CLUSTER ANALYSIS

UNIVERSITÀ
DEGLI STUDI
DI MILANO

STUDENT NAME: SAINEY MANGA
MAT: 943874
PROGRAM: MASTER'S IN DATA SCIENCE AND ECONOMICS
COURSE CORDINATOR: PROFESSOR GIANCARLO MANZI
COURSE: ADVANCE MULTIVARIATE STATISTICS

CLUSTER ANALYSIS

ABSTRACT:

Cluster analysis employs various algorithms to group similar groups of observations into different clusters using different approaches. It is the process of assigning data objects into similar groups (clusters), usually (but not always!) according to a defined distance measure (or proximity measure). (Lecture notes, Prof. Giancarlo).

In this project, we try to compare two methods of clustering namely K-means and Model based clustering. This is to allow us to sense the difference in the performance of the two methods on the same dataset.

The goal of this project is to identify and group amongst 167, countries that are poor and "needy" into clusters. The assumption is that world international organizations want to identify countries that might need help in policy formulation and possibly financial aid in a bit to remedy the long-term impact of the Covid-19 pandemic. We tried to relax the data description part till the next section.

We first begin with the K-means clustering method where we applied distinctively; the elbow method, the gap statistics and the silhouette method in order to find the optimal number of clusters to choose on our final cluster diagram. The reason is to explore the different number of optimal clusters generated by applying the different methods highlighted above. Both the elbow and the silhouette methods suggest a 5 optimal cluster as oppose to the gap statistics method which gave an output of 3 optimal clusters. In the end, we would go by the output of the common optimal clusters as we would later see that the results of the cluster diagram using five optimal clusters looks more appealing and perhaps more interpretable. At this point, we would relax the mathematical intuition behind the different methods applied until later in the next chapters as we would see why the different methods give different number of optimal clusters.

On the k-means cluster diagram, we have the 5 clusters as chosen before. We see that there is only one observation (Nigeria) in cluster 2, indicating that it is different from all other countries or perhaps could be an outlier, we would dilate more on that later. We have only 3 countries in cluster 3 where Malta, Luxemburg and Singapore are found. The likes of USA, Ireland and Qatar are in cluster 4 whilst the likes of Brazil, Argentina and Malaysia in cluster 1, and the likes of Angola, Afghanistan and Haiti in cluster 5 respectively. In a bit to make sense out of our clusters, we would construct a Pacoplot at the end of the project where we would interpret what our clusters really mean.

The second half of the project focuses on the algorithms of the Model based clustering. Unlike the k-means clustering, we do not have to determine the optimal number of clusters at the initial stage. We fit our data using the Mclust package in R, which automatically selects the best model and the optimal number of clusters to represent our data. Three clusters were given by the model selected through the BIC method. We will see that each datapoint is assigned a probability of belonging to the given clusters. Cluster 1 has 50 countries with the likes of Nigeria, Afghanistan, Angola and Chad. Cluster 2 has 74 countries with likes of Botswana, Egypt, Equatorial Guinea and China whilst cluster 3 has 43 countries with the likes of USA, UAE, Luxemburg and Brunei.

CLUSTER ANALYSIS

DATA DESCRIPTION:

The dataset named country-data is downloaded from https://www.kaggle.com/santhraul/country-data. The dataset consists of 167 observations representing 167 countries around the world, and 10 columns representing socio-economic and health attributes in those countries. The attributes are child-mortality, exports, health, imports, income, inflation, life-expectancy, total fertility and GDP. The column for country makes it 10, but we would make it a mere observation so we can handle the data for clustering. We are now going group countries into clusters based on the variables highlighted above.

MAIN RESULTS

K-MEANS CLUSTERING:

The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized. The standard algorithm is the Hartigan-Wong algorithm (Hartigan and Wong 1979), which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid:

$$W(Ck)=\sum xi \in Ck(xi-\mu k)^2$$

- *xi design a data point belonging to the cluster Ck*
- *μk is the mean value of the points assigned to the cluster Ck*

Each observation (xi) is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers μk is a minimum.

