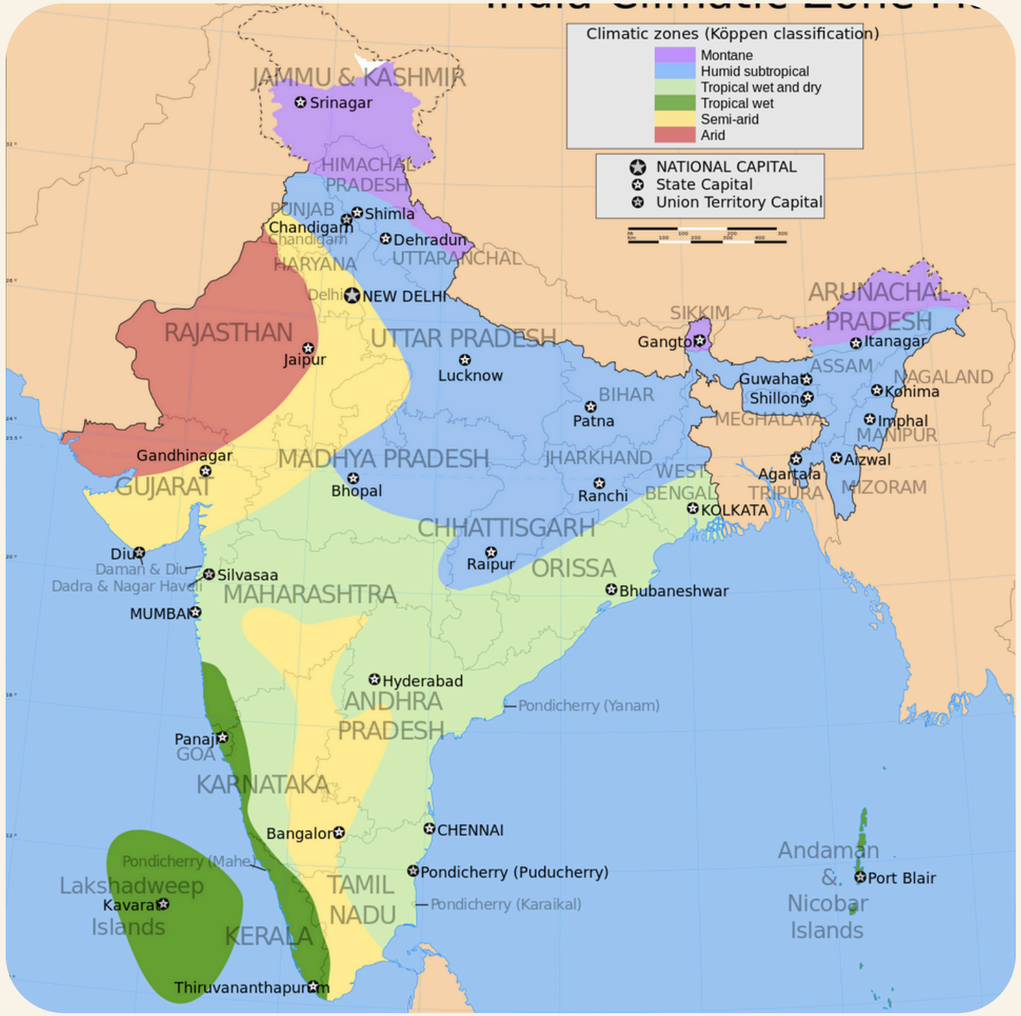
Cluster Analysis of pavement deterioration among climatic zones

Climatice zones of India

Objective: To divide the country into regions based on Climatic factors that affect pavement deterioration.

Features used for clustering:

- Average precipitation
- no. of wet days
- no. of high precipitation days
- Average daily maximum temperature
- no. of hot days
- Yearly maximum
- Average daily minimum temperature
- no. of cold days
- yearly minimum temperature
- average relative humidity



Data Collection

Required Climatic data of the last 10 years was downloaded from the Indian Meterological department, Pune website and NASA Data access portal.

The data for rainfall, minimum and maximum temperature was downloaded from IMD Pune website in the form of .grd files. These files were converted to usable .csv files using imdlib python library through the linked code found on Youtube

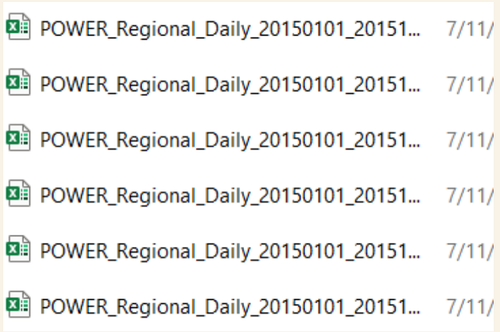
The humidity data was downloaded from the NASA Data access portal in the form of 6 .csv files with each having data of area of 10°x10° lat*lon. The data points were then adjusted to be similar with all other climatic data That was collected befire from IMD.

The raw data was then transformed and calculated to get final features for clustering.

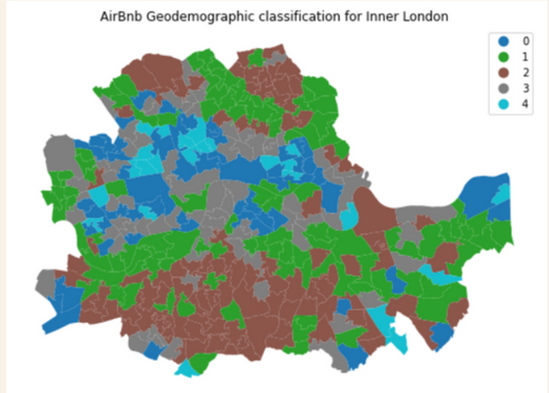
	A	B	C	D	E
		X	Y	1	2
37		73.5	8.5	32.87	32.52
40		76.5	8.5	31.04	30.94
41		77.5	8.5	32.04	31.97
42		78.5	8.5	31.11	30.93
71		76.5	9.5	31.27	31.18
72		77.5	9.5	30.67	30.67
73		78.5	9.5	29.44	29.3
74		79.5	9.5	30.61	30.16

X	Y	average	total	wetd
68.5	23.5	8.018193	2981.727	
69.5	21.5	7.921131	2944.821	
69.5	22.5	8.272256	3075.683	
69.5	23.5	8.361742	3109.377	
69.5	24.5	9.085993	3378.911	
69.5	27.5	10.88134	4047.351	
70.5	21.5	10.28052	3821.359	
70.5	22.5	8.719271	3241.752	

