Joaquín Herrera Ramos
A01207504
October 20th, 2020

# Finding climate analogs for a given location

## Introduction

The present project was developed during an internship at ClimateAI. The company is a Silicon Valley based startup that develops artificial intelligence solutions for more resilient agriculture. They currently sell their solutions to food and seed production companies.

This project is an extension of the work of one of the data scientists from the team, Christopher Cross, who worked on developing a more complete long-term climate report for our clients.

The piece of the report that will be extended in this project involves locating places in the world with a similar temperature and precipitation to a given location. The present project will be using K-means clustering algorithm to try to find the most similar locations to the given one.

## Relevance

This project helps to locate other places in the world with similar weather conditions (temperature and precipitation) to a given location. This might be useful for studying climate patterns around the world.

As climate change intensifies, for growers, it is imperative to have information for developing contingency plans. The objective of this project is to give ClimateAI's clients an understanding of what other places in the world could be a potential location for planting similar crops as they already have.

## Background

Christopher's work developed a paradigm for identifying climate analogs. These analogs are based on the Standard Euclidean distances in the yearly historical data between the location and the rest of the world [1].

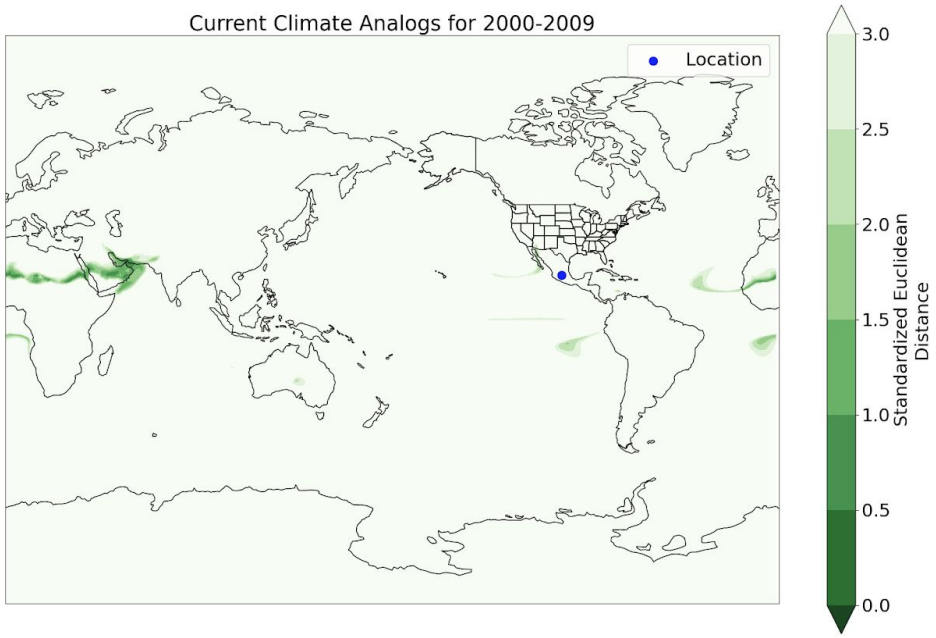
## Dataset

For the analogs, data from 2000 to 2009 downloaded from [2] is taken to compute the distances.

Joaquín Herrera Ramos
A01207504
October 20th, 2020

This project leverages this dataset that contains 55,295 locations (latitude and longitude) and uses Cross's work to calculate their corresponding standard euclidean distance. The smaller the standard euclidean distance (SED), the more similar the location is to the location that is being compared.
Given the nature of climate, there are some areas in the world that have similar climates, so the dataset ends up with many locations that are close together with a similar SED.

For example, given a location in Mexico (the blue dot), Christopher's algorithm computes the SED for all the 55,295 locations in the world. Then, using plot.contourf on the datarray, the SED is scaled to a 0 to 3 range. The locations with an SED close to 0 are the ones with the most similar climate. This locations are shown on a green scale:




**Problem Statement**

Since these green zones are groups of locations with similar SEDs, these can be treated as clusters. The objective is to find the centroids of these given clusters to provide a specific location that represents that climate cluster.

One thing to take into consideration is that the ocean should not be taken into consideration since we are looking for similar weather spots for planting and agriculture. Given this restriction there should be a preprocessing step to mask out any location that is on the ocean.

