TOPOLOGICAL DATA ANALYSIS ON U.S. PRECIPITATION DATA

EVAN MILLER

In this paper, 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.

1. Methods

1.1. **SW1PERS.**

1.1.1. *Sliding Window Point Clouds.* Our first goal is to represent cycles in our time series as a collection of points in a high dimensional space so that we may characterize their shape. Let $f(t)$ be a time series of N real values. The Sliding Window is a function $SW_{d,\tau}f(t)$ that sends $t \in \{1, \dots, N-\tau\}$ 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 Topological Data Analysis. The essential idea is that the periodicity of the time series is captured in the circularity of this sliding window point cloud, which we will compute using Persistent Homology. In order to examine periodicity for a cycle of a certain length, we set the window size to correspond to that cycle. In this paper, we examine cycles of 1 year and half a year, measured in weeks, so we look at windows of sizes 52 and 26 weeks.

1.1.2. *The Rips Complex and Homology.* Now that we have our point cloud, we compute the Rips Complex of this dataset. Computing the Rips Complex is fairly simple: for a distance ϵ, we construct the 1-skeleton by including the edge $\{x_i, x_j\}$ in the skeleton if $d(x_i, x_j) \leq \epsilon$ for x_i, x_j points in our point cloud. Then, we include a subset of k points as a k-simplex

Date: February 14, 2021.

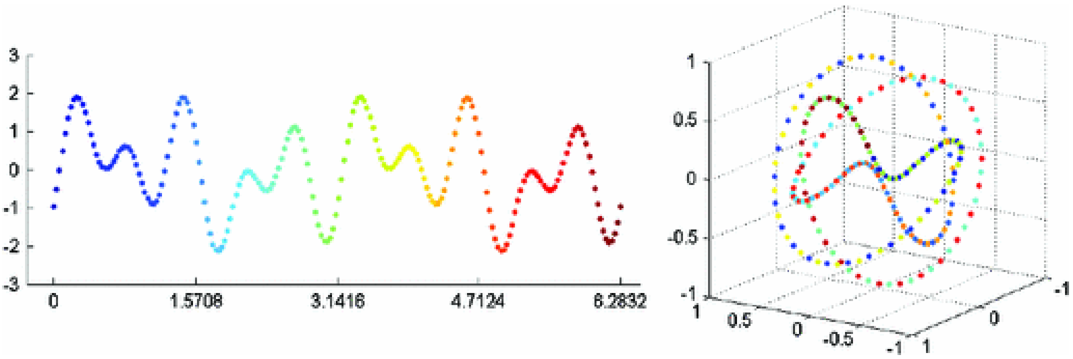

FIGURE 1. An example of a time series being mapped to a point cloud via sliding windows (from [P]).

if each of the points is connected by an edge and denote this collection of simplices K, a simplicial complex.

We now consider the space $C_k(K)$, which is the space of k-chains in K. Recall that a k-chain is a finite formal sum of k-simplices with coefficients from a field with a prime number of elements $\mathbb{F}_p$. We also define the boundary of a k-simplex by the formal alternating sum of the boundaries of $k-1$ dimensional faces of the simplex. We in turn define the boundary of $c \in C_k(K)$ as the sum of the $(k-1)$-simplices with their associated coefficients that construct it, except when $k=0$ we define the boundary to be the sum of the coefficients. To summarize,

$$\partial(c) = \sum_j \gamma_j \partial(b_j)$$

where $\gamma_j \in \mathbb{F}_p$, $b_j \in K$ is a k-simplex and ∂ is the boundary operator.

We now define $Z_k(K)$ to be the subspace of $C_k(K)$ where $\partial(c) = 0$, a k-cycle, for each $c \in Z_k(K)$. Define as well $B_k(K)$ as the subspace of $C_k(K)$ where $c \in B_k(K)$ is the boundary of a $k+1$-chain. One can check that the boundary of a boundary is 0, so we have that $B_k(K) \subseteq Z_k(K)$. Then, we have a quotient group $Z_k(K)/B_k(K) = H_k(K)$ which we call the kth simplicial homology group with $\mathbb{F}_p$ coefficients. Intuitively, this is the set of cycles in K that do not form a boundary, or in other words, the "holes" of K. The rank of these sets allows us to count these components. The only case that may not match our intuitive understanding of a hole is $H_0(K)$, which tells us the cycles of points we can construct that are not boundaries of chains of edges. This happens when we take chains of cycles that are not connected by edges. That is, $\operatorname{rk} H_0(K)$ is the number of path connected components in our complex.

As an example, consider the simplicial complex T in Figure 2. We have $\operatorname{rk} H_0(T) = 3$ as there are three connected components, $\operatorname{rk} H_1(T) = 1$ since there is one chain of 1-simplices (the empty triangle in the bottom right) which is not a boundary of a 2-simplex, and $\operatorname{rk} H_3(T) = 0$ since the only cycle of 2-simplices defines the boundary of a 3-simplex (the tetrahedron in the bottom left).

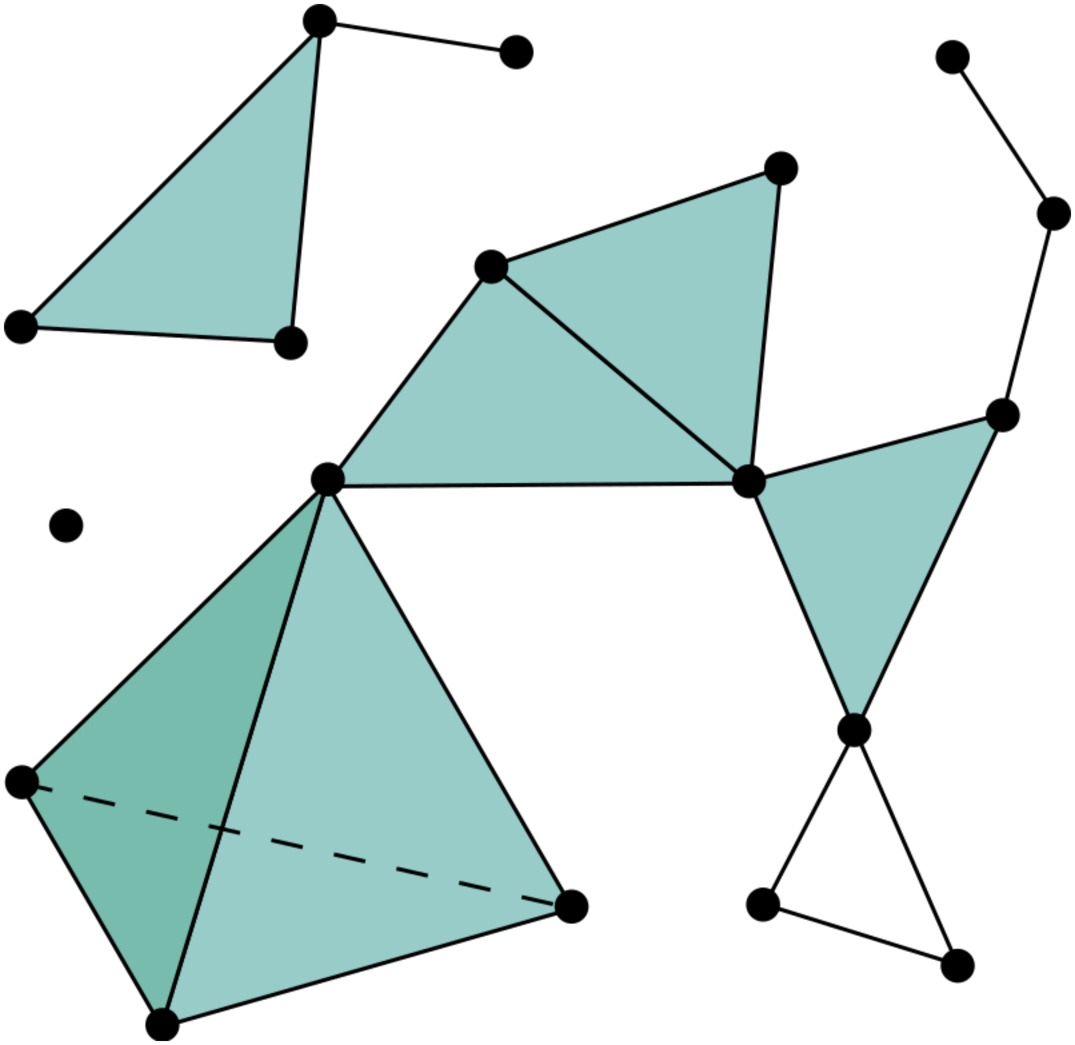

FIGURE 2. A simplicial complex consisiting of simplices of orders 1-3 (from [W]).

