E-ISSN 2281-4612
ISSN 2281-3993

Academic Journal of
Interdisciplinary Studies

Vol 6 No 2
July 2017

DE GRUYTER
OPEN

Research Article

Forecasting Sports Popularity: Application of Time Series Analysis

Ryan Miller[1]

Harrison Schwarz[1]

Ismael S. Talke[1]*

[1]*Department of Information Systems and Analytics,*
Miami University Ohio, Oxford
**Corresponding Author*

Doi: 10.1515/ajis-2017-0009

Abstract

Popularity trends of the NFL and NBA are fun and interesting for casual fans while also of critical importance for advertisers and businesses with an interest in the sports leagues. Sports leagues have clear and distinct seasons and these have a major impact on when each league is most popular. To measure the popularity of each league, we used search data from Google Trends that gives real-time and historical data on the relative popularity of search words. By using search volume to measure popularity, the times of year, a sport is popular relative to its season can be explained. It is also possible to forecast how sport leagues are trending relative to each other. We compared and discussed three different univariate models both theoretically and empirically: the trend plus seasonality regression, Holt-Winters Multiplicative (HWMM), and Seasonal Autoregressive Integrated Moving Average (SARIMA) models to determine the popularity trends. For each league, the six forecasting performance measures used in this study indicated HWMM gave the most accurate predictions.

Keywords: *Sports, NFL, NBA, Regression, Holt-Winter, Seasonal ARIMA Models and forecasting.*

1. Introduction

Time series analysis is an approach to forecasting commonly used in business to produce and improve point forecasts where regression falls short (Tsay, 2000). Time series forecasting is increasingly in demand due to its ability to predict events based solely on previously observed data of the given event (Donate et al., 2013, Omar et al., 2016).

Studies have also been done showing that early patterns found in web popularity reflect long-term interest in a topic (Szabo and Huberman, 2010). In other business studies, search engine popularity has been shown to reflect general popularity and interest in a specific product (Omar et al., 2016). Our models apply this interest assumption, using major sports leagues in the United States as our product.

Forecasting has been a growing trend in the world of sports, where it has been used in an attempt to predict outcomes of games (Spann and Skiera, 2009). Our analysis focuses on a separate and more general area within sports, the popularity of entire leagues. The average NFL team is worth $2.3 billion and the average NBA team is worth $1.25 billion (Ozanian, 2016, Baden hausen, 2016). With such large market values, even small changes in future popularity could have large business implications on marketing, social media promotion, and team value.

In order to model sport popularity, we pulled data from Google Trends. Google Trends is an analytical tool that allows users to compare the popularity of search terms over time. Google Trends

can be used to gain insights into popularity that may not otherwise be noticed, as shown in the recent 2016 presidential election (Rogers, 2016). Data is available from 2004 to the present, and we chose to use the full range of data available to us. In this study, we filtered the data down to popularity only in the United States. Using the SAS Time Series Forecasting System, we were able to develop adequate models to forecast popularity.

Several application of univariate time series models have been conducted since the introduction of the methods. To mention some, time series models have been used in modeling: airline passengers, chemical process reading, oil price, counterfeiting crime data and others (Tularam and Saeed , 2016; Anand and Ekata , 2012, and Box et al., 2008). However, note that the best model found varies depending on the applicability and nature of the data.

The objective of this study is to compare and contrast NFL and NBA popularity using univariate time series forecasting models in order to efficiently predict the trend popularity for and between the two leagues in the Unites States. We wanted to make a confident prediction about which league is growing faster. We believe sport's popularity is tailor made for time series forecasting. Sports have very distinct seasons, which allowed us to build a seasonality component and trend into our models. The paper is organized as follows: Section 2 briefly describes the data set used in this study. In Section 3, the materials and methodology used are discussed and presented. Section 4, presents the measures of forecasting performance. In Section 5, the main results and model comparisons are presented, and final concluding remarks are given in Section 6.

2. Data and Description

Our data was sourced from the Google Trends website. This data shows how the popularity of a term has changed over time in Google searches. We looked at the specific search terms "NFL" and "NBA". To see the scores relative to each other, we used the compare feature on the website. The data was available from December 2003 onward at the monthly level, giving us 153 observations at the time of writing. We filtered the data down to searches from only the United States. The trends are scored using a relative index of 0-100, with 100 being the point at which the most popular term being compared peaked in popularity. A value of 50 is 50% as popular as the peak. In model building we held the last 3 months data: June, July and August 2016 for model validation purpose and the remaining 150 to build the model. Descriptive statistics and other results are discussed in detail in Section 5.

