The role of policies and healthcare systems during the COVID-19 pandemic

Guillem Bonilla[a], Luis Rojo-González[a]

[a]*Universitat Politècnica de Catalunya, Barcelona, Spain*

Abstract

The global pandemic context we are living makes difficult whole live style we lived until the last year worldwide. It changed many of the activities we got used. However, what we could learn about this pandemic it is an important question to answer. On this way, we perform a multivariate analysis involving Clustering and Factorial analysis using number of infected, recover and death people since the first confirmed case in each country till the day of maximum number of observed cases for each kind of case, and link them with those variables related to healthcare systems reported by The World Bank. We find that there is an important outlier given by Unite States of America (US) and there are three different groups from each other. Also, we could draw a kind of map according to the way countries handle the situation and how their healthcare systems are. Thus, we can observe how variables behave through different groups where those related to healthcare systems and number of days to achieve the maximum number of cases are relevant only. Finally, groups are characterized by those: i) which their maximum level on cases are not achieved yet, ii) transition countries where healthcare systems are not relevant but their policies are, and iii) most damaged countries due to late and/or bad policies produced their current status whereas their healthcare systems make it worst.

Keywords: COVID-19, Clustering, Factorial analysis, Healthcare systems.

1. Introduction

At December 2019 coronavirus disease (COVID-19) emerged in Wuhan, China; was reported and few months after in most of countries we are living an pandemic of an (almost) unknown virus never observed before. Despite the drastic, large-scale containment measures promptly implemented by the Chinese government, in a matter of a few weeks the disease had spread well outside China, reaching countries in all parts of the globe (Fanelli and Piazza, 2020), countries which has been suffering human crisis as well as economical.

It has forced billions of people to shut themselves up in their homes given the lock-down suffered in almost all countries.

Thus, if we consider the entire world as an isolated system where any element may affects each other and therefore influencing the ensemble, many questions raises: Were we prepared to face a planetary challenge like this? Does government's handling way has a significant impact on virus spread? Are Spreading rhythm and healthcare system investment related?.

On this way, and following the questions stated above, this work aims to discover some patterns through analyzing how the so-called curves increase/decrease from one day to the next one by each kind of case and linking this behaviour with those variables related to investment on healthcare system getting a point of view from how the pandemic would evolves if we depend on healthcare systems.

This work is organized as follows: we describe the data we are working with in Section 2; Section 3 shows performed Cluster Analysis in order to check whether countries are similar in terms of their performance against the pandemic. Section 4 addresses dimension reduction issue applying Factorial Analysis to observe where countries are considering two latent factors: i) Pandemic behaviour and management from authorities and ii) Money invested on health-care systems; and analyzes how considered variables behave by each group to characterize them. Finally, Section 5 gives some conclusions and remarks towards what would we expect to happen of those countries whose first confirmed cases are more recently.

2. Dataset description

We use the time series reported by Johns Hopkins University (2020) for COVID-19 observed cases on daily-basis at may 23rd and The World Bank (2017) for healthcare systems variables.

a) Pandemic expansion:
   - Mean: The mean value of the increasing rate for each kind of case on a daily-basis.
   - Standard deviation: The mean value of the increasing rate for each kind of case on a daily-basis.
   - Days: The number of days since the first case was observed till the day of maximum observed cases.

b) Health investment:
   - Health exp per capita: Level of current health expenditure expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed during each year. This indicator does not include capital health expenditures such

Email addresses: guillembmontolio@gmail.com (Guillem Bonilla), luis.rojo.g@usach.cl (Luis Rojo-González)

as buildings, machinery, IT and stocks of vaccines for emergency or outbreaks.

