# Time Series Forecasting

Aditya N. Govardhan

**Abstract: In this report, time series data is forecasted using Autoregressive Integrated Moving Average (ARIMA) model. A equally intervaled univariate time series data is provided, which is used for forecasting the future values. It has been found that as compared to other methods of linear regression, ARIMA model is more flexible since it enables stochastic modelling because of the MA part. Proper selection of model parameters can lead to better forecasting, however the forecast is always as good as the available data. This report aims to put light on the preliminary design aspects to be taken into consideration while creating time series forecasts. These aspects can be applied to time series of varying complexity and form the basis of solving forecasting problems using non-linear techniques as well.**

***Index Terms*—forecasting, regression, stationarity, ARIMA, autocorrelation**

## I. Introduction

Time series forecasting is one of the important classes of regression problems in which future values are predicted based upon available timed data. Univariate time series forecasting is one of the representative problems of this class in which past data varies over only a single variable and is used to predict future values. Thus the forecasted value at time t depends upon the previous values of time t-1, t-2, t-3, ..., t-N. Models with linear parameters are used to carry out linear regression and can be used to forecast univariate data with good accuracy.

Autoregressive Integrated Moving Average (ARIMA) models are on the linear regression models which take advantage of ordered data in time series as well as provided flexibility due to its stochastic component. This report analyses the aspects of implementing ARIMA model for analyzing univariate time series data.

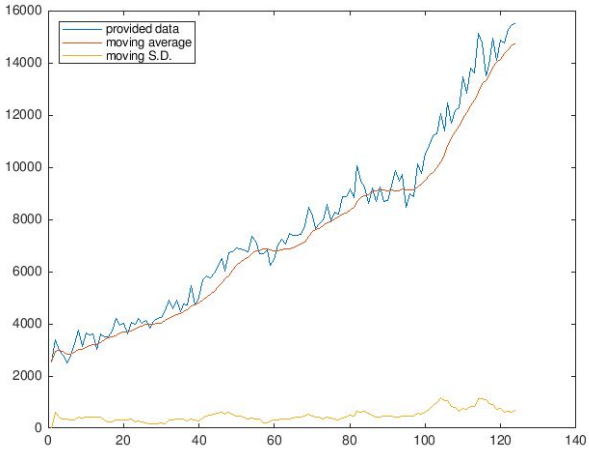


Fig. 1. Provided Data and its Characteristics

Aditya N. Govardhan is with the Arizona State University, Tempe, AZ 85281 USA (e-mail: Aditya.Govardhan@asu.edu).

## II. Data analysis and transformation

### A. Time series components and analysis

In this setup, 124 data points are provided which are not timed by default hence it can be assumed that it is evenly spaced with width of one day (other intervals are equally valid). The aim is to predict the next 20 values based on provided data. Line graph in fig. 1 with legend 'provided data' shows the variation of data.

A time series can generally be separated into the following components: (a) trend, (b) seasonal, (c) cyclic and (d) irregular. The trend component of a time series dictates how the graph increases or decreases continuously at a macro level. This trend could be linear, quadratic, logarithmic or some other variation. Seasonal component projects how data fluctuates within a given time period and this pattern is repeated for every time period. Cyclic component decides what pattern is followed in data over a macro level while irregular component dictates the random nature of data which is easily modelled using random processes.

General observation of provided data shows an increasing trend. This could also be observed from the moving average of fig. 1 over ten data points at a time. No specific seasonal or cyclic variation can be observed. However, irregular component is dominant which can be modelled using normally distributed values with zero mean and arbitrary standard deviation, since the moving standard deviation varies over a small range as shown in fig. 1.

### B. Stationarity and data transformation

Stationarity of data is a requirement for most of the linear prediction models. Stationary time series have (a) time independent mean (b) time independent variance and (c) correlation between any two instances is always equal given the same time difference is considered. Without stationarity, time series models can't be built. However, for practical purposes, strong stationarity is not required. If the time

series data is not stationary, it can be transformed to a weakly stationary series data set using various methods like detrending, differencing, log transformations and others.

The given time series data is clearly not stationary since the mean is increasing over time. To mitigate this, first difference is taken such that:

$$y'(t) = y(t) - y(t-1)$$

As a result, the differentiated dataset is reduced to 123 data points. Line graphs in fig. 2. indicate the dataset and its characteristics with first order of difference. As it can be seen, the mean has become fairly time independent and so has the standard deviation. The following section also comments about the autocorrelation of this dataset for varying time intervals. Thus obtaining a first order difference of the data provides us with fairly stationary data which can be used for time series modelling.

To verify the stationarity of obtained dataset, Dickey Fuller Test of stationarity can be carried out. An output value of 1 indicates that the unit root null hypothesis (i.e. the dataset is stationary) is rejected in favour of the alternative model (i.e. the dataset is stationary).

