F by Min for Clusters

Pseudo f-statistic and stability values: Stability is the solution that resulted in the same final segmentation for all different starting points or iterations. (Same ending point with different starting points then it will be trustworthy)

Usually we know that as the number of clusters increases, the pseudo F-statistic also increases.

If the pseudo F-statistic is decreased with the increase of the number of clusters, then it indicates that we probably asked for too many segments because we cannot continue to get separation of data points with more number of segments.

In our case, if we see the number of clusters increases from 6, the pseudo F-statistic value is decreased for cluster 7 and cluster 8. This implies that for cluster 7 and cluster 8 doesn't have good discrimination. Therefore we eliminate cluster 7 and cluster 8 solution as they have inferior discrimination.

Cluster 4 have 3 of its 10 iterations at one point and 7 of its iterations at another point. Cluster 5 also has 3 of its 10 iterations at one point and 7 of its iterations at another point. Cluster 6 has the same ending point for all iterations. If we compare the cluster 4 and cluster 5, cluster 5 is grouped closer when compared with the cluster 4. Average of pseudo f-statistic for cluster 4 is less than Average of pseudo f-statistic for cluster 5. This implies that cluster 4 have inferior discrimination and inferior stability when compared with cluster 5 and 6. Therefore we eliminate cluster 4.

Psuedo F-statistic and Stability comparision for cluster 5 and cluster 6:

minimum	PseudoF	Iter	Clusters
1,691	6,527.45	501	5
1,564	7,015.02	502	5
1,691	6,527.45	503	5
1,566	7,015.02	504	5
1,575	7,015.01	505	5
1,564	7,014.96	506	5
1,564	7,015.02	507	5
1,564	7,015.02	508	5
1,564	7,015.02	509	5
1,691	6,527.45	510	5
1,459	7,045.33	601	6
1,460	7,045.33	602	6
1,460	7,045.33	603	6
1,457	7,045.33	604	6
1,457	7,045.33	605	6
1,460	7,045.33	606	6
1,460	7,045.33	607	6
1,457	7,045.33	608	6
1,460	7,045.33	609	6
1,457	7,045.33	610	6

Average pseudo F-statistic value for cluster 5 is 6868.74. From the above table, we can see that for 3 iterations pseudo f-statistic value is around 6527 and for 7 iterations it is around 7015

Average pseudo f-statistic value for cluster 6 is 7045.33. For all the 10 iterations, the pseudo f-statistic value is same.

Since the pseudo f-statistic value is higher for cluster 6, we can say that cluster 6 has good discrimination. As the solution for all different iterations resulted in the same ending point for cluster 6, we can say that cluster 6 has more stability. Therefore cluster 6 has good discrimination and stability.

Size of smallest segment comparison for cluster 5 and cluster 6:

minimum	PseudoF	Iter	Clusters
1,691	6,527.45	501	5
1,564	7,015.02	502	5
1,691	6,527.45	503	5
1,566	7,015.02	504	5
1,575	7,015.01	505	5
1,564	7,014.96	506	5
1,564	7,015.02	507	5
1,564	7,015.02	508	5
1,564	7,015.02	509	5
1,691	6,527.45	510	5
1,459	7,045.33	601	6
1,460	7,045.33	602	6
1,460	7,045.33	603	6
1,457	7,045.33	604	6
1,457	7,045.33	605	6
1,460	7,045.33	606	6
1,460	7,045.33	607	6
1,457	7,045.33	608	6
1,460	7,045.33	609	6
1,457	7,045.33	610	6

For cluster 5 solution, largest value of size of smallest segment is 1691 and smallest value of size of smallest segment is 1564. Therefore maximum difference in the number of zipcodes in the smallest cluster is 127. This implies that from one iteration to other, 127 zipcodes are changing membership in and out of smallest cluster.

For cluster 6 solution, largest value of size of smallest segment is 1457 and smallest value of size of smallest segment is 1460. Therefore maximum difference in the number of zipcodes in the smallest cluster is 3. This implies that from one iteration to other, only 3 zipcodes are changing membership in and out of smallest cluster.

Since less number of zipcodes are changing membership for cluster 6 solution when compared to the cluster 5 solution, we can say that cluster 6 is the superior solution.

Summary:

Based on discrimination (pseudo F-statistic value), stability (number of times the iteration has the same ending point) and difference in the size of the smallest cluster, cluster 6 solution is superior to the cluster 5 solution. Therefore cluster 6 is the correct number of clusters that need to be selected.