Rain Prediction Using K-means Clusters

Kelly Tao
https://github.com/drwwoob/weather

Abstract

This report uses the k-mean clusters to analyze the data on the world-wide precipitation rate for the past 30 years. The goal is to find the similarities among areas, and group them together base on the learning, so that by giving that data of one place, we could predict on what will happen elsewhere.

1 Introduction

Everything on the Earth is connected. A heavy rain in South America could cause a heatwave in Europe, an eruption of an under sea volcano could bring storm to the other end of the world. With nowadays technology, we are still unable to predict precisely what will happen tomorrow. But even though some causes cannot be detected, the result surely can be. Thus, this model is build based on the data human are able to receive, and with the help of machine learning, look for the hidden connection among places.

1.1 Goal

With the limitation of data, it would be hard to create a model which can detect the exact change happening in one place, thus, the goal is mainly to separate areas out with the changes in their precipitation rate, and find possible important factors.

1.2 Hypothesis

With current studies supported, the image of how the model should look was firmly set. Before all, South Pole and North Pole should be itself as a whole, and not in the same group as any others, with their extreme conditions. The oceans should follow the same rules, with no barriers on the ocean that blocks the wind which blows the rain to move. On land, some special parts should be marked out as a different cluster, such as desert and the far North. Then, the map should be split into areas following the latitude, with their similar time shined by the sun.

2 Data Modify and Programming

The data provided are the monthly average precipitation rate in 2.0 degree latitude x 2.0 degree longitude global grid (360x180), from 1991.1 to 2020.1.

2.1 For Testing

For testing purposes, we split out the part from 1991.1 to 2014.12 as the training data, and 2015.1 - 2019.12 as the testing data. Therefore, the dimension of the training data became (360 * 180) * (12 * 24), and the dimension of the testing data would become (360 * 180) * (12 * 5), which could not fit together, so we've shrunk the into "the average precipitation in a month through out the 24 years", as (360 * 180) * 12, to insure the test from working.

2.2 Deviation

For studying the relationships, the similarity in the data itself is not enough; only when changes happens, can we see if another changes with it. Thus, we went for the change, or the deviations in the data.

2.3 coordinates

It is commonly know that the closer two places are, the more physically close they are, thus the more possibility that they could affect one another. Therefore, I added in coordinates to the last two digits to stronger their relationship.

2.4 Polynomial

Most of the deviation plot looks similar to a binomial, so we went for the direction of fitting a curve onto it. To do so, we used np.polyfit(), and tried to fit the data into both 2 degrees and 3 degrees.

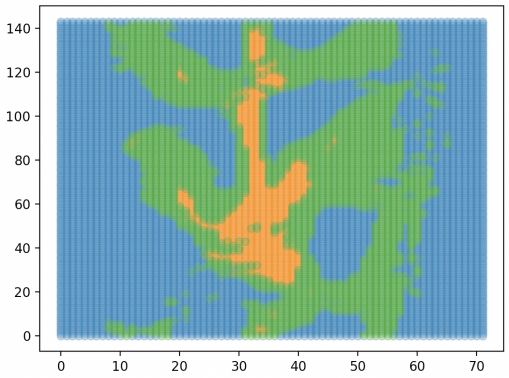
The factors of this method would be its coefficients. We did not add in weight, since it is unclear which factor is more important than others.

3 Result

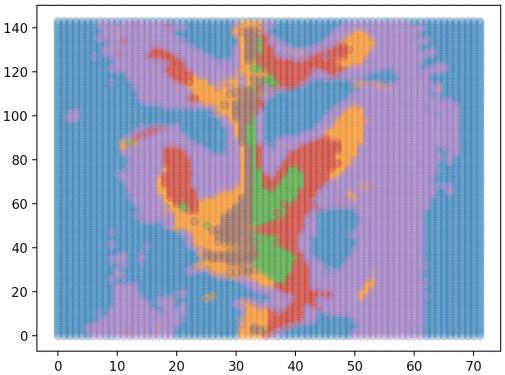
After getting the data ready, all that's needed to do is to fit it into the k-means clusters model provided by sklearn. With the trained model, we let it predict on the testing data, and drew graphs according to the coordinates of the data, colored in its own label.