- Health exp public pct: Share of current health expenditures funded from domestic public sources for health. Domestic public sources include domestic revenue as internal transfers and grants, transfers, subsidies to voluntary health insurance beneficiaries, non-profit institutions serving households (NPISH) or enterprise financing schemes as well as compulsory prepayment and social health insurance contributions. They do not include external resources spent by governments on health.
- Nurse midwife per 100,000: Nurses and midwives include professional nurses, professional midwives, auxiliary nurses, auxiliary midwives, enrolled nurses, enrolled midwives and other associated personnel, such as dental nurses and primary care nurses.
- Specialist surgical per 100,000: Specialist surgical workforce is the number of specialist surgical, anaesthetic, and obstetric (SAO) providers who are working in each country per 100,000 population.

3. Clustering

3.1. Partitioning clustering

Considered data is clearly characterized from what they measure. On this way, at first glance we would use Mahalanobis distance since variables are correlated, but we state Euclidean distance instead given two main reasons: i) variable selection was carried out, and ii) two merged data sets are balanced.

Thus, according to selected distance measure to use K-means algorithm (MacQueen et al., 1967) as well as Partitioning around Medoids (PAM) (Kaufman and Rousseeuw, 2009) can be performed; nonetheless, in order to get reliable results and enhance the interpretation of them we consider the last one as the algorithm to get desired groups. This algorithm use observations as representative element for each cluster we made. Thus, the might see how the groups are more or less. Then, once we decide which algorithm we will use, it is turn to decide how many groups we could recognize; for this, useful statistics are Silhouette (Rousseeuw, 1987), change on Total Error Sum of Squares (TESS) (Ward Jr, 1963) and pseudo-F (Milligan and Cooper, 1985) where the maximization of these is the goal; finally, we can consider the change in Total Within Sum of Square by changing the number of groups to support the decision.

Now, as we have state the measure for the distance and metrics to recognize how many clusters we could consider on a right way, we have to proceed in order to discard those possible outliers in the data, but how to discard them? this is an important issue which we are going to deal by using Factorial Analysis in next Section.

On this way, the implementations on R provided by Maechler et al. (2019) to get the PAM algorithm and Kassambara and Mundt (2017); Walesiak et al. (2011) to get the statistics; Table 1 shows the stated statistics for the first iteration where we realized *US* has to be considered as an outlier which adds noise to performed analysis, and Table 2 does the same for the second iteration by discarding this outlier.

Table 1: Performance indicator by changing the number of cluster to work with whole observations.

Cluster	ΔTESS	PseudoF	Silhouette
2	35.61	43.43	0.23
3	25.34	57.71	0.26
4	14.13	49.05	0.27
5	13.49	66.87	0.29
6	12.51	62.97	0.27
7	7.06	57.21	0.22
8	8.06	53.59	0.20
9	6.41	49.24	0.19
10	6.31	47.27	0.18

Table 2: Performance indicator by changing the number of cluster to work without those observations considered as outliers.

Cluster	ΔTESS	PseudoF	Silhouette
2	35.61	50.69	0.23
3	25.34	70.12	0.27
4	14.13	57.92	0.25
5	13.49	52.91	0.27
6	12.51	51.58	0.27
7	7.06	45.71	0.20
8	8.06	41.52	0.18
9	6.41	37.95	0.18
10	6.31	36.38	0.19

In the first iteration we can clearly see that a suitable number of groups would be five where medoids are given by *Honduras*, *Portugal*, *South-Korea*, *Spain* and *US* (notice that this last groups is only made by this country which is the ourlier). And, for the second iteration, the optimal number of groups would be three, where medoids are given by *Honduras*, *Portugal* and *Spain*.

Finally, Figure 1 shows how Total Within Sum of Squares for both analysis behaves in order to be helpful and which support the above decision due to kind of inflection points might be see at both number of groups stated above, nonetheless notice that the second iteration must be considered.

3.2. Significance checking

Then, the question is related to whether the groups are statistically different from each other. Thus, the typical procedure is checking multivariate normality assumptions, e.g. using Mardia test (Mardia, 1974; Korkmaz et al., 2014), to perform T^2 test for mean. The null hypothesis of Mardia test states that groups are different in terms of mean, skewness and kurtosis; nevertheless, in this case, Mardia test give us that none of them are multivariate normal distributed. As a consequence, we perform the Hotelling test based on pairwise permutation (Curran and Curran, 2018). The test provide us all pairwise groups are different from each other at almost 100% of confidence. On this

Figure 1: Screeplot for both instances. A1:=is the first analysis with whole observations. A2:=is the second analysis without those observations considered as outliers.

