Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.3, 2014

Modeling the Trend of Performance of the Manchester United Football Club in the 1960-2013 English Premiership

Linus Addae[1], Martin Owusu Amoamah[2], Iddrisu Wahab Abdul[*2] & Osei Antwi[2]
[1]Department of Mathematics and Statistics, Youngstown State University, UK
[2]Department of Mathematics and Statistics, Accra Polytechnic, Accra, Ghana
[*]E-mail of the corresponding author: perfectwahab@yahoo.com

Abstract
This research studied the trend of performance of Manchester United Football Club in the 1960-2013 English premiership seasons. The three main variables involved in this study are the number of games won, games drawn and games lost by Manchester United for the study period. But this study concentrated on the number of games lost since the objective of every manager of a team is to minimize loss and maximize win or draw. Thus the objective of this study is to develop a model for predicting the number of games that would be lost by Manchester United in future seasons using games played in the previous seasons. The data used for this study are secondary data obtained from sportamok and English premiership websites. The statistical technique used for this study is time series analysis. Specifically, Autoregressive Integrated Moving Average (ARIMA) model was used to develop a model for the number of games lost by Manchester United. The model was used to forecast for the next fifteen seasons. The model predicted that Manchester United will lose six (6) games for 2013/2014 season.

Keywords: Manchster United Football Club, English Premiership, Time Series Analysis, ARIMA.

1. Introduction

The English Premier League (EPL) is a professional league for association football clubs which was formerly called ”Football League” and dates as far back as 1888. The EPL was founded in February, 1992 as a result of the decision of the clubs in the Football League First Division to break away from the Football League.

The EPL before the 1995/1996 season consisted of twenty-two (22) teams that played in the premiership. The number of teams playing in the premiership since 1995/1996 season to date has been reduced to twenty (20) by the premiership board. Thus, before the 1995/1996 season, if n is the number of participating teams in the premier league, then each team played a total of $2(n-1)$ games, that is $2(22-1) = 42$ games for a particular season. Likewise, if n is the number of participating teams in the premier league, from 1995/1996 season to date, then each team played a total of $2(n-1)$ games, that is $2(20-1)= 38$ games for a particular season. These games are played on a home and away basis, once at their home stadium and once at that of their opponents.

The outcome of each game played is win, draw, and loss. There were two (2) points awarded for win, one (1) for draw, and zero (0) for loss until 1981. From 1981, a win was awarded three (3) points, one (1) for draw and zero (0) for loss. Teams are ranked by total points, then goal difference and then goals scored. At the end of each season, the club with the most points is crowned champion. If points are equal, the goal difference and then goals scored determine the winner. If still equal, teams are deemed to occupy the same position. The three lowest placed teams are relegated into the Football League Championship, and the top two teams from the Championship, together with the winner of play-off involving the third to sixth placed Championship clubs, are promoted in their place.

Manchester United Football Club, formerly known as Newton Heath LYR Football Club, is an English professional football club which was established in 1878 by the Carriage and Wagon department of the Lancashire and Yorkshire Railway depot at Newton Health. Manchester United by 1888 the club had become a founding

member of the Combination (league during the early days of English Football). The club was renamed as Manchester United Football Club on 24th April, 1902 and moved to Old Trafford in 1910 till date.

From the beginning of the club's official managerial records in 1892 to the start of the 20132014 seasons, the club has had nineteen (19) full-time managers. The most successful and longest serving of them all was Sir Alex Ferguson. Sir Ferguson served the club for more than 26years and won thirteen (13) Premier League titles, five (5) Football Association Challenge Cups (FA cups), ten (10) Community Shields cups, two (2) Union of European Football Associations (UEFA) Champions League titles, one (1) UEFA Cup Winners' Cup, one (1) UEFA Super Cup, one (1) Intercontinental Cup and one FIFA Club World Cup.

The objective is to study the number of games that were lost by Manchester United in previous seasons for the study period to develop a model that can predict number of games that would be lost in future seasons. Even though there are three outcomes of a game, every coach's objective is to minimize losses and maximize win or draw. Thus if Manchester United technical team knows before the start of any future season the number of games that would be lost, strategies will be adopted to minimize the losses and maximize win. This has motivated me to study losses instead of win or draw.

