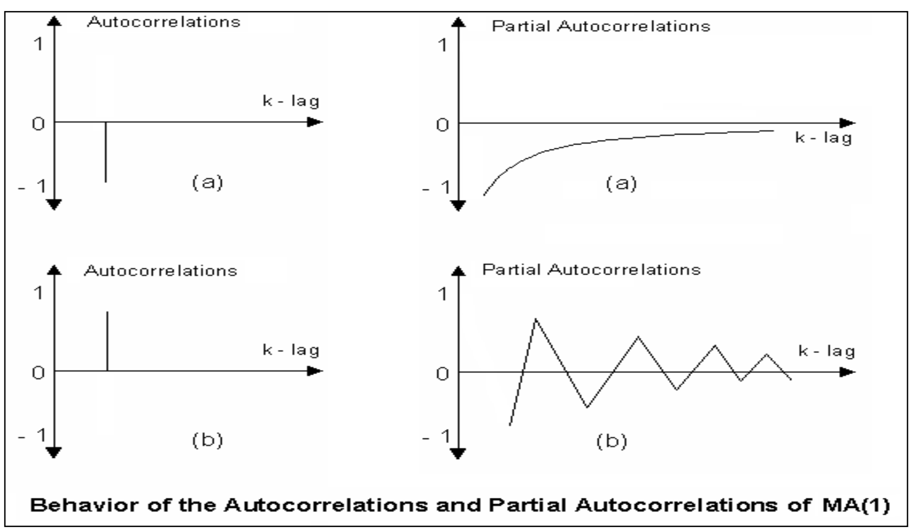

Figure 1.3: Typical autocorrelation and partial autocorrelation functions for stationary MA(1). After lag 1 the ACF =0 while the PACF dies off to 0 as an exponential function or damped sine wave

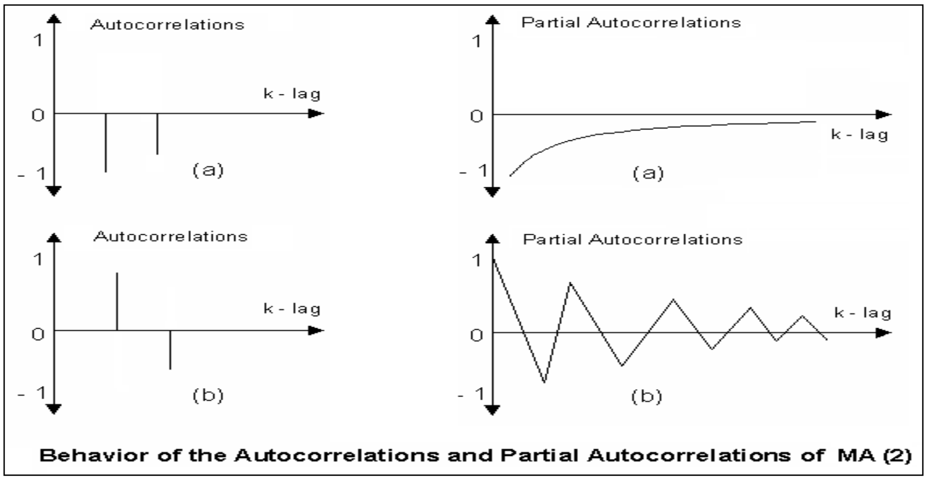

Figure 1.4: Typical autocorrelation and partial autocorrelation functions for stationary MA(2). After lag 2 the ACF =0 while the PACF dies off to 0 as an exponential function or damped sine wave

The autoregressive and moving average model (ARMA) model is a model that contains terms from both the autoregressive and moving average models. However, the ARIMA model is integrated, meaning that because it is a stationary model being fitted to a differenced series it has to be summed (or integrated) to provide a model for the originally nonstationary data. The ARMA model of order *p* and *q*, and the ARIMA model of order *p*, *d*, *q* are essentially the same except that the ARIMA model replaces y_t with the y_t series that has been differenced *d* times. Differencing involves computing the changes (differences) between successive observations of a given time series until it becomes stationary. For non-seasonal data, first order differencing is usually adequate enough to attain apparent stationary, so that the new series based on $\{x_1 \ldots x_N\}$ is expressed by $y_t = x_{t+1} - x_t = \nabla x_{t+1}$. Occasionally second-order differencing, if the trend is quadratic, is required using the operator ∇^2, where

$$\nabla^2 x_{t+2} = \nabla x_{t+2} - \nabla x_{t+1} = x_{t+2} - 2x_{t+1} + x_t$$

So the model for ARMA (*p*,*q*) or ARIMA (*p* ,0, *q*) is

[illegible]

where $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ are uncorrelated with one another.

The model for ARIMA (*p*, *d*, *q*) is

[illegible]

where $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ are uncorrelated with one another and $w_t = \nabla^d y_t$.

Pictured below table is a summary of the properties of the ACF and PACF for AR, MA and ARMA models. We notice that the behavior of the ACF and PACF are the exact opposite for the AR model and MA model.

Table 1.1

Summary of properties of ACF and PACF for Autoregressive, Moving Average and Mixed ARMA models [2]

	Autoregressive Processes	Moving Average Processes	Mixed Processes
Autocorrelation Function	Infinite(damped exponentials and/or damped sine waves) Tails off	Finite Cuts off after lag *q*	Infinite(damped exponentials and/or damped sine waves after first *q*-*p* lags) Tails off
Partial Autocorrelation Function	Finite Cuts off after lag p	Infinite (dominated by damped exponentials and/or sine waves) Tails off	Infinite(damped exponentials and/or damped sine waves after *p-q lags*) Tails off

1.4c. Other Models

A variety of different forecasting procedures are available and it is important to realize that no single method is universally applicable. The Box-Jenkins methodology does not consist of only one model, but rather a family of models. The time series models used in Box-Jenkins forecasting are called autoregressive integrated moving average (ARIMA) models. The class of ARIMA models is very large and some notable special forms of ARIMA models are exponential smoothing, autoregressive models, and random walk models. The Box-Jenkins method can be used to forecast discrete or continuous data. However, the data must be measured at equally spaced, discrete time intervals. Also, ARIMA models can only be applied to stationary series, which is a time series whose mean and variance are essentially constant throughout time, or a series which has been made stationary by differencing.

Exponential smoothing is a forecasting technique that attempts to track changes in a time series by using the newly observed time series values to update the estimates of the parameters describing the time series. In the smoothed form, the new forecast (for time *t*+1) may be thought of as a weighted average of the old forecast (for time *t*) and the new observation (at time *t*), with weight α given to the newly observed value and weight (1-α) given to the old forecast assuming $0 \leq \alpha \leq 1$. Thus $\hat{y}_t = (1-\alpha)\hat{y}_t + \alpha y_t$. The forecast for an exponential smoothing model is produced by an ARIMA (0, 1, 1) model with no constant term.

A random walk is a series whose first differences form a sample from a Normal distribution. That is, $y_t - y_{t-1} = \varepsilon_t$, with $\varepsilon_t \sim N(0,\sigma)$. This is exactly the form of an ARIMA (0,1,0) model with no constant term. If a nonzero constant term, θ_0, is used, the

ARIMA (0,1,0) model is then called a random walk with drift [5]. Random walks are characterized by extremely high autocorrelations. That is, adjacent observations are highly associated with each other. The ACFs will be high for the shortest lags (1 and 2) and decline slowly as the number of lag periods increase.

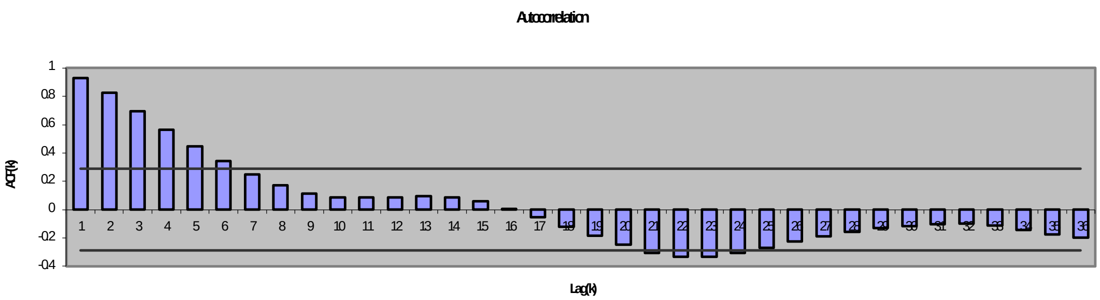

Figure 1.5: Autocorrelation Function for a Random Walk Series

Random walk ACF patterns are similar to trend ACF patterns. A series with random walk floats up and down over time whereas the ACF of a trend series shows a well-defined pattern in the ACF. Autocorrelations at low lags are very high, and decline slowly as the lag increases for a trended series.

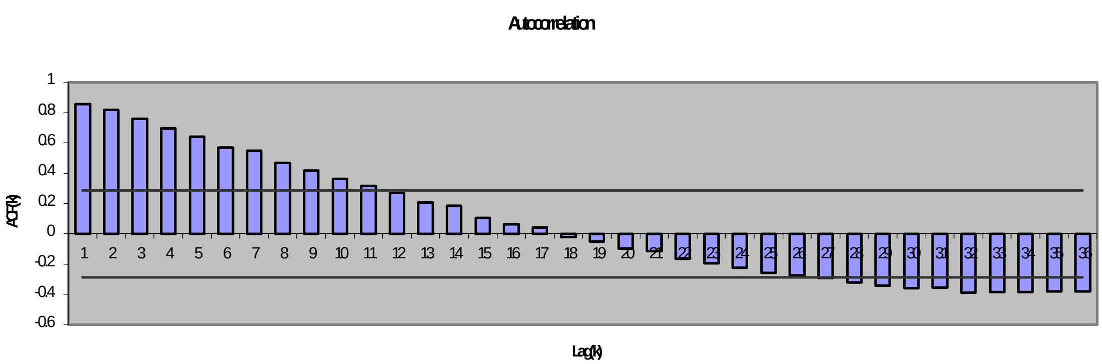

Figure 1.6: Autocorrelation function for a trended series

1.4d.Carrying out the Box-Jenkins procedure for a nonseasonal series

The Box–Jenkins forecasting methodology consists of four basic steps. The first step involves tentatively identifying a model by examining the behavior of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for the values of a stationary time series. The autocorrelation function is usually displayed as a graph of the sample autocorrelation coefficients evaluated for lag 1, lag 2, lag 3… lag k and graphed versus k.

Once the ACF and PACF have been calculated and the behavior of them has been examined to determine the number of autoregressive parameters (p), and/or the moving average parameters (q) and an appropriate model has been selected, then the next step is to use the historical data to estimate the parameters of the tentatively identified model. In theory, the parameters of the selected model can be generated through least squares. However, because it is sometimes the case that nonlinear least squares algorithms, which usually consist of a combination of search routines, that then need to be implemented, computer programs are a necessity to complete this step.

The third step in the Box-Jenkins modeling procedure is to perform a diagnostic check. A diagnostic check is carried out to validate the model, or possibly realize that the tentative model may need to be modified. For a model to be considered "good" it should have the following properties: the residuals should be approximately Normal, all the parameter estimates should have significant p-values, and the model should contain as few parameters as possible. The last step is to use the final model to forecast future time series values.

The following example in time series analysis is based on data that was collected at Broadbalk field at Rothamsted from 1892 to 1925. The data is made up of the count of the average yield of grain that was harvested annually from 1892 -1925. The Box-Jenkins procedure will be used to analyze this series. The following calculations and models are created using Minitab.

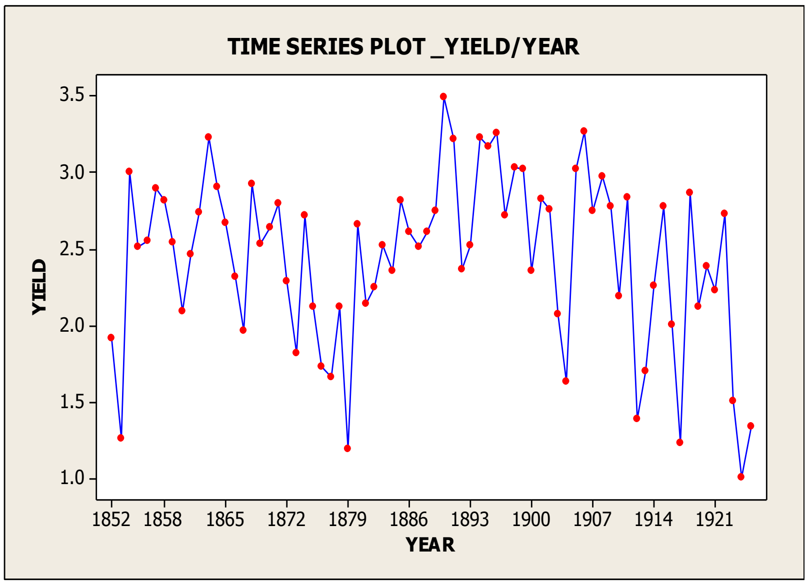

Figure 1.7: Time Series Plot of data

The first step in the process was to create a time series plot of the data, which displayed the daily average for each year of the yield of wheat, versus the years from 1852 to 1925. There are some outliers in 1853, 1879, 1916, and1924, but because the data goes back over 150 years ago and nothing could be found to explain the outliers, we cannot justify or eliminate the outliers.

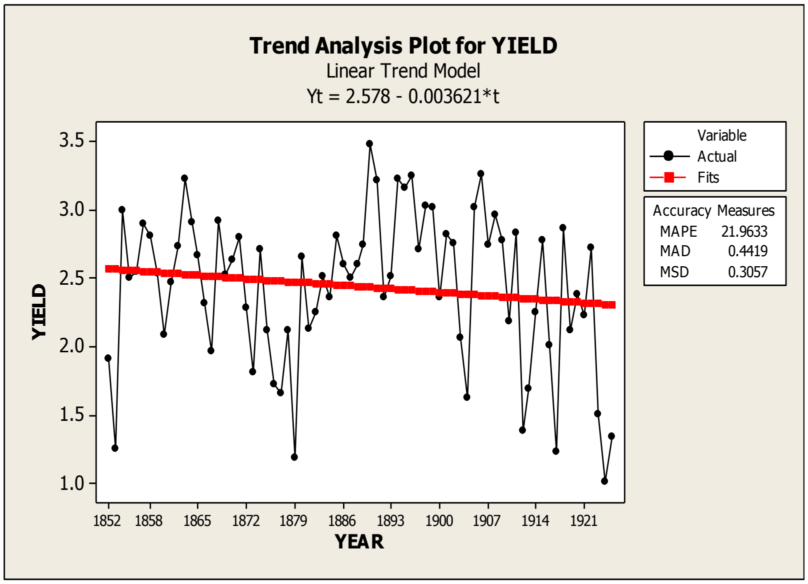

Figure 1.8: Time Series Plot of data with the trend modeled

The trend analysis displays that there is very little trend, so we compare the ACF of the original series to the ACF of a single differenced series to determine if differencing is necessary.

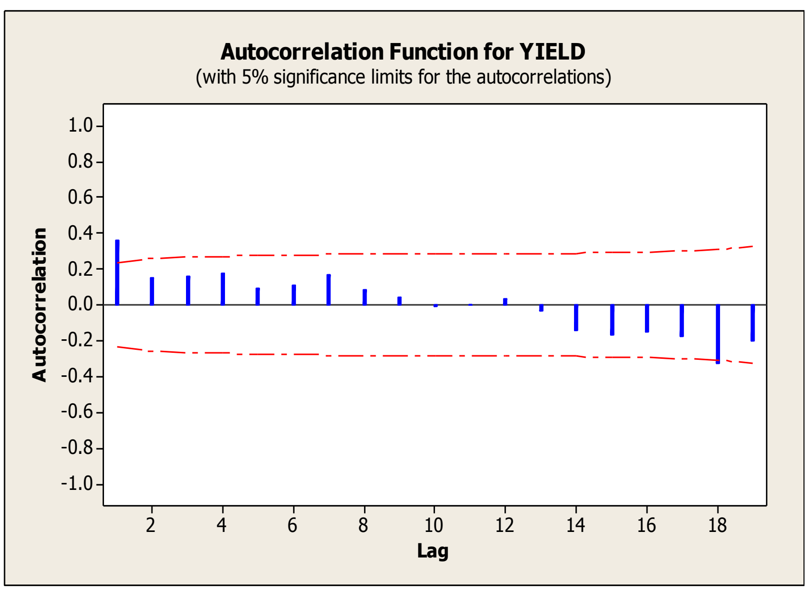

Figure 1.9: Autocorrelation Function of Yield/Year

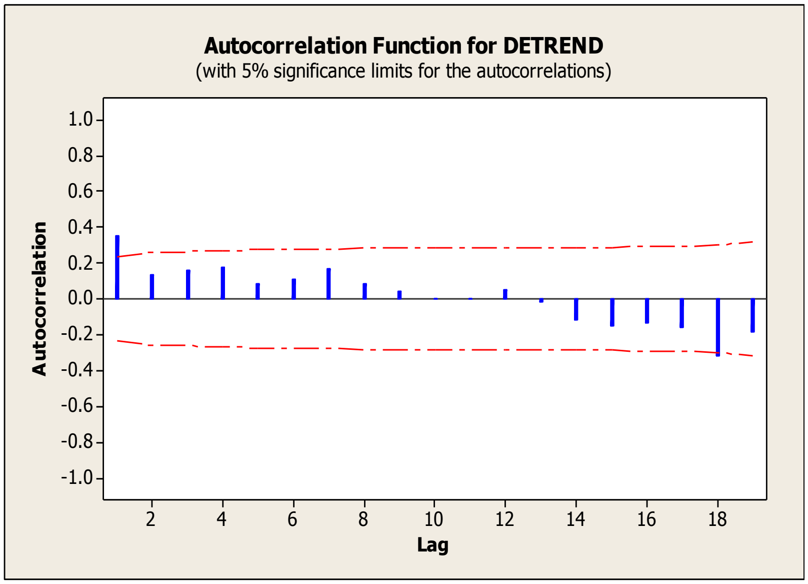

Figure 1.10: Plot of Autocorrelation Function of single differenced data

Examining the two autocorrelation functions (Figure 1.9 and 1.10) reveals that differencing does not cause the ACF to die off any faster. So we can use the series in its original form and assume it is stationary.

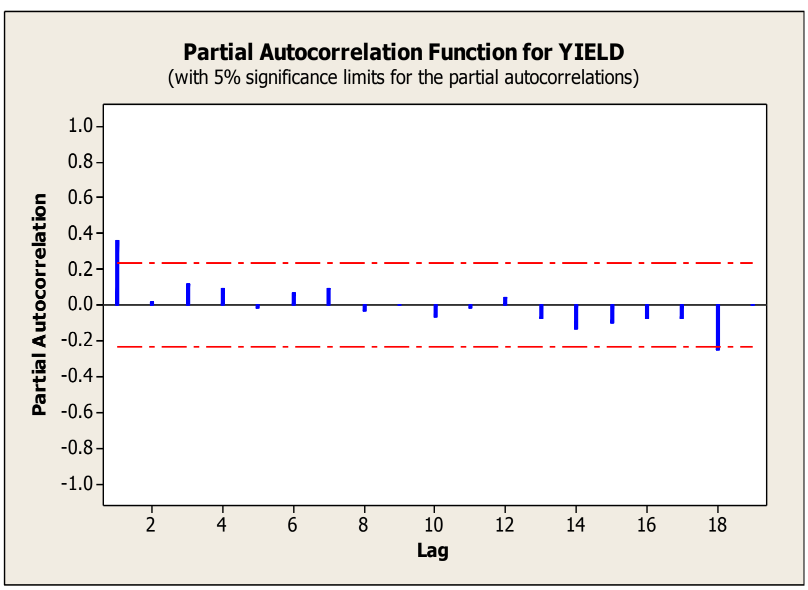

Figure 1.11: Plot of the Autocorrelation Function for the Yield/Year

Considering the behavior of the ACF and PACF, both trail off, so we let the proposed model be AR(1). Minitab produces the following output:

Table 1.2

Model output for Yield/Year data sampled yearly

ARIMA Model: YIELD

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0.3875	0.1121	3.46	0.001
Constant	1.48761	0.06102	24.38	0.000
Mean	2.42878	0.09963		

```
Number of observations:  74
Residuals:    SS =  19.7951 (backforecasts excluded)
              MS =  0.2749  DF = 72
```

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	4.3	19.2	29.3	40.0
DF	10	22	34	46
P-Value	0.935	0.634	0.699	0.719

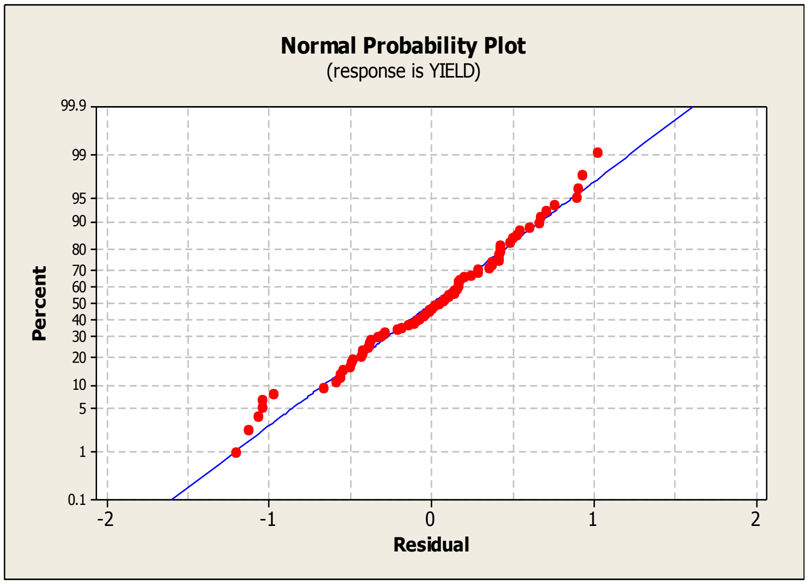

Figure 1.12: Histogram of the residuals

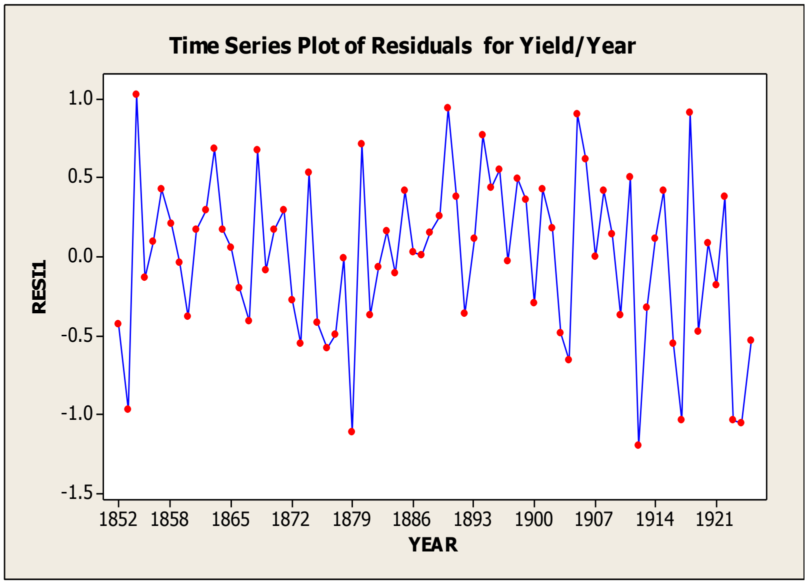

Figure 1.13: Time Series Plot of the residuals

From the output we are able to confirm the fact that all the parameters including the constant are significantly different from zero, because they have p-values that are significantly smaller than .05. Furthermore, the model contains minimal parameters. The model for this data is $y_t = 1.48761 + .3875 y_{t-1} + \varepsilon_t$. This model tells us that the predicted average annual yield was determined by .3875 of the previous year's yield plus 1.48761 and some random error. A probability plot of the residuals reveals that they are essentially Normal and the time series plot of the residuals contains only noise. These diagnostics indicate that a reasonable model has been found

1.4e. Carrying out the Box-Jenkins Procedure for a seasonal series

The previously mentioned process for the Box-Jenkins procedure was outlined assuming that the series was nonseasonal, but the process can also be extended to handle seasonal time series data. With seasonal data, not only does regular differencing have to be applied to make the nonseasonal part of the series stationary, but seasonal differencing has to be applied to the seasonal part to make it stationary as well. Seasonal and regular differencing are similar, except seasonal differences are taken over a span of *L* periods rather than one period.

$$\text{Regular differencing: } \nabla y_t = y_t - y_{t-1}$$

$$\text{Seasonal Differencing: } \nabla y_{Lt} = y_t - y_{t-L}$$

The number of seasonal differences used, *D*, and the number of regular differences, *d*, are needed when specifying the model. The seasonal part of the model also has its own autoregressive and moving average parameters with order *P* and *Q*, while the nonseasonal part are order *p* and *q* . Note seasonal parameters are the uppercase version of the nonseasonal parameters. To determine *P* and *Q*, the ACF and PACF are examined but only at the seasonal lags.

With seasonal models that also have nonseasonal terms a choice exists for how the seasonal and nonseasonal terms are to be combined into a single model. The terms can just be added together using an additive model, or a multiplicative model formed by applying the nonseasonal model to the terms from a purely seasonal model could be used. An additive model with one seasonal MA term and one nonseasonal MA term is written

$$y_t = \Theta_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \Theta_1 \varepsilon_{t-L}.$$

The multiplicative model is $y_t = \Theta_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} + \Theta_1 \varepsilon_{t-L} + \theta_1 \Theta_1 \varepsilon_{t-L-1}$.

After identifying the nonseasonal order (*p, d, q*) and the seasonal orders (*P,D, Q*) for a given series, a multiplicative model is denoted as an ARIMA $(p, d, q)(P,D,Q)^L$ model [5].

Once the appropriate model is identified, the diagnostic check is carried out the same way as it is for a nonseasonal series.

The following example in time series analysis is based on the data that comes from the classic Golden Gate Bridge example. The data is made up of the count of the average number of vehicles that crossed the Golden Gate Bridge daily, for each month, from the years 1967-1980 (Appendix 1). For this time series the Box-Jenkins procedure will be used to analyze the series and select an appropriate model that can be used for forecasting. The following calculations and models are created using Minitab.

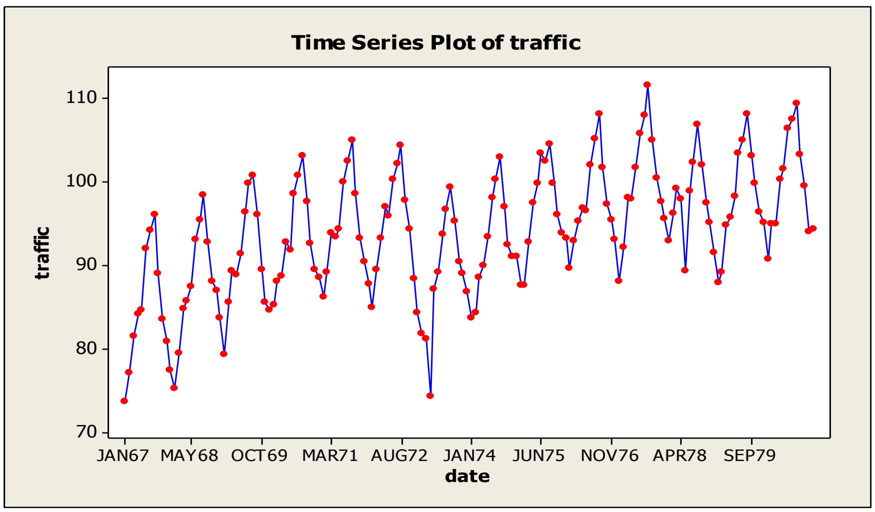

Figure 1.14: Time Series Plot of data

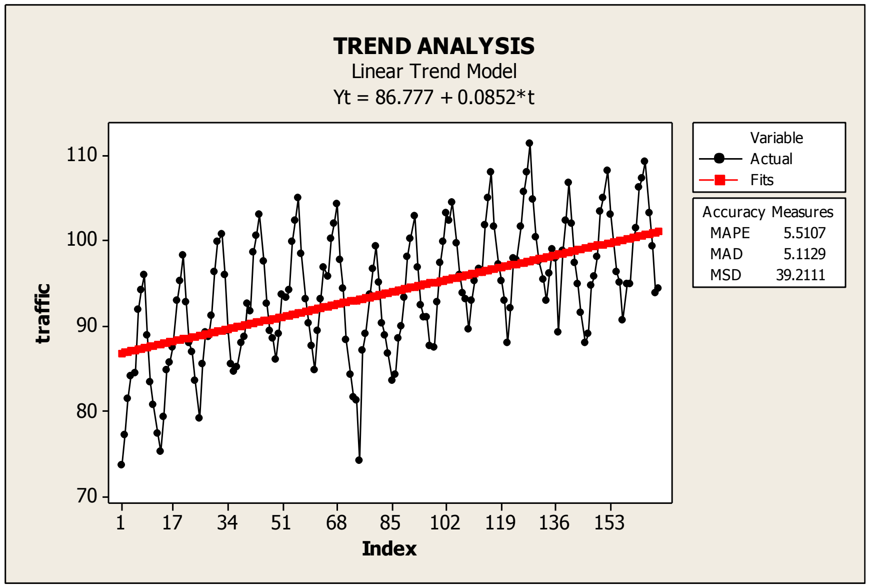

Figure 1.15: Time Series Plot of data with trend

The first step in the process was to create a time series plot of the data, which displayed the daily average for each month of the number of vehicles that crossed the bridge, versus the months from 1967 to 1979. From the plot, it is apparent there is an increasing linear trend and seasonal fluctuations because the increasing mean and pattern of peaks and valleys. In particular the seasonal variation is on a monthly basis, that is L=12. We also notice that there is an outlier at t=75. Because t is associated with a month and year, by doing additional research, it is discovered that there was an oil crisis in 1973, and this outlier was most likely a direct result of the oil shortage.

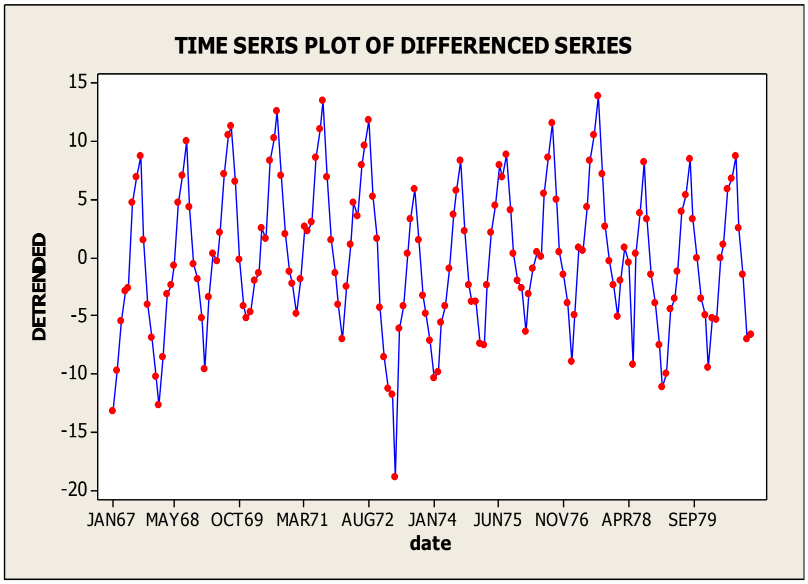

Figure 1.16: Time Series Plot of single differenced data

Because the original series is not stationary and the trend is linear, the first order differences are taken of the series. Using the differenced data produced “Time Series Plot of single differenced data” (Figure 1.16). The stationary series is used to calculate the ACF and PACF.

