UNIVERSITÀ DEGLI STUDI DI MILANO

LA STATALE

Department of Economics, Management and Quantitative Methods

Università degli Studi di Milano

Data science and Economics

STATISTICAL LEARNING PROJECT

UNSUPERVISED TECHNIQUES ON FAILED STATES INDEX

REPORT

Author:

MURAT AYDIN

Academic year 2020-2021

ABSTRACT

In today's highly interconnected world, pressures on one fragile state can have serious results not only for that state and its people but also for its neighbors and other states. A very practical example comes from middle east, Syria. Syria is a failed state which has been experiencing civil war that has been going on since 2011. This had immense repercussions not only for all of its neighbors and for the continental Europe but also the whole globe itself. Therefore, understanding what leads to such situations is vitally important. There are some other practical examples from today's world in this context. However, here our aim is to understand what are the specific features related to fragile states. To understand better the concept of fragile states, we use the data provied by fragilestatesindex.org(FSI) organized by The Fund for Peace(FFP), consisting 12 indices of 178 countries related to socio-economic and political dimensions of a given state. After a descriptive analysis of the dataset, we use Principal Component Analysis to reduce the dimensionality of the data and gain some intuition about the performances of each country. After that, we run K-means and Hierarchical clustering algorithms to group together the countries characterized by similar level of fragile state index. With PCA analysis we saw that there is one strong direction where failed states are clustered that is over first and fourth quadrant. Witk K-means, we identified 5 clusters where the identification of these clusters resulted in a very clear representation while with Hierarchical Clustering, we got 3 clusters with complete linkage method and 5 clusters again with the ward method.

1.Data Definition

FSI provides official and reliable sources for measuring the fragility of states. More specifically, the organization provides 12 different indices, grouped into four broad categories of social- economic- political measurements. These indices are critical tools in highlighting not only the normal pressures that all states experience, but also in identifying when those pressures are outweighing a states' capacity to manage those pressures. The FSI scores should be interpreted with the understanding that the lower the score, the better. Therefore, a reduced score indicates an improvement and greater relative stability, just as a higher score indicates greater instability. Each index assigns a maximum of 10 as a score. They are described in the methodology section fragilestateindex.org as follows:

1.1 Cohesion indicators

- Security Apparatus

This indactor considers the secuirt threats to a state, such as bombings, attacks and battle-related deaths, terrorism etc. It also takes into account organized crime and homicies.

- Factionalized Elites

This indicator considers the fragmentation of state institution along ethnic, class, clan, racial or religious lines, as well as brinkmanshio and gridlock between ruling elites.

- Group Grievance

This indicator focuses on divisions and schisms between different groups in society- particularly divisions based on social or political characteristics- and their role is access to services or resources, and inclusion in the political process.

1.2 Economic Indicators

- Economic Decline and Poverty(Economy)

This indicator considers factors related to economic decline within a country. For example, the indicator looks at patters of progressive economic decline of the society as a whole as measured by per capita income, GNP, unemployment rates, inflation, productivity, debt, poverty levels or business failures.

- Uneven Development (Economic Inequality)

This indicator considers inequality within the economy, irrespective of the actual performance of an economy. For example, it looks at structural inequality that is based on groups or based on education, economic status, or region.

- Human Flight and Brain Drain

This indicator considers the economic impact of human displacement for economic or political reasons and the consequences this may have on a country's development.

1.3 Political Indicators

- State Legitimacy

This indicator considers the representativeness and openness of government and its relationship with its citizenry.

- Public Services

This indicator refers to the presence of basic state functions that serve the people. On the one hand, this may include the provision of essential services, such as health, education, water and sanitation, transport infrastructure, electricity and power, and internet and connectivity.

- Human Right and Rule of Law

This indicator considers the relationship between the state and its population insofar as fundamental human rights are protected and freedoms are observed and respected.

1.4 Social Indicators

- Demographic Pressures

This indicator considers pressures upon the state deriving from the population itself or the environement around it. For example, it measures the population pressures related to food supply, access to safe water, and other life sustaining resources.

- Refugees

This indicator measures the pressure upon states caused by the forced displacement of large communities as a result of social, political, enviromental or other causes, measuring displacement within countries, as well as refuge flows into others.

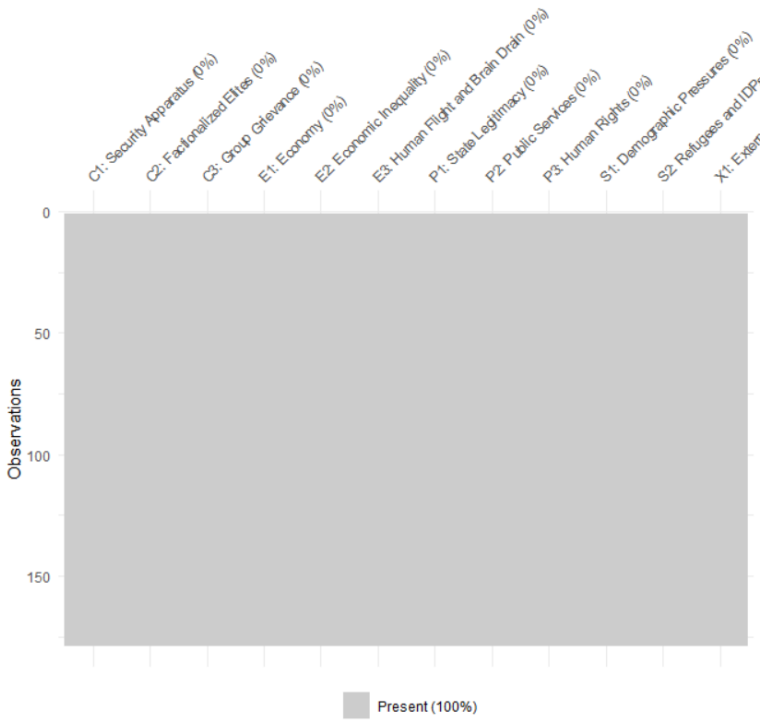
1.5 Cross-Cutting Indicators

- External Intervention

This indicator considers the influence and impact of external actors in the functioning- particularly security and economic- of a state.

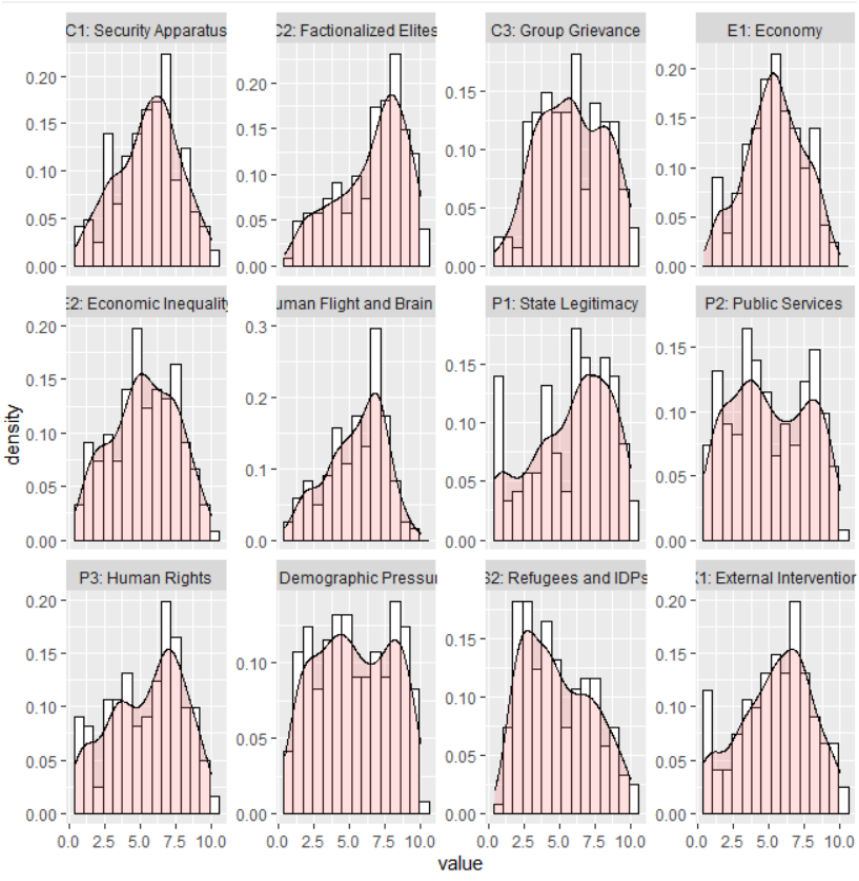
More information on how the data collected, measured and defined can be found here https://fragilestatesindex.org/methodology/.