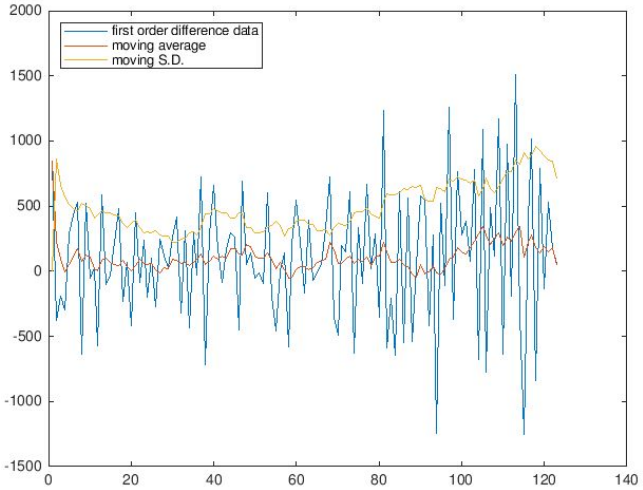


Fig. 2. Characteristics of First Order Differentiated Data

## III. Time series modelling and forecasting

Once the time series data is made stationary, it is ready to be modelled using a linear regression model. The model used in this forecasting is Autoregressive Integrated Moving Average (ARIMA) model. This model consists of three parts: (a) Autoregressive part, (b) Integrating part and (c) Moving Average part.

Autoregressive part can be explained conceptually using the following equation:

$$y(t) = \alpha y(t-1) + err(t)$$

It indicates that y(t) depends upon y(t-1) and the effect of this previous value is controlled by parameter α. This idea can be extended generally to the following equation:

$$y(t) = \alpha_1 y(t-1) + \alpha_2 y(t-2) + \ldots + \alpha_p y(t-p) + err(t)$$

Thus, autoregressive part is determined by p value and the coefficients $\alpha_1$, $\alpha_2$, …, $\alpha_p$ are estimated using the previous data. This is known as AR(p) model. Thus an AR(p) model is assumed to be a linear combination of p previous values along with a random error.

A similar argument considered for the moving average part. Conceptually a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations:

$$y(t) = err(t) + \beta_1 err(t-1) + \ldots + \beta_q err(t-q)$$

Thus, moving average part is determined by the value of q and $\beta_1, \ldots, \beta_q$ are estimated using the previous data. This is known as MA(q) model.

For integrator part, it's the order of difference that is required to make the data stationary. We have already determined the value of d while making the data stationary, since we are taking first order of difference, the value of d is set to 1.

Determining values of p and q parameters requires analysis of Autocorrelation Function graph (ACF) (shown in fig. 3.) and Partial Autocorrelation Function graph (PACF) (shown in fig. 4.). The point at which PACF graph falls below positive confidence interval (0.2 in our case) is considered to be the value of p and the point at which ACF graph falls below positive confidence interval is considered to be the value of q. A closer observation reveals that the value of p needs to be set at 4 and the value of q needs to be set at 2. Thus our ARIMA model is configured with parameters (p, d, q) valued (4, 1, 2).

Using this model, the parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2$ are determined using the estimate() function. Previous 124 data points are used for estimating these parameters.

Using the obtained model, the remaining 20 values are forecasted. Fig. 5. shows a graph of forecasted values.

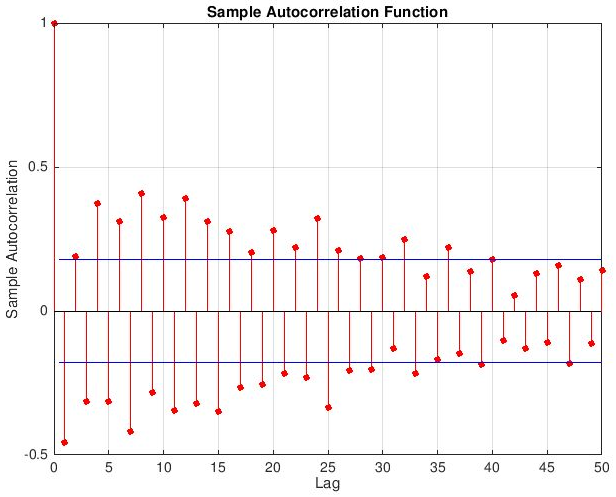


Fig. 3. Autocorrelation Function (ACF)

