

MACHINE LEARNING PROJECT 2: UNSUPERVISED ANALYSIS OF THE HUMAN DEVELOPMENT REPORT¹

M-SECUC, INFOY-112: MACHINE LEARNING

Muranovic Allan

¹ *Université de Namur*

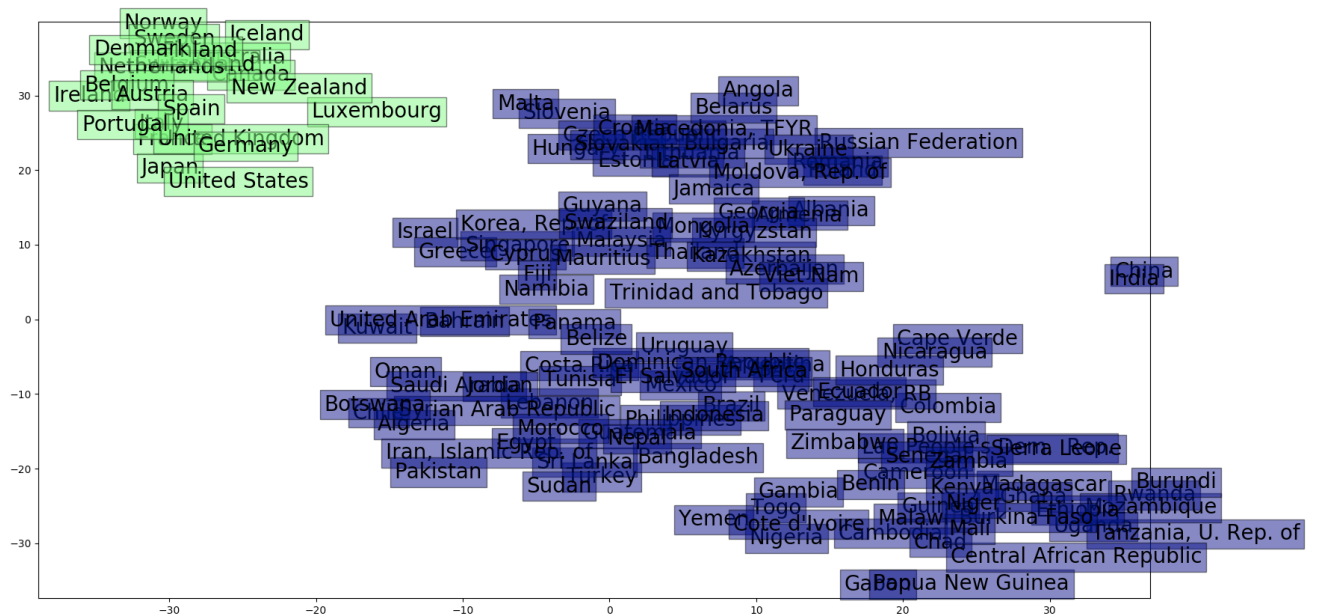
1 TASK 1 : 2 CLUSTERS

Clustering which does **not** contain enough clusters.

1.1 Most centered countries.

1. El Salvador
2. United Kingdom

1.2 *t*-SNE visualization.



1.3 Discussing the choice of 2 clusters.

We clearly see the separation into 2 clusters with disproportionate groups. Two clusters is therefore a good example of a choice of too few clusters to extract relevant information.

1.4 Interpretation of the clusters.

By containing only two clusters, we see 2 groups of different sizes. In the largest of the two groups, many countries are far from their El Salvador centre. Which indicates a bad cluster.

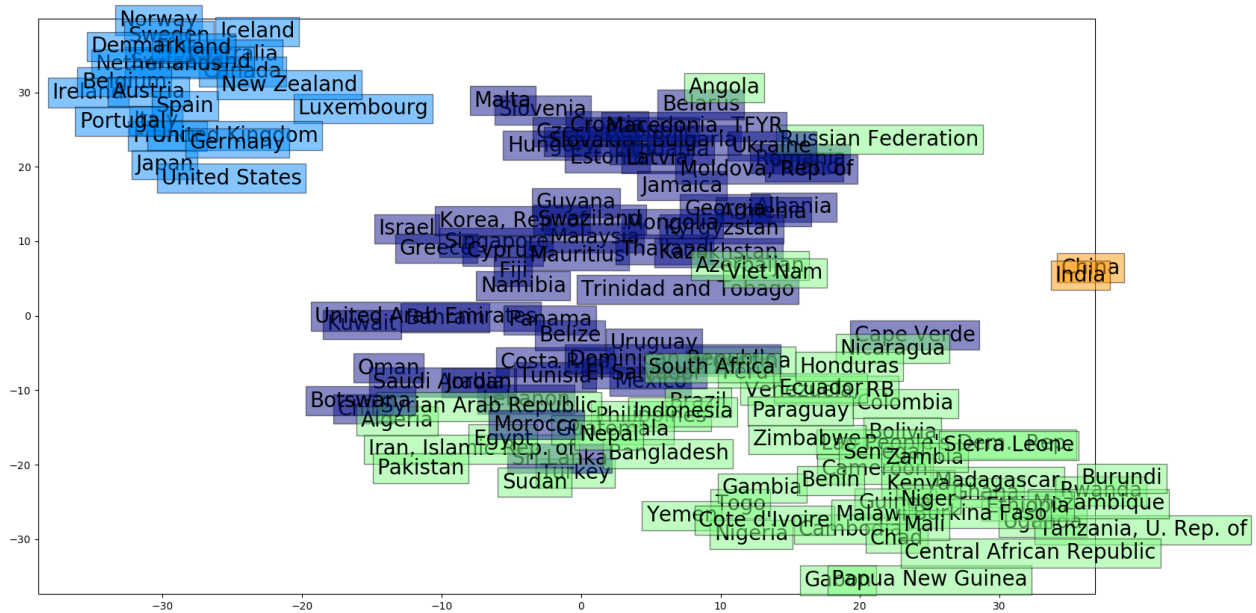
2 TASK 2 : 4 CLUSTERS

Clustering which **does** contain enough clusters.