**Proposed solutions**

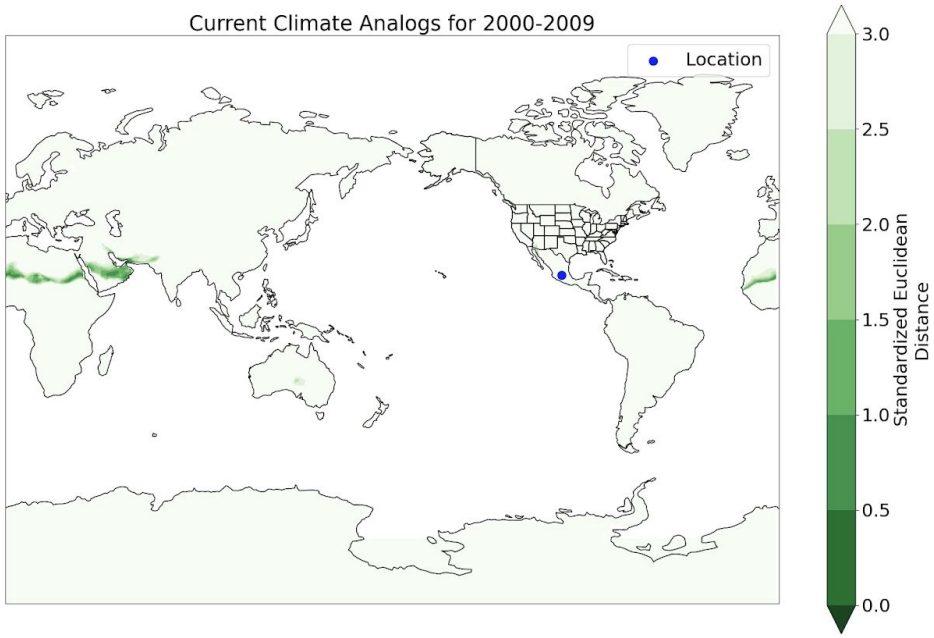
There are two proposed solutions that will be tested in this project.

1. A standard k-means solution that requires the binarization of the SED. If the SED is below a given threshold it will be considered 1, if above, it will be 0.
2. A weighted k-means implementation that uses a scaled and adapted version of the SED. The normal weighted k-means [4] algorithm takes into account a weight, the bigger it is, the more important it is for the cluster. Since the most interesting locations have lower SEDs, there needs to be a preprocessing that inverts these values.

**Data preprocessing**

First of all, a .nc file from NOAA PSL Climate Data Repository [3] is used to mask out the ocean. This leaves 18,478 data points from the 55,295 that were at the beginning. The new green zones look like this:

Joaquín Herrera Ramos
A01207504
October 20th, 2020

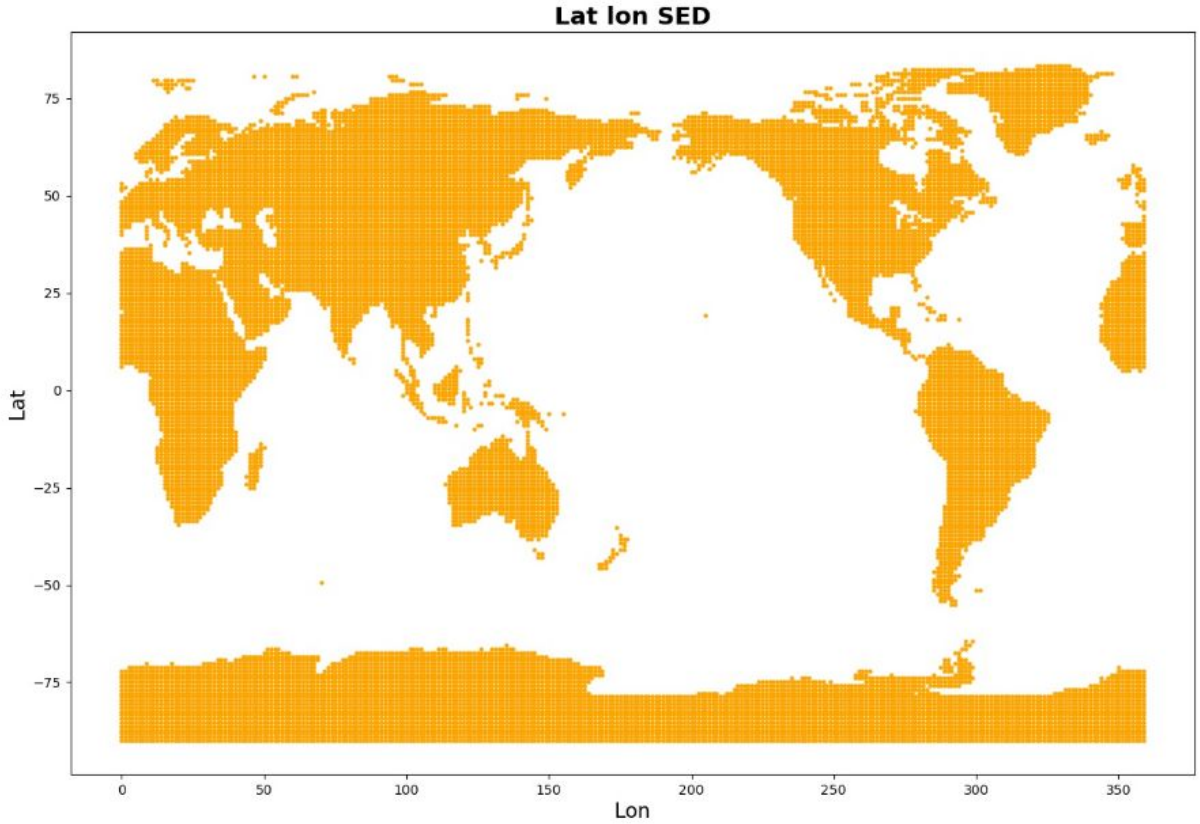



Before running any of the proposed solutions, a MinMaxScaler preprocessing algorithm was run on the Standard Euclidean Distance column.

Another preprocessing step was to drop 90% of the data. Only the top 10% of the points were kept (the points with the smallest SEDs). This means that from 18,478 data points, k-means will cluster only 1848 data points.

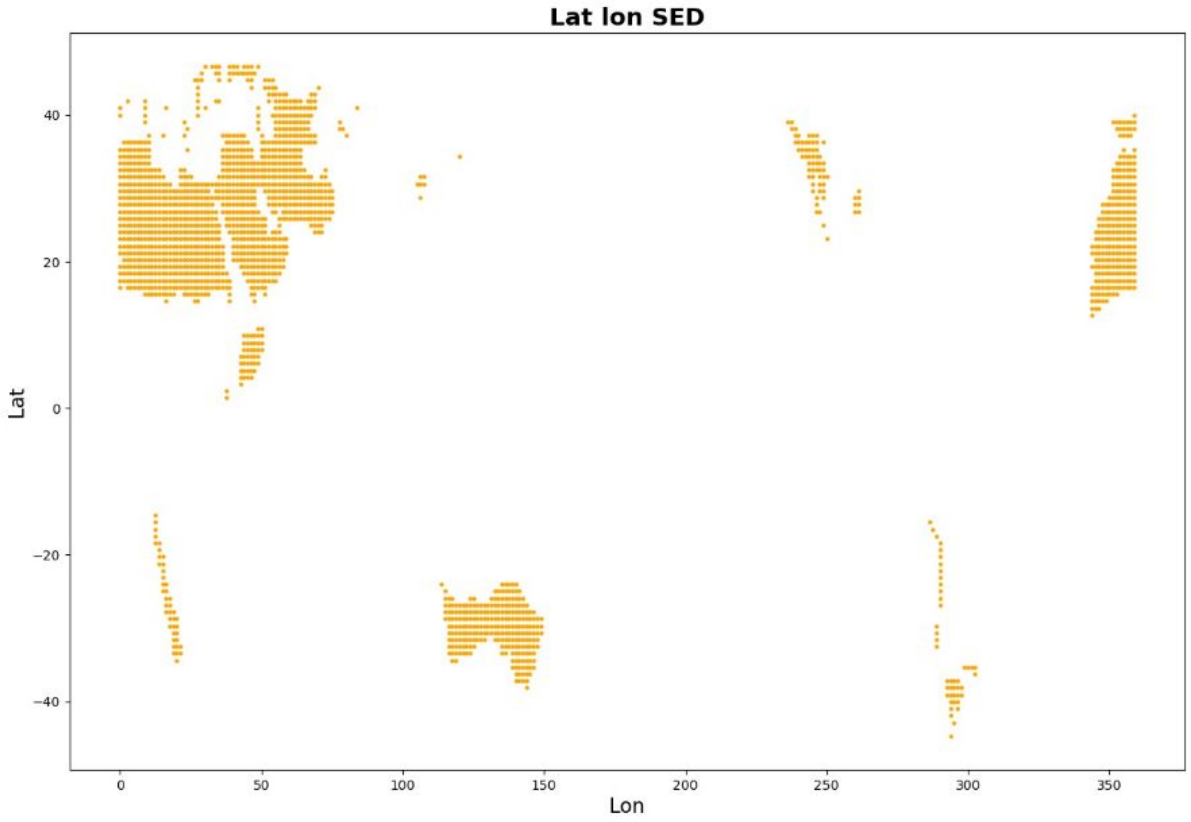
This is the map before dropping the 90% of the data, but after masking out the ocean:

Joaquín Herrera Ramos
A01207504
October 20th, 2020




This is how it looks with just 10% of the data:

Joaquín Herrera Ramos
A01207504
October 20th, 2020



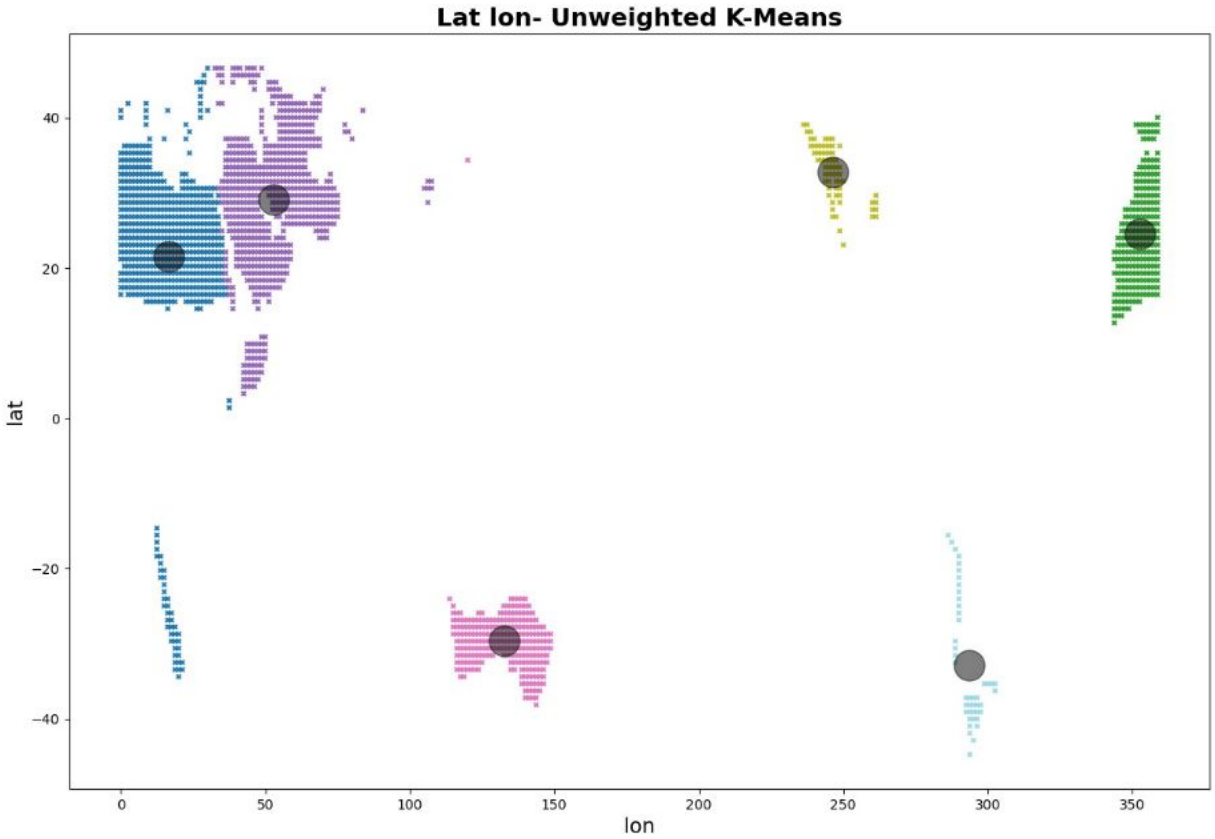

Clusters are starting to show. From this image, it was decided to use 6 centroids.

