Data Expeditions: Quantifying Periodicity in Weather

Welcome to the most epic expedition of your life—a data expedition!

About the dataset:
The National Climatic Data Center has provided a dataset of daily max and min temperatures from Jan. 1, 1980 to April 1, 2015 collected from RDU International Airport. *All throughout our basic unit of time is one day. Temperatures are measured in tenths of a degree Celsius.

(Points have been assigned to questions roughly based on some combination of difficulty and significance-- don't really worry about the points, as I'm not trying to judge you, instead do as much as you can and please attempt everything. A wrong answer that is well thought-out is worth more than you might expect.)

If you need any assistance in completing this assignment, please email me at hmg@math.duke.edu --I will be happy to answer your questions.

Begin: Part I

This type of data is referred to as a time-series, and analysis of it is aptly called time-series-analysis.

Question: Before looking at the data, what do you expect is the periodicity of the data? (1pt) Explain briefly why you believe this. (2pts)

Do Step 1: Plot the entire time-series of daily MAX temperatures. (3pts)

ALWAYS DO: Please remember to provide appropriate titles and axis labels, EVERY time you plot something.

Question: What is the range of the data (provide units)? (1pt)

Question: What does the periodicity appear to be from this plot? (1pt) Please explain how you determined this from the plot? (3pts)

Question: Do you see any features that you consider interesting or anomalous in the plot? (1pt)

Do Step 2: Plot the entire time-series of daily MIN temperatures. (3pts)

Question: How does this compare with the daily MAX temperatures? What is the periodicity? (3pts)

In class, we discussed the mathematical definition of (exact) periodicity.
Fill in the blank:

A function **f** is periodic with period **p** if _____ for every **t**. (2pts)

Question: Is temperature data periodic with respect to this definition? (1pts)
Explain the difficulties of using the strict definition with respect to this data. (3pts)

In light of the previous question, researchers have invented several notions of 'almost-periodicity.'

For example, a function **f** is (e,p)-almost-periodic, if there exists a positive number ***e*>0** and a positive number ***p*>0** such that for every ***t*** the set **{f(t+np)|n is any integer}** has diameter less than ***e***.

Question: Is temperature data almost-periodic under this definition with p = 1 year, and e = 5 degrees?<Hint: it is not, explain why> Can you find an explicit counterexample in the data?(4pts)

Optional question: Look-up another definition of almost-periodicity, and state whether temperature data satisfies that definition.

Optional question: Inventing Math: Think up your own definition of almost-periodicity. Does temperature data satisfy your definition of almost-periodicity?

Great Job, Key Going: Part II

Now, we will apply sliding window transformations to the time-series.

The sliding window transformation of a signal **f** is defined as:
SW_{M,d}(f)(t) = [f(t) f(t+d) . . . f(t+(M-1)d)]

Question: What is the domain of **SW_{M,d}(f)**? (1pt) What is the codomain? (1pt)

Recall from class the terminology for the parameters as you fill in the blanks: (3pts)
M is called the _____,
d is the _____, and
M*d is the _____.

In theory, we are sliding the window continuously and indefinitely.
In practice, as discussed in class, when performing SW-transformations on data, we must introduce a new parameter to discretize the 'sliding.' We call this parameter swShift, as you will see in the Matlab code provided to you.

We must also introduce a pair of parameters for the domain of the SW-transformation. Two parameters is exactly what is needed since the domain must be a finite interval [a,b] instead of the all the reals. Notice that the interval [a,b] can be specified by the two parameters **a** and **b** but it can also be specified by it's left

endpoint **a** and the length of the interval **b-a**. Analogously, we define the left endpoint is the 'startTime' and we call the length of the interval, the 'totalSlide'

Question: For this dataset we have daily temperature recordings over 35 years—what is the smallest possible value of **d** that we can use? (2pts) What is the smallest possible value of 'swShift'? (2pt) What would have to happen in order to be able to use an even smaller value? (2pts)

Enough theory—now it's time to compute.

First Construction:
Construct the sliding window point cloud from the segment of data between January 1981 and January 1983, with **M=360**, **d=1**, and **swShift=1**. Choose the value of 'startTime' so that the first window has left endpoint at January 1, 1981. Choose the value of 'totalSlide' so that the left endpoint of the final window has left endpoint at January 1, 1982.

