Application of Time Series Analysis: Research Trends and Prediction Evolutions

D.Vishwaksen Reddy[1], K.Pujitha Reddy[2], A.Aparna[3]
[1, 3] Dept of IT
[2]Dept of CSE
[1, 2, 3] Keshav Memorial Institute of Technology

Abstract- *In almost every scientific field, measurements are performed over time. These observations lead to a collection of organized data called time series. The purpose of time-series data mining is to try to extract all meaningful knowledge from the shape of data. Even if humans have a natural capacity to perform these tasks, it remains a complex problem for computers. In this paper we intend to provide a survey of the techniques applied for time-series data mining. The first part is devoted to an overview of the tasks that have captured most of the interest of researchers. Considering that in most cases, time-series task relies on the same components for implementation, and the study of Regression analysis ARIMA models, the relevant literature has been categorized for each individual aspects. Finally, the study submits various research trends and avenues that can be explored in the near future. We hope that this article can provide a broad and deep understanding of the time-series data mining research field.*

Keywords- Regression, ARIMA Models and forecasting, Research Trends.

I. 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 Trendscan 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.

1.1 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.

1.2 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 + S + I (Additive Model) -If seasonal fluctuation is constant
Y= T × S × I (Multiplicative Model) -If seasonal fluctuation is not constant Y

WhereT,S and I respectively are the trend, seasonality and Irregularity at time t. For a detailed discussion of time series models see (Box et al., 2008; Bower man 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.

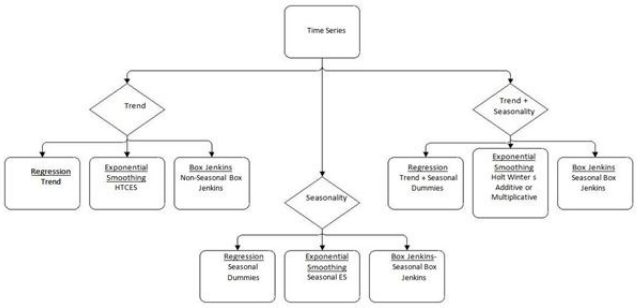


Figure 1: Forecasting Models

Model 1: Time series Regression Model

For the time series that exhibits trend and seasonality, the time series regression model fits
Additive Model (AM) $Y = T + S + \varepsilon$ - When Seasonal fluctuation is constant or
Multiplicative Model (MM) $Y = T \times S \times \varepsilon$ -When Seasonal fluctuation is not constant, whereis the error term (Irregularity or Noise term)

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

$(t) = F = (L + kT)S$

Where L is the level of the series, T is trend and S 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

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

$$Z = \psi + \sum\phi\ Z + \sum\Phi\ Z + \sum\ \theta\ \varepsilon + \sum\ \Theta\ \varepsilon + \varepsilon$$

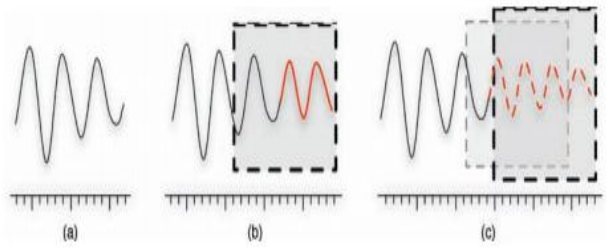
where ψ is an intercept term and depends if the series has a non zero mean or not, ϕ ,, θ , and Θ are the coefficients of the non-seasonal AR, the Seasonal AR terms, the non-seasonal MA, and the Seasonal MA terms respectively.

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. Forexample, d = 1 is first order non-seasonal difference and is calculated as $Z = Y - Y$, where Y and Y are observations at time t and $t - 1$ respectively, and D=1 is first seasonal difference and is calculated by $Z = Y - Y$. The Augmented Dickey Fuller (ADF) test is used for checking stationary condition.