**Running solution 1: unweighted k-means**

After the data preprocessing, standard, unweighted, sklearn k-means was run on the dataset as follows:

```
kmeans = KMeans(n_clusters = 6, random_state = 0, max_iter=1000)
```

Joaquín Herrera Ramos
A01207504
October 20th, 2020

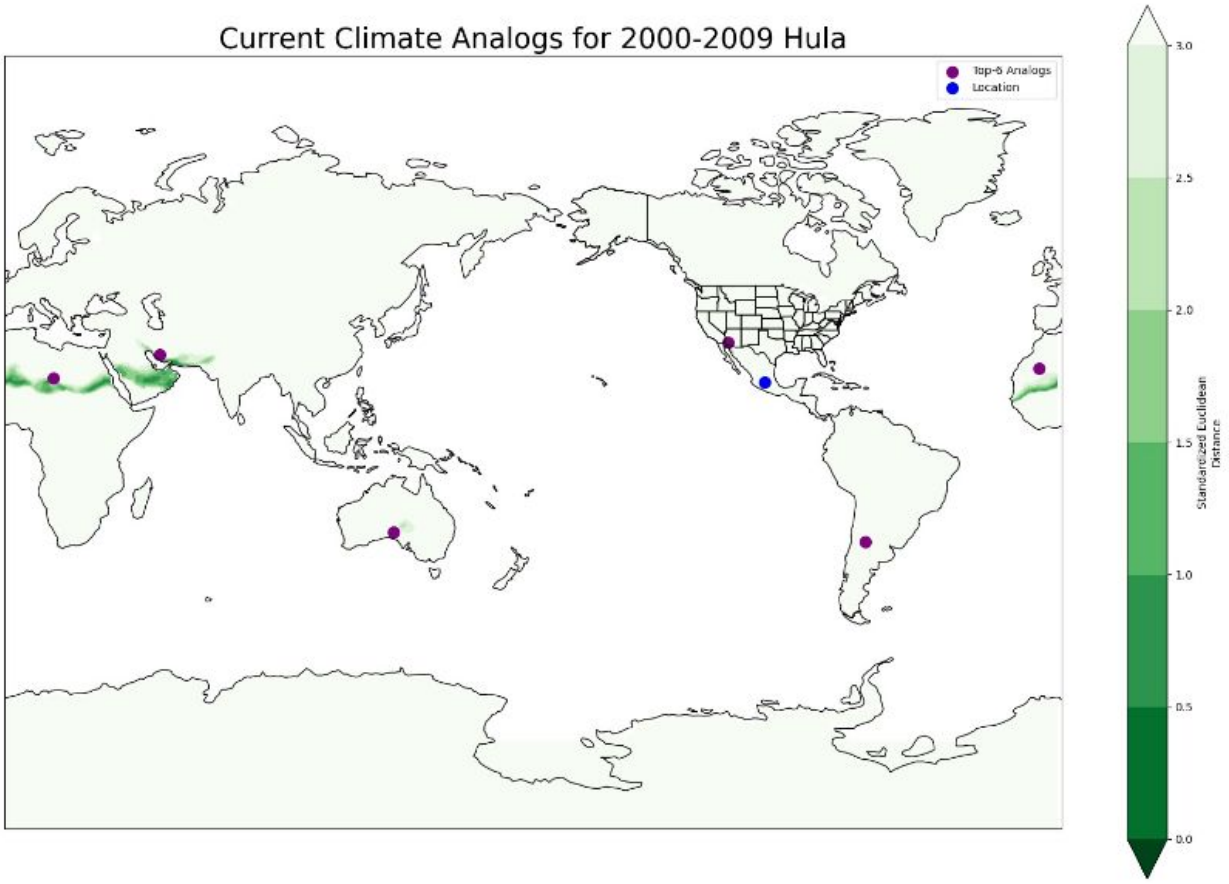



These are the selected centroids:

[[ 21.57376045  16.61968954]
 [ 24.6253685  352.4947479 ]
 [ 29.08695022  53.00233645]
 [-29.67826385 132.64616935]
 [ 32.83704188 246.46484375]
 [-32.83436446 293.35227273]]