1.1 Nas Handling

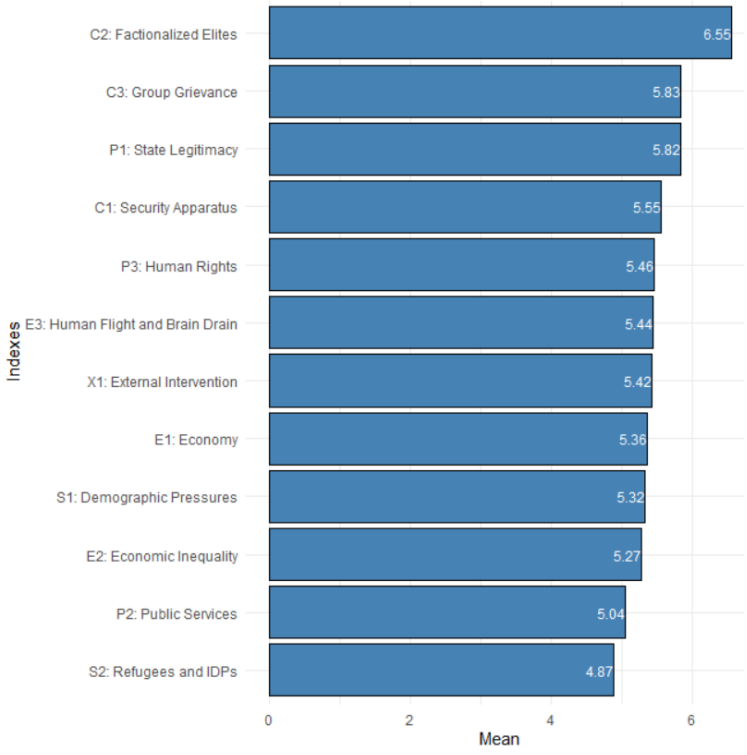


The plot above shows the distribution of NA values which we do not have any! Data is fully present.

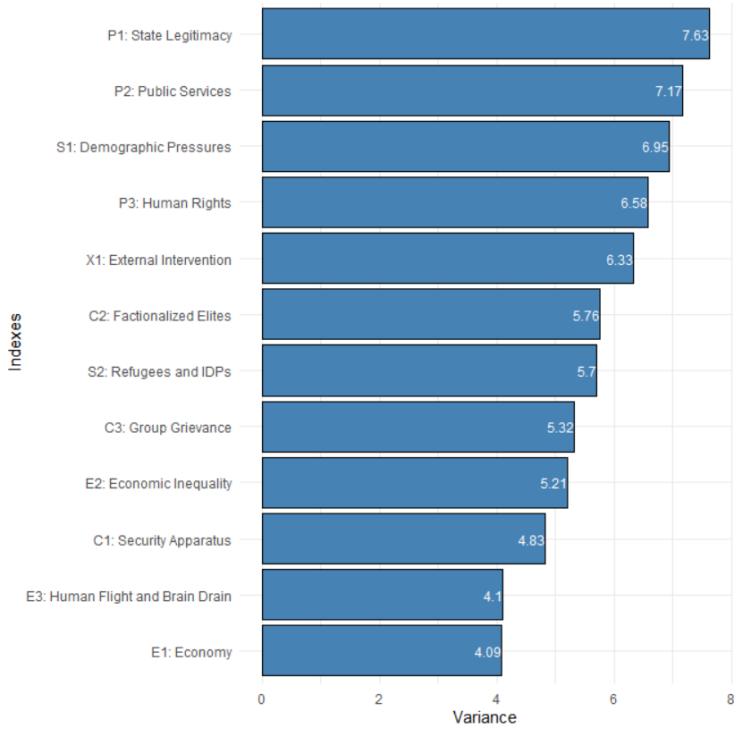
1.2 Descriptive Statistics



Notice that Factionalized Elites, State Legitimacy, Human Rights is slighlty left skewed meaning that most of the countries suffer from the fragmentation of state institutions along ethnic, religious etc. lines and from the representation problem of its citizens and from the protection of fundamental human rights, respectively. While The Refugees has a right skew meaning that there are a few countries in which refuge flow to external countries is high. Economic Inequality seems almost to be normally-distributed. This is actually quite intuitive since Economic Inequality is a global problem while it is extremely high in some countries. Overall, looking at the plots above we expect most of the countries to be fairly fragile.

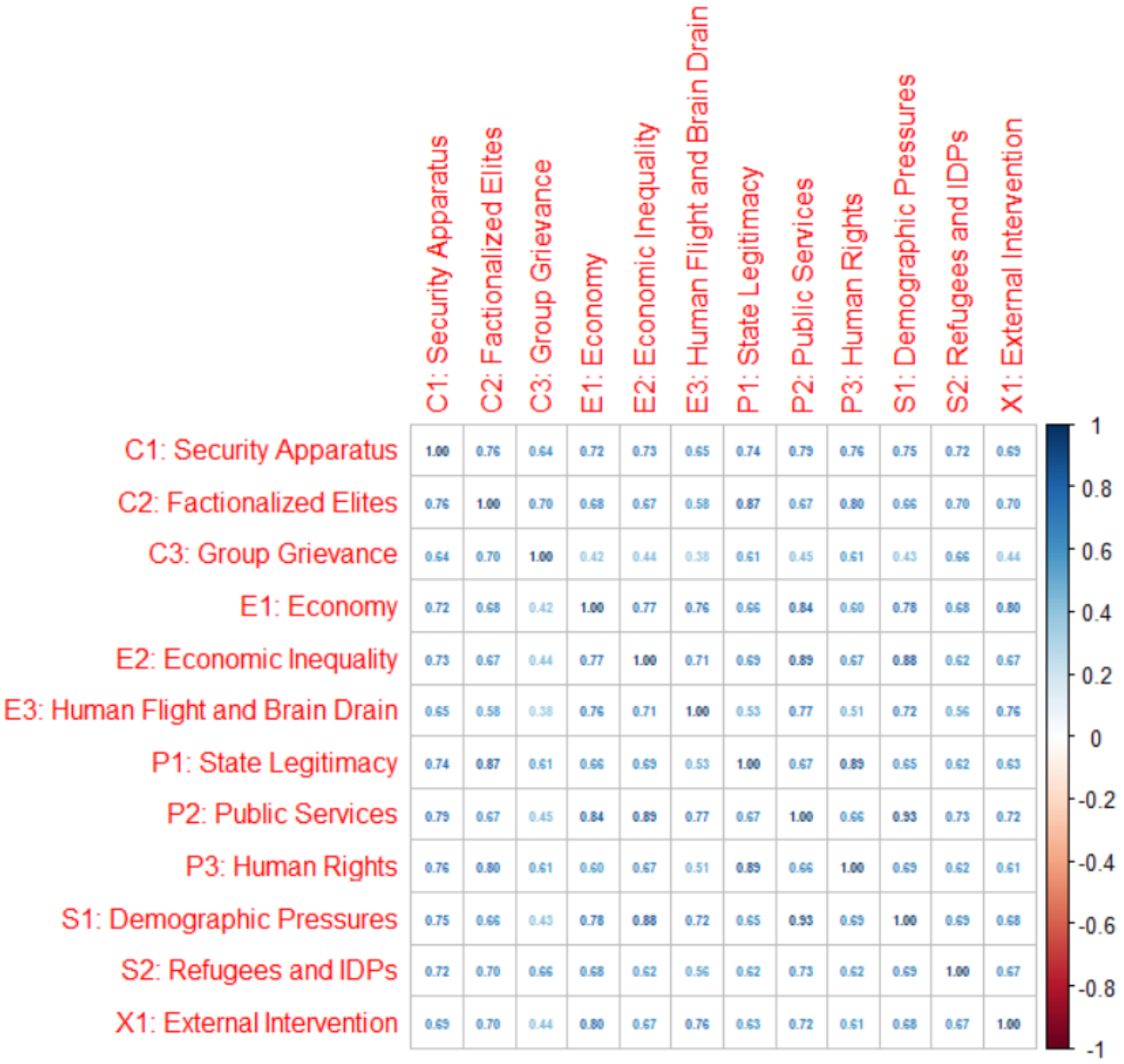


The plot might be more readable with respect to the previous one. As we can see, the mean value of Factionalized Elites is the highest as mentioned above. On average, most of the countries do not equally provide its services to the public. Factionalized Elites is followed by Group Grievance and State Legitimacy. Again, the democratic representation of people and equality in oppurtunies with respect to political processes is on average low. Refugees has the lowest mean as expected. As we already said above, most of the countries are expected to be fragile given that the measurement scale moves between 0 and 10 and the mean values of the indices are almost always higher than 5.