3. Materials and Methods

A time series is a sequence of observations measured at successive points in time. Generally, time series data consists of four components. These are trend (T), seasonality (S), cyclical (C) and Irregularity (noise) (I). To develop a forecasting model understanding these four components is crucial as it suggests which models to consider. The flow chart in Figure 1 shows the model depends on the time series components present in the data. This is similar to the idea that the type of data dictates the type of statistical models to be used. As is the case in most time series data the focus will be on T, S and I. That is, time series values at time t are often modeled as a function of these three components and depending on the seasonal fluctuation of the series, the model can be additive or multiplicative. That is,

$Y_t = T_t + S_t + I_t$ (Additive Model) -If seasonal fluctuation is constant

$Y_t = T_t \times S_t \times I_t$ (Multiplicative Model) -If seasonal fluctuation is not constant

where T_t, S_t and I_t respectively are the trend, seasonality and Irregularity at time t. For a detailed discussion of time series models see (Box et al., 2008; Bowerman et al., 2005; Box and Jenkins, 1980, and Montgomery et al., 2008). In this study three different models are considered, compared both theoretically and empirically. These are the time series regression model (Regression), Exponential Smoothing (ES) method, and seasonal ARIMA(p, d, q)(P, D, Q)$_m$ (SARIMA) models. The time series plot for monthly NFL and NBA data in Figure 2 exhibited trend and seasonality. As a result, in this study 3 univariate models will be presented: the Trend plus seasonality regression model, Holt-Winter Multiplicative Model (HWMM) and the seasonal

ARIMA(p, d, q)(P, D, Q)$_m$ (SARIMA) model. However, due to the non-constant seasonal variation present in the data, natural logarithmic transformation is used to stabilize the variation through out the three models.

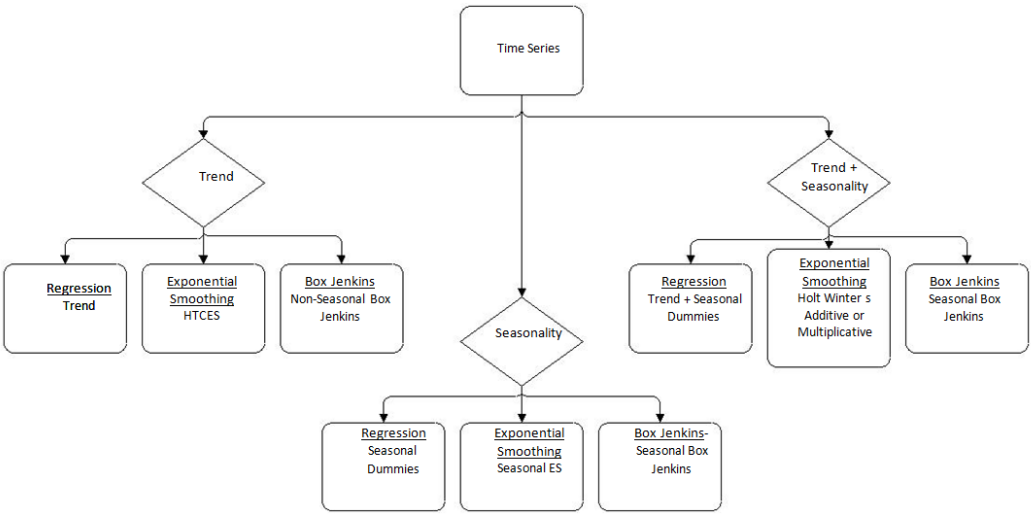

Figure 1: Forecasting Models

3.1 Model 1: Time series Regression Model

For the time series that exhibits trend and seasonality, the time series regression model fits

Additive Model (AM) $Y_t = T_t + S_t + \varepsilon_t$ - When Seasonal fluctuation is constant or

Multiplicative Model (MM) $Y_t = T_t \times S_t \times \varepsilon_t$ -When Seasonal fluctuation is not constant, where ε_t is the error term (Irregularity or Noise term)

For the seasonality factor seasonal dummies with december as the baseline is used and the trend can be linear or non-linear.

