BTP Evaluation 3 Report

Linear Regression on Nearest Neighbours

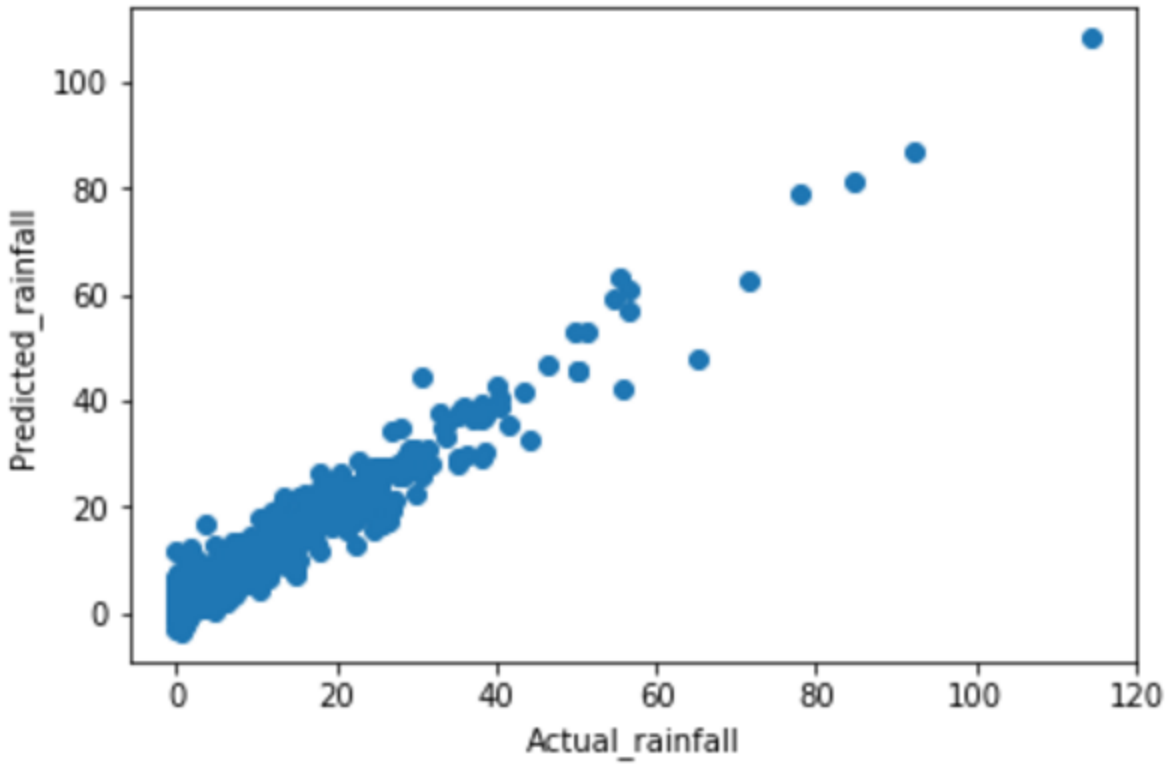
For each site we located neighbours within some fixed distance .Linear Regression was fit using rainfall data of the neighbours after some preprocessing (Removing the data points having 0 rainfall).

