STAT 520 Project Description – Fall 2017

NOTE: *The guidelines for this project have been borrowed heavily (with permission) from Dr. Joshua Tebbs, who has taught this course in previous semesters.*

The time has come for you (if you haven't done so already) to start working on your project for STAT 520. Remember that this project counts as 20 percent towards your final grade, so it is very important. By "project," I mean a written exposition that thoroughly describes the complete analysis of a data set. Of course, for us this will mean forecasting as well. After all, forecasting is an important part of time series analysis. It is important to think about the four major steps of analyzing a time series data set.

1. Model specification (Chapter 6)
2. Parameter estimation; i.e., model fitting (Chapter 7)
3. Model diagnostics (Chapter 8)
4. Forecasting (Chapter 9)

In the model specification phase, your goal is to come up with a small set of candidate models for your data. In the estimation phase, you are to fit the models. In the diagnostics phase, you are to check the adequacy of the model fits (and possibly revisit the model specification phase at this point, based on what you learn from the diagnostics). After diagnosing the model fits, you should choose a final model and forecast future observations.

Here are some guidelines for choosing a data set:

• Use other sources (e.g., data sets on line, other research projects, etc.) to find an interesting data set. The more interesting the better! You could also collect your own data (but you may want to get my blessing first if you do this).

• The responses Y_t should be continuous in nature (or at least somewhat continuous; just not binary or strikingly discrete). Also, make sure you completely understand the sampling frequency; e.g., are the data collected every day? every week? every month? every year? every minute?

• Choose a data set in an area you are interested in! You should be able to demonstrate a working knowledge of the subject area (this may mean you have to do a considerable amount of research to get to this level).

• Don't choose a data set that is very small (e.g., let's stay away from series where $n < 50$). Ideally, we want $n > 75$ or so, but this is just a guideline. Remember that many statistical methods we have discussed exploit asymptotic distribution theory, so we want to apply these methods to suitably lengthy data sets.

• In order to use the methods we have discussed in this class, it is best to choose *regularly spaced* data, meaning that there is essentially the same interval of time between each measurement.

• Please read the links under "Information on the Course Project" on the course web page, to see details about reading time series data into R and creating 'ts' objects.

• Because we won't discuss seasonal models until the end of the semester, you may want to consider staying away from series that exhibit seasonality. Exception: If you want to work with a data set that has seasonality, you may be forced to use the harmonic regression methods from Chapter 3 or you can read ahead and learn about seasonal ARIMA models (Chapter 10) on your own.

As a first step, I would recommend that you get your data and go through the methods in Chapter 6 to identify a small set of candidate models.

Another recommendation: It might be a good idea to "withhold" some of the data from your series towards the end of it so that you can compare your forecasts to the actual values of the process. For example, suppose that you have a series of length $n = 100$. Perform the specification, fit, and diagnostics on the first 95 observations and withhold the last 5. Then, when you forecast, you can compare your first

5 forecasts to the actual last 5 observations−this will give you an idea on how accurate/precise your forecasting is. Hopefully, your forecasts are "close" to the actual values in the series!

Outline of the written project (in this order):

• **Title page and abstract.** You must prepare a title page with an appropriate title and abstract. The abstract should go on the title page. An abstract is a very high-level written summary of the entire project. Main points and findings only. Aim for about 200 words.

• **Introduction.** This part introduces the reader to the data set and to the area to which it pertains. For example, if you are analyzing giraffe population growth data from southern Kenya, you should describe why this is an important problem to investigate and give the reader a review of pertinent background information about giraffes in Kenya. Basically, introduce the reader to the problem and why it is meritorious of investigation. This should be written at a very basic level (i.e., no mathematics or notation). Remember your reader may not know anything about the area in which you are writing. Aim for about 2 pages (if double-spaced; maybe about one page if single-spaced) here.

• **Model specification.** This is the "meat" of the paper and will be the longest in length. In this section, you want to describe, in clear detail, the data analysis used to specify your candidate models. Pretend as if you are taking the reader by the hand and leading him or her through your thought process which leads to your model selections. In doing this, however, try not to overdo the first-person writing. It can sound less professional and less authoritative if you continually write things like, "I tried this, and then I tried that ..."

• **Fitting and Diagnostics.** This part of the project should describe the model fitting and diagnostics techniques you used, with the goal of identifying a "final" model for forecasting. Identify also what possible deficiencies your final model has. Remember, no model is perfect.

• **Forecasting.** This section should describe the techniques you used to forecast future observations (see "Another recommendation" on the previous page). Why is forecasting important? What impacts could your forecasting have?

• **Discussion.** Here you want to offer a summary of what you did in the project and draw your main conclusions. Also, it is a good idea to discuss here other issues related to the data analysis. For example, does your analysis have any shortcomings or lack of generalizability? What were the main problems you encountered? It is OK if your final model is not "picture-perfect." Few are. Real life data analysis is often more difficult than textbook problems.

• **Bibliography.** Cite all references (including the original source of your data) carefully.

• **Appendices.** Use appendices to catalogue extra graphics/plots/output. You could also give the values of the series here as well (although this is not necessary). Basically, I use an appendix to house information that I want the reader to have access to, but feel that it would interrupt the flow of the main body of the paper.

Here is more general advice:

• The report may be single-spaced or double-spaced, but either way please make the formatting such that the report is easy to read. The abstract is usually single spaced.

• Break your report into sections. Each section should have a title. Use subsections (with titles) if necessary. Avoid subsubsections.

• Integrate R graphics and output into the written text as you see fit. For example, if you want to show me the time series itself (you better!), embed it into the written work. Look at the style of the way graphics are embedded into the text of our Cryer and Chan textbook as a guide. It is important to strike a balance here! You don't want too many graphics in the written text, because it may make your report seem "fragmented." Be judicious with your choices; don't be afraid to log ancillary graphs/plots/output in appendices.

• I am a very picky reader! I do not like errors in grammar, errors in the use of punctuation and capitalization, and spelling errors (I loathe these). These types of errors are clear signs of lack of interest

and of the author not paying attention to details. My advice: Edit and reread your written project at least 10 times. Get other people (not in statistics) to read your project and offer comments/feedback.
• Adopt a writing style that you are comfortable with. Written projects do not have to be terse and cold. I think the best writing makes me feel like the author is right next to me reading it. There are no "correct" writing styles, but there are certainly bad ones.
• Have fun! This is an opportunity for you to amalgamate all of your time series knowledge and apply it to a real problem. Show me what you have learned. Remember, the written project is how you disseminate your work. I think the most important part of a statistical analysis is clearly communicating it to others in writing.
• I don't have a specific target number of pages for the whole report. You should do enough to provide a full analysis of this data set, with attention paid to each of the sections listed above. However, you don't need to belabor things by including every possible method you've ever learned. That approach tends to result in a report that does not flow well and is a chore to read. Choose to include the methods of analysis that shed light on the fundamental questions you are trying to answer about the data set.

Due Date: Friday, December 8, 2017 by 3:00 p.m. You may hand me a hard copy, or you may email me a copy of your project as a Word document or pdf file.

Grading Scale:

Your report will be graded out of a total of 40 points, based on Writing, Analysis, and Context. For example:

Writing (out of 10 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist?

Analysis (out of 20 points): Were the chosen models, graphs, and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

Context (out of 10 points): Were the questions answered in terms of the variables of the data set? Have you attempted to frame your conclusions and interpretations in a subject-matter context rather than treating the data as simply a meaningless set of numbers? Have you provided some background information about the data set and why it is of interest?