3.2 Model 2: Holt-Winters Multiplicative Model (HWMM)

Unlike the time series regression models, ES methods use weighted average by assigning unequal weights by introducing smoothing constants. There are several ES methods, for example, for a series that has no trend and seasonality, Simple Exponential Smoothing (SES) model is used, which is analogous to the average model in time series regression, uses a smoothing constant weight that assigns unequal weight to the remote and recent observations. For a series that has a trend component, the Holt-Trend Corrected Exponential Smoothing (HTCES) model is used which is analogous to the linear trend model and unequal weight is assigned to the remote and recent observations and trend as shown in the flow chart. The Holt-Winters (HW) model is an ES method for modeling a series that exhibits trend and seasonality is a function of three components: the level, trend (growth or slope), and seasonality components. The HW model may be additive or multiplicative depending on the nature of seasonal fluctuation. In this study as our data has an increasing seasonal fluctuation only the HWMM model is considered. The k step ahead point forecast for HWMM model is given by

$$F_{t+k}(t) = F_t = (L_t + kT_t)S_{t+k-m}$$

Where L_t is the level of the series, T_t is trend and S_t is the seasonality factor at time t and m is 12 for a monthly data. The equations for the estimated level, growth rate (trend) and seasonal factor respectively are given below

$L_t = \alpha(Y_t/S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1})$
$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$
$S_t = \delta(Y_t/L_t) + (1-\delta)S_{t-m}$
Where $\alpha, \gamma,$ and δ are smoothing constants between 0 and 1.

3.3 Model 3: Seasonal ARIMA (SARIMA) MODEL

Box and Jenkins introduced the ARIMA models in 1970. This type of models encompasses three classes of models, the Autoregressive (AR), Moving Average (MA) and Autoregressive Moving Average (ARMA) Models. In this study the focus is on SARIMA models. The general shorthand notation for SARIMA model is ARIMA(p, d, q)(P, D, Q)$_m$ Where p = order of the non-seasonal AR term, q = order of the non-seasonal MA term, d = order of non-seasonal differencing P = order of the seasonal AR term, Q = order of the seasonal MA term, D = order of seasonal differencing and m = number of seasons per year for monthly data m = 12.

ARIMA(p, d, q)(P, D, Q)m (SARIMA) model is a function of both the lagged series and the random shocks and the equation of the model is given as

$Z_t = \psi + \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{j=1}^{P} \Phi_j Z_{t-jm} + \sum_{h=1}^{q} \theta_h \varepsilon_{t-h} + \sum_{l=1}^{Q} \Theta_l \varepsilon_{t-lm} + \varepsilon_t$

where ψ is an intercept term and depends if the series has a non zero mean or not, ϕ_i , Φ_j , θ_h, and Θ_l are the coefficients of the non-seasonal AR, the Seasonal AR terms, the non-seasonal MA, and the Seasonal MA terms respectively.

SARIMA models depend on the pattern of the autocorrelation and partial autocorrelation functions and are based on 5 steps: stationary, model identification, estimation, diagnostics and forecasting. If the original series is not stationary, non-stationarity is re-moved by identifying the type of differencing and order of differencing required. This can be just the non-seasonal difference or seasonal difference or mixture of both of order 1 or more until stationary is achieved. For example, d = 1 is first order non-seasonal difference and is calculated as $Z_t = Y_t - Y_{t-1}$, where Y_t and Y_{t-1} are observations at time t and $t-1$ respectively, and D=1 is first seasonal difference and is calculated by $Z_t = Y_t - Y_{t-m}$. The Augmented Dickey Fuller (ADF) test is used for checking stationary condition.

