




A statistical analysis of **State Fragility Index**

Data Science and Economics – A.Y. 2021-22
Joint Project for Advanced Multivariate Statistics Course
Marco Cazzola – 964573
Murat Aydin – 965334



DATA DEFINITION & RESEARCH OUTLINE

What is Fragile State Index about?

The Fragile State Index (FSI) is an annual ranking of world countries (here grouped according to their level of Human Development) based on the different pressures they face that impact their level of fragility.

How is the Index built?

Each country is evaluated on 12 key political, social and economic indicators, whose value vary between 0 and 10. The indicators can be grouped according to the field they are related to.

DATA SET FEATURES

COHESION INDICATORS

- ⊙ Security Apparatus
- ⊙ Factionalized Elites
- ⊙ Group Grievance

POLITICAL INDICATORS

- ⊙ State Legitimacy
- ⊙ Public Services
- ⊙ Human Rights and Rule of Law

ECONOMIC INDICATORS

- ⊙ Economic Decline and Poverty
- ⊙ Economic Inequality
- ⊙ Human Flight and Brain Drain

SOCIAL INDICATORS

- ⊙ Demographic Pressure
- ⊙ Refugees

CROSS-CUTTING INDICATORS

- ⊙ External Intervention

A decorative network graph in the top-left corner, consisting of various sized nodes (some solid grey, some hollow white with grey outlines) connected by thin grey lines. The graph is partially cut off by the top and left edges of the slide.

1.

BOOTSTRAP

Does fragility vary significantly across the world?

THE BOOTSTRAP

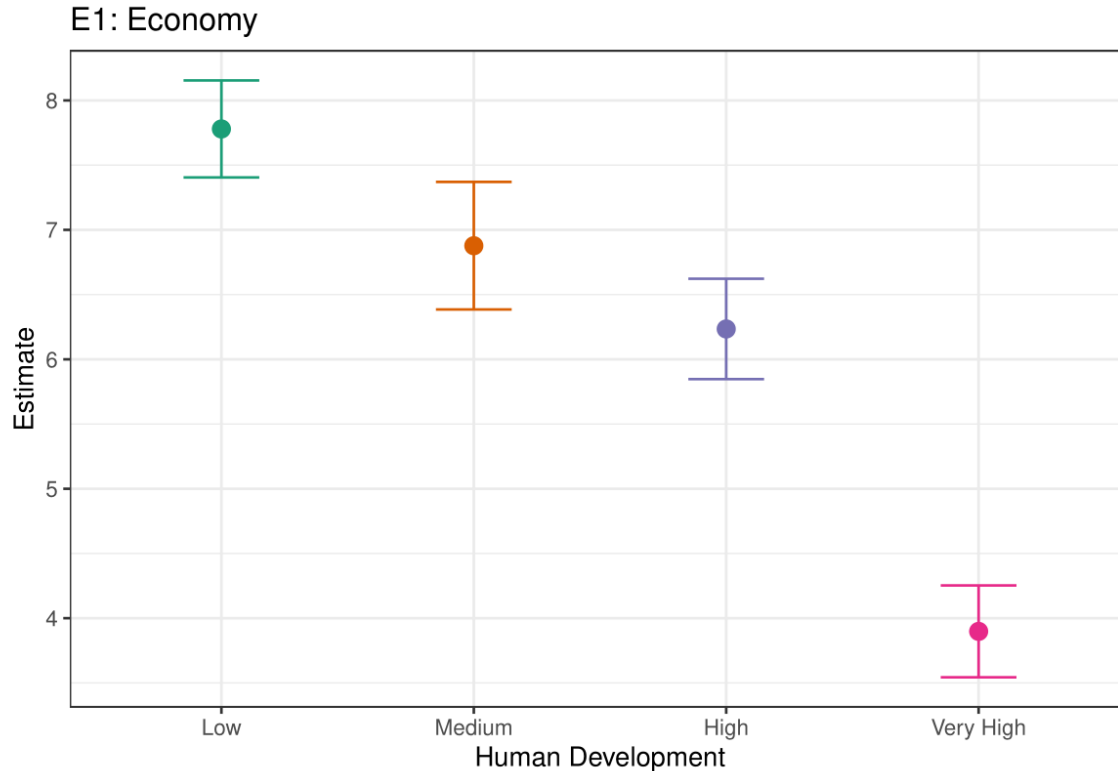
Why using the bootstrap?

Bootstrap is a very powerful statistical method which can be applied to a wide set of problems. The main advantage related to the bootstrap is that it does not require any strong assumption, differently from other methods such as ANOVA and MANOVA.

The application

We will use the bootstrap technique to estimate the means (and the related standard errors) for each group of countries, as defined by the Human Development Index, in order to test for significant differences among these groups.