2. Method

2.1 Data

The data used for this study are secondary data obtained from sportamok and English premiership websites. The data consists of several variables such as number of games played, games won, games drawn, games lost, goals scored, goals scored against and points accumulated at the end of the season.

Missing data are part of almost all research, and the resarcher must decide on how to estimate them. There is a missing data point for 1975 since Manchester United was relegated in 1974 and did not take part in the league. This missing data was estimated using SPSS. Since the data used for this project is time series data, SPSS has number of alternative ways to estimate or replace missing time series data. These methods include Mean series, Mean of nearby points, Median of nearby points, Linear interpolation and Linear trend at points. All the five methods of estimating or replacing missing values in SPSS for times series data works in efficient way. The study used linear trend at point method to estimate the missing value. This method replaces missing values with linear trend for that point. The existing series is regressed on an index variable scaled from 1 to n. Missing values are replaced with their predicted values.

2.2 Time Series Analysis

There are two methods for analyzing time series data. These are time domain methods and frequency domain methods. Time domain methods, according to Robert H. Shumway and David S. Stoffer [2] are generally motivated by the presumption that correlation between adjacent points in time is best explained in terms of dependence of the current value on past values. The time domain approach focuses on modeling some future value of a time series as a parametric function of the current and past values.

On the other hand, frequency domain methods assume the primary characteristics of interest in time series analyses relate to periodic or systematic sinusoidal variations found naturally in most data. These periodic variations are often caused by biological, physical, or environmental phenomena of interest. The study of periodicity extends to economics and social sciences, where one may be interested in yearly periodicities in such series as monthly unemployment or monthly birth rates. Time-series analysis is more appropriate for this study because the data are clearly explicit violation of the assumption of independence of errors. The errors are correlated due to the patterns over time in the data. Also, the patterns in the data may either be obscure or spuriously enhance the effect of an intervention unless accounted for in the model.

2.3 Components of Time series

Basically, time series data has four components. These are trend, cyclical, seasonal and irregular components. The trend is the long term pattern of a time series. A trend can be positive or negative depending on whether the time series exhibits an increasing long term pattern or a decreasing long term pattern. If a time series does not show an increasing or decreasing pattern then the series is stationary in the mean.

A cyclical movement is identified by up and down movement around a given trend. The duration of a cycle depends on the type of business or industry being analyzed. Seasonality arises when the time series displays regular fluctuations during the same month (or months) every year, or during the same quarter every year. Irregular patterns are unpredictable time series data. Every time series has some unpredictable component that makes it a random variable. In prediction, the objective is to model all the components to the point that the only component that remains unexplained is the random component.

2.4 Assumptions of Time series analysis

The assumptions of time series analysis include normality of the distribution of residuals, homogeneity of variance, zero mean and independence. The time series model is developed and then normality of residuals is evaluated in time-series analysis. The normal plot of residuals for the model is examined before evaluating an intervention. The normalized plot of residuals is examined as part of the diagnostic phase of modeling. The traditional square root, logarithmic, or inverse transformations are appropriate in the event of nonnormally distributed residuals.

After the model is developed, the plots of standardized residuals versus predicted values are examined to evaluate homogeneity of variance over time. If the width of the plots varies over predicted values, the data can be transformed. McCleary and Hay [1] recommend a logarithmic transformation to remedy heterogeneity of variance.

The time-series plot is scrutinized before and after adjusting for autocorrelation and seasonality to detect obvious outliers. According to Cryer [3] there are no concrete guidelines to determine how discrepant a case must be to be labeled an outlier in a time-series analysis. An outlier is dealt with in the usual manner by checking the original data for errors, deleting the observation, replacing the observation with an imputed value, and so on.