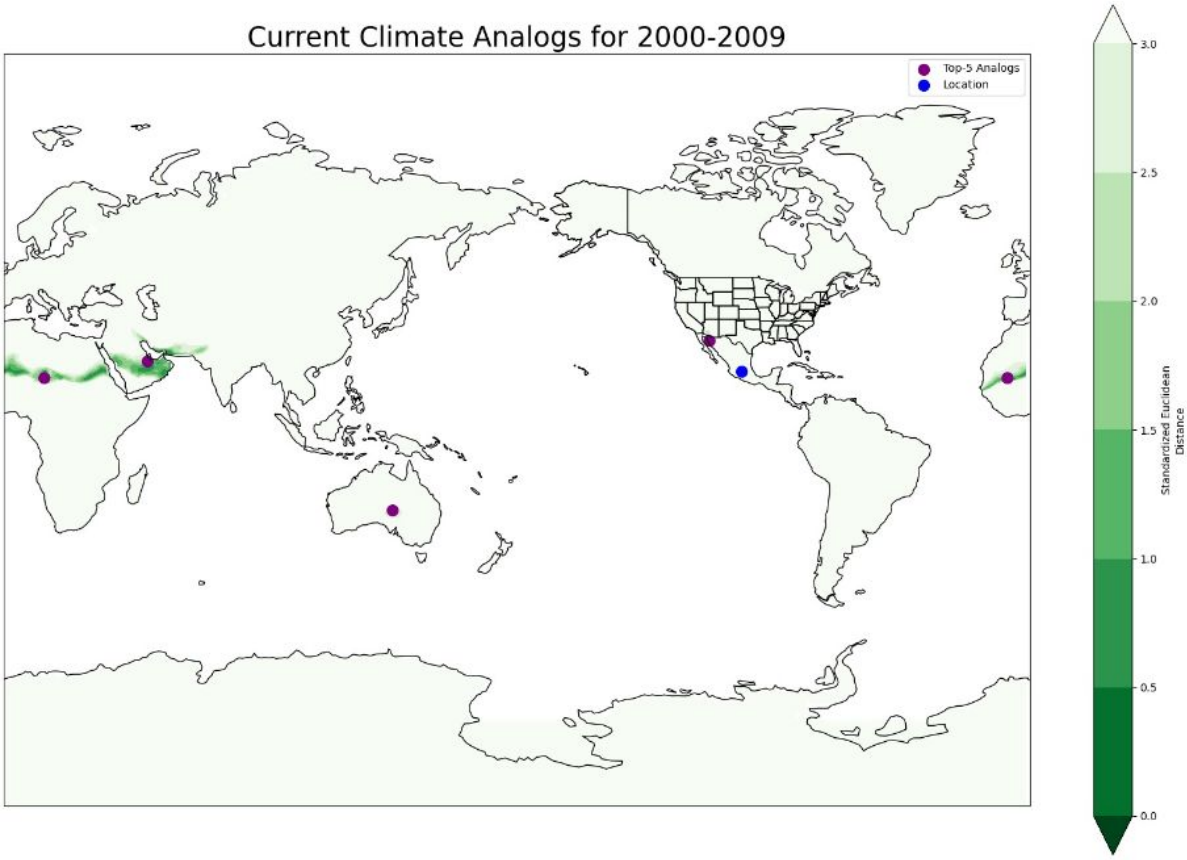
This is how the map looks like with the green contour:

Joaquín Herrera Ramos
A01207504
October 20th, 2020




From this map it can be seen that some of the purple dots are not in the greener areas. This may be improved by reducing even more the data. Instead of leaving 10% of the data, only 2% will be kept. From this, it was decided to use just 5 centroids instead of 6.