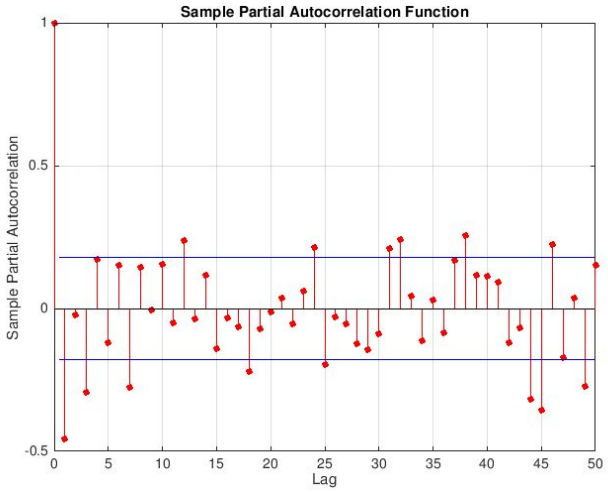


Fig. 4. Partial Autocorrelation Function (PACF)

## IV. Conclusion

Fig. 6. plots the given data (data1) versus estimated values by the model (ARIMA_data1). Fig 7. shows the forecasted value for different values of (p, d, q) parameters.

Components of time series and its stationarity play a major role in forecasting its values. Also, a parsimonious attitude towards selecting the number of parameters to be estimated is useful in building simple models. ARIMA models are linear regression models used in modelling complex time series. Its variation, Seasonal ARIMA (SARIMA) can also be used to take into consideration seasonality of data. These models find number of applications in financial domain. Finally, it has to be noted that non-linear modelling techniques can give better forecast values at the cost of complex algebraic equations.

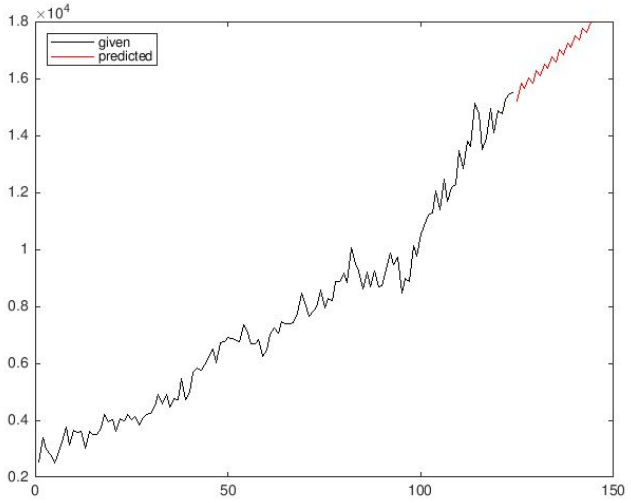


Fig. 5. Predicted Values (20 data points)

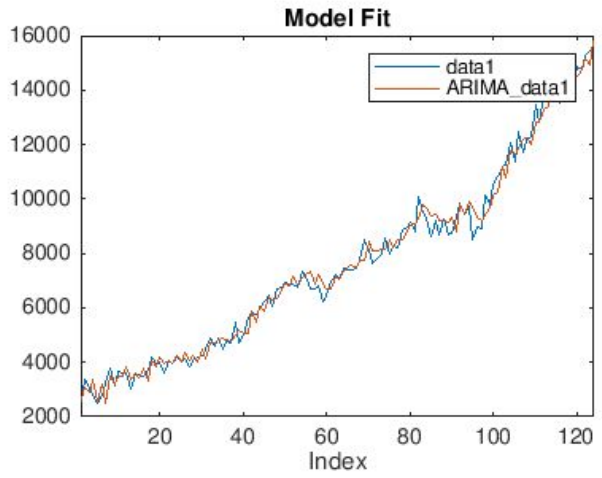


Fig. 6. Given Data vs. Estimated Data

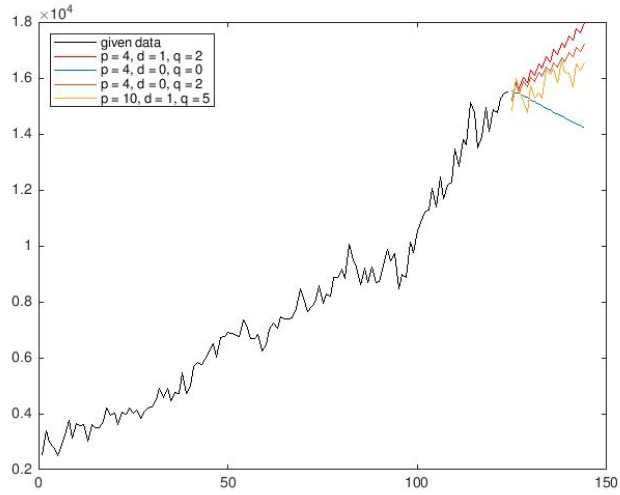


Fig. 7. Forecasts for different values of (p,d,q) parameters

## References

[1] MATLAB Econometric Toolbox Documentation

[2] Adhikari, R., & Agrawal, R.K. (2013). An Introductory Study on Time Series Modeling and Forecasting. CoRR, abs/1302.6613.