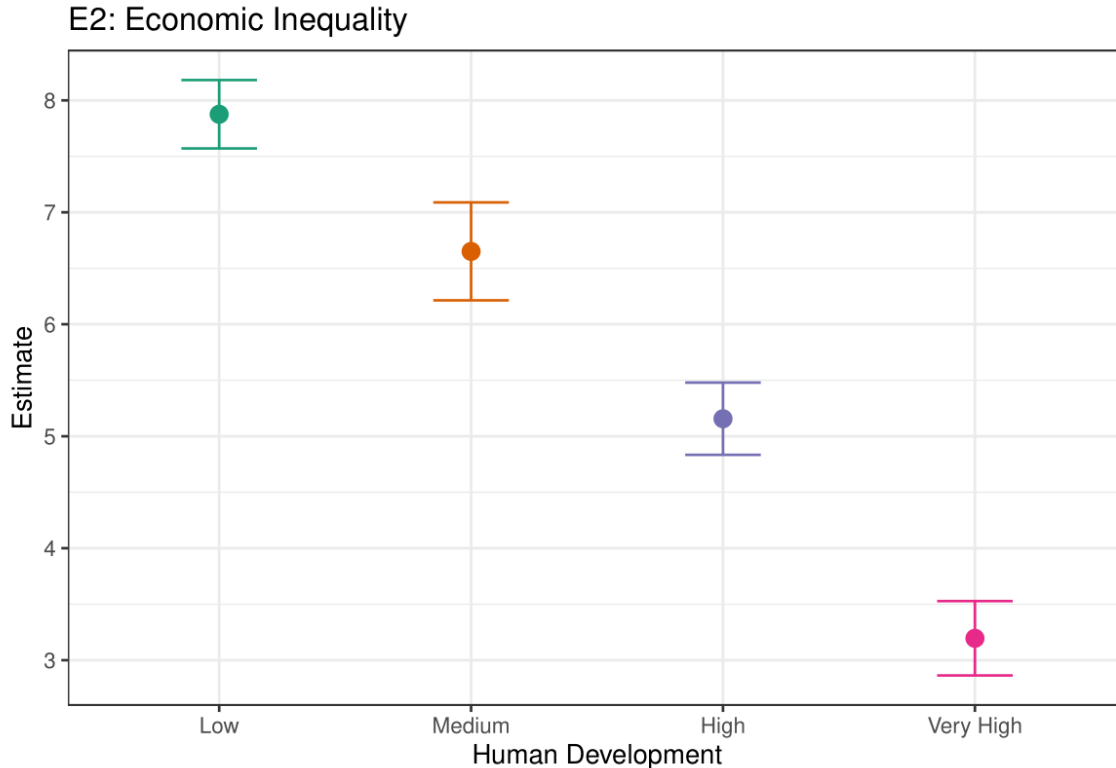
UNIVARIATE BOOTSTRAP



Very High HDI countries set themselves very far apart from the others.

On the contrary, the rest of the world suffer from general poverty and economic decline.

UNIVARIATE BOOTSTRAP

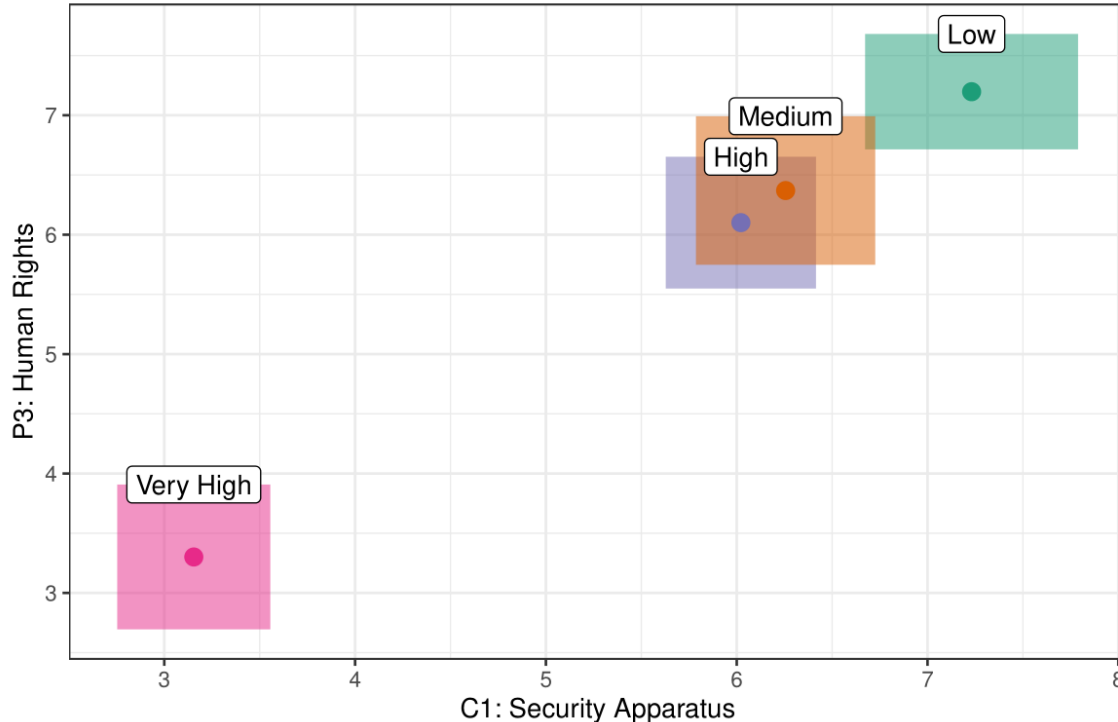


Very High group has the lowest inequality, while the very low group has the highest.

The largest variance is seen in the medium group and all the means are significantly different.

BIVARIATE BOOTSTRAP

C1: Security Apparatus vs P3: Human Rights



Once again, Very High HDI groups set themselves very well apart from the others.

While the differences between High and Low HDI countries are significant, they are not for Medium vs High and for Medium vs Low.

A decorative network graph in the top-left corner, consisting of various sized nodes (some solid grey, some hollow white) connected by thin grey lines, forming a complex web-like structure.

2.

DIMENSIONALITY REDUCTION

Is it possible to represent countries' differences on a reduced space?

