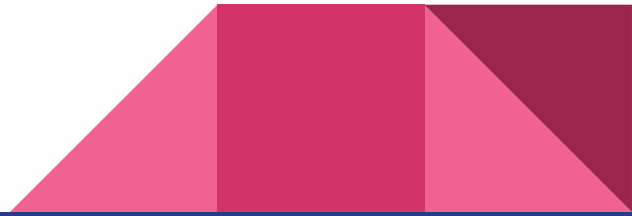
Clustering Districts on Basis of Spatial Variation of Attributes

Recap

- For each attribute, we have a curve that shows variation in attribute as distance from the district hotspot center increases
- We wished to cluster districts on the basis of these curves
- Aim - Districts with similar shapes of curves grouped together in same clusters
- 2 Methods of clustering -
  - K-Means
  - Dynamic Time Warping Agglomerative Clustering



Analysing Results - Choosing # of Clusters

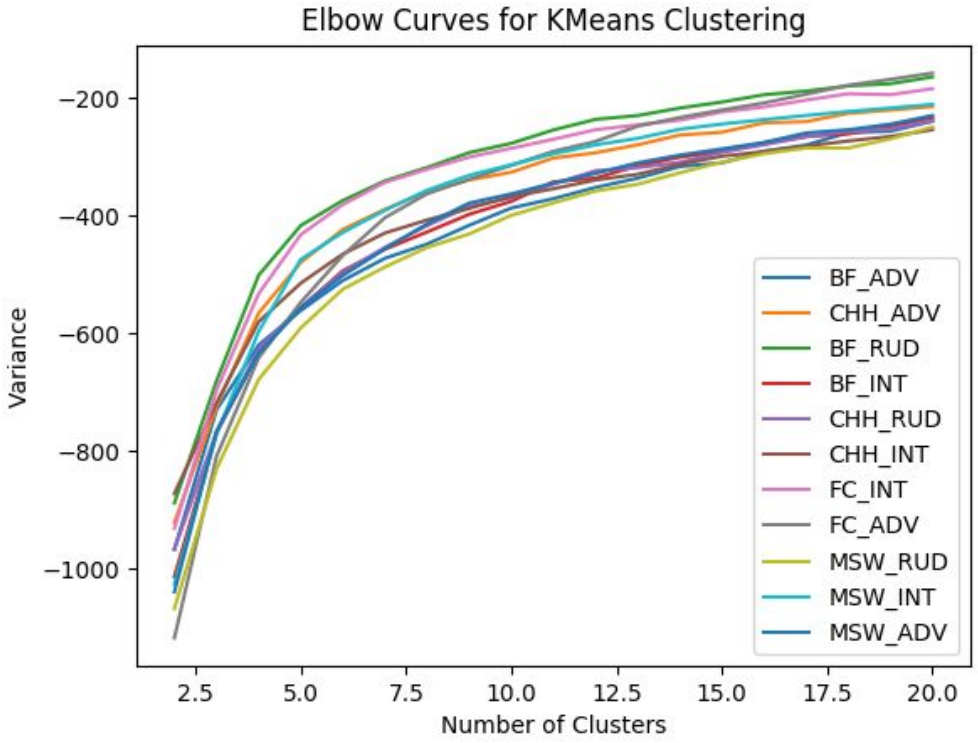
3 Common methods for determining number of clusters :-

- **Elbow Curve** (Only Applicable on KMeans)
- **Silhouette Analysis** (Applicable on both)
- **Gap Statistic** (Applicable on both)