2.5 ARIMA Models

An ARIMA (auto-regressive, integrated, moving average) time series model has three parameters: (p, d, q). The first letter p, represents the number of autoregressive terms in the model, d is the order of differencing and q is the number of moving average terms incorporated in the model. Identification of a time series model is the process of finding integer, values of p, d, and q that model the patterns in the data. When the value is 0, the element is not needed in the model. The middle element, d, is investigated before p and q. The goal is to determine if the process is stationary and, if not, to make it stationary before determining the values of p and q. A stationary process has a constant mean and variance over the time period of the study. If the process is not stationary, differencing the observations is one way to achieve stationarity. Differencing is done by subtracting the value of an earlier observation from a later observation.

Autoregressive models due to Robert H. Shumway and David S. Stoffer [4] are based on the idea that the current

value of the series, x_t can be explained as a function of p past values, $x_{t-1}, x_{t-2}, \dots, x_{t-p}$, where p determines the number of steps into the past needed to forecast the current value.

Generally, an autoregressive model of order p according to Robert H. Shumway and David S. Stoffer [5], denoted by AR(p), is of the form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

Where x_t is stationary, and $\phi_1, \phi_2, \dots, \phi_p$ are constants ($\phi_p \neq 0$). We assume w_t is Guassian white noise series with mean zero and variance σ^2_w. The mean of x_t in the formula above is zero. If the mean of x_t is zero, we substitute $x_t - \mu$ for x_t and the formula yields

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \dots + \phi_p(x_{t-p} - \mu) + w_t$$

And can be written as

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

Where $\alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$

Now, using the backshift operator, the AR(p) model can be written as

$$(1 - \phi_1\beta_1 - \phi_2\beta_2 - \cdots - \phi_p\beta_p) = w_t$$

which can be written as

$$\phi(\beta)x_t = w_t$$

Where

$$\phi(\beta) = 1 - \phi_1\beta_1 - \phi_2\beta_2 - \cdots - \phi_p\beta_p$$

As a special case, consider the AR(1) model,

$$x_t = \phi x_{t-1} + w_t$$

Iterating backwards k times, we have

$$x_t = \phi x_{t-1} + w_t$$

$$x_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t$$

$$= \phi^2 x_{t-2} + \phi w_{t-1} + w_t$$

$$= \phi^k x_{t-k} \sum_{j=1}^{k-1} \phi^j w_{t-j}$$

.

This indicates that, by continuing to iterate backwards and provided |φ| < 1 and x_t is stationary, then the AR(1) model can be represented as a linear process given by

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$$

2.6 The Autocorrelation function (ACF) and the Partial autocorrelation functions (PACF)

Models are identified through patterns in their ACFs (autocorrelation functions) and PACFs (partial autocorrelation functions). Both autocorrelations and partial autocorrelations are computed for successive lags in the series. The first lag has an autocorrelation between X_{t-1} and X_t, the second lag has both an autocorrelation and partial autocorrelation between X_{t-2} and X and so on. ACFs and PACFs are the functions across all the lags.

The ACF and PACF of the series are computed by the formulas

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

Where N is the number of observations, k is the lag. $\bar{X}$ is the mean of the whole series and the quantity in the denominator is the variance of the whole series.

The standard error of an autocorrelation is based on the squared autocorrelations from all previous lags. At lag 1, there are no previous autocorrelations, so ${r_0}^2$ is considered to be zero. Thus the standard error for the series is computed using

$$SE_{rk} = \sqrt{\frac{1 + 2\sum_{t=0}^{k-1} {r_t}^2}{N}}$$

The standard error of the partial autocorrelation is the same for all lags. It is computed using

$$SE_{pr} = \frac{1}{\sqrt{N}}$$

However, the standard error for partial autocorrelation is the same at all lags.

The criteria for judging which ARIMA model to use based on the computed values for both ACF and PACF at different lags for the number of games lost by Manchester United for the study period depends on whether the

computed values are significantly different from zero. For autocorrelation functions, if the computed value at some lag(s) is significantly different from zero, it is included in the ARIMA model. Similarly, if a partial autocorrelation at some lag is significantly different from zero, it is also included in the ARIMA model. The significance of full and partial autocorrelations is assessed using their standard errors. Even though autocorrelations and partial autocorrelations can be assessed numerically, the standard practice is to observe their plots for the series, in this case the number of games lost by Manchester United for the study period. The patterns displayed by these plots enable assessments to be made on the number of AR and MA terms to be included in the ARIMA model

3. Results

3.1 The behavior of games won over the seasons

The figure below illustrate the games won by Manchester United since the inception of the English premiership but the does not include the 2013/2014 season since the season is still underway.

