Stat 618: Experimental Design and Time Series
Spring 2019 *Last updated: Mar 23, 2019*

Course project

This document provides instructions for the course project in *36-618 Experimental Design and Time Series*.

This project focuses on the time series part of the course. We have seen during this course that a time series analysis typically involves the following steps:

1. Finding, downloading and cleaning up relevant time series data
2. Exploratory data analysis
3. Model identification
4. Model fitting
5. Model validation
6. Making inferences using the fitted model
7. Making forecasts using the fitted model

We have so far studied and practiced each of these steps separately. The goal of this project is for your to synthesize these skills by conducting a full time series analysis involving all of these steps with an actual real-life data set. An equally important goal is to learn to present the analysis results and conclusions both orally and in a written form.

This project should be done in groups of 2–3 students. The graded deliverables that are expected from the group are a) an oral presentation and b) a written report.

The timeline for this project is as follows:

- You must have communicated who is in your group and approved your dataset with Mikael by the end of the day on Apr 2.
- Project presentations will take place during the lecture time slots on Apr 23 and Apr 25.
- The deadline for submitting the written project report is Apr 28 at 11:59 pm.

This project will be graded as follows:

- Quality of the data analysis: 40%
- Quality of the oral presentation: 20%
- Quality of the written report: 40%

Your work will be graded holistically. In other words, there is no specific checklist of things we are looking for but we instead judge the quality and ambition of your work as a whole. Recall that this project constitutes 40% of your final course grade.

Specific instructions for each component of this project are given below.

1 Finding data

The groups are expected to analyze time series data of their own choosing. Plenty of suitable data sets can be easily found online. Please attempt to find a data set that piques the interest of the group members and that enables you to address a substantive scientific or business question. When deciding on the data set, you should observe the following guidelines:

- *The data must be time series data.* The data set may contain one or several time series.
- The data must be publicly available online. The data source must be identified in the final written report.
- Each group must analyze a different data set. If several groups wish to analyze the same data set, the first group to communicate their intention to work with these data to Mikael will get the priority.
- The data must consist of real-world observations.
- You may not use polished data examples from an R package, book, course or a tutorial. It should be the "real thing" in the sense of coming from an original data source in a raw format. If the data you find is already in an R format, then it is probably not suitable for this project.
- You *must* approve your choice of data with Mikael. You may do so during Mikael's office hour on Mar 25 (14:30-15:30), right after the lecture on Mar 26 (time constraints apply), during the office hour on Apr 2 (11:30-12:30) or via email. When you approve your data, you should also communicate to Mikael who is in your group.
- You must have approved your data set with Mikael by the end of the day on Apr 2.

Some potential data to consider are:

- Weather data
- Climate data
- Stock prices
- Commodity prices
- Economic indicators (GDP, employment rate,...)
- Environmental data (e.g. air quality data)
- Website usage data
- Customer data published by various companies
- Resource usage data from utility companies (e.g. power grid load data)
- Data available in online governmental data repositories (see, for example, `www.data.gov`)
- Data published alongside scientific papers
- ...

You can easily find examples of all of these by searching with Google. Note that you are by no means restricted to these options: you are free and encouraged to consider data of your choice as long as it satisfies the guidelines above.

The point of this part is to give you the experience of finding a relevant high-quality data set online. You may find that this is surprisingly challenging, but that's exactly the point here. Finding high-quality real-life data that comes in a desired format and with the right set of variables can be surprisingly difficult, but that is a challenge that you will almost certainly face once you start working on applied problems in the real world.

2 Data analysis

Once you have decided which time series data you are going to analyze, you should formulate a few scientific or business questions that you wish to answer using these data. You should then perform a time series analysis to answer these questions and produce a report of your analysis and conclusions. Your written project report should articulate which questions you set out to answer.

Your data analysis should consists of steps 1–7 given on page 1 of this document. You should use the tools and methods you have learned about throughout this course to carry out these steps. At each step, you should choose the tools and methods so that they are appropriate for the data and questions at hand. Here are some further guidelines for the data analysis part:

- It is compulsory that you identify and fit an ARMA (or ARIMA/SARIMA, if appropriate) model to your data.
- You must perform extensive checks for any model you fit. If the model does not seem to fit well, you should attempt to make changes to the model.
- Your analysis must include some inference (i.e., using the fitted model to draw conclusions about the dynamics of the time series process) and some prediction (i.e., forecasts into the future)
- You may use automatic tools, such as `auto.arima`, to identify models, but you must not use these tools as black boxes. This means that you should investigate the different options and tuning parameters implemented in such tools (instead of relying on the default options), make sure that the identified model makes sense and perform extensive model checking.
- If warranted by the problem at hand, you are free to use tools, methods, models and R packages that go beyond the materials we have covered in this course. In that case, you may need to do some background reading on your own. It is possible to receive full points for this project by performing a correct, high-quality ARIMA analysis. However, ambitious use of advanced techniques (for example, vector ARMA or GARCH models) will be seen positively when we grade the projects, provided that it is clear from your report and code that you have a solid grasp of each step of your analysis. If you use other tools than ARIMA models (e.g. autoregressive neural networks), you should compare the results with those from the compulsory ARIMA analysis.
- You should implement your data analysis in R. You must include your code as an appendix in your written report.

3 Oral presentation

The presentations will be 10–15 mins long (including questions) and will take place during the usual class times on Tue Apr 23 and Thu Apr 25. The exact duration will be determined once the number of groups is known. All group members must actively participate in the presentation. If for some reason you cannot make it during one of those classes, you must let Mikael know as soon as possible.

4 Written report

Your report should use the XIMRaD structure. That is, the report should contain the following sections:

- Executive summary
- Introduction
- Methods
- Results

- Discussion

You must also include an appendix that contains your R code. The maximum length of the report is 20 pages (excluding references and the code appendix).

You are free to produce the report using your favorite software, although the recommended option is to use R Markdown. A report template for R Markdown is available on the course Canvas page.

If you use outside references, you must cite them using the usual scientific and professional conventions.

Pay close attention to the quality of your plots, tables and writing. You must produce professional graphics that make appropriate use of axes labels, axes ranges, legends, colors, line types etc. Your tables must be easy to read and must have a professional look. Make sure to carefully proofread your text before handing it in. Remember that the quality of your report constitutes 40% of the project grade. It is best to start writing early so that you have enough time to polish everything before submission.

You must submit your report electronically in Canvas by 11:59 pm on Sun Apr 28.