LAB 2: DATA FILTERING AND NOISE REDUCTION

In this exercise, you will use Microsoft Excel to generate several synthetic data sets based on a simplified model of daily high temperatures in Boone. You will then apply several filtering techniques to your data to reduce noise. A key to this lab is that you use Excel in an efficient manner; otherwise, this will take a very long time to complete. The overarching purpose of this lab is three-fold: 1) Perform some quantitative data analysis 2) get results, and 3) understand the basic effects of moving window filters and stacking on noisy data. I provide a step by step guide below.

Part I: Creating a Temperature Model

1. Download the Excel file from the course website and save it to your flash drive. Open the Excel file. You will see that I have laid out the basic formatting to help you keep things organized in this lab exercise.

2. Column B, “Day Num”, is going to hold the day numbers for three years’ worth of days. Populate the rest of the column with a list of integers starting at 0 and ending after three years’ worth of days have been entered (i.e. 1095 should be the last entry). We will assume these are all non-leap years.

3. Weather.com states that the average daily high temperatures in Boone, NC range from 39 °F in January to 76 °F in August. To keep things simple, we will assume that temperatures vary sinusoidally with a wavelength of one year so they can be approximated with a sine function. Using this information, you can fit a sinusoidal function to the data. The general form of a sine wave of wavelength, λ, is:

\[ y = \sin \left( \frac{2\pi x}{\lambda} \right) \]

This will be the basis of your temperature model (column C), but this equation will not work as is. You will have to use your knowledge of mathematics to stretch and translate this equation until it produces a maximum of 76 °F and a minimum of 39 °F occurring on day 0, 365, 720, and 1095. Do not use a cosine function. Recall that all trigonometric functions in Excel are in radians. Excel has a function called PI() that you can use for the constant, π. While you are tweaking your sine function, it will help you a great deal to make a quick scatter 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 (first wavelength, then amplitude, then vertical shift, then horizontal shift) until you reach the desired result. Make certain that your result matches the appropriate wavelength, amplitude, and has the lowest daily temp (exactly 39 °F) on day 0, 365, 720, and 1095.

4. Once you have fit the equation to the provided data, make a scatter plot with straight line connectors and no symbols of your temperature model. Place this plot within your current sheet so you will be able to see the graph change as you tweak parameters later in this exercise.
Follow these specific plot formatting instructions:

a) Plot your temp model data with a solid black line at 2.25 pt. thickness.

b) Make the graph span 0-1095 days in the horizontal and 30-85 °F in the vertical direction.

c) Show minor tick marks on the vertical axis for each degree of temperature. Only label every 365 days on the horizontal axis.

d) Make the horizontal and vertical gridlines black and dotted @ 0.5 pt and have horizontal gridlines every 5 °F and vertical gridlines every half year.

e) Label both axes including units. All axes should use black text and lines.

f) Make a legend outside the plot on the right side and call this series “Temp Model”. You will need it later.

---

Part II: Creating a Synthetic Temperature Dataset

Although you now have a model to predict daily high temperatures in Boone, real data would never look like this due to noise/errors. In this exercise, we will only consider instrument error. Let’s assume that your boss is a cheapskate and wants to buy inexpensive temperature gauges that have a ±10 °F error. Your task is to determine what level of error is acceptable for predicting 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 sinusoidal trends. In other words, you just want to be able to see the wavelength of the data and the typical min/max temperatures.

1. Use the random number generator in Excel to add random noise to your temperature model to simulate measurements made by a device with a ±10 °F error. To do this, enter the value, 10, into cell A2. 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. Generate a whole column of random numbers for days 0-1095 in column D and make a quick plot of the data to see what the results are.

   a) What is the output range of RAND()?

   b) How can the range of RAND() be scaled? Recall, that RAND() is a function, so you can apply the same standard mathematical rules to translate and stretch RAND() as you would to any function.

   c) Once you have your random noise function setup to create ±10 °F noise, modify your RAND() equation so it only references cell A2. This way, when you type a new value into A2, your noise level will automatically change (your plot should automatically update). Be sure to test this by typing in several values into A2 and make sure that your plot produces the specified ± error range.

2. Once you have a working random noise function, use your random noise function to add ±10 °F of random noise added to your temperature model equation to create a realistic synthetic temperature data set. Call this series “Noisy Data” and add it to your temperature model plot with
Lab 2: Data Filtering and Noise Reduction

3pt sized circles with no fill and a 1pt blue outline. Do not plot a connected line; only plot the symbols.

3. Experiment with the error level in cell A2 to estimate the maximum level of noise that the temperature gauge can have and still produce a data set that has a visually resolvable annual sinusoidal trend. Make note of this max noise level. I will ask about this later.

4. After you have figured out the maximum acceptable ± error range, set cell A2 back to ± 10 °F error.

Part III: Quantifying Error in Noisy Data

Any scientist that were to look at your newly-created synthetic noisy dataset, could immediately tell that the data has noise/errors. In fact, all numerical scientific data has some level of error. Thus, quantifying error is of great importance in all sciences. In this section, we will use a simple measure of error to quantify the noise in our synthetic noisy dataset. We will later attempt to remove some of this error using digital filters and stacking.

