# INTRODUCTION TO TIME SERIES

## **What is a time series?**

A time series is a collection of observations made subsequently in time. It is an arrangement of statistical data in accordance with their occurrence in time.

When the observations or measurements are made on one variable then we have a univariate time series. When the variables are more, it is a multi variate time series.

Time series is applied in various fields such as;

* **Economics**. The stock market lists different stocks and shares of companies by time. Monitoring of various price indices as well as income of individuals, organizations and even countries.
* **Meteorology**. Weather phenomena such as rainfall, wind speed humidity, etc. can be monitored and described using time series.
* **Demography**. This is the scientific study of population and their characteristics. Population data can be studied using knowledge of time series.

## **Why study time series**

The objectives of time series are as follows;

1. **Description**. To understand various parameters attributing to the current state of the data. Some of these parameters are; trend and seasonal variations.
2. **Explanation**. When observations are taken on two or more variables, it may be possible to use the variables in one time series to explain the variation in another.
3. **Prediction**. Current observed values van be used to forecast future values and thus influence decision making.
4. **Control**. We may study a time series to understand the factors in play and therefore take measures to bring them into order. An example is in manufacturing where we a time series measuring quality of the process may be generated.

## **Characteristics of a time series**

1. Trend. Do the measurements tend to decrease or increase over time?
2. Seasonality. This refers to the regular repeating patterns of highs and lows related to calendar time such as seasons, quarters, months, days of the week etc.
3. Constancy of variance. Is the difference between observations constant?
4. Outliers. These are values way beyond the rest of the other. They may be too high or too low.
5. Abrupt changes to the series. Some time series may have abrupt changes due to sudden change in the variances of the variables.

## **Time domain models**

There are two basic types of time domain models;

* Models that relate the present value of a series to past values and past prediction errors - these are called ARIMA models (for Autoregressive Integrated Moving Average). We’ll spend substantial time on these.
* Ordinary regression models that use time indices as x-variables. These can be helpful for an initial description of the data and form the basis of several simple forecasting methods.

## **Some important terminologies in time series**

1. **Continuous time series**- A time series is said to be continuous when observations are made continuously in time. E.g., daily record of temperature of a location.
2. **Discrete time series-** When observations are taken only at specific times, usually equally spaced, then the time series is said to be discrete.
3. **Deterministic time series-** When successive observations are dependent, the future values may be predicted from first observations. A time series is said to be deterministic if it can be correctly predicted by first observation.
4. **Stochastic time series-** most time series are stochastic in that future values are only partly determined by past values, so that exact predictions are impossible and must be replaced by the idea that future values have a probability distribution which is conditioned by knowledge of past values.
5. **Stationary time series-** A time series is stationary if there is no systematic change in the mean, i.e. No trend if there is no systematic change in variation and if periodic variation is removed.

**Strict vs Weak stationary time series**

* A time series is strictly stationary if the joint distribution of;

Is the same as the joint distribution of;

*Where h is the distance between observations. (Lags)*

The probability of the process does not change with time.

* A time series is said to be weakly stationary if it satisfies the following properties:

1. The mean, is the same for all t.
2. The variance of Xt is the same for all t.
3. The covariance (and also correlation) between Xt and Xt-h is the same for all at each lag (1, 2, 3, …).
4. **Autocorrelation Function (ACF)-** Let Xt denote the value of a time series at time t. The ACF of the series gives correlations between Xt and Xt-h for h = 1, 2, 3, etc.

The equation above is as a result of weakly stationarity property.

1. **Partial Autocorrelation Function (PACF)-** It is a conditional correlation. It is the correlation between two variables under the assumption that we know and take into account the values of some other set of variables. For instance, consider a regression context in which y is the response variable and x1, x2, and x3 are predictor variables. The partial correlation between y and x3 is the correlation between the variables determined taking into account how both y and x3 are related to x1 and x2. The partial correlation is defined by the formula;

## **Autoregressive Integrated Moving Average Models (ARIMA)**

1. Autoregressive Model (AR)

These models use a linear model to predict the value at the present time using the value at the previous time. The order of the model indicates how many previous times we use to predict the present time.