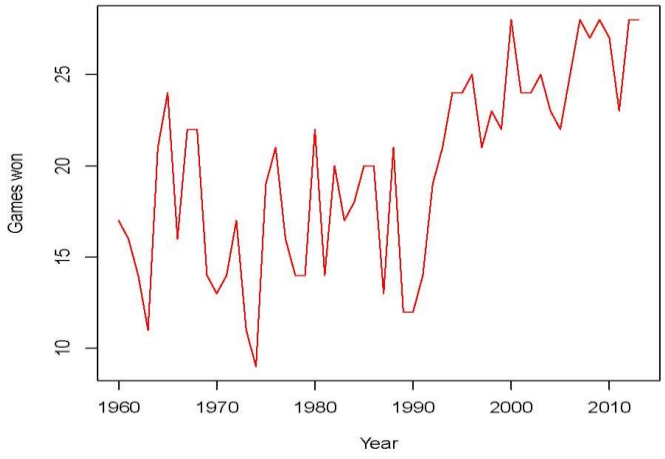

Figure 1: A time series graph showing games won by Manchester United for the seasons

From the graph above the behavior of games won by Manchester United over the seasons depict that of time series. The games won between 1960 and 1970 exhibits up and down movement. The games won from 1971 shows an in increasing trend until 1990 when it declined. The team again from 1991 experienced an increasing trend in the number of games won. Thus the number of games won by Manchester United over the seasons considered for this project shows an increasing trend.

3.2 The behavior of games drawn over the seasons

Observing the graph in Figure 2 indicates that, the number of games drawn by Manchester United for the seasons considered for this project seem to have a upward movement with no regular pattern from 1960 to 1980.The series decreased from 1981 through to 2013.This upward and downward movement occurred throughout the seasons for the number of games drawn. Thus the number of games drawn by Manchester United for the period considered for this project is a time series data with irregular movement.

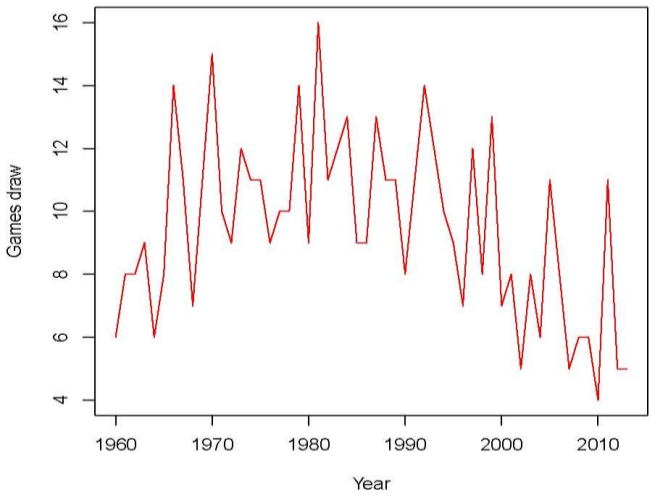

Figure 2: A time series graph showing games draw by Manchester United for the seasons

3.3 The behavior of games lost over the seasons

From the graph it can be observed that games lost by Manchester United over the seasons depict a downward movement with no apparent pattern. The series decreased in 1961 and increased between 1962 and 1963.This irregular movement transpires throughout the season for the number of games lost. Thus the number of games lost by Manchester United for the period considered for this project is a time series data with irregular movement