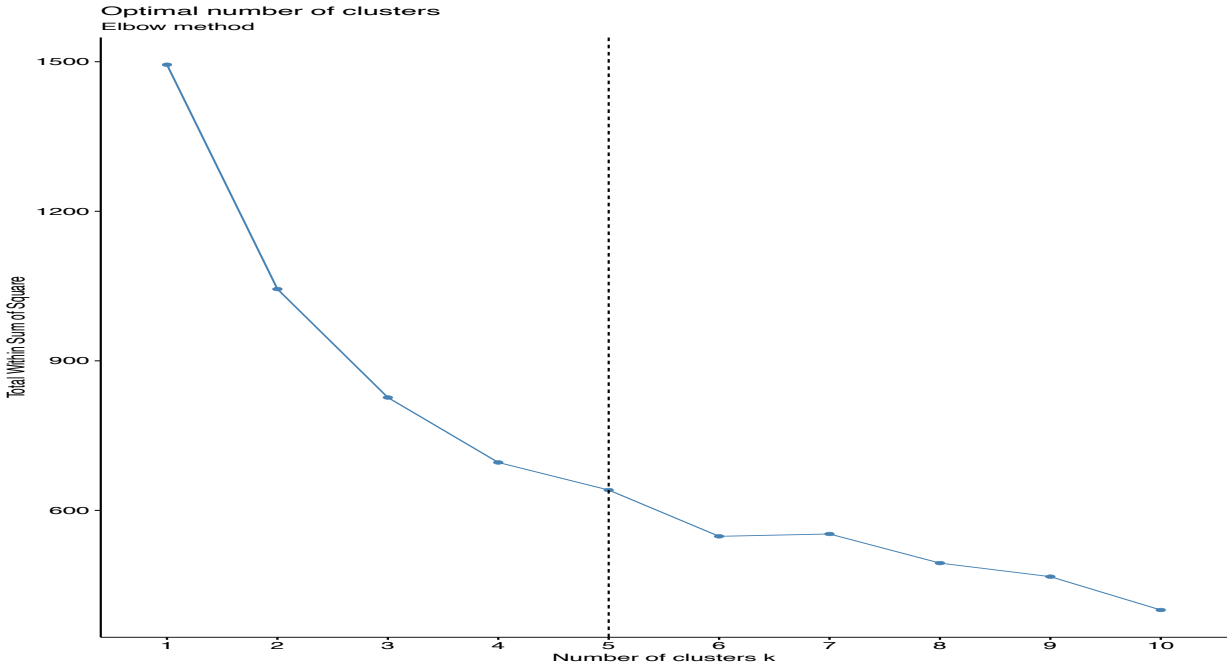
We define the total within-cluster variation as follow:

$$tot.withinss=\sum k=1kW(Ck)=\sum k=1k\sum xi\in Ck(xi-\mu k)^2$$

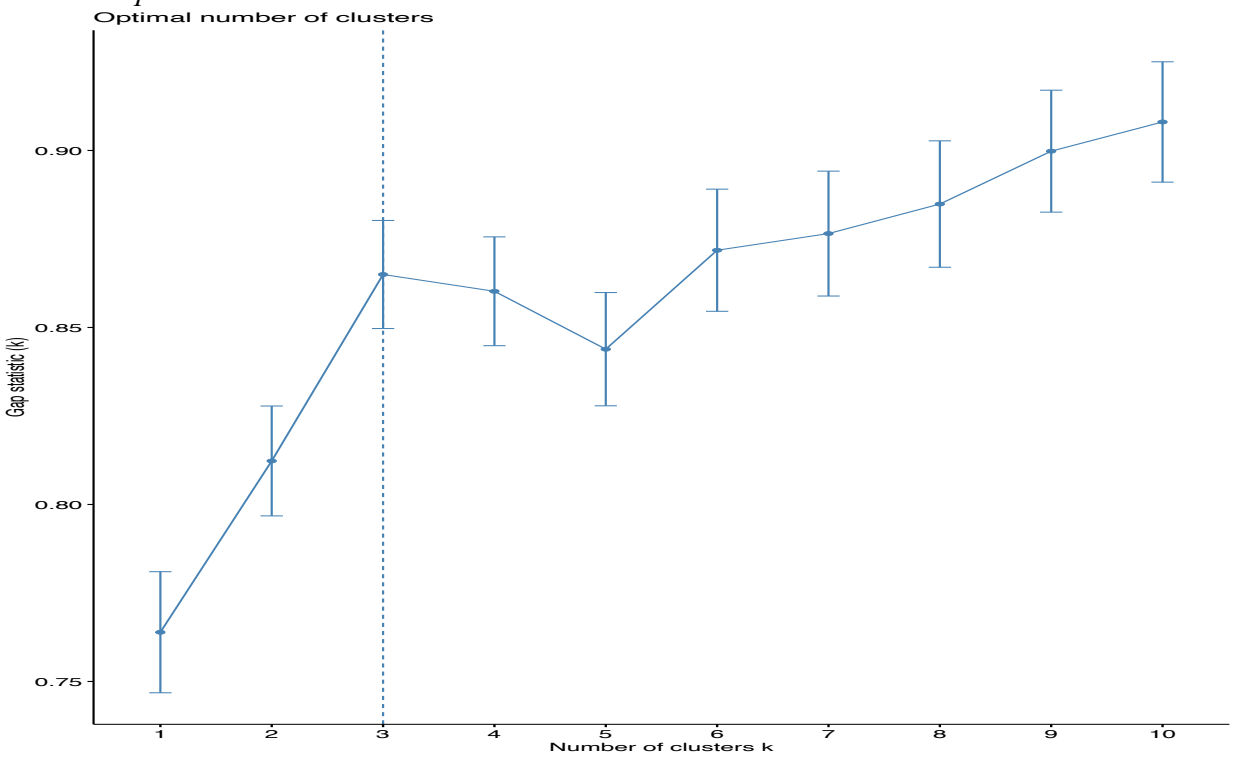
From here, we would move on to our results as stipulated before by identifying the optimal number of clusters using the three methods mention before.