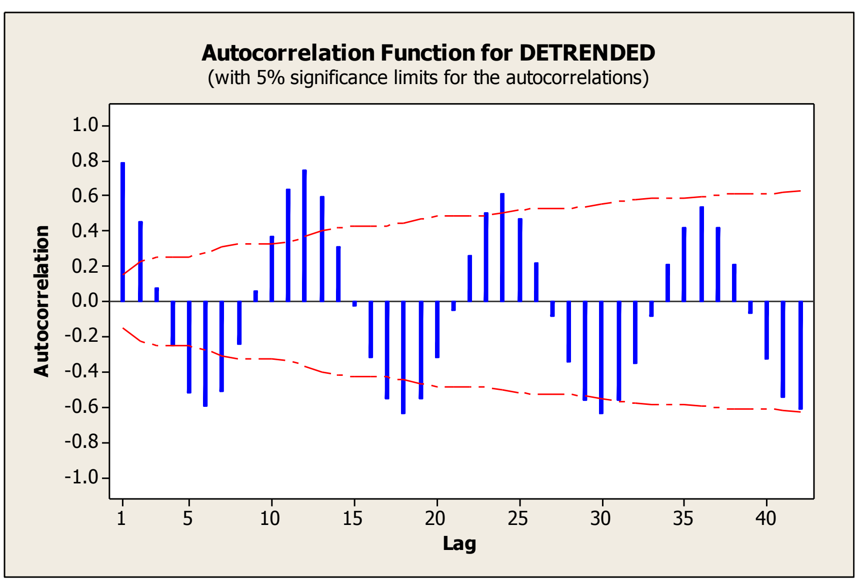

Figure 1.17: Plot of the Autocorrelation Function for the differenced series

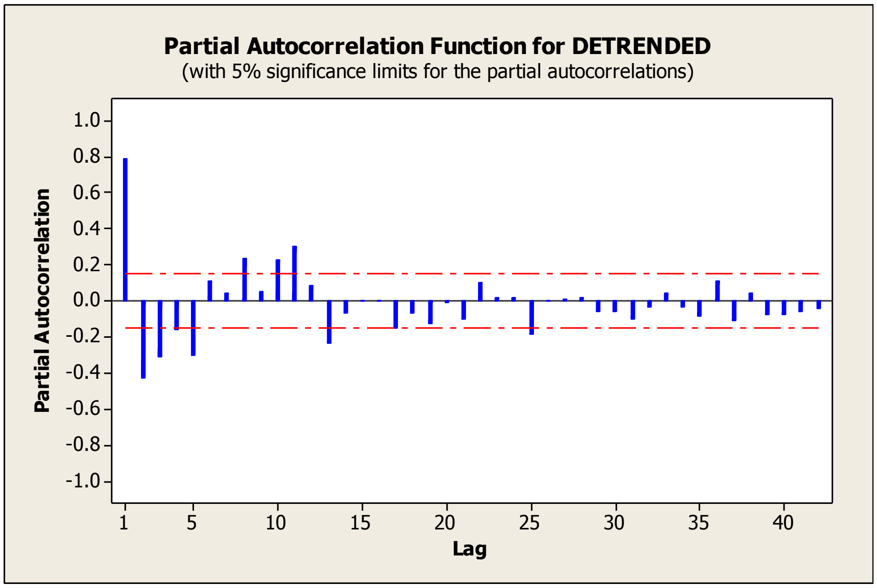

Figure 1.18: Plot of the Partial Autocorrelation Function for the differenced series

We note that there are significant positive correlations in the ACF at lags 8, 10, 11, 13, 25 and significant negative correlations at lags 5, 6, 7, 17, 18, 19, 30. Considering the behavior of the nonseasonal lags for the ACF and PACF, both trail off, so we let q=1, p=1, and d=1, because the series was already differenced once. Considering the behavior of the seasonal lags for the ACF and PACF , the ACF spikes once as the PACF dies off, so we let Q=1, P=0, and D=1. So the proposed model is ARIMA (1, 1, 1)(0,1,1) and for this model Minitab produces the following output:

Table 1.3

Minitab output for Golden Gate Bridge data

ARIMA Model: traffic

ARIMA model for traffic

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0.5128	0.1370	3.74	0.000
MA 1	0.8054	0.0944	8.53	0.000
SMA 12	0.9194	0.0562	16.36	0.000
Constant	-0.007533	0.005080	-1.48	0.140

Differencing: 1 regular, 1 seasonal of order 12

Number of observations: Original series 168, after differencing 155
Residuals: SS = 621.257 (backforecasts excluded)
MS = 4.114 DF = 151

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	3.4	8.1	12.5	30.0
DF	8	20	32	44
P-Value	0.909	0.991	0.999	0.946

Forecasts from period 156
95 Percent Limits

Period	Forecast	Lower	Upper	Actual
157	91.208	87.231	95.184	90.707
158	93.308	88.438	98.179	94.949
159	96.629	91.278	101.980	94.970
160	98.519	92.837	104.200	100.286
161	99.154	93.208	105.100	101.497
162	104.566	98.389	110.742	106.352
163	106.599	100.210	112.987	107.415
164	109.289	102.701	115.877	109.385
165	103.562	96.783	110.342	103.266
166	98.992	92.028	105.957	99.432
167	95.978	88.834	103.122	93.965
168	93.625	86.307	100.944	94.385

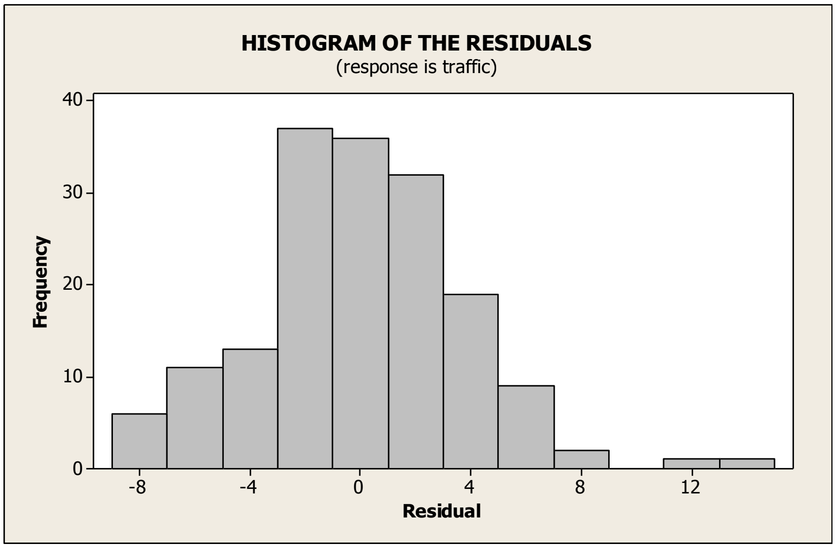

Figure 1.19: Histogram of the residuals

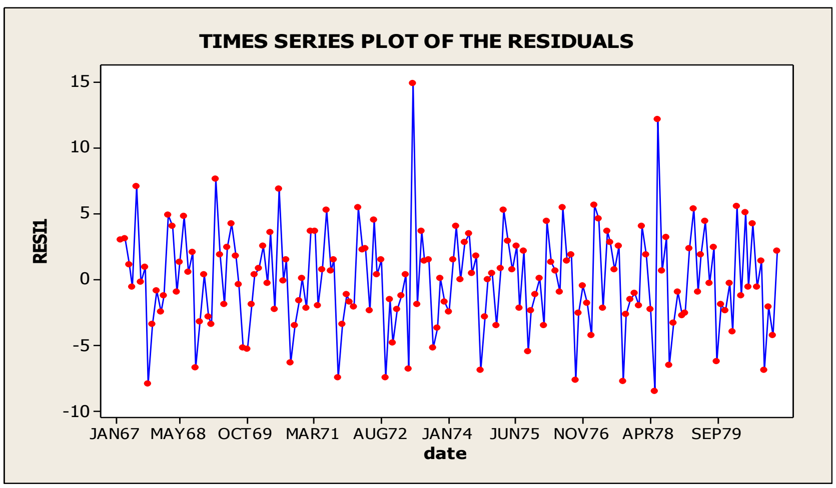

Figure 1.20: Time Series Plot of the residuals

From the output we are able to confirm the fact that all the parameters except the constant are significantly different from zero, because they have p-values that are equal zero. Furthermore, the model contains minimal parameters. The model for this data is $y_t = .5128y_{t-1} - .8054\varepsilon_{t-1} + .9194y_{t-12} + \varepsilon_t$. This means that the predicted number of cars to cross the bridge is made up of .5128 of the number of cars that crossed the previous month, minus .8054 of some random shock, which is then added to .9194 of number of cars that crossed a year ago and some random error.

A histogram of the residuals reveals that they are essentially Normal and a time series of the residuals contains only noise. We can conclude that we have produced an adequate model for forecasting future values of this series; however, we are uncertain how well this model will perform past the period of a year because this is the interval for which it was tested. It is suspected that the variability would increase as the forecast is extended to longer period of time.

1.5 Regression Methods

Regression analysis is a statistical method for estimating the functional relationship between a response variable and one or more independent predictor variables and can be applied to a set of *y* and *x* variables, whether or not they represent time series data.

However, there are regression models that have been adapted specifically for time series, where variables are not always independent. One of these involves using indicator variables to produce a model of seasonal data.

Time series regression models are useful when the parameters describing the trend, seasonal, or cyclic components of a time series are deterministic. It may also be useful to use a regression model when a linear model is not appropriate, a situation that can often occur in practice.

It can sometime be useful to describe a time series y_t by model $y_t = TR_t + SN_t + \varepsilon_t$ where TR_t denotes the trend in the time series *t*, SN_t, denotes the seasonal factor in time period *t*, and ε_t denotes the error term in the time period *t*. The above model says that the time series observation y_t can be represented by an average level that changes over time combined with random fluctuations.

When there is

$$\text{no trend: } TR_t = \beta_0$$

$$\text{linear trend: } TR_t = \beta_0 + \beta_1 t$$

$$\text{quadratic trend: } TR_t = \beta_0 + \beta_1 t + \beta_2 t^2 ,$$

where the $\beta's$ are the trend parameters. Supposing that there are L seasons per year, we will assume that SN_t is given by the equation: $SN_t = M_1 x_1 + M_2 x_2 + \ldots + M_{L-1} x_{L-1}$ where $M_1, M_2 \ldots M_{L-1}$ are used to denote the parameters for the monthly or quarterly seasonal component [6].

Seasonal patterns can also be modeled using indicator variables. The use of indicator variables to model time series is a regression procedure that has been specifically adapted to time series. A dummy or indicator variable can be used to incorporate qualitative information such as season of the year, geographical region, or occurrence of a specific event into a model. The indicator variables convert qualitative information into quantitative information by making use of a coding scheme. The most common coding scheme uses a 1 to indicate the occurrence of an event of interest and a 0 to indicate its nonoccurrence [6]. Using indicator variables is also a way of producing a model that reflects a change in slope; however, not all models using indicator variables will reflect a change in slope. In the latter situation the reason a simple trend line is not used is because the indicator variables approach gives a better estimate of the variation σ_ε in the series by using all the available data.

Although not mentioned when discussing the Box-Jenkins procedure, the autoregressive model is also a regression model. Autoregressive models are a method of fitting that is based on using lagged values of y_t as predictor variables in regression models. The parameters for regression models are also calculated using the least squares method and some type of computer package is most often used to calculate these parameters.

1.6 Conclusion

To a naïve mind, forecasting is simply putting data into the computer and letting the program do the work. This is not the case. Before being able to conceive the formation of a model, an in-depth analysis of the data has to be made. The computer is simply a tool to help the forecaster carry out the analysis in a shorter period of time. The analysis of time series has been carried out for decades and used in several different fields. The methods used to analyze a time series truly depend on the objective that needs to be met and the characteristics of the data. Time series can be analyzed solely for the purpose of determining the behavior of the series, i.e. recognizing trend or some seasonal or cyclic movement, and used strictly in that form to monitor behavior, or as a tool to see how occurrences in nature, the economy, or in the world could have possibly affected the data. They can be also analyzed with the purpose in mind of forecasting. In order to produce a model for a series, the forecaster first must recognize that there are many approaches to modeling and that no single method exists for every situation. The ARIMA models alone provide several options. The forecaster has to be familiar with the models that exist for each process, but also be aware of under what circumstances that procedure can be used. The Box-Jenkins approach is very useful in that it not only provides the forecaster with a class of models to choose from, but it also consists of a methodology that outlines how to forecast a series. It is again up to the forecaster to make the sure that the underlying assumed conditions are met to carry out the process and, more importantly, that the analysis work is done properly so that the most appropriate model is selected

Given the information provided in this chapter, the next step is to apply this content to stock prices. Our intent is not only to be able to create a model and predict values, but also through further research, do so in a more effective way. The basic question is to attempt to determine if there is some optimal data time frame that can be used to create an accurate forecast. For example, instead of having to collect data on a daily (or hourly) basis, would weekly or monthly observations provide models equally good (or better) for forecasting? Also, are stocks of the same industry equally predictable?

CHAPTER 2

Many empirical series, such as stock market prices, behave as though they have no fixed mean. However, they usually exhibit homogeneity in that one part of the series behaves much like any other part. Models that describe such homogenous nonstationary behavior can be obtained by assuming some difference of the process to be stationary. These are a class of models for which the *d*th difference is a stationary mixed autoregressive-moving average process. These models are called autoregressive integrated moving average (ARIMA) processes, which were introduced in Chapter 1.

This chapter looks at an initial analysis stock prices for Apple, Inc. (APPL), Microsoft Corp.(MSFT), Kroger Company (KR), Winn-Dixie Stores, Inc. (WINN), ASML Holding (ASML), Advanced Analogue Technologies, Inc. (AATI), PepsiCo, Inc. (PEP), and Coca-Cola Bottling Co. Consolidated(COKE). AATI is a company that develops and engineers of advanced power management semiconductors. AATI also offers a broad range of analog and mixed-signal circuits that play a critical role in system design and their product play a key role in the continuing evolution of feature rich convergent devices ASML is the world's leading provider of lithography systems for the semiconductor industry. They manufacture complex machines that are critical to the production of integrated circuits or chips. These eight stocks were chosen because they represent the leading companies in their perspective industries.

The closing prices for each stock going back ten years or from the date the company went public, whichever was reached first, were obtained from Yahoo! Finance. The companies of interest are two entities from the following industries: computers and software, grocery stores, semiconductor production, and soft drinks. The industries were chosen because they all offered different products to people all over the world. The data for each stock was examined on a weekly basis, and on the 1st and 15th trading day of the month: the closing price for each Tuesday (because Monday is often a holiday), the first trading day of the month, and the midway trading day of the month. Because there are 252 trading days on Wall Street in a year, when the word year is used in this context, it refers to this time interval.

It is important to note that the time frames that are being modeled can only be used to forecast values in the same range. That is the monthly data produces a forecast of monthly values. Extrapolation occurs when the values that are trying to be predicted are well beyond the range of the data provided.

Table 2.1

Data Dates

APPL	MSFT	KR	WINN	ASML	AATI	PEP	COKE
5/30/2000- 5/25/2010	5/30/2000- 5/25/2010	5/30/2000- 5/25/2010	11/28/2006 – 5/25/2010	10/2/2007- 5/25/2010	8/9/2005- 5/25/2010	5/30/2000- 5/25/2010	5/30/2000- 5/25/2010

We begin analyzing the stock prices by examining a time series for each interval. The following time series plots are of stock prices sampled weekly. The time series plots should reveal any apparent trend or seasonality,\ and indicate any drastic rise or fall in the price of the stock.

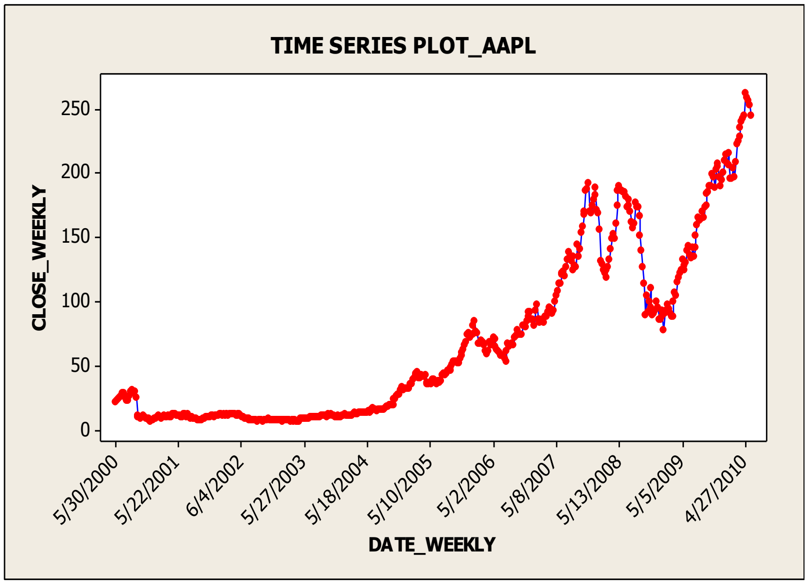

Figure 2.1: Time Series Plot of APPL sampled weekly

The time series indicated that there are possibly some events that influenced the price per share for Apple because of the inconsistency of the behavior of the stick. Through examining the release dates and announcements of products by APPL it was realized that an announcement or actual new release didn't necessarily have a significant change (positive or negative) in the price per share. Also the time series plot did not show any extreme outliers or drastic fluctuation in the price per share, when looking at the data sampled weekly.

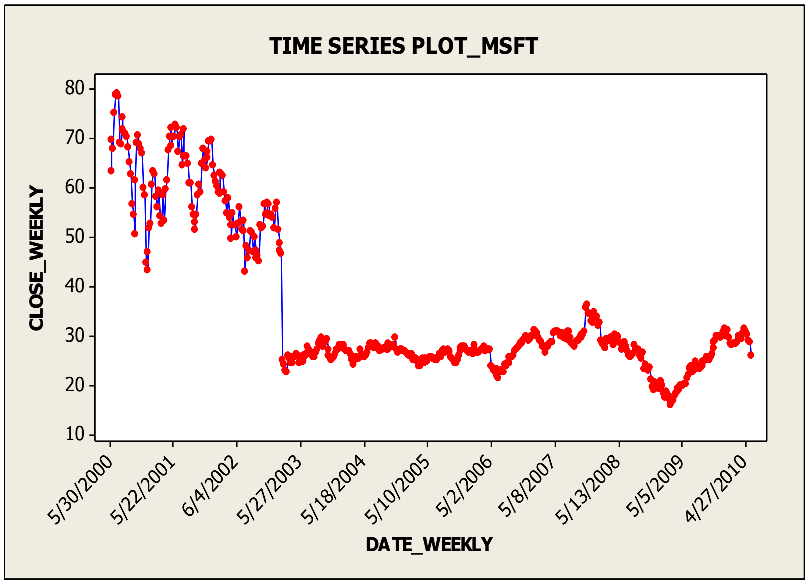

Figure 2.2: Time Series Plot of MSFT sampled weekly

Microsoft's drastic change in stock price that occurred on February 18, 2003 was due to a stock split that was authorized by the Board of Directors of Microsoft. The stock split was a two-for-one and was announced on January 16, 2003. A two-for-one split results in the price per share being decreased by half and stockholders owning twice as many shares. The price per share went from 46.44 on February 11, 2003 to 24.96 on the 18^{th}. Because we know exactly how this impacted the stock and at which point, we can justify eliminating the prices prior to February 18^{th}, 2003 and only modeling the resulting series. This manipulation produces the following time series plot.

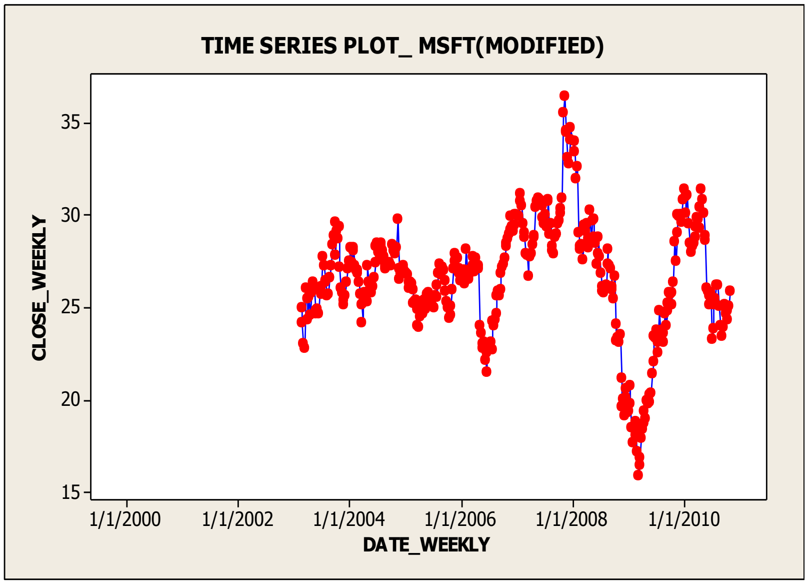

Figure 2.3: Time Series Plot of MSFT (modified) sampled weekly

This modified time series reflects the behavior of the stock without any known imposed manipulations. This series consists mainly of random peak and valleys without any significant change in the price per share.

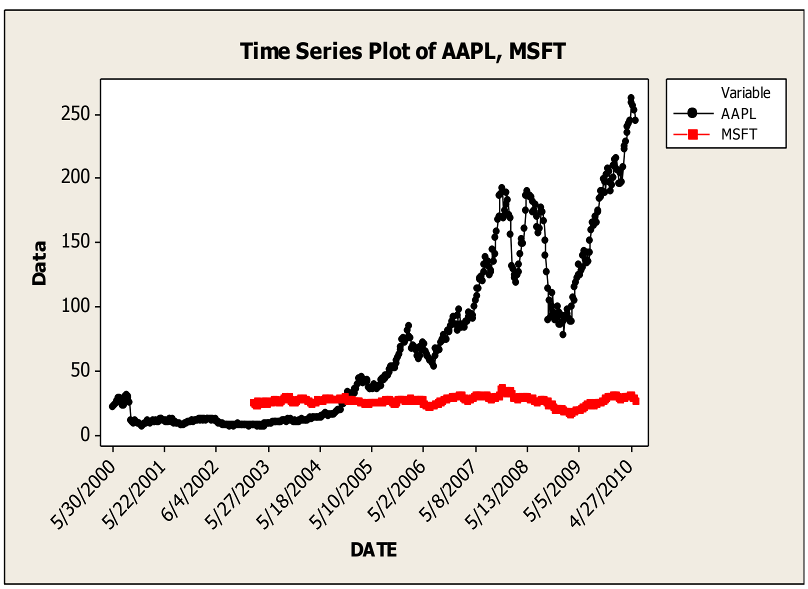

Figure 2.4: Time Series Plot of APPL and MSFT sampled weekly

Comparing APPL and MSFT on the same interval, we notice that while the price per share for both has continued to increase. APPL has done so at a much faster rate. In particular from February 18, 2003 to May 25, 2010 MSFT had an overall percent of increase of 44.5% versus APPL's being 3111.8%.

Table 2.2:

Comparison of the percent of change for the two (APPL and MSFT) stocks based on weekly data, in one-year intervals:

	APPL	MSFT
2/18/03-2/17/04	51.7%	08.1%
2/24/04-2/22/05	281.4%	06.1%
3/1/05-2/28/06	53.9%	06.3%
3/7/06-3/6/07	33.0%	02.9%
3/13/07-3/11/08	44.1%	09.6%
3/18/08-3/17/09	(25.0%)	(42.6%)
3/24/09-3/23/10	114.0%	66.8%
3/30/10-5/25/10	04.0%	12.4%

Table 3.2 shows that although at different rates, APPL and MSFT. The price per share for both increased and decreased during the same time frames.

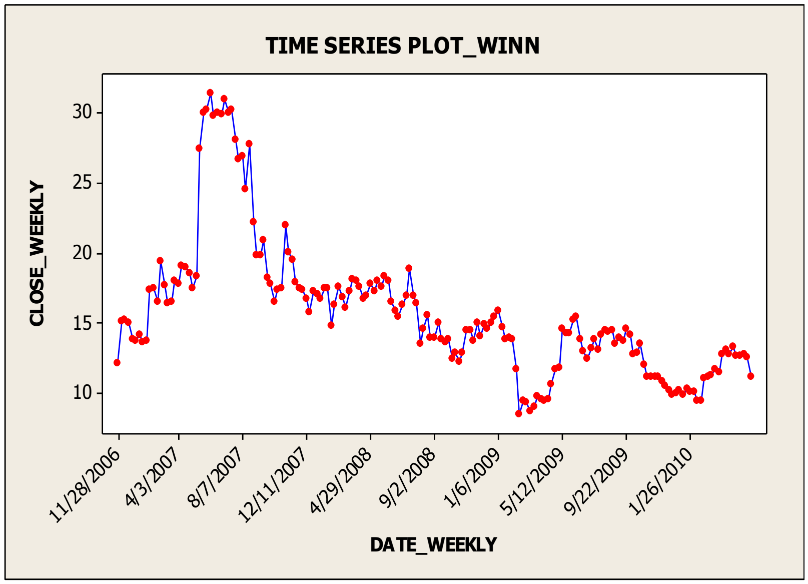

Figure 2.5: Time Series Plot of WINN sampled weekly

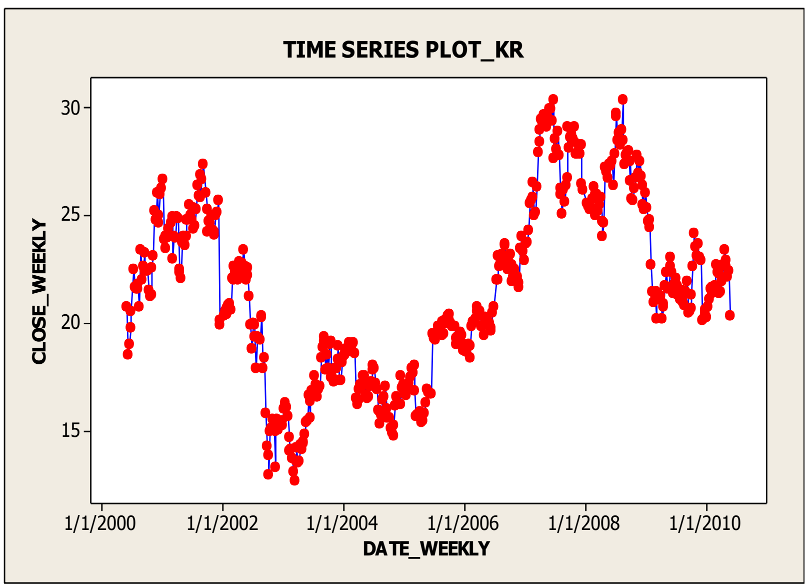

Figure 2.6: Time Series Plot of KR sampled weekly

The time series plots for WINN and KR have different starting points because Winn-Dixie went public after Kroger did.

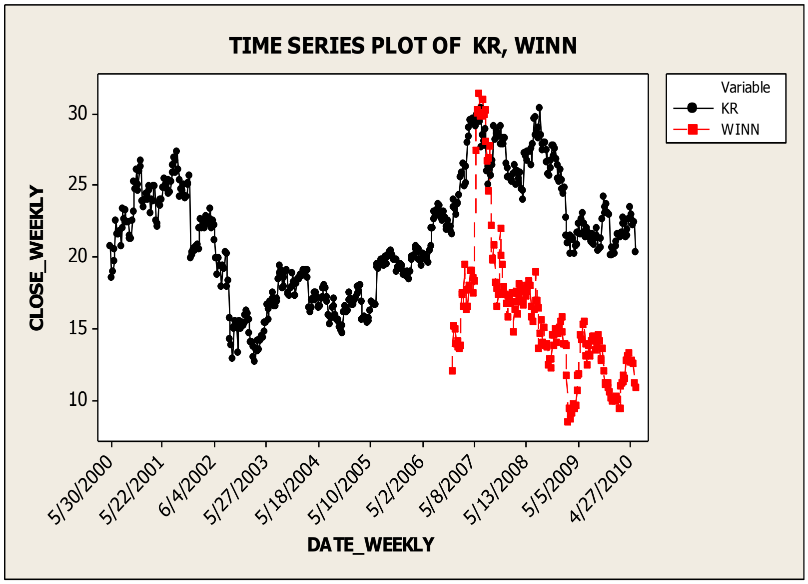

Figure 2.7: Time Series Plot of KR and WINN sampled weekly

Table 2.3

Comparison of the percent of change for the two stocks (KR and WINN) based on weekly data, in one-year intervals

	KR	WINN
11/18/06-11/27/07	28.7%	44.3%
12/4/07-12/2/08	(5.6%)	(18.8%)
12/9/08-12/8/09	(53.9%)	(31.8%)
12/15/09-5/25/10	.67%	9.7%

Comparing the percent change for KR and WINN on the same interval, we notice that the stocks behave similarly, although this is not so apparent from the individual time series plots. What is shown by time series plot is that since November 2006 the price per share for WINN has fluctuated within the 10 to 30 dollar price range, while during the same interval, KR has remained within a 20 to 30 dollar range. WINN is the more volatile of the two stocks.

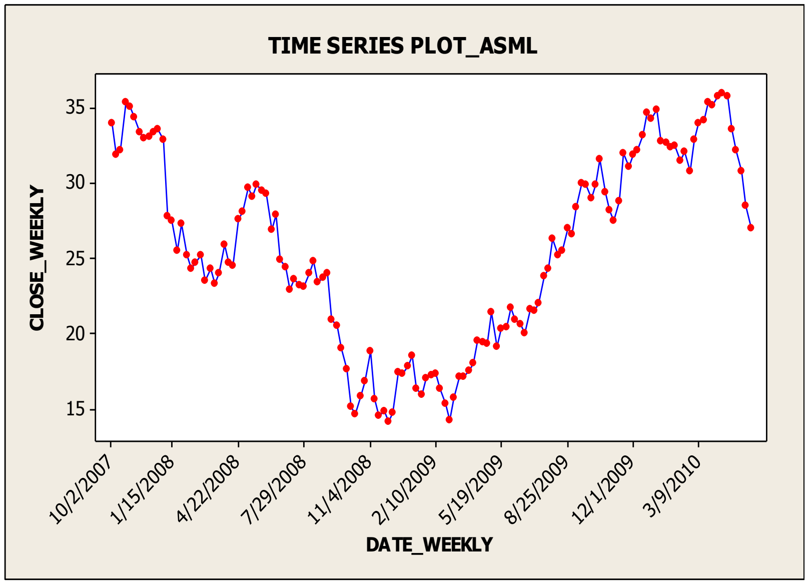

Figure 2.8: Time Series Plot of ASML sampled weekly

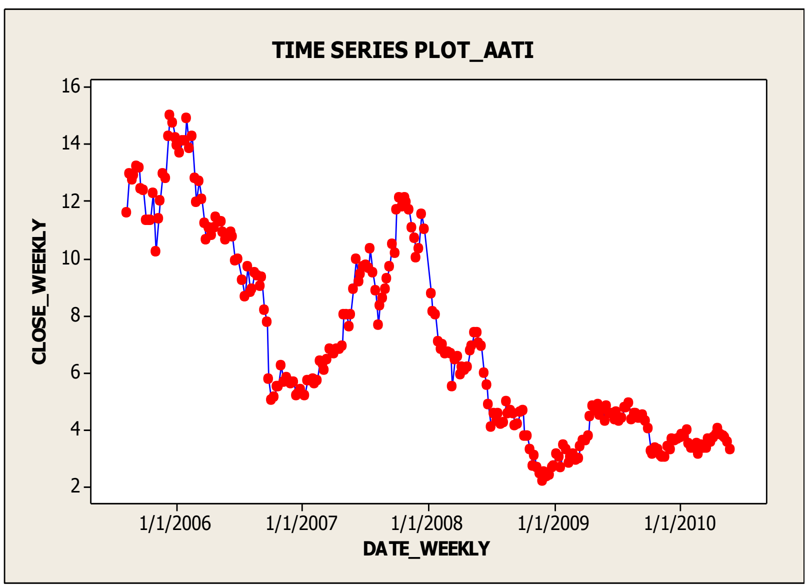

Figure 2.9: Time Series Plot of AATI sampled weekly

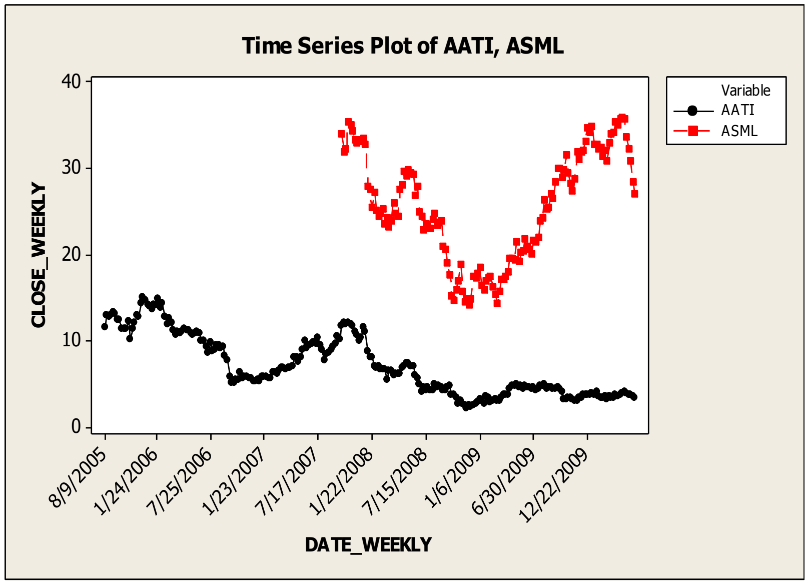

Figure 2.10: Time Series Plot of AATI and ASML sampled weekly

Table 2.4

Comparison of the percent of change for the two stocks (AATI and ASML) based on weekly data, in one-year intervals:

	AATI	ASML
10/2/07-9/30/08	(60.1%)	(48.1%)
10/7/08-10/06/09	(14.1%)	96.7%
10/13/09-5/25/2010	5.4	(14.5%)

Comparing the percent change for AATI and ASML on the same interval, we notice that the stocks only behaved the same way from 10/2/2007 – 9/30/2008. From 10/2/2007 – 5/25/2010, although price per share for AATI has remained between 11.68 and 3.31, this is nearly a 72% decrease in price per share for this time interval. This decline is most obvious when examining the individual time series plot for AATI (Figure 2.9). However, during the same interval ASML, although still on the decline, has only decreased by 20%

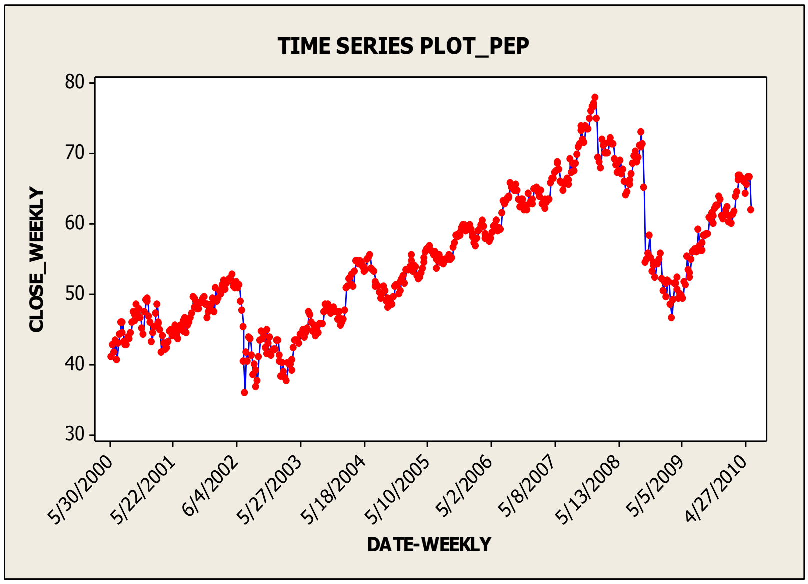

Figure 2.11: Time Series Plot of PEP sampled weekly

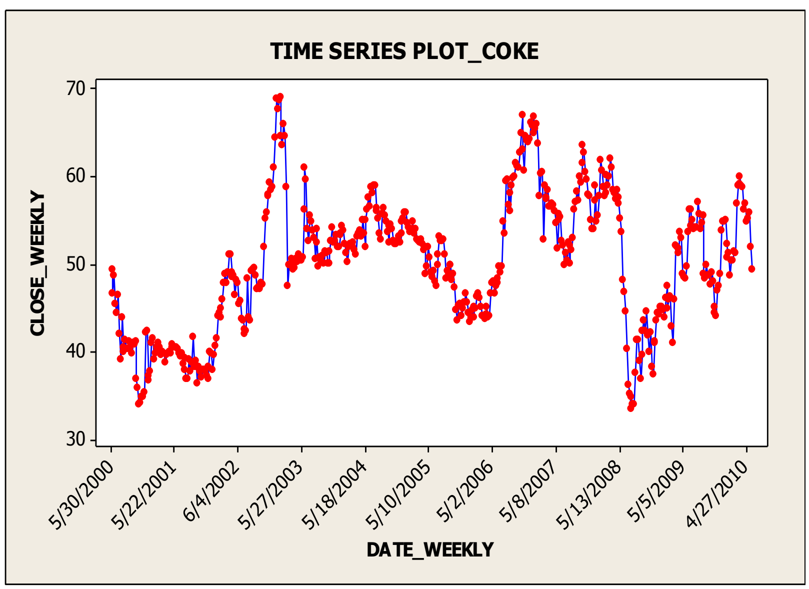

Figure 2.12: Time Series Plot of COKE sampled weekly

Table 2.5

Comparison of the percent of change for the two (PEP and COKE) stocks based on weekly data, in one-year intervals:

	PEP	COKE
5/30/00-5/29/01	9.6%	(18.0%)
6/5/01-6/4/02	13.6%	18.1%
6/11/02-6/10/03	(12.8%)	23.7%
6/17/03-6/15/04	23.0%	(7.3%)
6/22/04-6/21/05	.7%	(19.2%)
6/28/05-6/27/06	7.8%	0.0%
7/11/06-7/10/07	6.9%	(7.9%)
7/17/07-7/15/08	(1.3%)	(36.2%)
7/22/08-7/21/09	(15.2%)	55.2%
7/28/09-5/25/10	12.4%	(12.9%)

Comparing the percent change for PEP and COKE we notice that, for seven of the ten years, as COKE saw a decrease in price per share PEP experienced an increase during those same intervals. Looking at both time series plots individually shows that both stocks have consistently fluctuated up and down, with no seasonal variation.