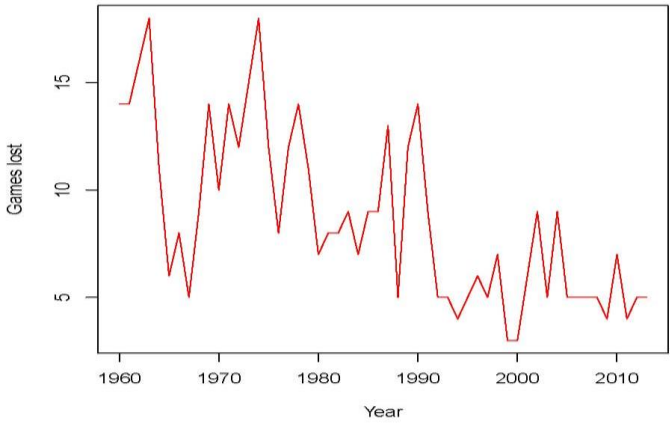

Figure 3: A time series graph illustrating games lost by Manchester United for the seasons

3.4 Identifying an Appropriate Model for games lost

This is a non-seasonal time series data and composed of only trend and irregular components. Decomposing this data requires separating the series into these components, which is estimating the trend component and irregular component. The trend component for non seasonal time series that can be studied using additive model has to be smoothing using an appropriate smoothing method such as a simple moving average. The games lost by Manchester United for the seasons is smoothing using simple moving average of order nine in order to separate trend from irregular component.

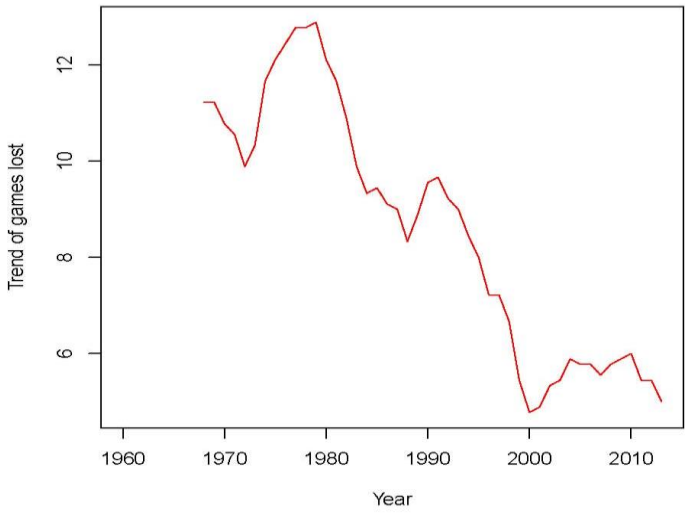

Figure 4: A time series graph showing the trend of games lost by Manchester United for the seasons

The graph above indicates that the number of games lost by Manchester United over the seasons shows decreasing trend. The games lost increased from 1970 to 1980 and exhibit a downward trend from 1981 until 2013.Thus the number of games lost exhibited decreasing trend.

The Autocorrelation Function (ACF), the Partial Autocorrelation Function (PACF) and Extended Autocorrelation Function (EACF) are examined to determine a suitable model for the series.

