LSTAT2170: Guidelines for the preparation of the data analysis project

The projects are individual: you can work together on screening and discussing the data if you have decided to work in a group of two but everybody has to prepare his or her own **individual report.**

You need to do an analysis of the real data you got, which follows the techniques studied in class and during the R-tutorial.

You need to find a model (or if interesting two equally competitive models), fully justify it (see list below) and apply it to do a prediction of roughly 10% of the length of the series into the future (or at least one cycle for seasonal data). Including the prediction intervals (which might be the most interesting part for the financial data).

The length of the report should not be more than 15 pages, but you can put additional figures and code into an annex. Pay attention to your figures being sufficiently large, and well scaled. For the oral exam you are supposed to bring a printed copy without any annotations.

Here are now some **guidelines** that should help you:

The report will be evaluated by checking if you have done the following:

* Preliminary presentation of your data (with comments, if appropriate, on some existing trend, seasonality, structural breaks, particular variance behaviour)
* Justification of applying a preliminary transformation (if appropriate)
* Intuitive motivation and statistical justification of a model with possible seasonalities and/or trends (if appropriate) – however ***don’t copy parts of the theory from the class material, this is not necessary.***
* Interpretation of the plots of empirical acf and pacf (together!) of the stationary component for preliminary choice of order of a possible collection of ARMA/SARIMA models to fit – **in case of a seasonality particular care needs to be taken in order to not overfit**
* In general try to choose *parsimonious* models
* Even if the seasonality is not very pronounced, use a seasonal ARIMA model **(SARIMA) in case of data with a** **monthly/yearly structure**
* Write down your complete final model(s) – not only the one for the stationary component – and do not forget to give values for all significant coefficients and the innovations variance
* Justification of choice of one or several models, estimation of its parameters, and proposal of a complete final model (or two), using the following tools:

Significance tests of the coefficients.

Automatic order selection criteria: comparison of AIC (and maybe log-likelihood to get an idea about the amount of penalisation).

Predictive power of a candidate model: use the one-step ahead prediction error « in sample », i.e. on the last 20% of your observations, to compare several models.

Analysis of the residuals of the model fit: interpretation of the plots of arima.diag, Portmanteau (Ljung-Box) test.

Analysis of the residuals: is there a correlation in the squared residuals? If yes, need of applying (G)ARCH-modelling on the innovations from an ARMA-fit. However, maybe these correlations are very weak. Compare the remark on financial data modelling below. In any case, non-financial data rarely need GARCH modelling.

* For modelling financial data with (G)ARCH: Work on the log-returns: log(X\_{t}) – log (X\_{t-1}), do not apply ARMA+(G)ARCH unless it should be really necessary: don’t get fooled by weak potentially significant appearing (P)ACFs of the log-returns:

For long time series, the confidence intervals for (P)ACFs become narrow, and potentially weak empirical (P)ACFs appear to be significant (if necessary choose wider confidence intervals on a higher level to verify).

If interesting, discuss the distributional properties of your data (normal, non-normal).

* Prediction of the time series of roughly 10% of the length of the original series beyond its end (or at least one cycle into the future in case of a seasonality): Plot of this prediction and of the prediction intervals. Explanation how these intervals were constructed. Interpretation, given the nature of the data.

Comparison (even “in sample”) with a nonparametric prediction method and discussion of potential differences.

In case of financial data, comparison of shape and length of *prediction intervals* at potentially different time points of the series (cf the discussion in class) - the values of the (point) predictor are not really the most interesting.

**Reminder**: *you have to submit your project on Moodle - please do not submit via e-mail*

*You find the deadline for submission on Moodle or in an e-mail sent to you.*

Plus: please do not hesitate to contact us if you have doubts about your model(s) that you intend to choose after having worked on your data, and this as early as possible before the submission deadline!