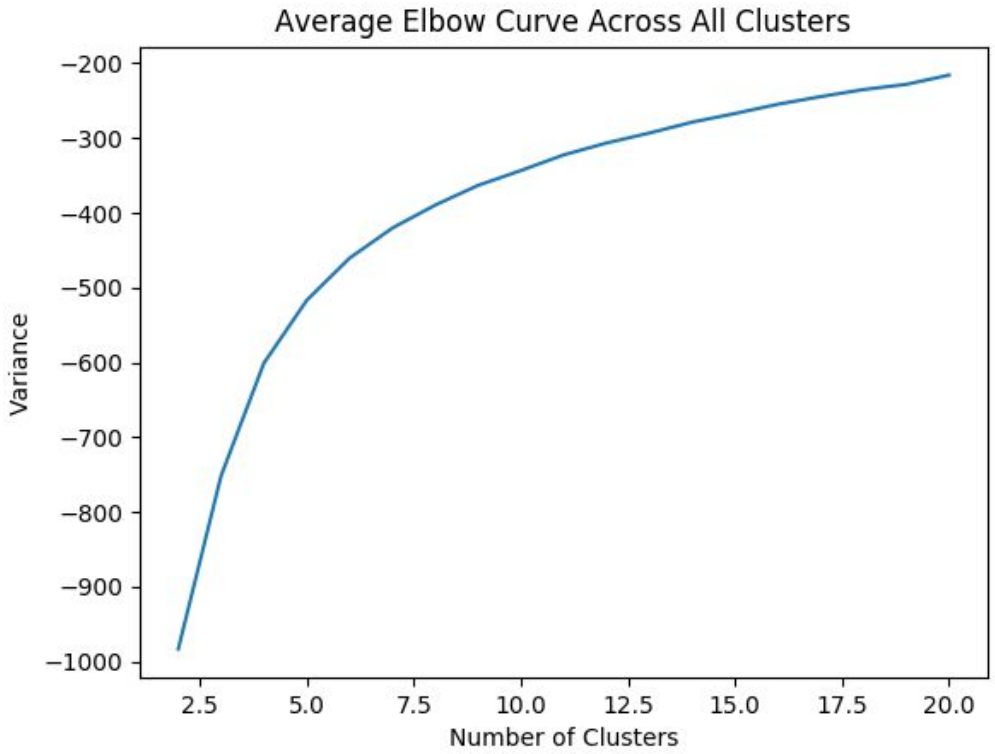
We will deal with Elbow Curve and Silhouette Analysis

Elbow Curve

Avg Elbow Curve Across All Attributes

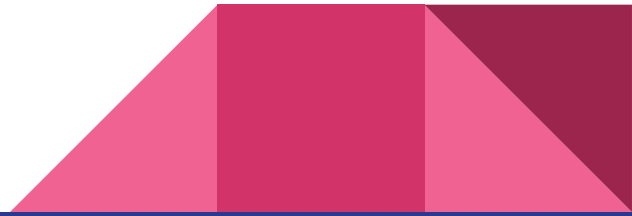


Elbow Curve For All Attributes



Results

- No particularly visible elbow
- A smooth, almost hyperbola like curve for all attributes
- Let’s look at other methods of analysis for choosing number of clusters