4. Forecasting Performance Measures

The above-discussed models will be fit for the monthly NFL and NBA data, compared, and then best model will be selected. The overall performance of the models fitted will be measured using the following 6 measures of forecasting performance:

1. Mean Square Error (MSE): $=\frac{\sum_{i=1}^{n}\epsilon_i^2}{n}$ where ϵ_t is the difference between the observed values and predicted value at time t is the error at time t. That is, $\epsilon_t = Y_t - F_t$
2. Root Mean Square Error (RMSE): $= \sqrt{MSE}$
3. Mean Absolute Error (MAE): $= \frac{\sum_{i=1}^{n}|\epsilon_i|}{n}$
4. Mean Absolute Percentage Error (MAPE): $=\frac{100\times\sum_{i=1}^{n}\frac{|\epsilon_i|}{|Y_t|}}{n}$
5. R-Square Adjusted (R2-Adj): = 1- $\frac{(n-i)(1-R^2)}{n-k}$ where n =number of observations and parameters including intercept and i= 1 with intercept term and 0 with no intercept term.
6. Amemiyas Prediction Criterion (APC): = $\frac{(n-k)MSE}{n+k}$

The best model is the one that has the highest (maximum) Adjusted R-square and the lowest (minimum) values in the remaining five criteria. In this study, the first 150 months data are used to build the models and the last three months data are used for validation purpose to see the prediction power of the best model.

5. Results and Discussion

SAS Time series forecasting statistical software is used to fit the Regression, HWMM and SARIMA models. Once the outputs are obtained from SAS, Figures 2 and 3 are produced using MATLAB (R2015a) mathematical software. The SARIMA models require stationarity, and the Augmented Dickey-Fuller test indicated the original series wasn't stationary. As a result both regular and seasonal differencing were needed for both series to achieve sationarity. The autocorrelation structure from the Sample Autocorrelation (SAC) and Sample Partial Autocorrelation (SPAC) was used to determine the order of the seasonal and non-seasonal models (p, q, P, and Q) of the SARIMA model. The non-seasonal and seasonal models are combined to find the final best and adequate model. The best Regression, ES and SARIMA models are shown in Table 1 and 2 for NFL and NBA series respectively.

Table 1 and 2 shows the prediction performance of each of the three-univariate time series models using the 6 forecasting measures. This is empirical comparison of the three models for the NFL and NBA data. In all the six performance measures as shown in bold-faced figures, the log-HWMM model was consistently found to be the best for both the sports. In comparison SARIMA model was the next best model after log-HWMM and its performance is almost as good as the log-HWMM and the regression model performed the worst. Natural logarithmic transformation is used to stabilize the variability. The forecasting performance measures reported in this study for the estimation and validation periods are given in original units and it's due to this reason that the unlogged errors might seem a bit higher.

A time series plot for each model is included in Figure 2 and this graph shows how the models performed over time when dealing with different aspects of the data, such as the increasing seasonal variation and capturing the pattern and forecasting the future. Moreover, Figure 2 also depicts graphical comparison of the model performance by comparing the accuracy of the forecasted and actual popularity index score for the NFL and NBA sports. Figure 2 results is inline with what was obtained in Table 1 and 2. That is, the superior performance of the log-HWMM model is clearly visible as the forecasted popularity indexes from this model are close to the actual popularity index score. Note that the forecasted popularity index values in Figure 2 are obtained from the respective models shown in Table 1 and 2.

Table 3 shows further comparison of the models using the validation performance of the models for the three periods (June-July-August 2016) both for the NFL and NBA. The bold faced result shows once again the interval width from the Log-HWMM for the three periods has narrower prediction intervals indicating this model provides the most accurate forecast and makes it the best model. Figure 3 shows the comparison between the actual and the predicted popularity score index, the shaded region shows the 95% Interval from Log-HWMM model for both NFL and NBA. Figure 3 clearly shows the Log-HWMM model is best for both sports and the forecasted and actual popularity score indexes consistently falls within the 95% confidence interval.

Table 4 exhibits the year-over-year increase in search popularity from the data in Google Trends. As the color indicators show, NFL's popularity is decreasing from 2015 to 2016 in June, July, and August, while NBA's popularity is increasing in those months

Table 1: Forecasting Measures For NFL Models

Model	R^2-Adj	MSE	RMSE	MAE	MAPE	APC
Log-HWMM	**0.951**	**26.181**	**5.117**	**3.174**	**11.302**	**27.250**
Log-ARIMA(0; 1; 1)(0; 1; 2)$_{12}$	0.946	30.007	5.478	3.652	12.794	31.351
Log-Linear. T + S Regression.	0.901	48.915	6.994	4.874	15.770	58.198