2.1 Most centered countries.

1. Cyprus
2. United Kingdom
3. Senegal
4. China

2.2 *t*-SNE visualization.



2.3 Discussing the choice of 4 clusters.

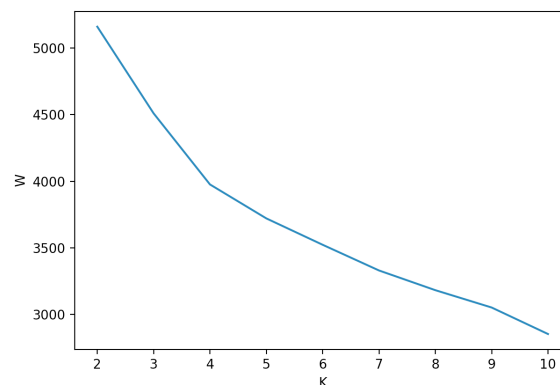


Figure 1: W to K evolution

K being the number of clusters (from 2 to 10), we observe on the graph above that once $K = 4$ passed, the other classes do not gain significant information. $K = 4$ is the last to gain so much. This is why the number of 4 clusters was chosen, it gives us enough information on the clustering data.

2.4 Interpretation of the clusters.

The largest of the groups in the figure in point 1.2 has been divided into 3 smaller groups. As for the smallest one, he hasn't been touched. We therefore have 4 clusters well distributed without countries too far from the centre.

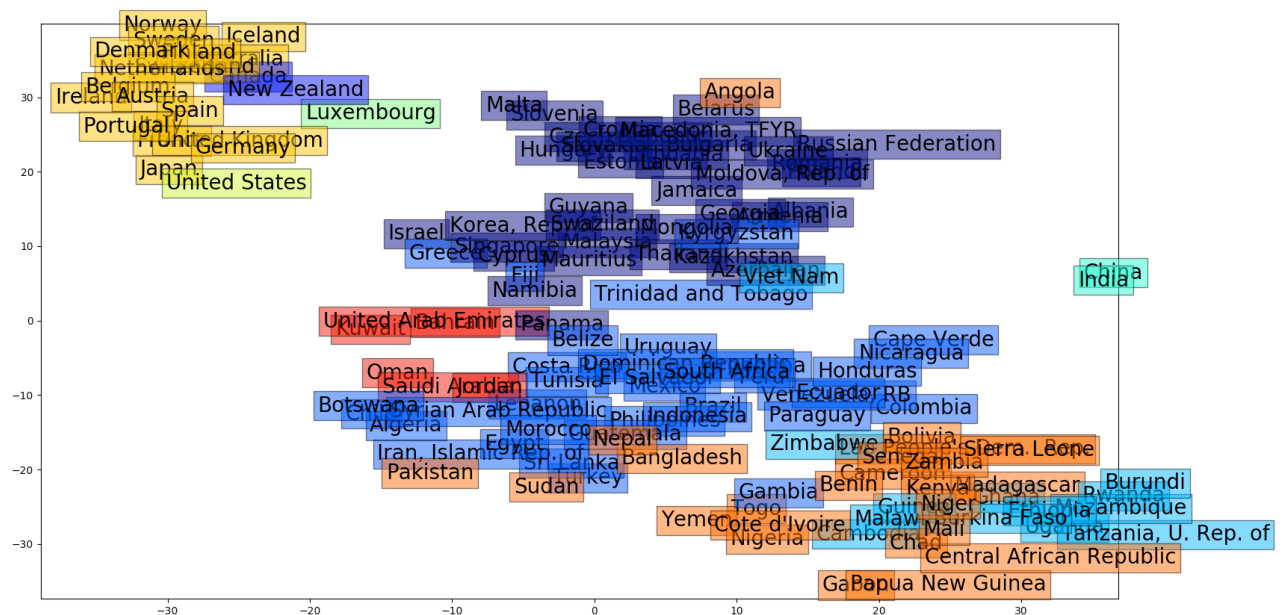
3 TASK 3 : 10 CLUSTERS

Clustering which does contain **too many** clusters.

3.1 Most centered countries.

1. Latvia
2. Canada
3. El Salvador
4. Guinea
5. China
6. Luxembourg
7. United States
8. United Kingdom
9. Senegal
10. Saudia Arabia

3.2 *t-SNE* visualization.



3.3 Discussing the choice of 10 clusters.

As we can see, there are 10 groups, including some containing only one country: the United States, Luxembourg. This is problematic, as the purpose of clustering is to form groups with the same characteristics and not to isolate certain elements. This indicates that the number of clusters is too high.

3.4 Interpretation of the clusters.

We note that the Orange Grouping has countries far from its centre, Senegal, which indicates relatively large differences between them. The two light blue groups, dark blue and orange, have the same problem. For the other groups, the distances to the centres are appropriate.