**Autoregressive model of order 1 (AR1)**

We write the statistical equation of the model as;

σThis model satisfies the following assumptions;

1. meaning that the errors are independently distributed with a normal distribution that has mean 0 and constant variance.
2. Properties of the errors wt are independent of x.
3. The series X1, X2, ... is weakly stationary.

This is essentially the ordinary simple linear regression equation, but there is one difference. In ordinary least squares regression, we assume that the x-variable is not random but instead is something we can control.

**Properties of AR (1) model**

**Mean**

*Proof*

Recall stationarity property that the mean is the same for all functions of Xt. Therefore, we have

Hence;

**Variance**

*Proof*

Recall stationarity property that the variance is the same for all functions of Xt. Therefore, we have

Hence;

**ACF**

*Where h is the lags*

*Proof*

To start, assume the data have mean 0, which happens when δ=0, and . In practice this isn’t necessary, but it simplifies matters. Values of variances, covariances and correlations are not affected by the specific value of the mean.

Let , the covariance observations h time periods apart (when the mean = 0). Let  = correlation between observations that are h time periods apart.

Covariance and correlation between observations one time period apart.

Covariance and correlation between observations *h* time periods apart

To find the covariance yh, multiply each side of the model for xt by xt−h, then take expectations.

If we start at y1, and move recursively forward we get;

But;

Therefore

The correlation is therefore given by;

1. Moving Average Models

A moving average term in a time series model is a past error (multiplied by a coefficient). Let ,meaning that the wt are identically, independently distributed, each with a normal distribution having mean 0 and the same variance.

The 1st order MA is written as;

The 2nd order MA is written as;

The qth order MA is written as;

**Properties of MA (1) Model**

**Mean**

*Proof*

**Variance**

*Proof*

**ACF**

And

For lags h≥ 2

*Proof*

Consider the covariance between xt and xt−h. This is;

)

When h = 1, the previous expression becomes

The previous expression = 0. The reason is that, by definition of independence of the

*for any k ≠ j.*

Further, because the wt has mean 0,

For a time series,

**Example**

Find the mean, variance and ACF of the MA (1) process below.

*Where*

**Solution**

Mean

Variance

ACF

*And 0 for lags h ≥ 2*

**Properties of MA (2) Model**

**Mean**

**Variance**

**ACF**

And

For lags h≥ 3

Proofs follow the same procedures as those for MA (1) models.

**Example**

Find the mean, variance and ACF of the MA (2) process below.

**Solution**

**Mean**

**Variance**

**ACF**

And

For lags h≥ 3

A key way to know which model to use is by observing the ACF and PACF plots.

For an AR process, the ACF decays exponentially towards zero across all lags some call it *tapering*. Its PACF on the other hand has its plot having significant spikes totaling to the order of the model, i.e. one spike for order 0ne, two for two and so on. The rest completely shut off. The process is reversed for MA processes. The figure below indicates ACF and PACF plots for AR (1), AR (2), MA (1) and ARMA (1,1). In conclusion, we can use ACF to determine the order of the MA process and PACF to determine the order of the AR process.

