Data Analytics Final Project

Ryan Kaplan, Prof. Thilanka Munasinghe
RPI Computer Science

Abstract and Problem Area

Cluster analysis is a technique for analyzing data through examining the groups, or "clusters", that tend to form within large data sets. If these groups were given it would be a simple classification task, however this technique is unsupervised, with no such labeled data points.(1) Instead, we group data points together in terms of similarity. I was interested in continuing my work on the California environmental research paper, which involved applying cluster analysis to several relevant datasets, by comparing two of the most prominent clustering techniques; k-means and hierarchical clustering. The idea I want to investigate would thus be, how do the two methods differ in how they cluster this data? Specifically, how do they differ both in terms of a hard cluster evaluation metric, and how do they differ in the maps they can be used to make.

Datasets

All three datasets contain coordinate information as one of each data point's features, so once we've assigned a cluster label to each point, we can examine them overlaid on a map of California, and see how different regions are classified.

- **Precipitation** consists of the date, latitude and longitude, and the average monthly precipitation in that area.
- **Soil Moisture** has the date, latitude/longitude, soil moisture density in kg/m^2, soil moisture availability percentage, as well as average surface temperature.
- **Solar Radiation** has date, latitude/longitude, and surface absorbed longwave radiation (LWGAB)

Methods

- **K-means:** Implemented using SSE. SSE, or Sum Squared Error, looks to minimize the total squared distance, or error, between each node and its cluster's centroid, and we want the number of clusters which produces the minimum SSE. Using k-means to actually generate *n* clusters, we initialize *n* randomly placed centroids, and assign every data point to the closest one, and determine our SSE. (2) Next, we set each centroid to the average value of the data points assigned to it, and repeat the process until our centroids stop changing, and take the *n* value with the lowest SSE.

- **Hierarchical Clustering:** Also an iterative process, however unlike k-means, while it is going on, the number of clusters is variable.(3) Initially, every single data point is its own cluster, and at each step, the two most closest clusters are merged. As with k-means, we determine the closest cluster using SSE between the elements of our cluster, and every other. Of course, if we keep going in this manner we eventually end up with a single cluster, however when the distance between clusters is large enough, we stop merging. We determine this distance the same way as before, and so end up with the same number of clusters for each dataset.

References:
1. Frades, I. & Matthiesen, R. Overview on techniques in cluster analysis. Bioinforma. methods clinical research 81–107 (2010)
2. Shahapure, K. R. & Nicholas, C. Cluster quality analysis using silhouette score. In
2020 IEEE 7th international conference on data science and advanced analytics (DSAA), 747–748 (IEEE, 2020).

3. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2 , 86–97 (2012).