We will use this notion of holes to measure the "circularity" of our point cloud. We expect that for a circular point cloud, we have one large hole and much smaller holes along the boundary of that hole. That is, we expect to find an element of $H_k(K)$ that is present for a relatively larger range of ϵ than any other elements. We call this range persistence, and we define it more rigorously below.

1.1.3. *Filtration and Persistence.* Now that we have a notion of the holes of our point cloud, we must find a way to quantify the "circularity" of our holes. Consider that as we increase ϵ, the resulting Rips Complex will contain all of the simplices of lower values of ϵ. Also, the complex resulting from $\epsilon = 0$ denoted K_0 is empty, and there is an m such that all $\epsilon \geq m$ causes the points of our point cloud to form a single simplex K_m. Then, we have a filtration $K_0 \subset K_1 \subset \cdots \subset K_m$ for our point cloud. We say that a homology class is born at K_i if it is in $H_k(K_i)$, but not in the image of $H_k(K_{i-1})$ under the inclusion $K_{i-1} \subset K_i$. This means that a cycle now exists that is not a boundary of a $k + 1$ chain in K_i. A homology class

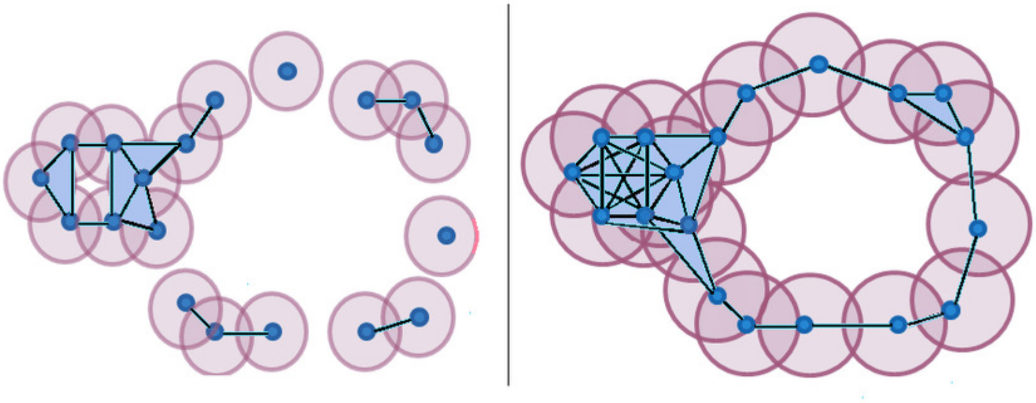

FIGURE 3. As the radii of the balls are increased, a homology class is born in the small circle of the complex on the left and dies at the complex on the right. A new homology class is formed on the large circle on the right (from [F]).

born at K_i dies at K_j if it is not in the image of $H_k(K_{i-1})$ under the inclusion $K_{i-1} \subset K_{j-1}$ but is in the image of $H_k(K_{i-1})$ under the inclusion $K_{i-1} \subset K_j$. This means that the cycle born at K_i is now the boundary of a $k+1$ chain in K_j. We call $j-i$ the persistence, or index persistence, of the homology class, and we measure the "circularity" of our point cloud by its maximum persistence for all of its homology classes. We then finalize a score from SW1PERS as $1-s$ where s is a function of the maximum persistence of H_1 over our filtration with values between 0 and 1. As an example, see Figure 3 (You may notice that the complex in this figure is in fact a Čech complex, but these lead to equivalent results as our cases.)

1.1.4. *Smoothsplining and Data Sampling.* To convert our weekly data into something easier to work with, we fit a curve to our data using the process of smoothsplining. Smoothsplining is a method of fitting a function $s(x)$ that is piecewise a polynomial to a set of data points by minimizing the following

$$p\sum_i (y_i - s(x_i))^2 + (1-p)\int (\frac{d^2 s}{dx^2})^2 dx$$

where (x_i, y_i) are the pairs of times and values of our time series at that time and p is a constant in $[0,1]$ that is optimized for the smoothest curve. For further details, see [S]. This allows us to resample any number of points we choose from a window of any size. It is shown in [P] that window size should correspond to the length of the cycle, and the number of samples taken from the window (the dimension of the embedding) should be as high as computationally possible.

1.2. **Fourier Analysis.**

1.2.1. *Defining the 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 implementation of the Discrete Fourier Transform, of which there are many, that is used in numpy [NP] defines the Discrete Fourier Transform as follows.

The key to understanding the Fourier Transform is to identify our finite range of time with the unit circle in $\mathbb{C}$. Begin by viewing a time series as a function whose domain is a finite collection of evenly spaced points on our time interval. We can then look at a compact interval $[0, n\Delta t]$, where Δt is the space between each time sample and n is the number of samples taken. We can "wrap" our compact interval of time around the unit circle, and end up with a way to identify our finite collection of points with points on the unit circle. Our time series is now a function that sends these points to a complex (or real, as is our case) 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 tells us that for any function f defined on our sequence,

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

where

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

is the Discrete Fourier transform. We then evaluate these sums in a computationally efficient way given by James W. Cooley and John W. Tukey [CJ] and graph each $|A_k|$ as the quills of the Fourier Transform.

1.2.2. *Interpreting the Discreet Fourier Transform.* From this, we can immediately see that the quills of the Fourier Transform are complex conjugates after $k = \frac{n}{2}$. Since we are primarily searching for annual and semi-annual precipitation cycles, we further limit our view to $k = 1, \ldots, 104$. This enhances visibility of quills drastically. To understand k as a unit of cycle length, divide k by 52. This number is the number of cycles in a year. Now, we can examine the periodicity of our data by looking directly at the graphs of $|A_k|$. For example, the station cluster in Figure 4 has a relatively large quill at $k = 52$, which corresponds to an annual cycle, while the station cluster in Figure 5 has no relatively large quills.

Further details about Fourier Analysis can be found in [FO] and [R].

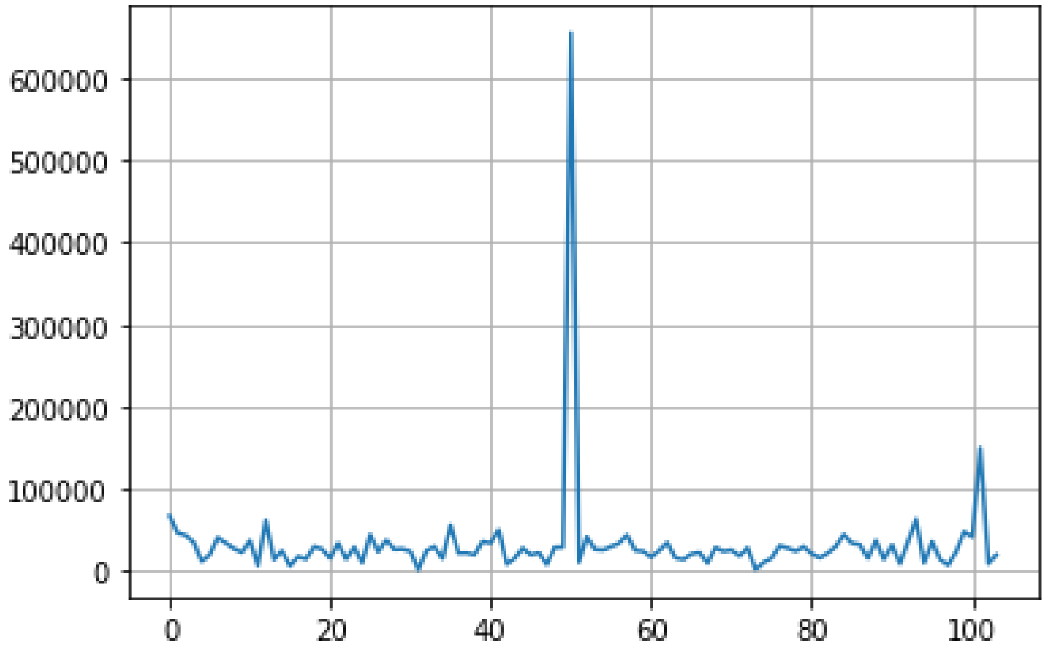

Figure 4. An example of a Fourier Analysis graph with a prominent quill at $k = 52$, corresponding to 1 year.

