TOPOLOGICAL DATA ANALYSIS ON U.S. PRECIPITATION DATA

EVAN MILLER
STUDENT | DEPARTMENT OF MATHEMATICS | CENTRAL MICHIGAN UNIVERSITY

INTRODUCTION

In this project, we quantify the periodicity in precipitation data from across the United States. We took data from the GHCN-Daily dataset [M1] and clustered the stations into 28 clusters using Kmeans and Ward clustering. Then, the data is analyzed with SW1PERS[P] and Fourier Analysis to determine the periodicity. We found that stations near the west coast had the best periodicity scores, and both SW1PERS and Fourier Analysis agreed on which clusters were highly periodic. This search is motivated by a need to give mathematical rigor to the patterns we see in meteorological data collecting, as well as a method of testing SW1PERS against a more conventional signal analysis method on a real world dataset.

DISCRETE FOURIER TRANSFORM

The goal of a Fourier Transform is to take a time series and map it to a frequency series. In this project, we examine weekly precipitation accumulation over time and try to recover cycles from the Fourier series. Our data is a discrete signal, so we apply a Discrete Fourier Transform.
The key to understanding the Fourier Transform is to identify our finite range of time with the unit circle in $\mathbb{C}$. Our time series is now a function that sends points on this circle to a real valued sequence. The space of these functions in fact form a vector space with an inner product given by

$$\langle (f(\frac{m}{n})), (g(\frac{m}{n})) \rangle := \frac{1}{n}\sum_{m=0}^{n-1} f(\frac{m}{n})\overline{g(\frac{m}{n})}$$

and an orthonormal basis given by

$$\{e^{2\pi i k\frac{m}{n}} \mid k = 0, \ldots, n-1\}$$

where f, g are functions defined on our sequence on the unit circle [W]. This defines for us

$$\langle (f(\frac{m}{n})), (e^{2\pi i k\frac{m}{n}}) \rangle = A_k = \frac{1}{n}\sum_{m=0}^{n-1} f(\frac{m}{n})e^{-2\pi i\frac{mk}{n}}$$

as the Discrete Fourier transform of a time series f.

SW1PERS

The Sliding Window is a function $SW_{d,\tau}f(t)$ that sends $t \in \{1, \ldots, N-\tau\}$ to points in $\mathbb{R}^{d+1}$. We define a sliding window by

$$SW_{d,\tau}f(t) = \begin{pmatrix} f(t) \\ f(t+\tau) \\ \vdots \\ f(t+d\tau) \end{pmatrix}$$

where τ is the step size between points in the window and $d+1$ is the dimension of the embedding. Applying this function to each point in our time series, we generate a collection of points seen in Figure 1 called the sliding window point cloud, to which we will apply Persistent Homology on a Rips Complex.
The Persistent Homology of a simplicial complex measures how long a "hole" lasts in a filtration of complexes on our point cloud. This is done by taking the set of all cycles of simplices and modding out those cycles that form a boundary of a higher dimensional simplex, leaving only the cycles that are "holes." We construct our filtration by increasing the distance of a Rips Complex. In the Rips Complex, we connect two pints by an edge if they are within the specified distance of each other, then we include a set of k points as a k complex if they are all connected to each other by exactly one edge. The change in distance between when a cycle that is not a boundary is created and when it becomes a boundary is the persistence of the homolgy class, which we use to quantify the periodicity of our time series. An example of this can be seen in Figure 2.

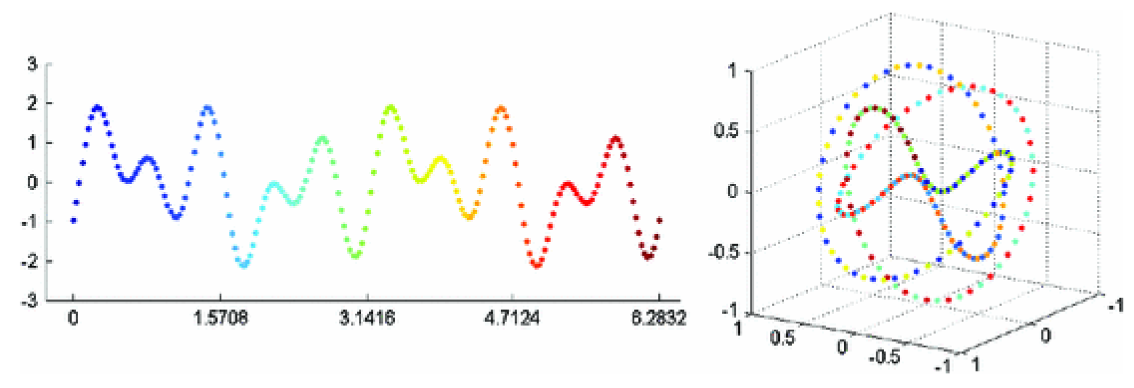


Figure 1: An example of a time series being mapped to a point cloud via sliding windows (from [P]).