i. *Elbow Method:*
   On the below diagram, we consider the optimal number of clusters as the traditional number at the bend knee or elbow which in our case is at cluster number 5. Therefore, our k=5. Indeed, one could argue choosing 4 as the optimal number of clusters or 3 as the optimal number of clusters, but in the end the cluster diagram would basically tell us a most preferred number of optimal number of clusters.

CLUSTER ANALYSIS



ii. The Gap statistic:

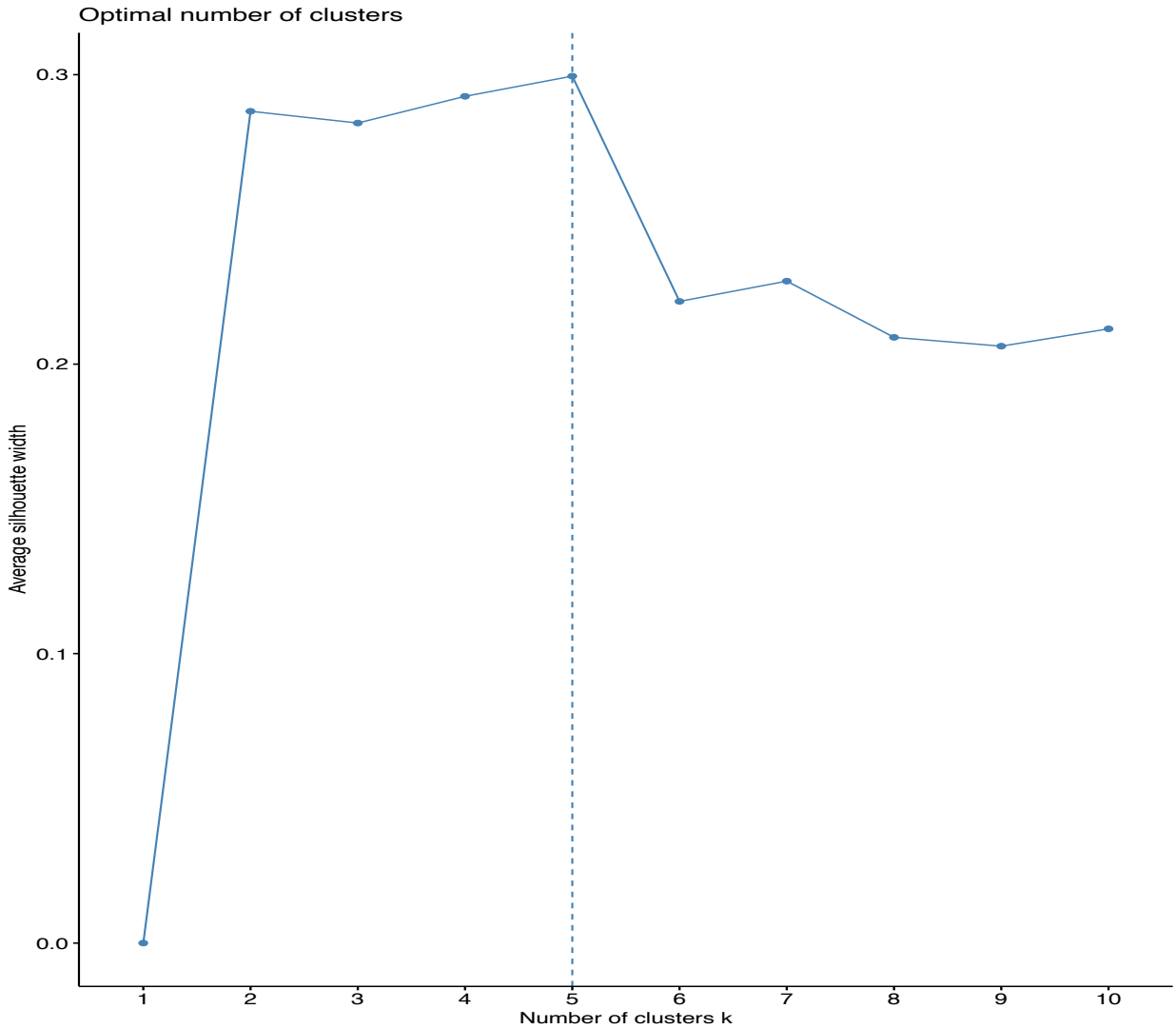


Although we applied a maximum of 10 optimal number of clusters on our R-code, but the results of the diagram still cuts at 3 as our optimal number of clusters, this is because the "gap stat()" function in R automatically recommends the optimal number of clusters from a given range of k>=2 clusters. We know that the gap statistic compares the total intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The gap statistic for a given k is given:

Gap(K) = log W (K) − log W_{unif} (K). if the data were uniformly distributed.

iii. The Silhouette method:

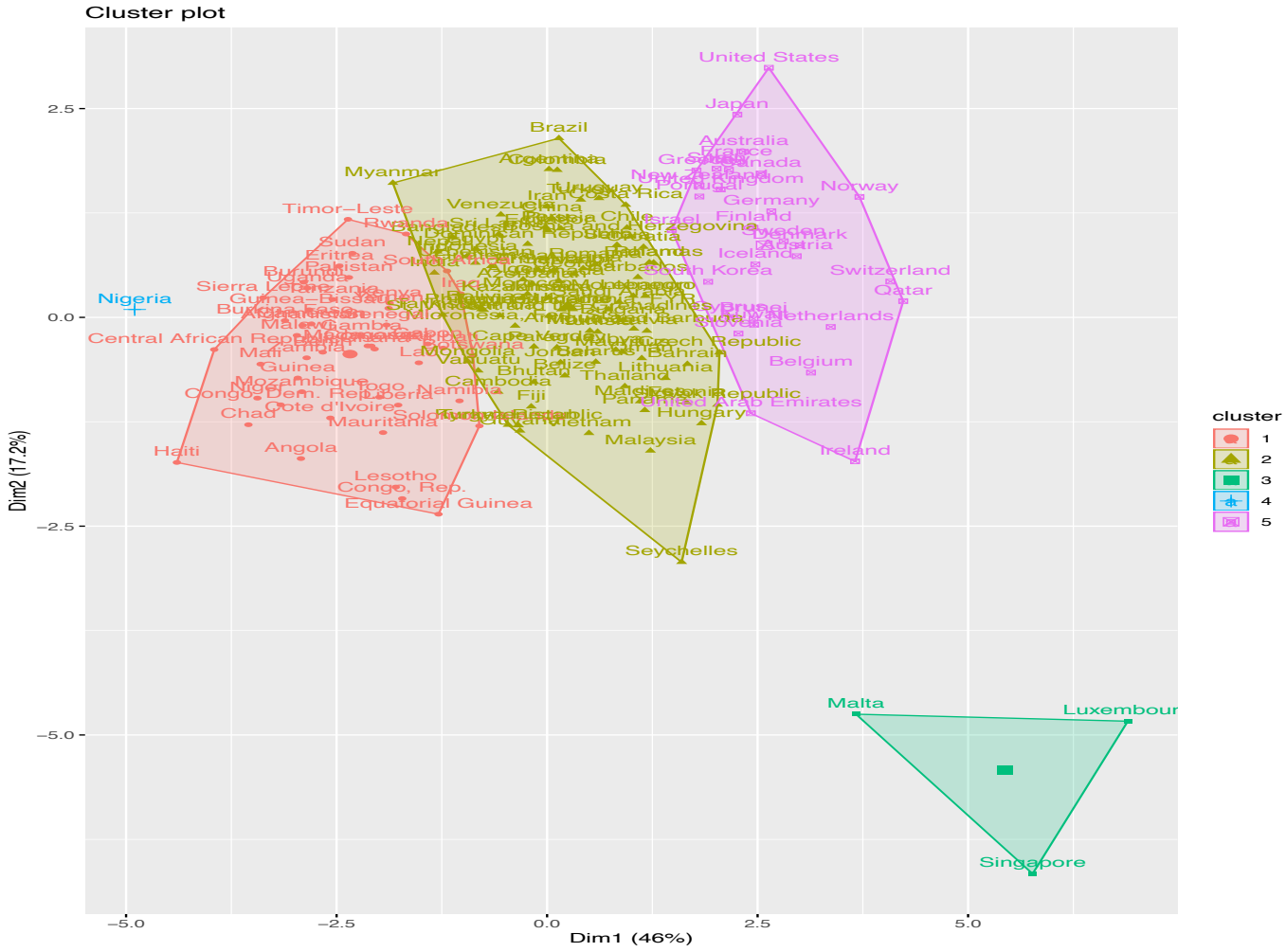
The average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k. Our below Silhouette diagram proposes 5 number of optimal clusters like our elbow method.