DIMENSIONALITY REDUCTION

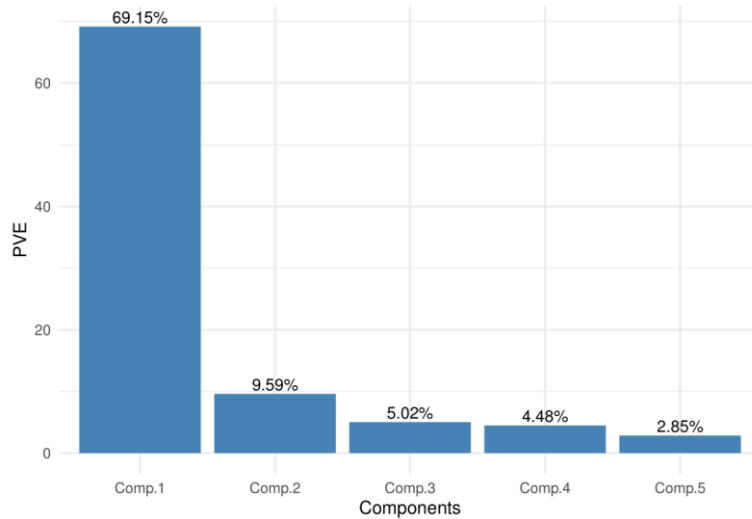
The goal of these techniques

Dimensionality reduction techniques are useful methods which allow to obtain a representation of the data in a reduced space. In particular, PCA takes as input the correlation matrix of the original variables, while MDS considers the proximity matrix of the observations.

The application

We will use the PCA and MDS to map countries in a reduced space, and see if their position in this reduced space is somehow correlated with their ranking according to the Human Development Index.

PRINCIPAL COMPONENT ANALYSIS

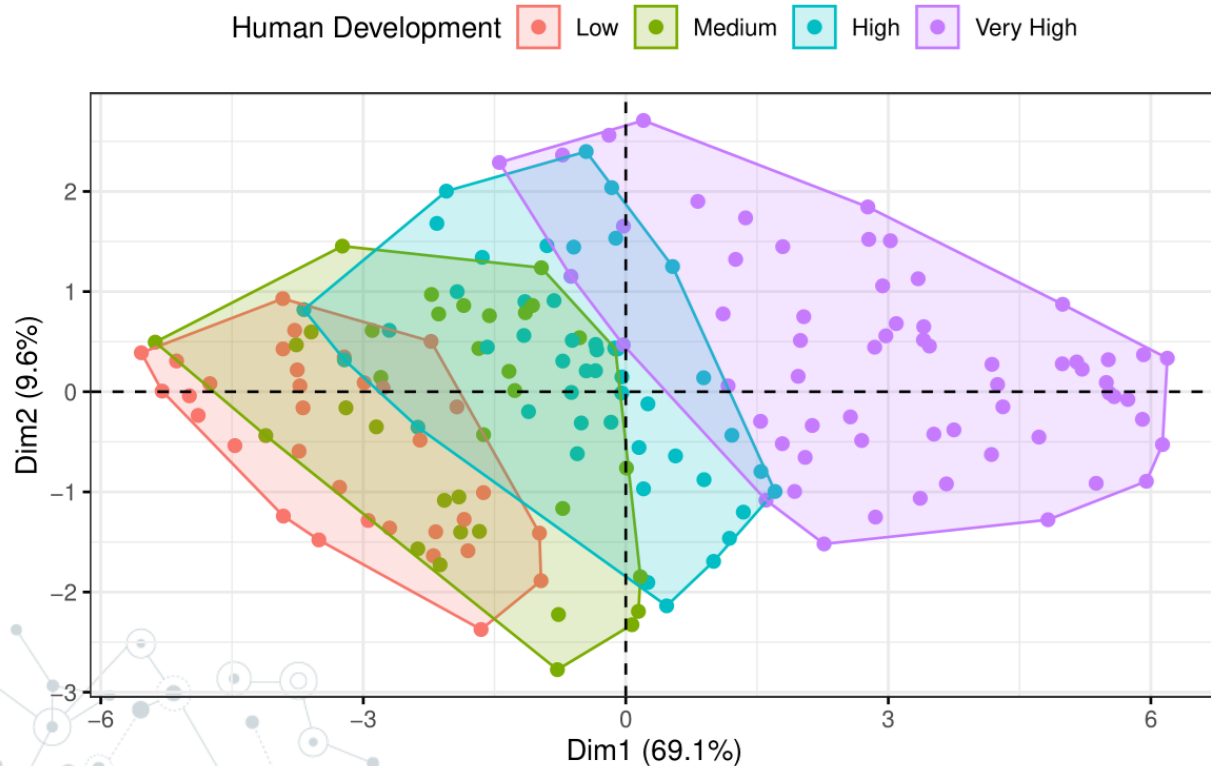


With just two components, we are able to explain ~ 80% of the data set variability.

The first component is strongly correlated with all the features but *Group Grievances*, which is more related to the second component.

	C1: Security Apparatus	C2: Factionalized Elites	C3: Group Grievance	E1: Economy	E2: Economic Inequality	E3: Human Flight and Brain Drain	P1: State Legitimacy	P2: Public Services	S1: Human Rights	S2: Demographic Pressures	X1: Refugees and IDPs	X1: External Inter
Comp.1	-0.88	-0.86	-0.66	-0.82	-0.86	-0.78	-0.85	-0.91	-0.83	-0.87	-0.81	-0.82
Comp.2	-0.08	-0.32	-0.55	0.35	0.22	0.38	-0.35	0.26	-0.37	0.23	-0.06	0.20
Communality	0.79	0.85	0.74	0.80	0.79	0.74	0.84	0.89	0.82	0.81	0.66	0.71