This plot shows the variability of indices across countries. Notice that Economy has the lowest variance which is very intuitive. This means that most of the countries has almost the same position with respect to Economic situation. Since, in today's world, economy has a global structure, the richness and poverty comes as a whole. Therefore, countries share similar positions with respect to progressive economic decline, inflation, debt etc.

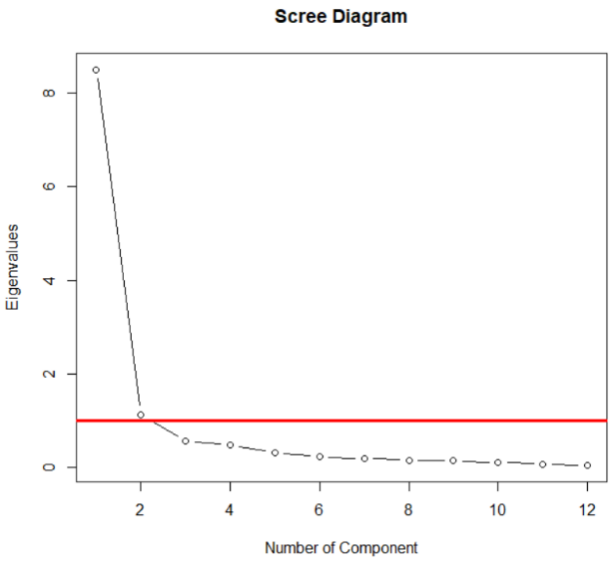
1.2.1 Multicollinearity

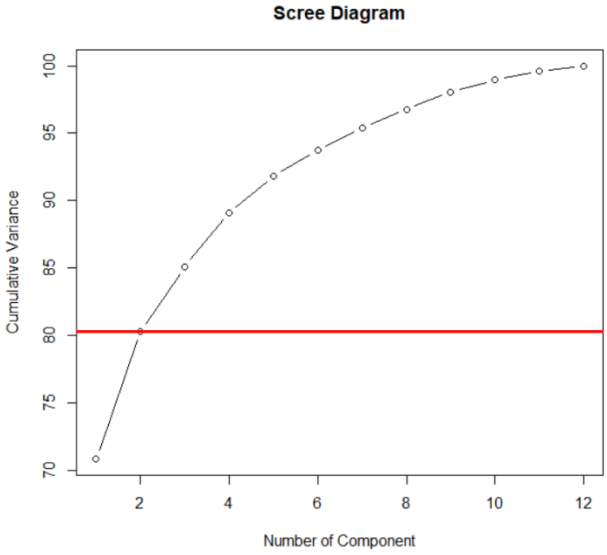


We have a high multicollinearity caused by the very structure of the dataset. This is expected when dealing with socio-economic data since they are highly related to each other by their nature. However, here understanding the causality is highly difficult. Economic inequality causes Security Apparatus or vice versa? We do not know it but it is very intuitive to understand why they are highly correlated because Economic Inequality in many cases might mean low education, under-development, failure in protection of human rights, equality in oppurtunies and corruption which all then might lead to Security Apparatus which is a perfect dimension to describe a failed state where terrorism is high and domestic security is very low. Where there is Security Apparatus, we of course would expect a group of people to secure their lifes with corruption: Security Apparatus has the highest correlation with Factionalized Elites which has the highest correlation with State Legitimacy which is again very intuitive because in countries where the government is owned constantly by a group of elites the legitimacy that this state has should be very low. We therefore expect to see the loading vectors of many of the variables focused on the same direction. In particular, looking at the multicollinearity plot, we could say that Group Grievance and Human Rights should set out themselves slightly away from others.

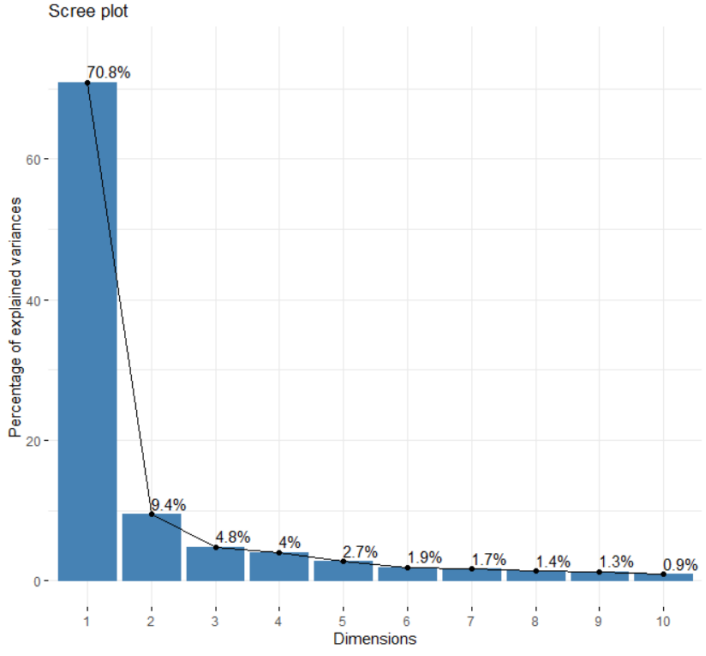
1.3 PCA ANALYSIS

Here, we want to figure out amount of information carried by each country namely its score, collecting all the correlated information contained in whole dataset and shrink it. This is done through the PCA analysis that responds to such aim by reducing the dataset dimension to a few components that are keeping most of the information contained in it. In particular, two is a desirable number of dimensions for graphical applications. The principal components are extracted through singluar value decomposition of the correlation matrix of the variables.. Each principal component is associated with an eigenvalue of that matrix and principal components associated with eigenvalues larger than 1 are able to explain more variance than only one variable. That is why we retain only the components whose eigenvalue is >1.





Above we have two plots showing the eigenvalues (left) and cumulative variance explanied (right). As we can see, we have two eigenvalues greater than 1 which amount to over %80 variance explained which is more than good for our analysis. We therefore retain only the first two components.



Dimensions	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	8.50171833525617	70.8476527938015	70.8476527938015
Dim.2	1.1320639290585	9.4338660754875	80.281518869289
Dim.3	0.579193794049563	4.82661495041303	85.108133819702
Dim.4	0.479396165765629	3.99496804804691	89.1031018677489
Dim.5	0.326291057316283	2.71909214430236	91.8221940120513
Dim.6	0.23185197825595	1.93209981879959	93.7542938308508
Dim.7	0.199844142740241	1.66536785616867	95.4196616870195
Dim.8	0.162973249867538	1.35811041556282	96.7777721025823
Dim.9	0.152073374334065	1.26727811945054	98.0450502220329
Dim.10	0.109955323834026	0.916294365283554	98.9613445873164
Dim.11	0.0792036487734216	0.66003040644518	99.6213749937616
Dim.12	0.045435000748608	0.3786250062384	100

As shown in the table(right), eigenvalues drop to under 1 right after the second

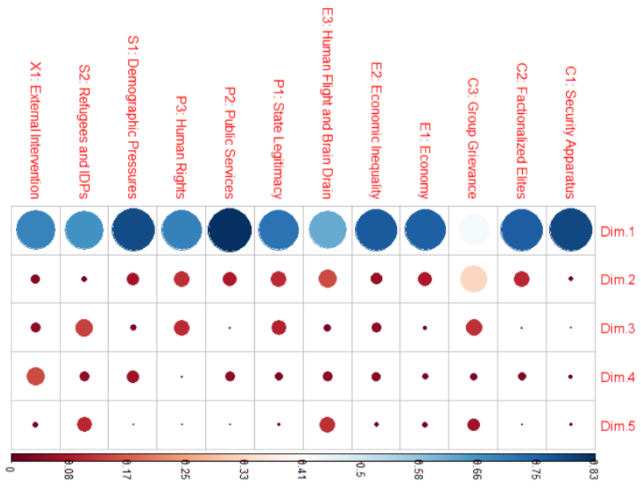
component. The sum of all the eigenvalues give a total variance of 12. The proportion of variation explained by each eigenvalue is given in the third column. For example, 8.50 divided by 12 equals to 70.84 or about %70.84 of the variation is explained by this first eigenvalue. The cumulative percentage explained is obtained by adding the successive proportions of variation explained obtained the running total.

