LAB 2 :: DATA FILTERING AND NOISE REDUCTION

The following lab utilizes the computer program, Excel. In this exercise you will generate synthetic data sets based on a simplified model of daily high temperatures in Boone and apply several filtering techniques to your data. A key to this lab is that you use Excel in an efficient manner; otherwise, this exercise may take a long time to complete. A formal typed write-up with referenced figures of your results is due by the beginning of the next lab period. The overarching purpose of this lab is two-fold: 1) Perform some quantitative data processing and get results, and 2) Make a professional interpretation and recommendation based on your results. This is more or less what consultants do.

Part I :: Noise & Resolution

1. Set up an Excel spreadsheet and leave column A blank for now. Title column B, "Julian Day" and populate the rest of the column with a list starting at 0 and ending after three years' worth of days have been entered (i.e. 1094 should be the last entry). Do not worry about leap years.

2. Weather.com states that the average daily high temperatures in Boone, NC range from 39^{o}F in January to 76^{o}F in August. We will assume that these temperatures vary sinusoidally (i.e. they can be mathematically represented by a sine function) and that day zero (Jan 1st) has the lowest daily high temperature. Label column C "Temperature (F)" and calculate the predicted daily temperature values using a sine function. Hint: a simple y = sin(x) will not work. The general form of a sine wave of wavelength, λ, is:

$$y = \sin\left(\frac{2\pi\, x}{\lambda}\right)$$

   You will have to use your knowledge of mathematics to determine the equation for a sine wave that has a maximum of 76 and a minimum of 39 with 39 occurring on day 0, 365, etc. Recall that like virtually all programs, the sin function in Excel is in radians. Do not use a cosine function. While you are tweaking your sin function, it will help you a great deal to make a plot to visually check your results. I recommend that you start out plotting $y = \sin\left(\frac{2\pi\, x}{\lambda}\right)$ and then tweak this equation one parameter at a time until you get the desired result. Make certain that your result matches the appropriate wavelength, amplitude, and has the lowest daily temp on day 0, 365, etc… **In your write-up, you should state the type of model you are using, your model's equation, and describe what each portion of the equation represents. Do not write equations in Excel format in your report! Your supervisor doesn't care how Excel works. He/She wants to see the equations written in normal mathematical notation.**

3. Make a scatter plot with straight lines of your predicted daily high temperatures. Place this plot in your spreadsheet so you will be able to see the graph change as you tweak parameters. Follow these specific instructions: Title the plot "Figure 1" and label both axes. Make the horizontal and vertical gridlines dotted @ 0.5 pt and have horizontal gridlines every 5^{o} F and vertical gridlines every half year. Make the graph span 0 - 1095 days in the horizontal and 30 - 85^{o} F in the vertical direction. Show minor tick marks on the vertical axis for each degree of temperature. Only label every 365

days on the horizontal axis. Leave the legend in the graph and call this series “Temp Model”. You will need it later. Plot your data with a dark red line at 2.25 pt. thickness.

4. Although you now have a model to predict daily high temperatures in Boone, real data would never look like this due to noise. Let’s assume that your advisor is a cheapskate and wants to buy temperature gauges that have a +/- 10^{o} F error. You task is to determine what level of error is acceptable if you are trying to predict the annual pattern of daily high temperatures. I.e. you do not care about accurately predicting each daily high temperature; you just want to be able to see the annual trends. Use the random number generator in Excel to add noise to your predictions simulating measurements made by a device with a +/- 10^{o} F error. To do this, make a cell that will hold your error level variable. Label cell A1, “Temp Error +/-”, and below this enter the number 10. Before you attempt to use the random number generator in Excel, you should first play with the RAND() function to figure out how it works. What is the output range of RAND()? How can the range of RAND() be scaled? How can you use RAND() to give a +/- error range? This should be described in your write up. Once you have a good understanding on RAND(), use RAND() to add a +/- 10^{o}F random noise to your data. Add this noisy data to your Figure 1 plot. Call this series “Noisy Data” and plot it with 3pt circles with no fill and a blue marker line color. Do not plot a connected line; only plot the symbols. Feel free to play with the error level in cell A2, but make sure that the plot you include in your write up has a +/- 10^{o} F error level. Be sure to discuss the following questions in your write-up.
    a) What kind of modeling are you doing? What are the advantages and disadvantages of this type of model? You should discuss this in your report.
    b) Given the +/- 10^{o} F error of the proposed temperature gauge, will the annual variation in daily high temperatures be detectable?
    c) At what level of error (in terms of +/-) does the annual pattern become completely obfuscated?
    d) What error level (given as +/-) would you recommend to your advisor and why?
    e) What is the minimum sampling rate (in days) needed to capture the annual variations in daily high temperatures? What sampling rate would you recommend to your advisor?
    f) Give one example of how this data set could become aliased and how you will avoid this in your study.

Part II :: Stacking vs. Digital Filters

Because you were so helpful to your advisor in determining the necessary precision of temperature gauges, your advisor has given you the task of determining what types of filtering will be most effective at improving your data quality of your noisy temperature data (i.e. maximizing the signal to noise ratio).

1. Assume that measurements are made daily (same as in part I) and that your instrument has a +/- 10^{o} F error. Test the overall effectiveness of a 3-point moving window, a 5-point moving window, and a weighted 7-point filter (use Equations. 3.1 & 3.2 on pg. 17 of your textbook), as well as stacking of 5 and 20 daily measurement data sets. These filters are described in your textbook in detail on pages 16-18. Determine which filter type will maximize the signal to noise ratio of your data by calculating the Root Mean Squared (RMS) error (see equation below). In your write-up you should:
   a) **Perform each technique in a separate sheet in Excel. Do each of the 3pt filter, 5pt filter and 7pt filter in separate sheets and put both stacks in the same sheet and on the same plot. Therefore, your stacking sheet should have 20 columns of noisy data.** Include the plots within each sheet, not as separate sheets. Make a plot of the result of each technique (i.e. plot your temperature model as a solid black line, the noisy data as small blue circles, and the filtered data as a dark red solid line). Make the plots in the same style as Figure 1. Title the plots Figures 2a-5a (Fig2=3pt filter, Fig3=5pt filter, Fig4=7pt filter, Fig5=stacking).
   b) For each filtering technique, also make plot of the residuals for the raw noisy data and the filter results (as solid lines with no symbols). Label these figures 2b-5b. The residual is the (result of the filter or noisy data) – (the temperature model). Be sure to describe and reference the residual plot in your write-up.
   c) Calculate and report the Root Mean Squared (RMS) error for each approach (do this in Excel column A). To do this, use the formula:

$$RMS_{error} = \sqrt{\frac{1}{n}\sum_{1}^{n} e^2}$$

   where *e* is the error (i.e. filteredData – TempModel) and n is the number of data points. Which method is most successful at reducing the RMS error?
   d) In your methods section of your write-up, **briefly** describe each filtering/stacking technique and how each works. Discuss the advantages & limitations of each approach.
   e) How is stacking different than the data filters? What must be done differently in your study if you are going to use stacking and how would you accomplish this? Do you recommend using stacking or filtering, or both? Why?
   f) Make your final recommendation. What filtering technique(s) do you recommend and why? Be creative, but make sure that you recommendations are reasonable.

Part III :: Final Report

Your results of this exercise should be presented as a typed formal lab report in the 3-4 page range (not including graphs). A formal lab report should have an introduction, methods, results, and conclusions section. Any and all figures must be labeled (e.g. Figure 1) and must be referenced directly in the text.