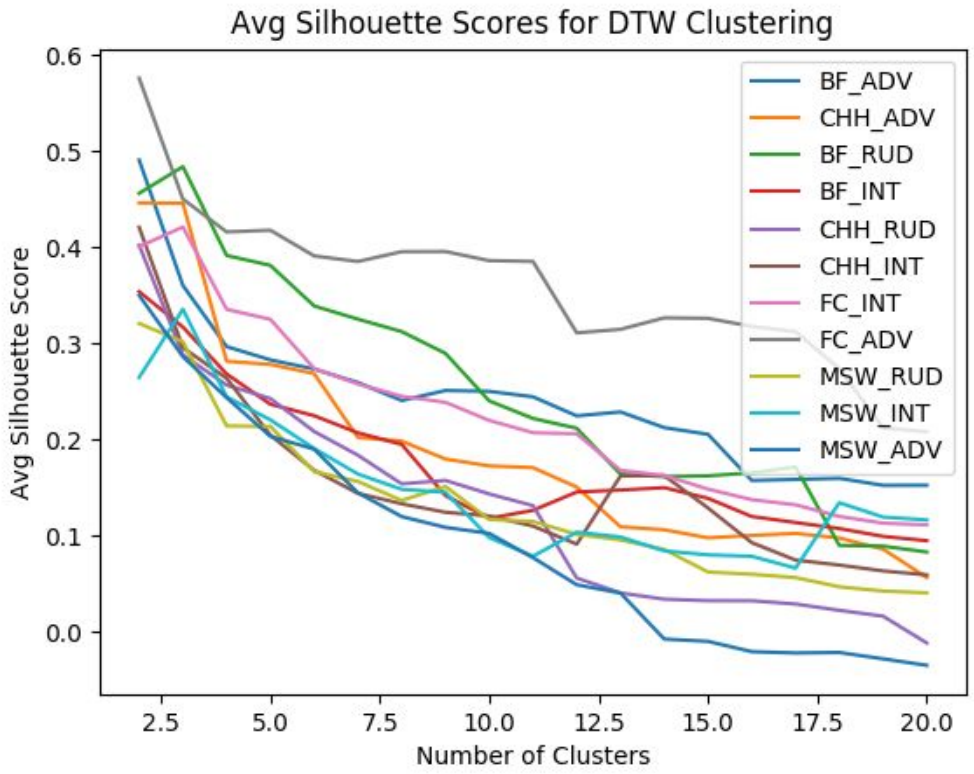


Silhouette Analysis

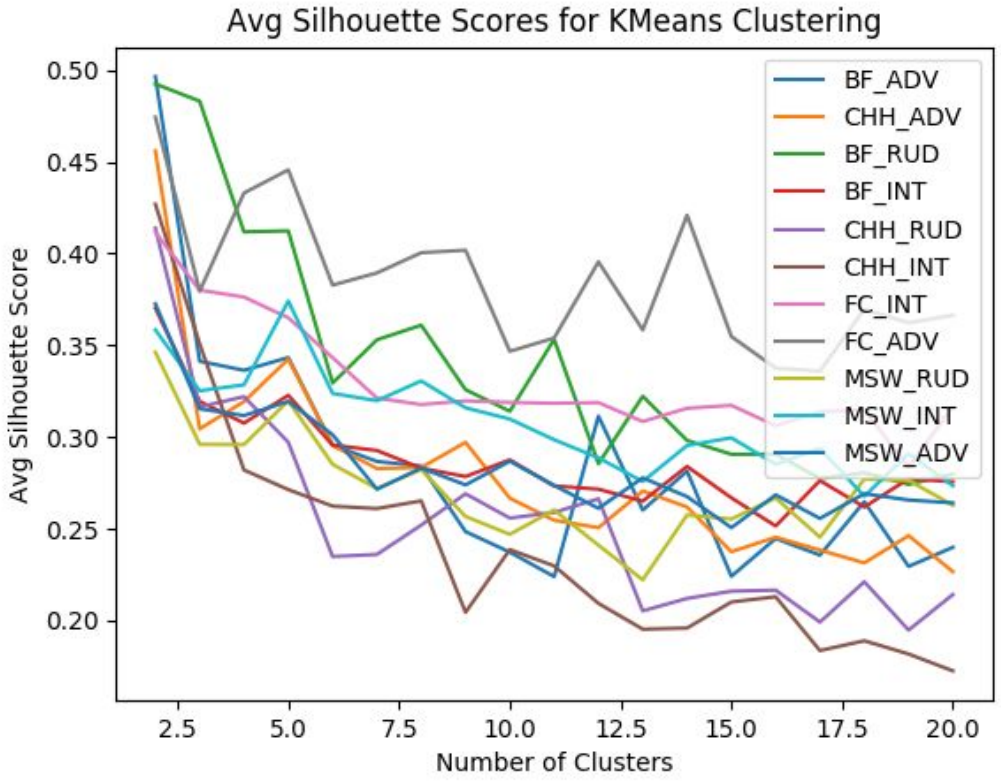
Silhouette Analysis

- Developed by Peter J. Rousseeuw (1987)
- Study of separation distance between clusters for choosing # of clusters
- Silhouette measure for 1 point in a particular cluster is a value **ranging from [-1,1]** indicating how far that point is from neighbouring clusters
- Values near +1 indicate point **far away** from neighbouring cluster
- Values near 0 indicate point **on or very close** to decision boundary
- Negative values indicate sample possibly **assigned to wrong cluster**
- Taking average across all points give Avg Silhouette Score for particular clustering. **The higher, the better.**