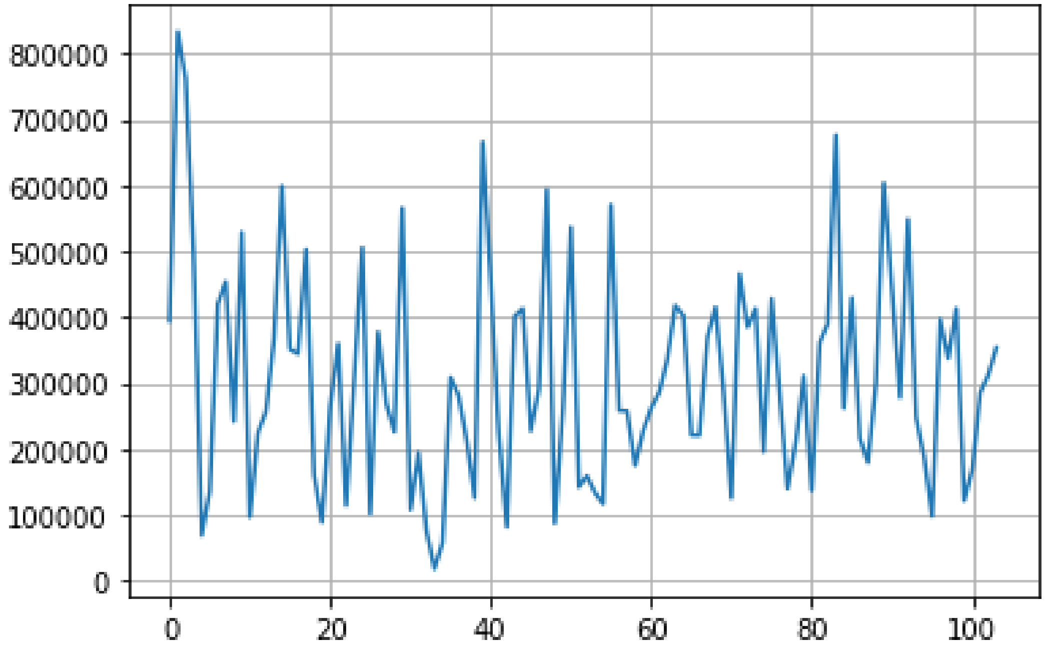

Figure 5. An example of a Fourier Analysis graph with no prominent quills.

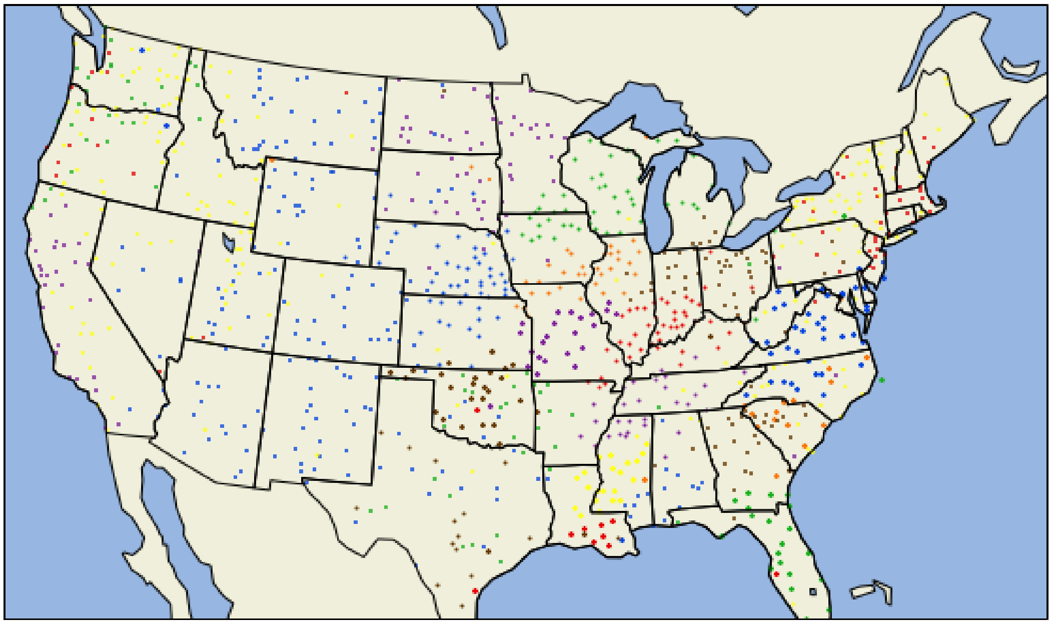

FIGURE 6. Stations grouped into 28 clusters using Kmeans. Each combination of color and point type is a different cluster.

1.3. **Clustering.** In this project, we are interested in quantifying the periodicity of precipitation data from groups of stations to make generalizations about the areas these groups cover. To that effect, we group our stations into groups of 14 and 28 stations based on similarity in their reported precipitation. We produce clusters using both Kmeans [K] and Ward [WA] algorithms on weekly precipiatation data from 1960 to 2010 and sum each week over each station to get a precipitation value for the cluster's region. A visual representation of these clusters is given in Figures 6 and 7.

2. Data

The Global Historical Climate Network Daily (GHCND)[M2] is a free online database hosted by the National Oceanic and Atmospheric Administration. It contains climate records from weather stations around the globe. These stations report a wide range of climate related data, from weekly temperature ranges to daily snow depth. The data is sourced through a variety of channels and compiled into a unified format. Each station has a unique file that contains lists of daily values for every month and data type pairing available from the station. These stations have descriptions in a metadata file, where one can determine the geographic coordinates of the station as well as any certifications it has from quality control organizations like the GCOS Surface Network [G], the U.S. Historical Climatology Network [H], or the U.S. Climate Refererence Network [C]. The GCOS Surface Network requires that participating stations report at least temperature and precipitation data. Furthermore, stations have to submit monthly CLIMAT reports to be verified. The U.S. Historical Climatology Network selects stations based on spatial coverage, record length, data completeness, and historical stability. Furthermore, older records have had minor corrections made to account for bias

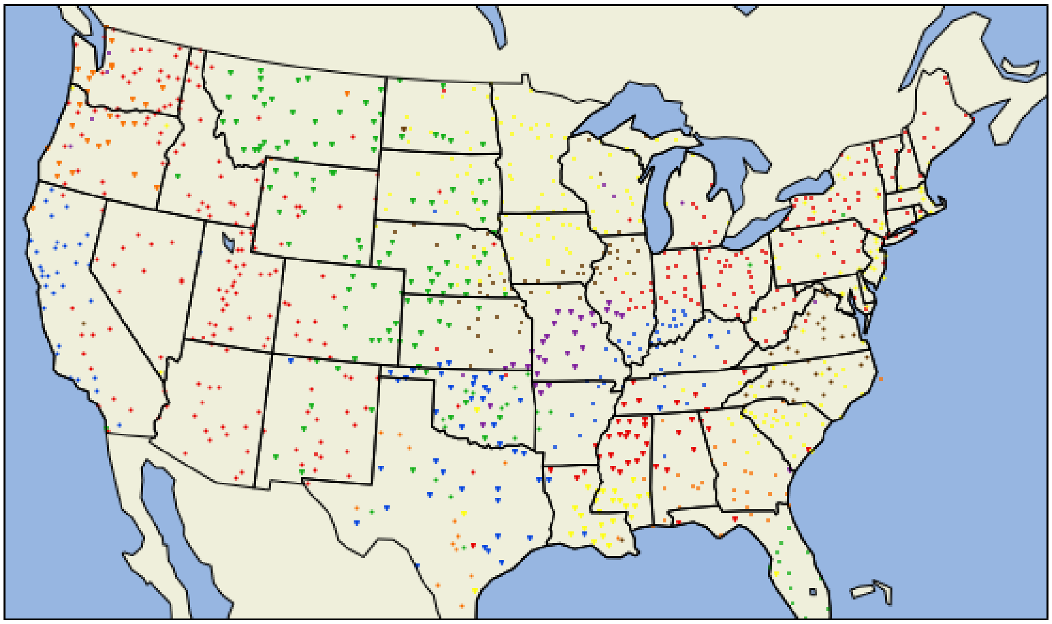

FIGURE 7. Stations grouped into 28 clusters using Ward. Each combination of color and point type is a different cluster.

from poor instrumentation. The U.S. Climate Refererence Network includes stations that receive regular calibrations to their equipment and a high standard of accuracy. These stations are all required to report both temperature and precipitation, as well as surface temperature, solar radiation, surface winds, soil conditions, and relative humidity.