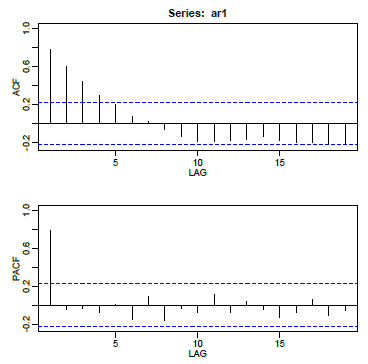


Figure : AR (1)

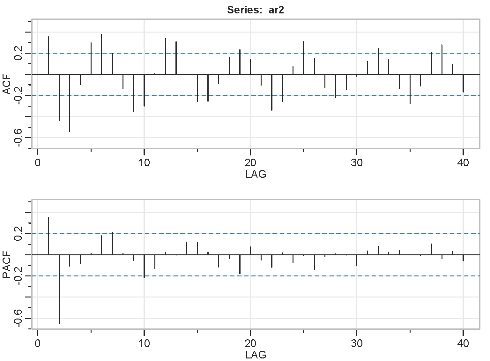


Figure : AR (2)

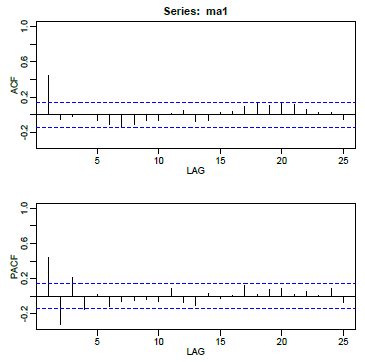


Figure : MA (1)

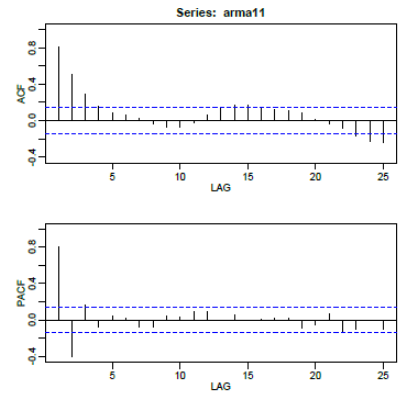


Figure : ARMA (1,1)

## **Fitting a time series in R**

Now that we’ve grasped the essentials of time series lets delve into fitting the model in R. R allows us to manipulate time series with the package **ASTSA,** previously known as **TSA**. There are two types of ARIMA models that we will fit in this section.

1. Non-Seasonal ARIMA
2. Seasonal ARIMA

**Non-Seasonal ARIMA Models**

As discussed earlier, the term ARIMA contains three parts;

AR- Autoregressive

I- Integrated part. This represents the differencing part.

MA- Moving Average part.

A model with (only) two AR terms would be specified as an ARIMA of order (2,0,0).

A MA (2) model would be specified as an ARIMA of order (0,0,2).

A model with one AR term, a first difference, and one MA term would have order (1,1,1). For the last model, ARIMA (1,1,1), a model with one AR term and one MA term is being applied to the variable Zt = Xt − Xt−1. A first difference might be used to account for a linear trend in the data.

The differencing order refers to successive first differences. For example, for a difference order = 2 the variable analysed is zt = (xt − xt−1) − (xt−1 − xt−2), the first difference of first differences. This type of difference might account for a quadratic trend in the data.

**Steps to Follow When Fitting the Model**

1. Generate a time series plot. From the plot, check out for the characteristics of a time series. Look for possible trend, seasonality, outliers, constancy of variance. The plot might give you insights on what to do. Below are some possible scenarios one may encounter.

You won’t be able to spot any particular model by looking at this plot, but you will be able to see the need for various possible actions. If there’s an obvious upward or downward linear trend, a first difference may be needed. A quadratic trend might need a 2nd order difference (as described above). We rarely want to go much beyond two. In those cases, we might want to think about things like smoothing. Over-differencing can cause us to introduce unnecessary levels of dependency (difference white noise to obtain an MA (1)–difference again to obtain an MA (2), etc.)

For data with a curved upward trend accompanied by increasing variance, you should consider transforming the series with either a logarithm or a square root

1. ACF and PACF. ACF plots best describe the MA order terms while PACF do describe AR order terms. As explained in part two, AR has clear significant spikes in the PACF while its ACF tapers to zero. MA does the opposite.