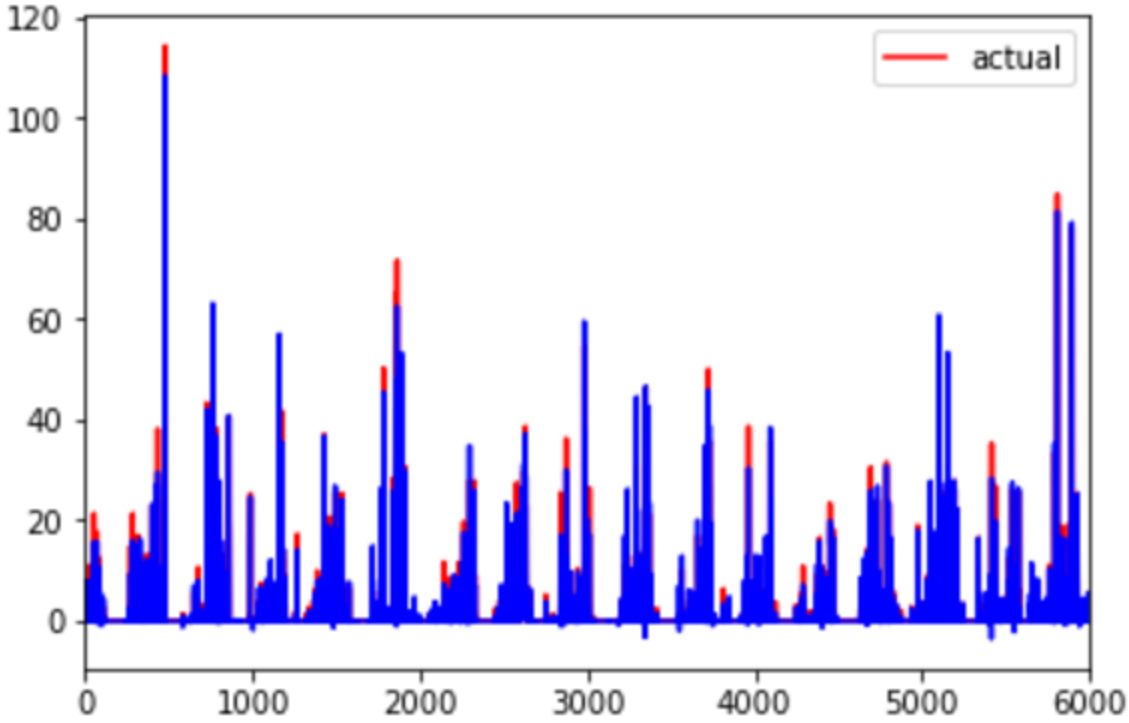
$$Y = a_1x_1 + a_2x_2 + ... + a_kx_k$$

$$x_1, x_2, \ldots, x_k$$

are spatially neighbouring sites .

Following were the results obtained -


Predicted_rainfall
Actual_rainfall
0
20
40
60
80
100
120


actual
0
20
40
60
80
100
120
0
1000
2000
3000
4000
5000
6000

R2 score was found to be [0.90-.95] for different sites .

We then wanted to see whether sites which are not neighbours also show similar patterns . For this we explored Spectral Graph theory and clustering .

Spectral Clustering

Assuming the rainfall sites as nodes of a simple undirected weighted graph,G(V,E) , where an edge is considered between two nodes i,j if correlation(i,j) >0.5 and weight of the edge is the correlation value . We also consider that there is no self loop in the graph To put it more formally ,

If A is the adjacency matrix of the graph ,

A[i][j] = correlation(i,j) if correlation(i,j) > 0.5
0 if correlation(i,j) < 0.5
0 if i=j

Degree matrix is a diagonal matrix containing the degree of every vertex, where degree of vertex 'v' is defined as sum of weights of edges incident on v . More formally,

D[i][j] = 0 if i != j
sum(values in A[i]) if i=j

Laplacian is calculated as follows -

L = D - A

The Laplacian matrix is a discrete analog of the Laplacian operator in multivariable calculus and serves a similar purpose by measuring to what extent a graph differs at one vertex from its values at nearby vertices.We could imagine an edge between 2 nodes being an arrow pointing each way and the edge is the sum of those two edges, which make a closed loop. Then the second derivative is a loop that goes out-in in both directions. So the "+2" in the middle is the two "out" arrows and the "-1"s are the "in" arrows. (By the way, all derivatives are like this: an x-y derivative is 4 arrows going

around a square.) So, the degree of a node is a count of out arrows; the adjacency elements are a count of in arrows; and the Laplacian adds them together to get the total count of second-order arrows which might satisfy our calculation of D-A .

We then Proceed to the eigen decomposition of laplacian -

$$L\,v = \lambda v$$

Before we move forward , let's look at some properties of Laplacian and its eigenvectors , eigenvalues -

