NEWSPAPER CIRCULATION VOLUME MODELING AND PREDICTION VIA TIME SERIES ANALYSIS

Kevin Yang
Duluth East High School
2900 East 4th Street
Duluth, MN 55812

Abstract

The purpose of this study is to analyze newspaper circulation volume data using time series analysis and develop an appropriate model used in the prediction of newspaper circulation volume. We focus on the prediction by studying the circulation data, modeling, and diagnostic checking so that short-term newspaper circulation can be predicted with reasonable accuracy. In this paper, the New York Times is chosen as our case study. The time series analysis techniques are used in our modeling; in particular, we focus on the autoregressive integrated moving average (ARIMA) model due to the non-stationary property of the data obtained. The models established are verified via residual analysis. Finally, based on the models developed, we present our prediction results together with some discussion. Our study indicates the potential and effectiveness of using the time series modeling in the prediction of newspaper circulation.

Introduction

Newspaper circulation is the number of newspapers a particular newspaper bureau distributes in an average day. This number includes both newspaper subscriptions and papers bought at newsstands. Subscription numbers are largely dependent on the population of a certain region and the newspaper's reputation (whether it is well-known or not). These two factors do not change drastically and, accordingly, the portion of newspaper circulation that is dependent on subscriptions does not change drastically, either. As a result, most of the fluctuations that are present in the circulation data are caused by changes in newsstand sales. Newsstand sales are dependent on the news that is covered in a particular issue, the weather (accessibility to newsstands), and economic factors. Newspaper circulation is often divided into two parts: daily circulation and Sunday circulation. Daily circulation is typically less than Sunday circulation because Sunday editions tend to have more sections and have coverage of the previous week's events, whereas daily editions only cover the previous day's events. Newspaper circulation tends to follow a seasonal cycle. The average circulation of the 6 months ending in March is always higher than the average of the 6 months of a year, ending in September. Reasons for this could be that many people leave to go on vacation during summers (weather being a major factor), and thus, causes people to suspend their subscriptions or stop buying from newsstands.

This paper focuses on the newspaper circulation prediction problem by studying the available circulation data, modeling, and diagnostic checking so that the short-term circulation volume can be reasonably predicted. Both daily circulation and Sunday circulation were considered in this study. To explore the feasibility and effectiveness of the proposed method, the New York Times is used as our case study. Since the circulation data can be considered as a collection of observations made sequentially in time and treated as a realization of a stochastic process, the newspaper circulation volume modeling with the time series techniques is used. Particularly, we focus on using the autoregressive integrated moving average (ARIMA) model due to the non-stationary property of the data we obtained. Time series prediction is challenging work due to many uncertainties. However, the analysis of

historical data can provide valuable insight and is essential for developing an appropriate model to predict near-term circulation volume. This paper presents the modeling and prediction results. Also our study indicates the potential and effectiveness of using the time series modeling process in the prediction of newspaper circulation. Furthermore, the modeling approach presented here can be easily modified and used in short-term newspaper circulation prediction for other urban areas.

The Circulation Data

The New York Times is the nation's third largest newspaper in terms of circulation, behind USA Today and the Wall Street Journal. It is owned by the New York Times Media Company, which owns and publishes 40 other newspapers worldwide, as well as many other media outlets. The daily and Sunday circulation data of the New York Times that was reported to the Audit Bureau of Circulations during the period 1998-2005 can be found in [1, 2] and they are shown in Fig. 1 and 2, respectively (i.e., a graph showing the observations against time). Note that in these two figures, each data point represents the average of the previous 6 months' data ending on the month shown in the graph. Analysis of this data through time series analysis can allow one to properly model the observed data and be able to make a prediction of future values. Newspaper circulation, for instance, can be treated as a marketing time series because it deals with sales figures over time. Time series analysis is used in many different areas (e.g., economics, finance, physical sciences, etc.). In economics, it can be used to predict unemployment figures by using past figures. In finance, it can be used to predict future prices of a stock, so a company or an individual can consider whether to buy, sell, or hold a stock. In the physical sciences, time series analysis can be used to make a hypothesis on future temperature trends, such as global warming. These are only a few of the many uses of time series modeling and prediction. The fundamental goal of time series analysis is to understand the mechanism that generates the observed data and, in turn, forecast future values of the series.