Table 2: Forecasting Measures For NBA Models

Model	R^2-Adj	MSE	RMSE	MAE	MAPE	APC
Log-HWMM	**0.950**	**10.019**	**3.165**	**2.055**	**9.510**	**10.428**
Log-ARIMA(0; 1; 2)(0; 1; 1)$_{12}$	0.943	11.873	3.446	2.301	10.770	12.405
Log-Quadratic T + S Regression.	0.908	16.943	4.116	2.750	13.024	20.431

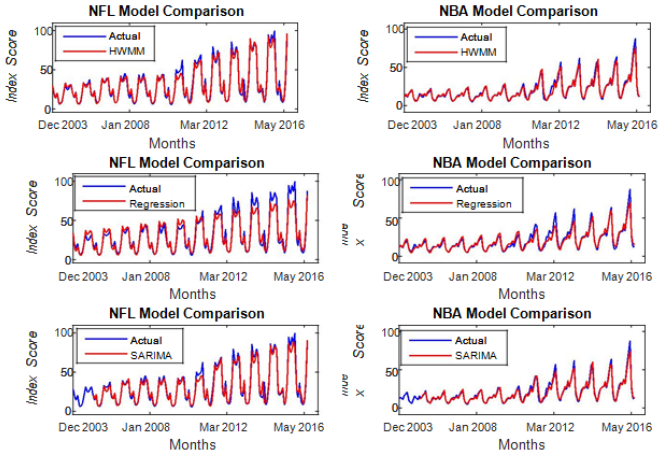

Figure 2: Actual Vs Predicted NFL & NBA Index Popularity from the Three Models

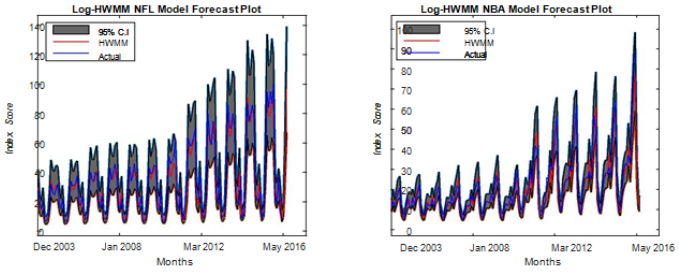

Figure 3: Actual Vs Forecasted Popularity & 95% C.I. from Log-HWMM Model for NFL & NBA

Table 3: NFL and NBA Models

NFL Models				
Models	Measures	Period 1	Period 2	Period 3
HWMM	95% C.I.	(9.4111,19.3043)	(28.0553,57.7699)	(67.2688,139.2015)
	Predicted	13.4787	40.2585	96.7673
	Actual	12	35	88
	Interval Width	**9.8932**	**28.0583**	71.9327
SARIMA	95% C.I.	(9.7791,20.8161)	(27.2807,58.2928)	(64.5793,138.5164)
	Predicted	14.2675	39.8782	94.5796
	Actual	12	35	88
	Interval Width	11.037	31.0121	73.9371
Regression	95% C.I.	(10.9155,24.7606)	(27.1639,61.6182)	(55.2615,125.3544)
	Predicted	16.4400	40.9120	83.2302
	Actual	12	35	88
	Interval Width	13.8451	34.4543	**70.0929**
NBA Models				
Models	Measures	Period 1	Period 2	Period 3
HWMM	95% C.I.	(20.4088,33.9777)	(10.0708,17.3751)	(9.0071,16.2908)
	Predicted	26.3333	13.2280	12.1133
	Actual	26	12	13
	Interval Width	**13.5689**	**7.3043**	**7.2827**
SARIMA	95% C.I.	(21.7784,38.0957)	(10.0189,19.5741)	(9.1977,18.4631)
	Predicted	28.8039	14.0040	13.0314
	Actual	26	12	13
	Interval Width	16.3173	9.5552	9.2654
Regression	95% C.I.	(20.1233,39.4610)	(11.7378,23.0174)	(10.7572,21.0944)
	Predicted	28.1795	16.4370	15.0638
	Actual	26	12	13
	Interval Width	19.3377	11.2796	10.3372

Table 4: Increase (Decrease) In Search Popularity For NFL and NBA

Time	NFL	NBA
June 2015 June 2016	7:7%	36:8%
July 2015 July 2016	7:9%	33:3%
Aug 2015 Aug 2016	7:4%	18:2%