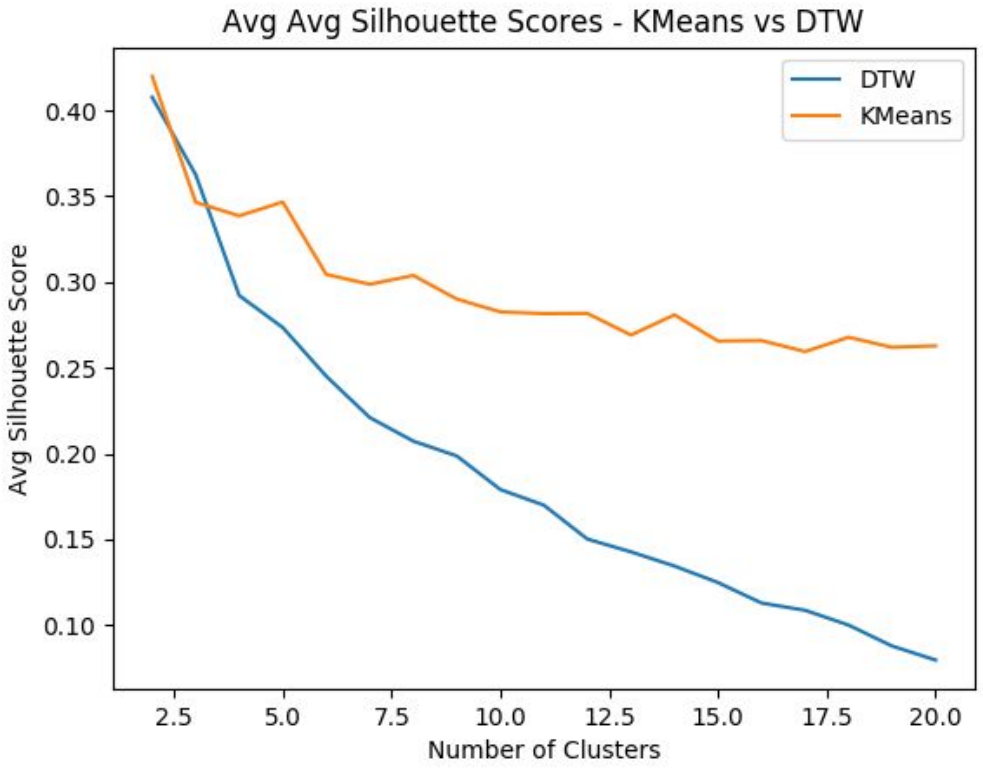
Avg Silhouette Score - DTW Clustering



Avg Silhouette Score - KMeans Clustering



Avg of Avg Silhouette Score - KMeans vs DTW



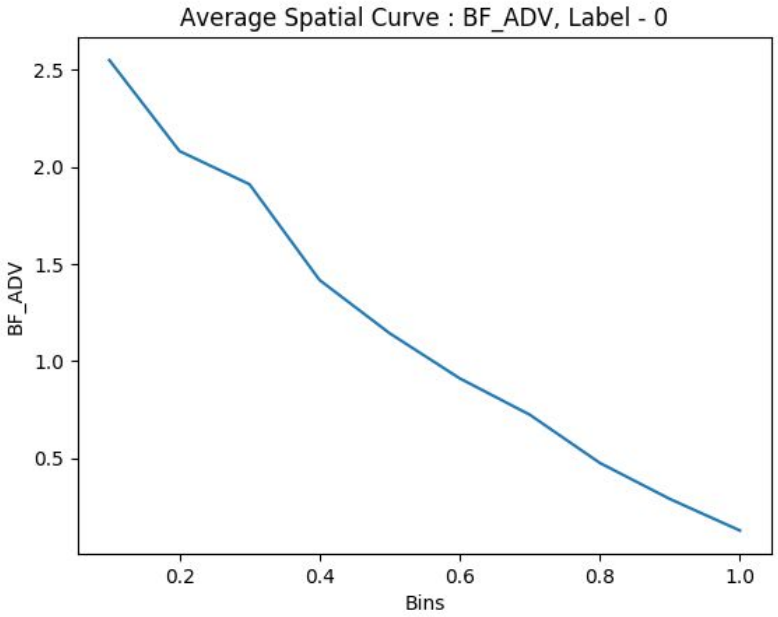
Results

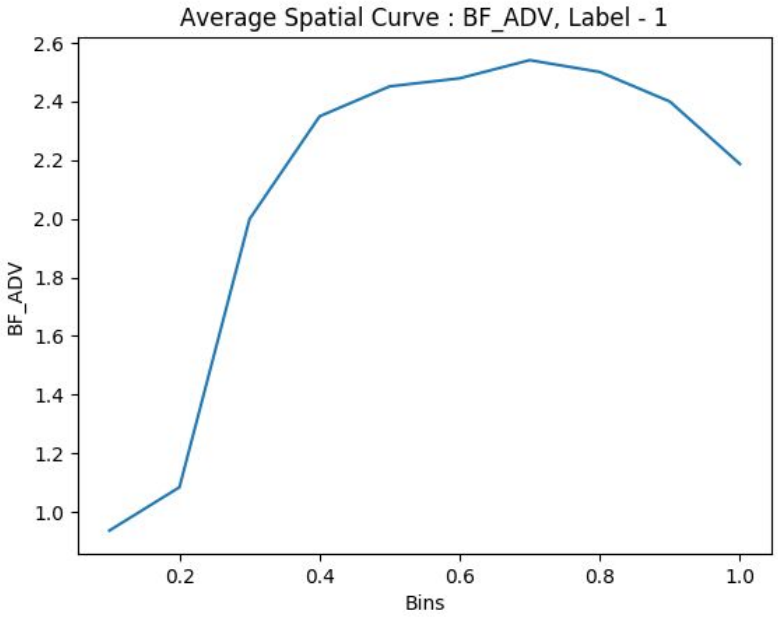
- The higher the avg silhouette score, the better
- For all attributes, **2 clusters gave highest avg silhouette score** (for both methods)
- Avg of avg silhouette score across all attributes (KMeans) - 0.419
- Avg of avg silhouette score across all attributes (DTW) - 0.407