Joaquín Herrera Ramos
A01207504
October 20th, 2020

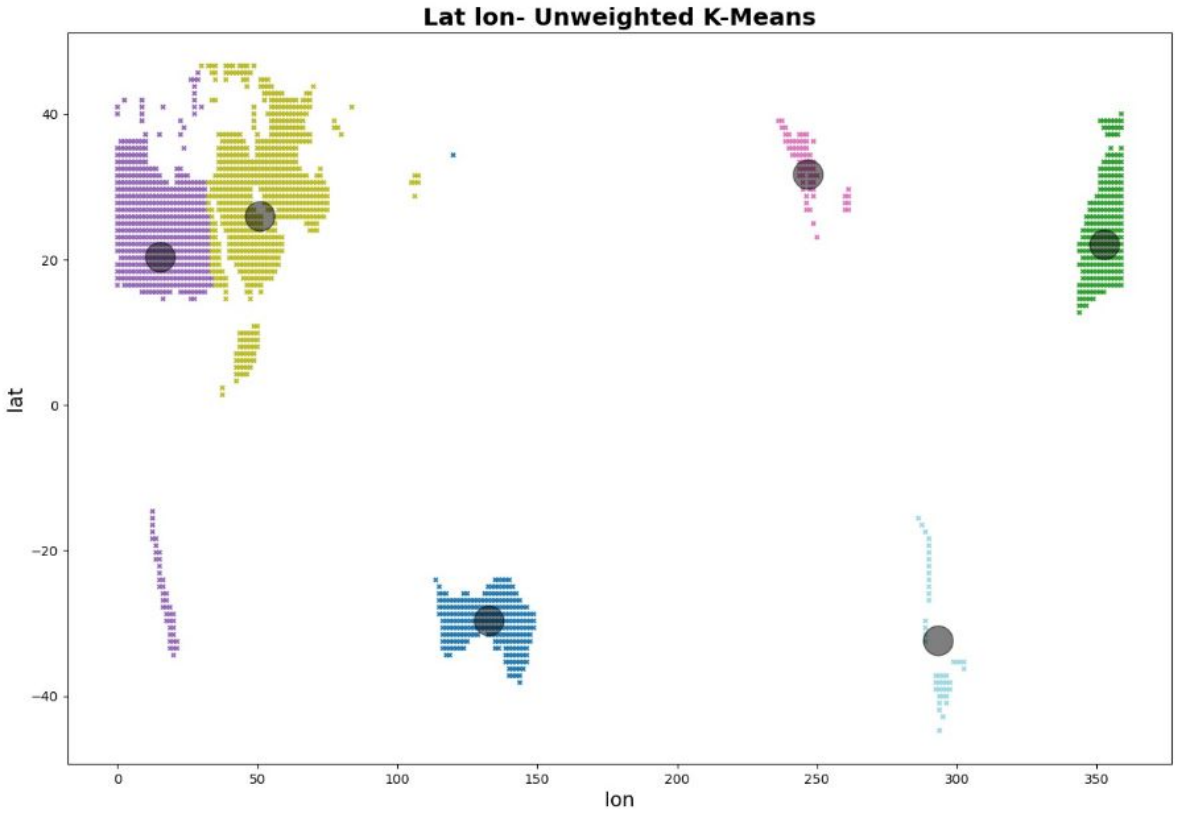



This shows that with some manual analysis, the centroids can be improved. The new centroids are a better representation of the climate analogs.

**Running solution 2: weighted k-means**

After the data preprocessing, a weighted implementation of k-means was run following Dey's medium tutorial [4].

