# Model Exploration- Sprint 1- Research Findings

In this document, we aim to cover the following topics in their corresponding sections:

1. What Geospatial Clustering Algorithm we will use & why
2. Any findings for APIs
3. Flowchart for possible model suggestion
4. Any relevant links

## 1. *Geospatial Clustering Algorithm*

Most relevant clustering: **Density Based Clustering.**
Read below for description grabbed from TowardsDataScience Article[1]

Density-based clustering works by grouping *regions of high density* and separating them from *regions of low density*. The most well known density-based clustering algorithm is the DBSCAN algorithm (Density-based spatial clustering with the application of noise).

The density is calculated by using two parameters which are as follows

1.EPS: This defines the neighborhood around the data point i.e if the distance between two points is less than or equal to eps then they are said to be neighbors

2. MinPts: THis defines the minimum number of data points that form a neighborhood. The size of the dataset and the value of MinPts are directly proportional.

DBSCAN algorithm visits every point and if it contains MinPts within eps then cluster formation starts. Any other point is defined as noise. This process continues till a density connected cluster is formed and then it restarts with a new point.

*Note:* K means method can also be used

## 2. *API Inspirations*

The API we are building will focus on the Population Density based on given coordinates. This will be done by a GET request that targets a specific URL, and from the URL gathers information on coordinates along with population density and stores it in a dictionary to be processed and used further as data.

[1] https://towardsdatascience.com/geospatial-clustering-kinds-and-uses-9aef7601f386

The information will be retrieved from the OpenDataSoft; a public database which gives access to each city of the world including their Population, Country, Coordinates and Region. The dataset can be found here.

## 3. *Flowchart for possible model suggestion*

-Identification of what we are trying to predict
-Selection of important factors - Feature engineering :
In order to build the model it is necessary to determine the important features that will be used for the prediction of the water shortage. Apart from the data that we already have, additional data can be used for the prediction model. There are many aspects of water resources vulnerability arising from various physical, social, economic, and environmental factors. Therefore, water risk probability requires the consideration of vulnerability, which means that additional factors should be taken into account. For example, climate change (e.g. drought) or integrated urbanization.
This can also be done by performing a statistical analysis of the data (graphs, tables,etc..)

***Supervised Learning***

**Logistic Regression**

We could consider the water shortage as a binary categorical variable (can occur or not). A logistic regression model can be used to describe the relation between the water shortage risk and its impact factors. The maximum likelihood estimation is used for parameter estimation.

**Random Forest**

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set. It runs efficiently in large data sets providing low bias.

***Unsupervised Learning***

**Spectral Clustering Based on K-nearest neighbours**

Assigning of data points to groups based upon on how similar the point data are. As spatial features, Latitude and Longitude make natural candidates for this algorithm.

(These are the suggested algorithms for the prediction model, but it is important to consider first which data will be used, how we will combine them and what is the *'key'* question.)

Flowchart (Needs further processing probably after determining the first two steps)

..