II. PROPOSED MODEL

2.1. Prediction

Time series are usually very long and considered smooth, that is, subsequent values are within predictable ranges of one another [Shasha and Zhu 2004]. The task of prediction is aimed at explicitly modeling such variable dependencies to forecast the next few values of a series. Figure 5 depicts various forecasting scenarios. Definition 3.10. Given a time series T = (t1,...tn), predict the k next values (tn+1,...,tn+k) that are most likely to occur. Prediction is a major area in several fields of research. Concerning time series, it is one of the most extensively applied tasks. Literature about this is so abundant that dozens of reviews can focus on only a specific field of application or family of learning methods. Even if it can use time-series representations and a notion of similarity to



A typical example of the time-series prediction task. (a) The input time series may exhibit a periodical and thus predictable structure. (b) The goal is to forecast a maximum number of upcoming datapoints within a prediction window. (c) The task becomes really hard when it comes to having *recursive prediction*, that is, the long-term prediction of a time series implies reusing the earlier forecast values as inputs in order to go on predicting.

evaluate accuracy, it also relies on several statistical components that are out of the scope of this article, for example, model selection and statistical learning. This task will be mentioned because of its importance but the interested reader willing to have further information may consult several references on forecasting [Brockwell and Davis 2002, 2009; Harris and Sollis 2003; Tsay 2005]. Several methods have been applied to this task. A natural option could be AR models [Box et al. 1976]. These models have been applied for a long time to prediction tasks involving signal denoising or dynamic systems modeling. It is, however, possible to use more complex approaches such as neural networks [Koskela 2003] or cluster function approximation [Sfetsos and Siriopoulos 2004] to solve this problem. A polynomial architecture has been developed to improve a multilayer neural network in Yadav et al. [2007] by reducing higherorder terms to a simple product of linear functions.

III. IMPLEMENTATION COMPONENTS

In this section, we review the implementation components common to most time-series mining tasks. As said earlier, the three key aspects when managing time-series data are representation methods, similarity measures, and indexing techniques. Because of the high dimensionality of time series, it is crucial to design low-dimensional representations that preserve the fundamental characteristics of a series. Given this representation scheme, the distance between time series needs to be carefully definedin order to

exhibit perceptually relevant aspects of the underlying similarity. Finally the indexing scheme must allow to efficiently manage and query evergrowing massive datasets. 4.1. Preprocessing In real-life scenarios, time series usually come from live observations [Reeves et al. 2009] or sensors [Stiefmeier et al. 2007] which are particularly subject to noise and outliers. These problems are usually handled by preprocessing the data. Noise filtering can be handled by using traditional signal processing techniques like digital filters or wavelet thresholding. In Himberg et al. [2001b], Independent Component Analysis (ICA) is used to extract the main mode of the series. As will be explained in Section 4.2, several representations implicitly handle noise as part of the transformation. The second issue concerns the scaling differences between time series. This problem can be overcome by a linear transformation of the amplitudes [Goldin and Kanellakis 1995]. Normalizing to a fixed range [Agrawal et al. 1995] or first subtracting the mean (known as zero mean/unit variance [Keogh et al. 2001a]) may be applied to both time series, however, it does not give the optimal match of two series under linear transformations [Argyros and Ermopoulos 2003]. In Goldin et al. [2004] the transformation is sought with optional bounds on the amount of scaling and shifting. However, normalization should be handled with care. As noted by Vlachos et al. [2002], normalizing an essentially flat but noisy series to unit variance will completely modify its nature and normalizing sufficiently small subsequences can provoke all series to look the same [Lin and Keogh 2005]. Finally, resampling (or uniform time warping [Palpanas et al. 2004a]) can be performed in order to obtain series of the same length [Keogh and Kasetty 2003]. Downsampling the longer series has been shown to be fast and robust.

Time series are essentially high-dimensional data. Defining algorithms that work directly on the raw time series would therefore be computationally too expensive. The main motivation of representations is thus to emphasize the essential characteristics of the data in a concise way. Additional benefits gained are efficient storage, speedup of processing, as well as implicit noise removal. These basic properties lead to the following requirements for any representation: —significant reduction of the data dimensionality; —emphasis on fundamental shape characteristics on both local and global scales; —low computational cost for computing the representation; —good reconstruction quality from the reduced representation; —insensitivity to noise or implicit noise handling. Many representation techniques have been investigated, each of them offering different trade-offs between the properties listed before. It is, however, possible to classify these approaches according to the kind of transformations applied. In order to perform such classification, we follow the taxonomy of Keogh et al. [2004] by dividing representations into three categories, namely nondata adaptive, data adaptive, and model based. 4.2.1. Nondata Adaptive. In nondata-adaptive representations, the parameters of the transformation remain the same for every time series regardless of its nature.

IV. RESEARCH TRENDS AND ISSUES

Time-series data mining has been an ever growing and stimulating field of study that has continuously raised challenges and research issues over the past decade. We discuss in the following open research issues and trends in time-series data mining for the next decade.