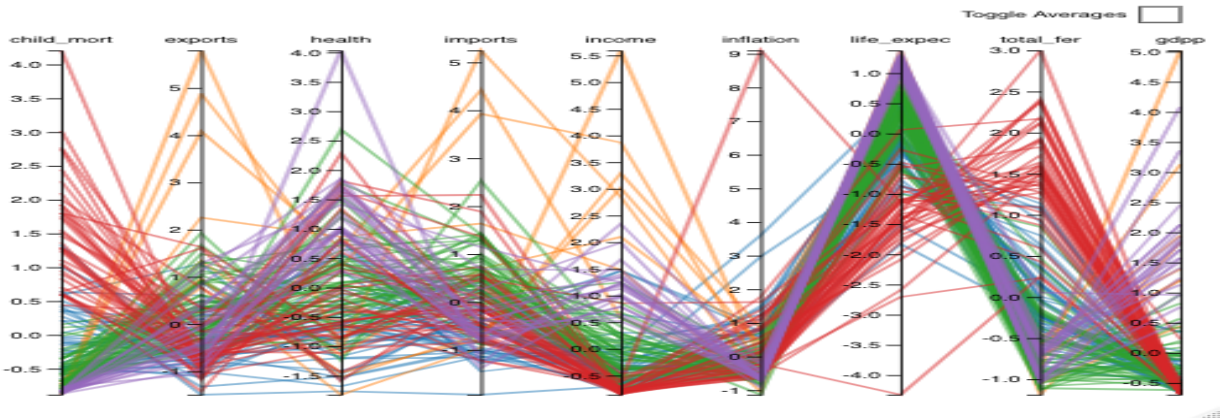
Now that we have diagnosed all the three methods in order to determine the number of optimal clusters we would use to plot our final cluster diagram, we would go by k=5 clusters as commonly proposed by both the elbow and silhouette method. Although Nigeria which seems to be an outlier when we use k=5 clusters, joins the likes of Haiti, Angola etc. when we use k=3 clusters. One could argue that k=5 could be more realistic as Nigeria is more similar with Angola and the likes in terms of their economic variable trends. K=3 clusters also includes USA and the likes of Singapore

in the same cluster. However, it could also be appealing to have Nigeria stand alone as it is the only country with the uppermost inflation rate and total fertility.

The K-means cluster plot



To make sense from our clusters, we would plot a Pacoplot to show why countries are group into the clusters they are.



CLUSTER ANALYSIS

From the two plots, we can deduce that members of cluster 1 have high child mortality, a very low health coefficient, low exports, relative imports, lowest income, low life-expectancy, high total fertility and a lowest GDP. We can classify this cluster as underdeveloped or more euphemistically, developing countries or poor countries if you like. They are countries that would need more aid or help base on our assumption. Cluster 2 follows cluster 1 in our ranking, making it our class of developing or middle-class countries. Cluster 3 and 5 are our developed countries as they are among countries with high GDP, life-expectancy, income, health, and a low total fertility, imports and inflation. Cluster 3 is not any different with cluster 1 except that it has the highest inflation rate and no other countries followed closely.

MODEL BASED CLUSTERING:

Unlike k-means, the model-based clustering uses a soft assignment, where each data point has a probability of belonging to each cluster. The data is considered as coming from a mixture of density. Each cluster k is modeled by the normal or Gaussian distribution which is characterized by the parameters:

- *μkμk: mean vector,*
- *∑k∑k: covariance matrix,*
- *An associated probability in the mixture. Each point has a probability of belonging to each cluster.*

The model parameters can be estimated using the Expectation-Maximization (EM) algorithm initialized by hierarchical model-based clustering. Each cluster k is centered at the means μkμk, with increased density for points near the mean. The Mclust package uses maximum likelihood to fit all these models, with different covariance matrix parameterizations, for a range of k components. The best model is selected using the Bayesian Information Criterion or BIC. A large BIC score indicates strong evidence for the corresponding model.
We now use the Mclust package to fit our data with a suitable model and automatically identify our optimal number of clusters.