PRINCIPAL COMPONENT ANALYSIS

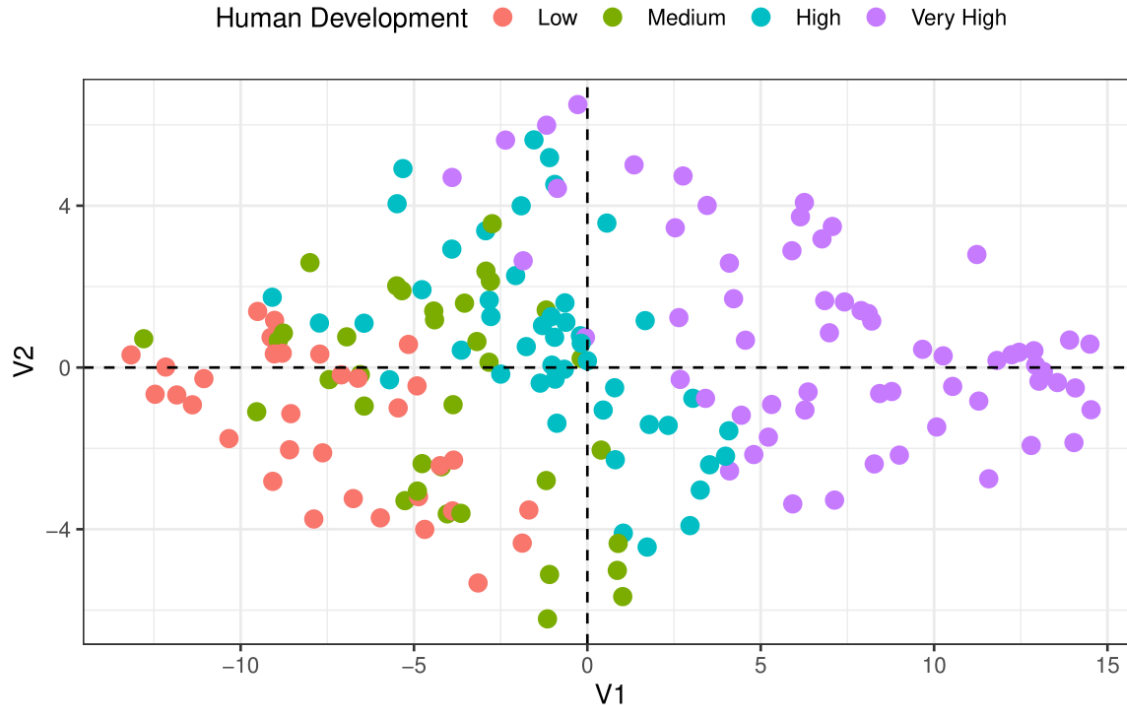


Not surprisingly, Very High HD countries are less fragile and less affected by Group Grievance.

Interestingly, there is much overlapping between all the three remaining groups.

MULTIDIMENSIONAL SCALING

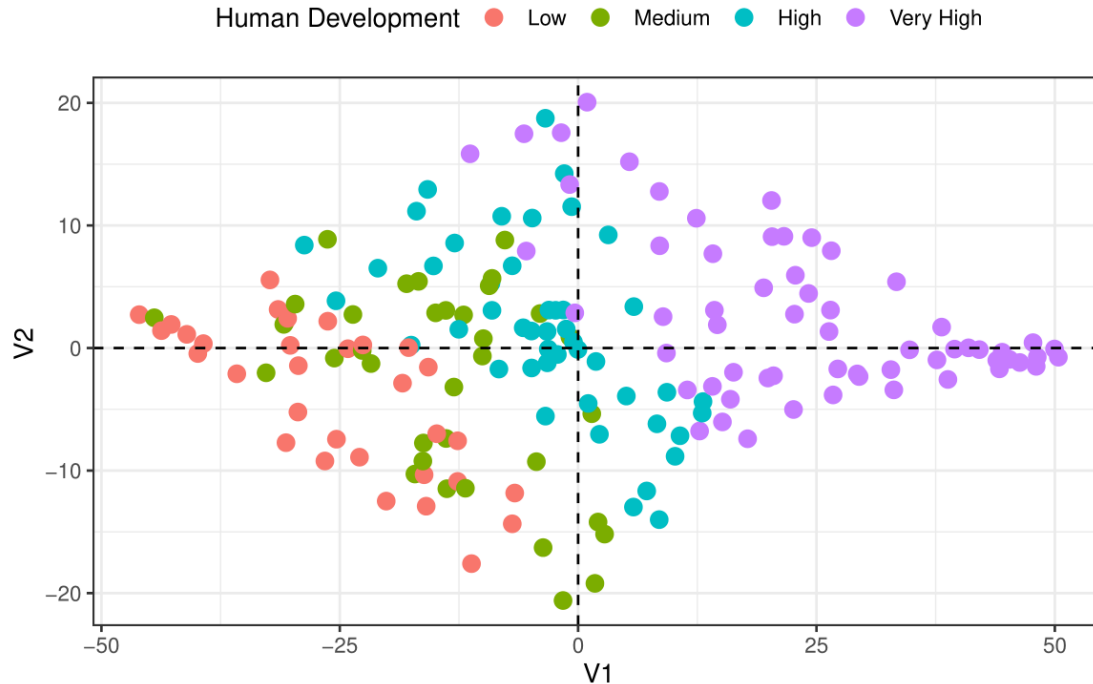
MDS with Euclidean distance



With Euclidean distance, multidimensional scaling returns exactly the same scores as PCA computed through the sample covariance matrix.