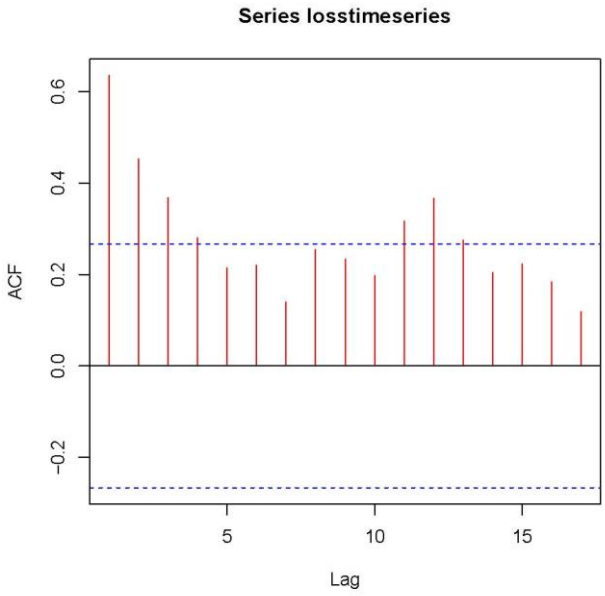

Figure 5: A plot dsecribing the ACF

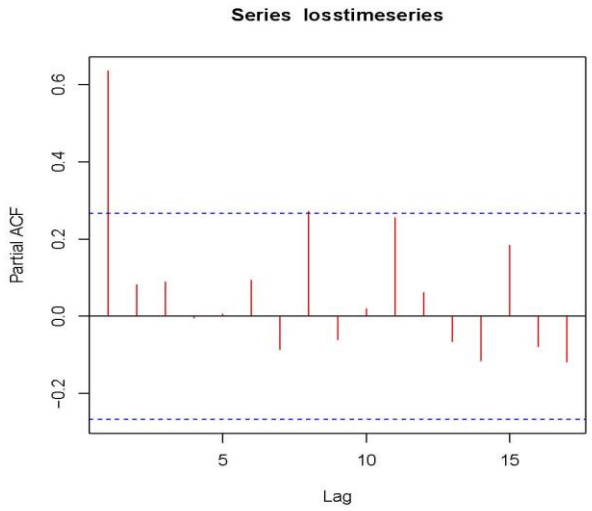

Figure 6: A plot showing the PACF

Observing figures 5 and 6 above, the PACF has cut off sharply at lag 1 and the ACF decay slowly. This suggests an Autoregressive model AR(1). Thus the ACF and PACF plots indicates ARIMA(1,0,0). There was no difference since the original series was stationary. The EACF diagram also suggest four models, that is ARIMA(0,0,1), ARIMA(2,0,1), ARIMA(1,0,1) and ARIMA(1,0,0). Thus these four models will be applied to the lost data and the best one selected. The best model must satisfy some key assumptions of the residuals. These assumptions are randomness, independence and normality of the residuals. The Akaike Information Criterion will also be compared for the various models.

3.5 Fitting the ARIMA(1,0,0) Model for the lost data

The best model selected among the four based on how well the model satisfied the criteria in the paragraph above is ARIMA(1,0,0). The model is of the form $\widehat{X}_t = \mu + \phi X_{t-1} + w_t$, where $\widehat{X}_t$ is the dependent variable, ϕ is the coefficient associated with the AR and w_t is the error term. The model parameters are estimated below.

Series: losstimeseries

ARIMA(1,0,0) with non-zero mean

Coefficients:

	ar1	intercept
	0.6552	8.7217
s.e.	0.1033	1.1514

sigma^2 estimated as 9.104: log likelihood=-136.54
AIC=279.08 AICc=279.56 BIC=285.05

The model parameter ϕ was estimated to be 0.6552 and μ was 8.7217. The standard error associated with ϕ and μ for the AR(1) model were estimated as 0.1033 and 1.1514 respectively. The Akaike Information Criterion (AIC) associated with the AR(1) was found to be 279.08. The Bayesian Information Criterion (BIC) was found to be 285.05.Thus the ARIMA(1,0,0) model was estimated to be $\widehat{X}_t = 8.7745 + 0.6552Y_{t-1} + w_t$

3.6 Checking the Validity of the Model

The model validity was checked by examining the residuals of the model. The assumption that the residuals should be random and normally distributed will be verified to confirm the validity of the model. The residuals of the model were plotted and the graph examined for randomness. The ACF plot of the rsesiduals was also examined as well as the normal probability plots.