Now we would like to determine if there is some trend and/or seasonal component or each of the stocks. To determine if there is any underlying trend or seasonality, an autocorrelation function (ACF) of the undifferenced series is analyzed. The autocorrelations for stationary series are large for low order autocorrelations but die out rapidly as lag length increases. If the series is trended, autocorrelations are large and positive for short lags, decreasing slowly as the lag increases.

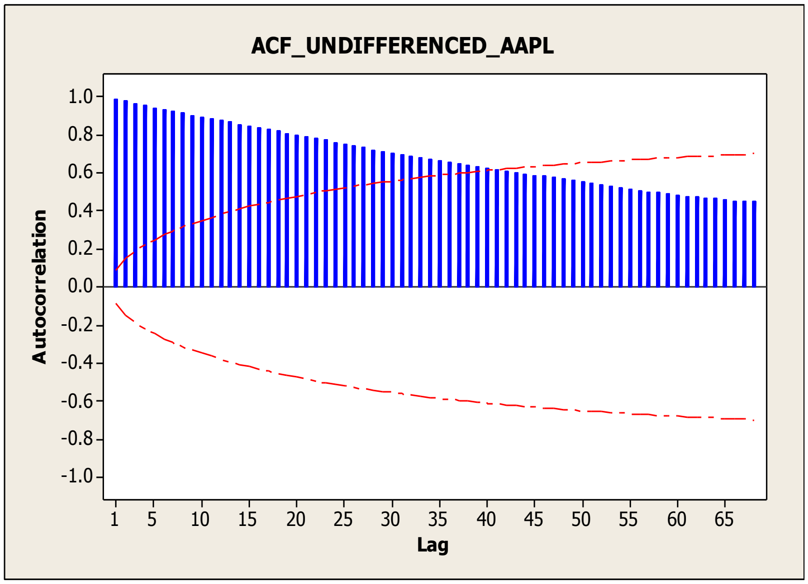

Figure 2.14: Autocorrelation Function for Undifferenced APPL

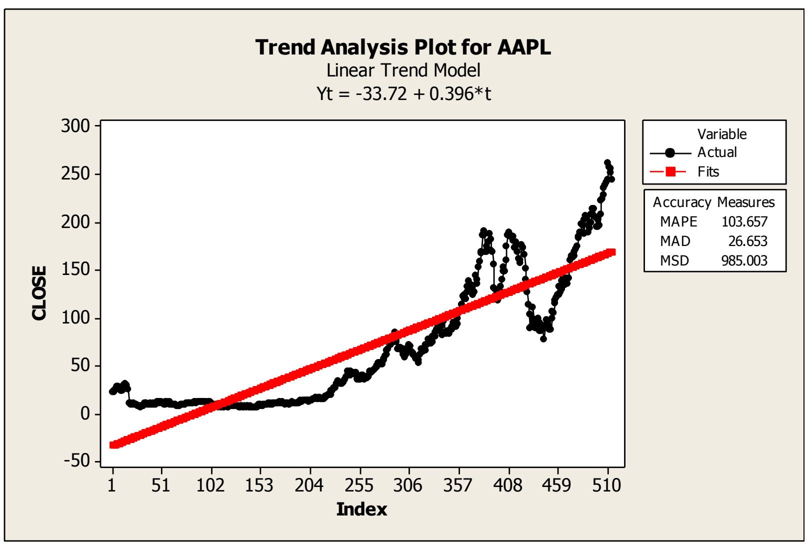

Figure 2.15: Trend Analysis for APPL

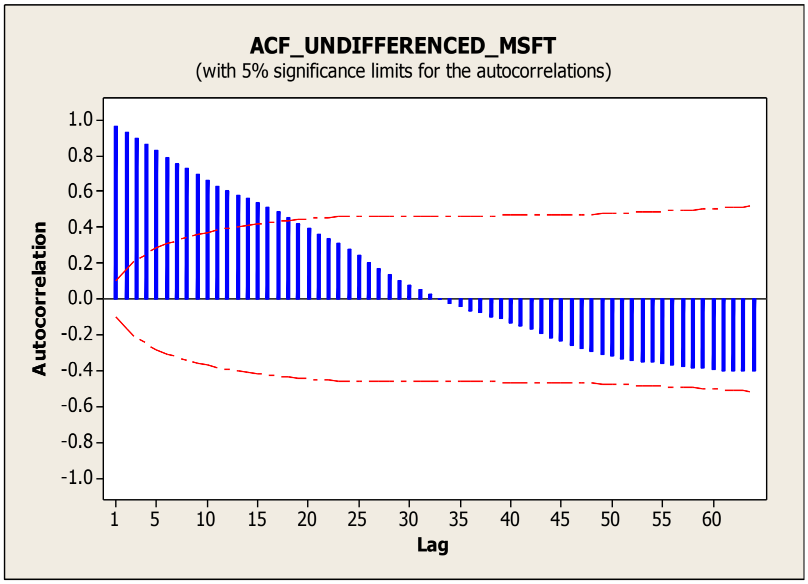

Figure 2.16: Autocorrelation Function for Undifferenced MSFT

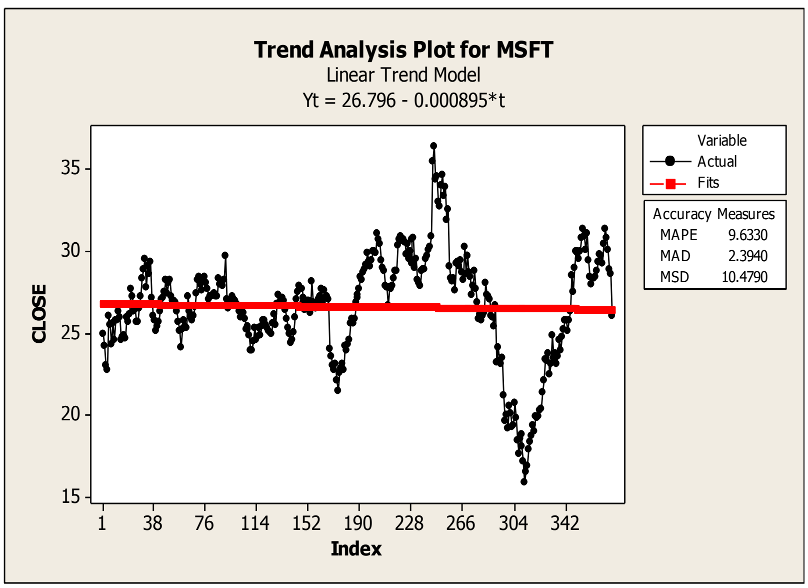

Figure 2.17: Trend Analysis for MSFT

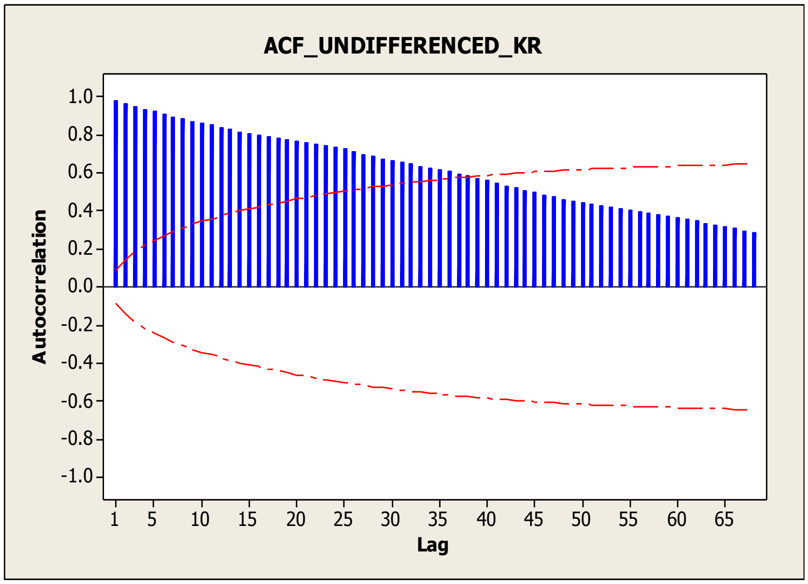

Figure 2.18: Autocorrelation Function for Undifferenced KR

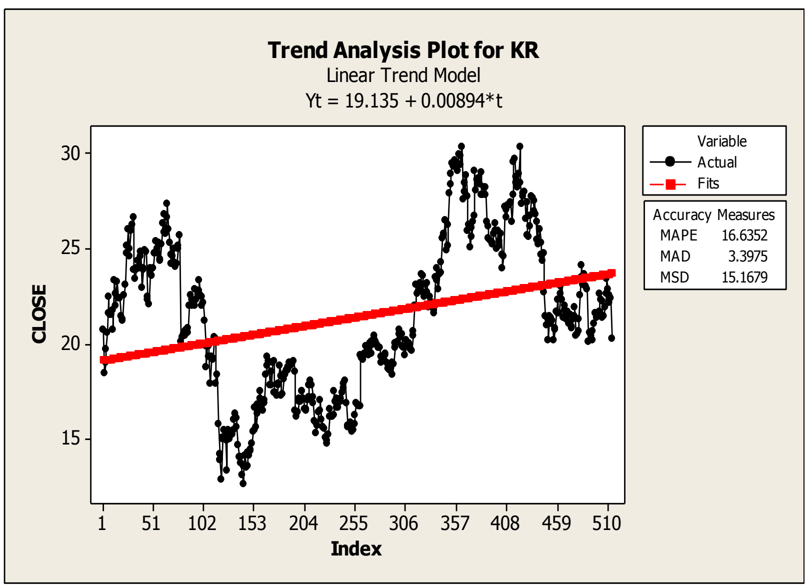

Figure 2.19: Trend Analysis for KR

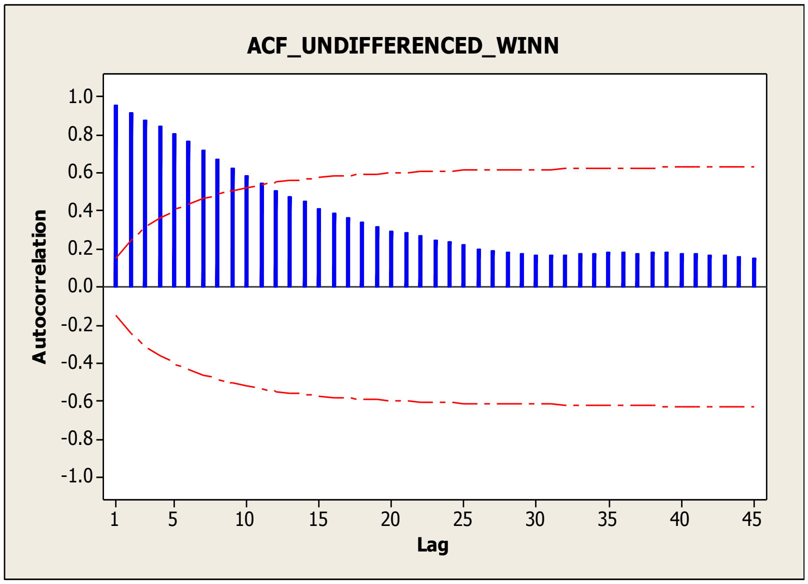

Figure 2.20: Autocorrelation Function for Undifferenced WINN

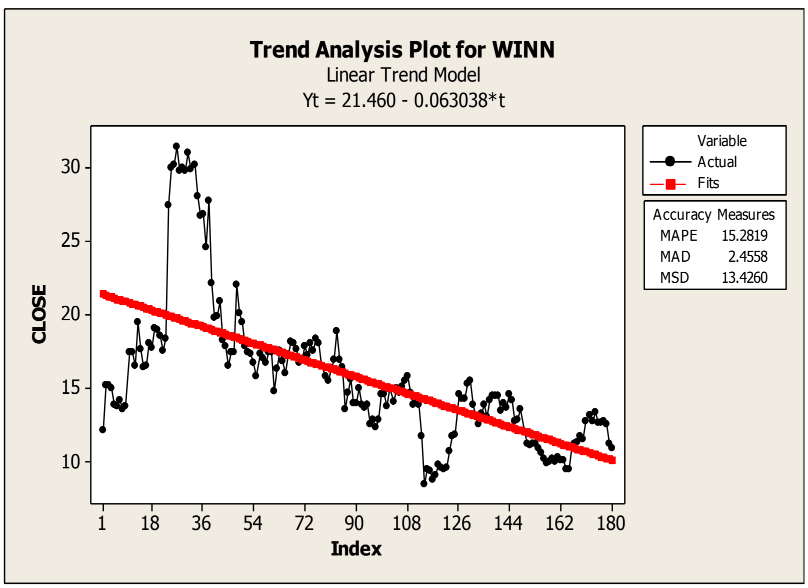

Figure 2.21: Trend Analysis for WINN

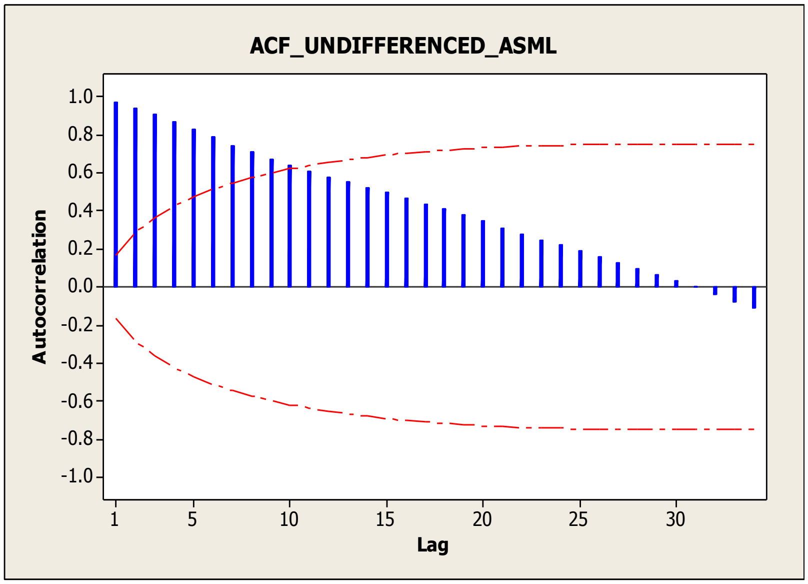

Figure 2.22: Autocorrelation Function for Undifferenced ASML

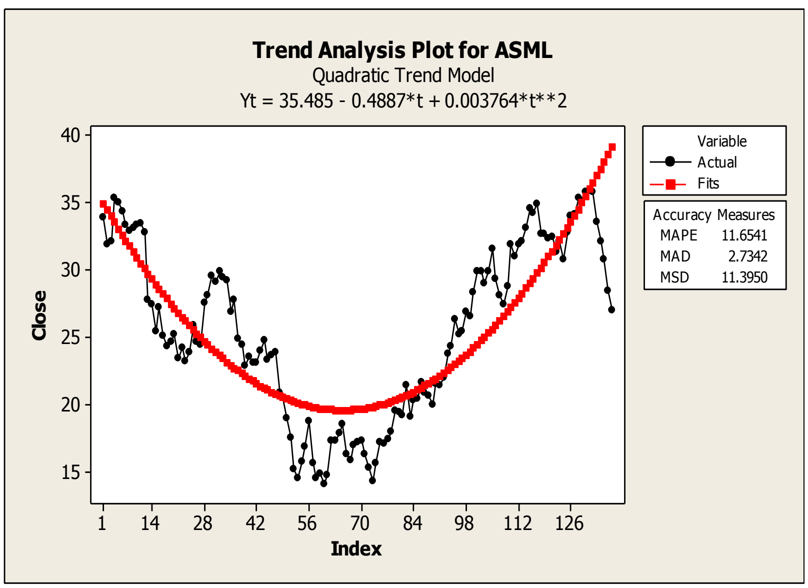

Figure 2.23: Trend Analysis for ASML

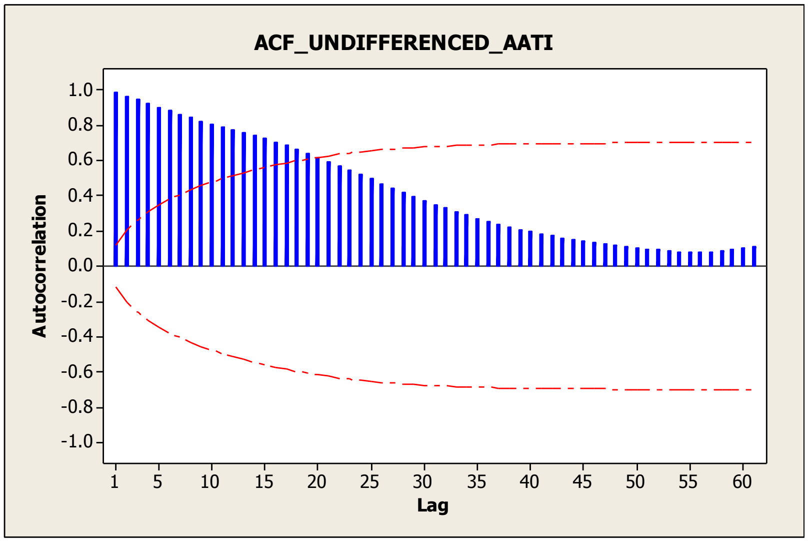

Figure 2.24: Autocorrelation Function for Undifferenced AATI

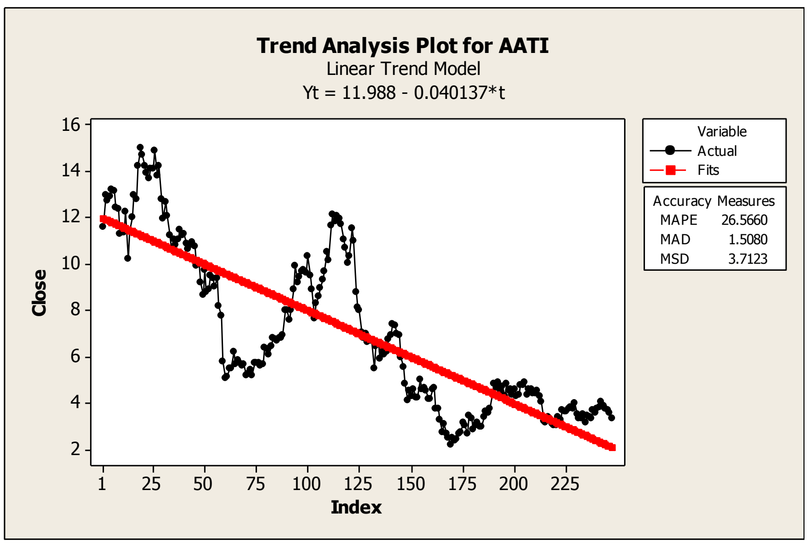

Figure 2.25: Trend Analysis for AATI

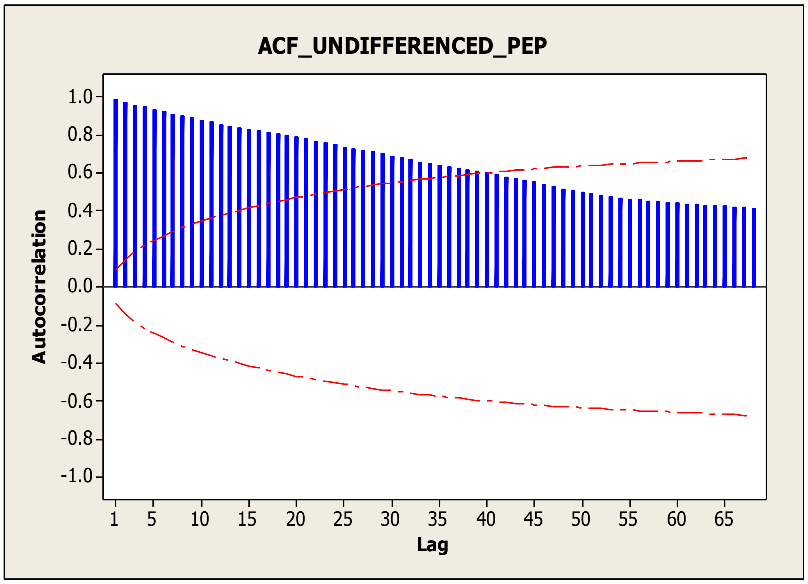

Figure 2.26: Autocorrelation Function for Undifferenced PEP

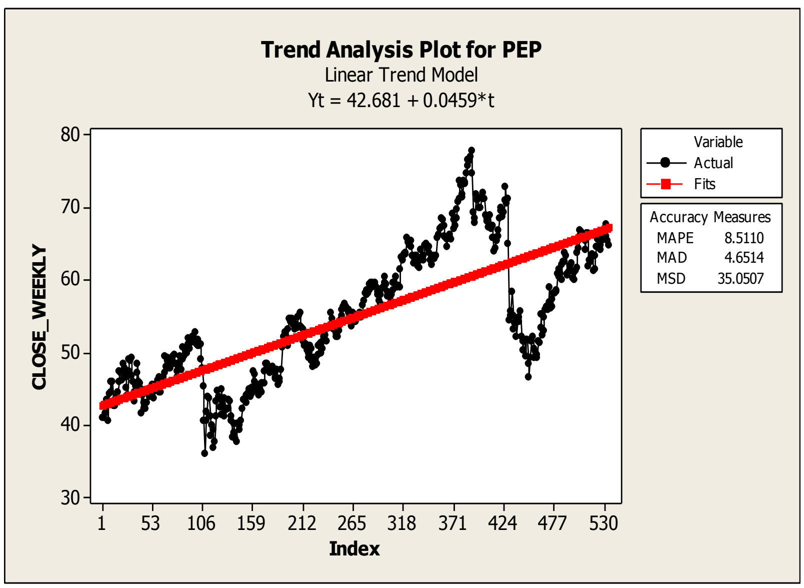

Figure 2.27: Trend Analysis for PEP

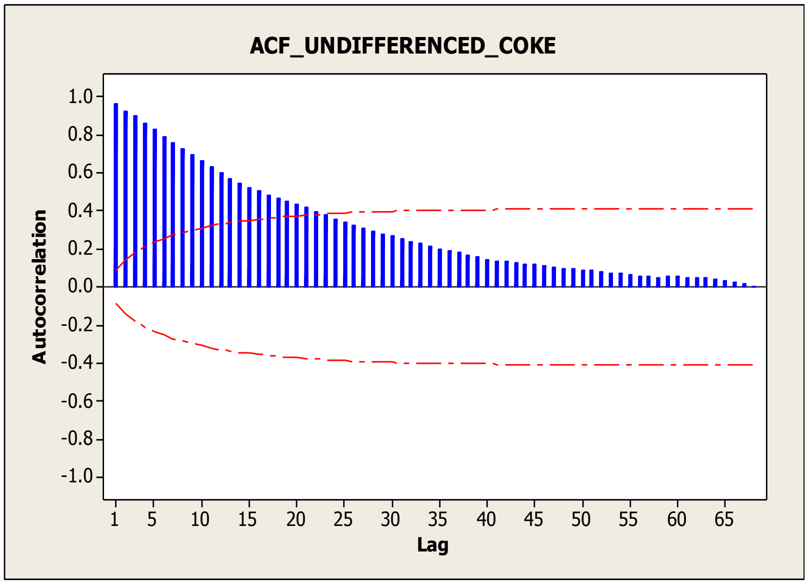

Figure 2.28: Autocorrelation Function for Undifferenced COKE

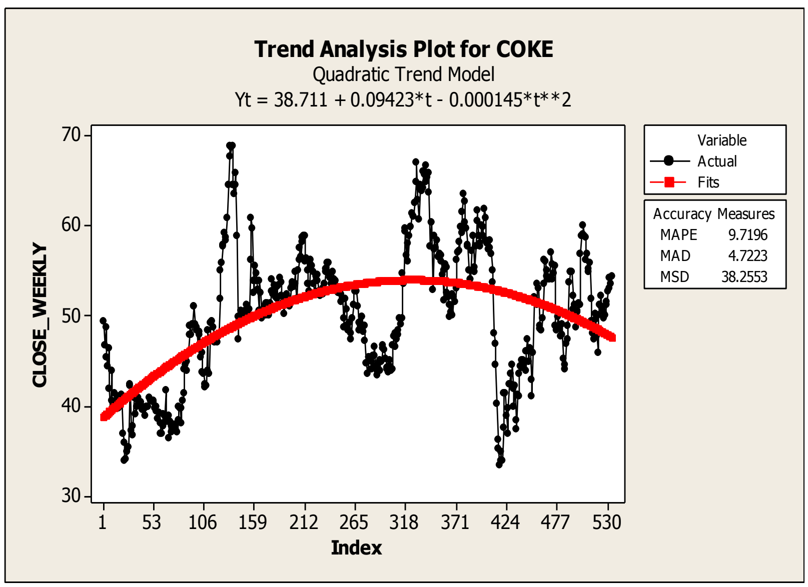

Figure 2.29: Trend Analysis for COKE

We again note that although Coke and Pepsi are in the same industry the two stocks behave very differently, this may largely be due to the frequent change in consumers taste and the constant introduction of similar products into the market.

Because the ACF for each of the stocks dies off slowly, we can conclude that all the stocks contain some trend. Once it is determined that the series is not stationary, then we model the trend. The trend in the series is modeled by simply fitting a general trend model to the series. This trend is most often linear or quadratic. This is represented in ARIMA (*p, d, q*) as $d = 0$ if there is no trend present, $d = 1$ if the trend is linear, and $d = 2$ if the trend is quadratic.

After developing a proposed model for the trend, we have to confirm that the chosen model is the most accurate by examining the measures of accuracy for fitted models. The mean absolute percentage error (MAPE), mean absolute deviation (MAD), and mean standard deviation (MSD), are three measures of accuracy that and the smaller they are the better the fit of the model. All three indicators measure the accuracy of the fitted time series values, but they express the results in different units. MAPE measure the accuracy of fitted time series values and expresses accuracy as a percentage. MAPE is calculated as $\frac{\sum_{t=1}^{n} \left| (y_t - \hat{y}_t) / y_t \right|}{n}$ x 100 where y_t equals the actual value, $\hat{y}_t$ equals the fitted value, and *n* equals the number of observations. MAD is expressed in the same units as the data and is calculated as $\frac{\sum_{t=1}^{n} \left| (y_t - \hat{y}_t) \right|}{n}$. And $MSD = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}$. Each of these three statistics can be used to compare the fits of the different models [5].

Once the trend has been modeled appropriately, we compute the autocorrelation and partial autocorrelation coefficients for the residuals of the trend model, or detrended data. The ACF and PACF are only useful indicators of the order *p* and *q* of the ARIMA (*p*,*q*,*d*) model if the series for which they are computed is stationary.

After deciding on a tentative model, it is essential to perform diagnostic checks to determine if the selected model contains the smallest possible number of parameters and that the chosen model is the most appropriate. The least squares criterion is used to obtain the estimates for the parameters, and by examining the p-value of the t-test, it can be determined whether or not the parameters are statistically significant (different from zero). In some instances, further analysis of the variances at different lags $k \geq p$, may indicate that an AR(p) is equivalent to ARIMA (1,1), but using the AR(p) would entail the use of additional parameters.

A good way to check the adequacy of an overall ARIMA model is to analyze the residuals obtained from the model, by calculating the autocorrelation and partial autocorrelation function for the residuals. Using the ACF and PACF we can examine a statistic that determines whether the first K sample autocorrelations of the residuals, considered together, indicate adequacy of the model.

One such statistic is the Ljung-Box statistic: $Q^* = n'(n'+2)\sum_{t=1}^{K}(n'-l)^{-1}r_l^{\,2}(\hat{a})$.

Here $n' = n-(d+LD)$ where n is the number of observations in the original time series, L is the number of seasons in a year (if seasonal variation is present), and d and D, respectively, the degrees of nonseasonal and seasonal differencing used to transform the original time series into a stationary time series. $r_l^2(\hat{a})$ is the square of $r_l\ (\hat{a})$, the sample autocorrelation of the residuals at lag l that is, the sample autocorrelation of residuals separated by lag of l time units [2]. Because the modeling process is supposed to account for the relationship between the time series observations, the residuals of the model should be uncorrelated. Hence, the autocorrelations of the residuals should be small, resulting in a small Q^*. A large value of Q^* indicates that the model is inadequate. We can reject the adequacy of the model under consideration by setting the probability of Type I error equal to α if and only if either of the following equivalent conditions hold

1. Q^* is greater $\chi^2_{[\alpha]}(K-n_p)$, the point on the scale of the chi-square distribution having $K-n_p$ degrees of freedom, such that there is an area of α under the curve of the distribution above the point. Here n_p is the number of parameters that must be estimated in the model under consideration.
2. P-value is less that α where p-value is the area under the curve of the Chi-Square distribution having $K-n_p$ degrees of freedom to the right of Q^*

The following chapter contains the result of applying the outlined process to the previously mentioned stocks.

CHAPTER 3

3.1. Results

Table 3.1 provides the confirmed model for each stock at each interval. Confirmed, meaning that the models contain only significant parameters and the residuals are both random and approximately Normal.

Table 3.1

Confirmed Models

	Weekly	1^{ST} Trading Day of the Month	15^{th} Trading Day of the Month
KR	ARIMA(1,1,0) $w_t = -.0879 w_{t-1} + \varepsilon_t$	AR(1) $y_t = 1.4002 + .9350 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 1.6845 + .9260 y_{t-1} + \varepsilon_t$
WINN	AR(1) $y_t = .36313 + .9732 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 1.9935 + .8575 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 1.2664 + .9083 y_{t-1} + \varepsilon_t$
ASML	AR(1) $y_t = .5347 + .9815 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 3.1738 + .8892 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 2.3721 + .9158 y_{t-1} + \varepsilon_t$
AATI	AR(1) $y_t = .9952 y_{t-1} + \varepsilon_t$	AR(1) $y_t = .9714 y_{t-1} + \varepsilon_t$	AR(1) $y_t = .9717 y_{t-1} + \varepsilon_t$
PEP	AR(1) $y_t = 1.5886 + .9732 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 6.0996 + .8984 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 2.2305 + .9595 y_{t-1} + \varepsilon_t$
COKE	AR(1) $y_t = 1.78 + .96 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 5.72 + .89 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 7.70 + 8643 y_{t-1} + \varepsilon_t$
MSFT	AR(1) $y_t = .98941 + .9625 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 4.0799 + .8456 y_{t-1} + \varepsilon_t$	AR(1) $y_t = 3.3369 + .8644 y_{t-1} + \varepsilon_t$
APPL	ARIMA(1,1,0) $w_t = .4783 + 1.070 w_{t-1} + \varepsilon_t$	ARIMA(0,1,0) $w_t = w_{t-1} + \varepsilon_t$	ARIMA(1,1,0) $w_t = .2930 w_{t-1} + \varepsilon_t$

To determine which interval is possibly most effective, we compare the mean square errors (MSE) for each model at the different intervals. The smaller the mean squared error the better the fit of the model. Amongst all the industries, weekly data resulted in a better model. Monthly data sampled on the first only produced a better model than from the 15^{th} with COKE, KR, and PEP.

Table 3.2

Mean Squared Error for each model

	Weekly	1ST Trading Day of the Month	15th Trading Day of the Month
KR	.659	2.149	2.575
WINN	1.774	8.759	6.063
ASML	2.006	10.152	8.399
AATI	.267	1.440	1.431
PEP	2.221	8.278	8.930
COKE	3.830	11.800	16.380
MSFT	.754	3.230	2.743
APPL	26.600	156.700	139.100

3.1 a: Kroger Company

The resulting models for Kroger, were ARIMA (1,1,0) for data sampled weekly and AR(1) for the data sampled both on the 1^{st} trading day of the month and 15th trading day of the month. Looking at the single differenced ACFs below, of the data sampled on the 1^{st} and the 15^{th}, we recognize that they are almost exactly the same. Because the ACFs are similar it is expected that the model for both series will also be the same. For both Figures 3.1 and 3.2, the autocorrelations at low lags are very high, and decline slowly as the lags increase. However, a well-defined pattern does exist because the ACFs continue to decrease and become negative; this is behavior that is expected of a trended series, not one that has already been differenced. This suggests that although trend is present, it is not significant. We conclude that the appropriate model is an AR(1). An AR(1) tells us that this week's (or month's) price is a function of last week's (month's) price plus some constant and error term.