MULTIDIMENSIONAL SCALING

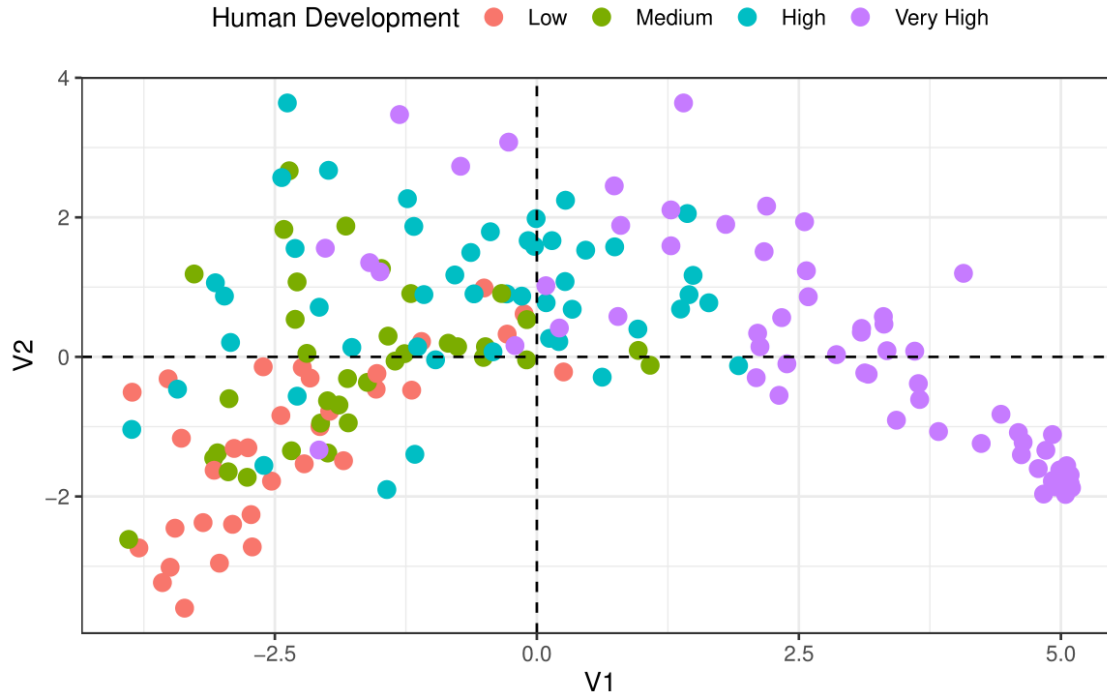
MDS with Manhattan distance



With Manhattan distance, observations appear to be a little bit less sparse, but there is not much difference.

MULTIDIMENSIONAL SCALING

MDS with Maximum distance



With Maximum distance,
the representation gets
really messy.

The best choice is
probably PCA (or MDS
with Euclidean distance
proximity matrix).

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting different levels of connectivity or importance. The lines are thin and gray, creating a mesh-like structure.

3. **CLUSTERING**

Grouping countries according to their differences

CLUSTERING METHODS

Different approaches

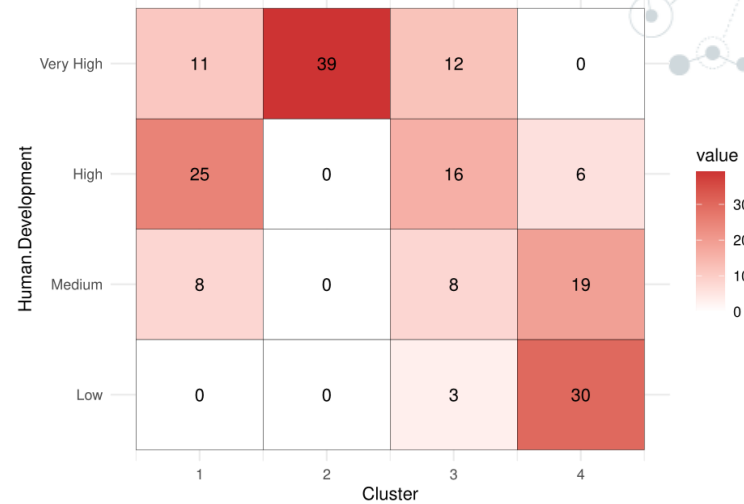
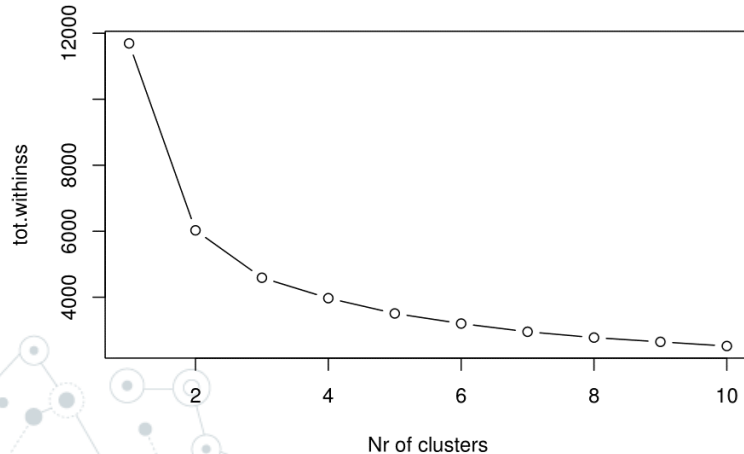
Methods like *K-means* and *hierarchical clustering* are based on heuristic and reasonable procedures, but they are not supported by a solid statistical model, differently from *model-based clustering*, which therefore could be used for inferencial purposes as well.

The application

We will now cluster observations using K-means and model-based clustering, comparing the results with the groups as defined by the level of Human Development to see if there is any correspondence between the two.