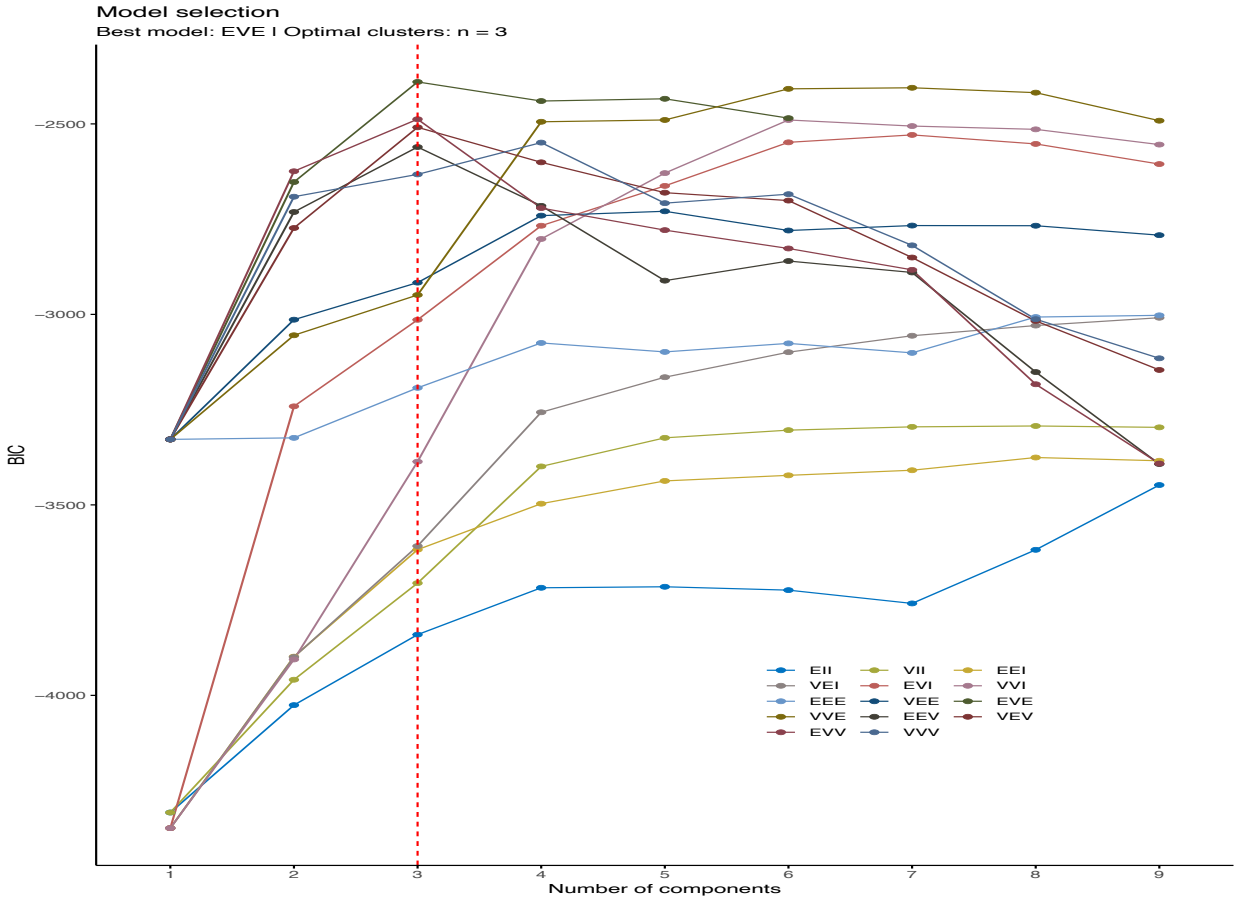
```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EVE (ellipsoidal, equal volume and orientation) model with 3 components:

 log-likelihood   n  df      BIC      ICL
      -964.6146 167  90 -2389.849 -2400.492

Clustering table:
 1  2  3
50 74 43
```

CLUSTER ANALYSIS

It can be seen that model-based clustering selected a model with three clusters. The optimal selected model name is EVE model. That is the three components are ellipsoidal with equal volume, shape, and orientation. The summary contains also the clustering table specifying the number of observations in each cluster where cluster 1 represents 50 countries, cluster 2 being the highest with 74 countries and cluster 3 with 43 countries.
Let us show the probabilities attached to at least each of the first 10 countries as we cannot show all the 167 countries.