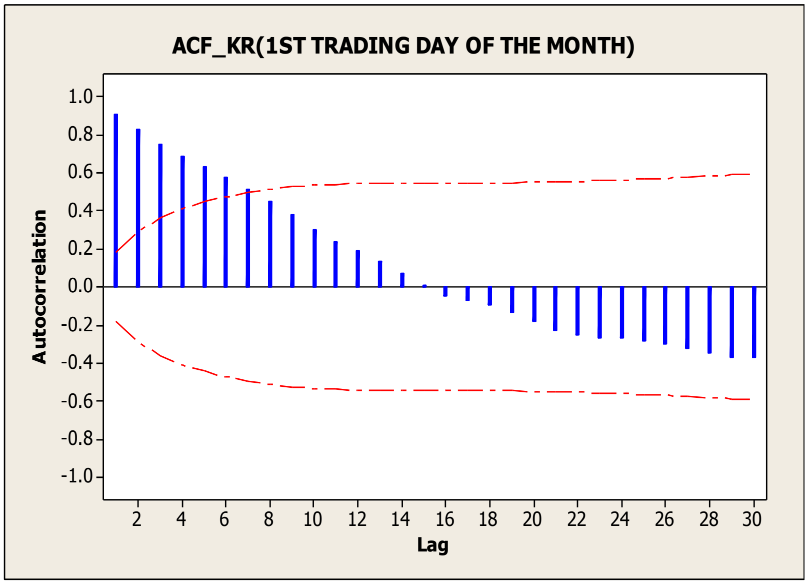

Figure 3.1: Autocorrelation function for single differenced KR sampled on the 1st trading day of the month

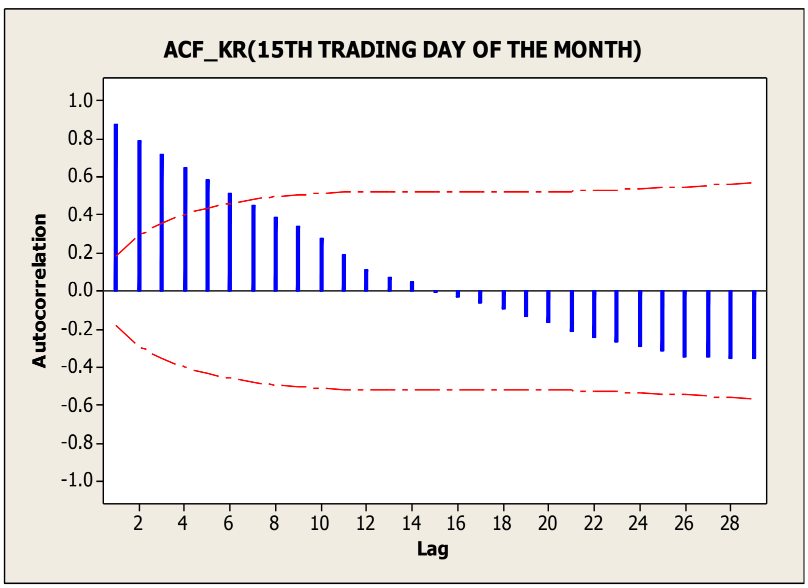

Figure 3.2: Autocorrelation function for single differenced KR sampled on the 15^{th} trading day of the month

Table 3.3

Partial Minitab output for KR sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type          Coef   SE Coef       T       P
AR    1     0.9350    0.0319   29.27   0.000
Constant    1.4002    0.1306   10.72   0.000
Mean        21.535     2.009

Number of observations:  126
Residuals:     SS =  266.418 (backforecasts excluded)
               MS =  2.149  DF = 124

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12      24      36      48
Chi-Square    11.3    17.7    24.2    28.9
DF              10      22      34      46
P-Value      0.335   0.725   0.893   0.977
```

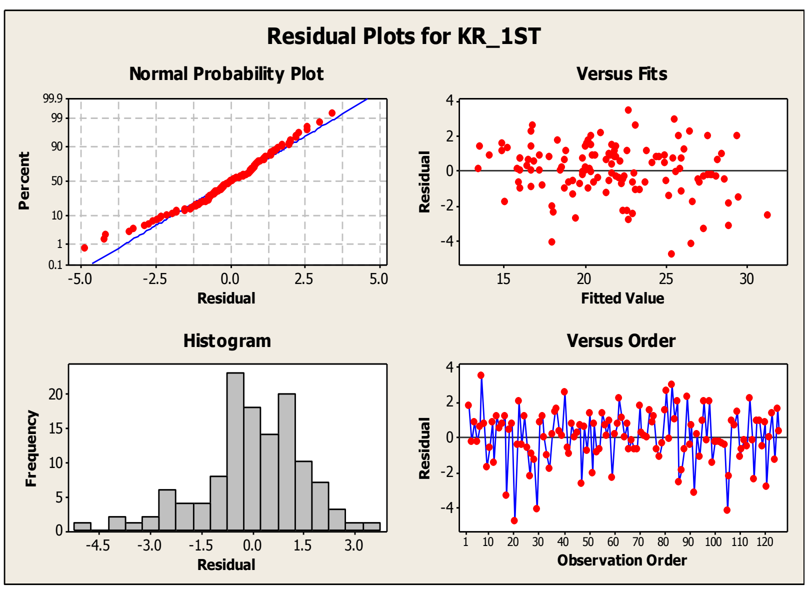

Figure 3.3: Four-in-One Residual Plots for KR sampled on the 1st trading day of the month

Table 3.4

Partial Minitab output for KR sampled on the 15th trading day of the month

```
Final Estimates of Parameters

Type         Coef  SE Coef       T      P
AR   1     0.9206   0.0359   25.63  0.000
Constant   1.6845   0.1454   11.59  0.000
Mean       21.220    1.831

Number of observations:  122
Residuals:    SS =  308.980 (backforecasts excluded)
              MS =  2.575   DF = 120

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12      24      36      48
Chi-Square     9.1    19.9    26.1    30.2
DF              10      22      34      46
P-Value      0.523   0.589   0.831   0.966
```

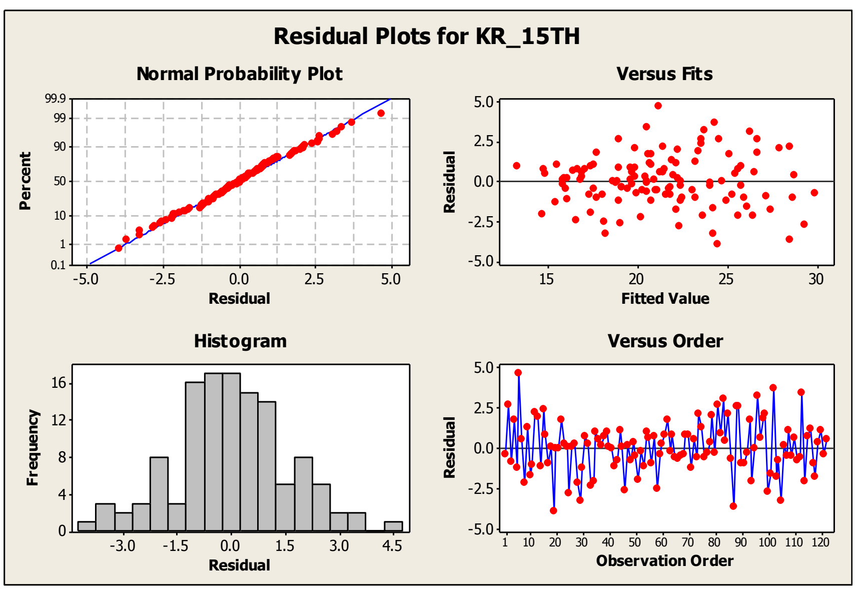

Figure 3.4: Four-in-One Residual Plots for KR sampled on the 15th trading day of the month

Because all the p-values are much less than .05 and the behavior of the residuals are Normal we can conclude that AR(1) is a reasonable fit for the data that was sampled on the 1st and 15th trading day of the month.

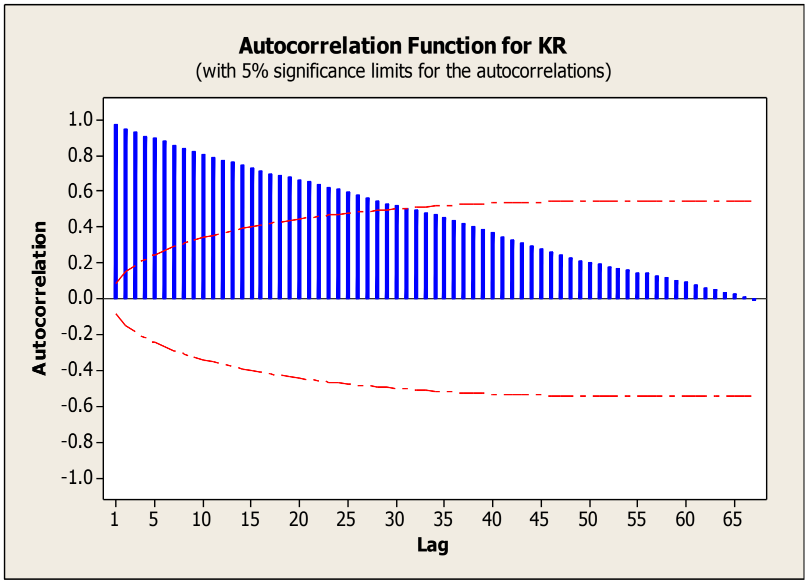

Figure 3.5: Autocorrelation Function for single differenced KR sampled weekly

The ACF for the series that was sampled weekly behaves differently than the two previously mentioned. They are all similar because the autocorrelations are highly correlated for lower lags and then decreases as the lags increase, but for this series the ACF continues to die off positively. Also compared to the ACF for the original series, differencing caused the ACF to die off faster. We can assume the model for this series should be ARIMA (1,1,0). The Minitab output also supports this claim. The MSE for the model ARIMA (1,1,0) was.659. Also the histogram and probability plot of the residuals suggests that the residual are Normal.

Table 3.5

Partial Minitab output for KR sampled weekly(with constant)

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1    -0.0879   0.0432  -2.03  0.042
Constant  0.00295  0.03514   0.08  0.933

Differencing: 1 regular difference
Number of observations:  Original series 535, after differencing 534
Residuals:    SS =  350.816 (backforecasts excluded)
              MS =  0.659  DF = 532

Modified Box-Pierce (Ljung-Box) Chi-Square statist
Lag            12     24     36     48
Chi-Square   13.9   25.1   43.7   49.5
DF             10     22     34     46
P-Value     0.179  0.292  0.124  0.337
```

Table 3.6

Partial Minitab output for KR sampled weekly

```
Final Estimates of Parameters

Type        Coef  SE Coef      T      P
AR   1  -0.0879   0.0432  -2.04  0.042

Differencing: 1 regular difference
Number of observations:  Original series 535, after differencing 534
Residuals:    SS =  350.821 (backforecasts excluded)
              MS =  0.658  DF = 533

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   13.9   25.1   43.7   49.5
DF             11     23     35     47
P-Value     0.240  0.345  0.150  0.375
```

Because the constant term is not significantly different from zero, the model was refit without a constant, resulting in the following output:

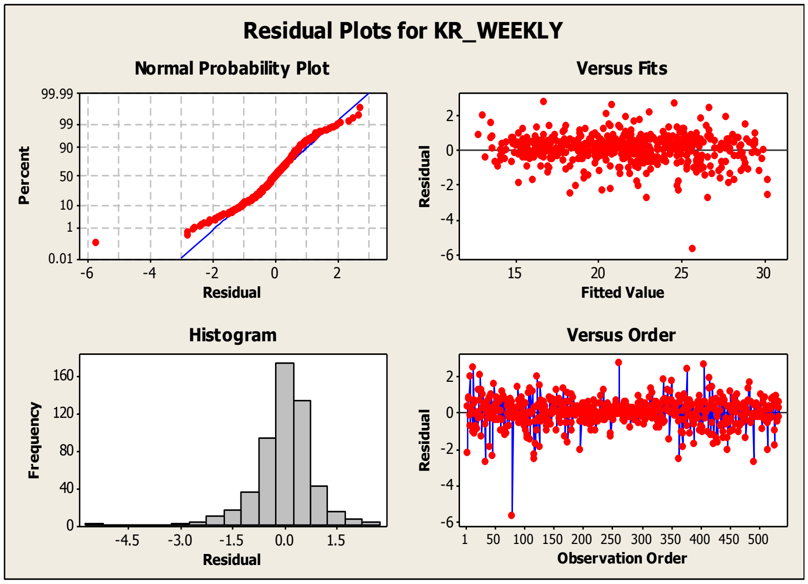

Figure 3.6: Four-in-One Residual Plots for KR sampled weekly

Comparing the MSE for each sampling interval, from Table 3.2, we can conclude that the model series for KR that was sampled weekly fits better than data sampled on the 1st or 15th trading day of the month.

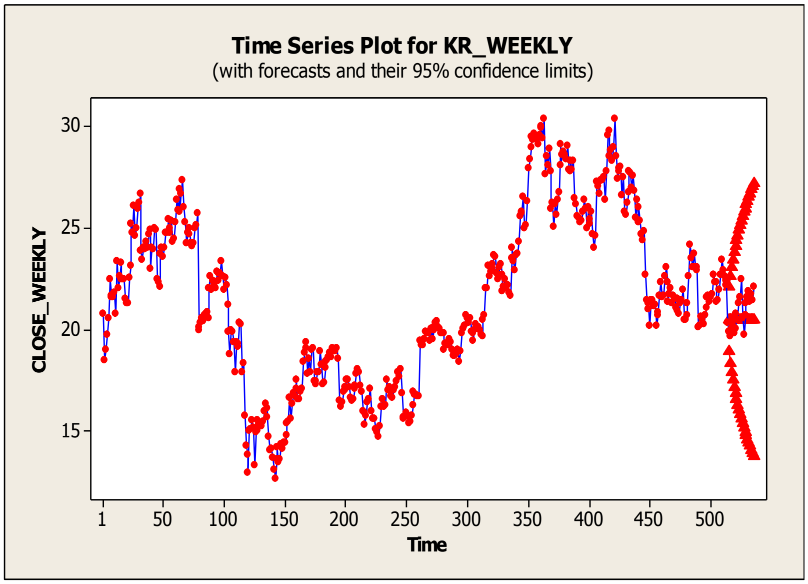

Figure 3.7: Time Series Plot of KR sampled weekly with forecasts

Table 3.7

Analysis of forecasts for KR sampled weekly (ARIMA (1,1,0))

Dates	Actual	Forecast
6/1/2010	19.87	20.48
6/8/2010	19.64	20.47
6/11/2010	19.99	20.47
6/22/2010	20.18	20.48
6/29/2010	19.81	20.48
7/6/2010	20.02	20.48
7/13/2010	20.78	20.48
7/20/2010	20.49	20.49
7/27/2010	21.28	20.49
8/3/2010	21.58	20.49
8/10/2010	22.43	20.49
8/24/2010	20.540	20.50
8/31/2010	19.730	20.50
9/7/2010	20.700	20.50
9/14/2010	21.260	20.51
9/21/2010	21.600	20.51
9/28/2010	21.880	20.51
10/5/2010	21.370	20.51
10/12/2010	21.690	20.51
10/19/2010	21.42	20.52
10/26/2010	22.070	20.52

The mean absolute percentage error (MAPE) for the forecasted values is .036292. When examining the forecasts for Kroger, for the data sampled weekly, we notice that the fits barely change. For the five month period the fits had a range of 20.48 – 20.52. However, the actual prices had a range of 19.64 – 22.43. There is not much change in the fits because the coefficient of the autoregressive term is -.0879 and the model does not contain a constant, so the forecast’s price is only -.0879 of the previous week’s price. The model is nearly ARIMA (0,1,0) ; this is most likely due to the random behavior in approximately the last 100 observations, compared to the prior observations that contained quadratic trend. Modeling the series using only the last 100 observation, we see that fits behave more like the actuals, compared to the series that contained all the data points. The confirmed model for this series is AR(1.)

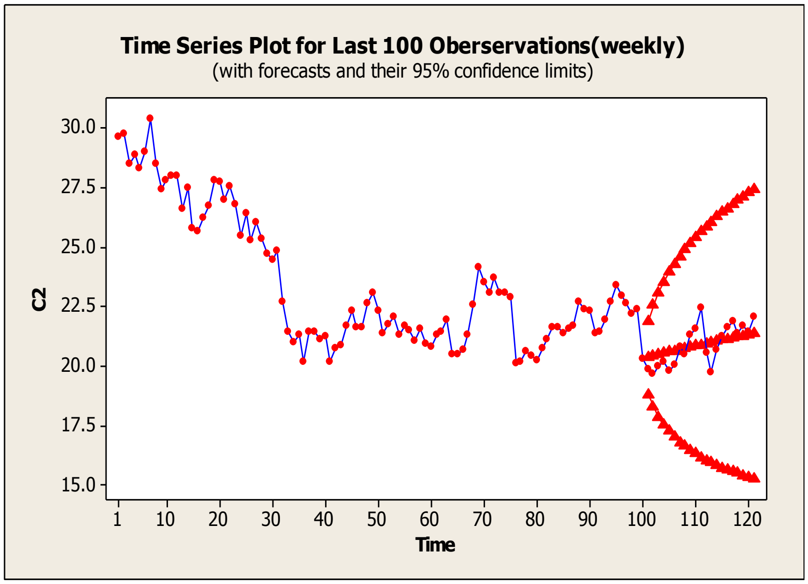

Figure 3.8: Time Series Plot of KR sampled weekly (last 100 observations) with forecasts

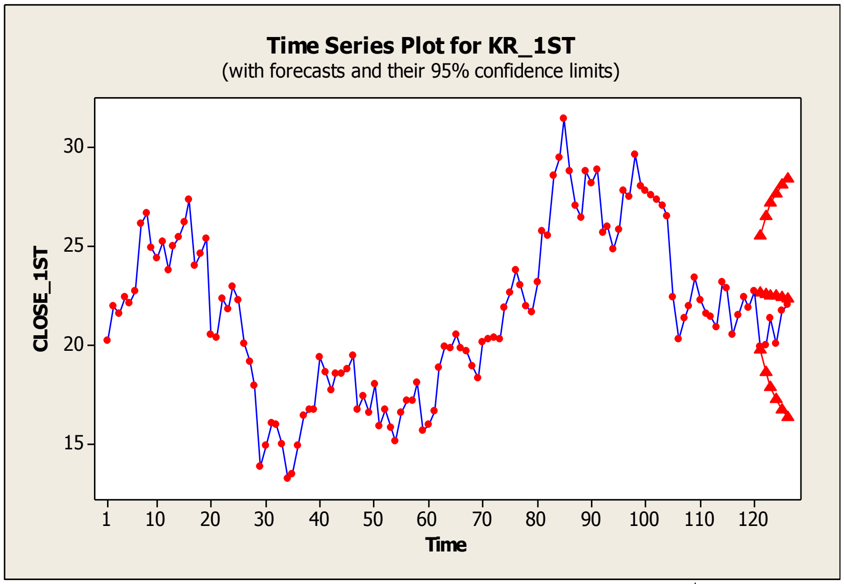

Figure 3.9: Time Series Plot of KR sampled on the 1st trading day of the month with forecasts

Table 3.8

Analysis of forecasts for KR sampled on the 1st trading day of the month (AR(1))

Date	Actual	Forecasts
6/1/2010	19.87	22.66
7/1/2010	20.01	22.59
8/2/2010	21.38	22.52
9/1/2010	20.06	22.46
10/1/2010	21.72	22.40
11/1/2010	22.01	22.34

The MAPE for the forecasted values is 0.081.The fits of the model show a slight downward trend, because only a portion (.9350) of last month's prices is used to predict the forecasts. This differs from the behavior of the actual values, which increased during this time frame. So although the model is an appropriate fit statistically, the results would not be desirable for an investor.

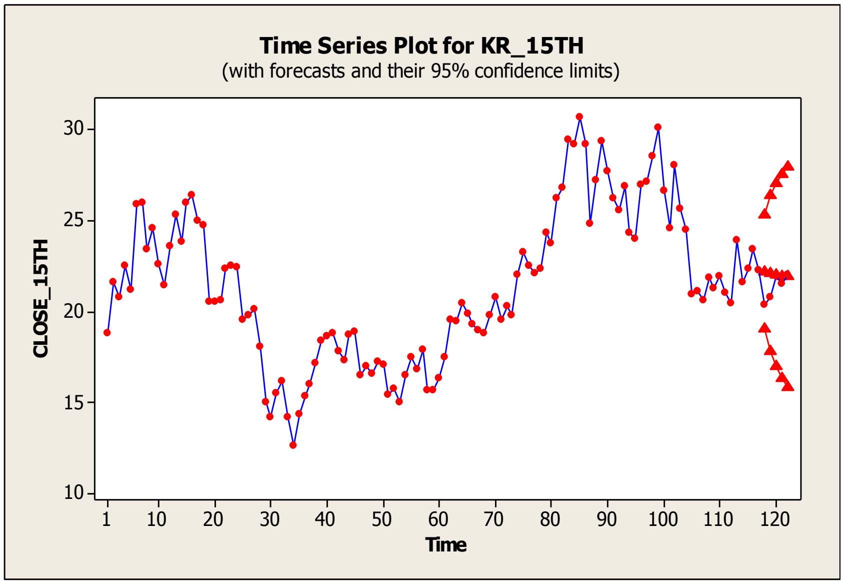

Figure 3.10: Time Series Plot of KR sampled on the 15th trading day of the month with forecasts

Table 3.9

Analysis of forecasts for KR sampled on the 15th trading day of the month (ARIMA(1,1,0))

Date	Actual	Forecasts
6/15/2010	20.38	22.159
7/15/2010	20.79	22.084
8/16/2010	21.92	22.016
9/15/2010	21.49	21.95
10/15/2010	22.01	21.89

The MAPE for forecasted values is .0361. Examining the actual versus the fits for the data sampled on the 15th trading day of the month, their behavior is similar to the data sampled on the 1st trading day of the month. The fits decreased, while the actual values increased. This model would also not be desirable for an investor.

3.1 b. Winn-Dixie Stores Inc.

The resulting models for Winn -Dixie, were all AR(1). Looking at the ACFs below, we recognize that they behave the same, except for the weekly data containing more lags.

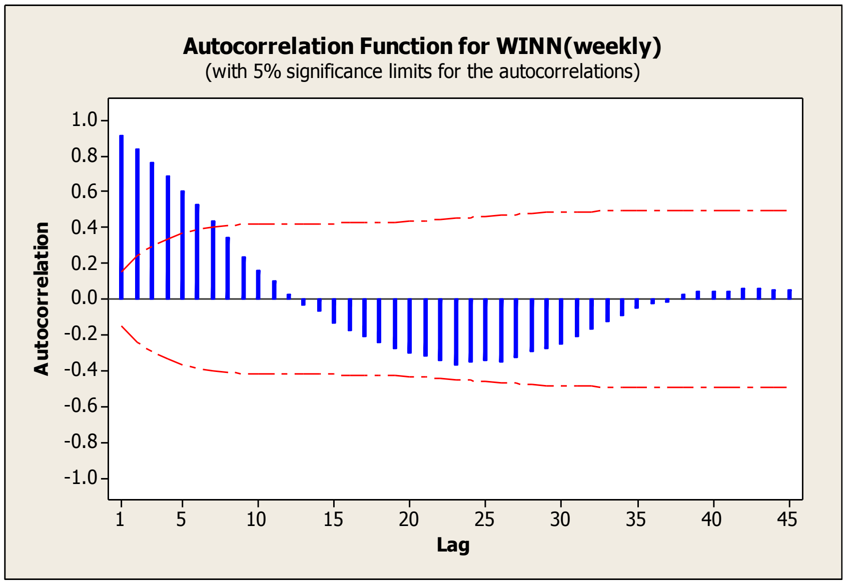

Figure 3.11: Autocorrelation Function for single differenced WINN sampled weekly

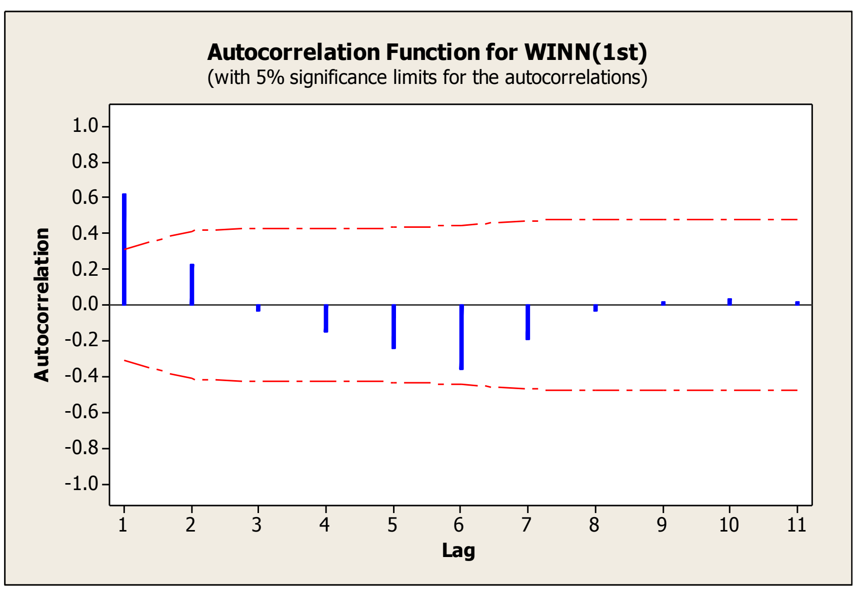

Figure 3.12: Autocorrelation Function for single differenced WINN sampled on the 1st trading day of the month

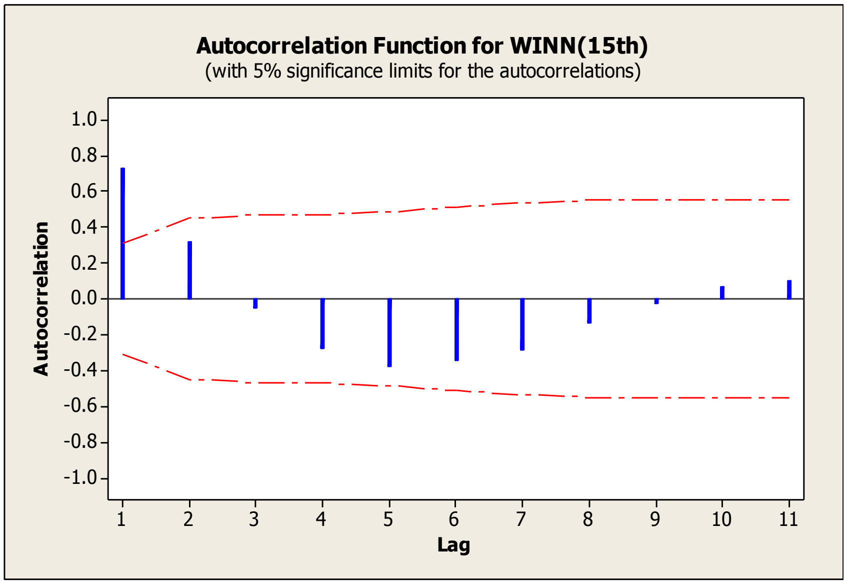

Figure 3.13: Autocorrelation Function for single differenced WINN sampled on the 15th trading day of the month

Table 3.10

Partial Minitab Output for WINN sampled weekly

```
Final Estimates of Parameters

Type          Coef   SE Coef      T      P
AR   1      0.9732    0.0183  53.09  0.000
Constant   0.36313   0.09815   3.70  0.000
Mean        13.549     3.662

Number of observations:  200
Residuals:    SS =  351.223 (backforecasts excluded)
              MS =  1.774  DF = 198

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square     9.8   31.9   48.1   59.5
DF              10     22     34     46
P-Value      0.460  0.079  0.055  0.087
```

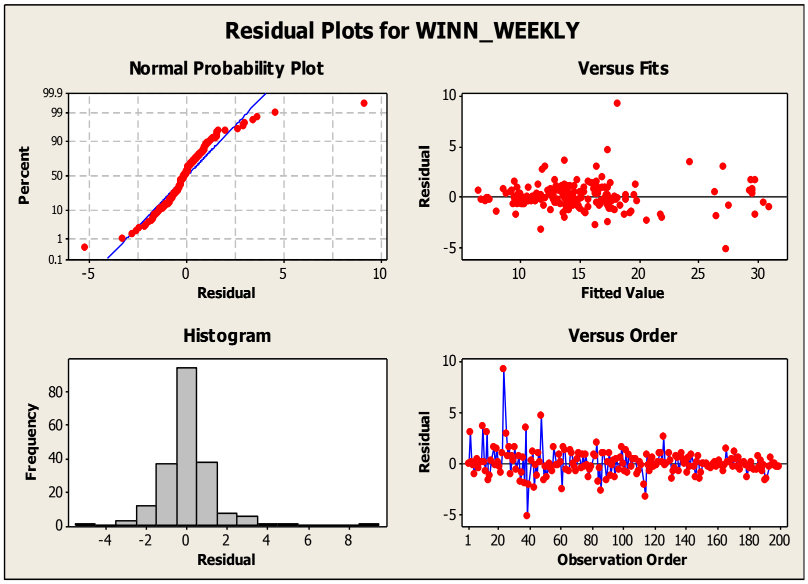

Figure 3.14: Four-in-One Residual Plots for WINN sampled weekly

Table 3.11

Partial Minitab Output for WINN sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.8575   0.0835  10.27  0.000
Constant   1.9935   0.4364   4.57  0.000
Mean       13.994    3.064

Number of observations:  48
Residuals:     SS =  402.904 (backforecasts excluded)
               MS =  8.759   DF = 46

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12      24      36  48
Chi-Square     4.7    14.9    20.1   *
DF              10      22      34   *
P-Value      0.913   0.868   0.972   *
```

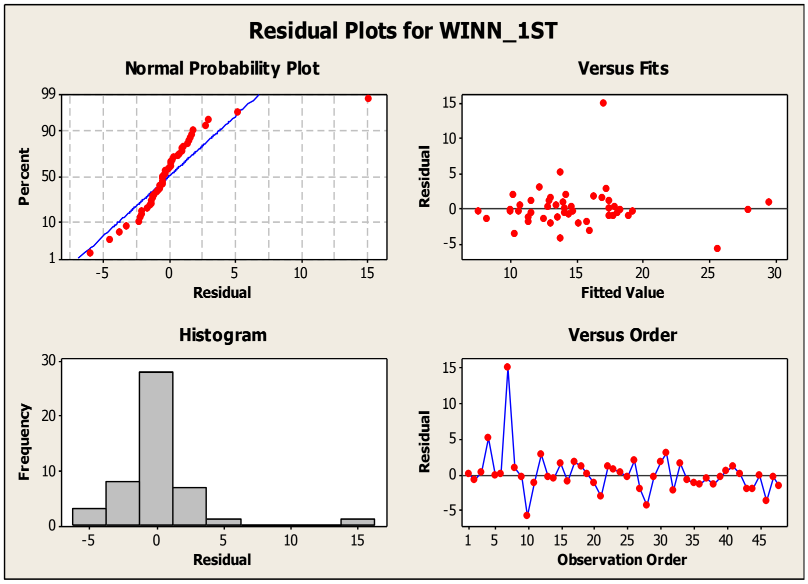

Figure 3.15: Four-in-One Residual Plots for WINN sampled on the 1st trading day of the month

Table 3.12

Partial Minitab Output for WINN sampled on the 15th trading day of the month

```
Final Estimates of Parameters

Type          Coef  SE Coef       T       P
AR    1     0.9083   0.0706   12.87   0.000
Constant    1.2664   0.3708    3.42   0.001
Mean        13.805    4.042

Number of observations:   47
Residuals:     SS =  272.844 (backforecasts excluded)
               MS =  6.063   DF = 45