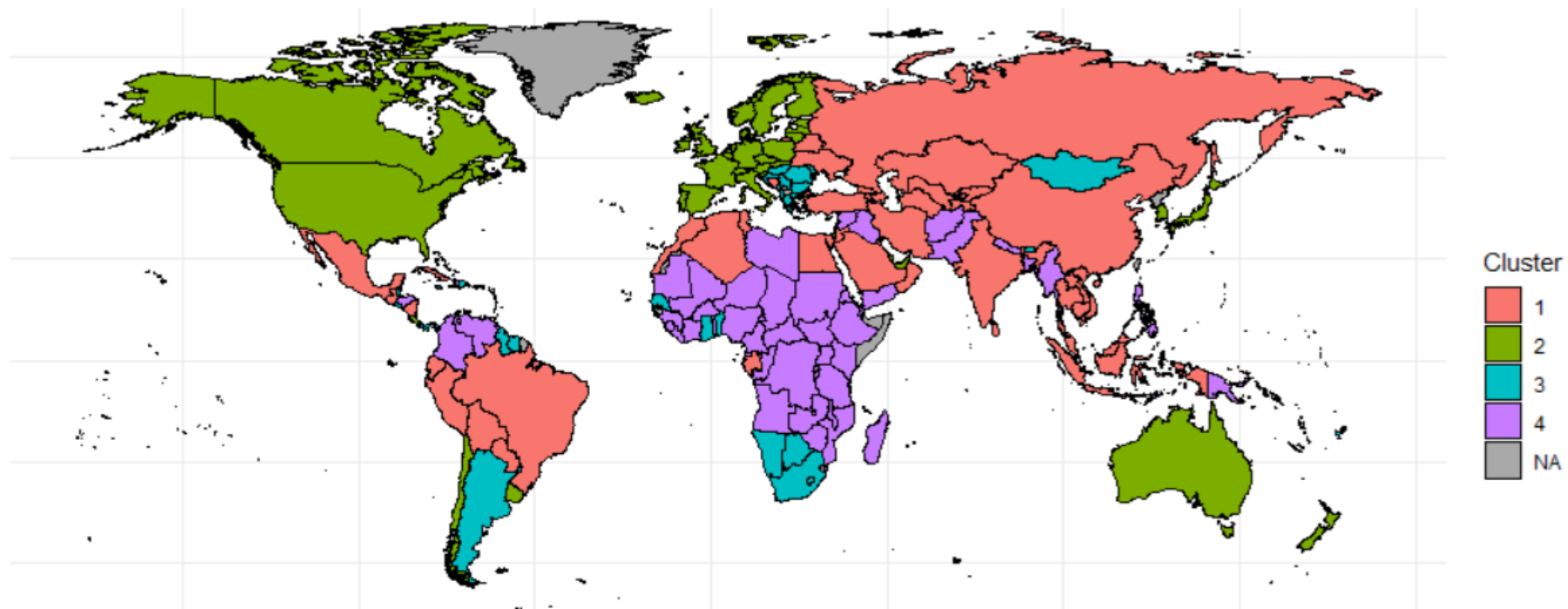
K-MEANS

Total within variance does not decrease much after 4 clusters, so we proceed with $K = 4$.

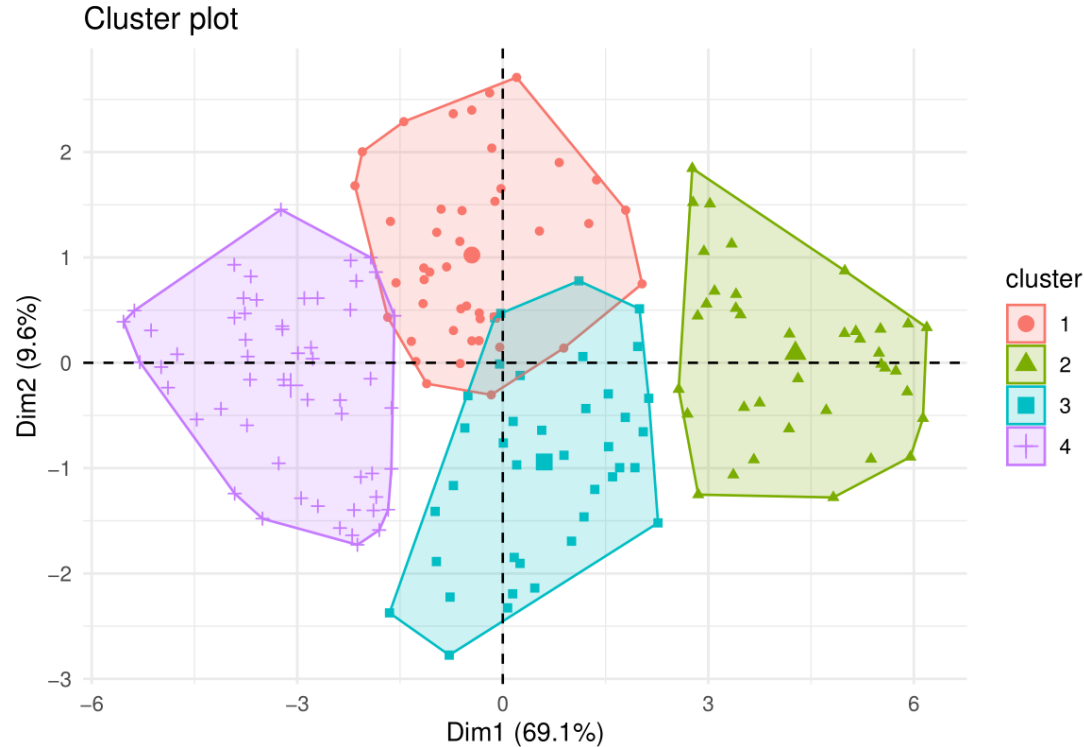


We find some correspondence between, for example, cluster 4 and Low HD countries, and cluster 2 and Very High HD countries.