- They are in the same ballpark region, so there is **little motivation of choosing DTW (a complicated clustering method) over KMeans (much simpler), according to this method**

Results of KMeans with 2 Clusters

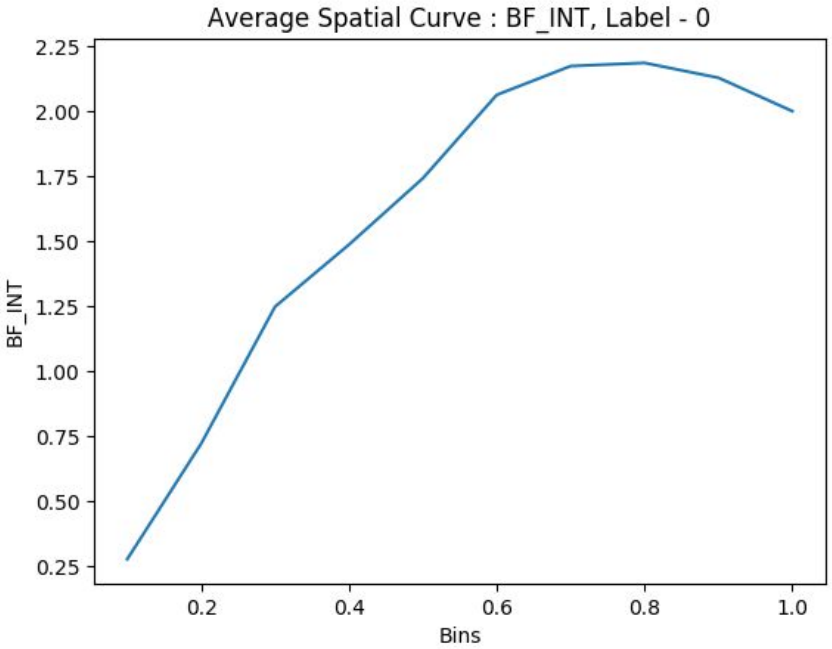
- For all attributes, I ran KMeans with 2 clusters, and then for each cluster, I plotted the average curve across all districts, ie, the curve obtained by taking the average across each district’s curve
- The following slides show the results