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12      24      36   48
Chi-Square    13.9    25.6    37.5    *
DF              10      22      34    *
P-Value      0.178   0.268   0.313    *
```

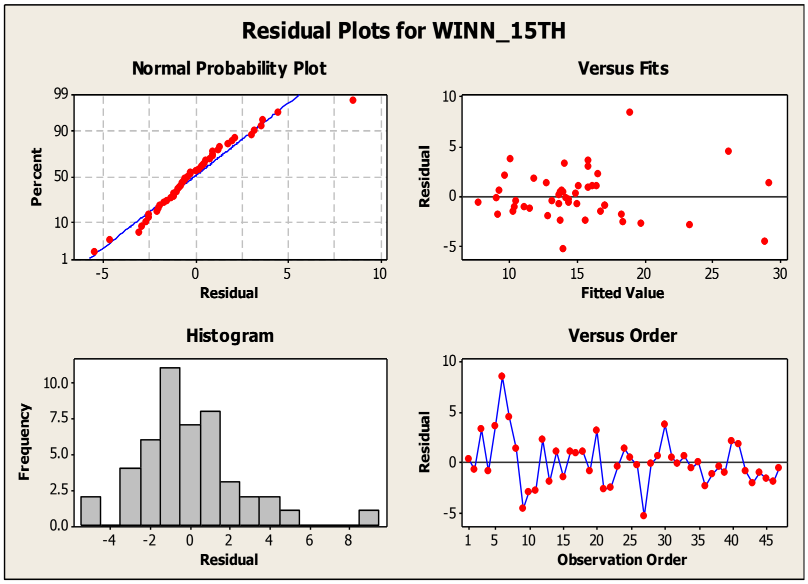

Figure 3.16: Four-in-One Residual Plots for WINN sampled on the 15th trading day of the month

The Minitab output confirms that an AR(1) is the appropriate model for each of the series. The small p-value, essentially zero, for the autoregressive parameter and the constant tell us that these parameters are significantly different from zero. Also, the histogram and probability plot of the residuals confirms Normality. But we do acknowledge that all series have a high outlier, which will be examined in more detail later in the paper. However, data that is sampled on a weekly basis produces the best model based on the MSE.

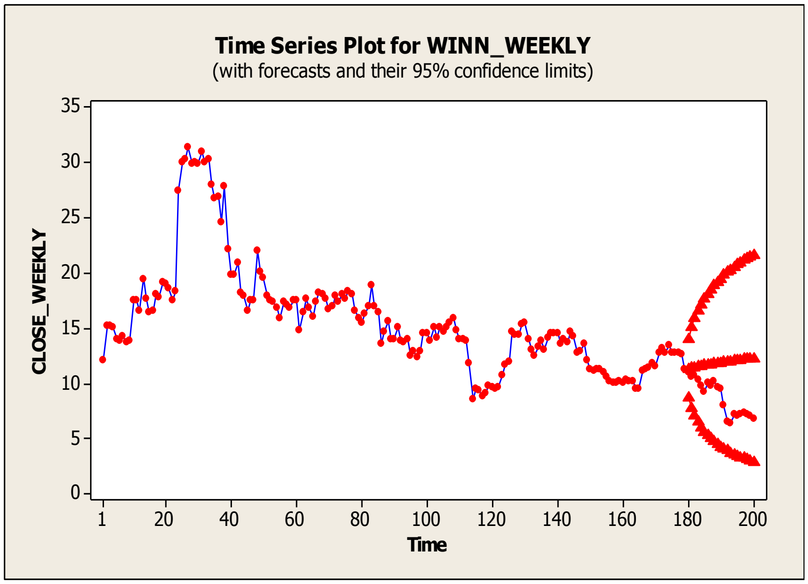

Figure 3.17: Time Series Plot for WINN sampled weekly with forecasts

Table 3.13

Forecasts analysis for WINN sampled weekly (AR(1))

Date	Actual	Forecasts
6/1/2010	10.93	11.31
6/8/2010	10.58	11.37
6/11/2010	10.72	11.43
6/22/2010	10.34	11.49
6/29/2010	9.74	11.54
7/6/2010	9.18	11.60
7/13/2010	9.97	11.65
7/20/2010	9.80	11.70
7/27/2010	10.22	11.75
8/3/2010	9.60	11.80
8/10/2010	9.54	11.84
8/24/2010	7.94	11.89
8/31/2010	6.56	11.93
9/7/2010	6.40	11.98
9/14/2010	7.17	12.02
9/21/2010	7.08	12.06
9/28/2010	7.17	12.10
10/5/2010	7.27	12.14
10/12/2010	7.11	12.18
10/19/2010	7.00	12.21
10/26/2010	6.79	12.25

The MAPE for the forecasted values is 0.424. Although the confirmed model for the data sampled weekly was a AR(1) the fits increased while the actual values decreased. This is because the confirmed model is $y_t = .36313 + .9732 y_{t-1} + \varepsilon_t$. Although most of last week's price (.9732) is used to forecasts this week's price, the forecasts increase because the constant in the model (.36313) is greater than nearly all the differences of consecutive decreasing actual values. So, this confirmed model would not be valuable for an investor, because the fits do not behave like the actual values. The same is true for the data that was sampled on the 1st and 15th trading day of the month. The models for those series were

$y_t = 1.9935 + .8575 y_{t-1} + \varepsilon_t$ and $y_t = 1.2664 + .9083 y_{t-1} + \varepsilon_t$.

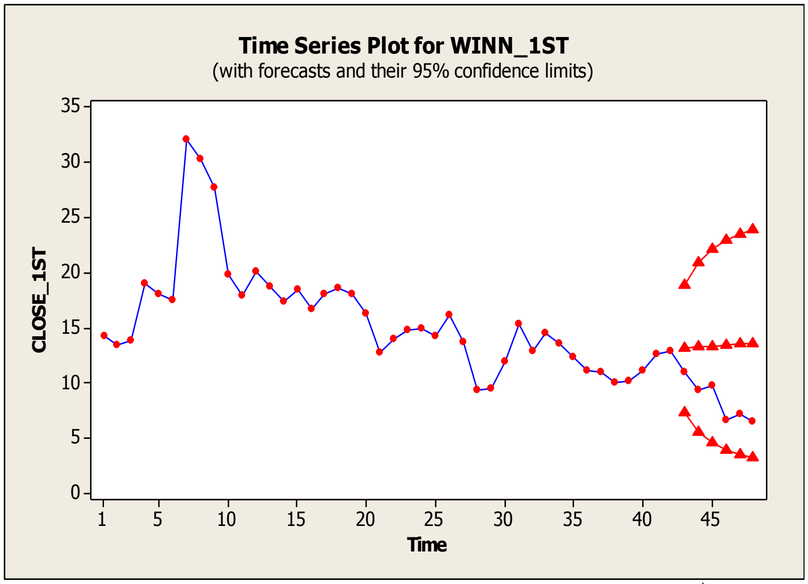

Figure 3.18: Time Series for WINN sampled on the 1st trading day of the month with forecasts

Table 3.14

Forecasts analysis for WINN sampled on the 1st trading day of the month (AR(1))

Date	Actual	Forecasts
6/1/2010	10.93	13.09
7/1/2010	9.33	13.22
8/2/2010	9.73	13.33
9/1/2010	6.58	13.42
10/1/2010	7.22	13.51
11/1/2010	6.53	13.58

The MAPE for the forecasted values is 0.662.

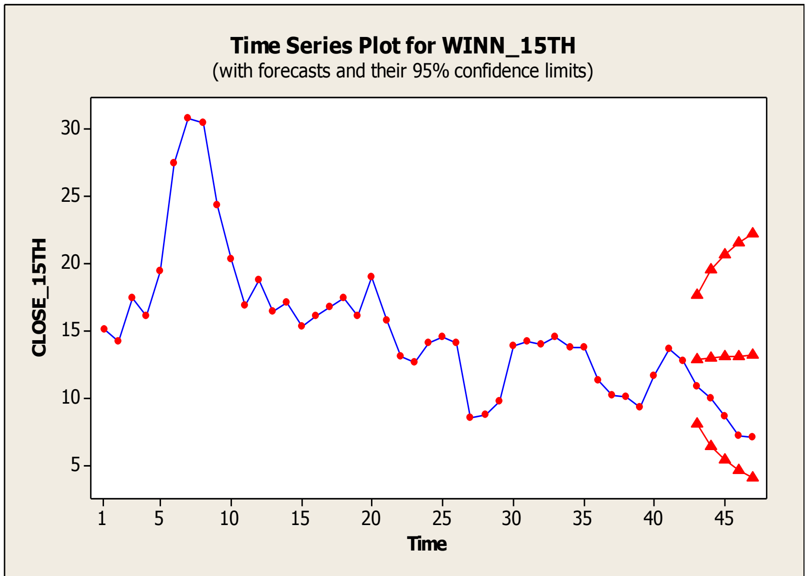

Figure 3.19: Time Series for WINN sampled on the 15th trading day of the month with forecasts

Table 3.15

Forecasts analysis for WINN sampled on the 15th trading day of the month (AR(1))

Date	Actual	Forecasts
6/15/2010	10.80	12.86
7/15/2010	9.93	12.94
8/16/2010	8.64	13.02
9/15/2010	7.16	13.09
10/15/2010	7.11	13.16

The MAPE for the forecasted values is 0.536.

3.1 c. ASML Holdings

The resulting models for ASML Holdings were all AR(1), although the results from the trend analysis in Chapter 2 may have suggested otherwise. The trend analysis in

Chapter 2 for the weekly data shows that the modeled trend for this series is quadratic. Also, the ACF for the original series compared to the series differenced twice shows that differencing the series twice does cause the ACF to die off faster. So we initially proposed that the appropriate model is AR(0,2,0), because the ACF of the twice differenced series behaves like a random walk.

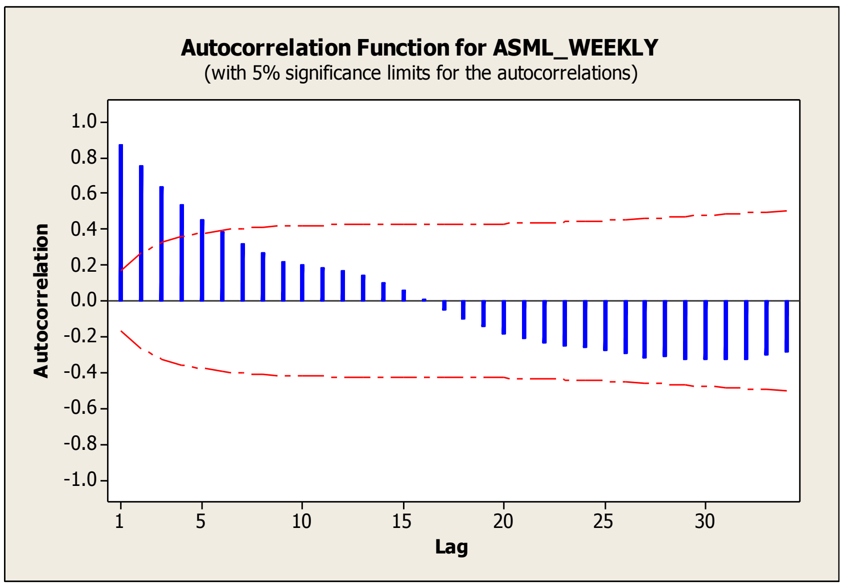

Figure 3.20: Autocorrelation Function ASML differenced twice sampled weekly

However the Minitab output for ASML sampled weekly for the model ARIMA (2,2,0), suggests that differencing twice causes us to overfit the model, because based on the behavior of the ACF no AR parameters should be significant.

Table 3.16

Partial Minitab Output for ASML sampled weekly (ARIMA (2,2,0))

```
Final Estimates of Parameters

Type          Coef   SE Coef      T      P
AR   1     -0.7127    0.0828  -8.60  0.000
AR   2     -0.2936    0.0828  -3.55  0.001
Constant   -0.0199    0.1341  -0.15  0.882

Differencing: 2 regular differences
Number of observations:  Original series 137, after differencing 135
Residuals:    SS =  320.662 (backforecasts excluded)
              MS =  2.429  DF = 132

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   18.3   31.3   51.2   75.1
DF              9     21     33     45
P-Value     0.032  0.069  0.022  0.003
```

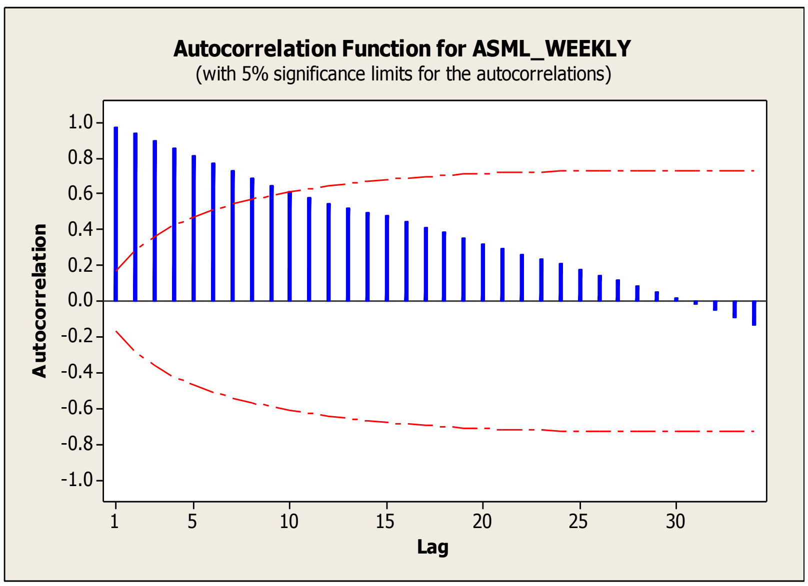

Figure 3.21: Autocorrelation function for ASML differenced once sampled weekly

So we then examine the ACF for the weekly data differenced once. The resulting ACF is that of a trended model. This is the same scenario that happened with KR and for reasons explained earlier, the appropriate model for the ASML data sampled weekly is

AR(1). For the data that was sampled on the 1st and the 15th, the resulting differenced ACFs are again trended series, so the model for these series should also be AR(1).

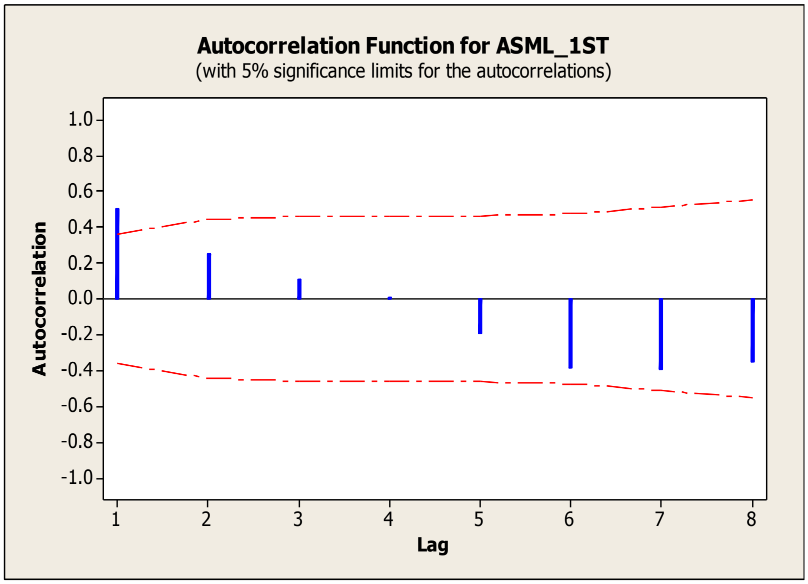

Figure 3.22: Autocorrelation Function for ASML differenced twice sampled on the 1st trading day of the month

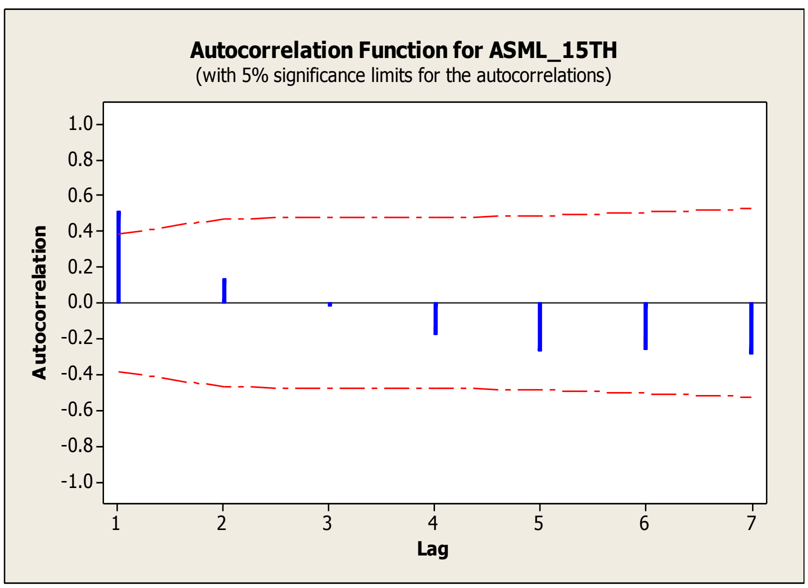

Figure 3.23: Autocorrelation Function for ASML differenced twice sampled on the 15th trading day of the month

Table 3.17

Partial Minitab Output for ASML sampled weekly (AR(1))

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.9815   0.0183  53.51  0.000
Constant    0.5347   0.1236   4.33  0.000
Mean        28.874    6.675

Number of observations:  158
Residuals:    SS =  312.909 (backforecasts excluded)
              MS =  2.006  DF = 156

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square     6.2   18.7   29.7   41.6
DF              10     22     34     46
P-Value      0.802  0.664  0.679  0.657
```

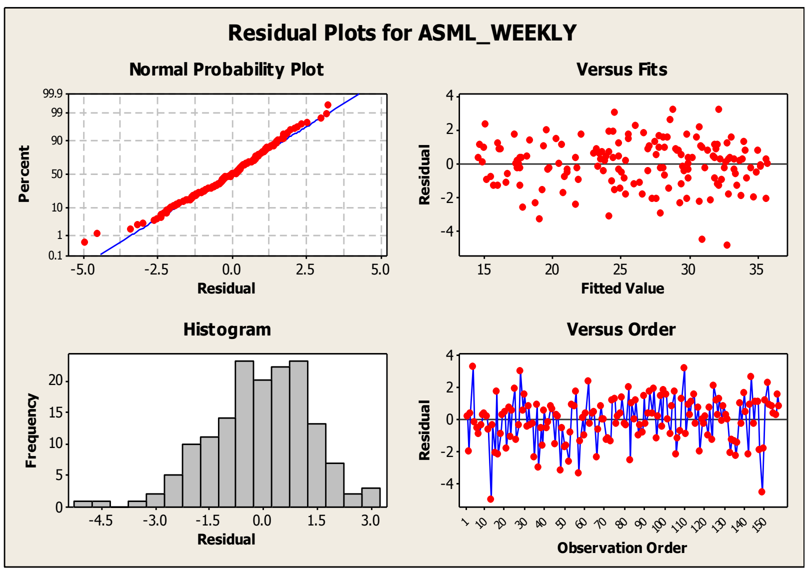

Figure 3.24: Four-in-One Residual Plots for ASML sampled weekly

Table 3.18

Partial Minitab Output for ASML sampled on the 1st trading day of the month (AR(1))

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.8892   0.0851  10.45  0.000
Constant   3.1738   0.5478   5.79  0.000
Mean       28.641    4.944

Number of observations:  38
Residuals:    SS =  365.462 (backforecasts excluded)
              MS =  10.152  DF = 36

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36   48
Chi-Square   14.0   28.3   34.5    *
DF             10     22     34    *
P-Value     0.172  0.167  0.446    *
```

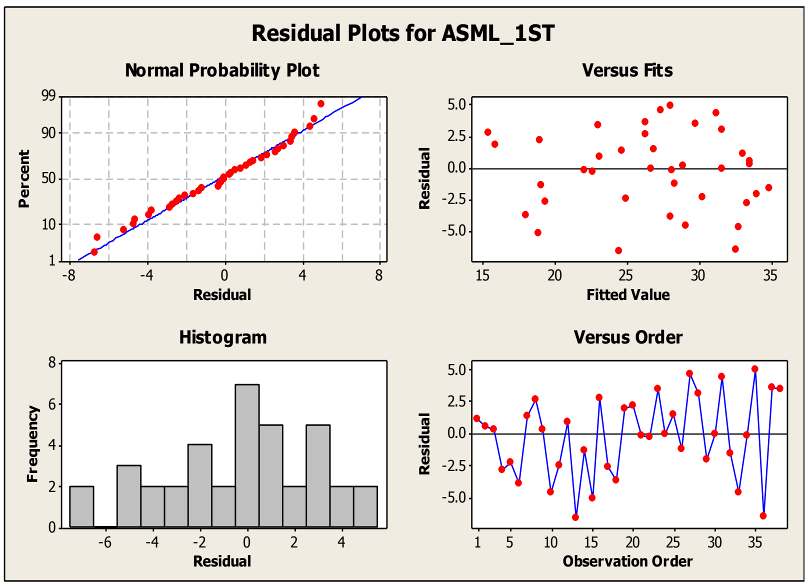

Figure 3.25: Four-in-One Residual Plots for ASML sampled on the 1st trading day of the month

Table 3.19

Partial Minitab Output for ASML sampled on the 15th trading day of the month (AR(1))

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.9158   0.0815  11.24  0.000
Constant   2.3721   0.5400   4.39  0.000
Mean       28.182    6.416

Number of observations:  34
Residuals:    SS =  268.768 (backforecasts excluded)
              MS =  8.399  DF = 32

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24   36   48
Chi-Square    6.2   18.9    *    *
DF             10     22    *    *
P-Value     0.798  0.650    *    *
```

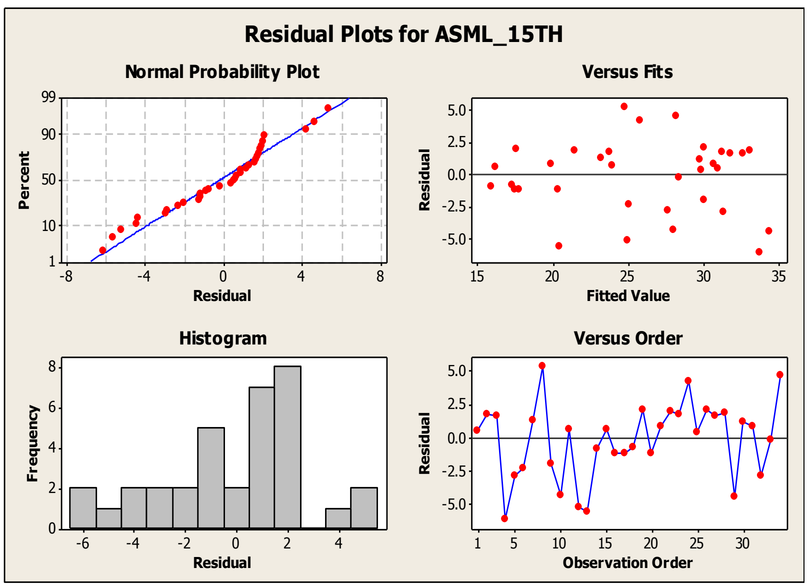

Figure 3.26: Four-in-One Residual Plots for ASML sampled on the 15th trading day of the month

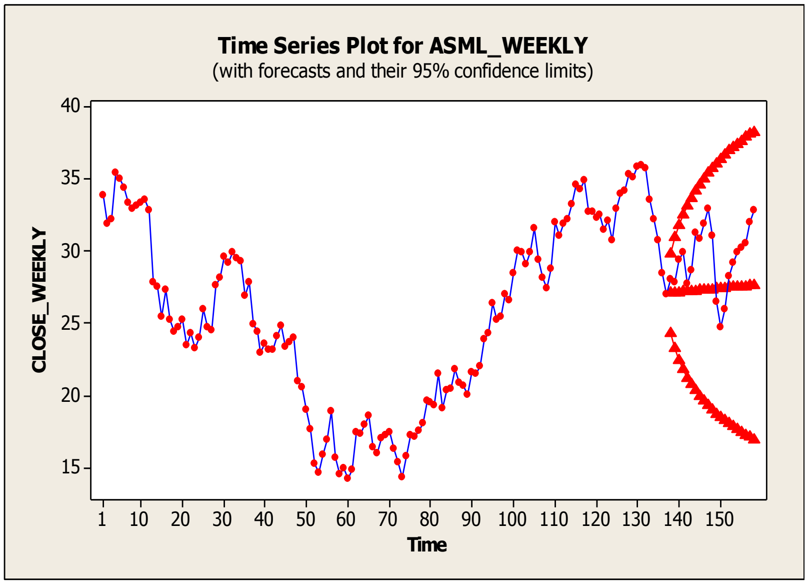

Figure 3.27: Time Series plot of ASML sampled weekly with forecasts

Table 3.20

Analysis of the forecasts for ASML sampled weekly

Date	Actual	Forecasts
6/1/2010	28.10	27.08
6/8/2010	27.82	27.12
6/11/2010	29.44	27.15
6/22/2010	29.92	27.18
6/29/2010	27.75	27.21
7/6/2010	28.68	27.24
7/13/2010	31.27	27.27
7/20/2010	30.93	27.30
7/27/2010	31.94	27.33
8/3/2010	33.01	27.36
8/10/2010	31.06	27.39
8/24/2010	26.50	27.42
8/31/2010	24.73	27.44
9/7/2010	25.98	27.47
9/14/2010	28.26	27.50
9/21/2010	29.21	27.52
9/28/2010	29.99	27.55
10/5/2010	30.31	27.57
10/12/2010	30.57	27.60
10/19/2010	32.05	27.62
10/26/2010	32.82	27.64

The MAPE for the forecasted values is 0.087.The fits continued to steadily increase, while the actual values also increased overall, but not without decreasing at some periods in the forecasting time interval and having a fair amount of variability. During this five month period the overall change in the fits was $ 0.56 or 2% whereas the percent change for the actual values was 15.8%. Although the overall behavior of the forecasts and actual values are both increasing the forecasts in no way reflect what happens during the forecasting period. This information could be vital to an investor because that behavior would indicate if they should sell the stock, in anticipation of a drastic change in price, or continue to stay invested to take advantage of the long-term behavior of the stock.

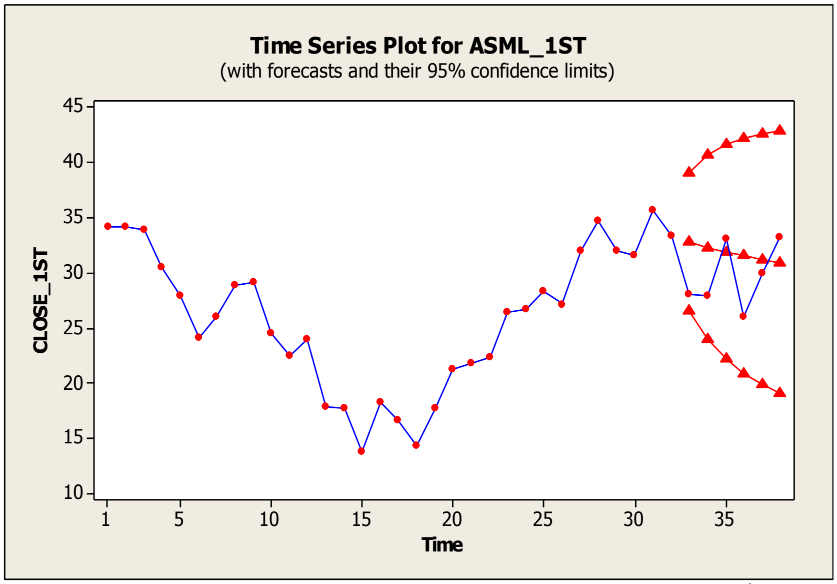

Figure 3.28: Time Series plot of ASML sampled on the 1st trading day of the month with forecasts

Table 3.21

Analysis of the forecasts for ASML sampled on the 1st trading day of the month

Date	Actual	Forecasts
6/1/2010	28.10	32.77
7/1/2010	27.94	32.31
8/2/2010	33.02	31.90
9/1/2010	25.99	31.54
10/1/2010	29.92	31.22
11/1/2010	33.27	30.93

The MAPE for the forecasted values is 0.113917. The forecasts for the weekly data suggested that stock would slowly, but steadily be on the rise for this five month period. However, the forecasts for the data sampled on the 1st and 15th trading day of the month suggest that the prices will decline. So, although we were able to confirm models for the data that were sampled on the 1st and 15th, those models would not be appropriate investing tools.

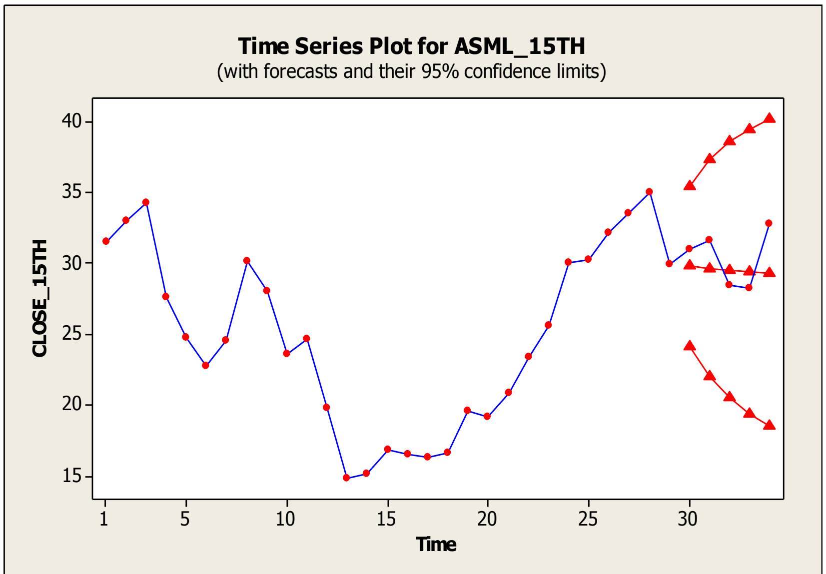

Figure 3.29: Time Series plot of ASML sampled on the 15th trading day of the month with forecasts

Table 3.22

Analysis of the forecasts for ASML sampled on the 15th trading day of the month

Date	Actual	Forecasts
6/15/2010	30.93	29.77
7/15/2010	31.59	29.64
8/16/2010	28.39	29.52
9/15/2010	28.17	29.40
10/15/2010	32.81	29.30

The MAPE for the forecasted values is 0.057914.The Minitab output confirms that an AR(1) is an adequate model for each of the series. The small p-value, essentially zero, for the autoregressive parameter and the constant tell us that these parameters are significantly different from zero. The plots for the analysis of the residuals suggest Normality and randomness. We can conclude that data that is sampled on a weekly basis produces the best model based on the MSE and the analysis of the forecasts.

3.1 d. Advanced Analogue Technologies, Inc.

The resulting models for Advanced Analogue Technologies, Inc. were once again all AR(1). Looking at the ACFs below, we recognize that they behave the same, except for the weekly data containing more lags. Again, the ACF for the differenced series are those of a trended series, so we conclude that differencing is not necessary and the lags are strongly correlated, so the appropriate model is an AR(1).

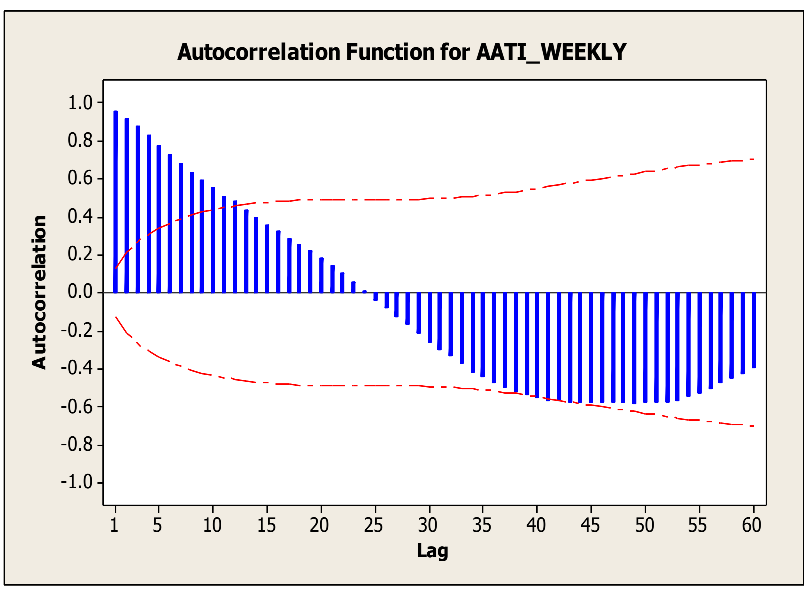

Figure 3.30: Autocorrelation Function for single differenced AATI sampled weekly

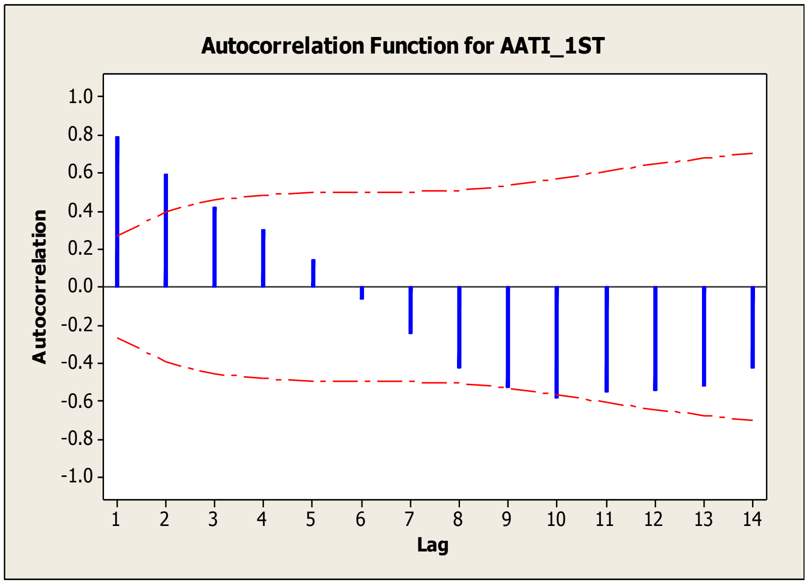

Figure 3.31: Autocorrelation Function for single differenced AATI sampled on the 1st trading day of the month

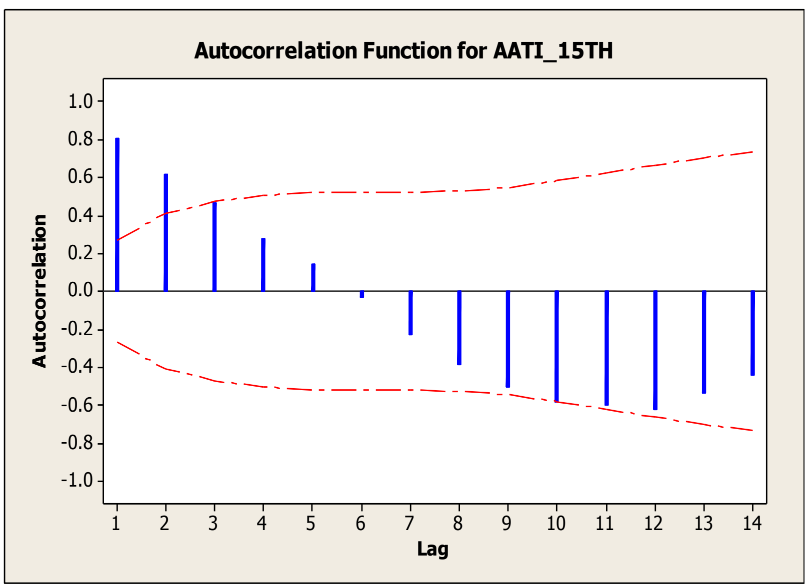

Figure 3.32: Autocorrelation Function for single differenced AATI sampled on the 15th trading day of the month

In the Minitab output below we fit the model without a constant because the results for the model with the constant indicated it is not significantly different from zero, which was reflected by the p-value for this parameters being larger than .05(nearly 1). So we refit the models and produce the following results.