1. To give a single quantitative estimate of how much error your synthetic data has, you should calculate and report the Root Mean Squared (RMS) error of the data from column D. RMS error is defined as

\[ \text{RMS}_{\text{error}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2} \]

where \( e \) is the error or residuals (i.e. Synthetic Data – Temp Model) and \( n \) is the number of data points. RMS error is mathematically similar to Standard Deviation (but is more general), and essentially provides a single number that represents the average error of an entire data set (of any size). Calculating RMS error will take two columns in our Excel sheet: Column E should contain the residuals/errors (which we want to plot) and Column F should contain the residuals squared. Complete the rest of the RMS error equation in cell A5, so it is easily visible when looking at the top of your data.

2. You may have noticed that although you applied a consistent ± 10 °F random noise, the synthetic data tends to look noisier near the peaks and the troughs of the temperature data. This is a visual artifact, and so when comparing noise/errors, geophysicists typically detrend data (remove slopes in data), or simply calculate and plot residuals. We have already calculated the residuals (in Column E), so let’s make a plot that shows the residuals.

For your residual plot, use the following formatting:

a) Make a scatterplot with straight line connectors and no symbols. Set the series name to “Raw”. Place this plot directly below your existing temperature plot and make this new plot the same width and about half the height as your existing plot. It helps if you line up the plots vertically, so you can make sure the widths are the same.
b) Plot the residuals using a blue line at 1.5 pt. thickness (same color blue as your blue circles in the previous plot).

c) Label the y-axis, “Residuals”, and the x-axis, “Days”, and include the appropriate units.

d) Use the same x-axis formatting as before. Make the y-axis go from -10 to 10 with tick marks every one degree and labels every five degrees.

e) Make a legend with a single entry, labeled “Raw”. This is a bit silly right now, but the legends on your plots will be very useful later.

If you did everything correctly, you should see a plot of random noise between -10 and 10 degrees. In the next sections, we will attempt to minimize these residuals.

---

**Part IV: Moving Window Filtering**

Because, earlier, you were so helpful in determining the necessary precision of temperature gauges, your boss has now given you a promotion. You are now tasked with determining what types of setup and analysis will be most effective at reducing noise in the temperature data (i.e. maximizing the signal to noise ratio). Oh yeah, and your mathematical and plotting prowess got you a nice salary raise, too!

1. Start by making a copy of your entire first sheet by right clicking near the bottom of your Excel window on the “Temp Model + Noise” tab and selecting “Move or Copy”. This way, your temperature model, noisy data, and your carefully formatted plots are all automatically copied. Nice! Name this sheet “Moving Window Filters”. Our noisy data is already plotted, residuals are calculated, and RMS error is already completed, so all we need to do now is apply two moving window filters to the noisy data and see how well they work. Make sure you are using the same error range as before (± 10 °F) for all subsequent tests in the remainder of this exercise.


3. In column G, create an equation that apply a standard 3pt moving window filter to the “Synthetic Data”. The moving window filter equations are described in your textbook in detail on pages 16-18 and were covered in lecture. Note that the 3pt filter will cause you to lose one data point off both the beginning and end of your data. Feel free to move the plots around in your sheet as you are typing new equations if the plots are in the way.

4. In Column H, calculate the 3pt residuals. Note that the residuals are the result of the 3pt filter minus the Temp Model.

5. In Column I, calculate the squared 3pt residuals.

6. Label Cell A6, “RMS Error 3pt”, and below this, calculate the RMS error for the 3pt filter. Keep in mind that you now have two less data points, so be careful!
7. Repeat the exact same process for a 5pt moving window filter. Make sure to add in an “RMS Error 5pt” label to Cell A8 and calculate the RMS error in Cell A9. Keep in mind that the 5pt filter has lost another two points compared to the 3pt filter.

8. Add the 3pt filter results (Column G) to the existing carefully-formatted maximum daily temperature plot already in this sheet. Call this series, “3pt filter”. Your temperature model should already be plotted as a solid black line with 2.25 pt. thickness and the noisy data as small blue circles with no fill and no connecting lines. Add the 3pt filter as a dark red solid line with 1.5 pt. thickness (no symbols). In the end make sure that all of the plotted data can be easily seen on the plot. If not, consider reordering the data series, by right clicking on the plot and selecting “Select Data”. In that menu, you can reorder what is on top of what.

9. Add the 3pt residuals to the residual plot and call the series “3pt filter”. Use the same blue line with 1.5pt thickness as you did in the previous question.

10. Add the 5pt results to the two plots using the same process as before. Use a line with 1.5pt thickness and an olive green or light green color of your choice for the 5pt filter results.

---

**Part V: Stacking**

Now we will perform a 3- and 5-stack on the noisy data to see which is better at removing noise. Don’t worry though because 95% of this work is already completed! Some Excel tricks will greatly speed this up.