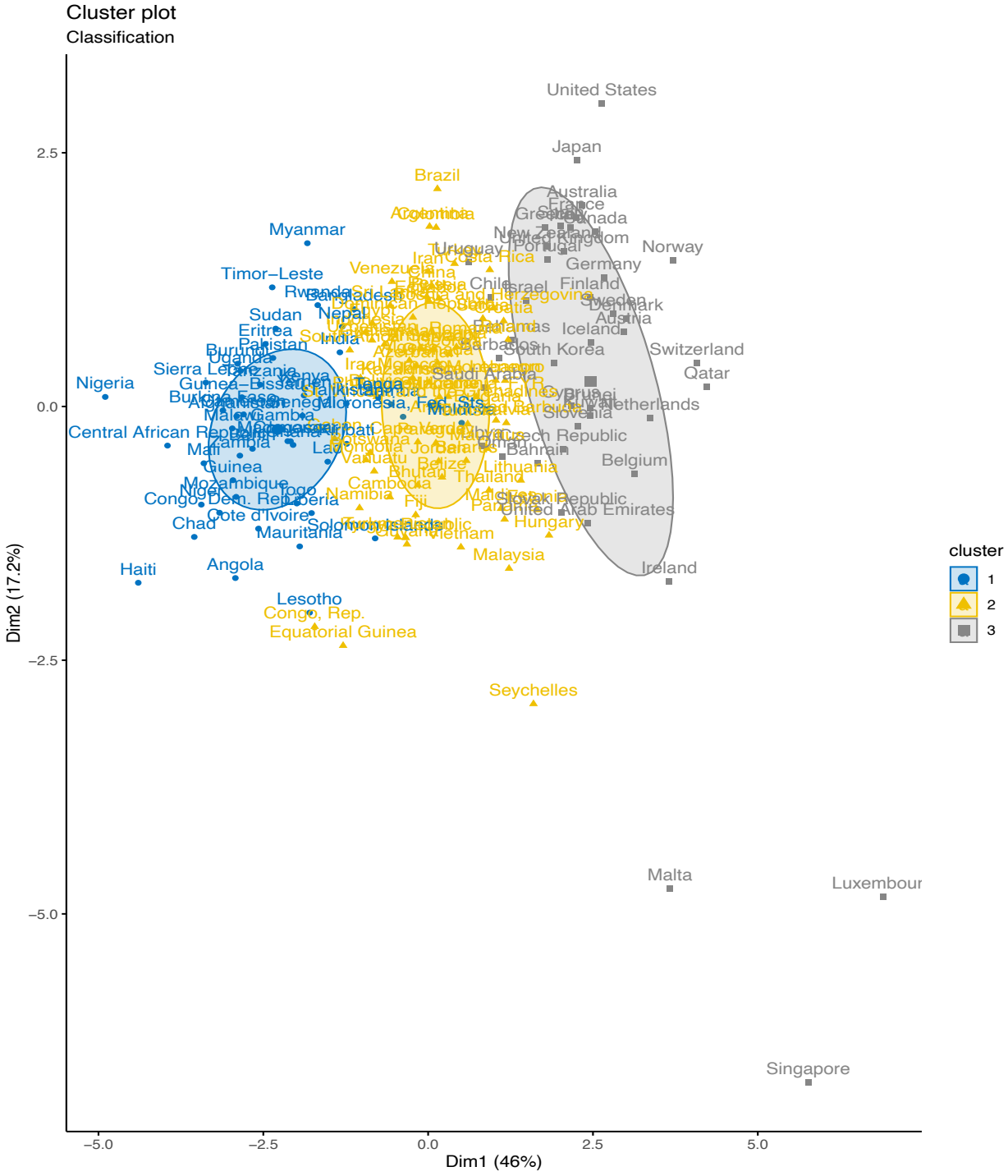
	[,1]	[,2]	[,3]
Afghanistan	9.999993e-01	7.297002e-07	3.730744e-88
Albania	6.123661e-09	9.999994e-01	6.122055e-07
Algeria	3.208510e-18	9.999647e-01	3.527618e-05
Angola	9.999998e-01	1.909461e-07	6.574346e-97
Antigua and Barbuda	2.344122e-41	9.997578e-01	2.422063e-04
Argentina	6.208698e-40	9.997110e-01	2.889617e-04
Armenia	3.144035e-02	9.685596e-01	1.056908e-10
Australia	0.000000e+00	1.293817e-83	1.000000e+00
Austria	0.000000e+00	9.376469e-56	1.000000e+00
Azerbaijan	7.005335e-33	1.000000e+00	8.673762e-13

We proceed to show the plot used by BIC values to choose the optimal number of clusters