Table 3.23

Partial Minitab Output for AATI sampled weekly

```
Final Estimates of Parameters

Type       Coef  SE Coef       T      P
AR   1   0.9981   0.0042  240.26  0.000

Number of observations:  268
Residuals:     SS =  71.1688 (backforecasts excluded)
               MS =  0.2665  DF = 267

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   14.1   22.6   34.3   52.6
DF             11     23     35     47
P-Value     0.225  0.484  0.500  0.267
```

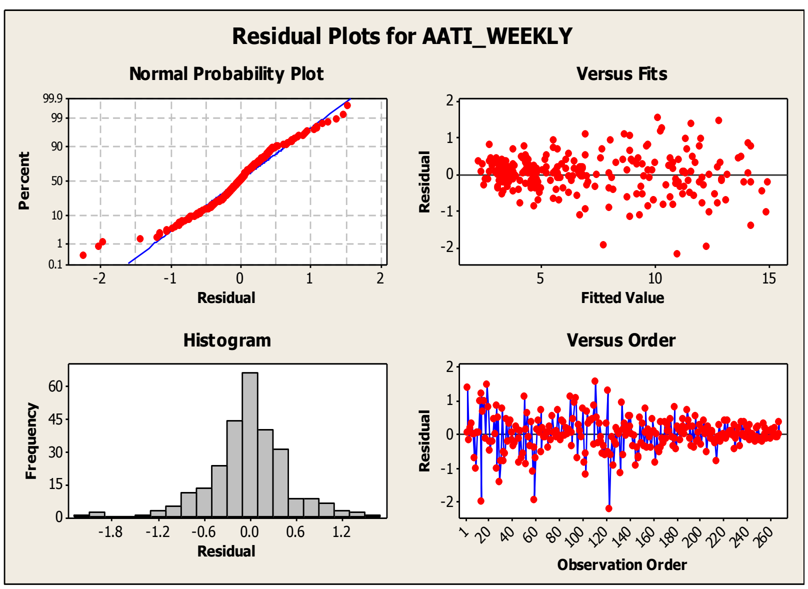

Figure 3.33: Four-in-One Residual Plots for AATI sampled weekly

Table 3.24

Partial Minitab Output for AATI sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type       Coef  SE Coef       T      P
AR    1  0.9978   0.0197   50.61  0.000

Number of observations:  63
Residuals:    SS =  90.2395 (backforecasts excluded)
              MS =  1.4555  DF = 62

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag              12      24      36      48
Chi-Square      9.7    20.3    31.4    33.9
DF               11      23      35      47
P-Value       0.557   0.626   0.645   0.923
```

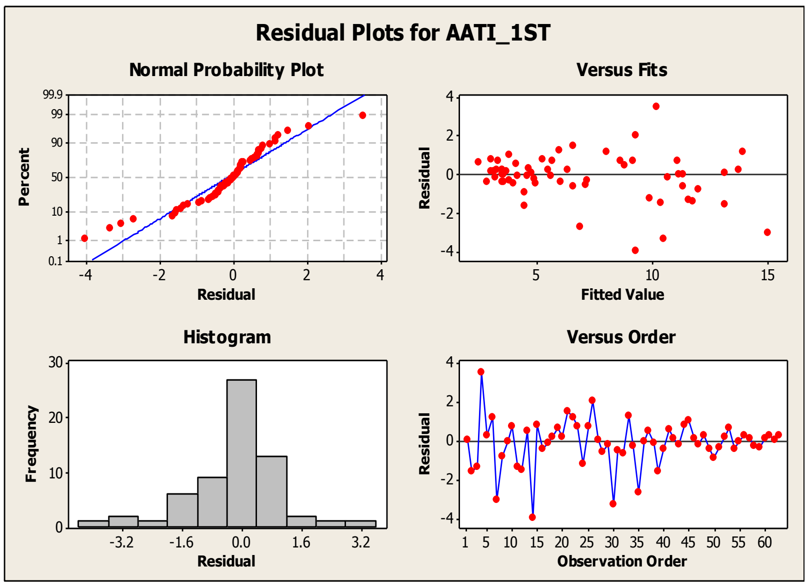

Figure 3.34: Four-in-One Residual Plots for AATI sampled on the 1st trading day of the month

The histogram and probability plots of the residuals (Figure 3.36) again show that there are three low outliers and also one high outlier. This causes skewness. Also, examining the residuals versus order we recognize that the variation is nonconstant. The residuals versus fits are relatively homoscedastic, compared to the order plot.

Table 3.25

Partial Minitab Output for AATI sampled on the 15th trading day of the month

```
Final Estimates of Parameters

Type       Coef  SE Coef       T      P
AR   1   0.9979   0.0193   51.69  0.000

Number of observations:  60
Residuals:     SS =  84.6610 (backforecasts excluded)
               MS =  1.4349  DF = 59

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   13.0   23.9   33.8   36.4
DF             11     23     35     47
P-Value     0.295  0.409  0.525  0.869
```

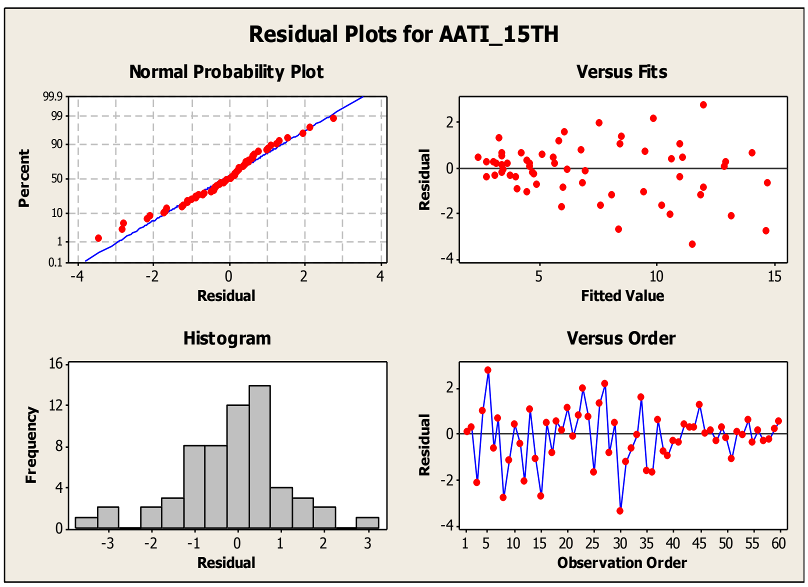

Figure 3.35: Four-in-One Residual Plots for AATI sampled on the 15th trading day of the month

The Minitab output confirms that an AR(1) is an adequate model for each of the series. The small p-value (essentially zero) for the autoregressive parameter tells us that the autoregressive parameter is significantly different from zero. Although the probability plot suggests Normality, there are several low outliers present. Also the plot of the residuals versus the fit does not appear to be totally random or dispersed evenly.

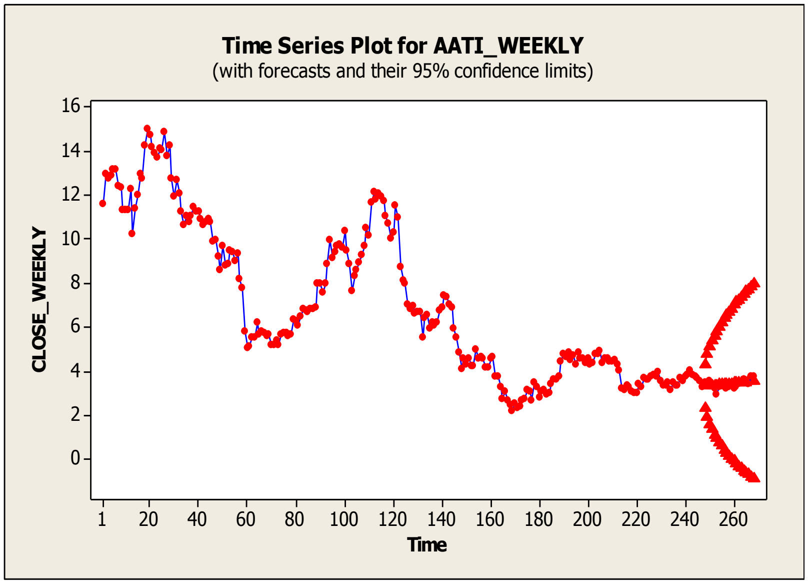

Figure 3.36: Time Series Plot for AATI sampled weekly with forecasts

Examining the time series above, we notice that the behavior is completely different, during different intervals in the series. For the first 1/3 of the series, from observations 1 to about 60, the price per share starts out increasing but then steadily decreases. For the second 1/3, the price gradually increases, but also begins to decrease. There is a steady cluster of observations present amongst the last one-third of the fitted values.

Table 3.26

Analysis of forecasts for AATI sampled weekly (AR(1))

Dates	Actual	Forecasts
6/1/2010	3.46	3.30
6/8/2010	3.34	3.30
6/11/2010	3.55	3.29
6/22/2010	3.37	3.29
6/29/2010	3.21	3.28
7/6/2010	2.97	3.27
7/13/2010	3.27	3.27
7/20/2010	3.42	3.26
7/27/2010	3.38	3.25
8/3/2010	3.24	3.25
8/10/2010	3.31	3.24
8/24/2010	3.37	3.24
8/31/2010	3.24	3.23
9/7/2010	3.30	3.22
9/14/2010	3.42	3.22
9/21/2010	3.46	3.21
9/28/2010	3.62	3.21
10/5/2010	3.53	3.20
10/12/2010	3.43	3.19
10/19/2010	3.78	3.19
10/26/2010	3.75	3.18

The MAPE for the forecasted values is 0.056.

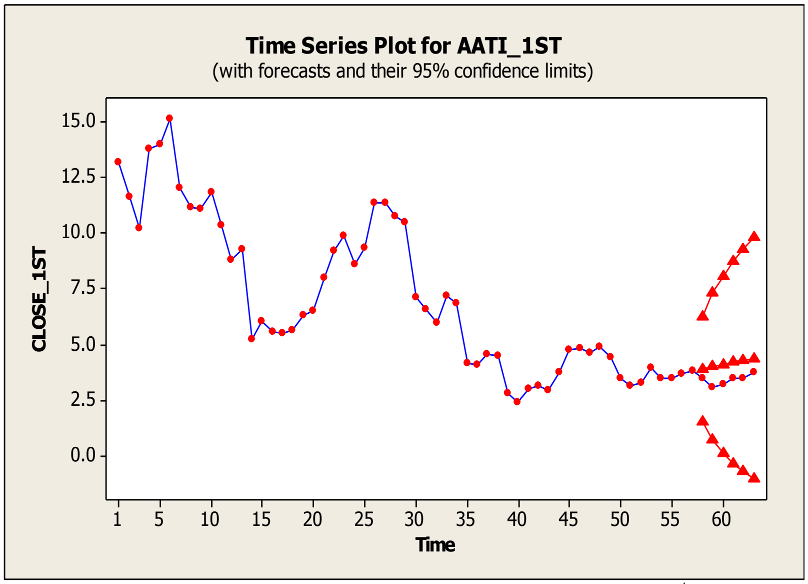

Figure 3.37: Time Series for AATI sampled on the 1st trading day of the month with forecasts

Table 3.27

Analysis of forecasts for AATI sampled on the 1st trading day of the month (AR(1))

Dates	Actual	Forecasts
6/1/2010	3.4600	3.45
7/1/2010	3.0900	3.45
8/2/2010	3.2200	3.44
9/1/2010	3.4700	3.43
10/1/2010	3.4800	3.42
11/1/2010	3.7200	3.42

The MAPE for the forecasted values is 0.049.

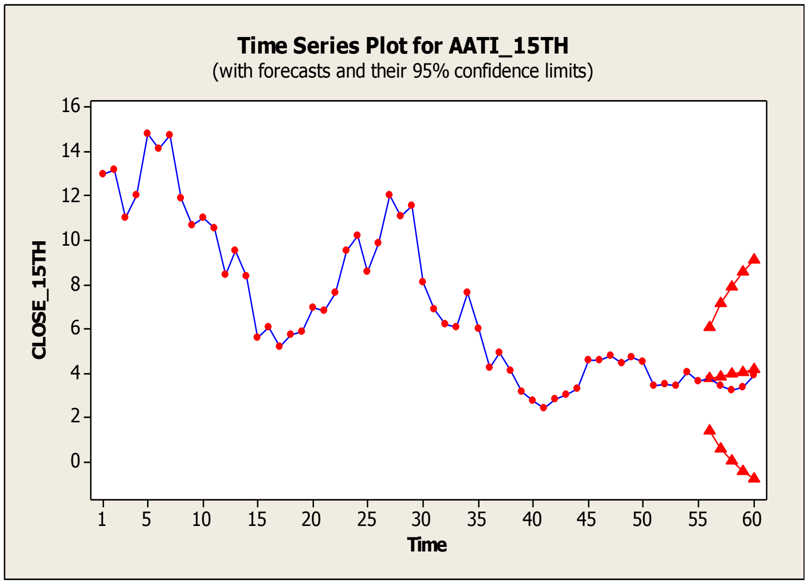

Figure 3.38: Time Series for AATI sampled on the 15th trading day of the month with forecasts

Table 3.28

Analysis of forecasts for AATI sampled on the 1st trading day of the month (AR(1))

Dates	Actual	Forecasts
6/15/2010	3.78	3.62
7/15/2010	3.43	3.62
8/16/2010	3.20	3.61
9/15/2010	3.39	3.60
10/15/2010	3.91	3.59

The MAPE for the forecasted values is 0.073.The models for the forecast for the series were, $y_t = .9981 y_{t-1} + \varepsilon_t$, $y_t = .9978 y_{t-1} + \varepsilon_t$, and $y_t = .9979 y_{t-1} + \varepsilon_t$, for the data sampled weekly, on the 1st trading day of the month and the 15th trading day of the month , respectively. The model for each series is very similar to one that would be produced for a random walk. A random walk model is of the form $y_t = y_{t-1} + \varepsilon_t$. So comparing this to an AR(1) model or $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$, for a random walk $\phi_0 = 0$ and $\phi_1 = 1$. And for the confirmed models the autoregressive parameters were .9981, .997 and .9979. These models indicate the next weeks or month's price is just this week's or month's price plus some random error. However, although the Minitab output confirms that the parameters are significantly different from zero, an analysis of the residuals the data sampled on the first and 15th of the month indicated that the data is not well modeled. Even though we realize that AR(1) does not produce an adequate model, this is the most appropriate model when compared to the results of any other ARIMA model for these series.

3.1 e. PepsiCo Inc.

The resulting models for PepsiCo. were all again AR(1). Looking at the ACF's below, we recognize that they behave the same, except for the weekly data containing more lags. The suggested model is AR(1) because the ACF for the differenced series was that of a trended series. From previous series, we know this implies that differencing isn't necessary and that the appropriate AR(1) model.

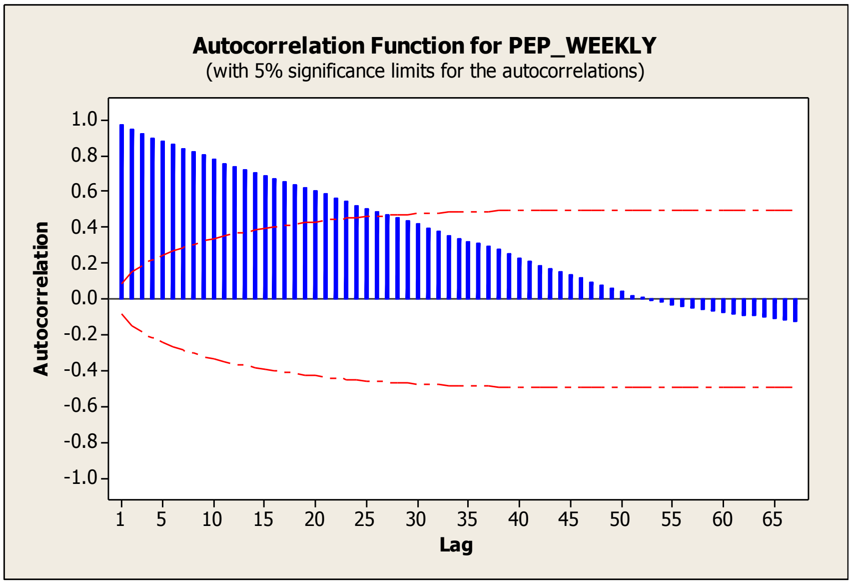

Figure 3.39: Autocorrelation Function for single differenced PEP sampled weekly

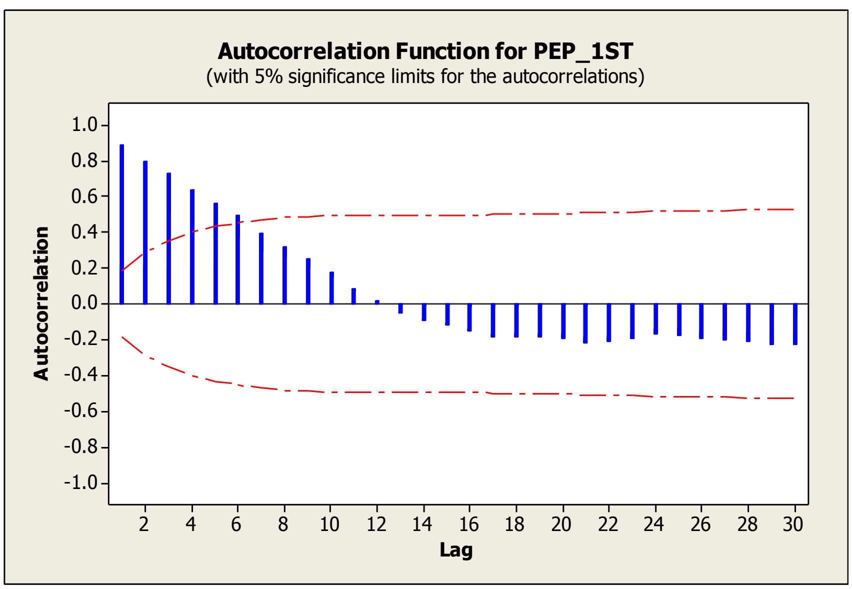

Figure 3.40: Autocorrelation Function for single differenced PEP sampled on the 1st trading day of the month

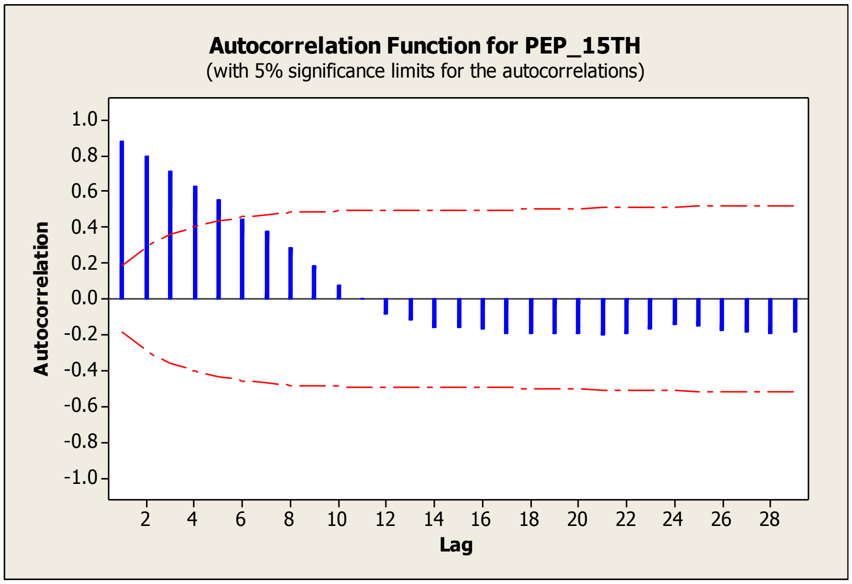

Figure 3.41: Autocorrelation Function for single differenced PEP sampled on the 15th trading day of the month

Table 3.29

Partial Minitab Output for PEP sampled weekly

```
Final Estimates of Parameters

Type          Coef  SE Coef       T      P
AR   1      0.9923   0.0067  147.63  0.000
Constant   0.42796  0.06253    6.84  0.000
Mean        55.920    8.170

Number of observations:  535
Residuals:    SS =  1102.73 (backforecasts excluded)
              MS =  2.07  DF = 533

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   15.7   26.0   42.7   55.4
DF             10     22     34     46
P-Value     0.108  0.250  0.146  0.161
```

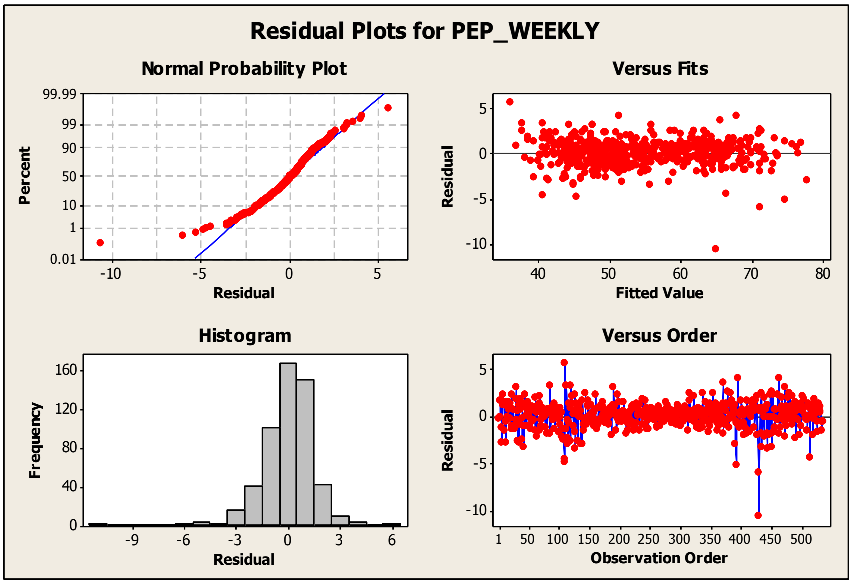

Figure 3.42: Four-in-One Residual Plots for PEP sampled weekly

Table 3.30

Partial Minitab Output for PEP sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.9666   0.0274  35.26  0.000
Constant   1.8214   0.2571   7.09  0.000
Mean       54.525    7.695

Number of observations:  126
Residuals:    SS =  1030.65 (backforecasts excluded)
              MS =  8.31  DF = 124

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12      24      36      48
Chi-Square   10.6    29.6    42.8    48.0
DF             10      22      34      46
P-Value     0.391   0.129   0.142   0.394
```

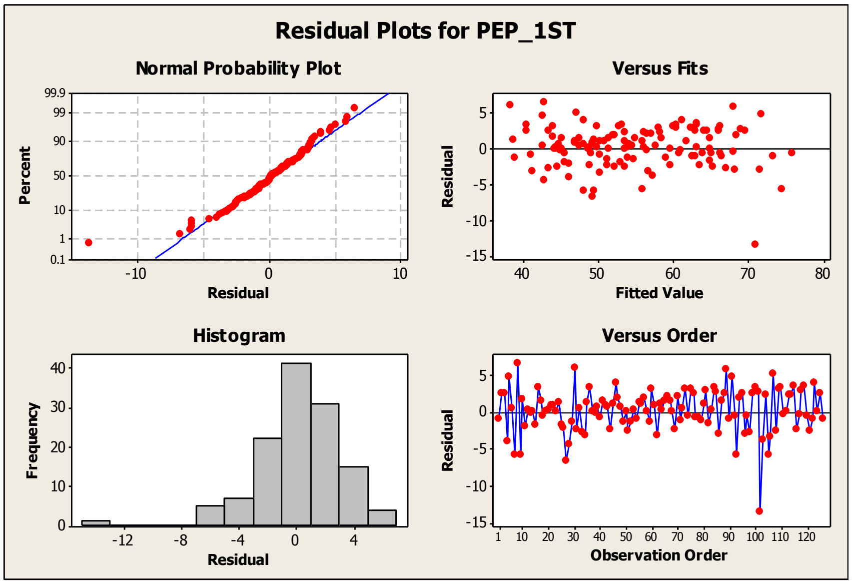

Figure 3.43: Four-in-One Residual Plots for PEP sampled on the 1^{st} trading day of the month

Table 3.31

Partial Minitab Output for PEP sampled on the 15th trading day of the month

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.9595   0.0293  32.77  0.000
Constant    2.2305   0.2707   8.24  0.000
Mean        55.077    6.684

Number of observations:  122
Residuals:     SS =  1071.83 (backforecasts excluded)
               MS =  8.93  DF = 120

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12      24      36      48
Chi-Square   15.3    24.6    29.9    37.9
DF             10      22      34      46
P-Value     0.123   0.317   0.669   0.798
```

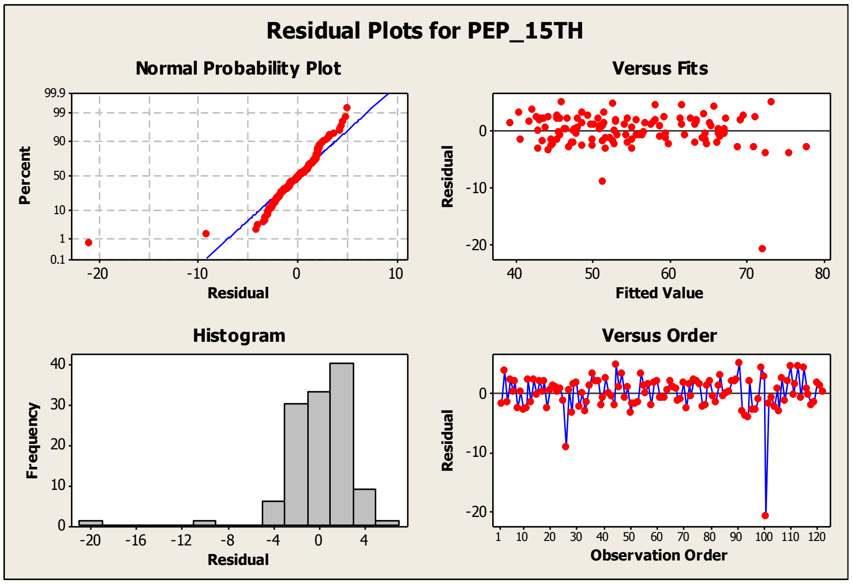

Figure 3.44: Four-in-One Residual Plots for PEP sampled on the 15th trading day of the month

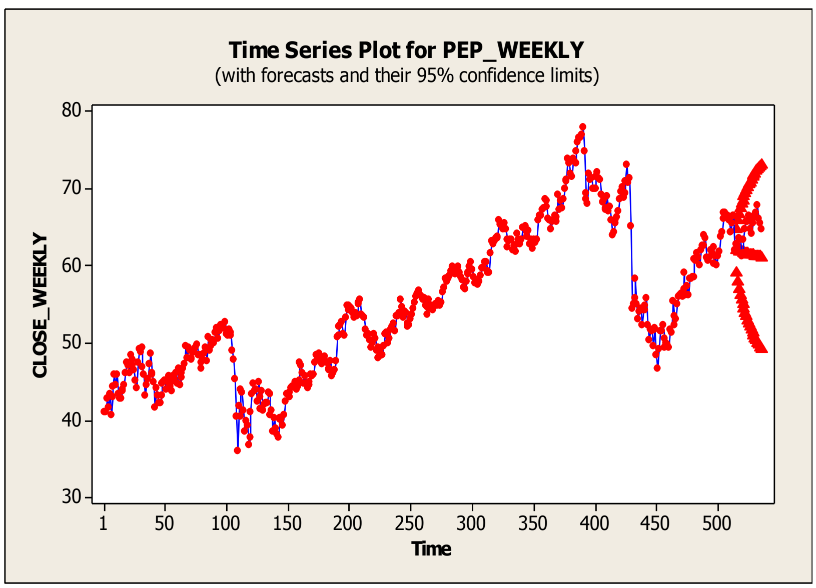

Figure 3.45: Time Series Plot for PEP sampled weekly with forecasts

The plots in the residual analysis reveal that there are at least two low outliers. From examining the time series plot above we notice that these are caused by the drop in price that occurred from 9/30/2008 to 10/7/2008, where the price per share dropped from $71.27 to $65.12 and then to $54.40 on the 10/14/2008. By doing some additional research these drops can be explained by events that occurred within the company. In October 2008, Pepsi announced that it would be redesigning its logo and re-branding many of its products by early 2009. Also, in late 2008, Pepsi overhauled their entire brand, simultaneously introducing a new logo and a minimalist label design. The redesign was comparable to Coca-Cola's earlier simplification of their can and bottle designs. Also in 2008 Pepsi teamed up with Google/YouTube to produce the first daily entertainment show on Youtube, Poptub. This daily show deals with pop culture, internet viral videos, and celebrity gossip. Poptub is updated daily from Pepsi. Because it is most likely that these events and the anticipation of them caused the price per share to drop drastically and then continue to decline, and more importantly this type of overhaul of a company is not something that occurs regularly, we explain the outliers and can justify eliminating all observations the observations prior to this point and only using the values after 10/14/2008 to model this series. However doing this does significantly shrink our data set.

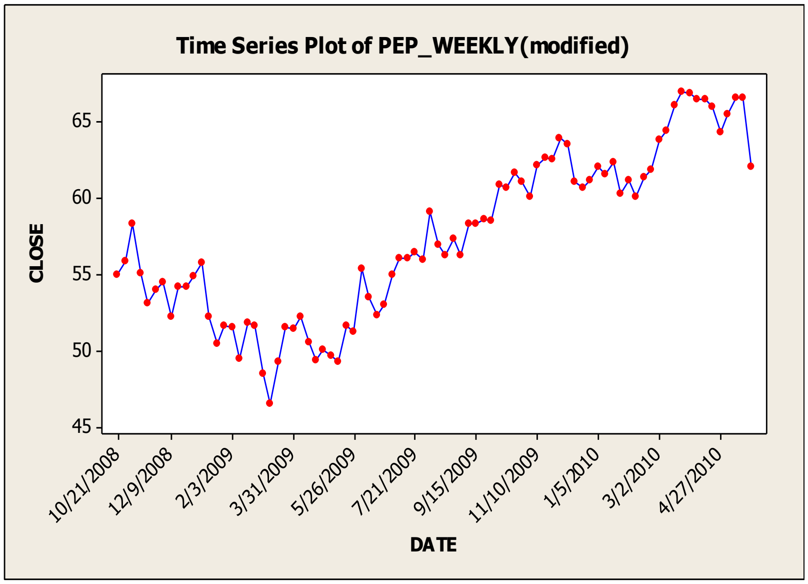

Figure 3.46: Time Series Plot for PEP sampled weekly (modified)

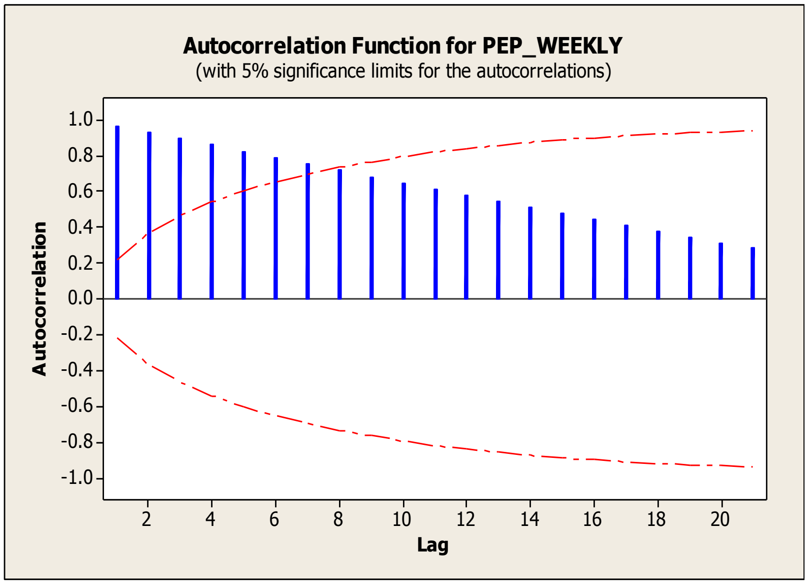

Figure 3.47: Autocorrelation Function for PEP sampled weekly differenced once(modified)

Table 3.32

Partial Minitab Output for PEP sampled weekly (modified)

```
Final Estimates of Parameters

Type          Coef   SE Coef      T      P
AR   1      0.9732    0.0256  38.00  0.000
Constant    1.5886    0.1459  10.89  0.000
Mean        59.368     5.451

Number of observations:  105
Residuals:    SS =  227.692 (backforecasts excluded)
              MS =  2.211   DF = 103