Once the possible models have been decided on, diagnosis is needed to settle on the best. R provides some estimates for comparison. Consider the following;

1. Look at the significance of the coefficients. In R, SARIMA provides p-values and so you may simply compare the p-value to the standard 0.05 cut-off. The ARIMA command does not provide p-values and so you can calculate a t-statistic:   
   If  
   , then the estimated coefficient is significantly different from 0. When n is large, you may compare  
   to 1.96.
2. Look at the ACF of the residuals. For a good model, all autocorrelations for the residual series should be non-significant. If this isn’t the case, you need to try a different model.
3. Look at Box-Pierce (Ljung) tests for possible residual autocorrelation at various lags.

After considering the above and you still have more than one model looking okay, then do the following;

1. Possibly choose the model with the fewest parameters.
2. Examine standard errors of forecast values. Pick the model with the generally lowest standard errors for predictions of the future.
3. Compare models with regard to statistics such as the MSE (the estimate of the variance of the wt), AIC, AICc, and SIC (also called BIC). Lower values of these statistics are desirable

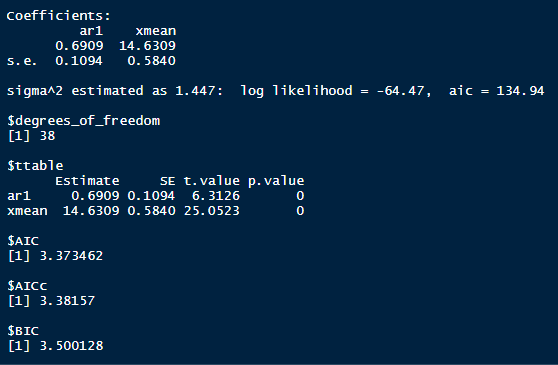
Once the SARIMA function has produced the output and you’ve settled on the best possible model, you have to write down the model equation. Let’s look at the example below for a time series of Lake Erie data. The full script is uploaded at the end of this section.

Figure : ARIMA (1,0,1) parameters for Erie Data

R uses the model below;

The estimated model can be written as

This is equivalent to;

The AR coefficient is statistically significant

Taking a look at the plots we have,

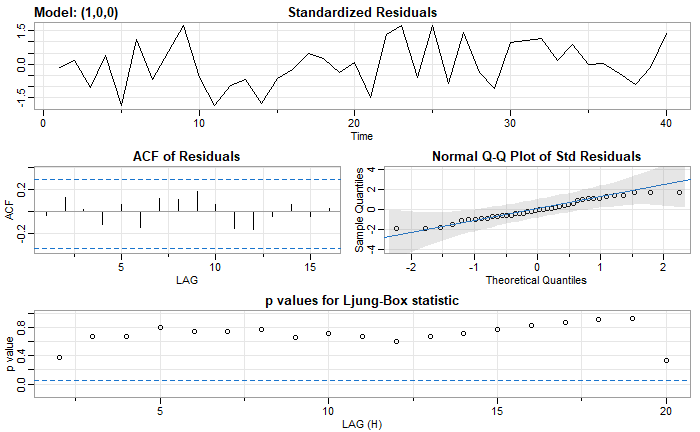


Figure : Diagnostics

* The time series plot of the standardized residuals mostly indicates that there’s no trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals shows no significant autocorrelations; a good result.
* The Q-Q plot is a normal probability plot. It doesn’t look too bad, so the assumption of normally distributed residuals looks okay.
* The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. These statistics consider the accumulated residual autocorrelation from lag 1 up to and including the lag on the horizontal axis. The dashed blue line is at .05. All p-values are above it. That’s a good result. We want non-significant values for this statistic when looking at residuals.
* All in all, the fit looks good.

To perform prediction, use the function; **SARIMA.FOR** as shown in the script. It produces estimates and plots for the series.