1.4 RESULTS

We first present the correlation circle between the principal components and the features. The correlation between a variable and a principal component (PC) is used as the coordinates of the variables on the PC. The representation of variables differs from the plot of the observations. The observations are represented by their projections but the variables are represented by their correlations.

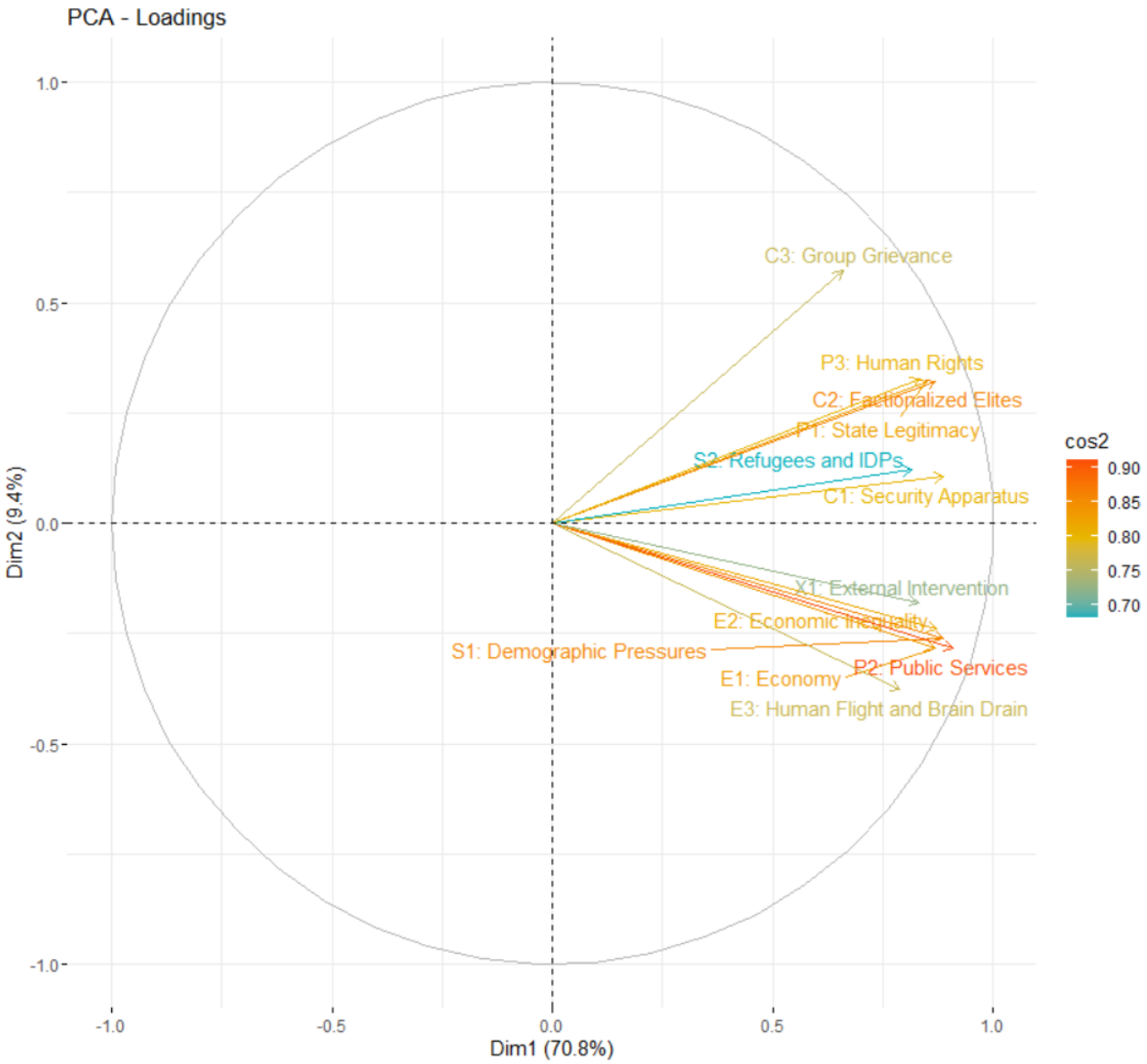
	Comp.1	Comp.2	Communality
C1: Security Apparatus	0.888	0.105	0.8
C2: Factionalized Elites	0.869	0.321	0.859
C3: Group Grievance	0.659	0.572	0.762
E1: Economy	0.868	-0.284	0.833
E2: Economic Inequality	0.872	-0.24	0.817
E3: Human Flight and Brain Drain	0.789	-0.378	0.765
P1: State Legitimacy	0.848	0.322	0.824
P2: Public Services	0.91	-0.285	0.91
P3: Human Rights	0.834	0.328	0.804
S1: Demographic Pressures	0.887	-0.262	0.855
S2: Refugees and IDPs	0.816	0.121	0.681
X1: External Intervention	0.831	-0.181	0.724

Here we see how our features are correlated with the given dimensions. The first and most important dimension is highly correlated with all the variables with the least being Group Grievance. In fact, we actually expected that when examining the correlation matrix at the beginning. Therefore, we can say being projected on the negative side of the X axis (recall that higher values imply more fragility) representing the first dimension will tell us how a country is away from being fragile. We expect EU countries to cluster around this happy cone. On the other hand, the second PC is only correlated with Group Grievance. So we recover a tiny bit of information lost with the first PC with respect to Group Grievance.



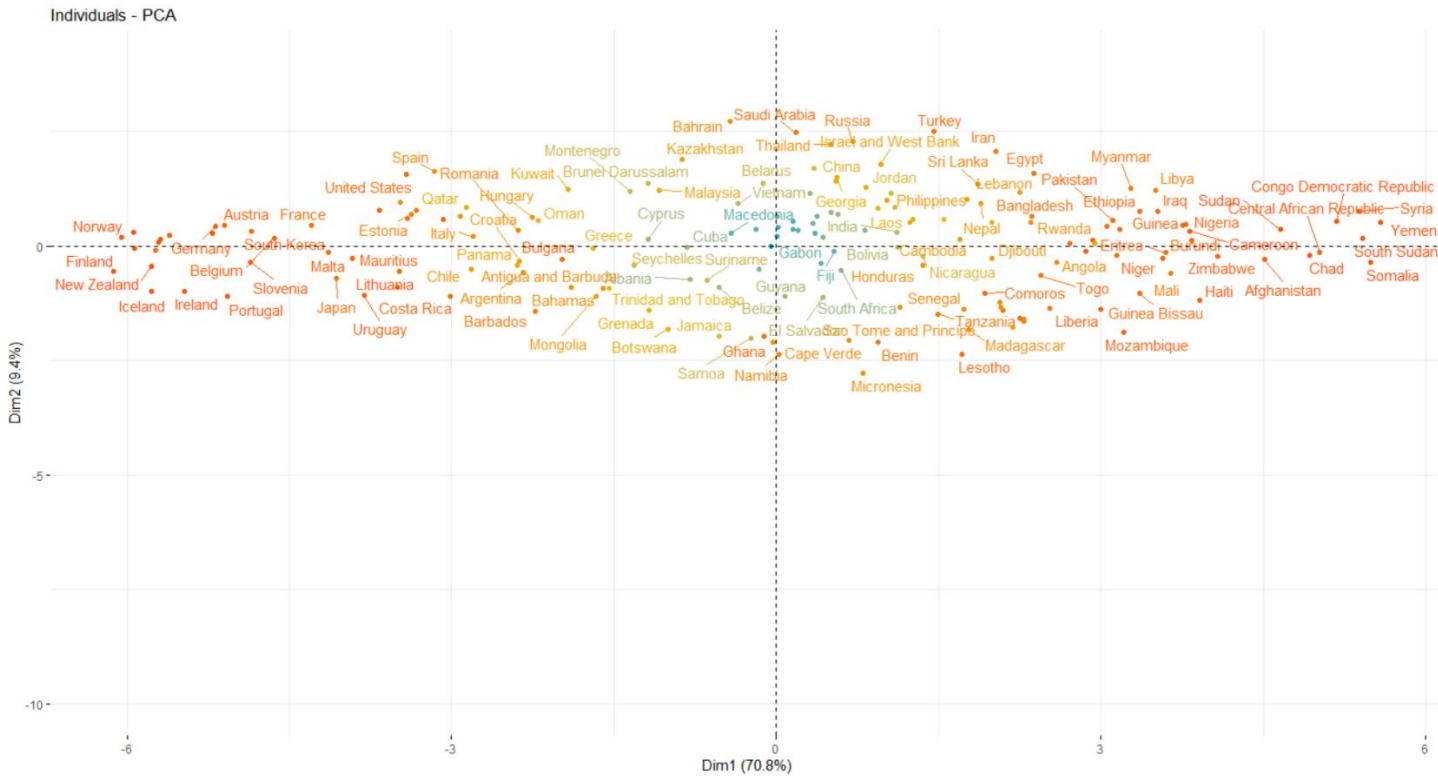
Notice that after the first component the strength of the relation with the variables decrease dramatically.