	X	Y	avghum
87	68.5	23.5	61.67446
69	69.5	21.5	67.25214
78	69.5	22.5	63.77786
88	69.5	23.5	51.87132
97	69.5	24.5	44.21612
122	69.5	27.5	33.35617
70	70.5	21.5	56.50129

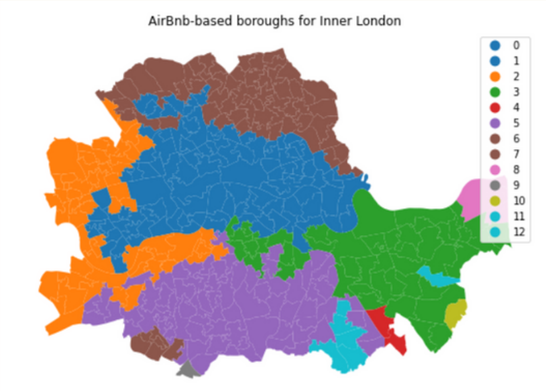


Spatially Connected Clusters: Uniting Data Points with Agglomerative Clustering



The final collected data was shorted down to 309 spatial points
The spatial weights of these points along with the climatic factors were inputed to a hierarchial clustering algorithm to form clusters

Regionalisation algorithm:



Regionalization is the subset of clustering techniques that impose a spatial constraint on the classification. In other words, the result of a regionalization algorithm contains areas that are spatially contiguous.

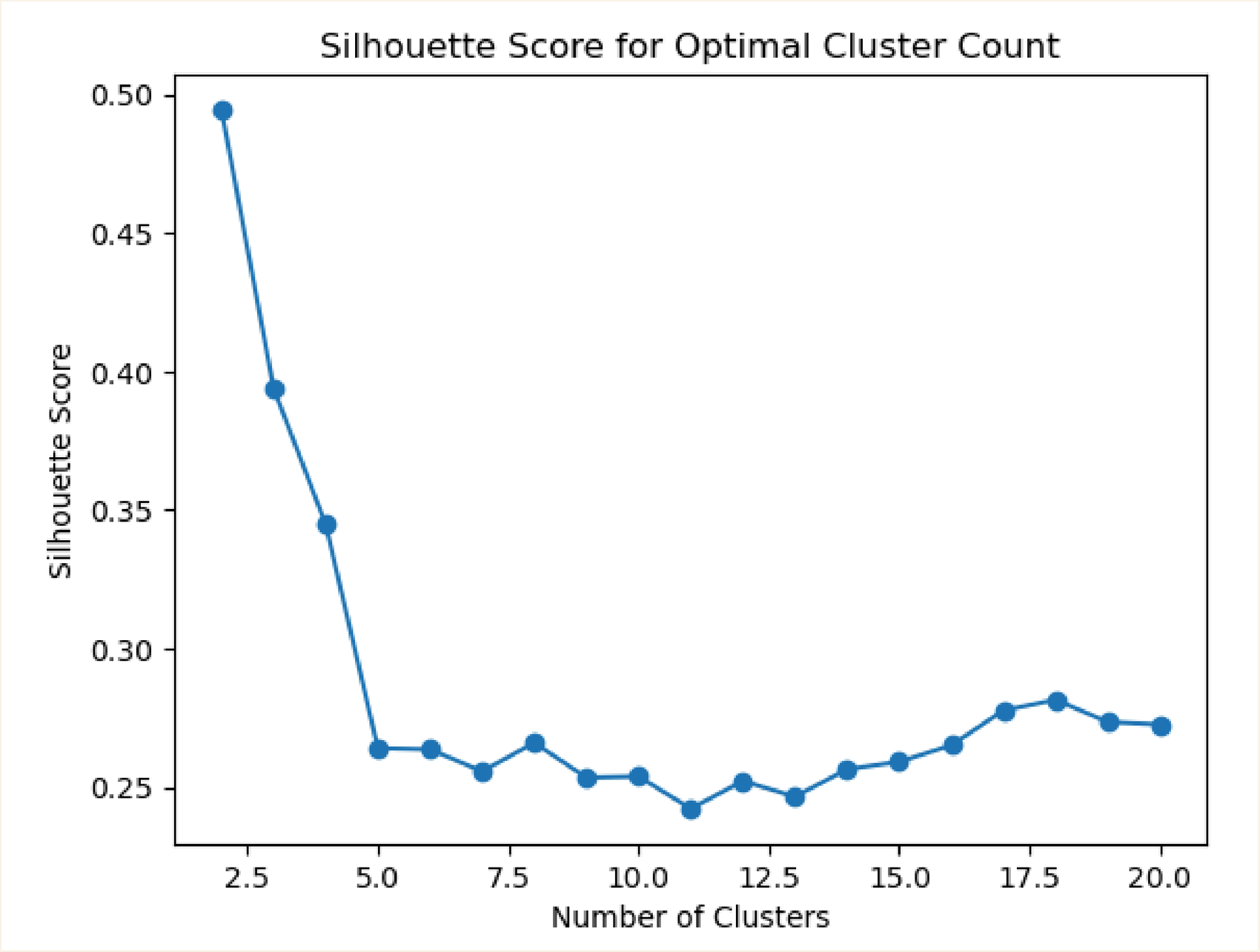
Agglomerative clustering with spatial connectivity constraint is one such regionalisation algorithm. It uses either spatial weights matrix or shape files to aggreagate areas that are neighbours to each other.

Here, **in our data**, we do not have spatial weights matrices or shape files. Rather, we have climatic data at known points. Hence, we use the geopandas and pysal libraries to convert latitudes and longitudes to geodataframe and then spatial weigths with queen criteria.

Reference here.