6. Conclusion

The main focus of this paper was to use time series data from Google Trends to predict the future popularity for the search terms, "NFL" and "NBA". The models we used were Log-HWMM, Log-SARIMA, and Log-trend plus seasonal dummies regression models for both the NFL and the NBA monthly data. The Log-HWMM model provided more accurate forecast compared to Log-SARIMA, and Log-trend plus seasonal dummies regression models for both the NFL and NBA search popularity. In addition to the forecasts, the actual search popularity data for each sports league is in line with recent news that NFL TV ratings are down (Gillette, 2016) and NBA TV ratings are up (Ben, 2016). This is clearly presented in Table 4. The forecasts could make a difference for NFL and NBA business interests due the huge amount of money involved with the leagues. Small percentage differences in popularity could mean thousands more people looking at ads, thousands more people buying merchandise, and thousands more dollars in profits. Businesses interested in advertising or marketing to or investing with either league may find these forecasts useful for deciding which sports league provides the greater short-term or long-term value. We would encourage further work using time varying models, such as the Generalized Autoregressive Conditional Heteroskedasticity (GARCH), and State Space Models.

References

Anand, K.S. and Ekata, (2012). Applicability of box Jenkins Arima model in crime forecasting: A case study of counterfeiting in gujarat state. International Journal of Advanced Research in Computer Engineering and Technology Volume 1, Issue 4, [Online] Available: http://ijarcet.org/?p=1309.

Badenhausen, Kurt., (2016). New york knicks head the NBA's most valuable teams at $3 billion. forbes. [Online] Available: http://www.forbes.com/nba-valuations/.

Ben, Cafardo, (2016). NBA on ESPN: Overnight ratings for Nov. 4 prime-time doubleheader up 42 percent espn. [Online] Available: http://espnmediazone.com/us/press-releases/2016/11/nba-espn-overnight-ratings-nov-4-prime-time doubleheader-42-percent

Bowerman, B.L., OConnell, R.T., and Koehler, A.B., (2005). Forecasting, Time Series, and Regression: an Applied Approach. 4th Edition.

Box, G.E.P. and Jenkins, G.M., (1980) Time Series Analysis: Forecasting and control. Upper Saddle River, NJ: Prentice Hall.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C., (2008). Time Series Analysis: Forecasting and control. 4th ed. New York: Wiley.

Donate, J.P., Li, X., Sánchez, G.G. et al., (2013) Neural Comput. & Applic. 22: 11. doi:10.1007/s00521-011-0741-0

Gillette, Felix (2016). NFL was a sure thing for tv networks. until now, Bloomberg,. [Online] Available: http://www.bloomberg.com/news/articles/2016-11-03/nfl-was-a-sure-thing-for-tv-networks-until-now.

Spann, M. and Skiera, B., (2009), Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. J. Forecast., 28: 55–72. doi:10.1002/for.1091

Motgomery, D.C., Jennings, C.L., and Kulachi M., (2008). Introduction To Time Series Analysis and Forecasting. 2nd Edition Wiley.

Hani Omar, Van Hai Hoang, and Duen-Ren Liu, (2016). "A Hybrid Neural Network Model for Sales Forecasting Based on ARIMA and Search Popularity of Article Titles," Computational Intelligence and Neuroscience, vol. 2016, Article ID 9656453, 9 pages, 2016. doi:10.1155/2016/9656453

Ozanian, Mike.(2016) N valuations. forbes,. Forbes 4 Sept. 2016. [Online] Available: https://www. forbes.com/nfl-valuations/#6eee8b4946cb/.

Rogers, James (2016). Did google search data provide a clue to trump's shock election victory? FOX News Network, [Online] Available: http://www.foxnews.com/ tech/2016/11/09/.

Szabo, G. and B.A. Huberman (2010). Predicting the popularity of online content. Communications of the ACM,

vol. 53, no. 8,pp. 80-88. [Online] Available: https://www.scopus.com/ record/display.uri?eid=2-s2.0-77955233534&origin=inward&txGid=0.

Tsay, Ruey S.,(2000). Time series and forecasting: Brief history and future research. Journal of the American Statistical Association, vol. 95, no. 450, 2000, pp. 638-43.

Tularam, G.A. and Saeed, T.,(2016) Oil-price forecasting based on various univariate time-series models. American Journal of operations Research, 6, 226-235. [Online] Available: http: //dx.doi.org/10.4236/ajor.2016.63023.