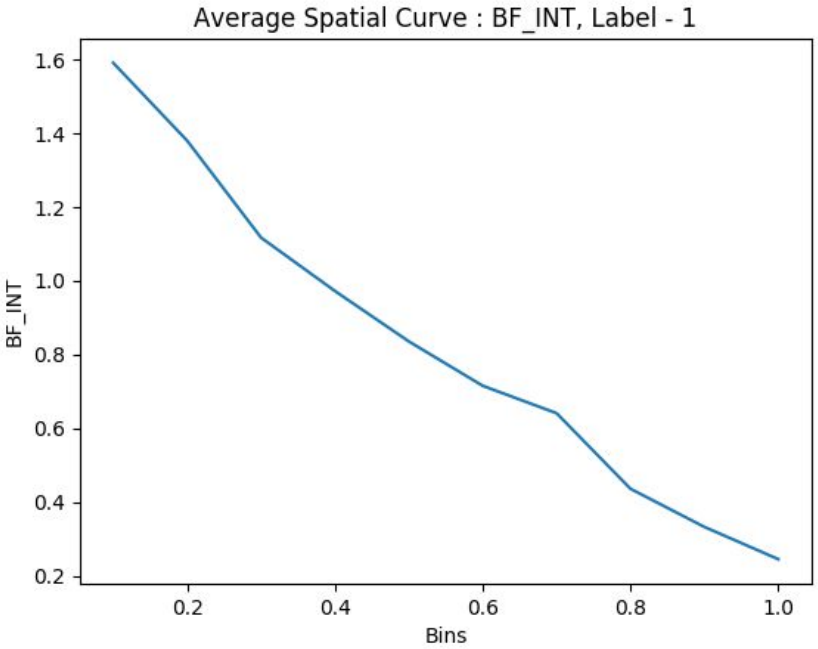
BF_ADV


Average Spatial Curve : BF_ADV, Label - 0
BF_ADV
2.5
2.0
1.5
1.0
0.5
0.2
0.4
0.6
0.8
1.0
Bins

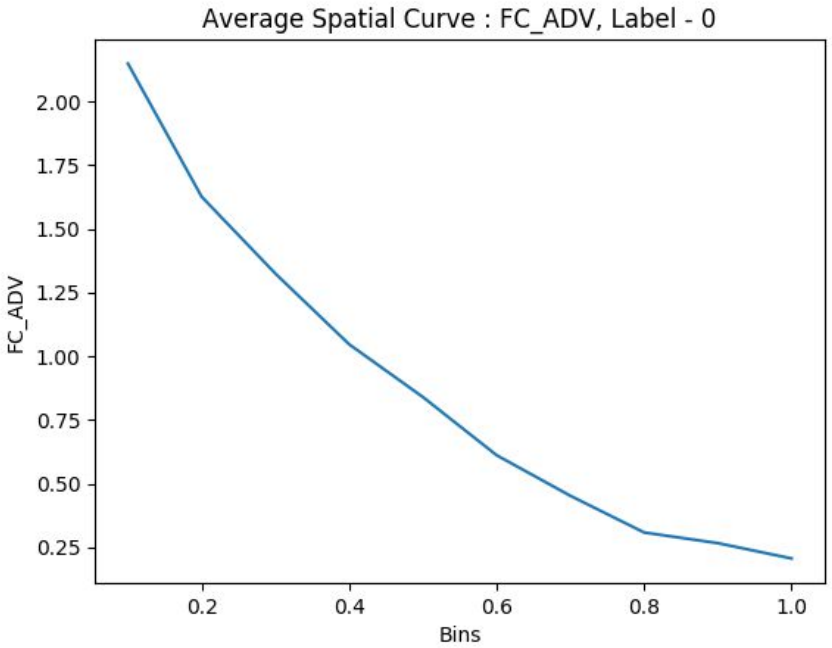

Average Spatial Curve : BF_ADV, Label - 1
BF_ADV
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.2
0.4
0.6
0.8
1.0
Bins

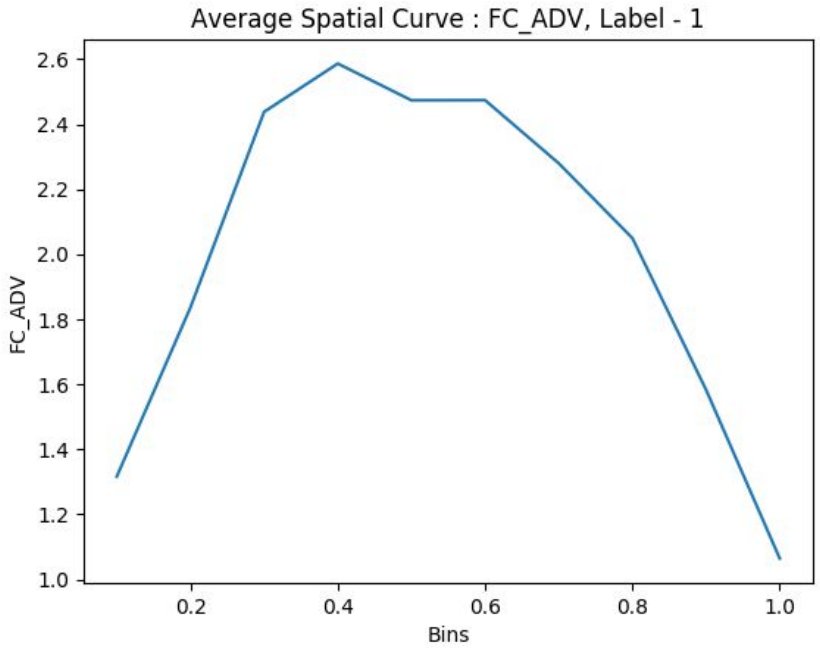
BF_INT


Average Spatial Curve : BF_INT, Label - 0
BF_INT
Bins
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.2
0.4
0.6
0.8
1.0