Compute the 0-dimensional and 1-dimensional persistent homology of this point cloud using Rips Collapse Algorithm.

Plot the 0-dimensional persistence diagram. (2pts)

Question: How many 0-dimensional persistence intervals are there? How does this number relate to the number of points in the point cloud? (2pts)

Question: Above which filtration value does the associated Rips Complex have exactly one connected component? (Hint: look at the 0-dimensional persistence intervals) (2pts)

Plot the 1-dimensional persistence diagram of the resulting point cloud and the sorted list of the persistence of all the 1-dimensional persistent classes. (2pts)

Interpretation:
Using your best judgment, answer the following:
We get few/many (circle one)'small' persistent 1-cycles,

and we get _____(exact number) large persistent 1-cycle(s).

Question: Give a reasonable interpretation of the relationship between the large persistent 1-cycle and the time-series. (3pts)

Next, repeat the first construction keeping all the parameters the same, except for the startTime. Choose 4 other startTimes(your choice), construct the point clouds, and plot the resulting persistence diagrams. (4pts)

Question: Compare the 5 persistence diagrams that you've computed so far. (2pts)

Question: If we keep all the parameters fixed in the first construction, but we change swShift from 1 to 2, then what is the relationship between the two resulting point clouds? (5pts)

In the first construction, we slide a window of size little less than a year (M=360,d=1) a total amount of 1 year, and we found one large 'loop' because of the yearly periodicity. But what would happen if we didn't slide the window a total of one year? What if we slide less or more than a year? We will explore this in the next part, and surprisingly we will find that we can actually discover the period.

Wait for the Punch Line. . . Part III

Second Construction – Varying the totalSlide
Construct a sequence of SW-point clouds as follows: Use M=360, d=1, swShift=1, and startTime corresponding to Jan 1, 1981, but vary the totalSlide from 80 to 800, incrementing by 20. This will result in _____(how many?) point clouds. Compute the persistence diagram for each.

Question: When we increase the totalSlide from k to k+20 keeping all the other parameters fixed, what is the relationship between the two point clouds? (3pts)

Question: What do you observe about the sequence of persistence diagrams? What do they have in common? How do they differ? (2pts)

Next, we will focus our attention on the most persistent 1-cycle in each of these point clouds and observe how they change.
Plot the birth and death times of the most-persistent 1-cycle as a function of totalSlide. (5pts) In the same figure, also plot the lifetime of each 1-cycle. (2pts)

Question: You will notice that the lifetimes, are small for a while, then it increases, and then levels off. Approximately when does the leveling off occur? (2pts) How does this compare with the period? (2pts)

So why does the leveling off occur? We can explain this intuitively by considering that as we increase the total slide amount we go around more of the major loop in the point cloud, and then as we increase the slide-amount beyond the period of the signal, since the windows are repeating themselves, we are going around the same major loop over and over again.

Question: Give your best attempt at explaining the jump in the birth time of the most persistent 1-cycle by considering the geometry of the point cloud. (3pts)

In conclusion, we observed how to use sliding windows and persistence to quantify the period in temperature data. Our key idea was to observe the evolution of the sliding window curve and to watch when it first closes. Due to the high

dimensionality of the point cloud and the special geometry of the curve, we used persistent homology to make these observations.

Welcome to the End!

1. **Please submit your writeup as a PDF.**
2. **Make sure to include the plots you made and the generating scripts.**
3. **Send the email to hmg@math.duke.edu and please use the subject line 'Data Expedition Submission'**
4. **Please save your PDF as LASTNAME_Expedition2016.pdf --because it keeps me organized when I look at everyone's work.**

Thank you for being a part of this Expedition!

'ProTips' for using MATLAB

ProTip#1: Using the 'F9' key: If you want to evaluate just a section of the script, highlight that section and hit 'F9'.

ProTip#2: Using 'tab' completion: Sometimes the names of variables and functions are long and tedious to type—after typing the first few characters of a variable hit the 'tab' key and voilà MATLAB will either fill out the rest of the name for you or will give you options of other possible completions.