To give the correct interpretation of the scores' value, we now plot the loadings in the component's space.



Since the first component explaines over %70 of the variance, ıt seems like there is almost one strong direction on which we will find the fragile states. As we expected from our multicollinearity analysis, the Factionalized Elites, State Legitimacy and Human Rights are almost showing the same direction and point. On the other hand, Group Grievance is slighlty set apart from others since its information is provided by two different dimension. We see that economic variables are clustered together on the fourth quadrant. Countries focused on that region are expected to perform particularly bad in terms of economy which has a specific definition in this context explained in the data definition part. Countries clustered on the first quadrant will be relatively performing good in economy with respect to the countries in the fourth quadrant but are performing bad with respect to Social, Cohesion and Political indicators. Therefore, the countries close to the (1,0) point will be those which are the most fragile. Because they are performing bad with respect to all the indices which are Cohesion, Economic, Political and Cross-Cutting. We therefore expect all EU and other developed countries to be clustered on the negative side of the X axis away from the directions shown above.

We now project the countries on these two dimensions.

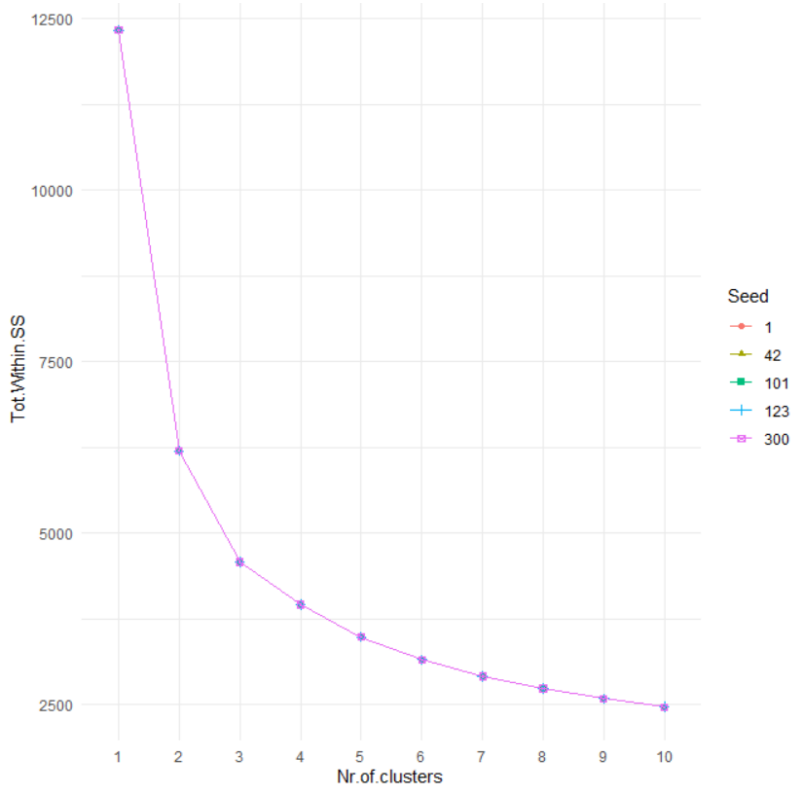


As expected, all of the European countries are on the negative side of the X axis clustered on the second and the third quadrant. In particular, Scandinavian and West European countries are located very close to each other significantly far away from being a fragile state. Expectedly, many Afrianca and Middle Eastern countries are located around (1,0) that is the unhappiest cone in this space. On the farthest point, we see Syria and Yemen : two countries where a civil war has been going on for years. Somalia is one of them, Sudan is accepted to be a failed state as well where there are lots of terrorist organizations. We can also see Afghanistan that is another failed state where more people die due to terrorism than natural death. About some countries on the first quadrant, we might take some interesting notes. For example, Russia, Israel, China and Turkey. These countries are doing good with respect to economic indicators but are performing bad with respect to Cohesion, Political and Social indiactors meaning the human rights, state legitimacy and group grievance. So, in these countries the division based on social and political characteritics is high as well as the fragmentation of state institutions along ethnic, racial or religious lines with also rule of law being low together with the failure of protection of fundamental human rights. Not entering in politics here, however, the problems about these countries are internationally known and seems to be approved by this analysis.

1.5 K-MEANS

K-means is a clustering algorithm in which the number of clusters is set a priori. Given the number of clusters, the algorithm seeks for the best partititon of individuals which minimizes the within cluster variance : observations within a cluster are expected to be similar, and so they should show a low level of variability. In our specific case, the within-cluster variation of a cluster is defined as the mean pairwise squared Euclidean distance between the observations that are part of the same cluster. The first step of K-Means algorithm consists in randomly assigning all the observations to a certain cluster, and then at each step we reduce the within-cluster variation

by changing the assigment. Therefore naturally, the output of K-Means critically depends on that initial starting random assignment. That is why we perform this random assigment with different seeds and see how the within cluster variation changes as we enlarge the number of clusters K.



Above plot shows the total within variance depending on number of clusters. It seems like all the seeds chosen provided a very similar result. We can see that after at the 5th cluster, the variability decreased by a huge margin however after that the variability became somehow stable. We therefore chose to perform K-Means with 5 clusters.