3.1 Raw data

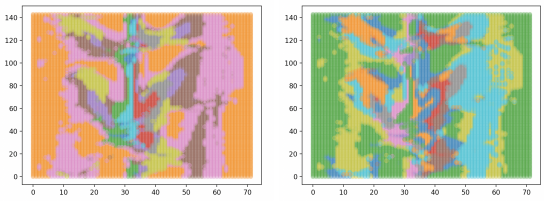
The result with the factors as each month's average showed a high potential as it separated land and ocean almost immediately at the number of clusters as only 3.



With the increment on the number of clusters, more details appeared on the result image. The clusters follow the latitude but not fully as expected, extending outward from equator.

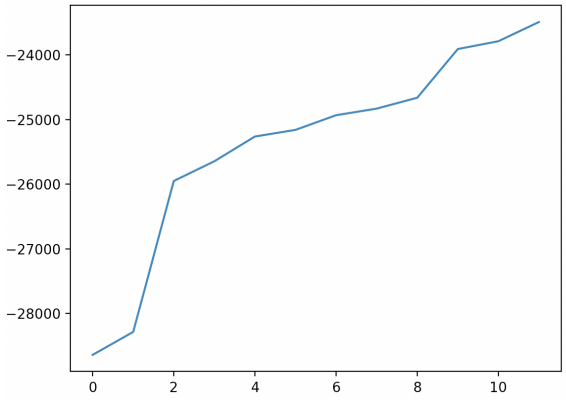


But as the number of clusters increasing, the model seemed to be over-fitting. In the graph on the left with the number of clusters 10, the clusters grouped the far North with part of the Atlantic ocean and part of Australia, which does not seems right. In the graph on the right with the number of clusters 14, where the area around the equator got grouped into the same group as the North Pole, possibly due to their lack of rain.



This model could give us an idea of what the graph should look like, but it depend too much on the basic amount of rain.

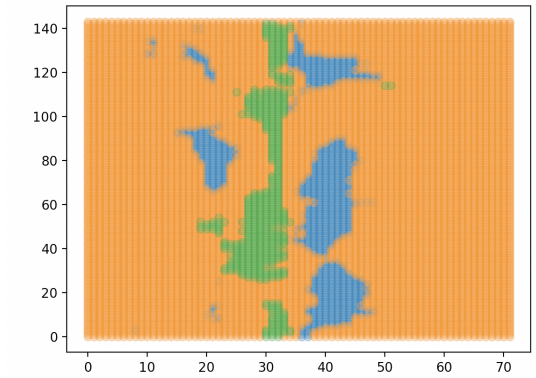
The score k-means got from the number of clusters 3 - 15 are very negative. One possible explanation of it is the wide range it got through out the map.



3.2 Deviation

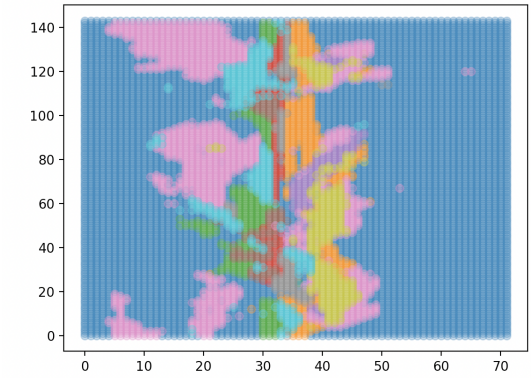
This model, beside from all models created, represent the closest map to our goal, though its score is not as cheerful.

The first insight of it is not as impressive as the raw data, for in the image of number of clusters = 3, instead of appearing in the shape of land, it only marked out partial equator and lands that are not connected.