Joaquín Herrera Ramos
A01207504
October 20th, 2020

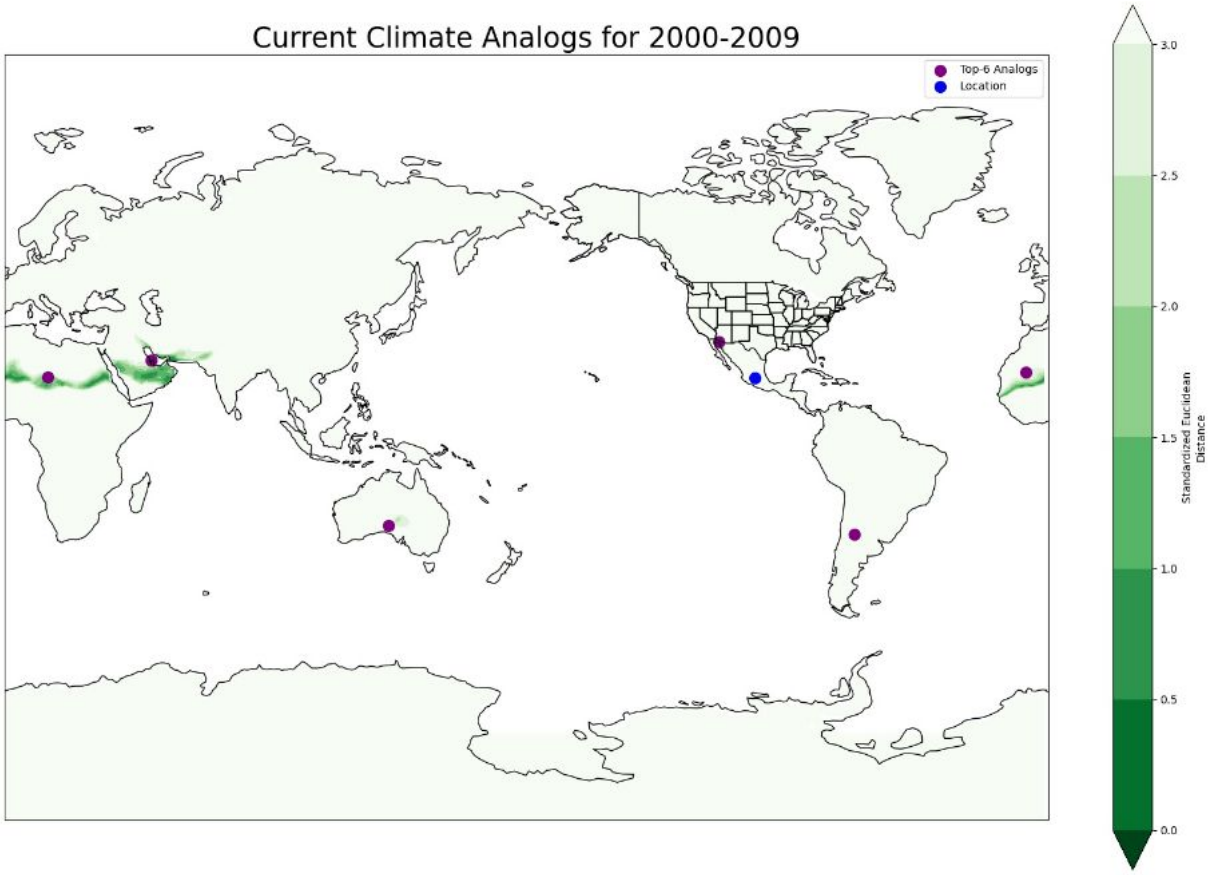



These are the selected centroids:

[[-29.59855374 132.6318726 ]
[ 22.10973649 352.46871127]
[ 20.42028639  15.13342405]
[ 31.77414684 246.54544323]
[ 25.95990227  50.85934949]
[-32.32524773 293.11178657]]

This is how the map looks like with the green contour:

Joaquín Herrera Ramos
A01207504
October 20th, 2020




Weighted k-means was more precise with 10% of the data compared to unweighted k-means on the same amount of data.

**Results interpretation**

Both k-means implementations gave good results but the weighted implementation turned out to give better results when there are more data points, this can be explained by the literal implementation of weighted k-means. If the clusters are large and no weight is provided, the geometric center will be the centroid, but if weight is taken into account, some differences might arise for the centroid selection, this gives the algorithm more robustness.

Since it desired to reduce manual analysis, weighted k-means is a better solution for this case.

**Testing and deployment**

After comparing these two clustering algorithms and selecting the weighted implementation, this logic was tested on other locations and then it was implemented in a Google Cloud Run instance. This allows Climate AI any desired amount of centroids (climate analogs) for any given locations.

**Further research**

This project could be further extended by testing other clustering algorithms like DBSCAN or a hierarchical approach. Other weather variables can be included too. In this project only precipitation and temperature were taken into account to compute the Standard Euclidean Distance, but adding more variables could result in centroids that have more similar weather conditions to the given location.

**Conclusion**

All the work and testing conducted for this project can be located in the github repository: [5].

This was an interesting project that can bring value to Climate AI's customers so they can make better decisions in the future and with the current weather changes.

Places in the world with similar weather conditions to a specific location can always come in handy, not necessarily for agriculture but also for tourism, migration studies and other sociological analysis.

**References**

1. Williams, John & Jackson, Stephen & Kutzbach, John. (2007). Williams JW, Jackson ST, Kutzbach JE. Projected distributions of novel and disappearing climates by 2100AD. Proc Natl Acad Sci USA 104: 5738-5742. Proceedings of the National Academy of
2. https://ncar-cesm-lens.s3-us-west-2.amazonaws.com/catalogs/aws-cesm1-le.json
3. NOAA, PSL Climate Data Repository. lsmask.oisst.v2.nc https://psl.noaa.gov/repository/entry/show?entryid=synth%3Ae570c8f9-ec09-4e8

9-93b4-babd5651e7a9%3AL25vYWEub2lzc3QudjIuaGlnaHJlcy9sc21hc2sub2lzc3QudjIubmM%3D

4. M. Dey, "Unsupervised Learning with Weighted K-Means", Mallika Dey Medium, 2019. [Online]. Available: https://medium.com/@dey.mallika/unsupervised-learning-with-weighted-k-means-3828b708d75d [Accessed: 27- Oct- 2020].
5. clustering analogs: https://github.com/ClimateAI/clustering_analogs