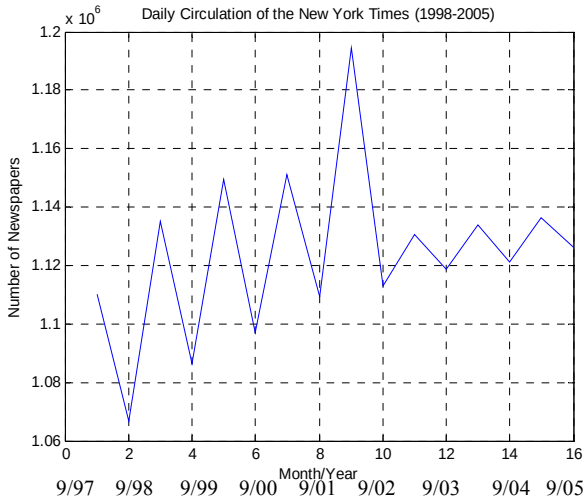

Fig. 1: Time plot of the daily circulation data.

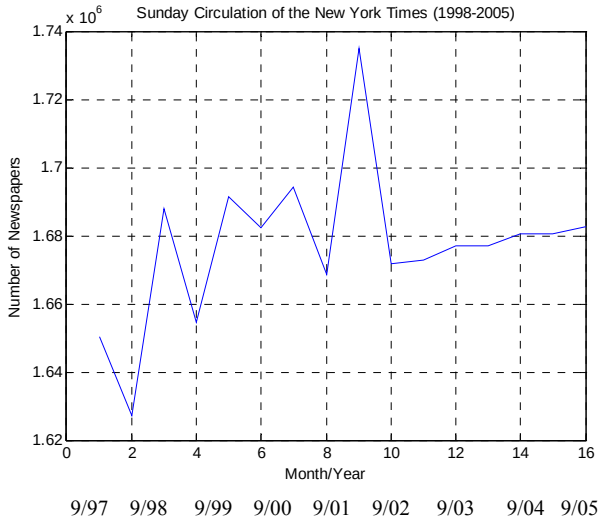

Fig. 2: Time plot of the Sunday circulation data.

Time Series Models

A time series is a collection of observations made sequentially in time. Any quantity recorded over time yields a time series. A time series model for the observed data, say $\{x_t\}$, is a specification of the joint distributions of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization. The term

time series can mean both the data and the process of which it is a realization. The fundamental aim of time series analysis is to understand the underlying mechanism that generates the observed data and, in turn, to forecast future values of the series. In this section, we briefly review the commonly used time series models. An excellent introduction to time series models can be found in [3].