To find the optimal number of clusters, the elbow method using **silhouette scores** as evaluation metrics is used.
The silhouette scores of all the climatic variables was calculated and the elbow was found at **5 number of clusters.**

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

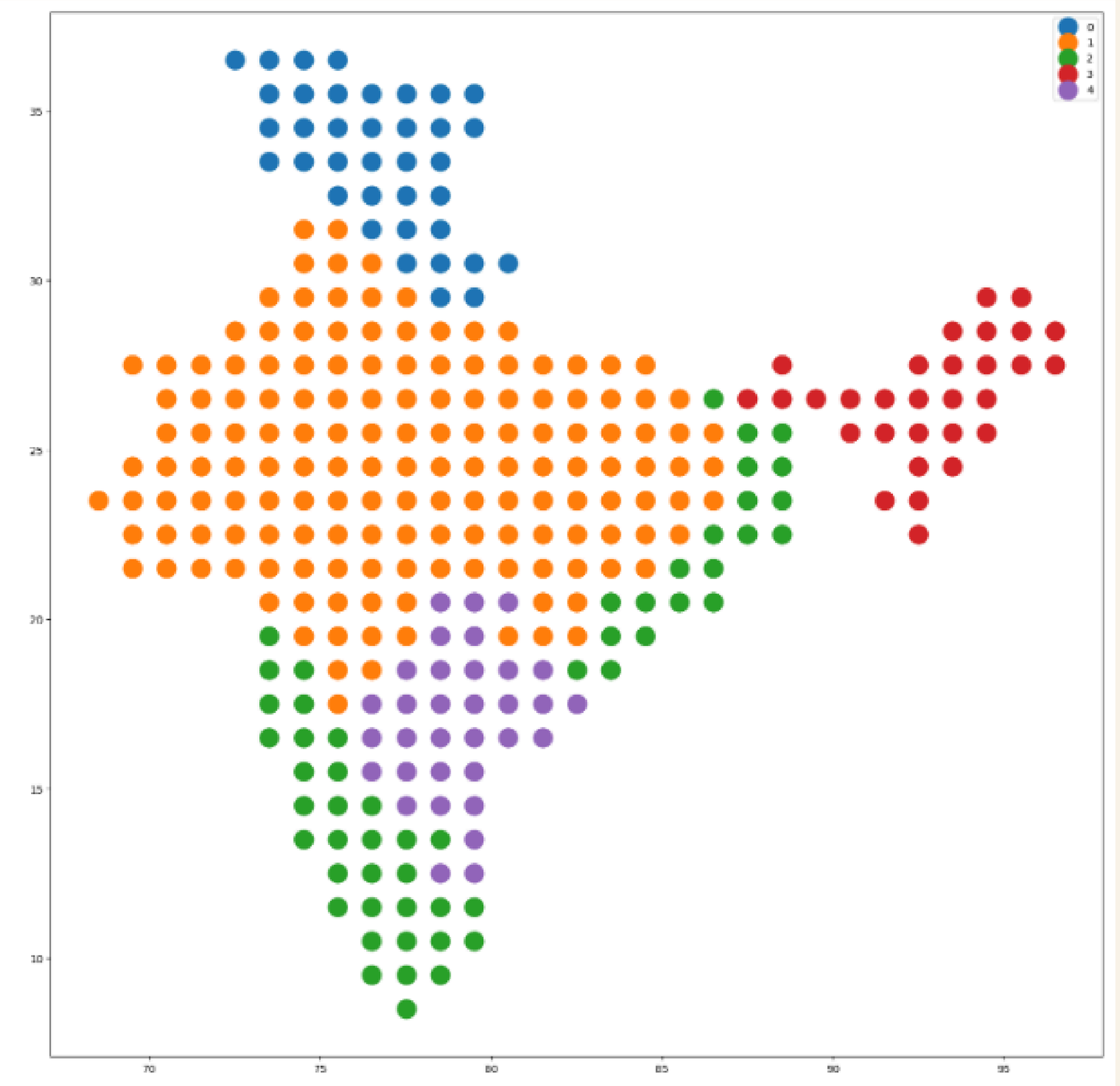


The following picture shows the cluster formation when rainfall, temperature and humidty are taken as features.

The **Davies-Bouldin Index and the Silhouette score** came out to be **1.1648** and **0.2661** for the above results.

This Davies Bouldin index represents the average 'similarity' of clusters, where similarity is a measure that relates cluster distance to cluster size. A model with a lower Davies-Bouldin index has a better separation between the clusters.

In our case, since the climatic data is continuous and irregular, the resulting Davies Bouldin Index is reasonable to be a little more than 1 and the silhouette score to be less than 0.5.



Preprocessing and Results

The features were standardized before being passed through clustering algorithm in order to ensue that there was no favour towards the features that naturally have larger magnitude like humidity. he difference in results with and withot standardisation are shown in the next slide.

0

	X	Y	average	wetdays	high_precip	maxaverage	yearlymax	hotdays	minaverage	yearlymin	avghum
count	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000
mean	76.689189	33.581081	15.056938	112.162162	28.567568	22.672967	34.060847	19.297297	12.962146	-1.239456	62.546940
std	2.092816	2.073354	3.371626	49.347134	18.403714	2.364418	2.350089	24.921950	2.516910	2.581675	6.803490
min	72.500000	29.500000	9.974820	17.000000	3.000000	19.233430	30.714497	0.000000	8.990645	-4.422795	51.871158
25%	75.500000	32.500000	11.153087	93.000000	4.000000	20.813515	32.218369	2.000000	11.226514	-3.058757	57.331925
50%	76.500000	33.500000	16.090432	124.000000	35.000000	21.630200	33.220177	4.000000	11.871618	-2.317508	61.434274
75%	78.500000	35.500000	17.924352	151.000000	44.000000	24.863992	35.806114	33.000000	15.220402	1.148496	67.624952
max	80.500000	36.500000	19.287686	167.000000	54.000000	27.002304	38.610123	82.000000	17.715110	3.603627	76.215904

1

average	wetdays	high_precip	maxaverage	yearlymax	hotdays	minaverage	yearlymin	avghum	sagg13cls
155.000000	155.000000	155.00000	155.000000	155.000000	155.000000	155.000000	155.000000	155.000000	155.0
11.806554	68.709677	20.83871	32.510431	43.776995	212.303226	21.073187	6.274766	50.996887	1.0
4.091051	23.339384	8.18653	1.013583	1.723511	27.229275	1.037090	2.227239	7.789554	0.0
5.139942	24.000000	4.00000	28.694117	38.417915	146.000000	18.074120	2.841949	33.356171	1.0
8.871994	52.500000	15.00000	31.917870	42.856134	198.000000	20.502214	4.655874	44.890640	1.0
10.311999	68.000000	20.00000	32.603501	44.330505	212.000000	21.117941	5.663892	51.349397	1.0
14.990567	83.500000	28.00000	33.160722	44.906582	226.000000	21.511383	7.792997	56.406764	1.0
21.085709	125.000000	45.00000	34.666935	46.454216	300.000000	23.591546	12.159250	68.526247	1.0