1. Start by making a copy of your moving window filters sheet by right clicking near the bottom of your Excel window on the “Moving Window Filters” tab and selecting “Move or Copy”. This way, your temperature model, noisy data, and your carefully formatted plots are all automatically copied. Nice! Name this sheet “Stacking”. Our noisy data is already plotted, residuals are calculated, and RMS error calculations are already completed, so all we need to do now is perform the stacking of 3 and 5 datasets to see how well they work. This shouldn’t take long.

2. Click at the top of Column E, “Residuals”. This should select the entire column. Right click in the same location and select insert. This should insert an empty column. If you look at your plots, all of your data and formatting should still be there. Repeat this three more times until you have four empty columns to the right of your “Synthetic Data (°F)” Column. Label the first cell in these empty columns “Synthetic Data 2”, “Synthetic Data 3”, “Synthetic Data 4”, and “Synthetic Data 5”. Copy and paste your equation from Column D over to these four columns to simulate having 5 total temperature gauges. Be careful to make sure that when you copied your equation over that you are still referencing the correct cells. If not, go back to your equation in Column D and add in dollar signs where needed to make sure that the source data locations do not move. In the end, Columns D-H should all yield different temperature predictions because of the random noise, but the overall seasonal patterns should be similar.

3. Columns K and N should now hold your 3pt and 5pt filters, which need to be changed into stacking. Rename the first cell in these columns to “3pt Stack” and “5pt Stack”. To perform a 3pt stack, just
average three of the “Synthetic Data” sets, and so on for the 5pt stack. The residual and residual^2 columns should still use the same formulas, but stacking does not lose any data points, so you should be able to quickly copy/paste in the formulas to fill things in.

4. The 3pt and 5pt RMS error calculations in Column A should be modified to divide by the total number of points, which is now 1096, like the original data. Recall that the moving window filters had only 1094 and 1092 total points each.

5. The plots should have automatically updated and be formatted identically to what we had before. Double check to make sure that you are plotting all of the stacking data. Update the series names to be “3pt Stack” and “5pt Stack”, and BAM! You are finished.

In the end, you should have:

a) A “Temp Model + Noise” sheet with a plot that you can use to determine how much error is acceptable along with a residual plot.

b) A “Moving Window Filters” sheet that has a plot of the predicted annual maximum temperatures with the noisy data and filter results. There should also be a residual plot of the raw noisy data and the two filter’s residuals. The sheet should also calculate the RMS error for the raw data and the two filters.

c) A “Stacking” sheet with the same items described in b) above.

| Part VI: What Plots Should I Hand In? |

Your results of this exercise should be presented a series of figures, each with a well-written caption below the figure. Keep in mind that a good geophysicist is always quantitative when discussing results (if possible). Each figure set plus caption should fit on a single page, but make sure the figures look professional and are not unusually small, distorted, or pixelated. The figures must be printed out in color and handed in as a hard copy stapled to the back of this assignment. Emailed assignments will not be accepted. Specifics are given below:

**Figure 1:** This should contain two plots and a single caption. 1a should contain the results of the moving window filters (temp model, synthetic data, and results after the filter). 1b should contain a residual plot of the raw synthetic data and the residuals of the data after the 3pt filter was applied. Be sure to mention any relevant and useful quantitative measurements of error in the caption. This caption is likely to be a bit on the long side (compared to a typical figure caption) as you will need to clearly describe what is plotted and what the results are.

**Figure 2:** Same as Figure 1, but for stacking.
After you have had a chance to look at all of your results, answer the following questions in the space provided. If your answer does not fit in the provided space, you should consider condensing it.

1) What is your modeled temperature equation and what do each of the parameters do? Do not write equations in Excel format. Your boss does not care about Excel. Only the science/math is relevant.

2) What kind of model is your temperature model? Choose from analog, analytical, conceptual, empirical, or numerical. What are the advantages and disadvantages of this type of model?

3) At what level of error (in terms of ±) does the annual pattern become completely obfuscated?
4) What temperature gauge error level (given as ±) would you recommend to your boss and why? Hint: there is no single correct answer here. Just make sure to briefly justify your reasoning.

5) What is the theoretical minimum sampling rate (in days) needed to capture the annual variations in daily high temperatures? (Hint: Think about the Nyquist stuff discussed in class)

6) What sampling rate would you recommend to your boss? Why? How you plan to avoid aliasing in your study? (Hint: There is no one right answer here, but there are incorrect answers)

7) Fill in the data table below with the RMS errors calculated from your synthetic data analysis. Include units, if appropriate.

<table>
<thead>
<tr>
<th></th>
<th>Raw Noisy Data (± 10 °F)</th>
<th>3pt Moving Window</th>
<th>5pt Moving Window</th>
<th>3 Stacked Data Sets</th>
<th>5 Stacked Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8) Which method is most successful at reducing error? Provide quantitative evidence.

9) What would need to be done differently in your study if you were going to use stacking vs. moving window filters and how would you accomplish this? (Hint: which one would be cheaper and why?)

10) What filtering technique(s) for reducing noise do you recommend and why? Feel free to be creative, but make sure that you recommendations are reasonable.