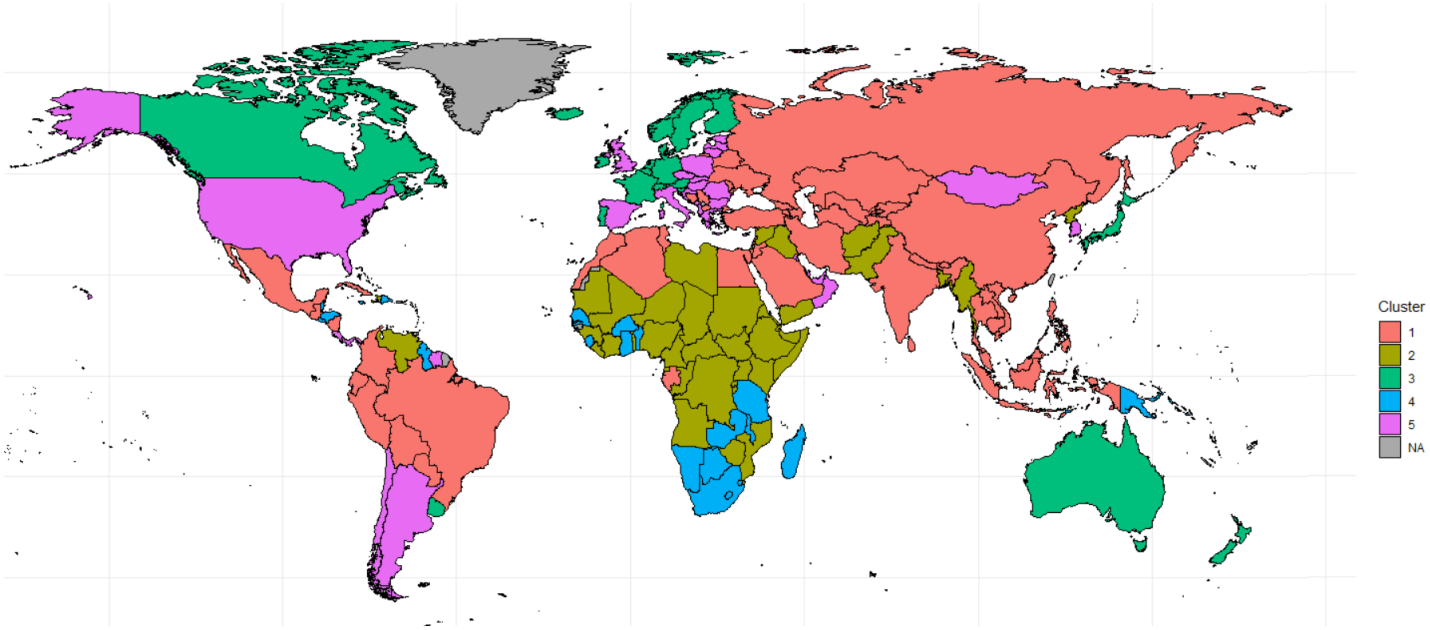
	1	2	3	4	5
C1: Security Apparatus	6.17733963880338	7.98683610731162	2.19608057355417	5.85254135013012	3.80543623664911
C2: Factionalized Elites	7.91944230769231	8.80546153846154	2.28636363636364	6.26206896551724	4.97777777777778
C3: Group Grievance	7.13237394141126	7.84550126570722	2.9144620433263	4.44827586206897	4.65249000990426
E1: Economy	5.14476907217148	7.4954500903623	2.3226792959829	6.79294058030694	4.0457177127824
E2: Economic Inequality	5.24947678614629	7.68821512551249	1.66668219014109	6.81883207518847	3.64369284949427
E3: Human Flight and Brain Drain	5.53818362576781	6.93752891416313	2.15242281778034	7.2551317715462	4.20565724370638
P1: State Legitimacy	7.18847840970523	8.63264574600641	0.968197702803305	5.42836804891549	4.1011754190585
P2: Public Services	4.58887971046433	8.49511805545231	1.24006109722476	6.93013525794423	2.77070151592409
P3: Human Rights	6.82675374358758	7.93683983898471	1.40939283152529	5.14733874961293	3.51445704297672
S1: Demographic Pressures	5.02255585350403	8.39346190611289	1.73810921446728	7.24819573706449	3.05452761615676
S2: Refugees and IDPs	5.06346153846154	7.94698628226746	2.54657405816966	4.71379310344828	2.8190452323911
X1: External Intervention	5.46925626687597	7.78851336788661	1.23960840020542	6.75283469951379	4.25556536440064
Indicator Cluster Means	5.94341424121593	7.99604651985239	1.89005282179535	6.13753801677143	3.82052033510183

Here is the mean values of each cluster with respect to a given indicator. Recall that higher values imply higher fragility. We see that in the second cluster we have the highest mean cluster values, therefore Cluster 2 should be containing the most fragile states, should be mainly middle eastern countries. Almost all the clusters differ significantly from eachother except the first and the fourth cluster. They have similar values for indicator means, the difference is that Cluster 1 countries are slighlty performing better than the Cluster 4 with respect to Economy while doing bad with respect to Human Rights. We therefore expect to have for example China, Russia, Turkey

and Israel in the Cluster 1 as we discussed in PCA analysis. The second most fragile group is the Cluster 4 which we expect to see African countries if we recall their projections from the PCA result. Based on this table, we can conclude that EU and some other developed countries should be in Cluster 3. Cluster 5 is the most similar cluster to Cluster 1. In Cluster 5 where we expect to see eastern European countries where there still is the problem of Factionalized Elites and Group Grievance.

	1	2	3	4
2	0.665152213			
3	0.340034621	0.005186834		
4	0.735402110	0.684988208	0.309824958	
5	0.653205979	0.318358192	0.686828642	0.616093980

Here is the table showing a similarity measure between the clusters. As we can see, Cluster 1 and 4 is very similar to each other as Cluster 5 and Cluster 3.



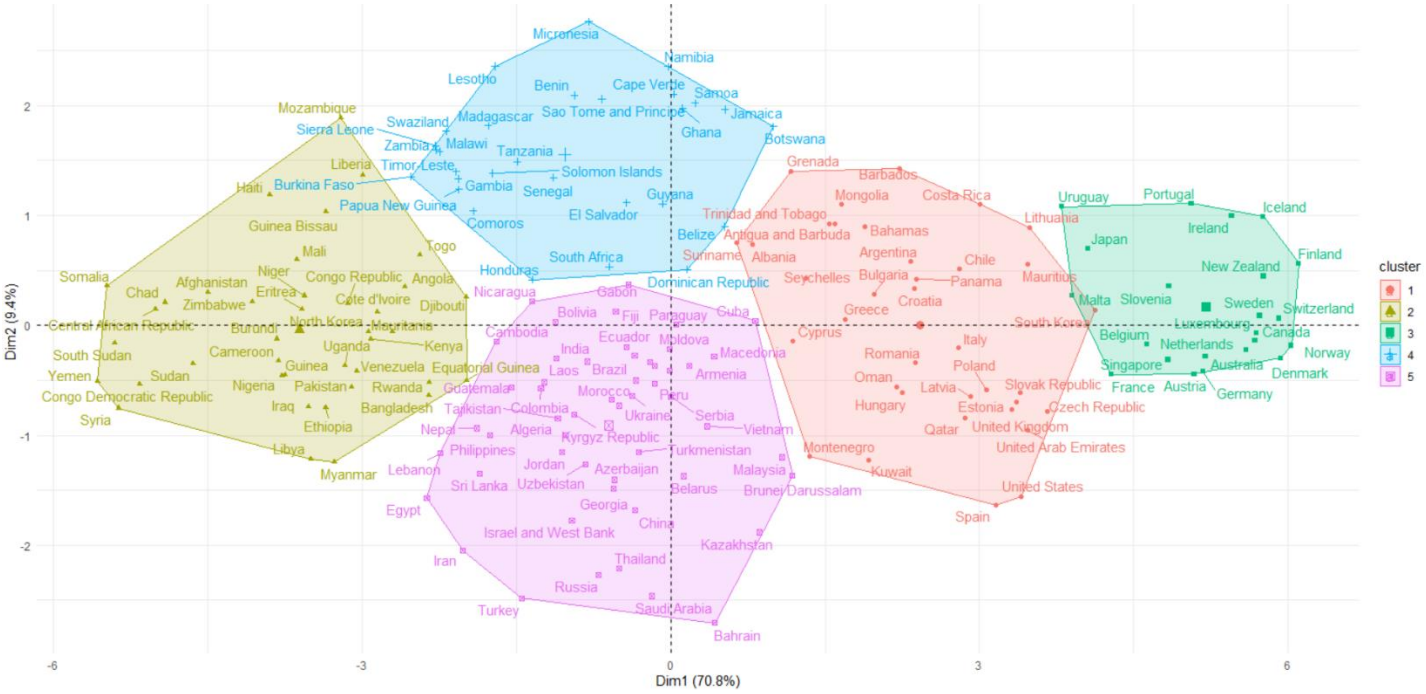
Cluster 1: Contains countries from almost all the continents. As we expected, Turkey, Russia, China and Israel falls into this cluster. In this cluster, we have the countries that are doing relatively good with respect to economic indicators but performing bad with respect to Human Rights, Factionalized Elites, Group Grievance and State Legitimacy. These countries fragility seem to emerge from the problems that they have with their own citizens. Therefore, State Legitimacy is low. Notice that these countries generally have long lasting governments. Turkey has been ruled by the same government for almost 20 years same as Russia. Israeli government has been in power for the last 12 years. As for China, they are de facto under communist regime where the group of people governing country never changes. When we go back and see again the Cluster means, we can understand why these countries fall in a Cluster where Factionalized Elites, Group Grievance and State Legitimacy has high values and low with respect to economic indicators. These are developing countries, except Israel, so having a better performance in economy but a bad performance with respect to indicators mentioned above.