From this dataset, we select those stations that have a quality certification and are located in the United States. Gaps between precipitation measurements are also measured, and those stations with over 8 sequential missing weeks are removed from consideration. As a further measure, data from 1960 to 2010 is used in this project, so stations that either started after January 1, 1960 or ended before December 31, 2010 are removed as well. This leaves a set of 1061 stations that covers the United States with the densest areas in the eastern half of the country as depicted in Figure 8.

Daily precipitation data is taken from each of these stations and grouped into weeks. To account for extra days coming from the last day of the year and the extra day in February on leap years, an eight day week is formed by appending the extra day to the preceding week. When data is missing, we fill the gap with a thirty year average of precipitation from the same day. Daily precipitation amounts are summed over each week and rerecorded as a weekly value. After this, we are left with a smooth signal of uniform length from each station to perform our analysis with.

3. RESULTS

The following table includes the SW1PERS scores for our Ward and Kmeans clusters for an annual cycle.

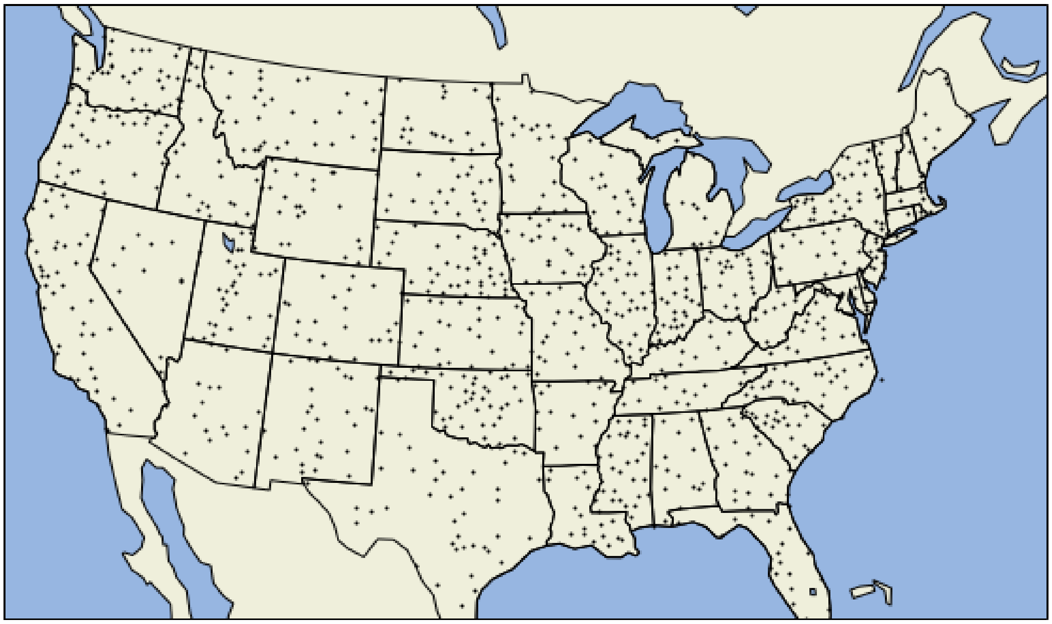

FIGURE 8. The set of all stations in our dataset after cleaning.

id	Kmeans	score	Ward	score
0		0.101928		0.199827
1		0.759343		0.247740
2		0.822254		0.629730
3		0.059355		0.038898
4		0.922969		0.556887
5		0.742610		0.843636
6		0.128635		0.575833
7		0.791889		0.398114
8		0.646822		0.158928
9		0.573878		0.893462
10		0.211225		0.043105
11		0.809996		0.611116
12		0.180933		0.193820
13		0.168364		0.045841
14		0.066144		0.160054
15		0.590608		0.138847
16		0.092759		0.957116
17		0.277278		0.164702
18		0.070092		0.281301
19		0.211887		0.510337
20		0.268669		0.626629
21		0.716011		0.054393
22		0.914671		0.795991
23		0.242086		0.680759
24		0.682648		0.847786
25		0.079037		0.059001
26		0.758512		0.056135
27		0.537304		0.740129

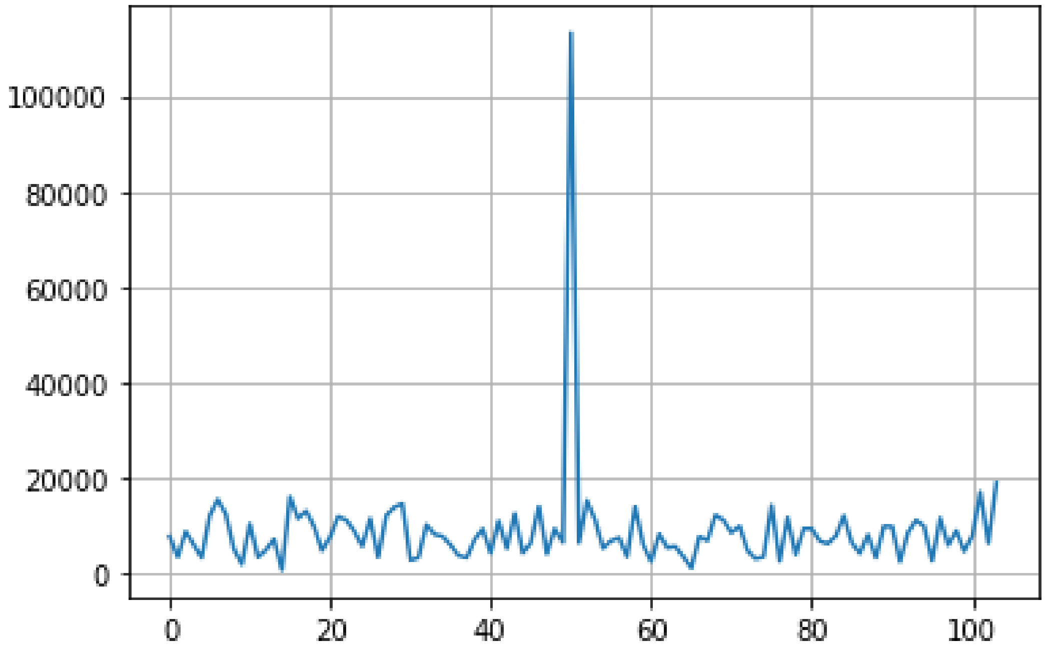

Figure 9. Cluster 22's Fourier Transform graph from Kmeans

Overall, SW1PERS and the Discrete Fourier Transform had similar results when used to detect year long cycles with only a few exceptions. Clusters 22 and 26 of our Kmeans clusters as well as clusters 16 and 27 of our Ward clusters had a significant quill at $k = 52$ in our Fourier Transform graph (seen in Figures 9, 10, 11, and 12), but high SW1PERS scores. On the other hand, cluster 8 of our Ward clusters did not have a significant quill at $k = 52$ in our Fourier Transform graph (seen in Figure 13), but low SW1PERS scores. In Figure 14, 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.

Figures 16, 15, 18, and 17 illustrate the geographic locations of clusters that achieved benchmarks of below 0.5 and 0.25 SW1PERS scores.

Half year cycles produced more consistent SW1PERS scores, with nearly all stations scoring around either 0.01 or 0.7. While the half year cycles discovered in Fourier Analysis corresponded to low SW1PERS scores, many other stations that did not have significant half year quills were discovered to be periodic through a low SW1PERS score.

Taking Kmeans 26 and Ward 8 as samples of our exceptions, we can look at our raw signal to check for patterns and irregularities. Random 2 year samples of these signals can be found in Figures 19, 20, 21, 22. For comparison, a sample from a series that both agree is periodic is given in Figure 23.

Furthermore, SW1PERS analysis on 10 year subsequences of our stations show that periodicity of the cluster can vary drastically over time. The SW1PERS scores for Ward cluster 8 is given below.