MAP



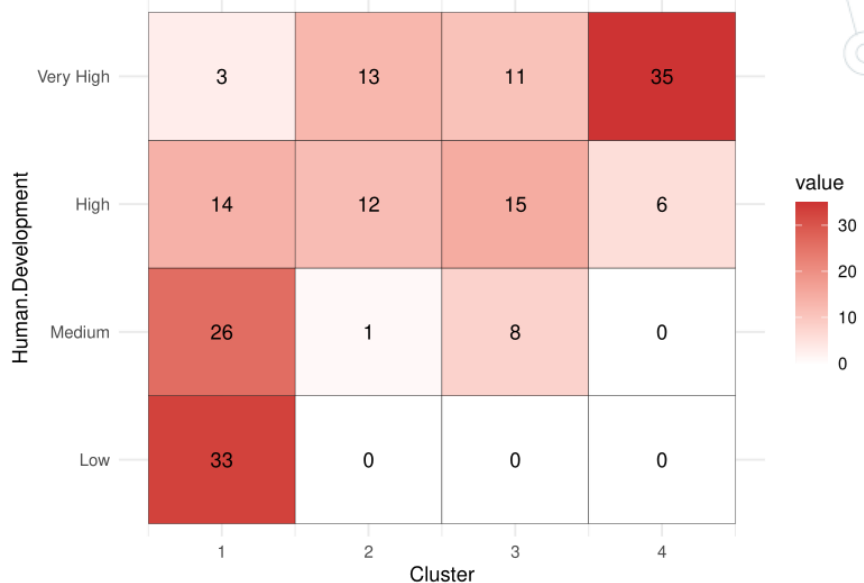
K-MEANS



While groups based on HDI were distributed along the horizontal axis, K-means identifies two clusters along the vertical dimension.

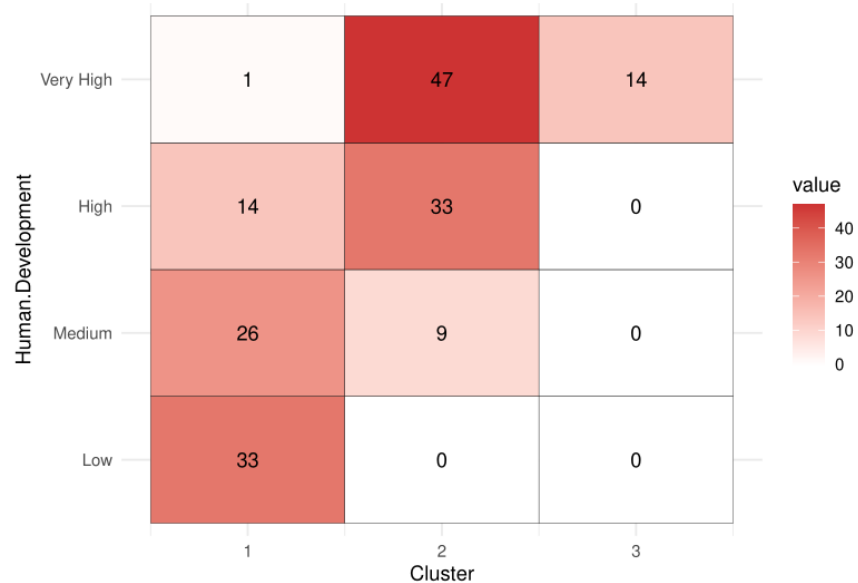
MODEL-BASED CLUSTERING

The algorithm suggested to use 4 clusters, but while this approach works well for Low HD countries (all grouped in cluster 1), it does not work for 3 Very High HD countries, that are put in the Low HD group. These countries are Barbados, Chile and Panama.