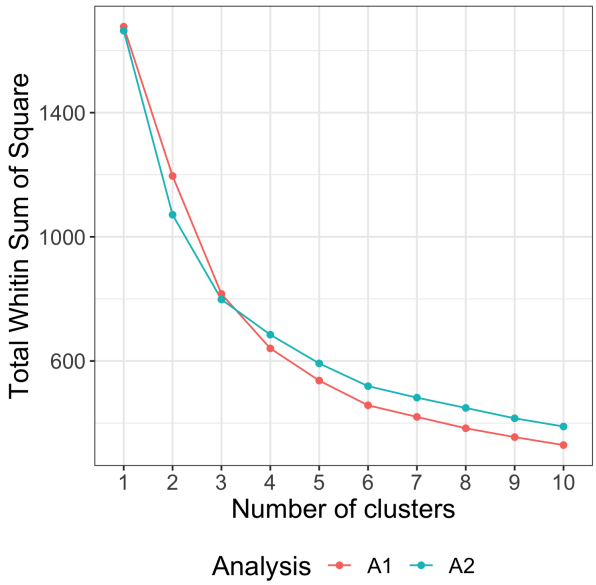
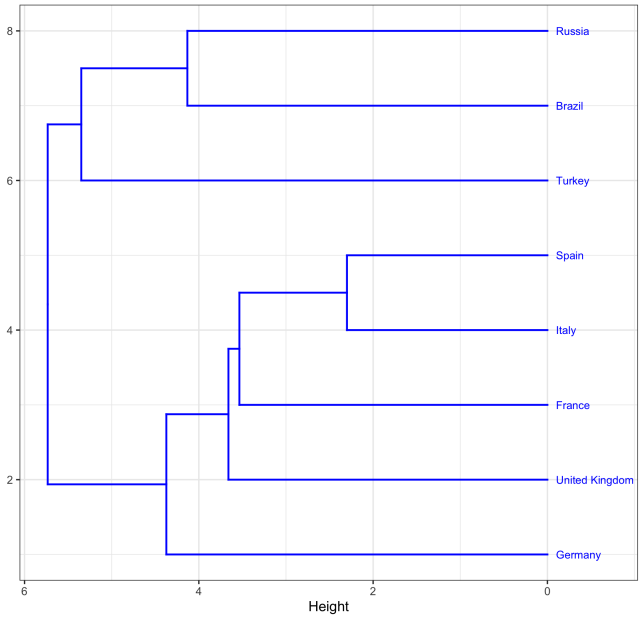


Figure 2: Dendogram for group three using single linkage given a cophenetic correlation of 81.93%.



way, we can claim that all groups are statistically different from each other and there are not artificial groups.

3.3. Hierarchical clustering

Usually, hierarchical clustering (Ward Jr, 1963) is applied as a first exploratory step when we do not have prior information regarding the dissimilarities structure of the data. However, we have applied it after determining the optimal number of cluster by PAM and check groups differences significance with the aim to dive in our knowledge about similarities of the countries within each group. We could see how they are inside as a hierarchical structure to identify some useful patterns that allow us to provide more accurate descriptions of these groups. In particular, we are interested in that group we consider the most affected, given prior information, by the pandemic.

On this way, Figure 2 shows the hierarchy of this group using Single linkage, with a 80% of cophenetic correlation coefficient (Sneath et al., 1973), where we can clearly see we could make a sub-group considering Spain, Italy, France, UK and Germany; countries which are currently the most damaged countries by the ongoing pandemic. And as we expected since our prior information, Spain and Italy are the most similar ones where the dissimilarity measure is the smallest one of the entire group.