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12      24      36      48
Chi-Square   13.4    18.2    26.2    37.4
DF             10      22      34      46
P-Value     0.204   0.694   0.830   0.813
```

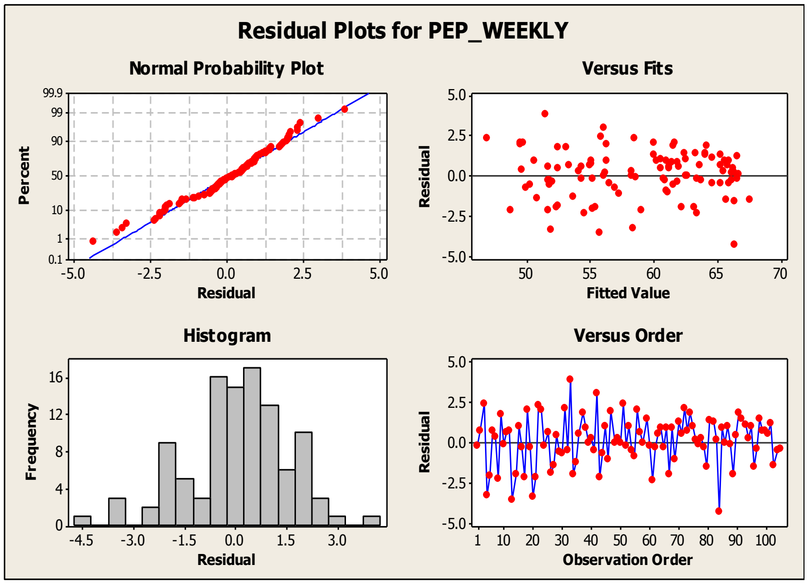

Figure 3.48: Four-in-One Residual Plots for PEP sampled weekly(modified)

Although there are still outliers present the residuals are appear both Normal and random. The confirmed model is still AR(1) but using the modified data we did not produce a better model, based on the MSE. For the original series it was 2.07 and for the modified series it is 2.211. But the MAPE for the forecasted values for the modified series was significantly better, 0.003309 compared to 0.04979. Because we are modeling these series in order to forecast stock prices, using the original series would be best, assuming that the assumptions for Normality and randomness are met. The same results were true for the data sampled on the 1st and 15th trading day of the month. The MAPE for the forecasts for the 1st trading day of the month was 0.037581 for the modified series and for the original series it was 0.03186.

Table 3.33

Analysis of forecasts for PEP sampled weekly (AR(1)) for the modified series

Data	Actual	Forecasts
6/1/2010	62.76	61.92
6/8/2010	62.65	61.85
6/11/2010	63.56	61.79
6/22/2010	63.30	61.72
6/29/2010	61.23	61.66
7/6/2010	61.64	61.60
7/13/2010	63.43	61.54
7/20/2010	64.73	61.48
7/27/2010	65.69	61.42
8/3/2010	65.77	61.37
8/10/2010	66.53	61.31
8/24/2010	64.78	61.26
8/31/2010	64.18	61.21
9/7/2010	65.48	61.16
9/14/2010	65.98	61.11
9/21/2010	66.46	61.07
9/28/2010	66.78	61.02
10/5/2010	67.76	60.98
10/12/2010	66.08	60.93
10/19/2010	65.41	60.89
10/26/2010	64.7900	60.85

The MAPE for the forecasted values is 0.003309.

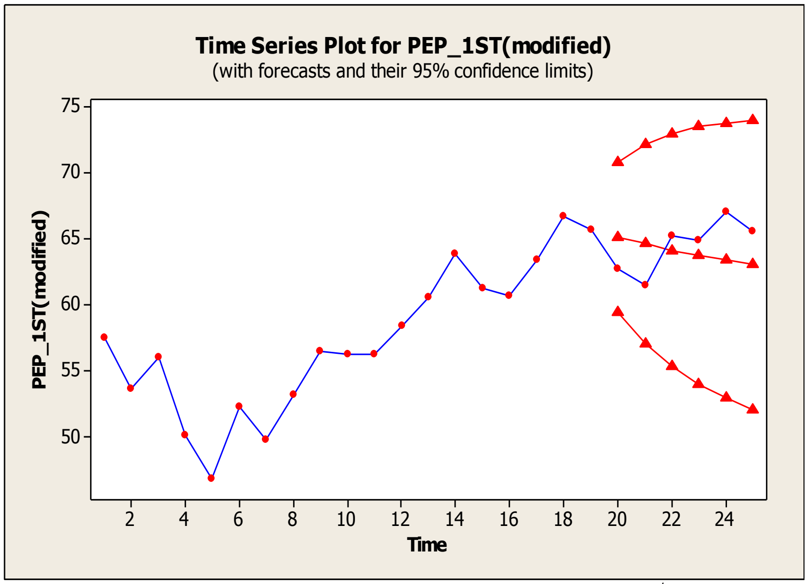

Figure 3.49: Time Series for PEP sampled on the 1st trading day of the month with forecasts (modified)

Table 3.34

Analysis of forecasts for PEP sampled on the 1st trading day of the month (AR(1))

Dates	Actual	Forecasts
6/1/2010	62.76	65.12
7/1/2010	61.52	64.61
8/2/2010	65.27	64.14
9/1/2010	64.89	63.73
10/1/2010	67.00	63.36
11/1/2010	65.55	63.02

The MAPE for the forecasted values is 0.037581.

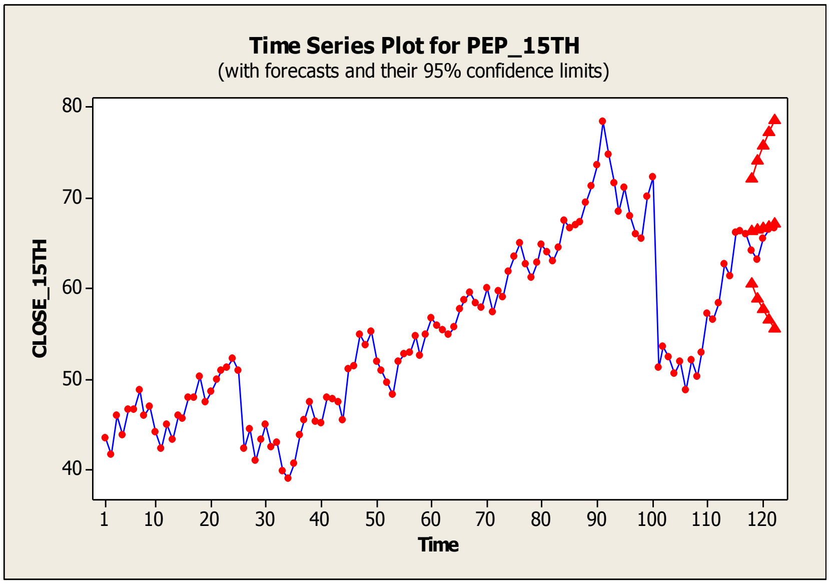

Figure 3.50: Time Series for PEP sampled on the 15th trading day of the month with forecasts

Table 3.35

Analysis of forecasts for PEP sampled on the 15th trading day of the month (AR(1))

Dates	Actual	Forecasts
6/15/2010	64.24	66.33
7/15/2010	63.16	66.52
8/16/2010	65.43	66.71
9/15/2010	66.50	66.91
10/15/2010	66.68	67.10

The MAPE for the forecasted values is 0.02354. The fits for all the models except the data that was sampled on the 15th trading day of the month showed a steady decline, although the actual prices increased overall. This is most likely do to the major drop in price that occurred in October 2008. This event did not affect the data sampled on the 15th trading day of the month as much, although the overall behavior is still very similar. The MSE tells us that the data sampled weekly produces a better model than the data sampled on the 1st and 15th trading day of the month. As an investing tool the data sampled on the 15th trading day of the month would be the most appropriate, because the fits behave most like the actual values and it has the smallest MAPE.

3.1 f. Coca-Cola Bottling Co. Consolidated

The resulting models for Coca-Cola were again all AR(1). For all three intervals, even after the series were differenced, the ACFs still displayed trend. So this implies that AR(1) is the appropriate model. We also note that although ARIMA (0,1,0) or an random walk would have also produced an adequate model, the resulting residuals were not Normal.

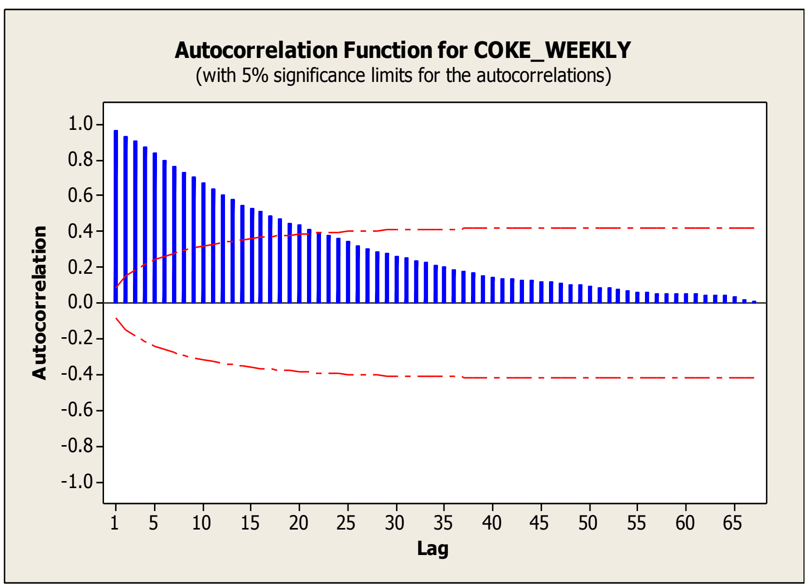

Figure 3.51: Autocorrelation Function for single differenced COKE sampled weekly

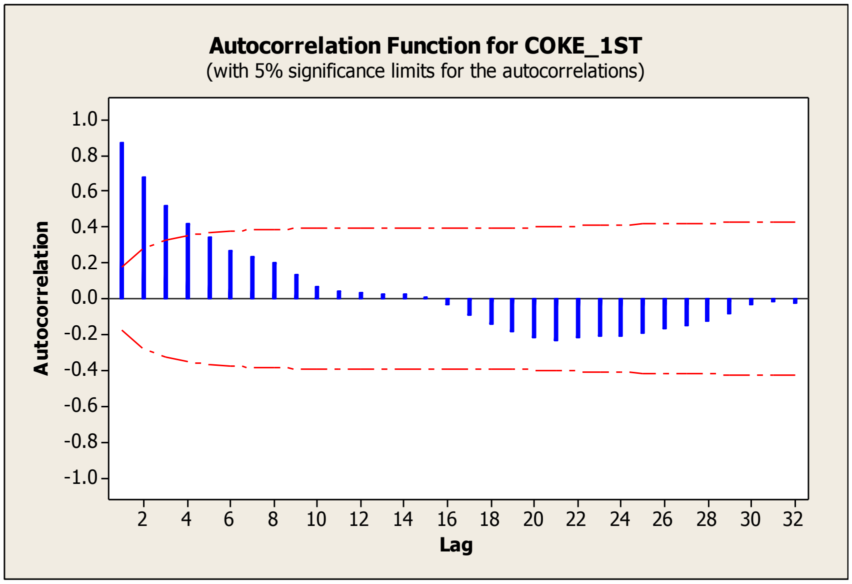

Figure 3.52: Autocorrelation Function for single differenced COKE sampled on the 1st trading day of the month

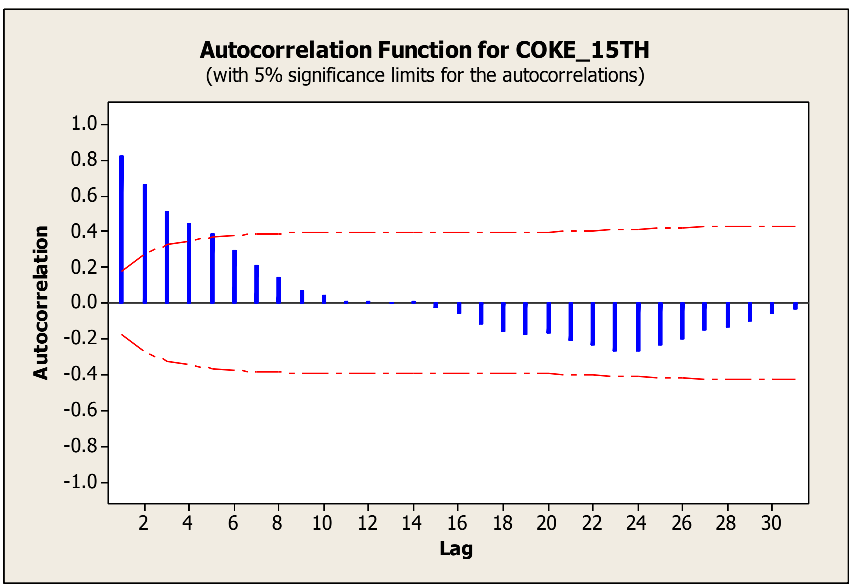

Figure 3.53: Autocorrelation Function for single differenced COKE sampled on the 15th trading day of the month

Table 3.54

Partial Minitab Output for COKE sampled weekly

```
Final Estimates of Parameters

Type           Coef   SE Coef       T       P
AR    1      0.9646    0.0115   84.05   0.000
Constant    1.78021   0.08461   21.04   0.000
Mean         50.273     2.389

Number of observations:  535
Residuals:     SS =  2040.00 (backforecasts excluded)
               MS =  3.83  DF = 533

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12      24      36      48
Chi-Square    17.1    29.2    39.9    47.7
DF              10      22      34      46
P-Value      0.073   0.139   0.224   0.404
```

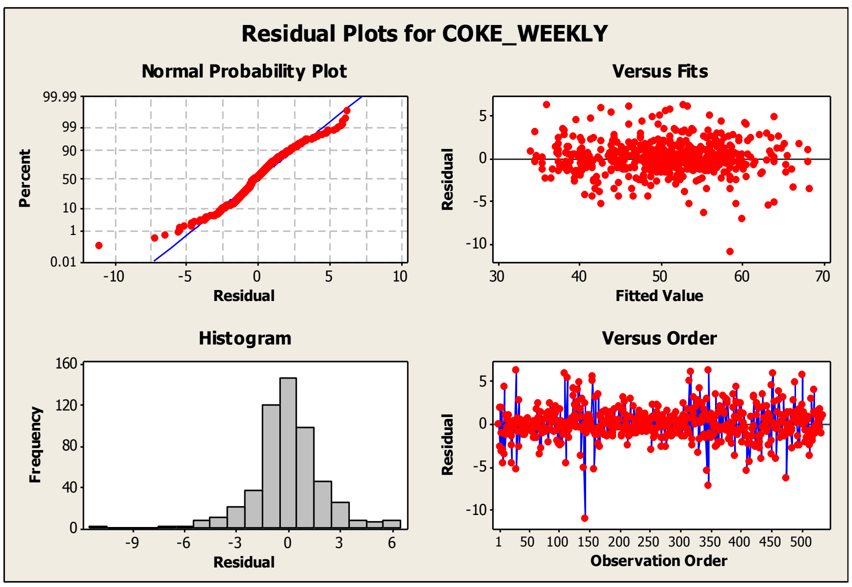

Figure 3.54: Four-in-One Residual Plots for COKE sampled weekly

The probability plot of the residuals shows that there is a low outlier present. This outlier is due to the drop in price on 12/18/2007, when the price per share dropped from $58.91 to $54.91. However there was nothing that could be found to explain this drop in price.

Table 3.37

Partial Minitab Output for COKE sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.8861   0.0418  21.19  0.000
Constant    5.7208   0.3060  18.69  0.000
Mean        50.227    2.687

Number of observations:  126
Residuals:     SS =  1463.07 (backforecasts excluded)
               MS =  11.80  DF = 124

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag              12      24      36      48
Chi-Square     17.8    25.6    32.7    36.0
DF               10      22      34      46
P-Value       0.059   0.269   0.531   0.855
```

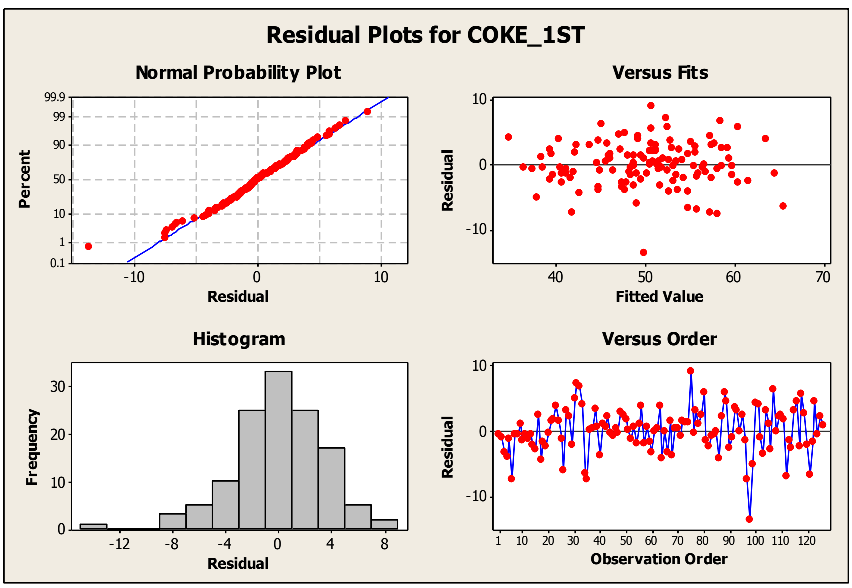

Figure 3.55: Four-in-One Residual Plots for COKE sampled on the 1st trading day of the month

Again the Normal plot of the residuals shows that there is a low outlier present. This outlier is due to the drop in price on 5/1/2008, when the price per share dropped from \$58.20 to \$49.83, but there wasn't anything that could be found to explain this drop in price.

Table 3.38

Partial Minitab Output for COKE sampled on the 15th trading day of the month

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.8462   0.0488  17.33  0.000
Constant    7.7007   0.3664  21.02  0.000
Mean        50.086    2.383

Number of observations:  122
Residuals:     SS =  1965.03 (backforecasts excluded)
               MS =  16.38  DF = 120

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square     8.8   18.6   25.5   31.7
DF              10     22     34     46
P-Value      0.546  0.671  0.854   0.94
```

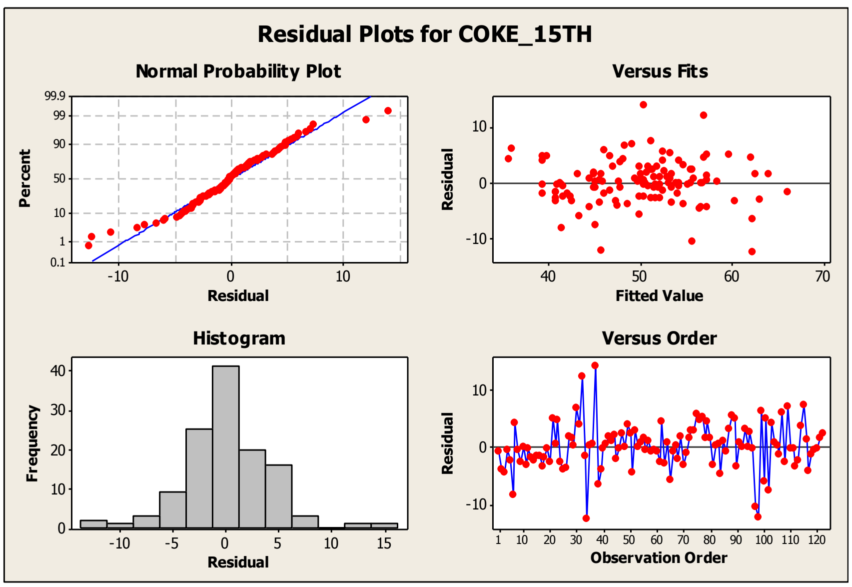

Figure 3.56: Four-in-One Residual Plots for COKE sampled on the 15th trading day of the month

For this data there are three low outliers and two high outliers. A possible cause of the low outliers could be a result of the major drop in price that occurred from 5/15/2008 to 7/15/2008. During this period the price per share dropped from 56.66 to 45.03 then to 33.47. Again we are not able to offer a possible cause.

The Minitab output confirms that AR(1) is the best model for each of the series. The small p-value (less than .05) for the autoregressive parameter and the constant tell us that both parameters are significantly different from zero. Also the Normality Probability plot and the histogram of the residuals confirm approximate Normality. For this stock the weekly data's MSE was smaller than the one for data that was sampled on the 1st and the 15th trading day of the month, so we can conclude that weekly data will produce the best model.

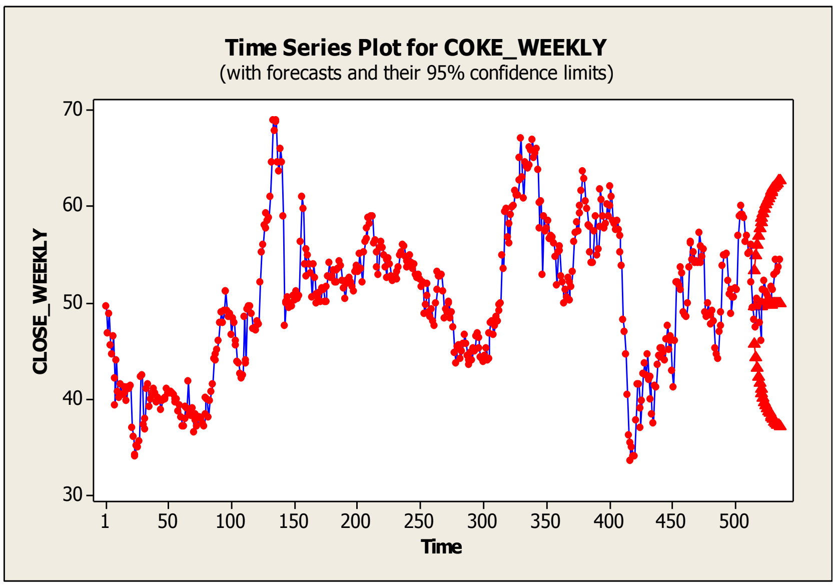

Figure 3.57: Time Series Plot for COKE sampled weekly with forecasts

Table 3.39

Analysis of Forecasts for COKE sampled weekly (AR(1))

Dates	Actual	Forecasts
6/1/2010	48.13	49.49
6/8/2010	47.40	49.52
6/11/2010	50.32	49.55
6/22/2010	49.55	49.57
6/29/2010	47.84	49.60
7/6/2010	45.92	49.62
7/13/2010	50.07	49.65
7/20/2010	51.34	49.67
7/27/2010	52.26	49.69
8/3/2010	51.14	49.71
8/10/2010	50.84	49.73
8/24/2010	50.44	49.75
8/31/2010	49.73	49.77
9/7/2010	50.08	49.79
9/14/2010	51.63	49.80
9/21/2010	51.23	49.82
9/28/2010	52.79	49.84
10/5/2010	54.33	49.85
10/12/2010	53.12	49.87
10/19/2010	53.60	49.88
10/26/2010	54.46	49.89

The MAPE for the forecasts is 0.03726.

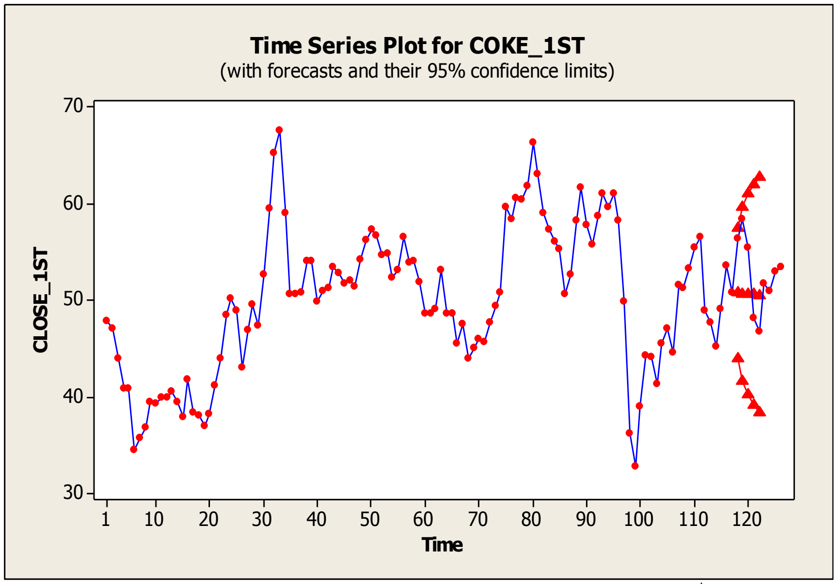

Figure 3.58: Time Series for COKE sampled on the 1st trading day of the month with forecasts

Table 3.40

Analysis of Forecasts for COKE sampled on the1st trading day of the month (AR(1))

Dates	Actual	Forecasts
6/1/2010	48.13	54.79
7/1/2010	46.72	54.27
8/2/2010	51.68	53.81
9/1/2010	50.93	53.40
10/1/2010	52.94	53.042
11/1/2010	53.44	52.72

The MAPE for the forecasts is 0.06755.

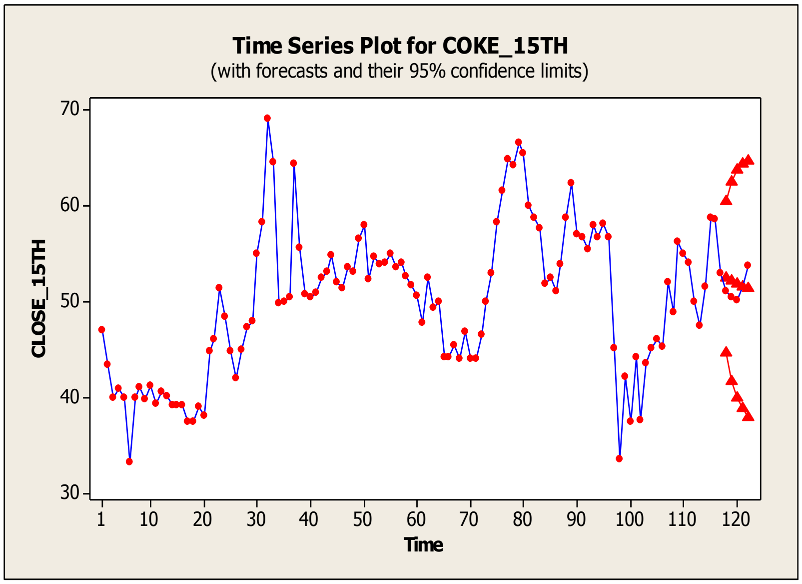

Figure 3.59: Time Series for COKE sampled on the 15th trading day of the month with forecasts

Table 3.41

Analysis of Forecasts for COKE sampled on the1st trading day of the month (AR(1))

Dates	Actual	Forecasts
6/15/2010	51.09	52.48
7/15/2010	50.40	52.12
8/16/2010	50.11	51.80
9/15/2010	51.52	51.54
10/15/2010	53.70	51.32

The MAPE for the forecasts is 0.02798. For all three of the intervals the forecast overall behaved similar to the actual values. Although the data sampled weekly produced the best MSE, the MAPE indicates the forecasts were better for the data sampled on the 15^{th} trading day of the month.

3.1 g. Microsoft Corporation

Again the ACFs for the differenced series displayed trend so the appropriate model is AR(1) for all three intervals. The Minitab output as well as the plots for the residuals analysis confirm that AR(1) is the appropriate model and Normality and randomness of the residuals are satisfied.

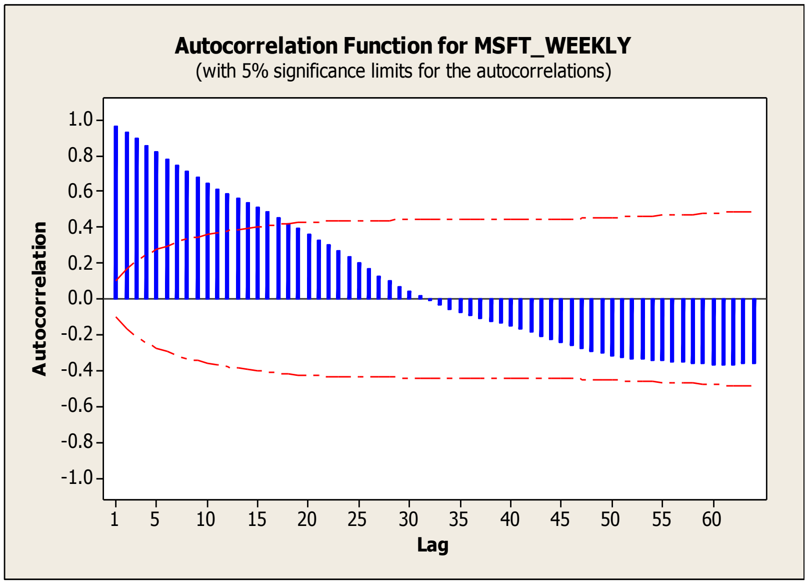

Figure 3.60: Autocorrelation Function for single differenced MSFT sampled weekly

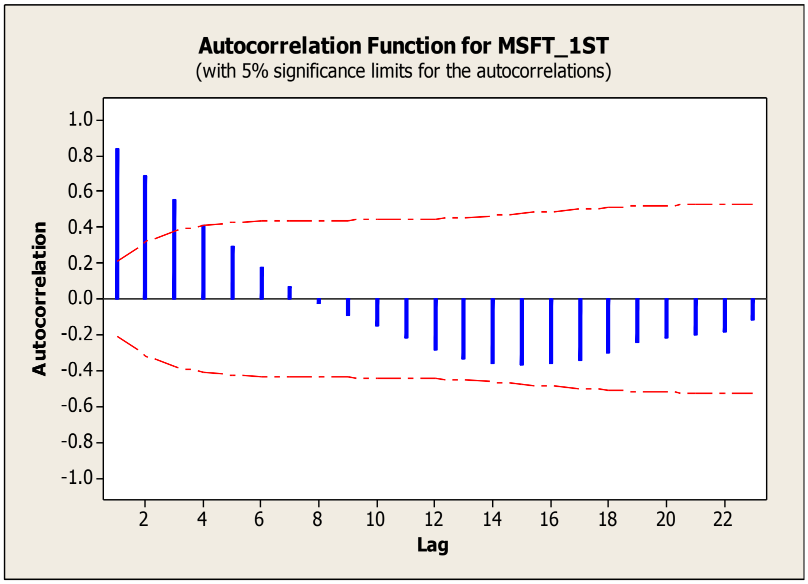

Figure 3.61: Autocorrelation Function for single differenced MSFT sampled on the 1st trading day of the month

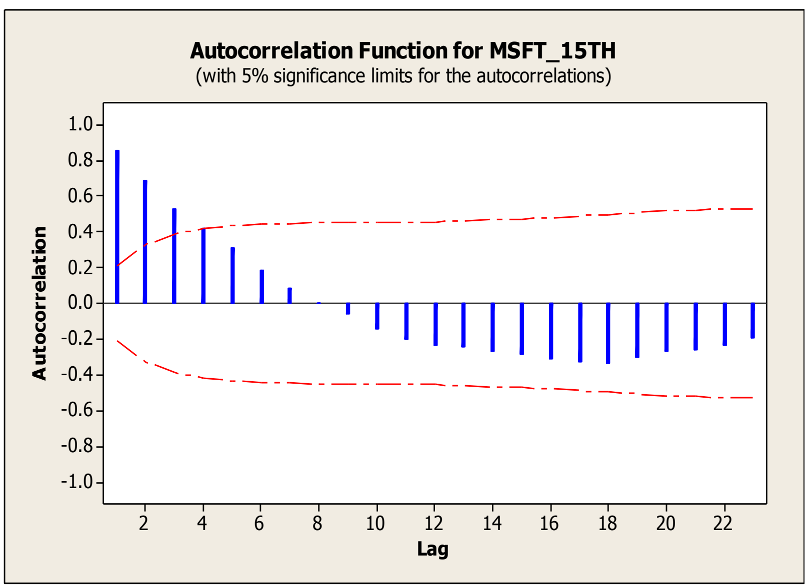

Figure 3.62: Autocorrelation Function for single differenced MSFT sampled on the 15th trading day of the month

Table 3.42

Partial Minitab Output for MSFT sampled weekly

```
Final Estimates of Parameters

Type          Coef   SE Coef      T      P
AR   1      0.9625    0.0137  70.24  0.000
Constant   0.98941   0.04362  22.68  0.000
Mean        26.412     1.164

Number of observations:  397
Residuals:     SS =  297.893 (backforecasts excluded)
               MS =  0.754  DF = 395

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square     8.6   16.7   37.1   50.2
DF              10     22     34     46
P-Value      0.572  0.778  0.327  0.310
```

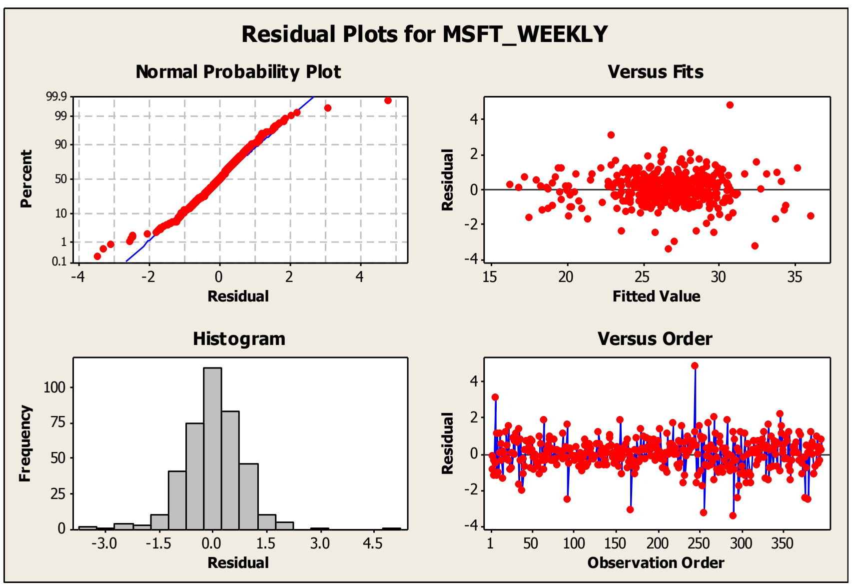

Figure 3.63: Four-in-One Residual Plots for MSFT sampled weekly

The probability plot and histogram reveal that there are low and high outliers for the data sampled weekly and high ones for the data sampled on the 1st and 15th trading day of the month. The time series plot does not show any major drops or increases greater than a few dollars which would lead us to offer an explanation for the outliers.

Table 3.43

Partial Minitab Output for MSFT sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type          Coef   SE Coef       T       P
AR   1      0.8456    0.0562   15.05   0.000
Constant    4.0799    0.1864   21.89   0.000
Mean        26.422     1.207

Number of observations:  93
Residuals:     SS =  293.914 (backforecasts excluded)
               MS =  3.230   DF = 91

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12      24      36      48
Chi-Square     2.8    13.6    25.5    32.2
DF              10      22      34      46
P-Value      0.985   0.914   0.854   0.939
```

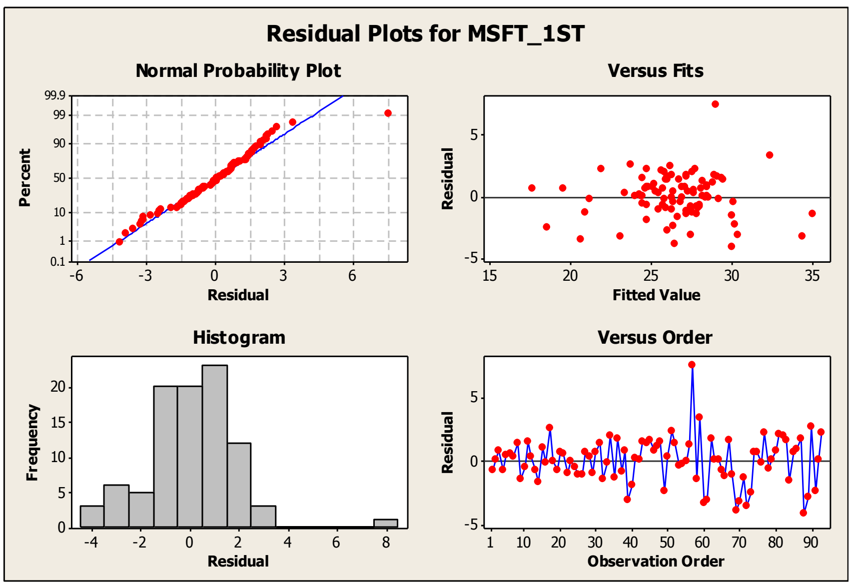

Figure 3.64: Four-in-One Residual Plots for MSFT sampled on the 1st trading day of the month

Table 3.44

Partial Minitab Output for MSFT sampled on the 15th trading day of the month

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.8644   0.0541  15.99  0.000
Constant    3.5569   0.1750  20.33  0.000
Mean        26.222    1.290

Number of observations:  90
Residuals:    SS =  241.354 (backforecasts excluded)
              MS =  2.743  DF = 88