- *L* is symmetric.
- L is positive semi-definite , i.e all λ's >= 0

  This is true , because ,

  $$\lambda_i = v_i^T L v_i$$

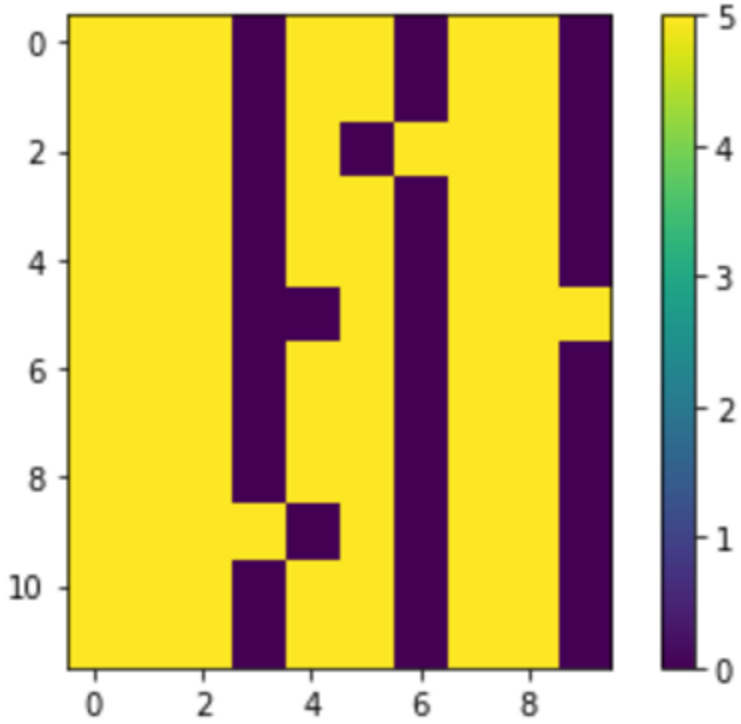
  L can be written as,

  MM^T where M is the incidence matrix therefore,

  $\lambda_i = (Mv_i^T)Mv_i$, inner product of a vector with itself , hence positive

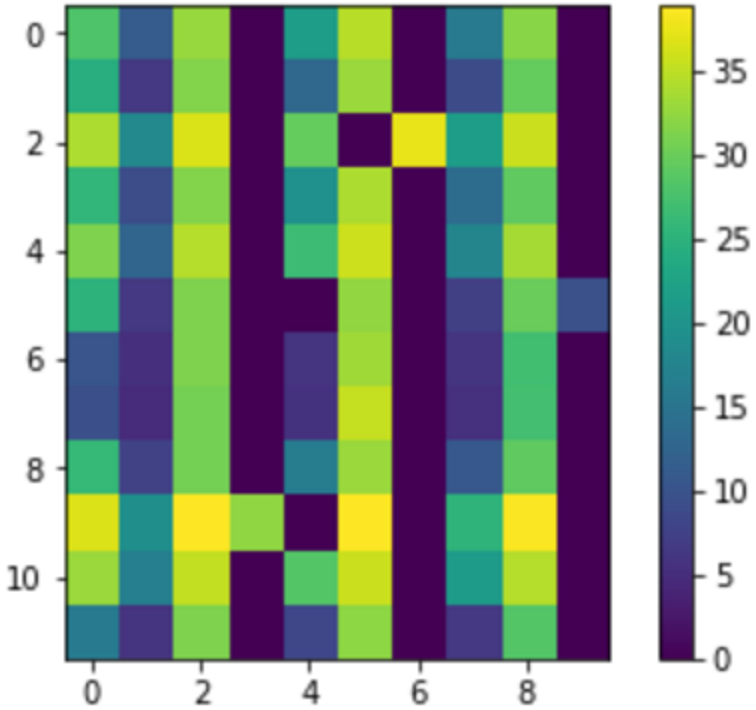
- First eigenvalue is 0 because the first eigenvector [1 1 1 1.. 1] satisfies Lv = 0
- The second smallest eigenvalue of *L* is the algebraic connectivity of *G* and approximates the sparsest cut of a graph.

When we plotted first eigenvector as a heatmap of weather sites represented in a grid-



