0.3.4.7 Exploratory analysis: time series problem

Recall that a time series problem has data consisting of a column of time-dependent numbers, Y. Time is an independent variable. The time variable can be explicit or implicit. If the data is not evenly spaced, the time variable must be provided explicitly. The model is of the type

$$yt = f(t) + error$$

and can be based on the time domain or the frequency domain. The result is a good prediction model / equation that relates Y to previous values of Y.

We have already said in 0.2.5.4 that time series are characterized by 4 important elements:

- Level: refers to the basic average value of a time series as if it were a straight line
- Trend: This means that changes in the data move up or down in reasonably foreseeable trends or patterns.
- Seasonality: This means that there may be seasonal variations that repeat over a certain period of time, such as every day, week, month or year.
- Variability: also called "noise" or "volatility", refers to random variations that do not fall into any of the other three previous categories.

The recommended exploratory analysis techniques are: **Autocorrelation Plot, Run Sequence plot, Spectral Plot, Complex Demodulation Amplitude Plot, Complex Demodulation Phase Plot, ARIMA Models.**

We have already talked about Run Sequence Plot in 0.3.4.1.1 as a scatterplot of Yi versus i (in this case t) and its ability to reassure on fixed variation.

0.3.4.7.1 Autocorrelation plot

It is a commonly used tool to check the randomness in a data set, that is, their non-dependence on time [AUT-1]. This randomness is ascertained by calculating autocorrelations for data values at varying time intervals. If random, such autocorrelations should be close to zero for all delay separations. If not random, one or more autocorrelations will be significantly different from zero.

In general, the **correlation coefficient** between two values in a time series is called the autocorrelation function (ACF). The ACF for a time series yt is given by Corr (yt, yt-k), where the value of k is the time interval considered and is called lag. A lag 1 autocorrelation (i.e., k = 1 in the above) is the correlation between values over one period of time. More generally, a lag k autocorrelation is the correlation between values that have a distance of k time periods. ACF is a way to measure the **linear relationship** between an observation at time t and observations at previous times.

Data that has significant autocorrelation is not random. However, data that does not show significant autocorrelation can still show non-randomness in other ways. Autocorrelation is just one of many possible measures of randomness.

The autocorrelation graph can provide answers to the questions:

- Is data random?
- Is an observation related to an adjacent observation?

- Is the observed time series white noise?
- Is the observed time series sinusoidal?
- Is the observed time series autoregressive?
- What is an appropriate model for the observed time series?
- Is the model Y = constant + error valid and sufficient?

An autocorrelation plot has the Rh autocorrelation coefficient on the vertical axis and the time lag h on the horizontal axis, with h = 1,2,3 ...

The Rh autocorrelation coefficient, which is between -1 and +1, is calculated as

$$Rh = Ch/C0$$

where Ch is the autocovariance function:

$$C_h = \frac{1}{N}\sum_{t=1}^{N-h}(Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$

and C0 is the variance function:

$$C_0 = \frac{\sum_{t=1}^{N}(Y_t - \bar{Y})^2}{N}$$

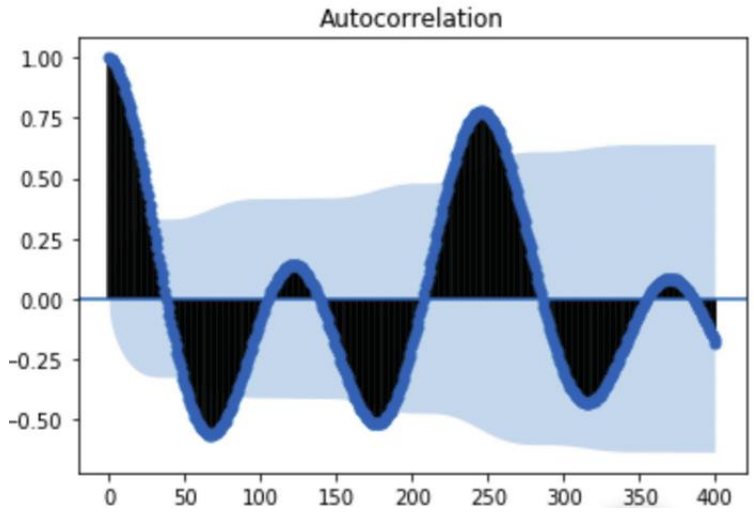

Figure 1: Time series autocorrelation

The autocorrelation graph can show randomness:

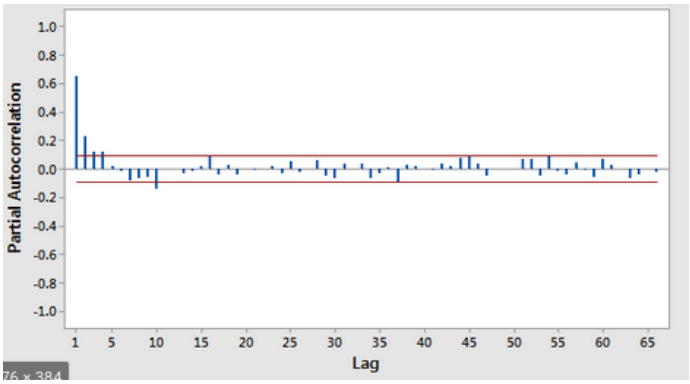

Figure 2: Autocorrelation and randomness

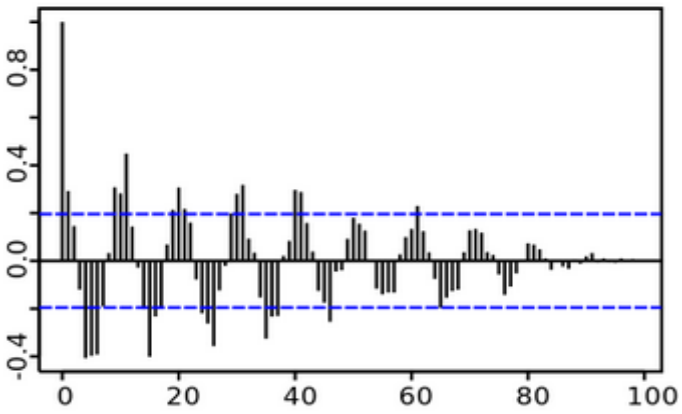

Figure 3: Autocorrelation and sinusoidal pattern

or it can show a sinusoidal pattern:

or it can show a high autocorrelation.

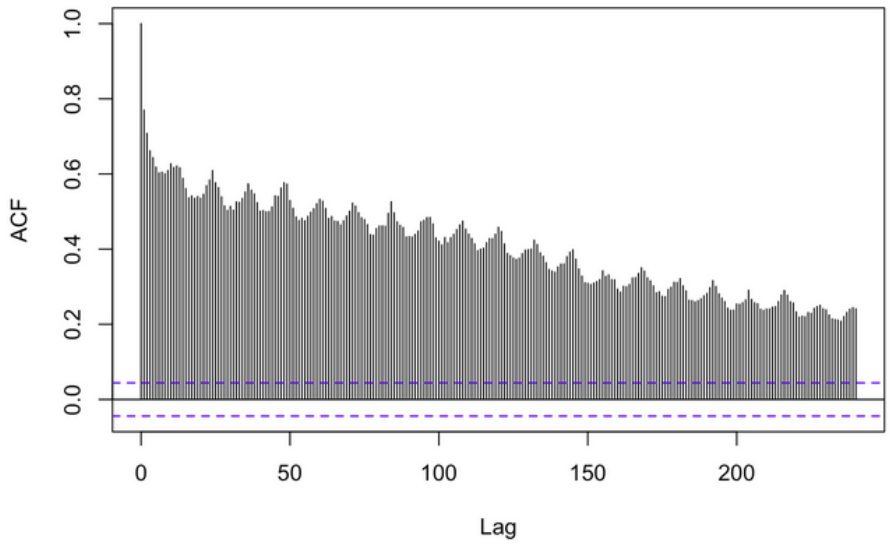

Figure 4: High autocorrelation for k =1

Often the autocorrelation graph shows a horizontal zone corresponding to the 95% confidence interval.

0.3.4.7.2 Spectral plot

It is a graphic technique to examine the presence of a cyclic structure in the time series [SPA-1] and

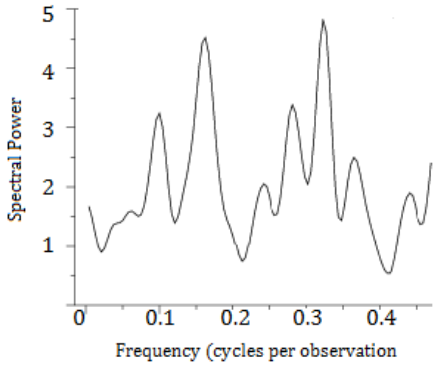

Figure 5: Spectral plot

to deduce its frequency, using frequency spectrum analysis techniques.

Spectral analysis studies the frequency spectrum contained in discrete and uniformly sampled data. The Fourier transform is the main tool that reveals the frequency components of a signal based on time or space by representing it in the frequency space and specifying a time series as a function of the trigonometric components.

An obvious model (as a Fourier series) would be:

$$Y_t = a\cos(ft) + b\sin(ft) + e_t$$

$$Y_t = R\cos(ft + d) + e_t$$

or in an equivalent way:

$$Y_t = R\sin(ft + d) + e_t$$

or:

where R is the amplitude of the variation, f is the frequency of the periodic variation and d is the phase.

Frequency is measured in cycles per unit of time where unit time is defined as the distance between 2 adiacent points. Period and frequency are reciprocal: for example a period of 12 corresponds to a frequency of 1/12 (or 0.083). A frequency of 0 corresponds to an infinite cycle while a frequency of 0.5 corresponds to a cycle of 2 adiacent data points. Equally spaced time series are inherently limited to detecting frequencies between 0 and 0.5.

The trend should typically be removed from the time series before applying the spectral plot. Trends can be detected by a run sequence plot.

A spectral plot consists of a:

- Vertical axis: Smoothed variance (energy or power)
- Horizontal axis: Frequency (cycles per observation)

and can be used to answer the following questions:

- How many cyclical components are there?
- Is there a dominant cyclic frequency?
- If there is a dominant cyclic frequency, what is it?

To learn more, see https://www.amazon.com/dp/B0CPX48CGR

or

https://www.amazon.com/dp/B0CR726PDW