Cluster 2: This cluster basically contains the failed states. In this Cluster, the mean of the all the within-cluster indicators mean is almost 8. That's a very big number given that maximum is the 10. These are countries where there are civil wars going on for years. Here we have Syria, Yemen, Iraq, Afghanistan, Sudan etc.

Cluster 3: This cluster is composed by the least Fragile States. Many Western European countries together with Singapore, Canada, New Zealand, Japan is here.

Cluster 4: This cluster is mainly containing Afrian countries and this is second most fragile group. They are somehow similar to the Cluster 2 but recall that the mean for the Cluster 2 is 8 while here it is 6.1. So even if they are somehow similar in terms of fragility, Cluster 2 countries are way more fragile and most of them are even failed states. This difference shows us that we actually made a good choice by chosing 5 Clusters because otherwise we could not capture this difference.

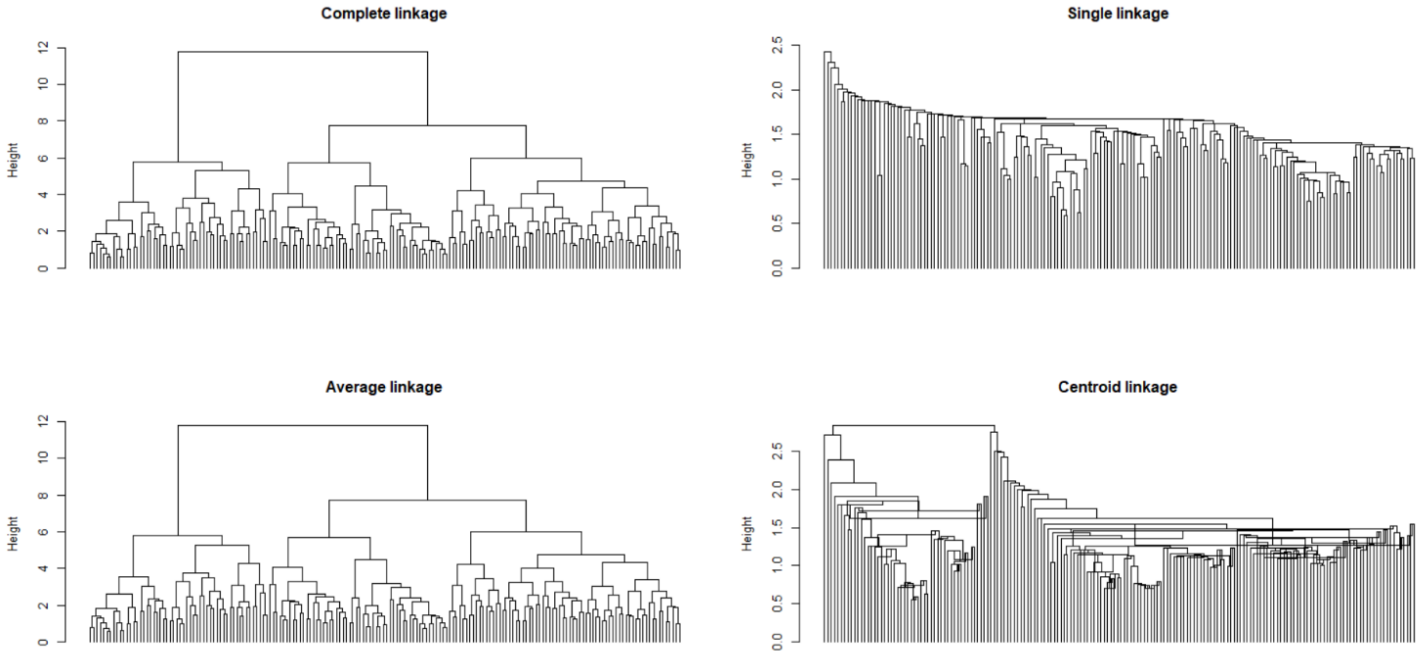
Cluster 5: This Cluster has Italy, USA and some Eastern European countries. These countries are doing good with respect to Public Services, Democratic Pressures, Human Rights. However, they suffer from Economic indicators (recall the national debts of these countries), Group Grievance, Factionalized Elites and Human Flight and Brain Drain.



Here we see the clusters of countries in the reduced principal component space. Notice that here principal components multiplied by a minus so the presentation is simply reversed with respect to the PC map we drawn above. At first glance, we actually see that 5 as a choice of number of Cluster seems to be working good! There are not overlapping among the Clusters. As we can see Cluster 3 and 5 are next to each other. They were in fact very similar. On the other hand, we see Cluster 2 on the left part of the plot where they are highly correlated with all the indicators' loading vector's drawn above and it contains basically the failed states.

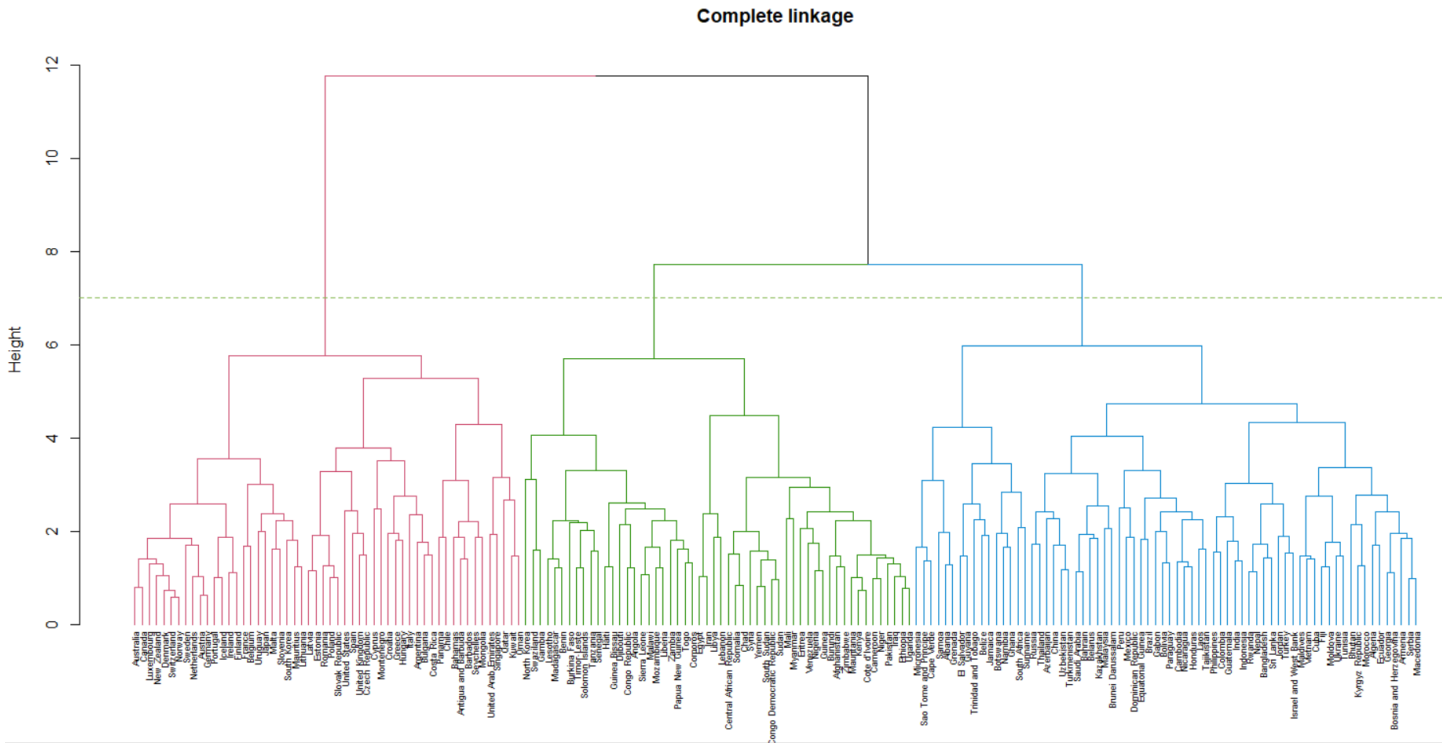
1.6 HIERARCHICAL CLUSTERING

In contrast to K-Means, Hierarchical Clustering method does not depend on the initial definition of the number of clusters. Instead, they require the user to specify a mesaure of dissimilarity between th group of observations. Hierarchical Clustering has an added advantage over K-Means clustering in that it results in an attractive tree-based representation of the observations, called dendrogram. A dendrogram provides a very interpretable complete description of the hierarchical clustering in a graphical format. In agglomerative approach, the clustering starts at the bottom where each observation represents a separate cluster and it recursively combines the two nearest clusters together until all the observations are grouped in a single cluster. In order to perform clustering, we first compute the distance matrix with distances for each pair of observations. Secondly, the algorithm needs us to specify the measure of distance between clusters in order to decide the rules for clustering. We perform hierarchical clustering of the observations using four different distance measures and compare the results by plotting the corresponding dendrograms. Here an important note to be made is that, before clustering the dataset was scaled, subtracting the mean to each variable and dividing by its standard deviation : such step ensures variables comparability and is indispensable.