2

average	wetdays	high_precip	maxaverage	yearlymax	hotdays	minaverage	yearlymin	avghum	sagg13cls
54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.0
14.524318	124.166667	40.203704	31.748674	38.102261	174.203704	21.992035	12.299358	69.644079	2.0
5.697192	33.652918	19.372593	1.236932	2.852381	67.007851	1.553061	3.400588	6.024755	0.0
4.249553	62.000000	14.000000	27.938279	33.048790	13.000000	17.661298	6.778946	54.324671	2.0
10.608614	103.500000	29.000000	31.116057	35.798718	138.500000	20.572546	10.252766	66.647652	2.0
15.318159	120.500000	36.000000	32.008716	38.019247	197.500000	22.118846	11.715906	69.841651	2.0
18.203120	136.750000	45.750000	32.445105	40.147143	215.500000	22.941892	13.763930	72.985264	2.0
31.110820	239.000000	94.000000	34.161010	44.619335	271.000000	25.192642	20.331005	80.204062	2.0

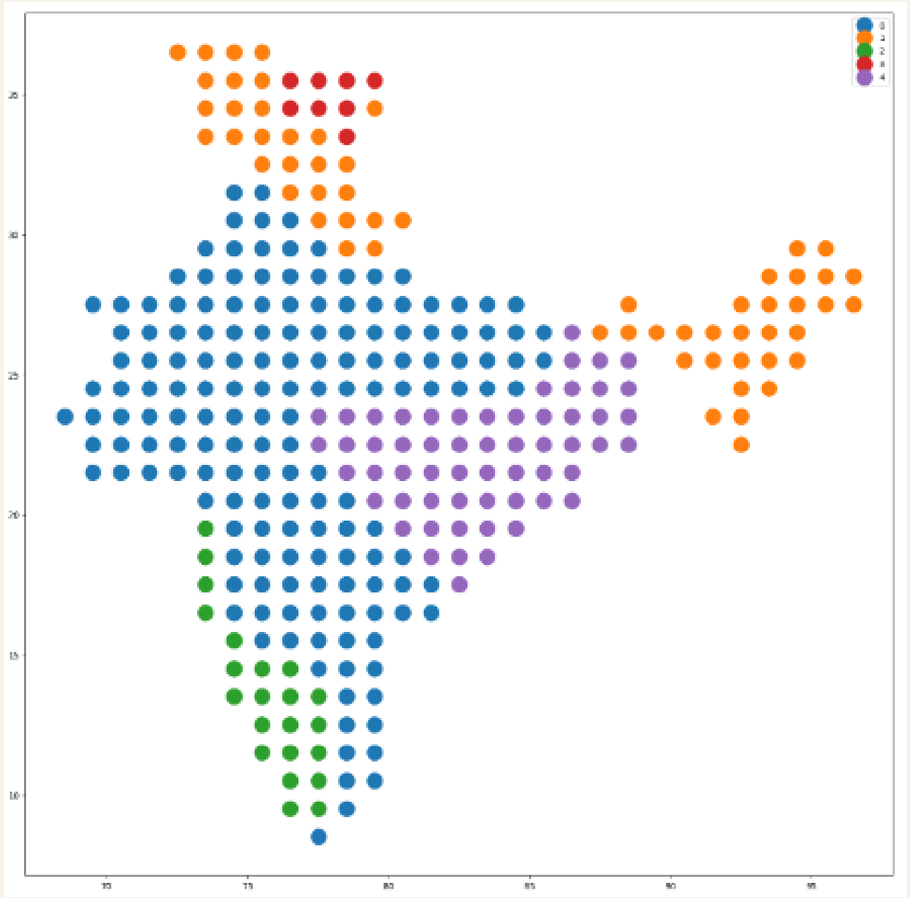
3

	X	Y	average	wetdays	high_precip	maxaverage	yearlymax	hotdays	minaverage	yearlymin	avghum
count	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000
mean	92.800000	26.533333	21.252708	174.833333	66.266667	29.031193	35.511650	73.500000	20.000692	6.766897	78.144574
std	2.336369	1.731719	6.717118	23.769124	22.869847	1.028810	1.452774	38.658318	0.633802	0.549699	5.032754
min	87.500000	22.500000	12.494647	130.000000	31.000000	26.868128	32.741028	12.000000	18.872981	5.395269	62.205651
25%	91.500000	25.500000	16.549008	157.500000	52.250000	28.584845	34.377373	48.250000	19.501826	6.628497	76.952067
50%	93.000000	26.500000	19.675250	171.000000	62.500000	28.893473	35.284666	60.500000	20.039592	6.839505	78.140116
75%	94.500000	27.500000	25.386980	196.250000	80.500000	29.924488	36.736464	113.000000	20.595096	7.104780	80.253372
max	96.500000	29.500000	40.947692	225.000000	113.000000	30.309222	38.215714	124.000000	20.906215	7.583978	87.620000