Time series modeling assumes that the value of the series at time t (i.e., X_t) depends only on its previous values and on a random noise. Therefore, if this dependence of X_t on the previous p values is linear, then X_t can be represented by $X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ldots + \Phi_p X_{t-p} + Z_t$, where $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_p)$ are the model parameters called the autoregressive (AR) coefficients and Z_t is the disturbance at time t. The process $\{Z_t\}$ is usually modeled as an independent and identically distributed (iid) white noise with zero mean and variance σ^2. That is, $E[Z_t] = 0$, $E[Z_t^2] = \sigma^2$ for all t, and $E[Z_t Z_s] = 0$ if $t \neq s$, where E[.] means the expectation. The process $\{X_t\}$ is said to be a moving average process of order q if X_t can be written as $X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$ are the moving average (MA) coefficients. In the above, p and q are the orders of AR(p) model and MA(q) model, respectively. By combining the AR and MA parts, we get a mixed autoregressive moving average (ARMA) process of order (p, q). That is, $X_t - \Phi_1 X_{t-1} - \Phi_2 X_{t-2} - \ldots - \Phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}$, and this defines the ARMA(p,q) model. By introducing the back shift operator B, i.e., $B^i X_t = X_{t-i}$, then the ARMA (p, q) model can be simplified as $\Phi(B) X_t = \theta(B) Z_t$,where $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$. Even though in practice most time series we faced are non-stationary, the stationary ARMA model can still be generalized to incorporate a special class of non-stationary time series models. For instance, if the observed time series is non-stationary, we can difference the series with X_t replaced by $(1-B)^d X_t$ where $(1-B) X_t = X_t - X_{t-1}$, $(1-B)^2 X_t = (1-B) X_{t-1} = X_t - 2X_{t-1} + X_{t-2}$, etc. This operation is called differencing the time series. The ARMA model then becomes $(1-B)^d \Phi(B) X_t = \theta(B) Z_t$, which is called the autoregressive integrated moving average (ARIMA) model and expressed as ARIMA(p, d, q). In other words, any ARIMA(p, d, q) series can be transformed into an ARMA(p, q) series by differencing it d times and, thus, the analysis of an ARIMA process does not pose any difficulty as long as we know the number of times to difference the series. Clearly, the ARIMA process constitutes of three parts, an autoregressive part (AR), a differencing part (I), and a moving average part (MA). The differencing part is used to convert a non-stationary series into a stationary series. It removes the trend from the data.

In time series analysis, it is very important to calculate the sample autocorrelation function (ACF) from the observed data of a given stationary process. Given n data points $\{X_1, X_2 \ldots X_n\}$ in a time series, the autocorrelation coefficient at lag k is defined as follows:

$$r_k = \frac{\sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n} (X_t - \bar{X})^2}$$

where $\bar{X}$ is the mean value and k (= 1, 2, 3 …) is known as the lag (which is how many times the data sequence is shifted for comparison). Apparently, r_k represents the amount of correlation between $\{X_t\}$ and $\{X_{t+k}\}$, or a measure of the strength of the linear relationship between $\{X_t\}$ and $\{X_{t+k}\}$. If $r_k = \pm 1$, the correlation will be linear. However, if $r_k = 0$, then there is no relationship between $\{X_t\}$ and $\{X_{t+k}\}$. The ACF provides a useful measure of the degree of dependence among the values of a time series at different times, and for this reason they play an important role when considering the prediction of future values of the series in terms of past and present values. To find an appropriate model for the data observed we use the correlograms. A correlogram is a graph showing the time series ACF values against the lag h. From observing a correlogram sometimes we can get important information about the time

series. For example, is the series stationary? If it is stationary, then is it AR(p), MA(q) or ARMA(p, q)? What can be the order, i.e, the values of p and q for the series? It is known that for a series that fits MA (q) model, its correlogram should show a sharp cut-off after h > q, that is, the ACF becomes zero if h > q, a special feature of MA processes. If the correlogram doesn't cut-off sharply and on the contrary, it decays either exponentially or sinusoidally or both, then it may suggest that the time series either an AR (p) or ARMA(p, q) type. In this case the correlogram doesn't provide much information about the order of the series. So, we pursue the partial correlogram (i.e., partial autocorrelation function PACF vs. lag h) to see any additional information can be extracted to find the proper order p. The partial correlation between $\{X_t\}$ and $\{X_{t-k}\}$ is the correlation between the two with all variables $\{X_{t-1}, X_{t-2} \ldots X_{t-(k+1)}\}$ fixed. It can be shown that the partial ACF of an AR(p) process "cuts off" at lag p. Note that sample correlation functions do not always resemble the true correlation functions, in particular, when the number of data observed is small. Therefore, it should always be used with caution.

Another type of ARMA order selection is based on the so-called information criteria. The idea is to balance the risks of under fitting (i.e., selecting the orders smaller than the true orders) and over fitting (i.e., selecting orders larger than true orders). This is done by minimizing a penalty function, and the two commonly used functions are: $\ln \sigma^2+2(p+q)/n$ (i.e., the Akaike's Information Criterion (AIC)) and $\ln \sigma^2+(p+q) \ln (n)/n$ (i.e., the Bayesian Information Criterion (BIC)), where σ^2 is the estimated noise variance and n is the length of the data. For details regarding AIC and BIC criteria and order selection, please refer to any standard time series analysis books (e.g., [3, 4]).