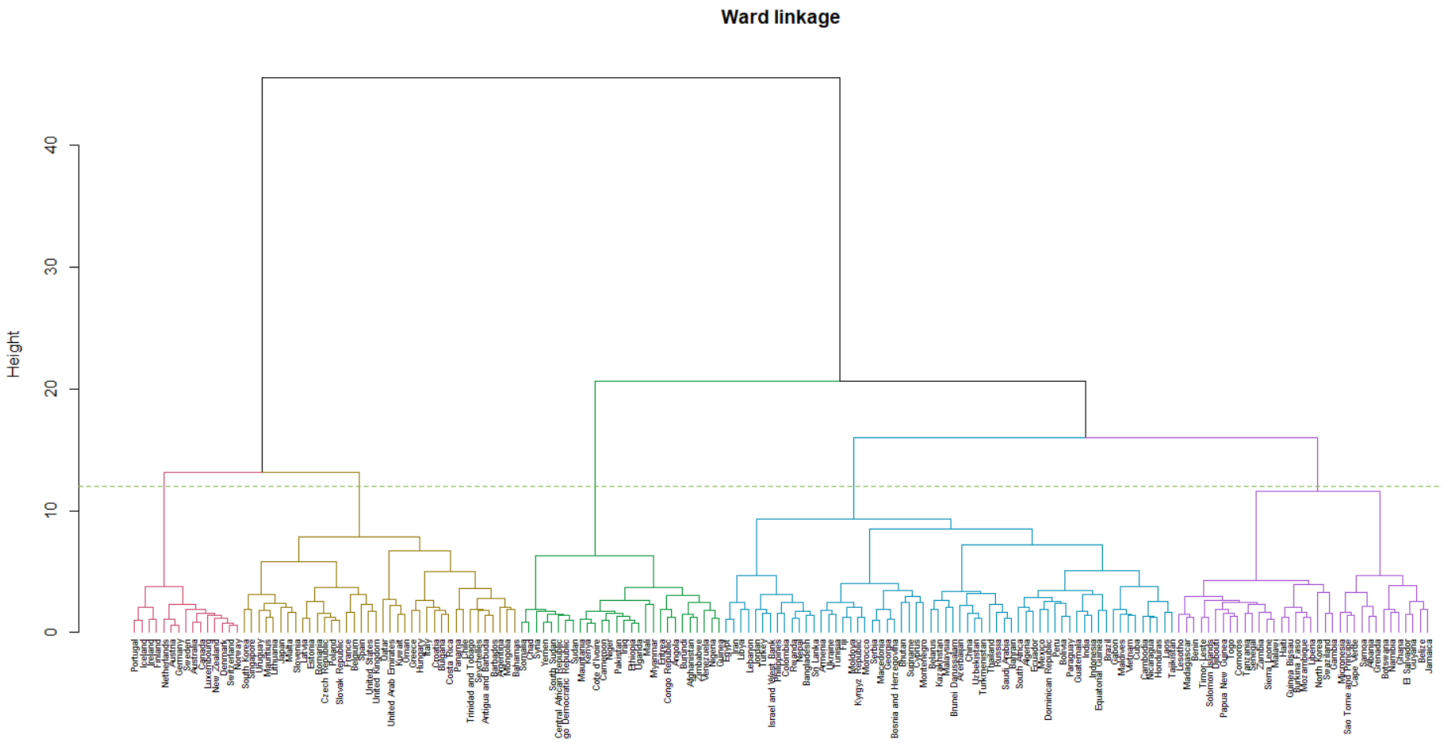


1.6.1 Selecting the agglomerative method

We can notice that the dendrograms obtained with single linkage and centroid linkage are strongly unbalanced and seem to be less helpful in identifying the clusters. The Complete and Average approach, on the other hand, looks much more interpretable. They seem to work better for this data, identifying well separated groups. Nevertheless, acknowledging the presence of outliers among our observations we also try Ward Agglomerative method which could be more appropriate in our case.



Complete linkage method which measures the maximal inter-cluster dissimilatriy seems to suggest three main clusters. The first cluster (left) contains almost all the EU countries without separating them eastern or western, with USA, UK and some other rich countries. The second cluster (middle) on the other hand includes the most fragile states: Syria, Yemen, Libya, Sudan are in this Cluster. The third Cluster (right) seems to have medium-fragile countries where we have China, Russia, Turkey, Israel which were always grouped together by all the algorithms we used so far. So, with complete linkage method, we can say that we have highly-fragile, medium-fragile and low-fragile groups which are second, third and first clusters respectively.



On the other hand, Ward method seems to suggest 5 clusters. This looks better for our dataset. We can see, on the very left (the first cluster) the western European countries. Complete linkage used to divide Europe in three clusters however having two clusters for EU sounds more plausible. The third cluster (green) is the most fragile or namely failed states including almost all the countries where there are serious problems. In the blue cluster we have again Russia, Turkey, China and Israel while last cluster is generally composed of African countries. Based on this results, we would prefer to go on with ward method since it is presenting a more compact and clear result.

1.7 CONCLUSION

In PCA analysis, we had two main dimensions where countries are clustered. In particular, there was one main direction where failed states grouped together. In this direction, we had highest scores for Security Apparatus, Factionalized Elites, Economic Inequality, State Legitimacy, Democgraphic Pressures. Countries that are having higher values in those variables turned out to be the super-fragile ones.

With K-Means clustering, we saw that with 5 clusters we had a very good representation of the countries. In these clusters, a small group of non-fragile states(Cluster 3) which is mainly consituted by Scandinavian and West European countries. In another Cluster(Cluster 5), we have slighlty fragile states which include Easter European countries with USA and Italy. This result was robust with respect to the PCA anaylsis where we had a similar distribution of the countries on the loading space. For example, Italy and Chile turned out to be in Cluster 5 and they were projected next to each other on the PCA loading space. In Cluster 2, we had the failed states which included Middle East and African countries where there has been civil war, terrosim and high poverty going on for years.

Lasty, we applied hierarchical clustering. In particular, with complete linkage method, there turned out be 3 clusters. However, this was slightly an over-simplification where we could not

get detailed representation. For example, the first cluster contained one of the most least-fragile states, Norway and Grece that is a way more fragile than Norway given the PCA and K-Means result. They were substantially far away from each other on the loading space projection of PCA anaylsis, also in K-Means clustering they were in another cluster. There are many other examples showing this simplification. To get a better representation, we used the Ward-Method where we get a clearer and a more robust representation with respect to the PCA and K-Means. Ward method provided a very similar result to PCA and K-Means. Here, we again had Greece and Italy in another cluster next to the West European and Scandinavian Countries. The Cluster of China, Russia, Israel and Turkey was the same like in the PCA and K-means. We therefore decided on the ward method.

In Conclusion, we can say that there are five types of countries. Those that are highly developed with respect to all the variables, those that are developed but having problems with respect to economic indicators(dept, inflation etc.). Another Cluster is newly-developed countries where there are problems about economy, state legitimacy, human rights(Russia, China, Turkey). The Fourth Cluster is highly fragile states. In this cluster, we have high values for all the variables but they are not as high as the last and the fifth cluster which is basically the failed states.