CLUSTER ANALYSIS

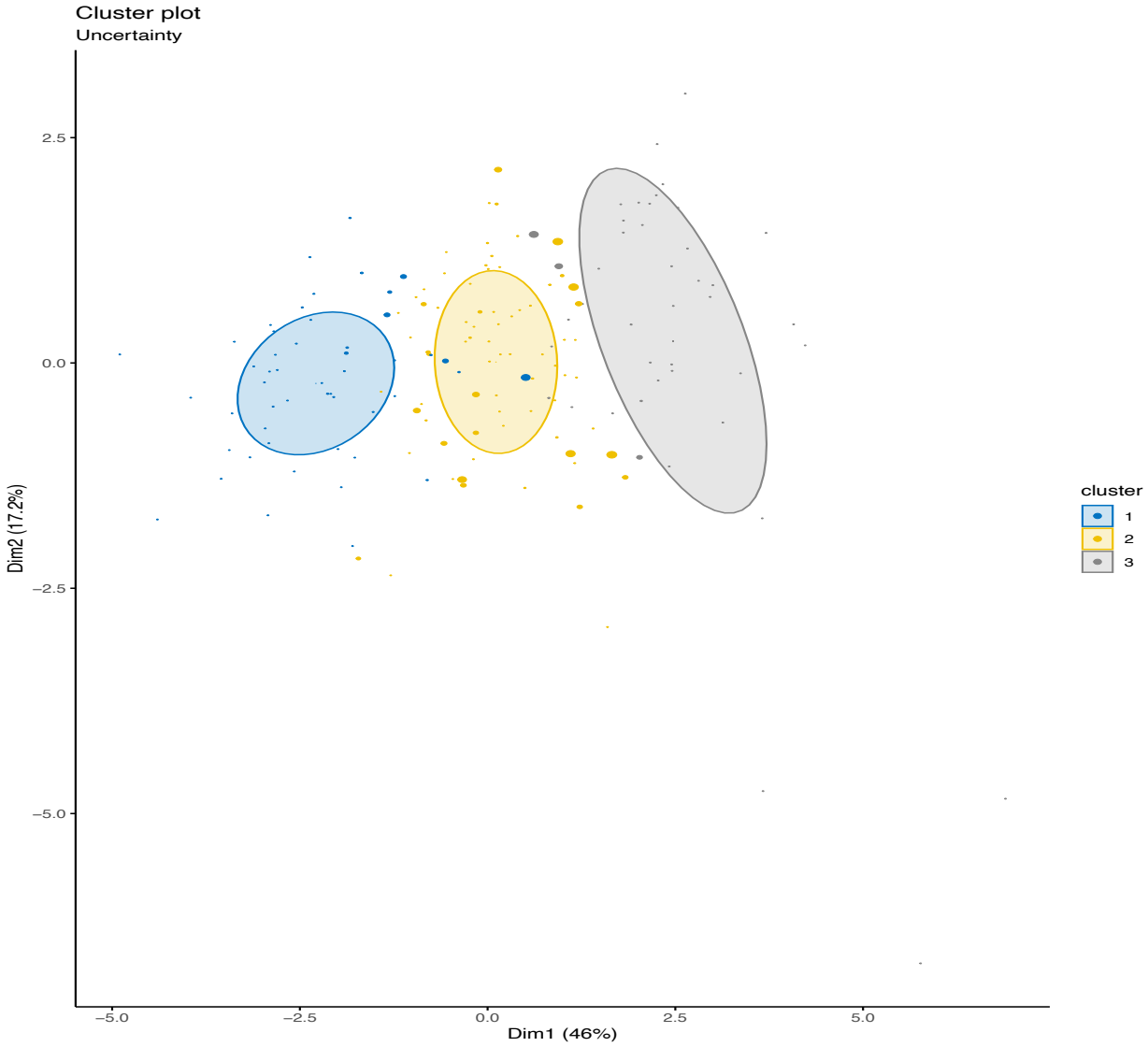
We have three optimal clusters generated from the diagram with the EVE being the best model chosen by the EM algorithm.



From the cluster plot above, cluster 1 includes the likes of Nigeria, Angola, Lesotho, Pakistan and so on. Cluster 2 includes Equatorial Guinea, Malaysia, Hungary, Brazil, Cambodia and so on. Finally, cluster 3 includes USA, Singapore, Malta, Norway, Qatar and so on. The beautiful thing

about model-based clustering is that each of these countries is assigned to its cluster with the highest probability.

The cluster below is generated based on uncertainty, the observations are assigned to clusters with uncertainty. The larger the symbols, the more uncertain the observations are as members of that cluster. This could even be noticed in the corresponding probabilities of each observation as some observation have very tiny differences in their probabilities of belonging to certain clusters. We refer to cluster 2 where we have a good number of observations with larger uncertainty of belonging to cluster 2.



CLUSTER ANALYSIS

CONCLUSION:

We have seen that the difference of results generated by our two methods is necessitated by our choice of optimal clusters in k-means clustering algorithm, otherwise, our results would have been much similar. The two methods would have produced somewhat similar results had we used the number of clusters recommended by the Gap statistic method. However, we know that cluster 3 in the model-based clustering has the same observation as cluster 3 and 5 of our k-means clustering. Cluster 2 and 1 of the model-based has similar observations as cluster 1, 2 and 4 of the k-means clustering, although some slight difference could be noticed between cluster 1 and 2 of the model-based with other counterparts. Some countries belong to the same cluster here but were in different clusters in the previous and vice-versa.

Reference:

Lecture notes, Prof Giancarlo Manzi, University of Milan.

https://www.datanovia.com/en/

An introduction to applied multivariate data analysis with R, **Brian S. Everitt,** *Institute* of *Psychiatry, King's College London, UK*