Data Analysis and Modeling

Based on the time plots shown in Figs. 1 and 2, one needs to determine what time series model will be appropriate. To begin, it is often necessary to make a non-stationary time series stationary, so that its statistical properties do not change over time. To fit a time series model to the data, we need to transform the raw data into a "well-behaved" form suitable for analyzing and modeling. In other words, the transformed data can be modeled by a zero-mean, stationary type of process. That is, the trend and mean value must be removed from the circulation data. The time plot helps us determine whether the process is stationary. If not, then the series is processed to make it stationary. A special type of filtering, which is particularly useful for removing a trend, is simply to difference a given time series until it becomes stationary. Differencing is an effective way to remove trend and seasonal components in a time series. In addition, it is sometimes used to change a non-stationary time series into a stationary time series. Figures 3 and 4 show the zero-mean, differenced circulation data for the daily and Sunday cases. In Fig. 4, 2^{nd} order differencing was necessary for the Sunday newspaper circulation study. Since the trend in Figs. 3 and 4 is no longer visible and the series seems stationary, further differencing will be unnecessary. Note that the interpretation of the data and the model fitted are by no means unique; often there can be several equally valid interpretations consistent with the data. Experience and good judgment also play an important role in the modeling process.

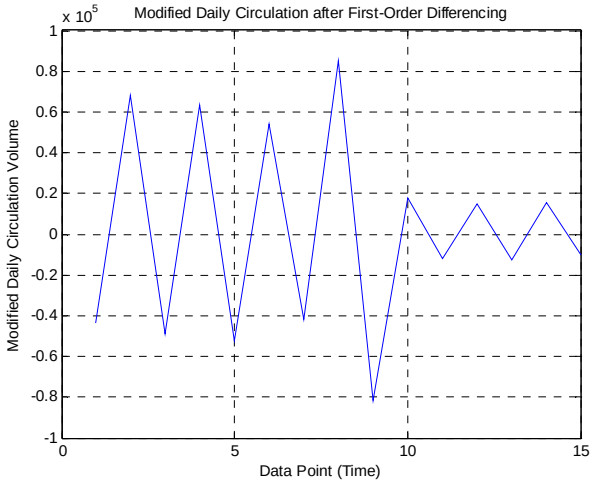

Fig. 3: The zero-mean, differenced daily circulation data.

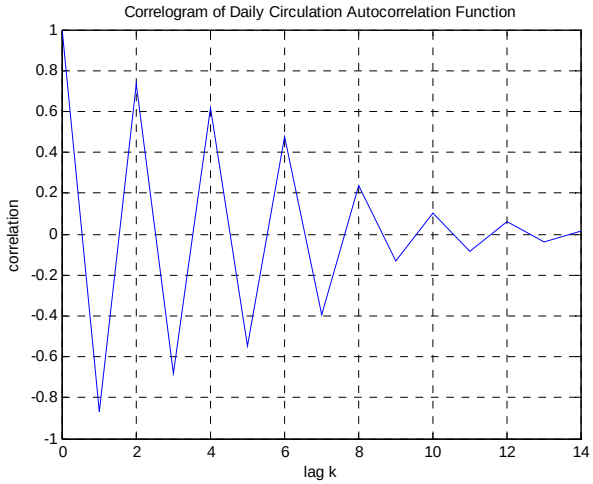

Fig. 5: The correlogram of the data shown in Fig. 3.