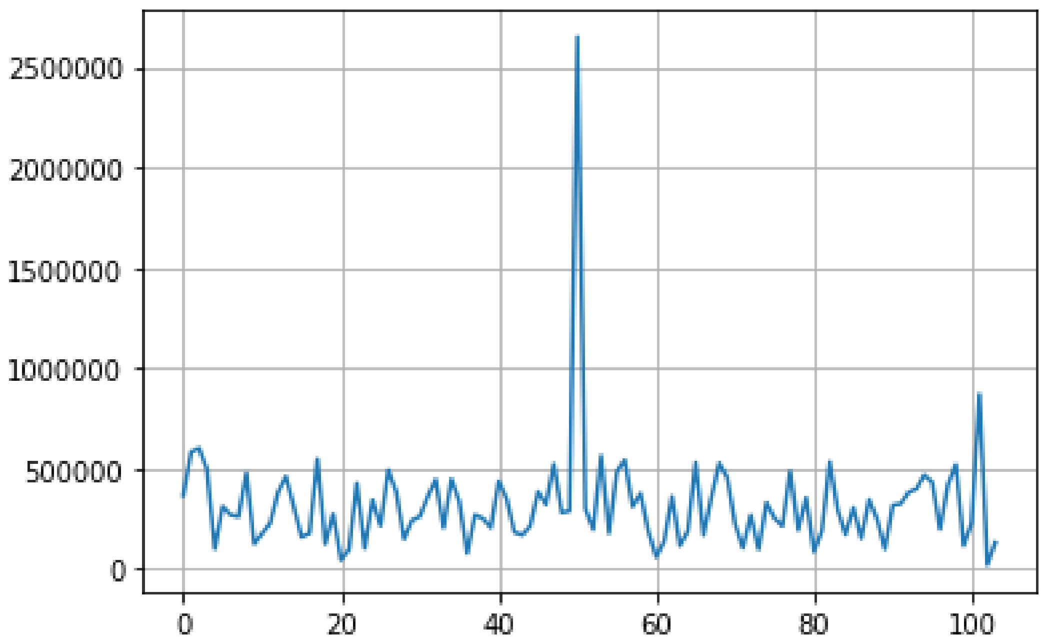

FIGURE 10. Cluster 26's Fourier Transform graph from Kmeans

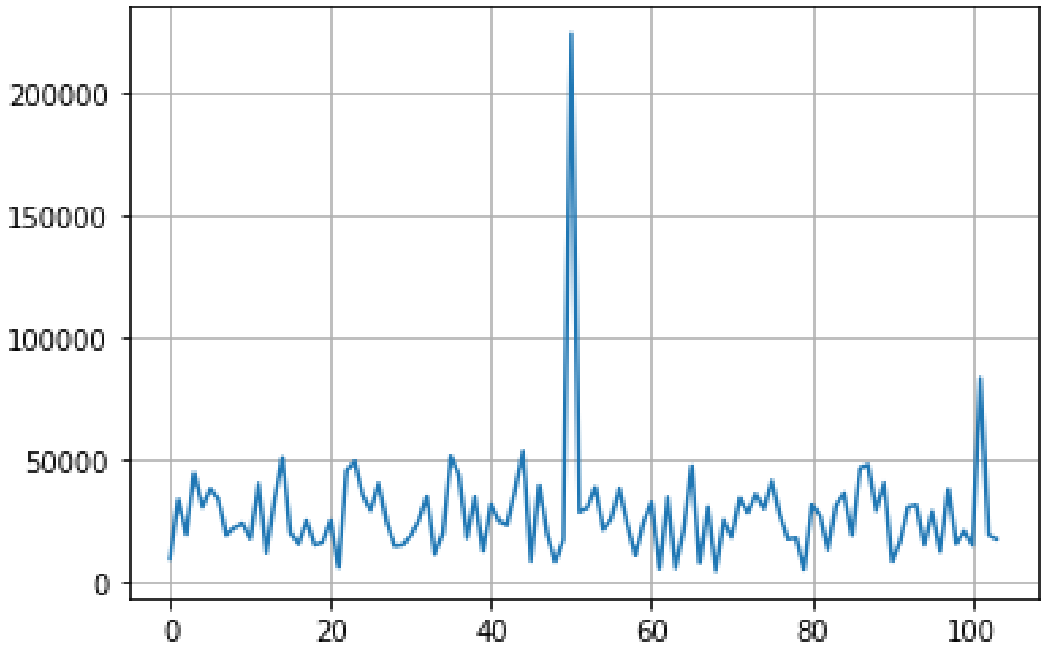

FIGURE 11. Cluster 16's Fourier Transform graph from Ward

years	score
1960s	0.152528
1970s	0.796802
1980s	0.680483
1990s	0.285848
2000s	0.767930

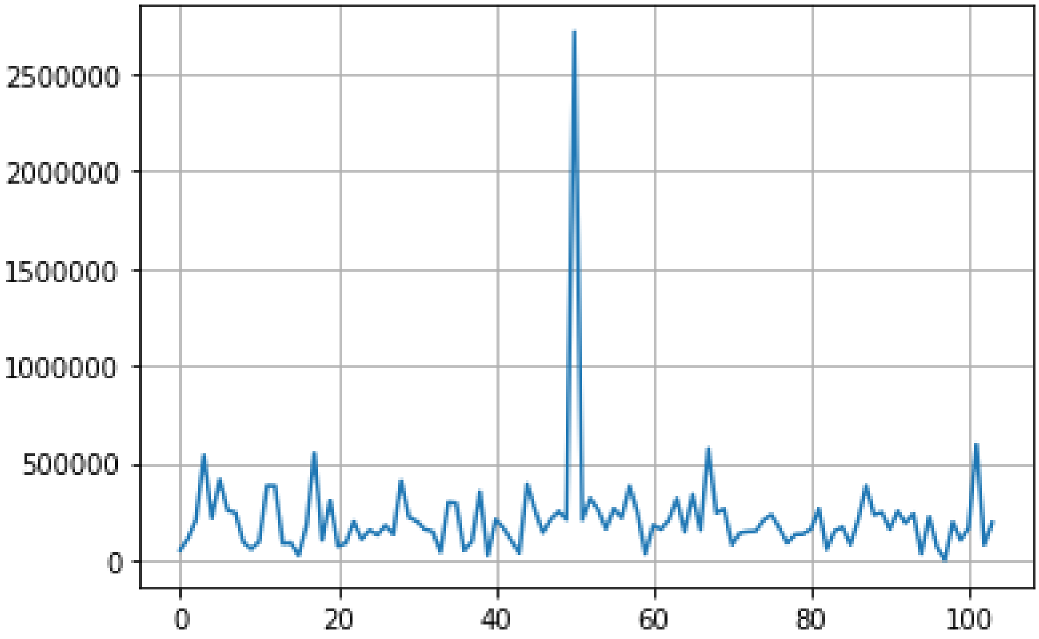

Figure 12. Cluster 27's Fourier Transform graph from Ward

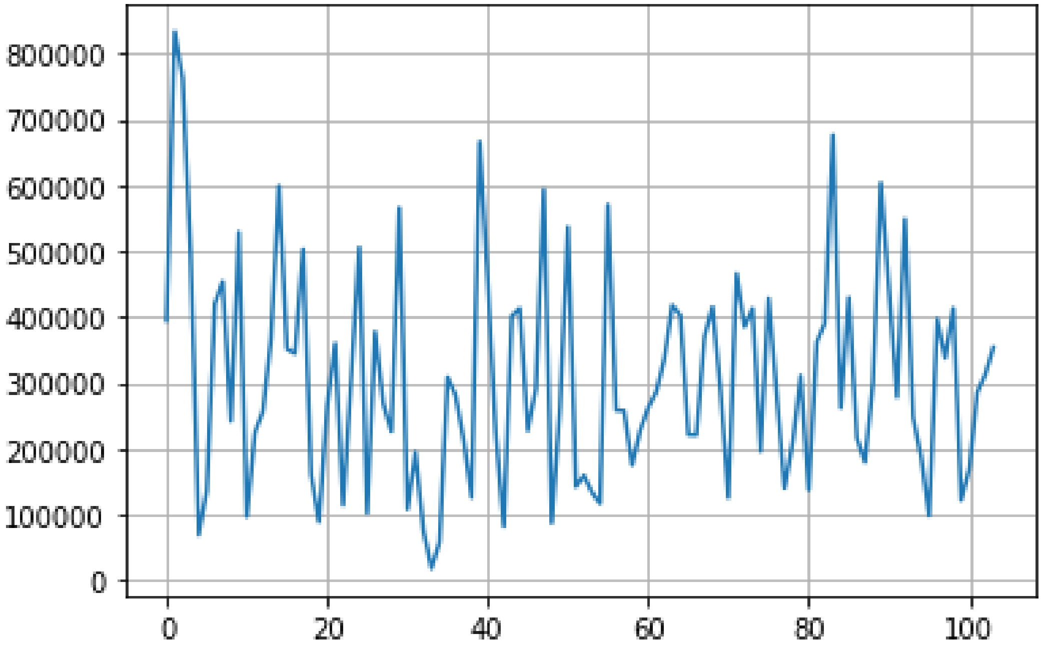

Figure 13. Cluster 8's Fourier Transform graph from Ward

4. Conclusion

As an evaluation of SW1PERS, we find that SW1PERS adequately quantifies periodic signals that are detected by Fourier Analysis. Furthermore, we can see from the raw signals

