Cairo University
Faculty of Engineering
Computer Engineering Department
CMPN451 - Big Data Mining

Time Series Analysis

Lab Objectives:

By the end of this lab, students should be able to:

1. Visualize a time series, together with its main components (trend component, seasonal component, and noise component).
2. Transform a non-stationary series to a stationary one using logarithms and differencing.
3. Choose the best ARIMA model (the one having least AIC and BIC values) and fit the data to it.
4. Understand the results displayed by ARIMA model.
5. Forecast the time series values for the next few days/months/years.
6. Visualize forecasts and their standard errors.

Lab Requirements:

1. Review the tutorial of Time Series Analysis. The slides are uploaded on the drive and the video of the tutorial is uploaded on the course's playlist on YouTube.
2. Run the R script once to know what is going on before answering the questions.

Lab Instructions:

1. This is a non-coding lab. You will write no new code.
2. You will have to run the script, observe what is happening, and write down your conclusions.

Lab Deliverables:

1. A document (pdf document) containing clear answers to the lab questions.
2. Don't send the R script.

Lab Questions:

1. **After inspecting the time series:**
   a) What is the time range (start and end) of this time series?
   b) How many values are there in this time series?
   c) What is the time interval separating between each two consecutive values in the time series? (monthly, weekly, daily, yearly, etc)

2. What does the parameter *frequency* mean ? Why did we set it to 12?

3. **After visualizing the time series:**
   a) Add a neat plot of the generated time series.
   b) Do you think there is a trend in the time series? If yes, then what is the degree of the trend (i.e. is it linear, quadratic, etc.)?
   c) Do you think there is a seasonality in the time series?

4. What does the function **stl** do? Add a neat plot of the plot generated.

5. **Back to the original time series**. The time series expects the data to be **stationary** which means the following:
   – **Mean is constant:** There should not be any increasing or decreasing trend in the data whether linear, quadratic or higher order trend.
   – **Variance is constant:** The time series should not diverge as it progresses.
   – **Auto-correlation structure is constant:** There are no values in the time series that are correlated with other previous values. (For example, sales of Jan. 2003 will be correlated in some sense with the sales of Jan. 2004 due to seasonality effects).

Let's investigate the original time series:
   a) Is this time series stationary?
   b) Mention the reasons behind your answer.

6. After **differencing** the time series:
   a) Add a neat plot of the time series after differencing.
   b) Does the time series become stationary? (Yes/No)
   c) If no, which of the three conditions mentioned above are still not satisfied?
   d) How does differencing help to make a time series stationary?

7. After applying **logarithm** to the time series, compare visually (5) and (7) and you will know the answer to the following questions:
   a) Add a neat plot of the time series after applying logarithm.

b) Does the time series become stationary? (Yes/No).
c) If no, which of the three conditions mentioned above are still not satisfied?
d) How does applying logarithm help to make a time series stationary? [This is a new piece of information never told in the lecture or the tutorial].

8. After applying both **differencing** and **logarithm** to the time series:
   a) Add a neat plot of the final time series.
   b) Does the time series become stationary? (Yes/No).
   c) If no, which of the three conditions mentioned above are still not satisfied?

9. **After fitting an ARIMA Model with the logarithm of the time series:**
   a) Does the time series passed to the ARIMA model successfully satsify the requirement of a stationary time series? Why?
   b) Inspect the summary of the model.
      The output of the ARIMA model is: (p,d,q) (P,D,Q)[S]
      What does the (p, d, q) mean?
      What do you think the (P, D, Q)[S] relate to? (no details are needed).
   c) Do you think that ARIMA model achieved the requirements of (9-a) internally (i.e. it operated on the stationary version of the time series)? If yes, how did it happen briefly? You don't need to give any mathematical proofs or so. You just need to observe the ARIMA model output (9-b) and you will get it.
   d) What do you think will be more suitable for the case of forecasting the tractor sales, an autoregressive (AR) model or a moving average (MA) model? Why?

10. After changing **trace = True**:
    a) How is the best model selected? What is the information criterion used in selecting the best model? [Mention only the name] [Check the help]
    b) What other information criteria are there that can be used as well? [Mention only the names][Check the help]
    c) Do we seek to get the minimum value or the maximum value of this criterion?

11. What is the meaning of n.head = 36?

12. **After forecasting and plotting the future values:**
    a) According to your observation, does this forecast work well?
    b) Add a neat plot of the generated time series.

13. After using **TSPred library**:
    a) Does this library generate a similar plot to (12)? Why?
    b) Add a neat plot of the generated time series.

14. What happened when we tried to forecast the tractor sales for an extended or longer time range? What do you notice?