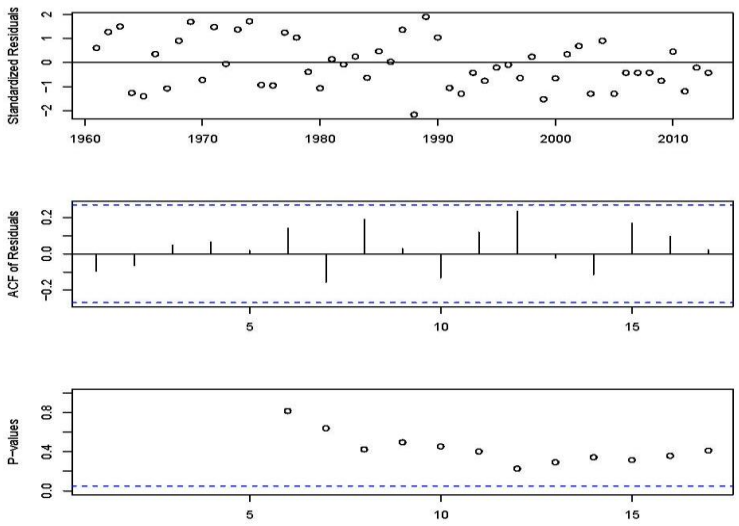

Figure 7: Residuals plots

The standardized residuals plot above shows the residuals are randomly distributed around zero. This indicates that the residual of the number of games lost by Manchester United over the seasons have constant variance.

The ACF plot of the residuals also suggests that there is no autocorrelation among the residuals of the number of games lost by Manchester United for the period considered for this project. It is an indication that there was no information lost and that every information was accounted for by the model.

3.7 Forecasting the number of games that will be lost using the model

Table 1: The predicted values for the next fifteen seasons

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2014	6.283057	2.416149	10.14996	0.36913134	12.19698
2015	7.123778	2.500678	11.74688	0.05335637	14.19420
2016	7.674661	2.762479	12.58684	0.16212700	15.18719
2017	8.035626	3.004421	13.06683	0.34106164	15.73019
2018	8.272148	3.190695	13.35360	0.50073685	16.04356
2019	8.427128	3.324254	13.53000	0.62295559	16.23130
2020	8.528680	3.416635	13.64072	0.71048232	16.34688
2021	8.595221	3.479244	13.71120	0.77100991	16.41943
2022	8.638822	3.521158	13.75649	0.81203044	16.46561
2023	8.667391	3.549003	13.78578	0.83949231	16.49529
2024	8.686112	3.567413	13.80481	0.85773695	16.51449
2025	8.698378	3.579545	13.81721	0.86979916	16.52696
2026	8.706416	3.587526	13.82531	0.87774904	16.53508
2027	8.711682	3.592768	13.83060	0.88297799	16.54039
2028	8.715133	3.596208	13.83406	0.88641276	16.54385

The ARIMA model developed for the number of games lost is used as a predictive model for making forecasts for future values of the number of games that will be lost by Manchester United in future games for various seasons. The predicted values for the next fifteen seasons are shown in Table 1.

The ARIMA model gives a forecast of the number of games that would be lost by Manchester United for the next fifteen seasons as well as 80% and 95% prediction intervals for those predictions. The last observed value in the original time series that is the number of games lost for 2013 was 5, and the ARIMA model gives the forecasted value of the number of games that would be lost for 2014 as 6.28 which is approximately 6 games.

4. Conclusion

The number of games lost by Manchester United Football Club for the study period depicts a downward trend. The ARIMA (1, 0, 0) was the model developed for predicting the future games that would be lost by Manchester United Football Club. All the model assumptions were satisfied. There was no autocorrelation among the residuals. The residuals were independent and also exhibited constant variance. The residuals were also normally distributed. The number of games that would be lost for Manchester United Football Club in 2013/2014 season was predicted to be six (6).

References

[1] McCleary, R., Hay, R.A. Jr. (1980). *Applied Time Series Analysis for the Social Sciences*. Beverly Hills and London: Sag, 328 pp.
[2] Shumway, R.H., Stoffer, D.S. (2005). *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer-Verlag New York, www.ici.ro/ici/revista/sic2000_4/art14.htm
[3] Cryer, J.D. (1986). *Time Series Analysis*. Duxbury Press, ISBN: 0871509636, 9780871509635
[4] Shumway, R.H., Stoffer, D.S. (2010). *Time Series Analysis and Its Applications with R Examples*. Third Edition, Springer Texts in Statistics. Springer-Verlag New York
[5] Shumway, R.H., Stoffer, D.S. (2008). *Time Series Analysis and Its Applications with R Examples*. Second Edition, Springer Texts in Statistics. Springer-Verlag New York