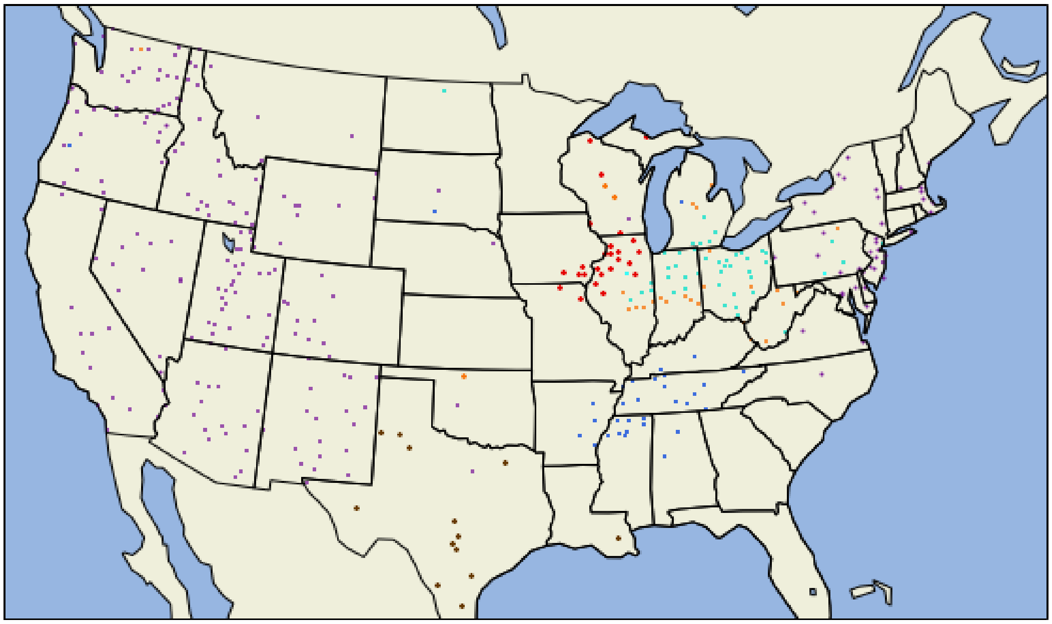

Figure 14. Clusters from both Ward and Kmeans whose SW1PERS scores did not correspond to their Fourier Analysis

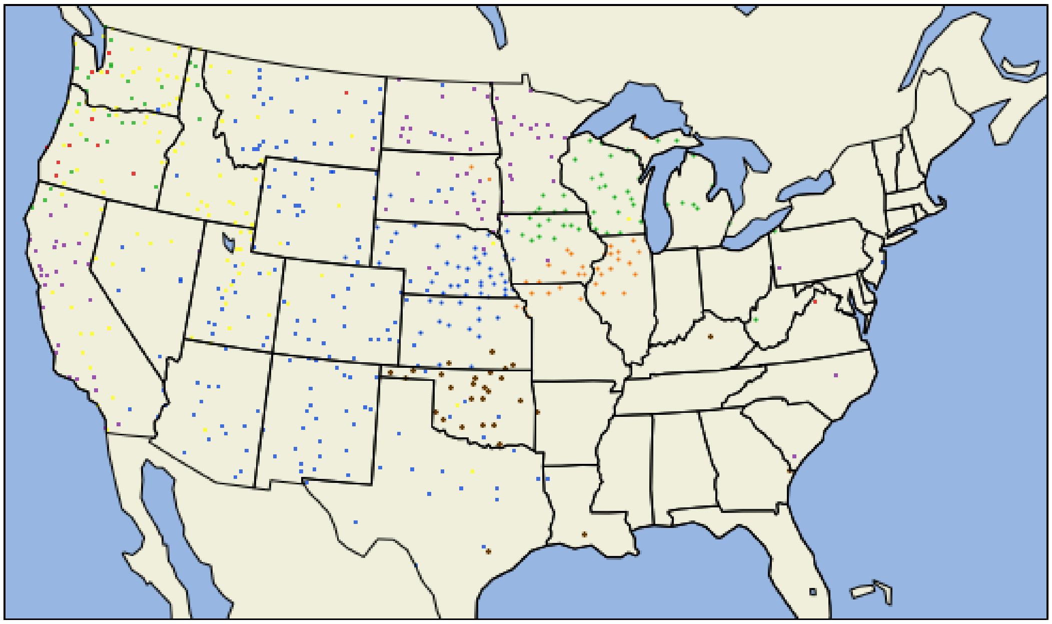

Figure 15. Clusters from Kmeans with a SW1PERS score under 0.5

(such as 23) that are determined to be periodic by Fourier Analysis usually experience a wet season in the middle of a year and a dry season on either end, an oscillatory pattern. The pattern is no longer oscillatory when looking for a half year cycle, so Fourier Analysis starts to fail to find periodicity. However, SW1PERS is shown to effectively pick up on these signals with irregular patterns.

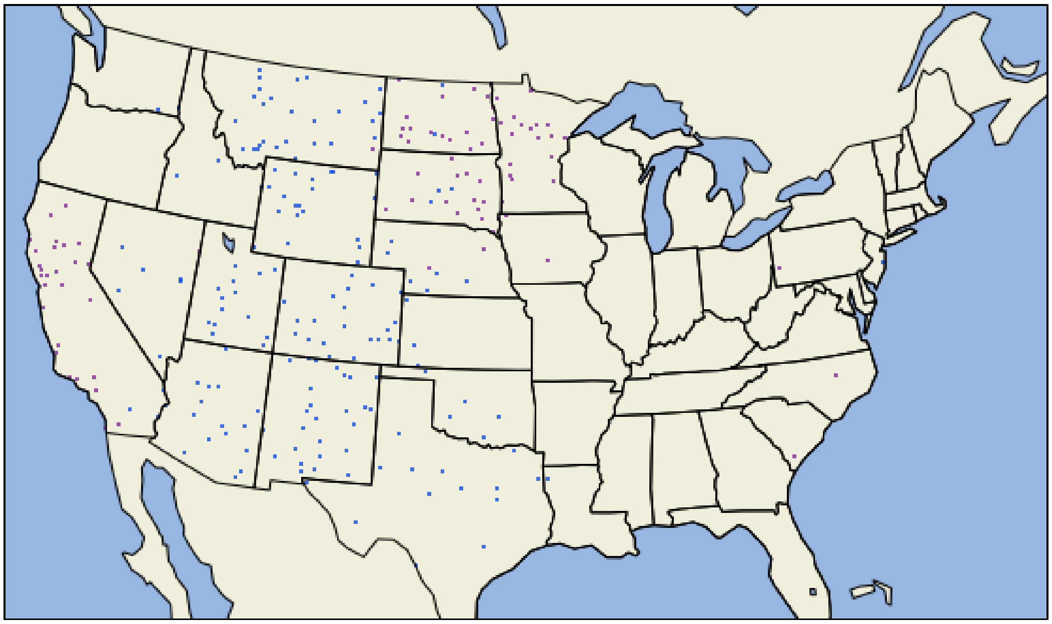

Figure 16. Clusters from Kmeans with a SW1PERS score under 0.25

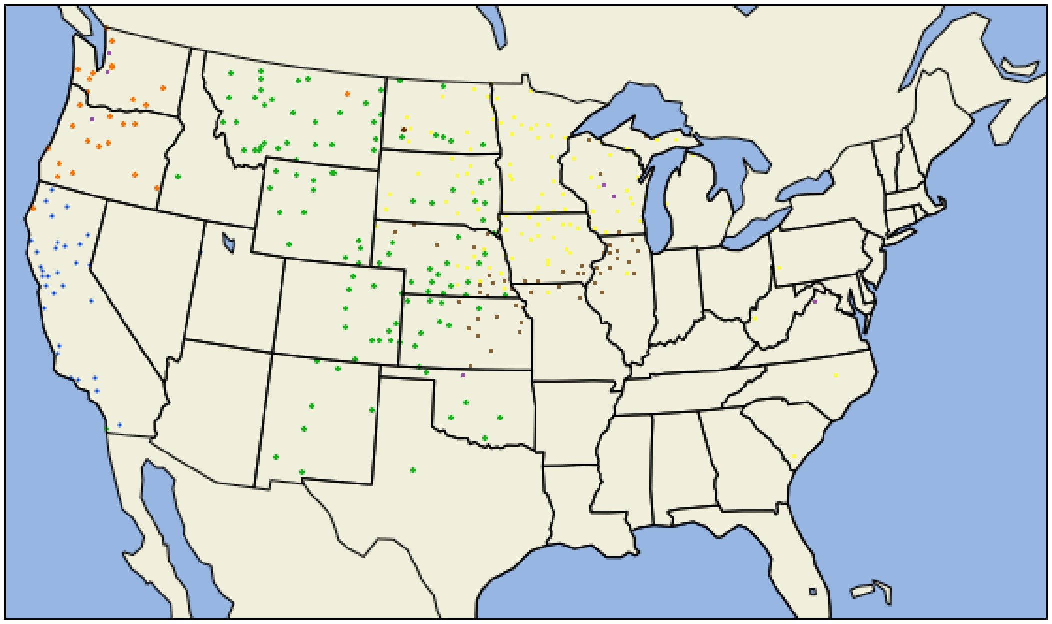

Figure 17. Clusters from Ward with a SW1PERS score between 0.5 and 0.25

There are many promising extensions of these results. Firstly, our clusters could be correlated with known climatic regions to justify our clustering. Second, it is shown in Table 3 that cluster periodicity can change over time. We conjecture that this is due to the evolution

