Math 265, Spring 2016, Final Project

Due Monday, 5/16/16. **No penalty** will be imposed if the project is submitted by noon on Monday, 5/23/16. No projects will be accepted after that time. Projects should be printed (no electronic versions), and may be submitted to me in person, slipped under my office door, or submitted to the math department office (and ask them to place the document in my mail box) any time before the due date(s).

Find on the course Piazza page five data sets.

Your assignment for the first four data sets is to forecast the next thirteen values, along with providing prediction intervals for these forecasts. Of course in order to accomplish this, you will need to take many steps, including (but not limited to) plotting ACF/PACF values, looking at the periodogram, fitting a variety of models and looking at AIC/etc., checking diagnostics, and many other things.

For **each** of these first four data sets, you should present an "Executive Report" of one page containing only the pertinent information (which might include a description of the project and, among other things, the chosen model, parameter estimates, the forecasts, an indication of how accurate these estimates/forecasts are, perhaps both graphically as well as having a table of values). In "real life" this is something that would be read by a non-technical person, and should be written with that in mind: They don't have a clue what "ARIMA" means or what "α" is, etc. You are not to be teaching them. You are giving them results that they can act upon to improve their business. This executive report would be followed by a technical appendix for each data set. In order to receive full credit, you will need to give a description of what you did (including only a **very few well chosen** graphs, etc., and explaining what you are looking for and what you find at each stage of the investigational process). This should be no longer than 4 pages for each data set (shorter is better, as long as it is complete and thorough). Thus, each data set should have its analysis presented in no more than 5 pages, hopefully less. In real life you may work on a project for months or years, but your reader will usually only glance through the report for a minute or two. Do not give all the minutia of your work, only provide the most important results. Do NOT EVER give pages and pages of raw data (or pages and pages of numbers of any other sort). Explain clearly the meaning and purpose of any numbers or figures that you do provide.

Your grade will be based upon the presentation style, the quality of the forecasts, and the thoroughness of the investigation of the data. Note that a person can always make prediction intervals that are extremely small by simply over fitting a model (which then in reality does not usually perform well). So, narrow prediction intervals do not necessarily mean quality forecasts.

Keep in mind the university policy on academic integrity (http://info.sjsu.edu/static/catalog/integrity.html). Anyone found to have searched the internet for solutions, discussed the data sets or their analysis with classmates or other students, or in any way violated the letter or the spirit of the academic integrity policy will be subject to penalties which may include failure of the class and/or expulsion from the university

Date sets one through three:

The first three data sets are artificial. They are created from a true (and known to me) ARIMA(p,d,q) process (possibly with transformations, or maybe not). When I created the data sets, I actually simulated a longer data set and thus know the "truth" for the next thirteen observations that you are to predict.

Data set four (deposits):

The fourth data set is real/observed data. For this data set… During a warming period on Earth, glaciers melt and the melted ice, now water, carries silt and sand downstream from the mountainside beside the glacier and is deposited in a layer usually at the floor of a valley below the glacier. The warmer the season, the more ice melts and the more sand and silt is carried downstream. In the winter, this layer of sediment hardens, and is thus differentiable from the layer from the previous year and the subsequent year. Thus, looking at these layers of silt and sediment can give an indication of the temperature changes from one year to the next. Our data is the thickness of the sedimentary deposits each year over a period of 624 years at one particular location. Of course there is random variation; two years with the same temperature will not necessarily leave exactly the same amount of sediment at the valley floor. Not only that, the amount of variability in the thickness of the sedimentary layer is related to the temperature, e.g., in a relatively cold year there will not be much melt and thus not much sediment, so the variability in the thickness of the sedimentary layer will be small. In a warm year there will be a higher level of melt, and there is more variation in the thickness of the sedimentary layer (if you plot the data against time/index, you will see that the smaller values are tightly packed near the bottom of the plot, while higher values are not so dense). This is an indication of non-stationarity (the covariance between two observations not only depends on the time between the observations, but upon their values), and thus the need to in some fashion transform the data before making an ARIMA model (with the transformation being "undone" afterwards to make predictions on the scale of the original data). Note that as you make predictions for "future" values of the thickness of the sedimentary value, you are in some sense making predictions of future weather patterns.

Data set five (HomePrice):

This problem is much more difficult than the earlier problems. Do what you can. As graduate students you are expected to be able to extend ideas that you have learned about in classes and be able to research/explore new ideas and methods which you may not have been taught. A portion of the purpose of this exercise is to explore what you are capable of as an independent researcher.

You should make sure that you have completed the analysis on the first four data sets before working seriously on this data set. The point value for your analysis of this data set is the same as that of any one of the previous data sets, but the analysis is much more difficult. It is not worth spending a lot of time on this section if you have not already completed the earlier sections.

The fifth and last data set is the S&P/Case-Shiller Home Price Index (feel free to read http://en.wikipedia.org/wiki/House_price_index and https://en.wikipedia.org/wiki/Case-Shiller_index). This index takes January 2000 to be a "base-line" and on that date the index is given a numerical value of 100 (as a percent), and then compares the value of homes at other times. The resulting "value" is an overall average throughout the USA. If you imagine the value of a home in January 2000 and multiply it by the index value for a certain date, you get the value for that date. For example, if you imagine a house which might have sold for $500,000 in January of 2000, and see that the index value for January 2016 is 175.42, you might assume that the value of the house on that date might be about $500,000* 175.42% = $500,000*1.7542 = $877,100. Again, this is showing the average increase/decrease throughout the USA and might not be typical for any particular area within the USA. Most people believe that there was a housing bubble in the 2000's where prices increased too rapidly until prices became unsustainable at which point the bubble burst and prices fell (feel free to read about such ideas at http://en.wikipedia.org/wiki/Housing_bubble , http://en.wikipedia.org/wiki/Financial_crisis_of_2007 and/or http://en.wikipedia.org/wiki/United_States_housing_bubble). Of course the rapid increase in housing prices was associated with greater volatility in the prices of homes (larger than typical swings in home prices). The provided data set gives the monthly index values from January 1987 through January 2016.

To read the data into R, after downloading the data to your computer you probably want to write something like

```
read.table("C:\\  …(your path)… \\HomePrice.txt", header=TRUE)
```

Your task is to again write a five page report, the first being an executive summary and the remaining four being a technical appendix. I will give a little more leeway for this one; the appendix may be up to six pages if you feel the need. You are to include the following points in your report:

- Based on your analysis, when did the housing bubble begin?
- Based on your analysis, when did the bubble burst (reach its peak and begin its precipitous fall)? This might be one point in time or an interval of time.
- Based on your analysis, have the effects of the bursting bubble returned us to the levels where one might have expected them to be had there been no bubble, or are we still feeling the effects of the bubble (had any trends before the bubble began simply continued in an ordinary fashion, without a bubble and its after effects, would housing prices be where they are now? Higher? Lower? Are there continuing effects from the bubble as of January 2016?)
- For a person who is currently renting, do you think right now (or January 2016) is a good time to buy a home? Why or why not?

Of course you should state how you come to your conclusions based on your analysis. Solutions or answers without supporting analysis will receive zero credit. If you would like to, feel free to search for other, related, data, and incorporate that data into your analysis. Or not.