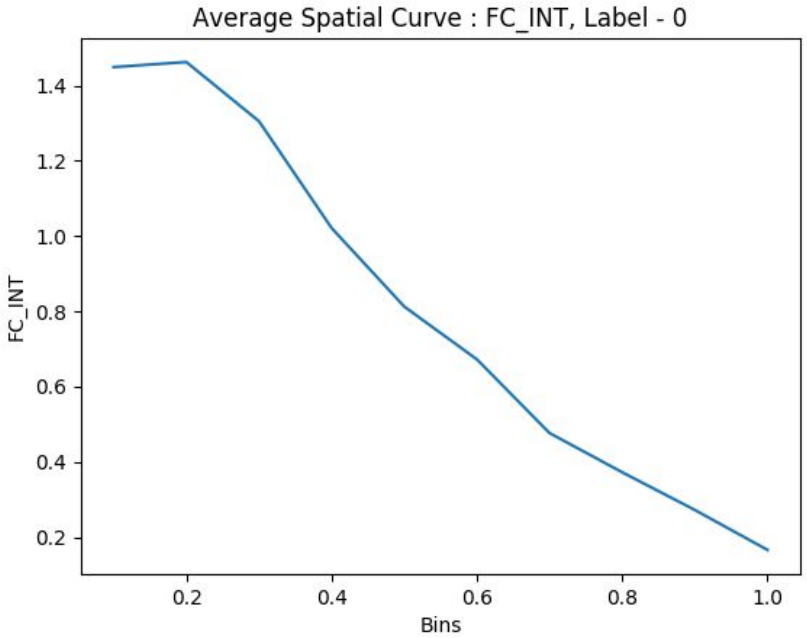

Average Spatial Curve : BF_INT, Label - 1
BF_INT
Bins
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0

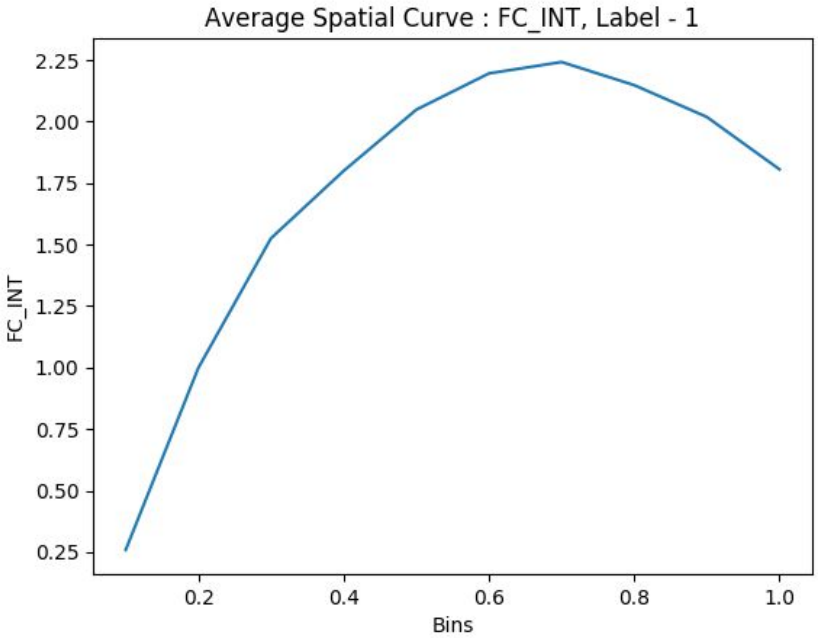
FC_ADV


Average Spatial Curve : FC_ADV, Label - 0
FC_ADV
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.2
0.4
0.6
0.8
1.0
Bins