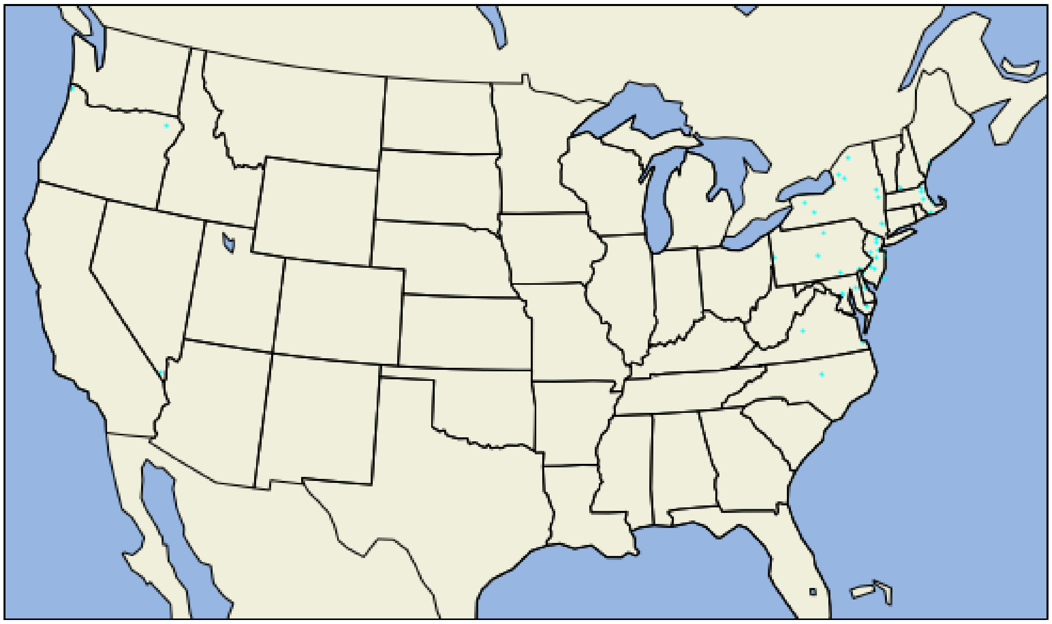

Figure 18. Clusters from Ward with a SW1PERS score under 0.25

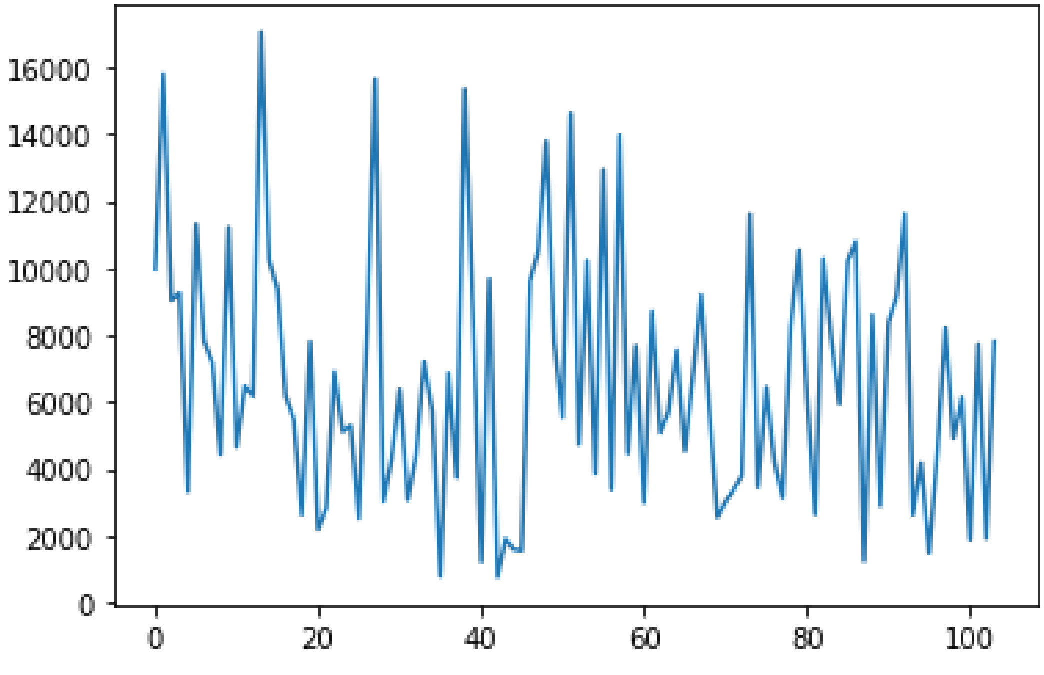

Figure 19. Raw signal from Ward cluster 8. This sample was taken over 1960 to 1962.

of climatic regions over time, and allowing the stations included in a cluster to change dynamically may help improve our scores. Third, other cycle lengths have not been investigated, and these could provide insight into greater patterns in precipitation data.

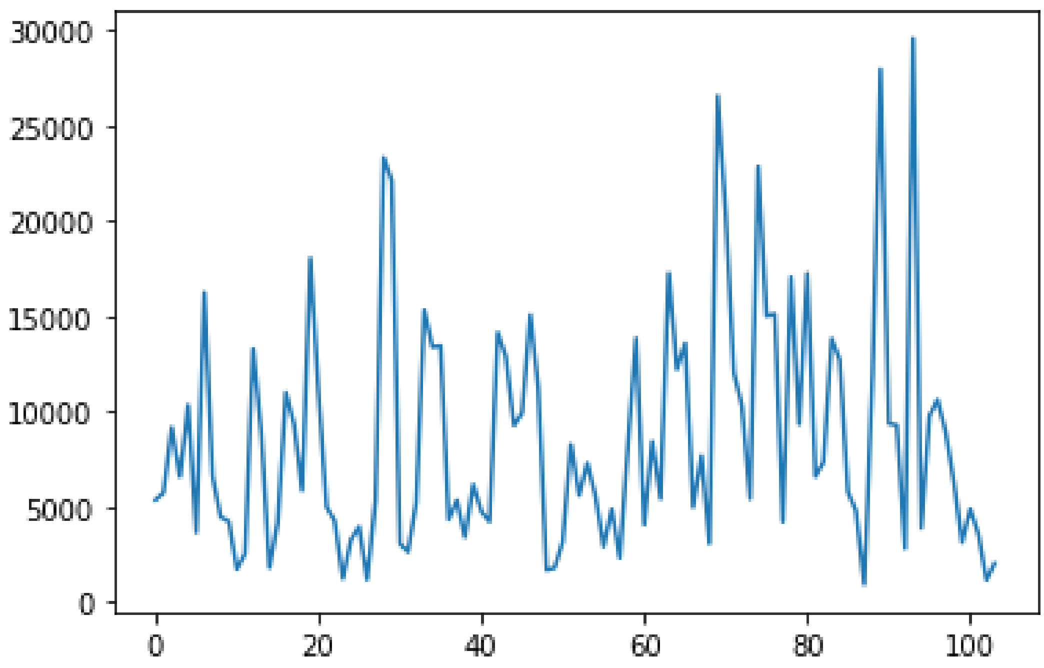

Figure 20. Raw signal from Ward cluster 8. This sample was taken over 1960 to 1962.