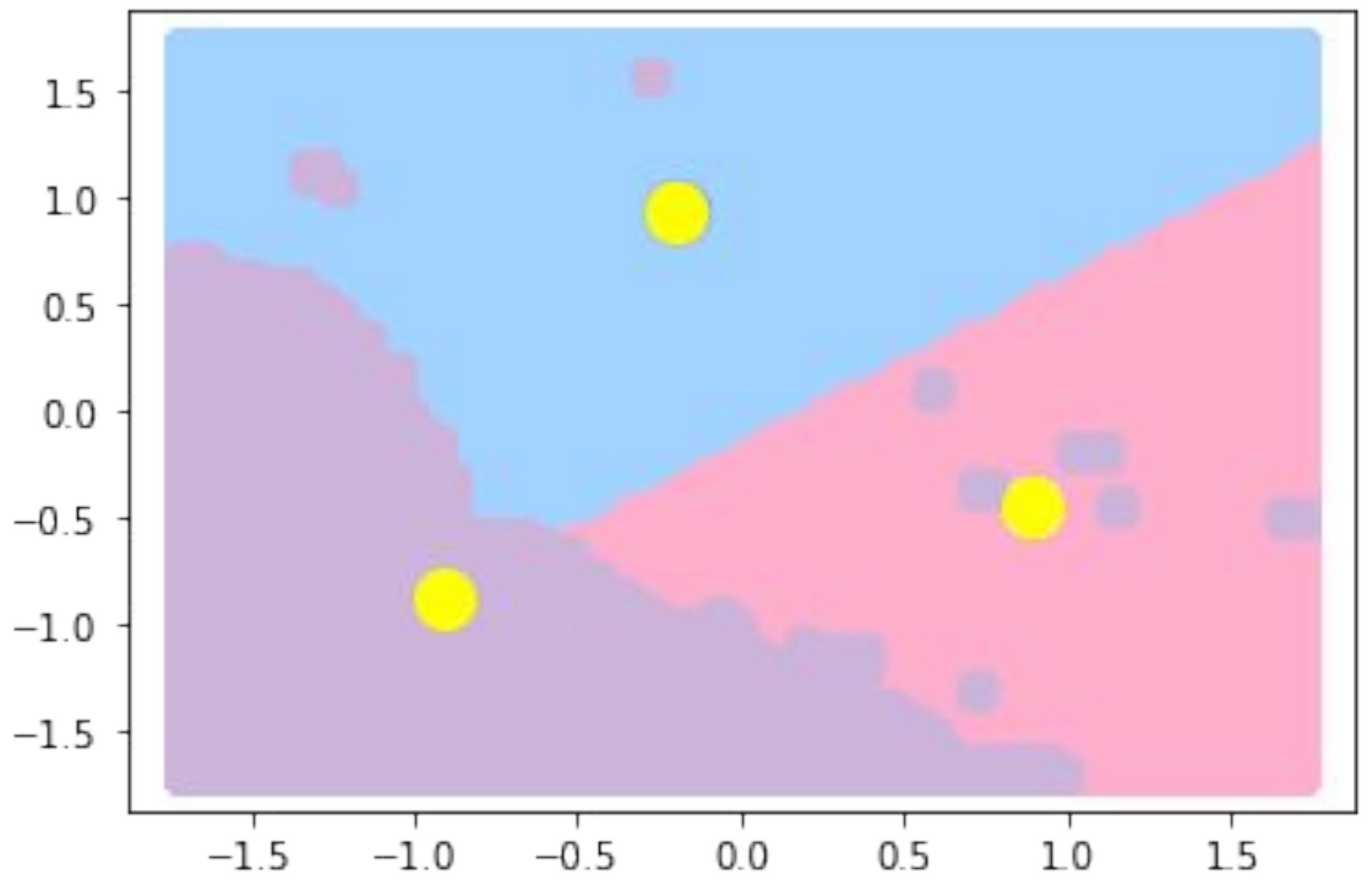

Figure 1: Example of k-means clustering on soil moisture data. The yellow dots represent our 3 centroids.

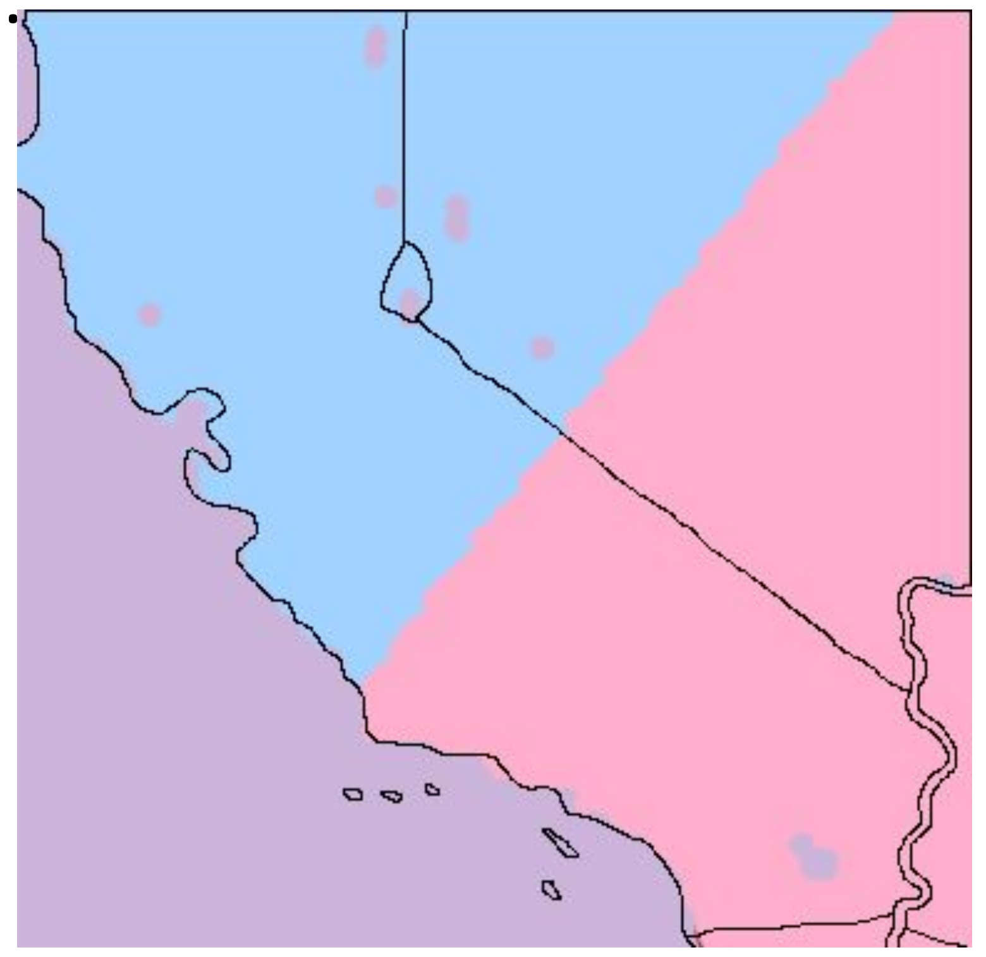

Figure 2: California map with soil moisture clusters overlaid. This is possible since datapoints include coordinates.