4. Factorial analysis

Once we have decided how many groups are and which country belongs to each one of them, we work with a suitable dimension reduction method which enables us to visualize how data would be in 2D.

In this case, we use Factorial analysis, with its implementation in R given by Revelle and Revelle (2015), based on the prior knowledge that there are two latent factors related to: i) Pandemic behaviour and ii) Investment on healthcare systems.

In order to decide how we have to apply this analysis, it must be necessary to compute the degrees of freedom. As we are working with a data set with 13 variables and considering 2 latent factors, the analysis has degrees of freedom equal to $1/2((13-2)^2-(13+2)) = 53 > 0$ which implies the solution is not unique. Therefore, it is reasonable to use the *varimax* rotation to enhance the visualization (Härdle and Simar, 2015). The cumulative variance explained by the 2D representation is approximately 70%, so it can be assumed representative. We also accepted the hypothesis test that 2 factors are good enough.

This analysis conjoint with the representation in 2D such as Figure 3 shows, enable us to claim that: i) such as we state above, there is an outlier which corresponds to US, ii) we get well differentiated structure between the groups of countries, iii) first latent factor is the magnitude of the damage suffered by the countries, and iv) second latent factor is the per capita investment on healthcare.

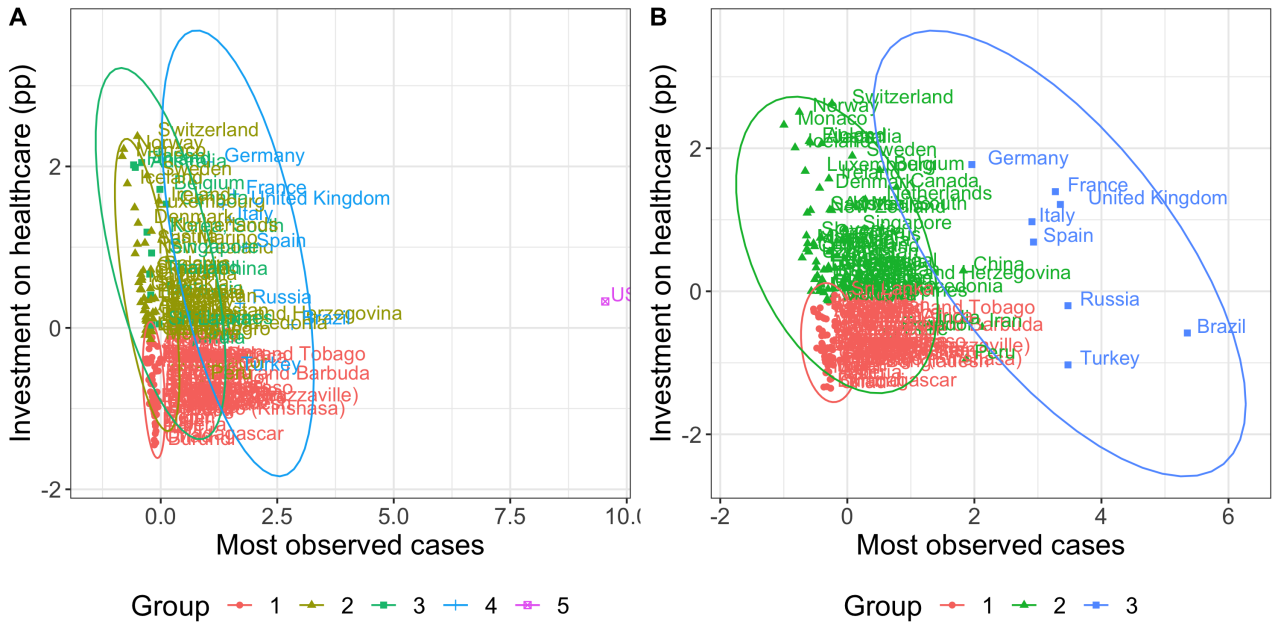
4.1. Variable analysis and world scenario

Finally, we perform a last analysis in order to enhance the conclusions we have till now. This analysis is related to how variables behave through groups made in Section 3 which we check they are different from each other.