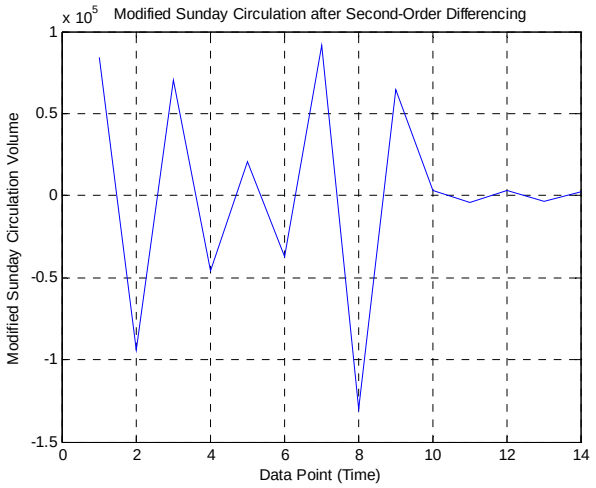

Fig. 4: The zero-mean, second-order differenced Sunday circulation data.

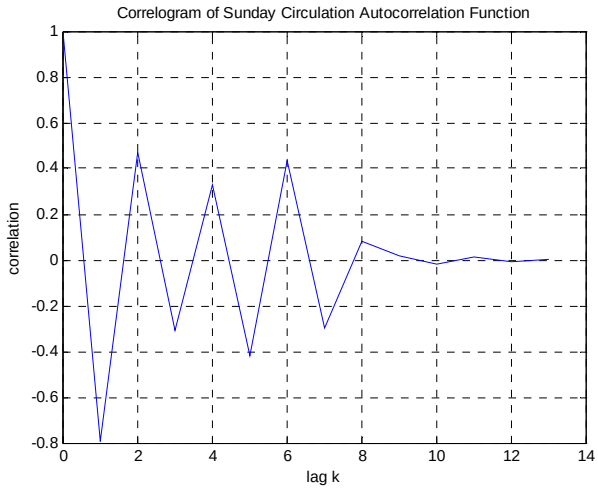

Fig. 6: The correlogram of the data shown in Fig. 4.

To assess the degree of dependence in the time series data and to select a model for the data that reflects this, we further examine the correlation function of the data. Figures 5-8 show the correlogram and partial correlogram of the differenced data for both the daily and Sunday circulation data. By inspecting these figures, we found that both ACF and PACF don't sharply cut-off to zero, which indicates that the appropriate model should be of the ARMA (p, q) type. Therefore, the ARIMA (p, d, q) model should be used for the original raw data because a differencing operation was conducted (i.e., $d \neq 0$).

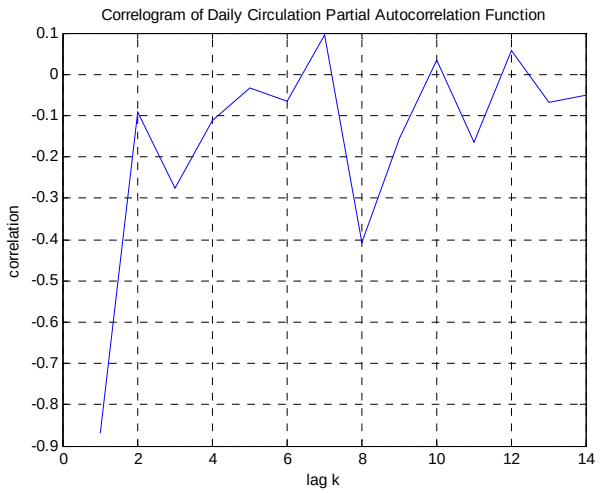

Fig. 7: The partial correlogram of the processed data shown in Fig. 3.

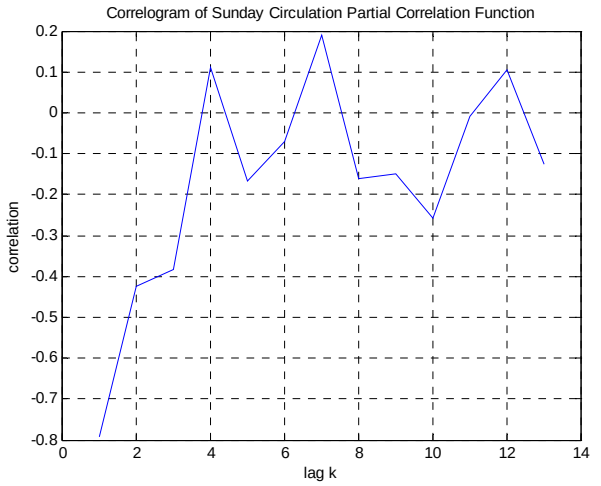

Fig. 8: The partial correlogram of the processed data shown in Fig. 4.

