Stat 618: Time Series and Experimental Design
Spring 2022 *Last updated: March 20, 2022*

Course project

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

The 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 and/or forecasts using the fitted model

We have so far studied and practiced each of these steps separately. The goal of this project is for you 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 2–4 students (it would be ideal if most groups had 3 students). My hope is that all of you have already found a group—if not, please let Mikael know as soon as possible.

The graded deliverables that are expected from each group are a) an oral presentation and b) a written report.

The timeline for the project is as follows:

- *Data approval:* You must have communicated who is in your group and approved your data set with Mikael by the end of the day on March 30.
- *Midpoint check-in:* At least one person from each group must briefly report on the group's progress during Mikael's office hours on either April 11 or April 13.
- *Presentations:* Project presentations will take place during the lecture time slots on April 19 and April 21.
- *Report submission:* The deadline for submitting the written project report is April 24 at 11:59 pm.

This project will be graded as follows:

- Quality of the data analysis: 50%
- Quality of the oral presentation: 20%
- Quality of the written report: 30%

Your work will be graded holistically. In other words, since each project will have its own unique data set with its own unique data analytic goals, there is no specific checklist of things we are looking for but we instead strive to judge the quality, correctness and ambition of your work as a whole. Recall that this project constitutes 35% of your final course grade.

Specific instructions for each component of the 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 your group and that enables you to address a substantive scientific, business or policy 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. In the past, the most interesting projects have worked with data involving multiple 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, a textbook, a 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 hours 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 March 30.

Some potential data to consider are:

- Weather and climate data
- Stock prices
- Commodity prices
- Economic indicators (GDP, employment rate,...)
- Environmental data (e.g., air quality data)
- Epidemiological data (e.g., COVID data)
- Website usage data
- Customer data released by various companies
- Resource usage data from utility companies (e.g., power grid load data)
- Data available in online data repositories (see, for example, `zenodo.org` or `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 choosing as long as they satisfy the guidelines above.

The goal of this part is to give you experience of finding a relevant high-quality data set online. You might find this to be surprisingly tricky, but that's exactly the point here. Finding accessible 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 also face in your job after graduation.

2 Data analysis

Once you have decided which time series data you are going to analyze, you should formulate a few scientific, business or policy questions that you aim to answer using these data. You should consider both inferential and predictive questions, as appropriate. You should then perform a time series analysis to answer these questions and produce a presentation and a report of your analysis and its conclusions. Your written project report should articulate clearly which questions you set out to answer.

Your data analysis should consists of steps 1–6 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 as part of your analysis.
- You must perform extensive checks for any model you fit and describe those checks in your written report. If the model does not seem to fit well, you should attempt to make changes to improve 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 automated 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 meticulous, high-quality ARIMA analysis. However, ambitious use of more advanced techniques (for example, vector ARMA, time series regression, GARCH,...) 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 basic 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 Midpoint check-in

At least one person from each group must briefly check-in during Mikael's office hours on Week 15 (i.e., April 11 or April 13). During the check-in, the group should briefly explain their work so far, discuss any unresolved challenges and describe the plan for the remaining work. It can be useful to show plots and R code during the check-in but the group does not need to prepare slides or other materials for the check-in. Please contact Mikael for alternative arrangements if no one from your group can attend one of Mikael's office hours during Week 15.

4 Oral presentation

15 mins

The presentations will be 10–15 mins long (including questions) and will be done in person during the usual class times on Tuesday, April 19 and Thursday, April 21. The exact duration will be determined once the number of groups is known. All group members should actively participate in the group's presentation and *all groups should follow all presentations*. Alternative arrangements can be made for groups or group members who, due to extenuating circumstances, may not be able to present during the

usual class time. If you believe that you may be in this situation, you should let Mikael know as soon as possible. You will be asked to share your slides with Mikael before your presentation. The quality of your visual materials (slides, figures, etc.) will be taken into account when grading the presentation.

5 Written report

Each group will hand in one report written jointly by all group members. The report should contain the following sections:

- Executive summary (max. 1 page)
- Introduction
- Methods
- Results
- Discussion

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

You are free to produce the report using your favorite software or online platform. A recommended option is to use R Markdown. A report template for R Markdown will be provided on the course Canvas page. Alternatively, Overleaf is an excellent online tool for collaborating on LaTeX documents. You may also consider using, for example, Google Docs, as long as you make sure that the final report has a professional look and feel to it.

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, aspect ratios, 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 30% of the project grade. It is best to start writing early so that you have enough time to polish everything before submission. Remember, high-quality writing always takes longer than you might expect!

You must submit your report electronically in Canvas by 11:59 pm on Sunday, April 24.