On this way, Figure 4 shows density plots for each variable through different groups, where we can clearly see there are many of them with same distribution no matter the groups; but there are others with pretty different shapes.

Those variables that distribute different according to each group are: 1:=Health exp per capita, 2:=Health exp public pct, 6:=Elapsed days since first confirmed case, 7:=Elapsed days

Figure 3: Factorial analysis plot for first two factors, **A** shows the whole countries in the study meanwhile **B** shows the same analysis but discarding US which is considered an outlier.



since first death case, 8:=Elapsed days since first recovered case, 9:=Nurse midwife per 100,000, and 13:=Specialist surgical per 100,000.

Thus, it is clearly to see that those variables which make groups different from each other are those related to healthcare systems investment and, on the other hand, the number of days at which countries achieve their pick on each kind of cases. In other words, we could discard those variables related to the curve growing rhythm.

If we do so, we could interpret the groups as follows:

Group 1: corresponds to those countries which are not too damaged (yet) due to pandemic but they have not enough time suffering this pandemic and also not have good numbers in healthcare system investment.

Group 2: corresponds to those countries that are not too damaged due to pandemic but it is given their healthcare system and policies during this time which can be translated to a good handling from authorities.

Group 3: corresponds to those countries that are too damaged due to pandemic where late and bad policies generated their current status and their healthcare system make it worst.