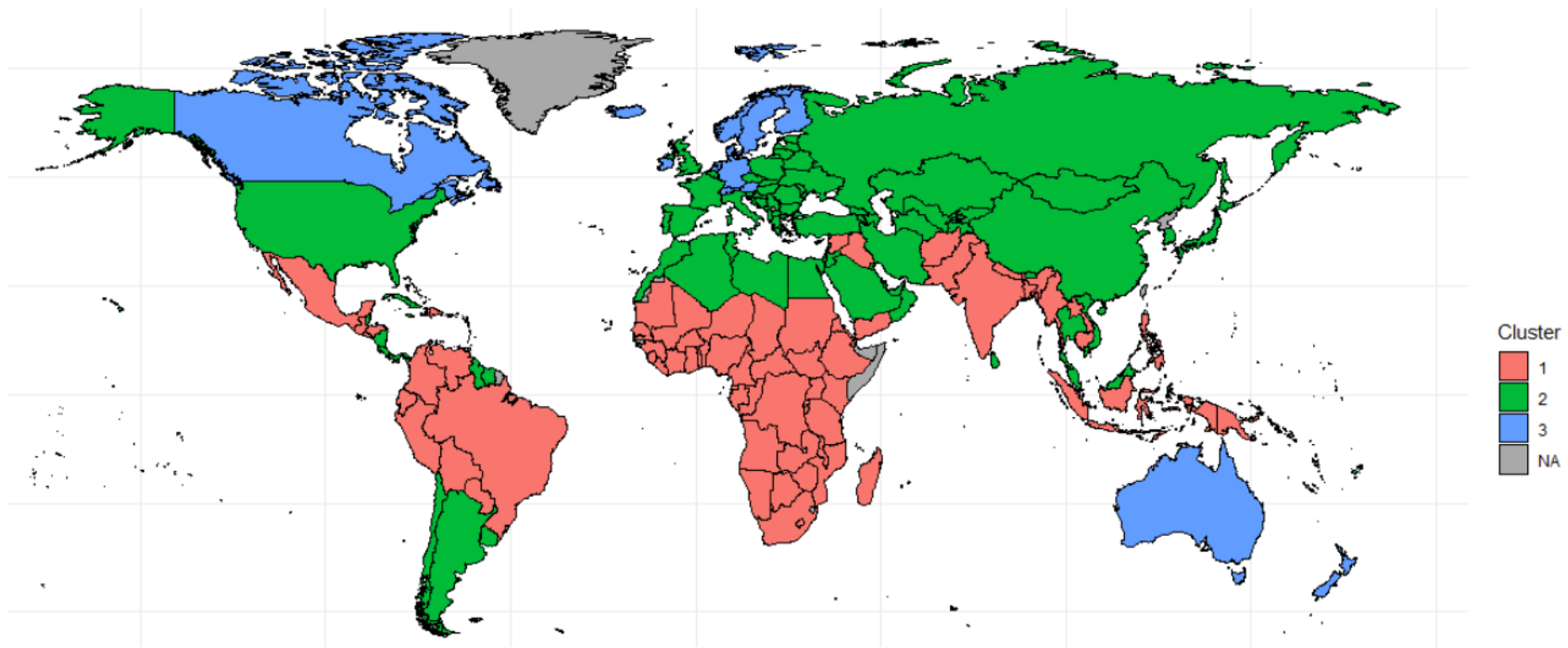


MODEL-BASED CLUSTERING

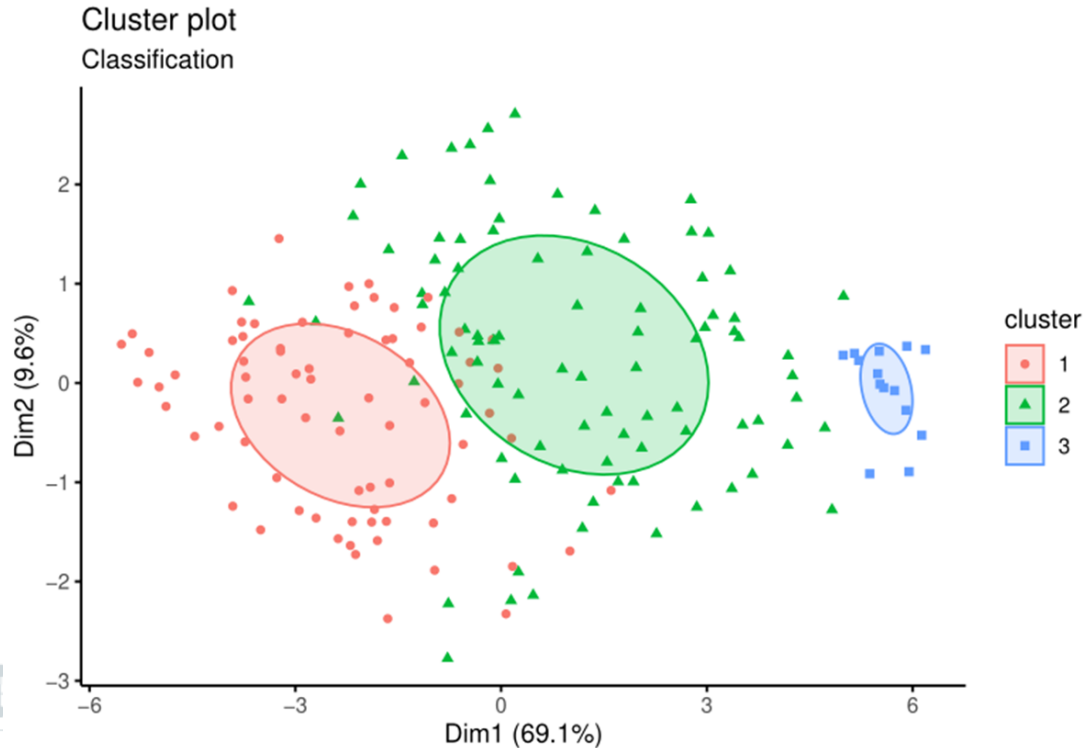
With three clusters, we obtain better results: all Low HD countries are still in cluster 1, together with most Medium HD countries; cluster 2 is related to High HD countries and cluster 3 to (Very) Very High HD countries.



MAP



MODEL-BASED CLUSTERING

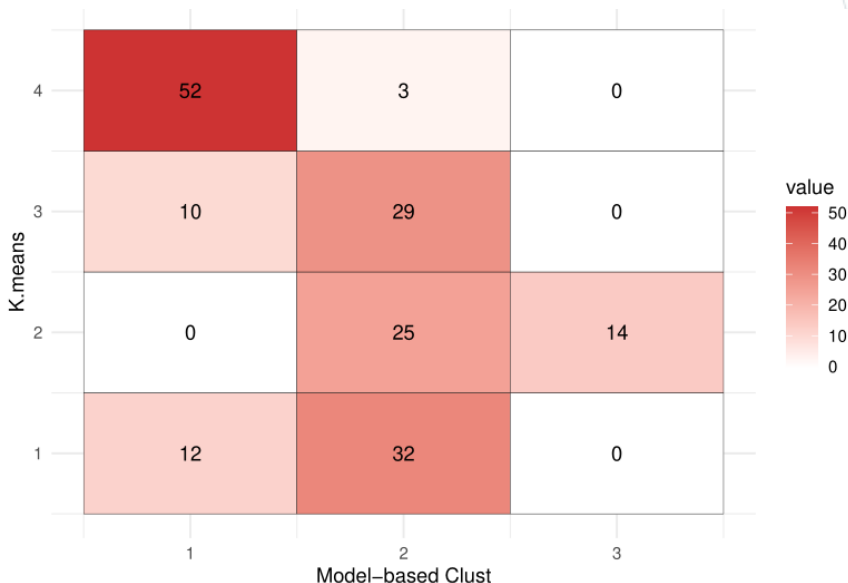


The grouping here is more similar to the one made according to HDI, even though we have three groups instead of four and the observations are more sparse.

COMPARISON

There is also much correspondence between the clusters obtained through the two different methods.

Even though model-based clustering gave simpler results, K-means clusters seem to be more balanced and informative.



Thanks for watching!



DATA SOURCES

- © *Fragility State Index:*
<https://fragilestatesindex.org/>
- © *Human Development Index:*
<http://hdr.undp.org/en/composite/HDI>

REFERENCES

- © *An Introduction to Statistical Learning*, James G. et al., 2013, Springer
- © *An Introduction to Applied Multivariate Analysis with R*, Everitt B. & Hothorn T., 2011, Springer