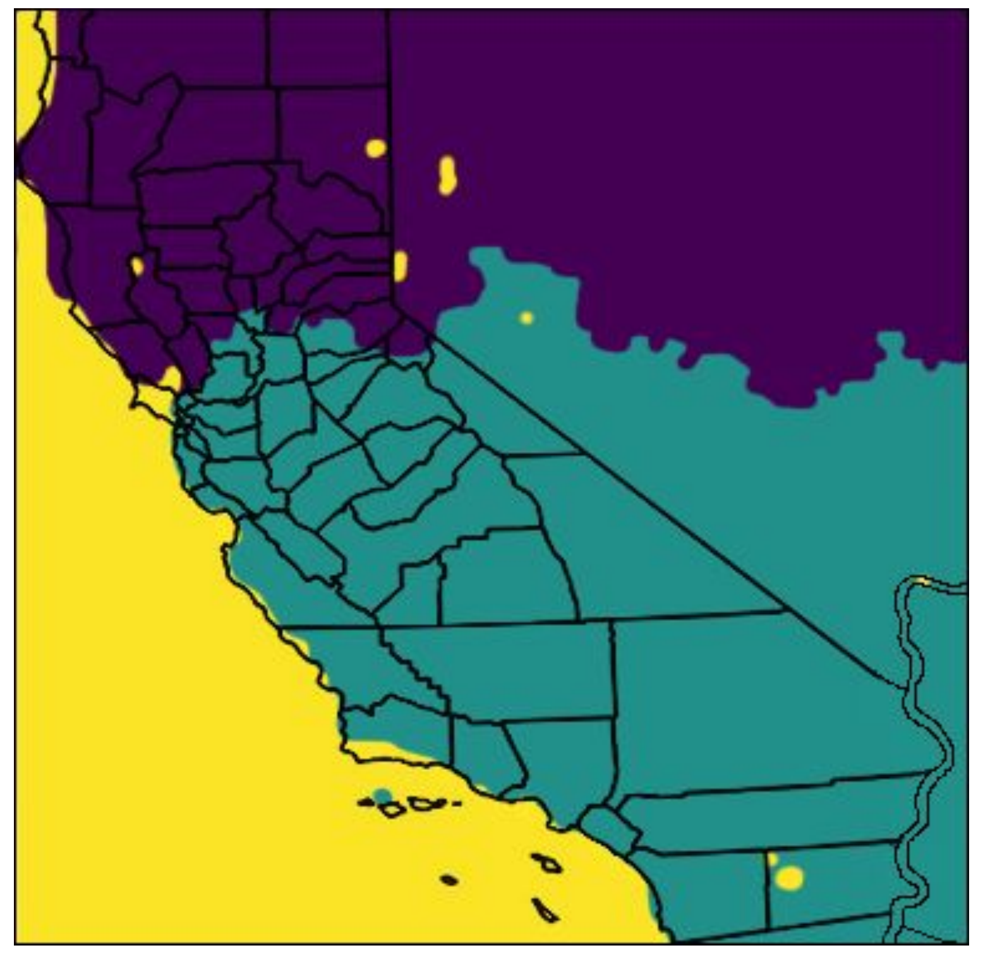

Figure 3: California map with soil moisture clusters, instead created using hierarchical clustering. Note the similarity, though the boundary is much rougher.

Model Performance

Dataset	Type	Distance metric	Score
Precipitation	k-means	Euclidean	0.5897…
Precipitation	k-means	Manhattan	0.5897…
Precipitation	hierarchical	Euclidean	0.3199…
Precipitation	hierarchical	Manhattan	0.3443…
Soil Moisture	k-means	Euclidean	0.1055…
Soil Moisture	k-means	Manhattan	0.1055…
Soil Moisture	hierarchical	Euclidean	0.2954…
Soil Moisture	hierarchical	Manhattan	0.2994…
Solar Radiation	k-means	Euclidean	0.1066…
Solar Radiation	k-means	Manhattan	0.1066…
Solar Radiation	hierarchical	Euclidean	0.2145…
Solar Radiation	hierarchical	Manhattan	0.2243…

We can score our model's performance separating each dataset using Silhouette score.

Conclusion

It seems that overall, hierarchical clustering did a better job on our environmental datasets than k-means. I must admit, I did not expect to find a clear winner, thinking that given the same dataset, particularly ones without anything crazy going on like we find here, we would get very similar results between the two methods. Even without the mathematical interpretation, I find that the rougher, more organic maps it generates are more true to reality than the clean, crisp border lines often formed by k-means. One thing that I am not sure of, is why the distance metric affected the two methods so differently. It seems logical that if it wouldn't affect one then it wouldn't affect the other, and vice-versa. I would be interested in looking deeper into the methods to justify this discovery, however that is outside the scope of this project.