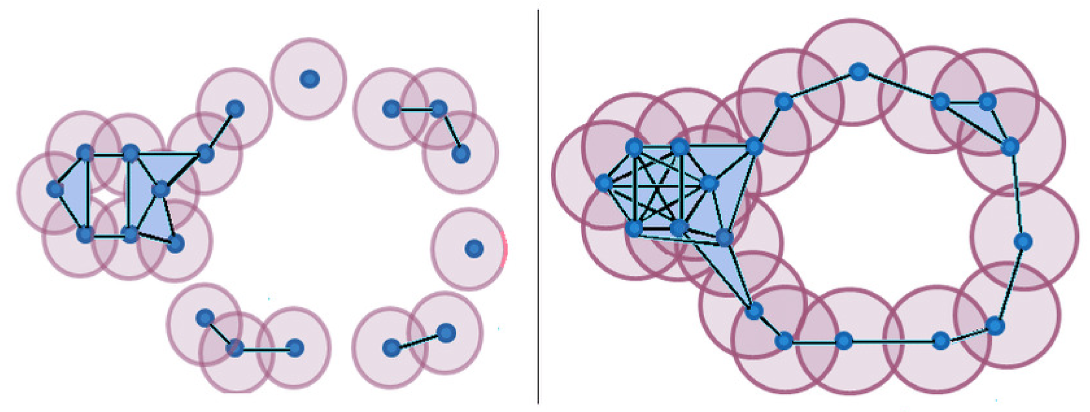


Figure 2: As the radii of the balls are increased, a homology class is born in the small circle of the complex on the left and dies at the complex on the right. A new homology class is formed on the large circle on the right (from [F]).

REFERENCES

[F] Gonzalez, Georgina & Ushakova, Arina & Sazdanovic, Radmila & Arsuaga, Javier. (2019). Prediction in cancer genomics using topological signatures and machine learning.

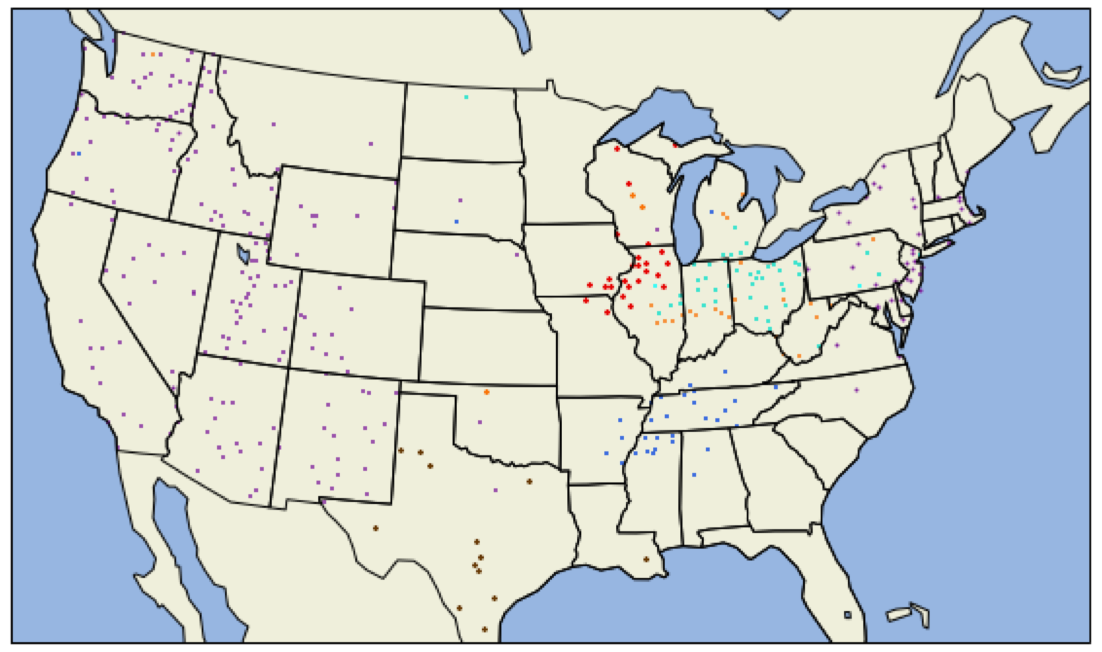
[P] Perea, J.A., Harer, J. *Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis*. Found Comput Math 15, 799-838 (2015). https://doi.org/10.1007/s10208-014-9206-z

[M1] Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston, 2012: *An overview of the Global Historical Climatology Network-Daily Database*. Journal of Atmospheric and Oceanic Technology, 29, 897-910, doi.10.1175/JTECH-D-11-00103.1

[W] Jordan Watts, "Applied Fourier Analysis", 2020 (unpublished notes).

RESULTS

Overall, SW1PERS and the Discrete Fourier Transform had similar results when used to detect year long cycles with only a few exceptions. In the figure below, we can see that these clusters are mostly around the Great Lakes region, with one other cluster spanning the Rocky Mountain and West Coast regions. We also discovered that SW1PERS was superior for detecting half year cycles.



NEXT STEPS

1. Our clusters could be correlated with known climatic regions to justify our clustering.
2. Cluster periodicity can change over time, which we conjecture is due to the evolution of climatic regions over time, so allowing the stations included in a cluster to change dynamically may help improve scores.
3. Other cycle lengths have not been investigated, and these could provide insight into greater patterns in precipitation data.

CONTACT INFORMATION

Email mille7em@cmich.edu