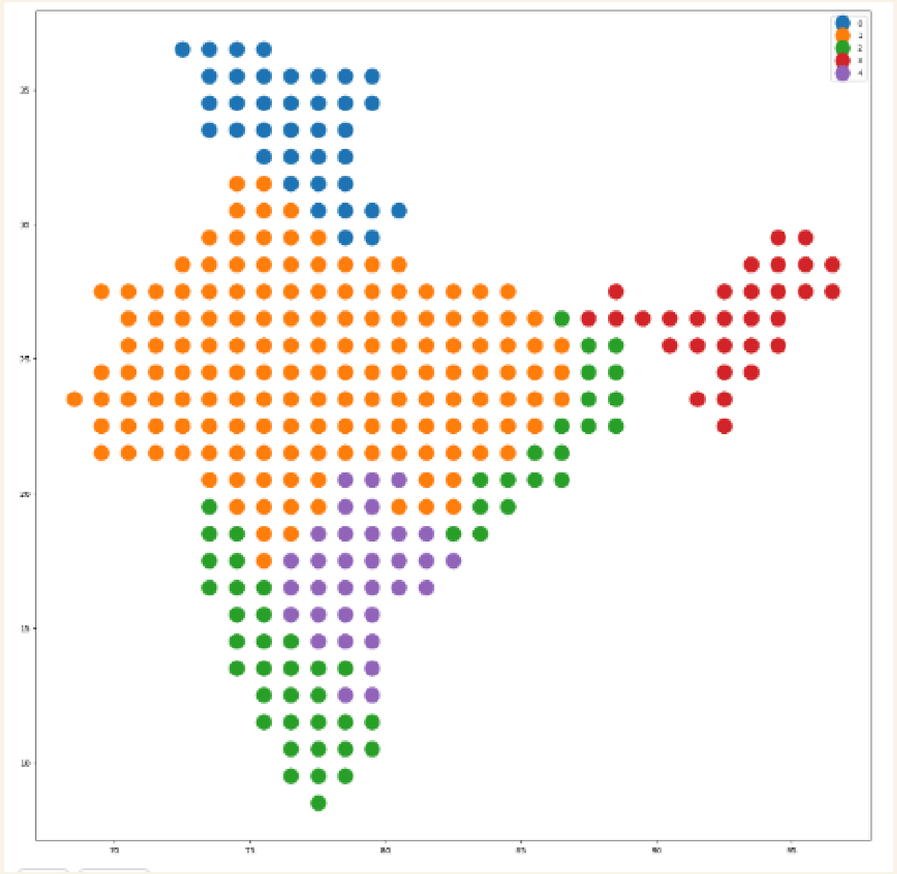
4

	X	Y	average	wetdays	high_precip	maxaverage	yearlymax	hotdays	minaverage	yearlymin	avghum
count	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000
mean	79.075758	16.924242	7.294249	96.757576	27.151515	33.989343	43.508978	262.484848	22.940391	11.807396	61.150344
std	1.521313	2.121767	2.454142	13.339393	8.832190	0.751997	2.290202	30.037603	0.895069	2.308418	7.173224
min	76.500000	12.500000	3.489025	70.000000	9.000000	32.035634	38.142365	189.000000	21.598161	7.557801	50.767418
25%	78.500000	15.500000	5.503795	87.000000	23.000000	33.608489	42.195511	251.000000	22.245184	10.412966	55.084884
50%	79.500000	17.500000	7.012450	94.000000	27.000000	34.268149	44.115711	273.000000	22.724657	11.724752	60.792877
75%	79.500000	18.500000	8.874582	107.000000	33.000000	34.450392	45.190971	284.000000	23.586974	13.013834	66.999836
max	82.500000	20.500000	14.014413	124.000000	48.000000	35.108548	46.874161	308.000000	24.517815	17.091764	75.440000

0-blue
1-orange
2-green
3=red
4-violet



without standardising



with standardising

The after results seem to be favourable when compared with the IRI clusters as their distribution was also simlar in the mid-region.

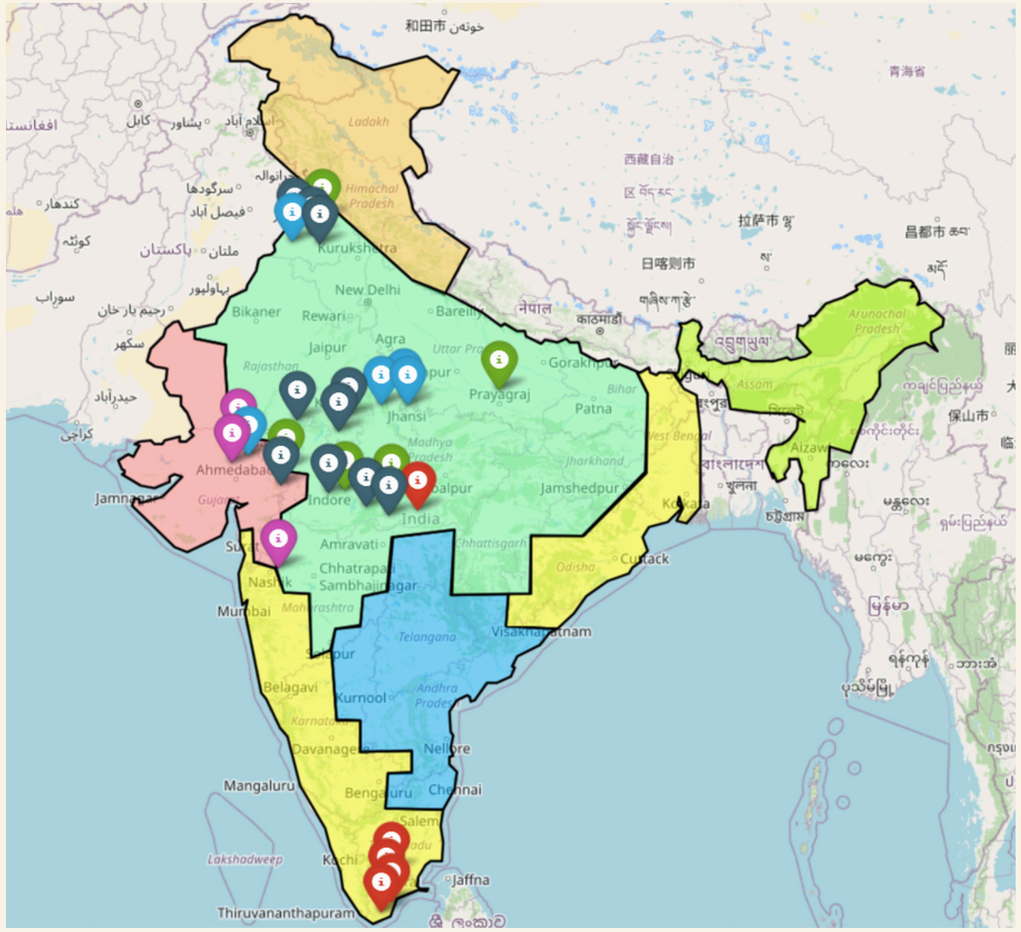
Validation of IRI difference with pavements climatic zone

The next part of the project deals with analysing how IRI is behaving with respect to the climatic zone that pavement falls in.

To find out the patterns of pavement deterioration with respect to the climate, we first cluster all the IRI data we have into different ranges. The process is elaborated in the upcoming slides.

Steps followed:

- Cleaning and transforming raw IRI data
- Assigning the IRI data to categorical variables
- Performing statistical analysis to find out patterns
- Clustering the categories together according similar IRI-difference range
- Comparing the locations of similar categories with obtained climatic zones



Data Preprocessing

The data was provided by MORTH,India and NHAI. The data consisted of section-wise IRI measurements for a set of national highways of India over multiple dates.

The IRI measures cannot be directly compared to find out patterns in pavement deterioration. Hence the difference of the IRI measurements of two dates for the same section of road is consired to be a factor. the change observed over an year of time in different regions is compared for our analysis.