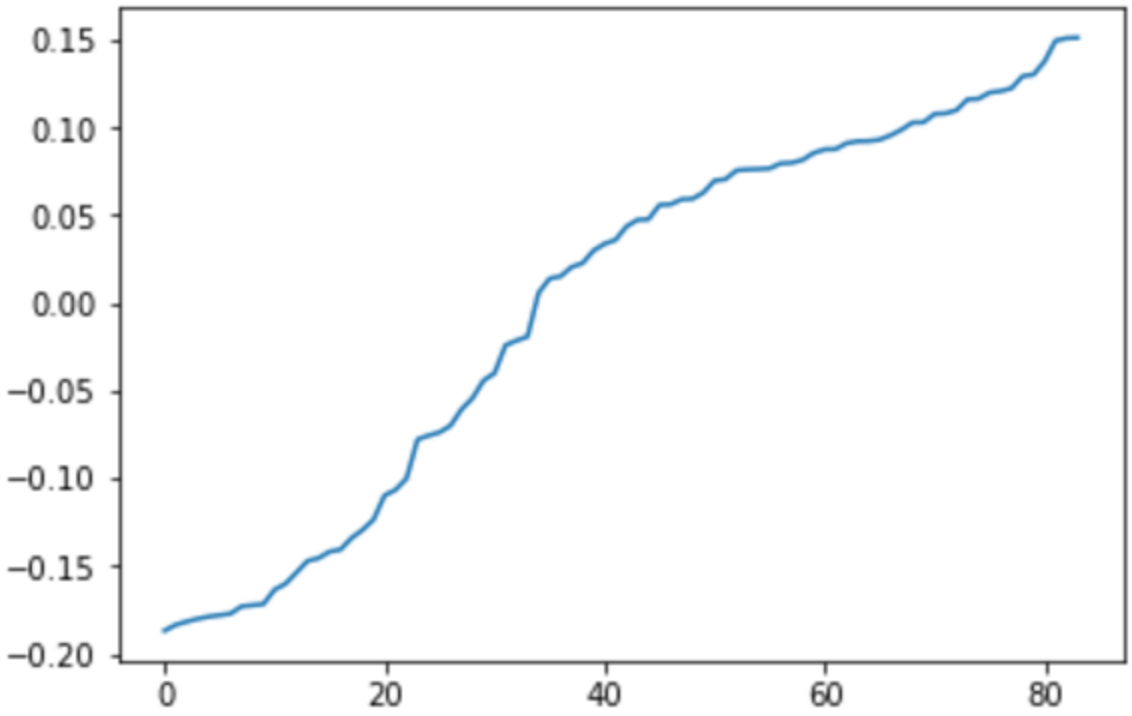
Dark Blue region depicts there were no weather sites at those places while all other regions form 1 connected component.This leads to the conclusion that the graph proposed by us has 1 connected component .

2nd eigenvector looked similar to this

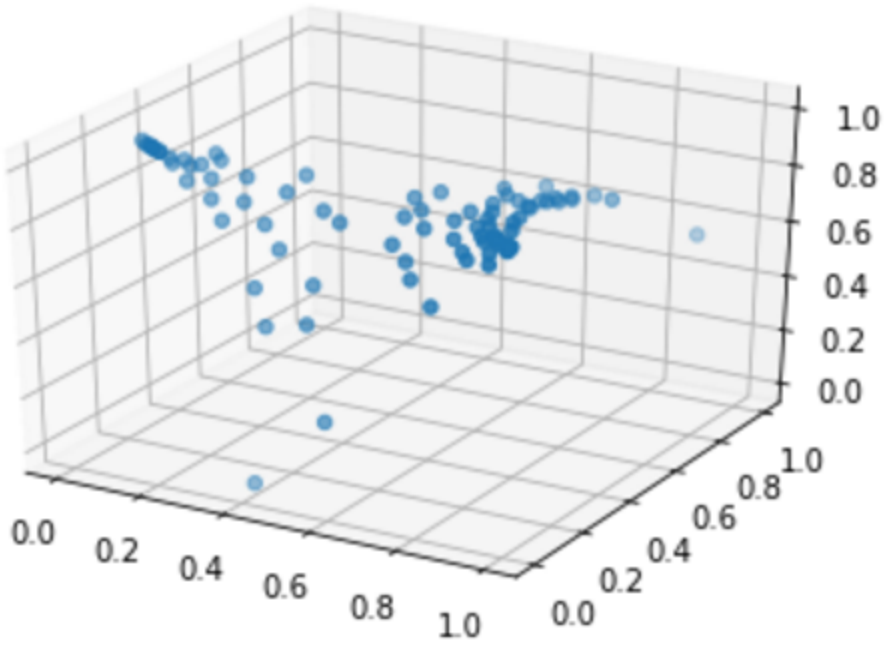


We can see two regions forming,one with yellowish-green shade and other with bluish shade.This depicts two components separated out in the graph and approximates the sparsest cut in the graph.