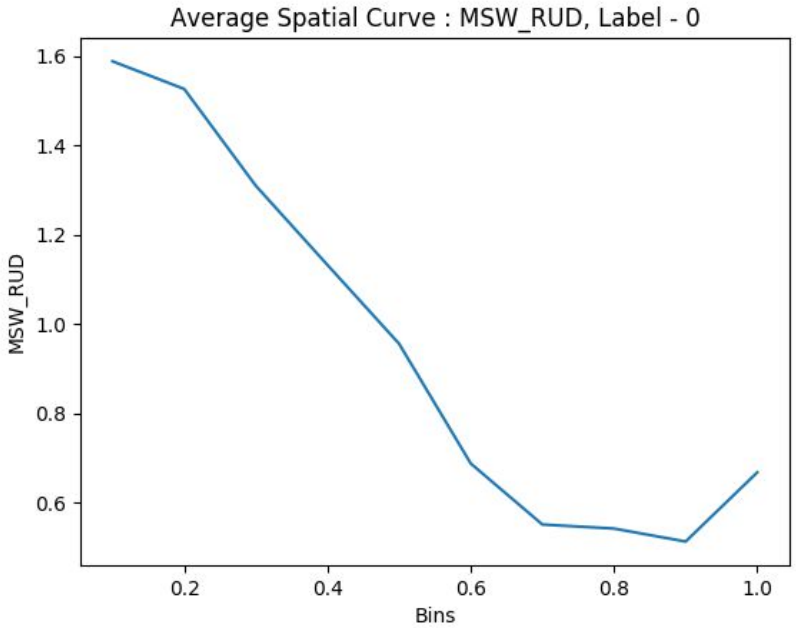

Average Spatial Curve : FC_ADV, Label - 1
FC_ADV
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.2
0.4
0.6
0.8
1.0
Bins

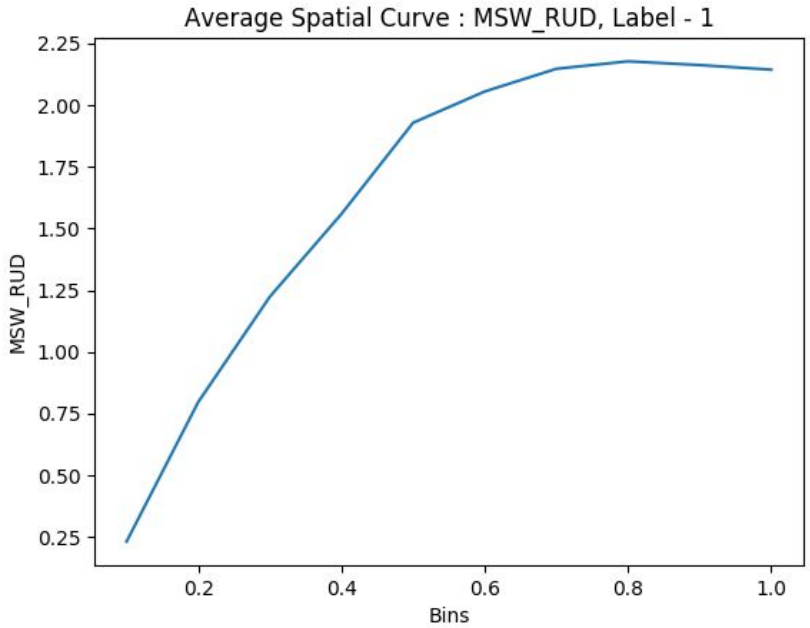
FC_INT


Average Spatial Curve : FC_INT, Label - 0
FC_INT
Bins
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0


Average Spatial Curve : FC_INT, Label - 1
FC_INT
Bins
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.2
0.4
0.6
0.8
1.0

MSW_RUD


Average Spatial Curve : MSW_RUD, Label - 0
MSW_RUD
Bins
1.6
1.4
1.2
1.0
0.8
0.6
0.2
0.4
0.6
0.8
1.0


Average Spatial Curve : MSW_RUD, Label - 1
MSW_RUD
Bins
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.2
0.4
0.6
0.8
1.0

Summary

- 2 Clusters explains the data the most - according to silhouette analysis
- With two clusters, the algorithm is generally differentiating between districts where the spatial parameter is increasing and the districts where the spatial parameter is decreasing
- But not always, for certain clusters, there is a rise and then a fall/a fall and then a rise.