	NH Number(New)	District	Start Chainage(km)	End Chainage(km)	Lane Number	IRI-15	IRI-16	IRI-17	16-15diff	16-17diff
0	NH0752	Jhalawar	89.517	89.6	L1	2.41	5.19	5.98	2.78	0.79
1	NH0752	Jhalawar	89.600	89.7	L1	2.41	4.29	5.87	1.88	1.58
2	NH0752	Jhalawar	89.700	89.8	L1	2.48	4.46	4.56	1.98	0.10
3	NH0752	Jhalawar	89.900	90.0	L1	3.24	3.50	3.78	0.26	0.28
4	NH0752	Jhalawar	90.200	90.3	L1	3.25	3.31	3.41	0.06	0.10
...	...	...	...	...	...	...	...	...	...	...
282	NH0138	Tuticorin	18.900	18.8	R2	1.52	2.62	2.79	1.10	0.17
283	NH0138	Tuticorin	22.100	22.0	R2	2.00	2.52	3.01	0.52	0.49
284	NH0138	Tuticorin	38.100	38.0	R2	2.59	2.68	2.82	0.09	0.14
285	NH0138	Tuticorin	39.700	39.6	R2	2.26	2.38	2.99	0.12	0.61
286	NH0138	Tuticorin	32.800	32.9	L2	2.30	2.51	3.40	0.21	0.89

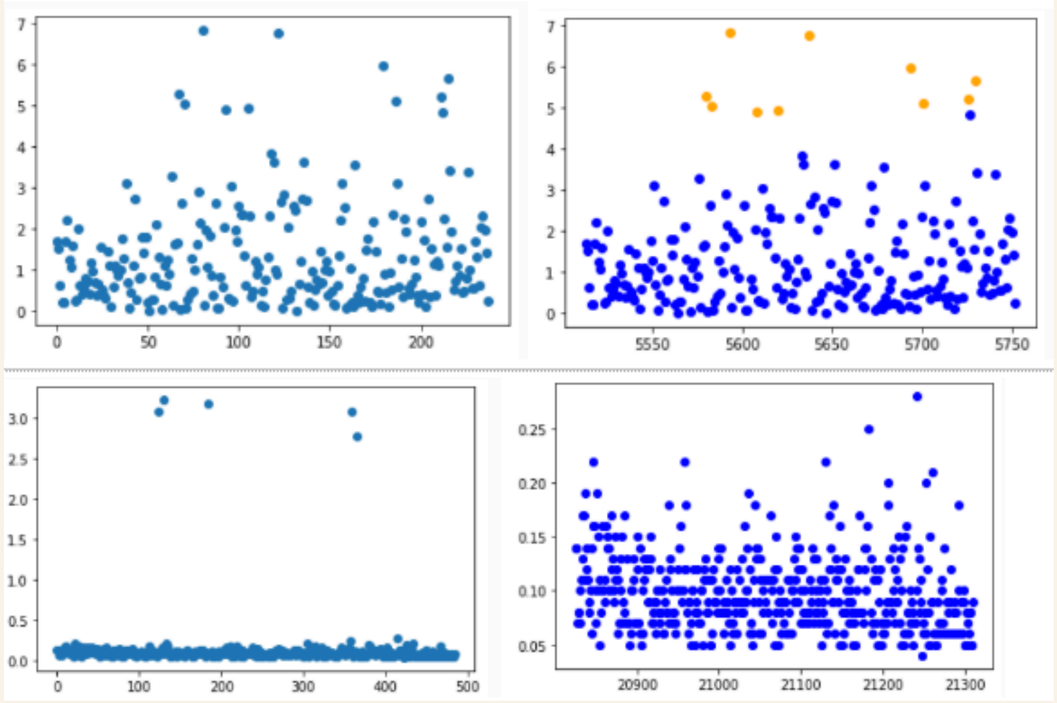
Data cleaning

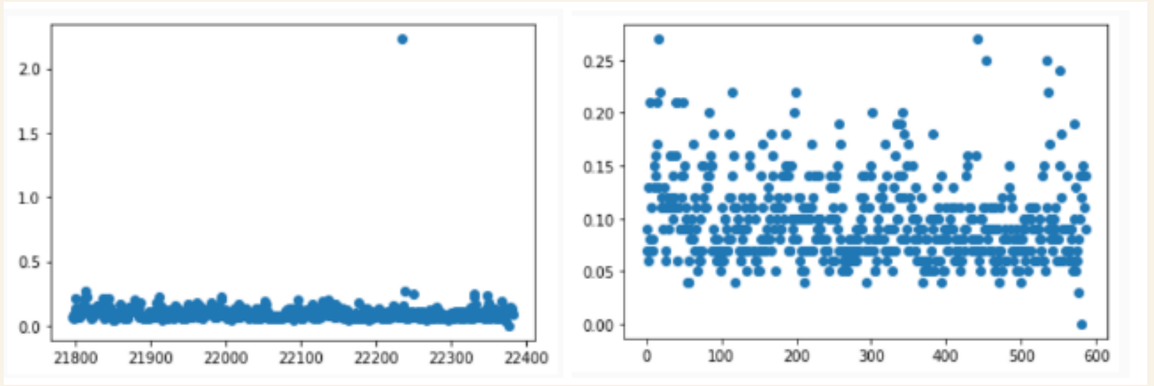
All the measurements of IRI differnce greater than 3 are ignored as they are practically rare and might be wrong measurements.

Outliers are detected and removed using DBSCAN clustering algorithm

The figure shows the type of points that are being considered outliers. (orange ones are outliers and fig2 shows the data after removal.

The range of IRI-difference where the section doesn't have much points are being considered outliers. Such range is found out for every section and removed.





Statistical understanding of the data

In order to cluster the data in a useful way, we have to consider the geographical locations too.