The models were selected based on the minimization of the AIC information criterion mentioned earlier. Since the Hannan-Rissanen (HR) algorithm [4] is a very effective way in determining ARMA model parameters, we used this HR procedure [5] to find the parameter values. For the daily circulation data shown in Fig. 3, we found that the ARIMA (1, 1, 1) model with $\Phi_1 = -0.848434$ and $\theta_1 = 0.0971008$ seems to generate the best results with the minimum AIC = 20.2149 among those candidate models selected. Similarly, the appropriate model for the Sunday circulation data in Fig. 4 was found to be of ARIMA (1, 2, 1) with $\Phi_1 = -0.557954$ and $\theta_1 = -0.966033$ with the corresponding AIC = 8.18584.

Model Validation

After fitting a model to a given set of data, the model needs to be examined to see if it is indeed an appropriate model. If the model is a "good" one, then its residuals should be random and close to zero. There are several ways of checking if a model is "good". One commonly used approach to diagnostic checking is the examination of residues. That is, the residues can be treated as a time series and the properties and correlogram of the residues (i.e., the autocorrelation coefficients of the residues at different lag k) can be studied. Therefore, the residuals, which are generally defined as the difference between the observed and fitted values (error), are checked. For a good fit, the residual time series should be close to an iid zero-mean white noise. If the residues, say $\{y_1, y_2, \ldots, y_n\}$, is a realization of such an iid sequence, then about 95% of the sample autocorrelations should fall between the bound $\pm 2/\sqrt{n}$, where n represents the length of data points [4]. A detailed analysis of residuals from ARMA processes can be found in [3]. To verify the models given in Data Analysis and Modeling, we conducted residual analysis. The correlogram of the residuals from the ARIMA (1, 1, 1) model for the daily circulation data and the ARIMA (1, 2, 1) model for the Sunday circulation data are shown respectively in Figs. 9 and 10. From these figures, one can see that correlation coefficients of the residue time series are fairly small and they fall within the bounds $\pm 2/\sqrt{n}$. Data points that fall within these bounds can be considered as "close to zero." Thus, we have no reason to reject the hypothesis that the set of data constitutes a realization of a white noise process. Therefore, these two models will be used in the prediction of the circulation volume. Thus, by changing the form of the standard ARIMA equation to match the p, d, and q values and by plugging in the parameters Φ and θ, the following models,

where $\{X_t\}$ is the circulation volume time series, will result. Notice that the mean value, which was removed earlier, is added back into the models on the right side.

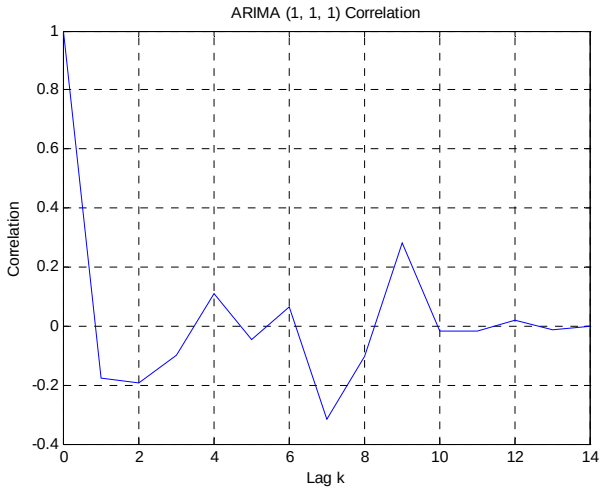

Fig. 9 The correlogram of the residual time series.