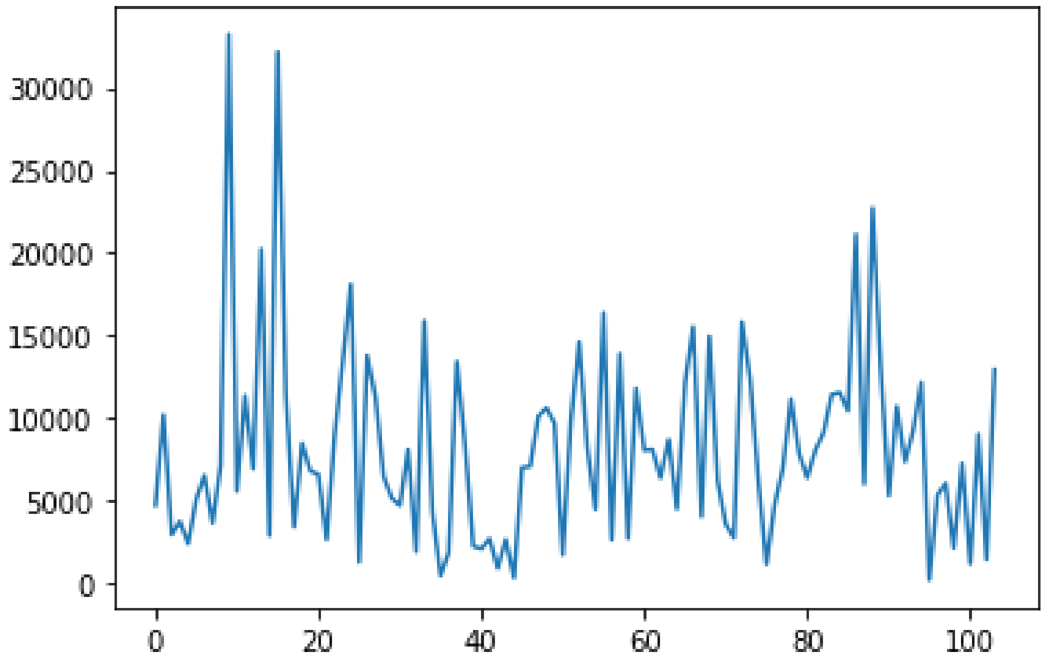

Figure 21. Raw signal from Kmeans cluster 26. This sample was taken over 1960 to 1962.

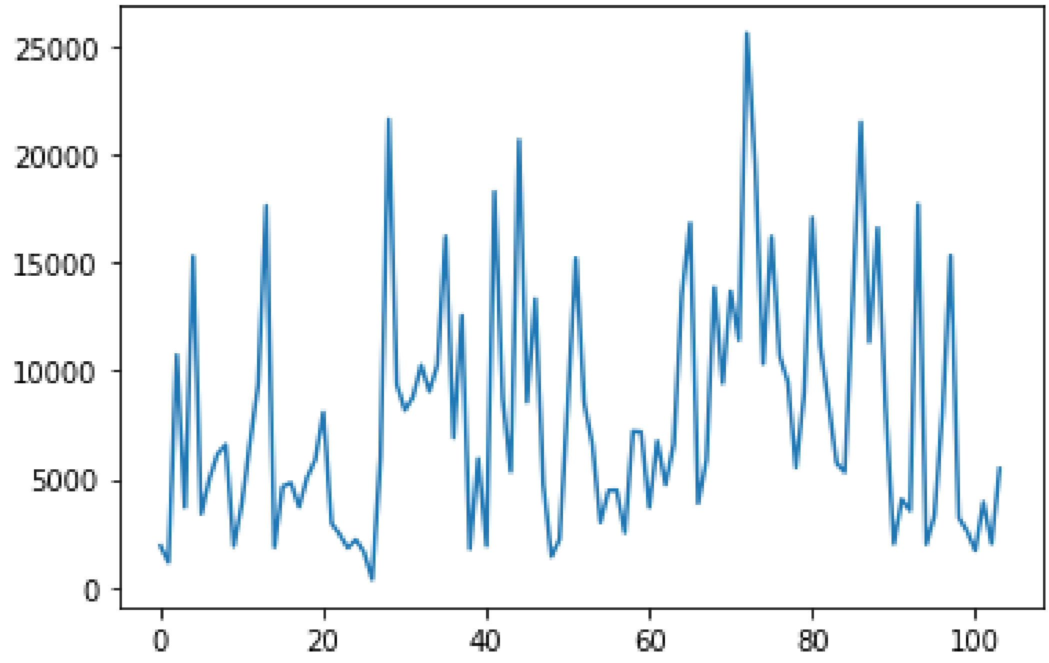

Figure 22. Raw signal from Kmeans cluster 26. This sample was taken over 1960 to 1962.

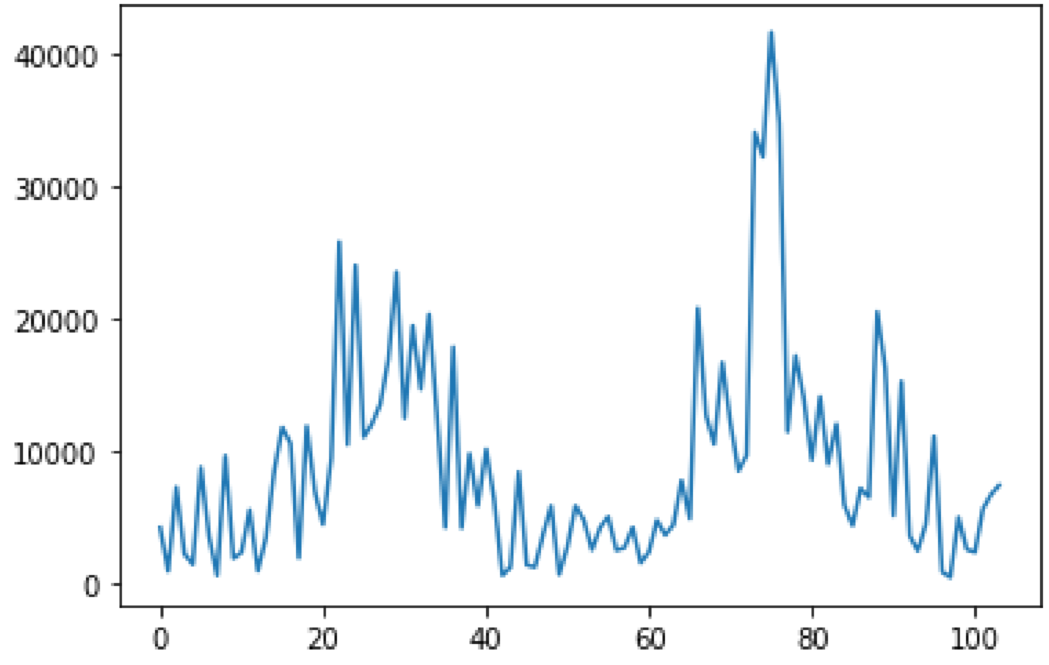

Figure 23. Raw signal from Ward cluster 13. This sample was taken over 1960 to 1962.

Acknowledgements. The author thanks the Central Michigan University Honors Program for the funding and resources required for this research. They also thank their advisor Dr. Jordan Watts for suggesting this topic and guiding them throughout the project.

References

[K] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[WA] https://scikit-learn.org/0.15/modules/generated/sklearn.cluster.Ward.html

[S] https://www.mathworks.com/help/curvefit/smoothing-splines.html

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

[W] https://en.wikipedia.org/wiki/Simplicial_complex

[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

[M2] Menne, M.J., I. Durre, B. Korzeniewski, S. McNeal, K. Thomas, X. Yin, S. Anthony, R. Ray, R.S. Vose, B.E.Gleason, and T.G. Houston, 2012: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3.26 https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/

[NP] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357-362 (2020). DOI: 0.1038/s41586-020-2649-2.

[CJ] Cooley, James W., and John W. Tukey, 1965, *An algorithm for the machine calculation of complex Fourier series*, Math. Comput. 19: 297-301.

[FO] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications, 2nd Edition*, Wiley Interscience, 1999.

[R] Stephen Roberts, "Signal Processing & Filter Design"; lecture notes, 2003 `http://www.robots.ox.ac.uk/~sjrob/Teaching/sp_course.html`.

[G] https://gcos.wmo.int/en/networks/atmospheric/gsn

[H] https://www.ncdc.noaa.gov/ushcn/introduction

[C] https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/us-climate-reference-network-uscrn

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

Department of Mathematics, Central Michigan University, Mount Pleasant, MI 48859

Email address: `mille7em@cmich.edu`