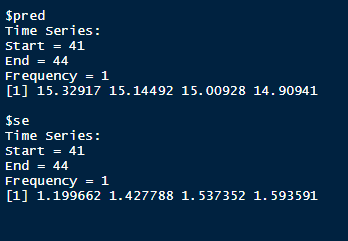


Figure : Prediction parameters

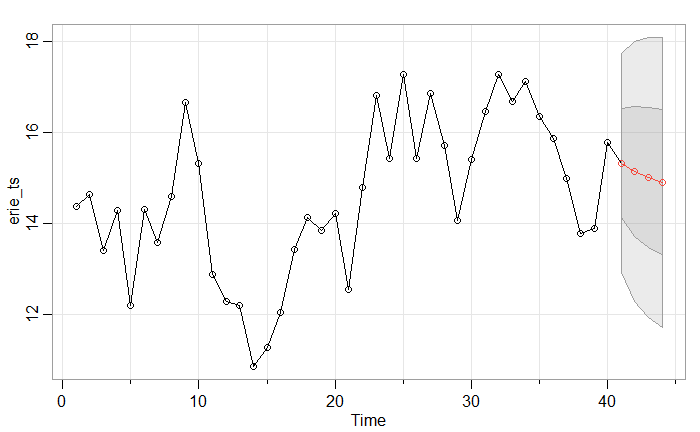


Figure : Plot for Erie time series with prediction outputs

Erie time series script

**Seasonal ARIMA Models**

Seasonality in a time series is a regular pattern of changes that repeats over S time periods, where S defines the number of time periods until the pattern repeats again. For example, there is seasonality in monthly data high values occur particular months and low values particular months. In this case, S = 12 (months per year) is the span of the periodic seasonal behaviour. For quarterly data, S = 4 time periods per year.

In seasonal ARIMA model, seasonal AR and MA terms predict using data values and errors at times with lags that are multiples of S (the span of the seasonality).

With monthly data (and S = 12), a seasonal first order autoregressive model would use xt−12 to predict xt.

A seasonal second order autoregressive model would use xt−12 and to predict xt−24.

A seasonal first order MA (1) model (with S = 12) would use wt−12 as a predictor. A seasonal second order MA (2) model would use wt−12 and wt−24.

**Differencing**

Seasonality usually causes the series to be nonstationary because the average values at some particular times within the seasonal span (months, for example) may be different than the average values at other times. For instance, sales of cooling fans will always be higher in the summer months.

**Seasonal differencing** is defined as a difference between a value and a value with lag that is a multiple of S.

With S = 12, which may occur with monthly data, a seasonal difference is;

The differences (from the previous year) may be about the same for each month of the year giving us a stationary series.

With S = 4, which may occur with quarterly data, a seasonal difference is;

Seasonal differencing removes seasonal trend and can also get rid of a seasonal random walk type of nonstationary.

**Non-seasonal differencing**

If trend is present in the data, we may also need non-seasonal differencing. Often (not always) a first difference (non-seasonal) will “detrend” the data. That is, we use the equation below in the presence of trend.

**Differencing for Trend and Seasonality**

When both trend and seasonality are present, we may need to apply both a non-seasonal first difference and a seasonal difference. That is, we may need to examine the ACF and PACF of

The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the model is

*Where;*

*p is the non-seasonal AR order*

*d is the non-seasonal differencing*

*q is the non-seasonal MA*

*P is the seasonal AR order*

*D is the seasonal differencing*

*Q is the seasonal MA*

*S is the time span of repeating seasonal pattern.*

The model is fitted in the same method as discussed in the non-seasonal model. However, once coefficients are obtained, the model is just written as stated by the coefficients. The is no need for computation of the constant as we did in the non-seasonal ARIMA model.

In the script below, ARIMA model is fitted to a data that contains beer consumption in Germany. The task was accomplished by the help of my friends; Erick Odhiambo, Yunus Mire Mohammed, Alex Ekai and Paul Wafula.

Beer data script.

This is the end of this series. Time series is a wide field, I hope this has provided a brief overview about this field. Read more about time series in the references stated below.