Cluster Analysis for Dail Voting Data

The data under investigation records the voting behaviour of 166 deputies in the Dail during 23 votes since the start of 2016. The aim of the investigation is to examine whether the politicans fall into clusters, and to explore and quantify possible voting patterns within the clusters. The data was converted into a binary format where '1' denotes a 'yes' and where 'abstain' and 'no' are denoted by '0'. The choice of this approach puts the emphasis on 'yes' votes, and downplays the relative importance of abstentions – a bias which is worth keeping in mind.

The initial analysis showed that the number of 'yes' responses for each vote varied from a low of 17 (out of 166 deputies) for votes #7 and #20, through to a high of 82 for vote #5. The mean, therefore, ranged from 0.1024 to 0.494. The individual deputies voted 'yes' an average of 5.68 times, and their voting behaviour in terms of their number of 'yes' votes, is shown in **table 1** below. It is noticable that 16 deputies never voted 'yes' in the time frame concerned; conversely, 1 deputy voted 'yes' in 14 of the 23 votes. A naïve clustering, based on the table, suggests one group of 58 deputies who voted 'yes' four times or less, a second group of 89 who voted 'yes' between 5 and 9 times, and a final, smaller group of 19 who voted 'yes' on ten or more ocassions.

#'yes' votes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# Deputies	16	13	11	11	7	15	21	14	20	19	5	4	7	2	1

Table 1 – Voting behavior of the 166 deputies by number of 'yes' votes

The data were then clustered using k-means clustering. Initially, two clusters were defined. This was based on 20 random starts to reduce convergence to a local minimum. It resulted in one cluster of 89 deputies and another of 77. There were large disparities within the cluster means for each of the 23 votes. The between sum of squares / total sum of squares came to 35.6%.

By plotting the within cluster sum of squares values against K, there is a significant fall in the former between 2 and 3 clusters and a smaller drop between 3 and 4 clusters - suggesting that 3 is an appropriate number of clusters.

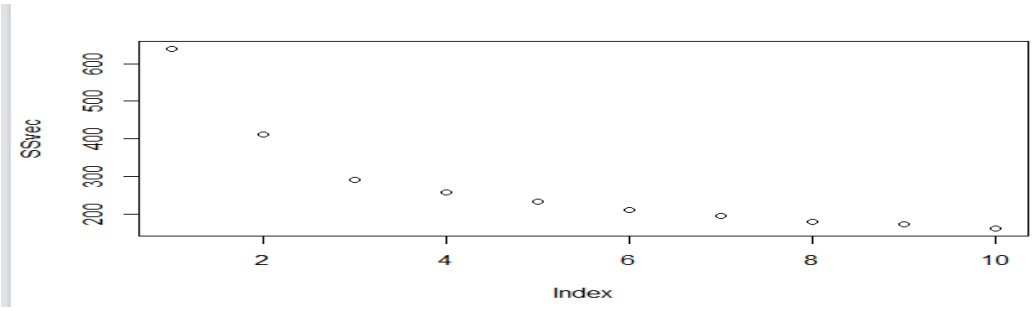

Figure 1 – Within cluster sum of squares for each value of K up to 10.

With 3 clusters, the between sum of squares / total sum of squares is now 53.4%. The first cluster of 61 deputies has a large number of 0's (14) combined with a significant number of means > 0.8 (4). This suggests a trend towards not voting or abstaining much of the time, combined with high participation in votes # 4,5,21 and 22. A second cluster of 75 deputies has a low number of 0's (2) combined with few means > 0.8 (2). The exception to this is their high participation in votes # 1,2 and 14. A final group of 30