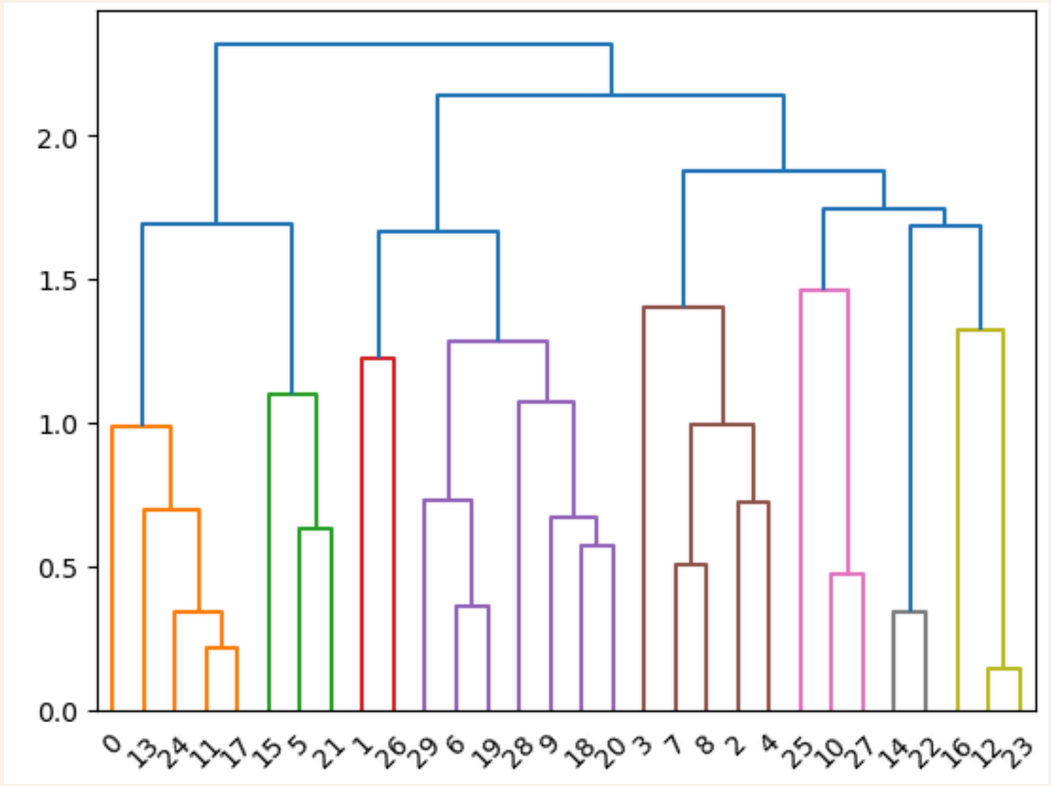
To choose from the above two, similarity is being checked in the two cases. A **null hypothesis** is proposed that there are no significant differences between the means of the groups. The null hypothesis is rejected when the p-value is less than 0.05

The Kruskal test(alternative to ANOVA test) was conducted for two cases

- single Highway, multiple districts. data of each NH$\cap$Di is taken as a group and there are j such groups (0<i<j) {p-value average =
- One district, multiple highways, data of each D$\cap$NHi is taken as a group and there are j such groups (0<i<j)

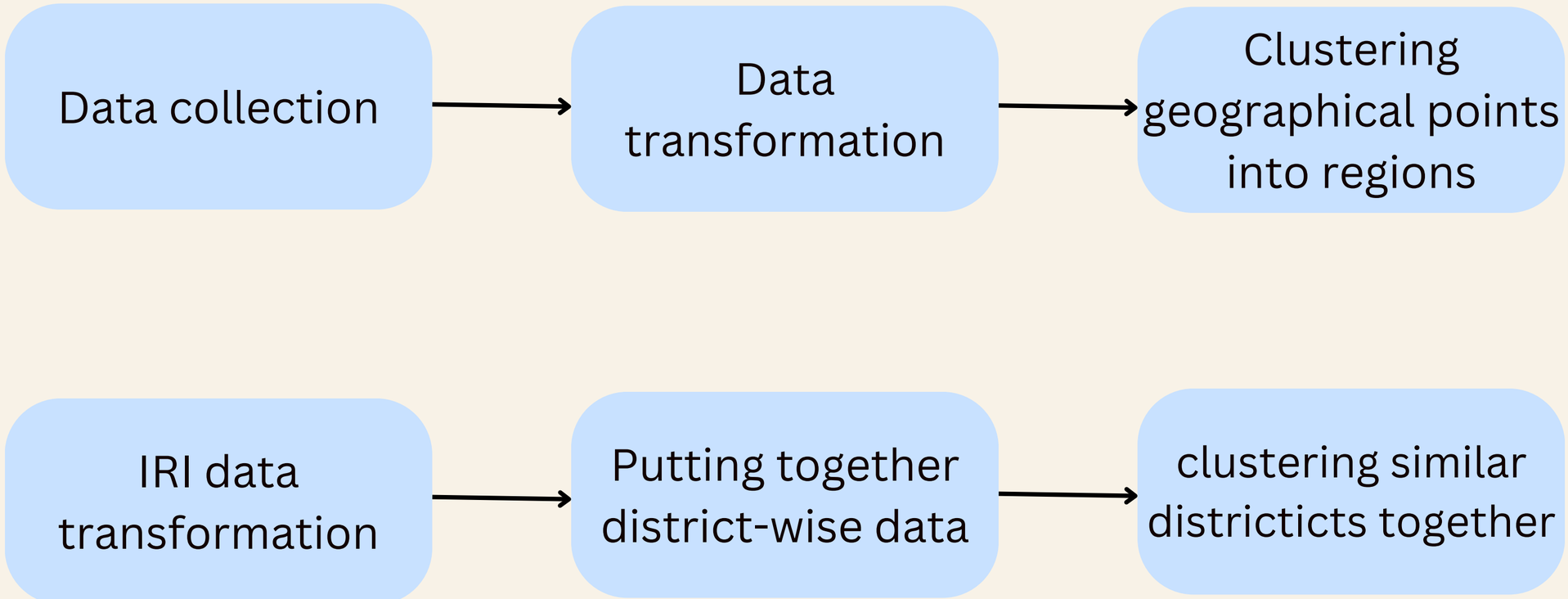
The results showed that the data within a district are more similar than in the first case. So, district wise clustering is chosen

Clustering districts with similar ΔIRI ranges



Cluster 1: [0, 5, 11, 13, 15, 17, 21, 24]
Cluster 2: [1, 6, 9, 18, 19, 20, 26, 28, 29]
Cluster 3: [2, 3, 4, 7, 8]
Cluster 4: [10, 25, 27]
Cluster 5: [12, 14, 16, 22, 23]

- Kruskal test is performed on all the districts' data and found out that they are all significantly different
- A pairwise t-test which is Conover test is performed for all the datasets and the p-value matrix is obtained
- The p-value matrix is being passed into the hierarchial clustering algorithm with multiple types of linkage methods.
- The average of p-values of each cluster obtained from different methods is used as evaluation metrics.
- It is found out that the highest p-value is obtained in case of average linkage method.
- The datasets in each cluster are visualised and finally choosen for comparing with climatic zones.