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square     7.6   12.2   19.8   27.2
DF              10     22     34     46
P-Value      0.672  0.953  0.975  0.987
```

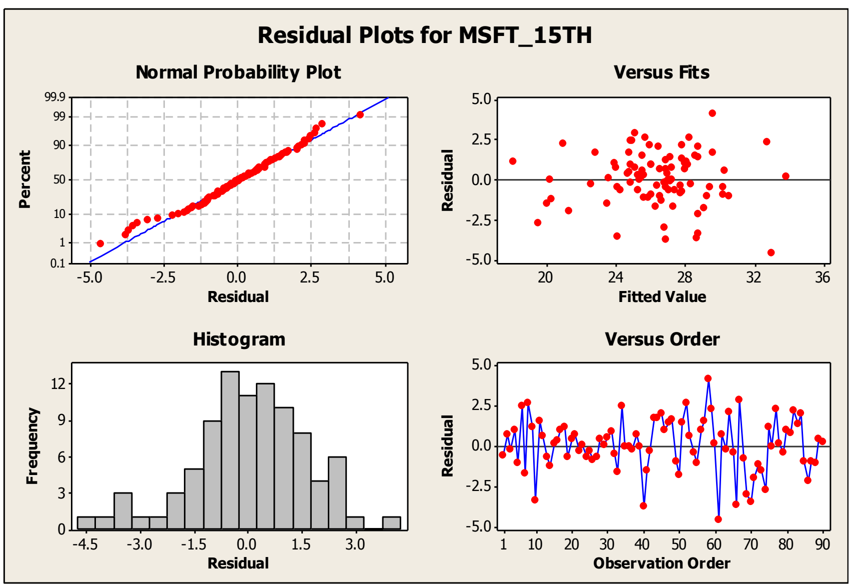

Figure 3.65: Four-in-One Residual Plots for MSFT sampled on the 15th trading day of the month

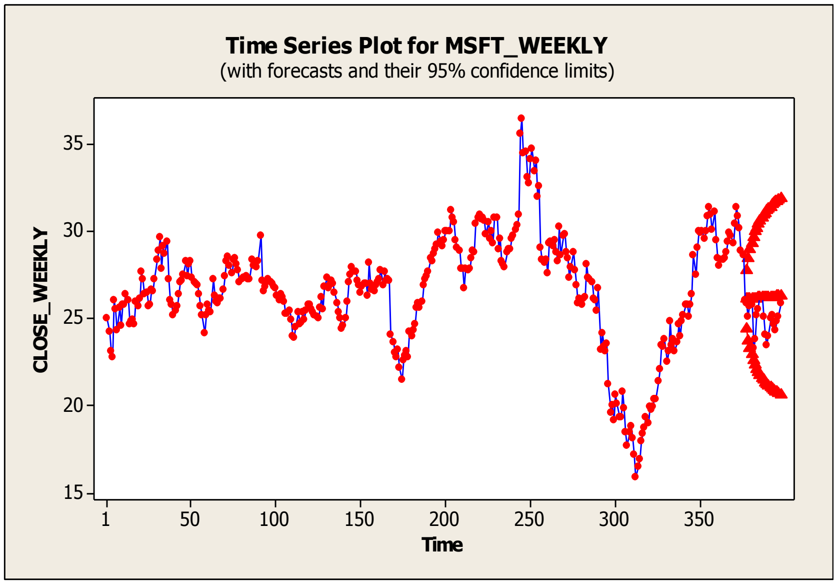

Figure 3.66: Time Series Plot for MSFT sampled weekly with forecasts

Table 3.45

Analysis of forecasts for MSFT sampled weekly (AR(1))

Date	Actual	Forecasts
6/1/2010	25.89	26.08
6/8/2010	25.11	26.10
6/11/2010	25.66	26.11
6/22/2010	25.77	26.12
6/29/2010	23.31	26.13
7/6/2010	23.82	26.14
7/13/2010	25.13	26.15
7/20/2010	25.48	26.16
7/27/2010	26.16	26.17
8/3/2010	26.16	26.18
8/10/2010	25.07	26.19
8/24/2010	24.04	26.20
8/31/2010	23.47	26.20
9/7/2010	23.96	26.21
9/14/2010	25.03	26.22
9/21/2010	25.15	26.23
9/28/2010	24.68	26.23
10/5/2010	24.35	26.24
10/12/2010	24.83	26.25
10/19/2010	25.10	26.25
10/26/2010	25.90	26.26

The MAPE for the forecasts is 0.05032.Although the forecasts do not fluctuate like the actual prices did, they follow the same overall behavior and at some points intersect. The forecasts are very similar to the actual prices.

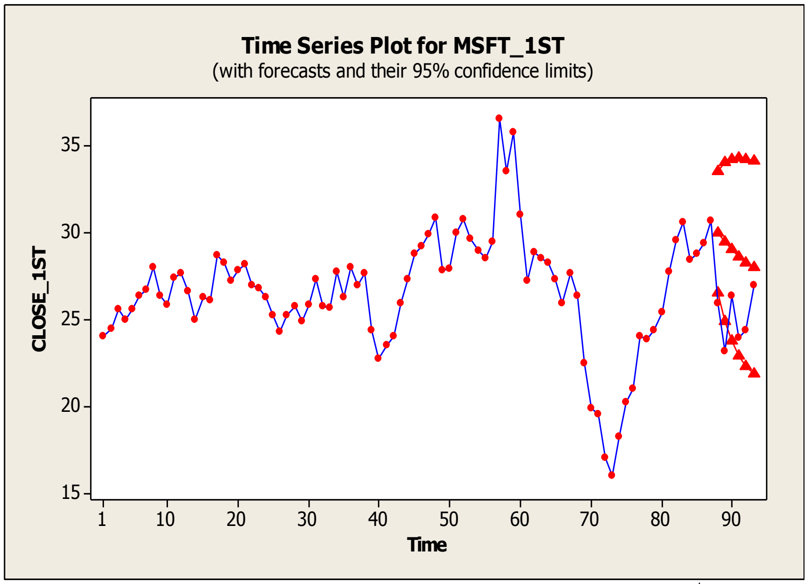

Figure 3.67: Time Series for MSFT sampled on the 1st trading day of the month with forecasts

Table 3.46

Analysis of forecasts for MSFT sampled on the 1st trading day of the month (AR(1))

Date	Actual	Forecasts
6/1/2010	25.89	30.01
7/1/2010	23.16	29.46
8/2/2010	26.33	28.99
9/1/2010	23.90	28.59
10/1/2010	24.38	28.26
11/1/2010	26.95	27.98

The MAPE for the forecasts is 0.15431.

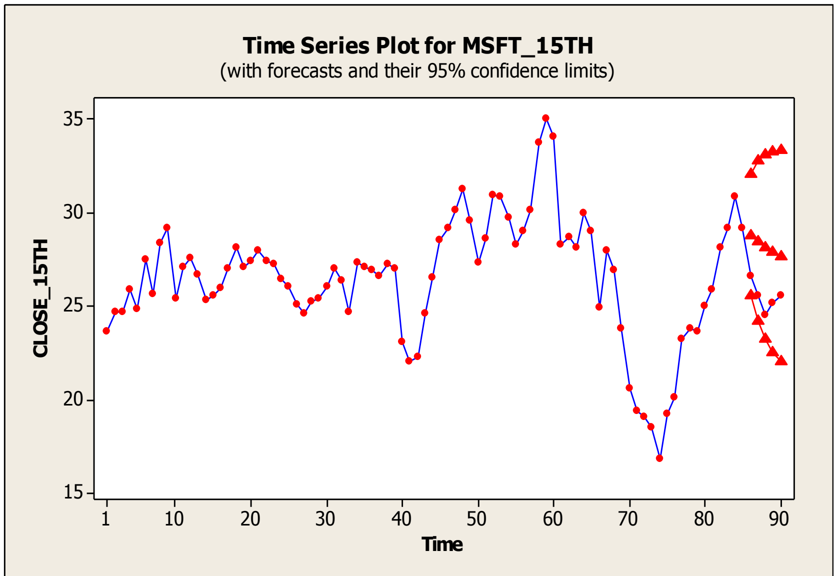

Figure 3.68: Time Series for MSFT sampled on the 15th trading day of the month with forecasts

Table 3.47

Analysis of forecasts for MSFT sampled on the 15th trading day of the month (AR(1))

Date	Actual	Forecasts
6/15/2010	26.58	28.80
7/15/2010	25.51	28.45
8/16/2010	24.50	28.15
9/15/2010	25.12	27.88
10/15/2010	25.54	27.66
6/15/2010	26.58	28.80

The MAPE for the forecasts is 0.12474. For the data sampled on the 1st and 15th trading day of the month the forecasts and the actual values both decreased, but the forecasts are quite a bit higher than the actual values. Also a few of the actual values are not contained within the 95% confidence interval. As a, result the MAPE for the forecasts for these two sampling intervals are significantly higher than the MAPE for the data that was sampled weekly.

3.1 h: Apple Inc.

The confirmed models for Apple sampled weekly and on the 15th trading day of the month is ARIMA(1,1,0) and for data sampled on the 1st trading day of the month the model is ARIMA(0,1,0). We notice that although the ACFs for the data sampled on the 1st and the 15th behave very similarly, the resulting models for the series are different. This is because the PACF for the series sampled on the 15th trading day of the month has a significant negative spike at lag 2, which is not present in the PACF for the data that was sampled on the 1st trading day of the month.

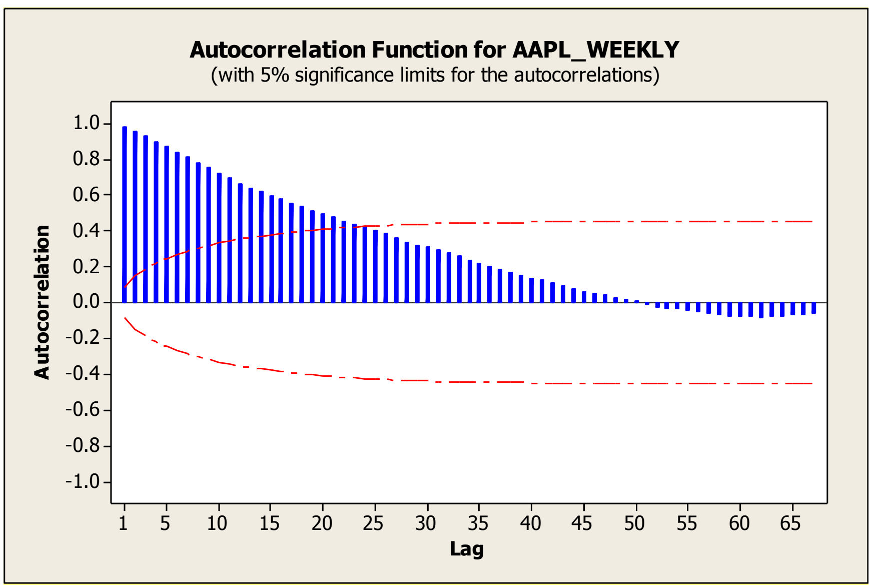

Figure 3.69: Autocorrelation function for single differenced APPL sampled on the 1st trading day of the month

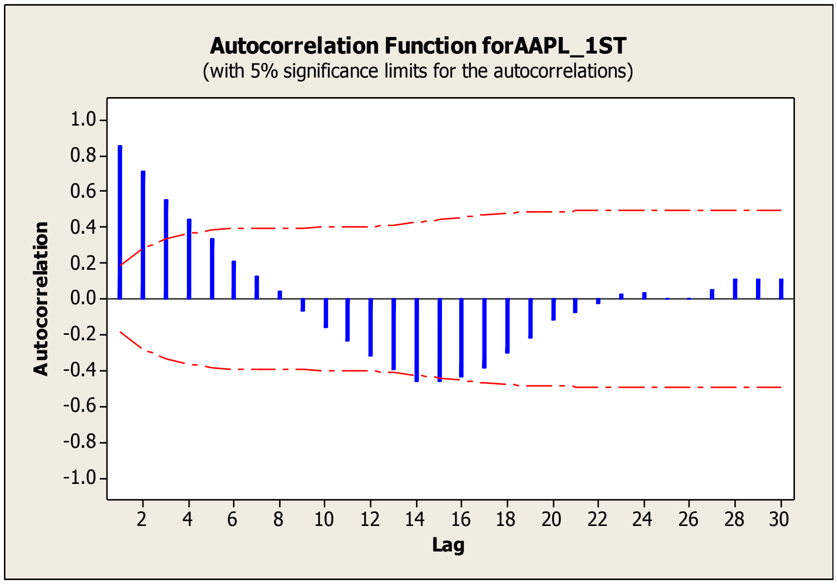

Figure 3.70: Autocorrelation function for single differenced APPL sampled on the 1st trading day of the month

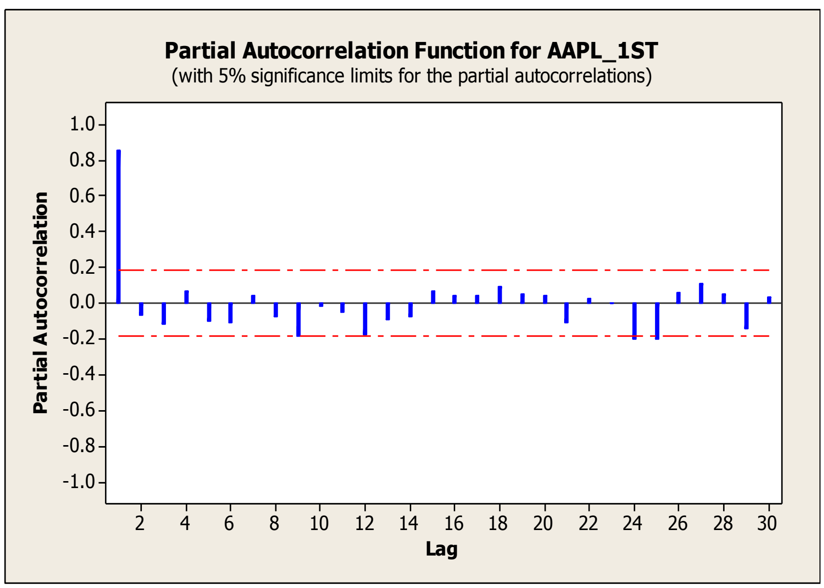

Figure 3.71: Partial autocorrelation function for single differenced APPL sampled on the 1st trading day of the month

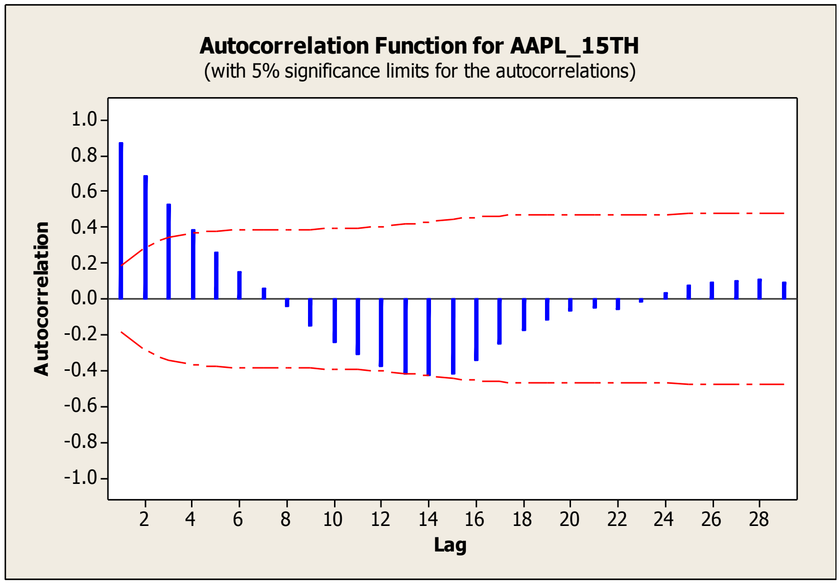

Figure 3.72: Autocorrelation function for single differenced APPL sample on the 15th trading day of the month

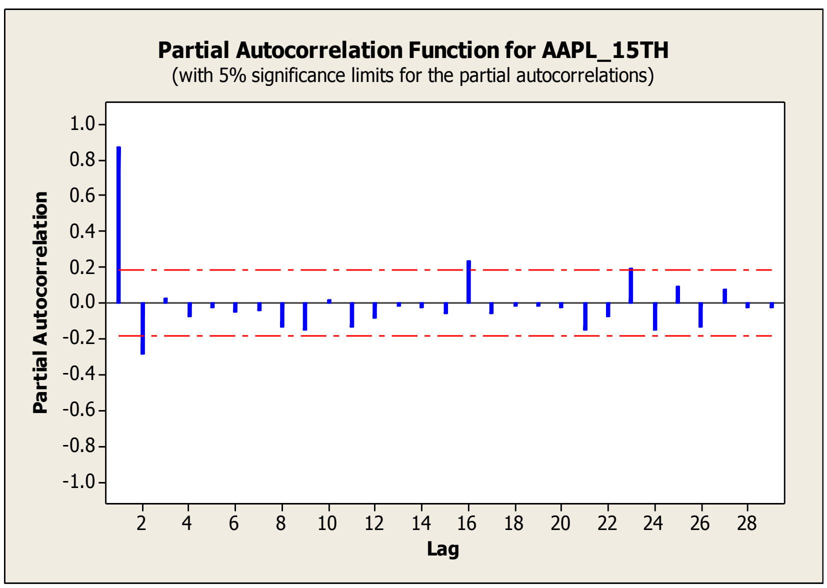

Figure 3.73: Partial autocorrelation function for single differenced APPL sampled on the 15th trading day of the month

Table 3.48

Partial Minitab output for APPL sampled weekly

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.1070   0.0431   2.48  0.013
Constant    0.4783   0.2234   2.14  0.033


Differencing: 1 regular difference
Number of observations:  Original series 535, after differencing 534
Residuals:    SS =  14174.7 (backforecasts excluded)
              MS =  26.6  DF = 532


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12      24      36      48
Chi-Square   11.7    31.8    49.6    59.7
DF             10      22      34      46
P-Value     0.303   0.081   0.041   0.085
```

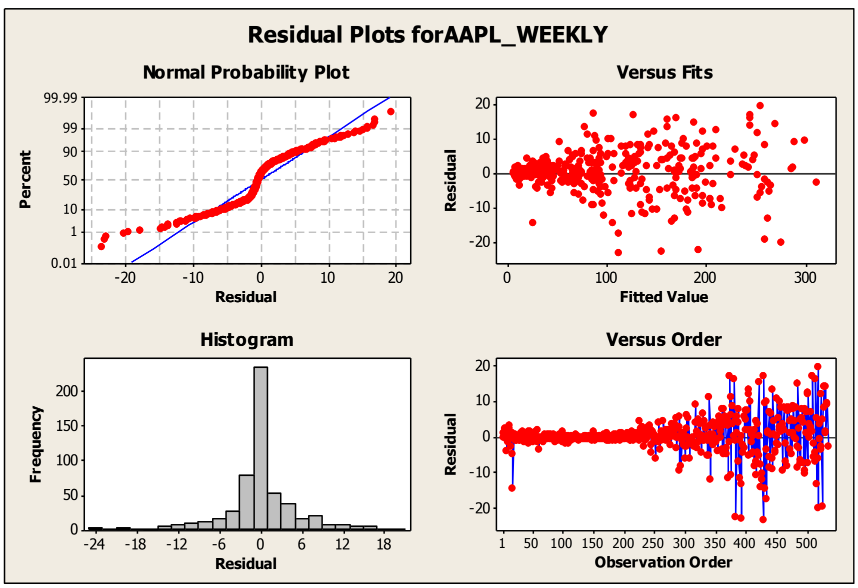

Figure 3.74: Four-in-One Residual Plots for APPL sampled weekly

The probability plot and the plot of the residuals versus order indicate that the residuals are not Normal and that the variation is not constant. The same is true for the data sampled on the 1st and 15th trading day of the month. But the Minitab output confirms that all the parameters in each model are significantly different from zero. So we tested other ARIMA models to determine if there may be a more appropriate model.

However, no other models produce results with significant estimates as well as residuals that are Normal and random. Also, although the p-values are not significant at the seasonal lags, they are low so we test to see if a seasonal component is present. So because the assumptions are violated we cannot trust the p-values of the t-test.

Table 3.49: Partial Minitab output for APPL sampled on the 1st trading day of the month

```
Final Estimates of Parameters

Type          Coef   SE Coef      T      P
AR   1      0.0699    0.0908   0.77  0.443
Constant     2.109     1.120   1.88  0.062

Differencing: 1 regular difference
Number of observations:  Original series 126, after differencing 125
Residuals:    SS =  19271.8 (backforecasts excluded)
              MS =  156.7  DF = 123

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square    8.6   20.9   39.3   41.0
DF             10     22     34     46
P-Value     0.572  0.527  0.244  0.682
```

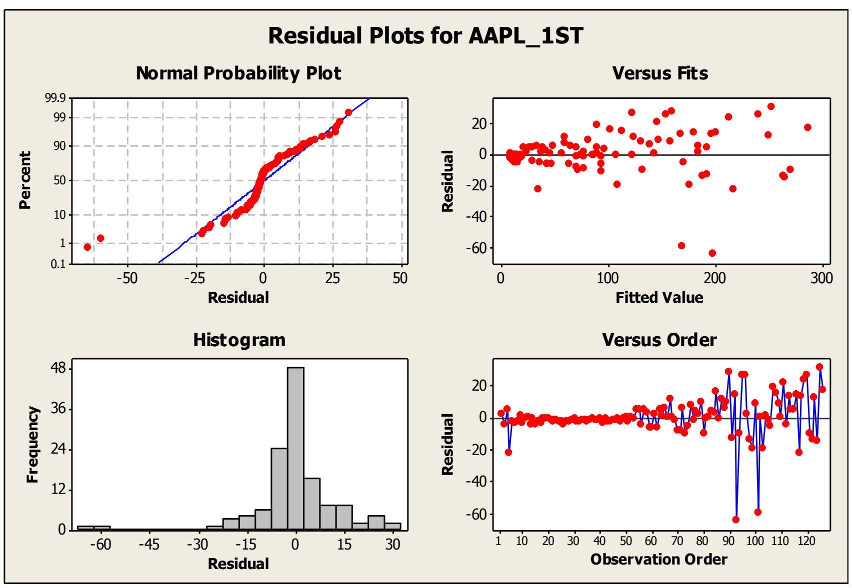

Figure 3.75: Four-in-One Residual Plots for APPL sampled on the 1st trading day of the month

Table 3.50

Partial Minitab output for APPL sampled on the 15th trading day of the month

Type	Coef	SE Coef	T	P
AR 1	0.2930	0.0928	3.16	0.002
Constant	1.813	1.073	1.69	0.094

```
Differencing: 1 regular difference
Number of observations:  Original series 122, after differencing 121
Residuals:     SS =  16554.9 (backforecasts excluded)
               MS =  139.1  DF = 119
```

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	9.3	20.2	31.1	32.2
DF	10	22	34	46
P-Value	0.500	0.569	0.611	0.939

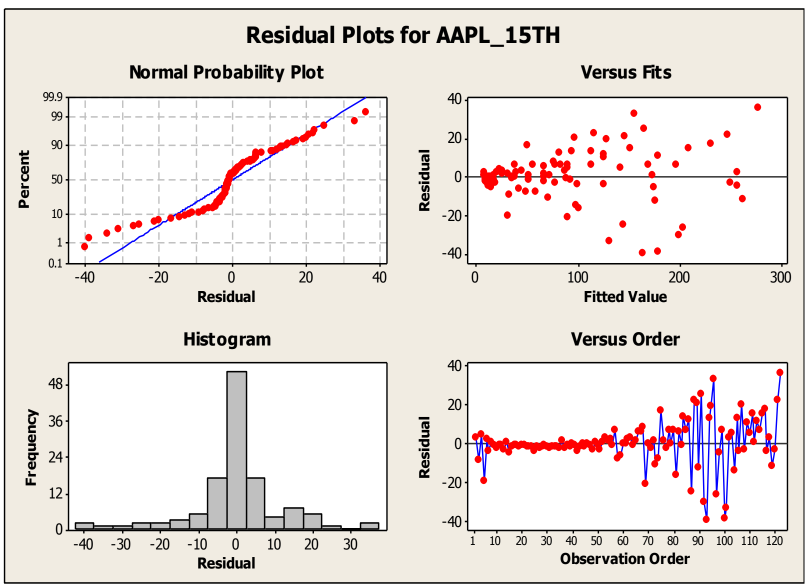

Figure 3.76: Four-in-One Residual Plots for APPL sampled on the 15th trading day of the month

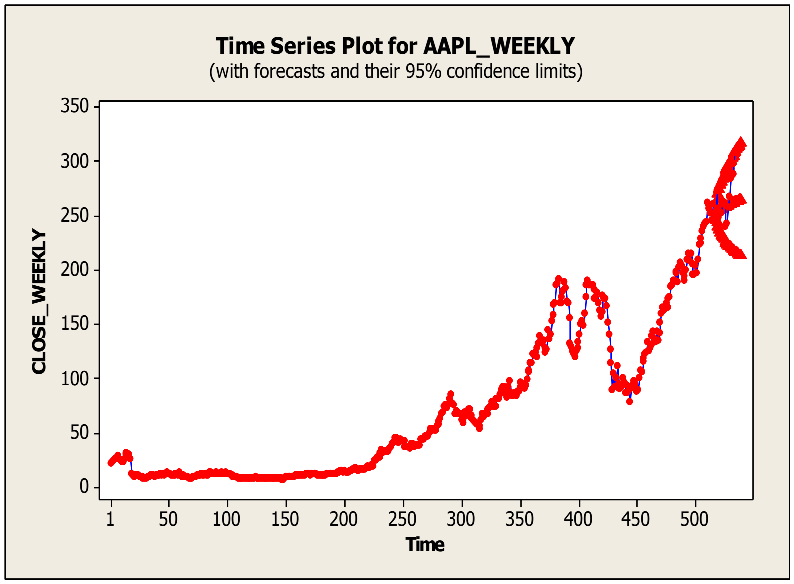

Figure 3.77: Time Series Plot of APPL sampled weekly with forecasts

Table 3.51

Forecasts analysis for APPL sampled weekly

Dates	Actual	Forecasts
6/1/2010	260.83	254.44
6/8/2010	249.33	255.01
6/11/2010	253.51	255.55
6/22/2010	273.85	256.09
6/29/2010	256.17	256.63
7/6/2010	248.63	257.16
7/13/2010	251.80	257.70
7/20/2010	251.89	258.23
7/27/2010	264.08	258.77
8/3/2010	261.93	259.30
8/10/2010	259.41	259.84
8/24/2010	239.93	260.37
8/31/2010	243.10	260.91
9/7/2010	257.81	261.45
9/14/2010	268.06	261.98
9/21/2010	283.77	262.52
9/28/2010	286.86	263.05
10/5/2010	288.94	263.59
10/12/2010	298.54	264.12
10/19/2010	309.49	264.656
10/26/2010	308.05	265.20

The MAPE for the forecasts is 0.05128.

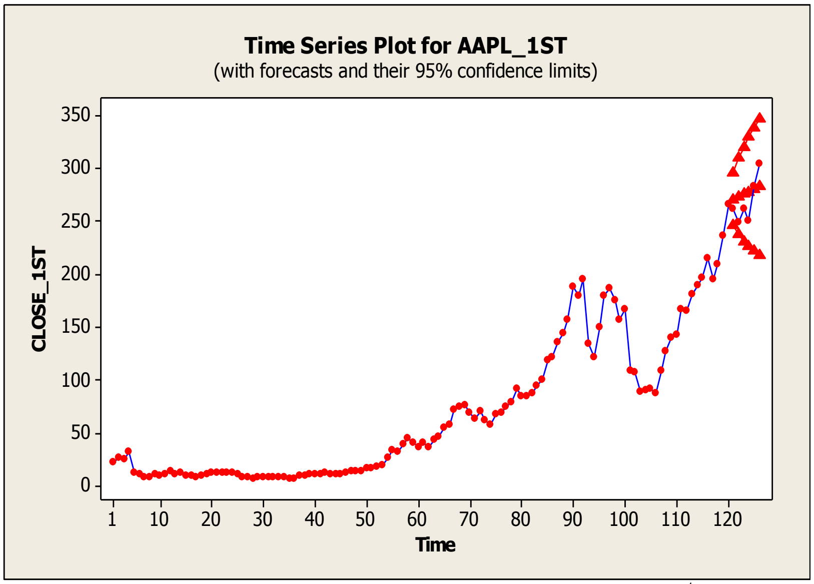

Figure 3.78: Time Series Plot of APPL sampled on 1st trading day of the month with forecasts

Table 3.52

Forecasts analysis for APPL sampled on the 1st trading day of the month

Dates	Actual	Forecasts
6/1/2010	260.83	270.58
7/1/2010	248.48	272.99
8/2/2010	261.85	275.29
9/1/2010	250.33	277.54
10/1/2010	282.52	279.81
11/1/2010	304.18	282.07

The MAPE for the forecasts is 0.06304.

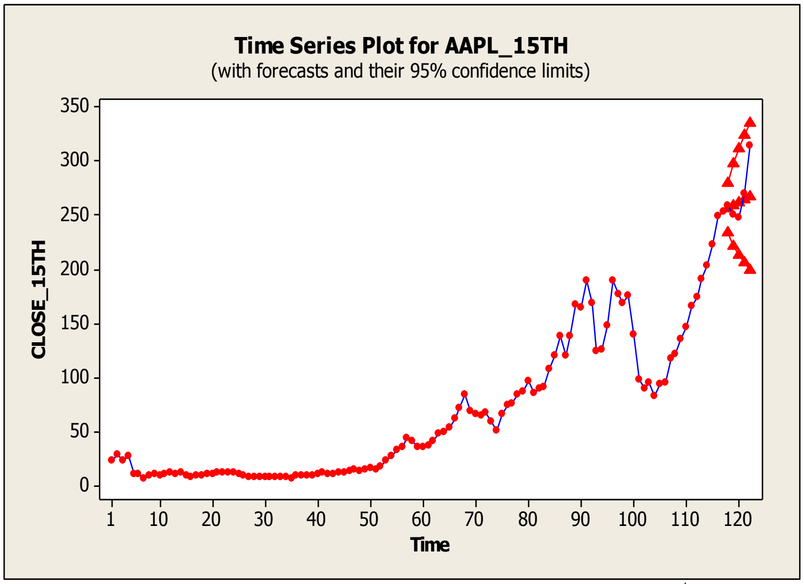

Figure 3.79: Time Series Plot of APPL sampled on1st trading day of the month with forecasts

Table 3.53

Forecasts analysis for APPL sampled on the 15th trading day of the month

Dates	Actual	Forecasts
6/15/2010	259.690	257.069
7/15/2010	251.450	259.834
8/16/2010	247.640	262.458
9/15/2010	270.220	265.040
10/15/2010	314.740	257.069
6/15/2010	259.690	259.834

The MAPE for the forecasts is 0.06125. Although all the forecasts are not far off from the actual values the residual analysis indicates that that these models are not adequate. So investor should be cautious when using this model, if they desire to use it at all. It was mentioned that the behavior of Apple's price per share has been overall constantly fluctuating, without any exact indicator present. But it is not unreasonable to assume that the behavior off this stock is more a result of speculation, then an actual event taking place. For example, the release of the newest IPod, IPhone, or MacBook is anticipated by the consumer months in advance. This anticipation could also be reflected in the market. However, there is no way to be exactly sure, but we do realize that this type of activity does influence the price per share.

CHAPTER 4

I. Conclusion

Although it is possible to forecast values based on an ARIMA (0,1,0) model, it would be not wise to do so, especially when using the model as an investment tool. The danger in using a random walk model to forecast is that we have no way of estimating ε_t within the model $y_t - y_{t-1} = \varepsilon_t$, with $\varepsilon_i \sim N(0,\sigma)$, which is the model for a random walk. Although this result isn't desirable for prediction it is makes sense with respect to the type of data we modeled. We are attempting to produce a model based solely on previous stock prices, which are correlated but can be heavily impacted by endless possible extraneous factors. From a statistical stand point broad aggregates such as national consumption, income and savings are not considered and make it hard to produce a model based only on previous stock prices. The resulting models for nearly all the series were AR(1) models either in a differenced or undifferenced form . For AATI, although the models were all AR(1) they behaved very similar to a random walk. This was the case because the AR parameter was very close to one and the constant was zero. Although the overall class of models were similar amongst the series the parameters for the confirmed models varied. The stocks belonging to the same class of ARIMA models only indicates that the ACF and PACF for the stocks behave similarly. But the analysis also revealed that even the though the ACF and PACF may behave similarly overall, differences within each particular series can heavily impact whether or not the model will be adequate for forecasting the series. Within the same stock there were also instances when the models varied greatly. This reveals the interval at which you choose to sample is very important. The adequacy of the forecasts and the model was determined by examining residual plots and analyzing how the actual prices compared to the predicted values

When examining the ACFs and PACF for the stocks, with regard to industry, there is no similarity in behavior among stocks within the same industry. The only similarities are those that exist because the stocks could be fitted using the same type of ARIMA model.

The aim of this paper was to address the following questions: using the Box-Jenkins approach, is there an ideal model and sampling interval when looking at modeling stock prices, with regard to specific industries. The sampling intervals that were explored were weekly, the 1^{st} trading day of the month and the 15^{th} trading day of the month. The companies of interest are two entities from each of the following industries: computers and software, grocery stores, semiconductor production, and soft drinks. This paper that addressed this issue was my first exploring time series analysis, its techniques, and with particular attention to the Box-Jenkins Approach. Then we carried out analysis of the stock prices for Apple, Inc. (APPL), Microsoft Corp.(MSFT), Kroger Company (KR), Winn-Dixie Stores, Inc. (WINN), ASML Holding (ASML), Advanced Analogue Technologies, Inc. (AATI), PepsiCo, Inc. (PEP), and Coca-Cola Bottling Co. Consolidated(COKE). The results of these analyses allowed us to determine which ARIMA model was most appropriate for each stock and interval and decide if there were some similarities between industries.

The results revealed stocks do not behave a certain way based on the industry they are in. Their behavior has more do to with that particular company and how much their stock prices are influenced by factors that cannot be quantified. But if given the option it is best to fit a model based on data that has been sampled weekly. To determine which interval was the best we examined the mean average percentage error for the forecasts for interval.

Table 4.1

MAPE for stocks on each interval

	Weekly	1st trading day of the month	15th trading day of the month
KR	.037	0.081	0.0361
WINN	0.424	0.662	0.536
ASML	0.087	0.114	0.058
AATI	0.0586	0.049	0.073
PEP	0.003	0.038	0.024
COKE	0.037	0.068	0.028
MSFT	0.050	0.154	0.125
APPL	0.051	0.063	0.061

The interval with the lowest MAPE for each stock is its respective suggested sampling interval was the data sampled weekly, except for PEP and AATI. There was no particular interval that performed better all of the time, even when comparing within each industry, the best sampling interval varied by stock.

II. Further Directions

The use of the Box-Jenkins Approach to forecast stock prices could be made better by incorporating covariates into the models such as; the introduction of product by an outside company, events that may be occurring in politics, natural disasters, and speculations that are being made about the company in the market. A covariate is a secondary variable that can affect the relationship between the dependent variable and other independent variables of primary interest, and our primary variable is time. By addressing the fact that there are outside components that influence the price per share, we can modify the model to try and anticipate the effect these events will have on the forecasts.

In addition to incorporating covariates we could also expand on the topic by looking at more stocks within each industry to determine if the behavior that we observed was the norm, or an exception. Also examine whether or not the behavior in different industries are codependent.

REFERENCES

[1] Bowerman, B.L. and O'Connell R.T. (1987). Time Series Forecasting: Unified Concepts and Computer Implementation. 2nd Edition. Boston: Duxbury Press.

[2] Box, G.E, Jenkins, G.M., and Reinsel, G.C.(1994). Time Series Analysis: Forecasting and Control. 3rd Edition. New Jersey: Prentice –Hall Inc.

[3] Chatfield, C. (1989). The Analysis of Time Series: An Introduction. 4th Edition. New York: Chapman and Hall.

[4] Duckworth, W.M., McCabe, G.P., Moore, D.S., and Sclove, S.L. (2003). The Practice of Business Statistics: Using Data for Decisions. New York: W.H. Freeman.

[5] Farnum, N.R. and Stanton, L.W. (1989). Quantitative Forecasting Methods. Massachusetts: PWS-KENT.

[6] Fokianos, K. and Kedem, B. (2002). Regression Models for Time Series Analysis. New York: Wiley.

[7] Hill B. and Kendall M.G. "The Analysis of Economic Time-Series- Part I: Prices." Journal of the Royal Statistical Society Vol. 116. No. 1 (1953): pp.11-34.