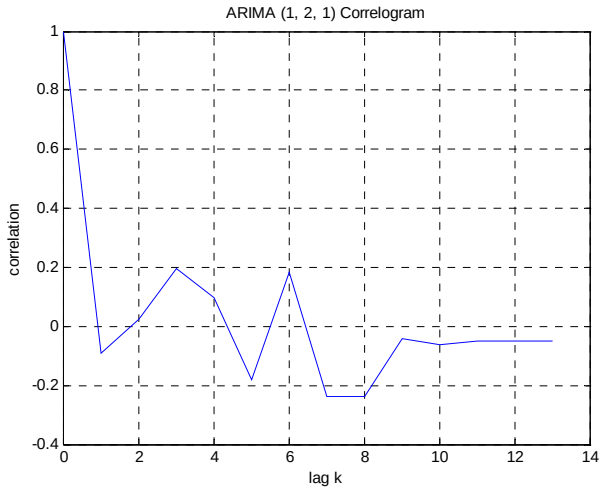

Fig. 10 The correlogram of the residual time series.

Daily Circulation Time Series Model:

$(1-B)(1+0.848434B)X_t = 1.124x10^6 + Z_t + 0.0971008Z_{t-1}$

Sunday Circulation Time Series Model:

$(1-B)^2(1+0.557954B)X_t = 1.677x10^6 + Z_t - 0.966033Z_{t-1}$

Circulation Prediction

Based on the modeling of the daily and Sunday circulation time series data, one can now use these models to make a prediction of future volume. That is, using the models developed and up-to-date circulation data, future circulation can be predicted. However, if future circulation were predicted, we would be unable to check it because The New York Times has not released the circulation data for the 6-month period ending in March 2006 at the time this paper was written. Thus, we would be unable to verify if the model is a good one if we used it to predict future data points. So, in order to have an actual figure to compare the predicted value to, we decided to use the time series models that we chose to predict known data points. Because p = 1 in both models, only one previous known data point is necessary for the purpose of prediction, even though the time series at current value also indirectly depends on its previous values (since $\{X_t\}$ is dependent on $\{X_{t-1}\}$ and $\{X_{t-1}\}$ is dependent on $\{X_{t-2}\}$ and so on). The zero-mean, differenced circulation data Y_t $\{t = 1, 2, 3\ldots\}$ was used together with the model to perform the one-step-ahead prediction. Notice that Y_t is defined as $(1-B)X_t - 1.124x10^6$ for daily circulation, $Y_t = (1-B)^2 X_t - 1.677x10^6$ for Sunday circulation, and X_t represents the actual newspaper circulation volume time series.

Figures 11 and 12 show the actual circulation data (zero-mean, differenced) $Y_{t,}$ $(i = 1, 2, \ldots)$ compared to predicted circulation data. The Best Linear Predictor algorithm in the Mathematica software[5] was used to calculate these predicted values. Clearly, the predicted results Y_t can be converted to the actual X_t in the reverse operations (i.e., inverse differencing and addition of the mean value). From these two figures, we see that the predicted values closely

mirror the actual values, thus, these models can be considered reasonably "good", considering only 16 data points were used. These models can be vastly improved with more data, but it would be difficult to obtain, as discussed below.

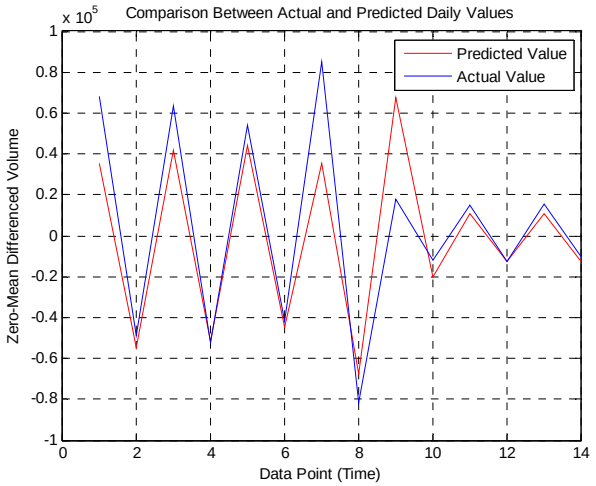

Fig. 11: Result comparison for the daily circulation study.