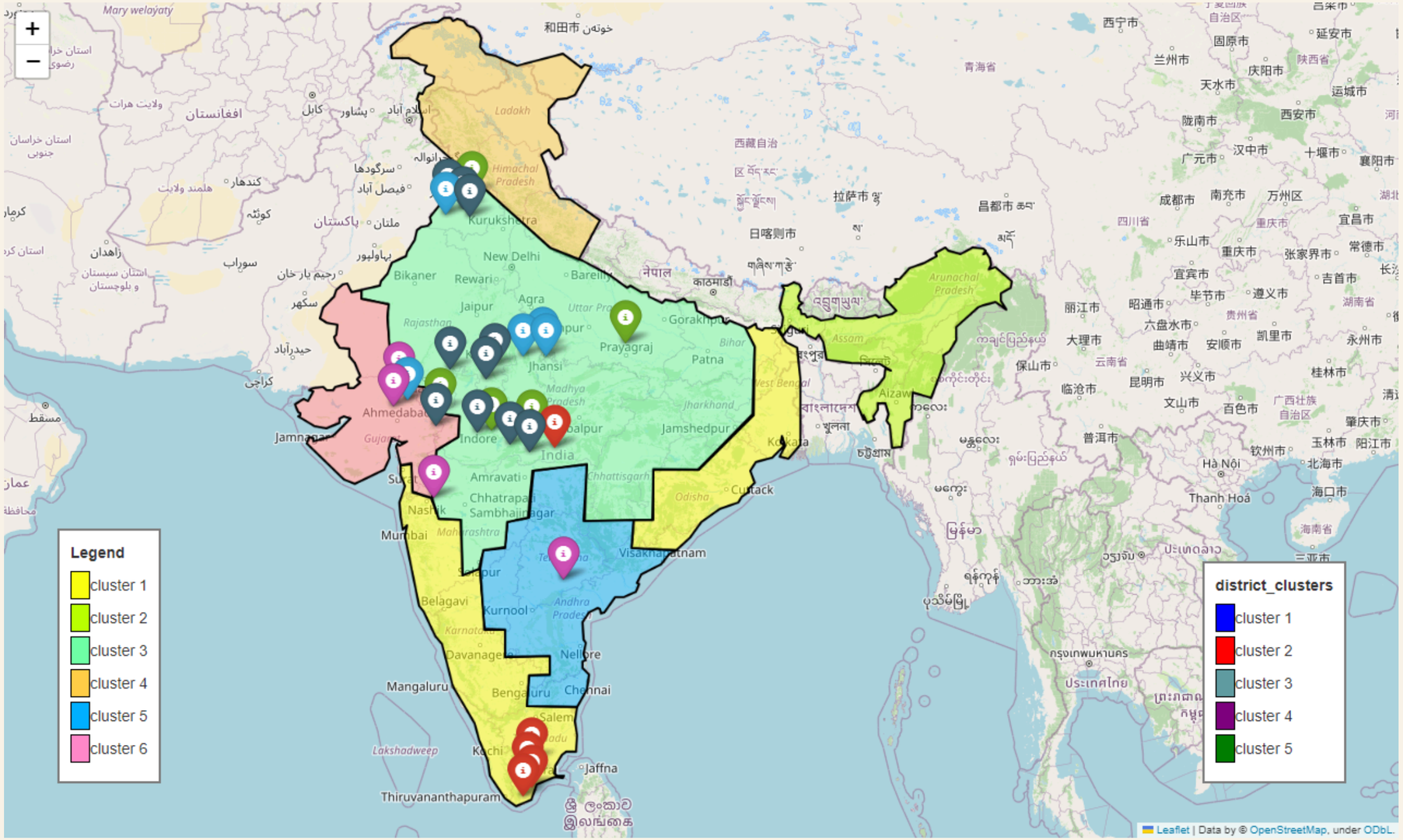

Data collection
Data transformation
Clustering geographical points into regions
IRI data transformation
Putting together district-wise data
clustering similar districticts together

Plotting the clusters and climatic zones together

The districts are plotted with the colour corresponding to the cluster they belong on the map that contains climatic zones.

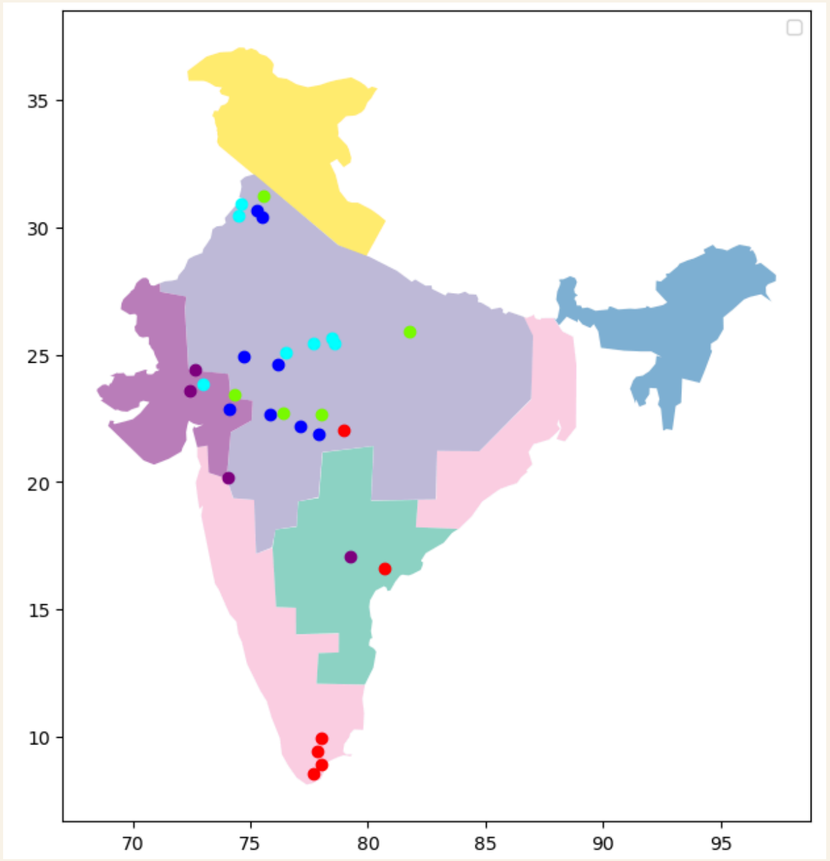
A pattern can be seen that the districts that belong in cluster 1,3 and 5 are lying in the north and eastern parts of the country and others in southern parts.

It is to be noted that both the lusters 2 and 4 show very less deterioration compared to other clusters.



Project Proposal

- Borcelle Studio -


35
30
25
20
15
10
70
75
80
85
90
95