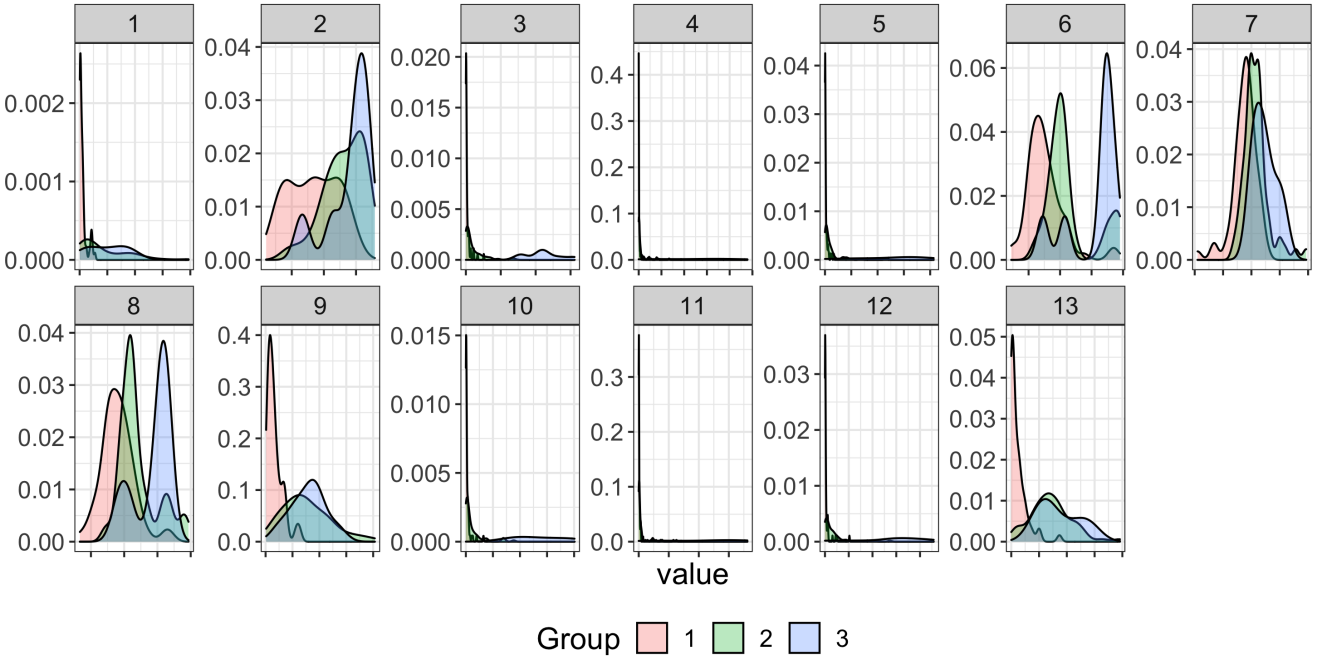
5. Conclusion

The aim of this work is to characterize countries against ongoing pandemic considering historical records related to healthcare systems investment and how the pandemic spread looking for patterns that enable us to see we were on a right way in terms of health systems according to money. To achieve it, and considering a Factor Analysis, the first latent factor is related to the information previous to the pick day for each country and, as second latent factor, those related variables to healthcare systems investment; both with the aim to link the way the pandemic behaves throughout those countries and whether it is related to their healthcare systems.

On this way, by applying the different clustering and dimension reduction methods, three main groups have been identified which contain countries with similar performances against the ongoing pandemic. Furthermore, policies adopted by each country as well as their investment on healthcare systems are relevant in how the pandemic behaves. Despite of the huge presence of countries belonging to group three in Europe is still and open question, policies adopted by different European countries have been similar but those countries of group two are there because of they have a more robust healthcare system. Nonetheless, countries of group three, such as Spain and Italy, are those that have suffered the most which could be given by late policies or even by the way people in these countries interact with each other, i.e. physical contact.

In terms of healthcare systems investment it seems to make the difference to group two and three; but the most important aspect is the way countries deal with the pandemic. This, because of we got that the number of days until countries reached their picks in each kind of observed cases distribute different by each group and also the per capita expenditure in health do including the number of professionals in health. This clarify the landscape towards the ongoing pandemic does not forgive anyone. Quality on healthcare systems are long-term measures which clearly help to both accelerate the recovery time and avoid the death probability. Even so, measures and policies adopted are pretty much necessary to stop the pandemic spread given the in-

Figure 4: Density plot for variables according to each group. Encoding is as follows: 1:=Health exp per capita, 2:=Health exp public pct, 3:=Mean confirmed cases by day, 4:=Mean death cases per day, 5:=Mean recovered cases per day, 6:=Elapsed days since first confirmed case, 7:=Elapsed days since first death case, 8:=Elapsed days since first recovered case, 9:=Nurse midwife per 1000, 10:=Standard deviation confirmed cases by day, 11:=Standard deviation death cases by day, 12:=Standard deviation recovered cases by day, and 13:=Specialist surgical per 1000.



fection rate and late actions would collapse the healthcare systems as good as they are.

References

Curran, J., Curran, M.J., 2018. Package 'hotelling' .

Fanelli, D., Piazza, F., 2020. Analysis and forecast of covid-19 spreading in china, italy and france. Chaos, Solitons & Fractals 134, 109761.

Härdle, W., Simar, L., 2015. Applied multivariate statistical analysis. 4 ed.. Springer. chapter 12. pp. 364–366.

Johns Hopkins University, 2020. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. `https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`. Online; accessed 20 may 2020.

Kassambara, A., Mundt, F., 2017. Package 'factoextra'. Extract and visualize the results of multivariate data analyses 76.

Kaufman, L., Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. volume 344. John Wiley & Sons.

Korkmaz, S., Goksuluk, D., Zararsiz, G., 2014. Mvn: An r package for assessing multivariate normality. The R Journal 6, 151–162.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA. pp. 281–297.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source).

Mardia, K.V., 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhyā: The Indian Journal of Statistics, Series B , 115–128.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Revelle, W., Revelle, M.W., 2015. Package 'psych'. The Comprehensive R Archive Network .

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65.

Sneath, P.H., Sokal, R.R., et al., 1973. Numerical taxonomy. The principles and practice of numerical classification.

The World Bank, 2017. World Development Indicators: Health systems. `http://wdi.worldbank.org/table/2.12#`. Online; accessed 20 may 2020.

Walesiak, M., Dudek, A., Dudek, M.A., 2011. clustersim package.

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58, 236–244.