Stream analysis. The last years of research in hardware and network research have witnessed an explosion of streaming technologies with the continuous advances of bandwidth capabilities. Streams are seen as continuously generated measurements that have to be processed in massive and fluctuating data rates. Analyzing and mining such data flows are computationally extreme tasks. Several papers review research issues for data streams mining [Gaber et al. 2005] or management [Golab and Ozsu 2003]. Algorithms designed for static datasets have usually not been sufficiently optimized to be capable of handling such continuous volumes of data. Many models have already been extended to control data streams, such as clustering [Domingos and Hulten 2000], classification [Hulten et al. 2001], segmentation [Keogh et al. 2003a], or anomaly detection [Chuah and Fu 2007]. Novel techniques will be required and they should be designed specifically to cope with the ever flowing data streams.

Convergence and hybrid approaches: A lot of new tasks can be derived through a relatively easy combination of the already existing tasks. For instance, Lian and Chen [2007] proposed three approaches, polynomial, DFT, and probabilistic, to predict the unknown values that have not fed into the system and answer queries based on forecast data. This approach is a combination of prediction (refer to Section 3.5) and query by content (refer to Section 3.1) over data streams. This work shows that future research has to rely on the convergence of several tasks. This could potentially lead to powerful hybrid approaches.

Data mining theory and formalization. A formalization of data mining would drastically enhance potential reasoning on design and development of algorithms through the use of a solid mathematical foundation. Faloutsos and Megalooikonomou [2007] examined the possibility of a more general theory of data mining that could be as useful as relational algebra is for database theory. They studied the link between data mining and Kolmogorov complexity by showing

their close relatedness. They conclude from the undesirability of the latter that data mining will never be automated, and therefore stating that "data mining will always be an art". However, a mathematical formalization could lead to global improvements of both reasoning and the evaluation of future research in this topic.

V. CONCLUSION

After almost two decades of research in time-series data mining, an incredible wealth of systems and algorithms has been proposed. The ubiquitous nature of time series led to an extension of the scope of applications simultaneously with the development of more mature and efficient solutions to deal with problems of increasing computational complexity. Time-series data mining techniques are currently applied to an incredible diversity of fields ranging from economy, medical surveillance, climate forecasting to biology, hydrology, genetics, or musical querying. Numerous facets of complexity emerge with the analysis of time series, due to the high dimensionality of such data, in combination with the difficulty to define an adequate similarity measure based on human perception.

As for most scientific research, trying to find the solution to a problem often leads to raising more questions than finding answers. We have thus outlined several trends and research directions as well as open issues for the near future. The topic oftime-series data mining still raises a set of open questions and the interest of such research sometimes lies more in the open questions than the answers that could be provided.

REFERENCES

[1] Antunes, C. and Oliveira A. (2001) Temporal Data Mining: An Overview, in: Proceedings of the Workshop on Temporal Data Mining at the 7th International Conference on Knowledge Discovery and Data Mining, ACM Press, San Francisco, California, pp. 1-13.

[2] Chung, L., Fu, T. C. and Luk, R. (2004) An evolutionary approach to pattern-based time series segmentation, IEEE Transactions on Evolutionary Computation, IEEE Press, Vol. 8, Issue 5, pp. 471-489.

[3] ZHANG, X., WU, J., YANG, X., OU, H., AND LV, T. 2009. A novel pattern extraction method for time series classification. Optimiz. Engin. 10, 2, 253–271

[4] POVINELLI, R., JOHNSON, M., LINDGREN, A., AND YE, J. 2004. Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. Knowl. Data Engin. 16, 6, 779–783.

[5] 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).

[6] 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).

[7] 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-overnightratings-nov-4-prime-time doubleheader-42-percent Bowerman, B.L., OConnell, R.T., and Koehler, A.B., (2005).

[8] Forecasting, Time Series, and Regression: an Applied Approach. 4th Edition. Box, G.E.P. and Jenkins, G.M., (1980)

[9] Time Series Analysis: Forecasting and control. Upper Saddle River, NJ: Prentice Hall. Box, G.E.P., Jenkins, G.M., and Reinsel, G.C., (2008).

[10] Time Series Analysis: Forecasting and control. 4th ed. New York: Wiley. Donate, J.P., Li, X., Sánchez, G.G. et al., (2013)

[11] 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),

[12] 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).

[13] 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.