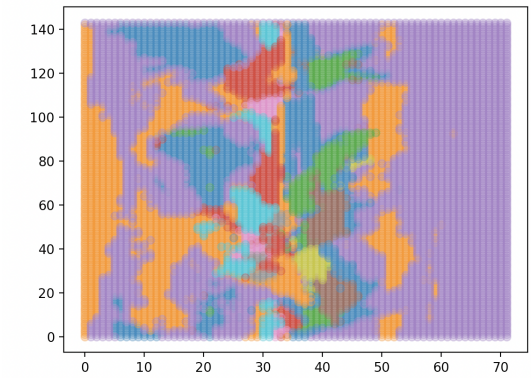


But as the number of clusters increased, the map seems much better than with the raw data. In the image of number of clusters = 10, though some are still not continuous, the Poles and the oceans are still not included into any other group.

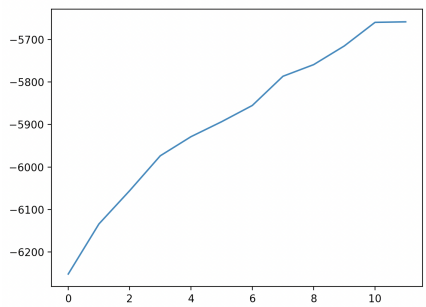
The area around the equator was also marked out individually, fit very much to common sense.



As the number of clusters reached 15, some shape seem lost around India and its surrounding, but more details are shown. One interesting thing on this map is that it grouped North Pole and South Pole together, though they are not anywhere close to each other, nor should their change in precipitation be close. Another phenomenon we observed is that there is an area in Africa that is in the same group as North Asia, where after comparing to map, is the Sahara and Siberia. Indeed, the two places are both somewhat close to a desert, but them being marking out as an individual group still gave me the understanding that this model could not separate as much "acute changes".

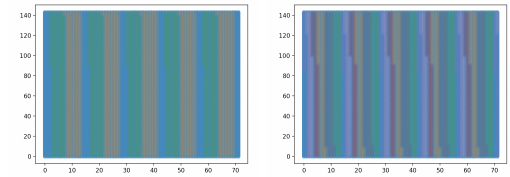


The score of this model is also better than the raw data one. Being still not as good, it decreased to one fourth of the previous one, though we suspect it is only due to the decrease in the scale of number.



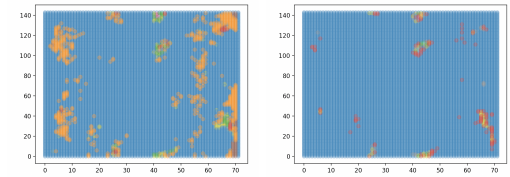
3.3 Coordinates

The test with coordinates did not go as good. We suspect the factors of coordinates took too much weight, since they are the most measurable factor within all other factors in the data.



3.4 Polynomial

The model for polynomial is also an failed one. The images for 15 clusters in either a two degree or three does not present any noticeable difference. We think it is classed by our lack of griping the important factors in the polynomial, but as we have explained previously, we are not sure about it. The fitting model loss too much detail on the data, and what the outcome represent was unclear.



4 Conclusion and Reflection

Overall, the model created something similar to what we've expected, but not precise enough for a prediction. For further improvements, adding data on others areas such as surface temperature and moisture's could possibly adding in more accuracy to the model, but we failed to do so with the lack of data within our data set, which contains too much NaNs for the model to gain the right data.

References

Xie, P., and P.A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bull. Amer. Meteor. Soc., 78, 2539 - 2558.

Morice, C.P., Kennedy, J.J., Rayner, N.A. and Jones, P.D., 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset. Journal of Geophysical Research, 117, D08101, doi:10.1029/2011JD017187

GISTEMP Team, 2018:GISS Surface Temperature Analysis (GISTEMP). NASA Goddard In-

stitute for Space Studies. Dataset accessed 20YY-MM-DD from NOAA/PSL'ss website