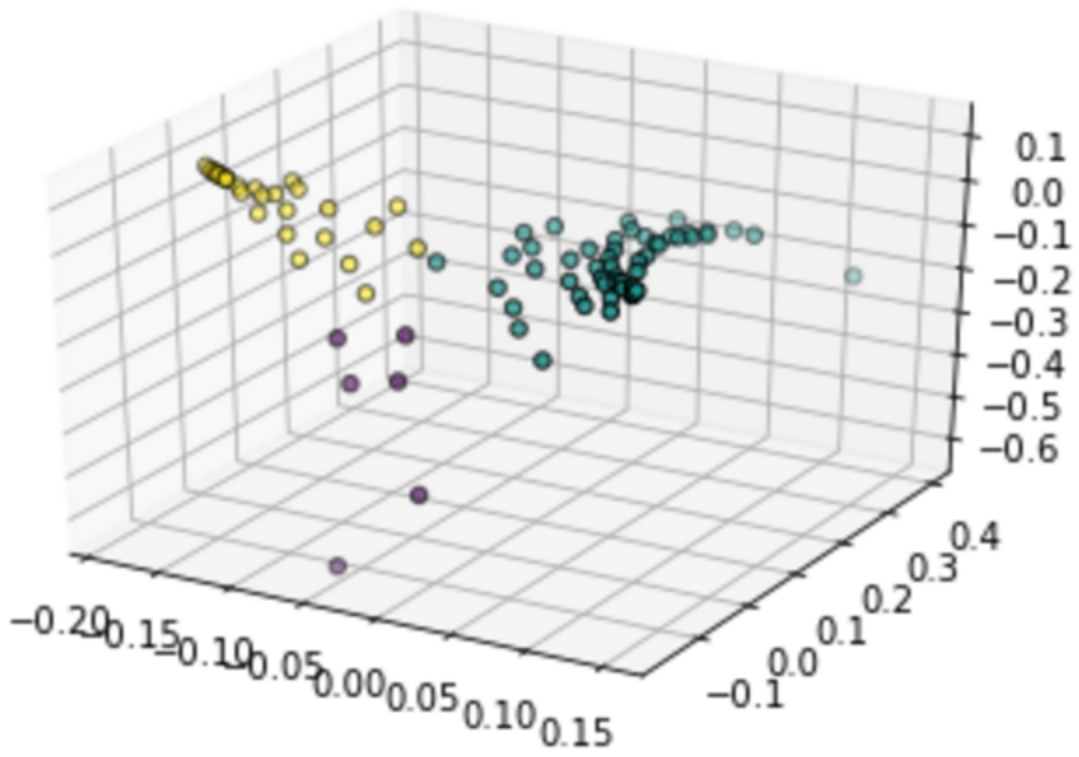
When the values of this eigenvector were sorted and plotted we could see a zero-crossing which depicts two regions in the graph.



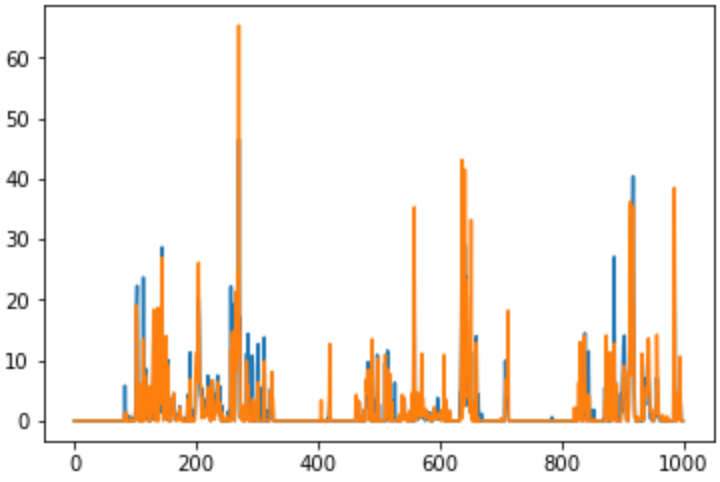
We now take the 3 smallest eigenvalues and plot the eigenvectors corresponding to those in 3D space .



And when we run K-means clustering on this projection we get 3 clusters .



Note that the sites falling under one cluster may not be spatially neighbours but they depict similar rainfall pattern. This is confirmed when we actually plot rainfall patterns of two sites far away from each other but falling under one cluster .



These two sites are not spatially neighbours but show almost similar rainfall pattern. Hence, we were able to establish spatial dependency for far-away neighbours as well.

We will be using neighbours based on these clusters to predict and test it with actual values .