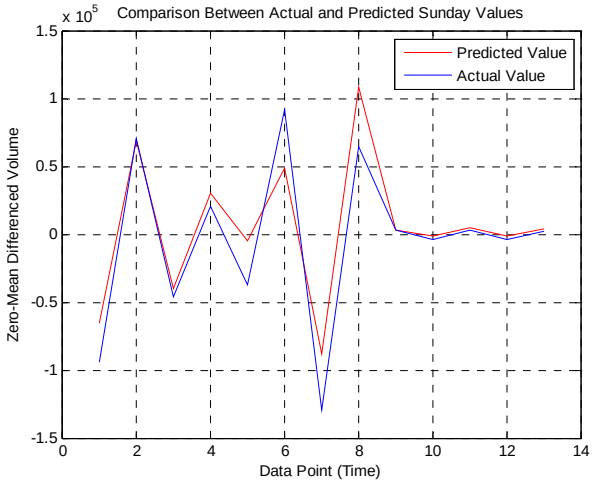

Fig. 12: Result comparison for the Sunday circulation study.

(a) Possible improvement (model refinement). The most obvious way of improving this model is to use more data. With more data, one can find trends that may have lasted for longer than the eight years of data used. In addition, it can also be used to create a better model. For instance, with more data points, the bounds, which are defined as $\pm 2/\sqrt{n}$, will be smaller and as a result, be more helpful in choosing the best model.

However, obtaining additional data would be difficult and costly to do. The Audit Bureau of Circulations, which contains large amounts of circulation data, is not public and is only available to newspaper companies and academic institutions, more specifically, universities.

(b) Possible sources of error. It is very difficult to predict future newspaper circulation with high accuracy because future values are only partially dependent on past values. One possible source of error is the amount of the data used. Since only eight years of newspaper circulation was used, any trends that lasted more than eight years cannot be determined. Because of this, it would be inappropriate to use this model to determine circulation far into the future.

Some other possible causes for errors would be a major news development that would encourage more people to purchase newspapers, adverse weather conditions that make newsstands inaccessible, or economic changes. These should be considered as outliers. However, we decided not to remove these outliers. The small amounts of circulation data was one of the major drawbacks to this study and outlier removal would only exacerbate the problem.

Conclusion

In this paper, we study the modeling and prediction of the newspaper circulation data via the time series techniques. The obtained data is treated as a realization of a time series stochastic process. Time series analysis is then used for modeling of the circulation volume data. In particular, we focus on the ARIMA model due to the non-stationarity of the observed data. A differencing technique is used to remove the trend from the data. The data correlation via the correlogram and partial correlogram are further examined to determine appropriate model type. In addition, the Hannan-Rissanen procedure is

used to determine the model order and also estimate the model parameter values. The model validation is performed via the residual analysis. Finally, the time series models are used to predict circulation volume and the results are presented. Our study indicates that these models can predict newspaper circulation volume with reasonable accuracy. Additionally, it indicates the potential and effectiveness of using the time series modeling in the prediction of newspaper circulation.

Acknowledgment

The author would like to thank Professor Jiann-Shiou Yang, from the University of Minnesota Duluth, for his helpful comments and suggestions on improving this paper.

References

1. New York Times Company: Investors: Circulation Data. The New York Times Company. 2006.<http://www.nytco.com/investors-nyt-circulation.html>.

2. The New York Times Company Annual Report 2004. The New York Times Company. 2004.

3. C. Chatfield, The Analysis of Time Series. Chapman & Hall/CRC. 2004.

4. P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting, 2nd ed., New York, NY: Springer, 2002.

5. Mathematica Time Series, Version 1.3.1. Wolfram Research, Inc. 2003.

Biographical Information

Kevin Yang is a senior at East High School in Duluth, Minnesota. He has won numerous academic awards and was recently designated as a Minnesota Scholar of Distinction in Mathematics. He plans on studying either engineering or biology in college.