deputies combine a degree of abstention / no votes (6) with high participation in a number of votes, such as votes # 12,13 and 14. The mean is > 0.8 for 7 votes.

```
Cluster means:
       [,1]      [,2]      [,3]     [,4]      [,5]      [,6]       [,7]       [,8]       [,9]      [,10]     [,11]
1 0.70666667 0.68000000 0.0000000 0.960000 0.9600000 0.5733333 0.00000000 0.00000000 0.00000000 0.00000000 0.0000000
2 0.09836066 0.09836066 0.1639344 0.147541 0.1639344 0.2786885 0.06557377 0.01639344 0.01639344 0.01639344 0.0000000
3 0.00000000 0.00000000 0.5000000 0.000000 0.0000000 0.1333333 0.43333333 0.83333333 0.83333333 0.80000000 0.8666667
       [,12]      [,13]      [,14]     [,15]     [,16]     [,17]      [,18]      [,19]      [,20]     [,21]
1 0.00000000 0.00000000 0.00000000 0.6400000 0.0000000 0.7733333 0.00000000 0.00000000 0.00000000 0.8533333
2 0.01639344 0.04918033 0.09836066 0.1803279 0.1311475 0.1475410 0.01639344 0.03278689 0.01639344 0.0000000
3 0.93333333 0.93333333 0.90000000 0.3000000 0.7333333 0.1666667 0.30000000 0.53333333 0.53333333 0.0000000
       [,22]    [,23]
1 0.82666667 0.000000
2 0.01639344 0.147541
3 0.00000000 0.400000
```

Figure 2 – Cluster means for each of the 23 votes (k=3, nstarts=20)

The silhouette plot of the data, based on the squared Euclidean distance, shows that clusters 1 and 2 have significant (but not high) average silhouette (>0.5). Both have long 'tails', where observations are similar to observations from another cluster. Cluster 3 has a below average silhouette, indicating weaker clustering. There are two observations with negative silhouette, one in each of clusters 1 and 3, indicating that the obseration should really be in another cluster.

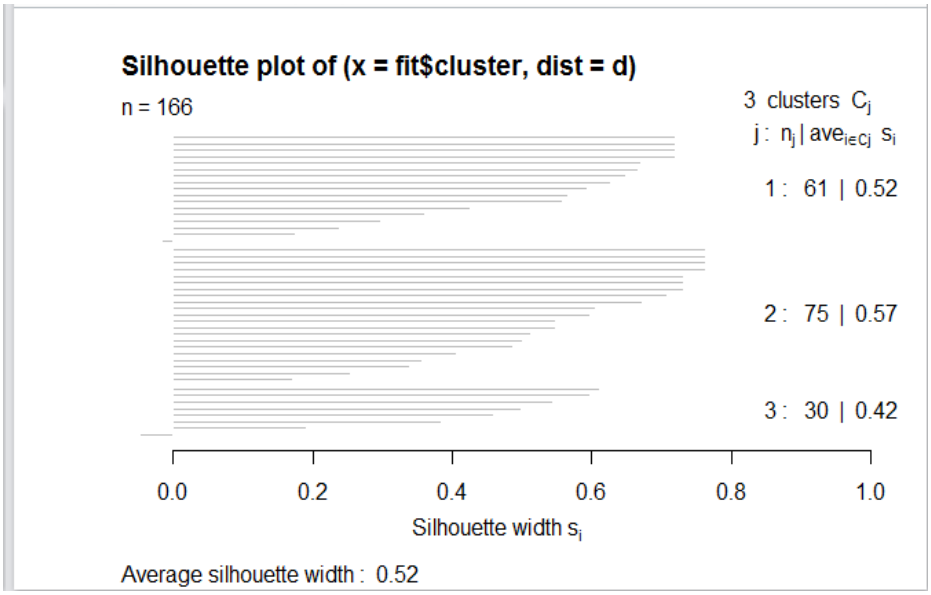

Figure 3 – Silhouette plot of k-means solution

The data was also analysed using the k-medoids algorithm with binary distance, and k=3. The results of the two approaches are shown in **figure 4**. From this we can see that k-medoid cluster # 1 is similar to k-means clusters # 2 and 3. K-medoid cluster #2 is similar to k-means clusters #1 and 2, and k-medoid cluster 3 is similar to k-means cluster 2. Finally, there is a Rand Index of 0.736 suggesting a high degree of agreement between the two algorithms. The Rand Index, corrected for agreement by chance, ('crand' in the e1071 package in r) is somewhat lower at 0.459.

```
    1  2  3
1   0 75  0
2  22 23 16
3  30  0  0
>
```

Figure 4 – Comparison of k-means and k-medoids findings