Analysis of Covid-19 in South America: Relationship between public Health damages and Economical Impact in the region

Santiago Pozo Ruiz, Genesis Carrillo Granizo
School of Mathematical and Computational Sciences
Yachay Tech University
Urucuqui, Imbabura, Ecuador

***Abstract*—In the present work, we explain some of the techniques used to analyze a data-set containing variables of covid-19 and we try to compare them and to find some relation with the economic variables like GDP or Unemployment Rate. Besides, we try to use cluster analysis to find relationships between individuals (countries), it will make us able to lead to conclusions on which countries had better performance in handling with the pandemic with the best results in terms of public health and economic outcomes.**

***Index Terms*—SARS CoV2, Multivariate Statistical Analysis, PCA, Clustering, Biplot**

I. Introduction

A trending topic nowadays is, definitively, the Covid-19 (Coronavirus Disease), this affection, caused by the severe acute respiratory syndrome (SARS-CoV-2) is a respiratory disease which was detected in first place in December 2019 in Wuhan, China, its supposed origin are bats, and humans were infected by zoonosis, however, its real origin is still in discussion. Since then, the virus has spread to 215 countries and territories, infected more than 13.4 million people by July 14, 2020, killed approximately 580230 people according to Worldmeters [2], and put more than half of world population in lock-down to avoid more infections and the collapse of the sanitary systems. Even though, the pandemic has came across different epicenters, being China the first, Europe after that, followed by the USA, by today, Latin America, and specially, South America, has become the epicenter of the pandemic, and the situation in the region is becoming worse every day, for example, Peru has the highest crude fatality rate in South America (371 per million), followed by Chile (370 per million) and Brazil (349 per million). However, the highest number of deaths took place in Brazil (74,262), followed by Peru (12229) and Chile (7069). In fact, Brazil is the second most affected country in the world, having performed a lot less testing than countries like USA, India or the Russian Federation (first, third and fourth most affected countries in the world), this is a common factor among the countries in the region. That makes us think about the possibility of disastrous situation happening in Latin America, without even noticing it.

Nonetheless, the great majority of countries in the region are now thinking about reopening its economies after the lock-down, some leaders, like the ones from Brazil and Mexico showed themselves even reticent about the benefits of a lock-down, that is mainly because those weak economies will not handle a very extended paralyzation of the economy for an extended time. That could be happening due to several factors which are common in the region, like poverty, extreme poverty, people who live from informal economy, among others. Those factors have also made the control of the pandemic in the region extremely difficult, and lock-downs have not showed actually good results, being infections more common nowadays than three months ago, where lock-downs began in most of the countries in the region.

Then, some analysis are necessary to see how countries, mainly in the fields of economy and public health, will evolve in the next months. Those analysis will make us able to take better decisions based on previous data which aroused from previous decisions. The approach made in this work tries to analyze data of public health outcome from the countries in South America, as the economic outcome, trying to compare them and to visualize the effects on the pandemic in the economies of the countries in the region.

II. Materials and Method

A. Data-set

The data used in this analysis was retrieved from two sources; Our World in Data, which is a big repository that contains a lot of useful data resources, among them we found the Corona-virus Pandemic (COVID-19) [3], in this enormous data-set, we found some interesting variables like total cases by country, total tests by country, stringency index, hand-washing facilities, among others. We used some those variables for our analysis on public health. We also used data from Worldometer [2]

Besides, we used data from the International Monetary Fund (IMF) [1] in our analysis, we retrieved information about the GDP growth rate, that means, growth or recession, we also used information about the unemployment rate. Both variables, from 2019, 2020 and 2021.

B. Method

We divided all the data we collected into two groups of them in order to take advantage of them in their analysis. The first group are just public health's related variables, such as number of infected per million people, number of deaths per million people, number of test per thousand people, stringency index, among others, Here, we aim to analyze each country merely depending on its public health outcome, and grouping them depending on that factor. On the other hand, the second group is formed by variables from both types, related with public health and related to economy, we are adding variables like GDP per capita, or the GDP growth or recession of the countries by the end of 2020.

Before the beginning of our analysis, we had to filter our data due to the big amounts of data we found on the "Our World in Data" data-set, this process was carried out in the R software.

We used multiple techniques to analyze and lead to conclusion from the collected data, and the same process was applied to the two data groups. First, we performed cluster analysis to identify which countries would have a similar outcome in both scenarios, the economics and the public health, to determine the optimal number of clusters from the data, we used two different techniques, the elbow method, and group dissimilarity which in both cases provided the same result in terms of the number of clusters. After determining the optimal number of cluster, "partitioning" was performed. Then, we carried out hierarchical clustering on the same criteria to verify the results obtained with partitioning.

Later, we performed an analysis on biplot generated from PCA in order to compare the relations between variables an individuals and its contributions to the principal factors of the model. The variation of the biplot used was the Classical Biplot (PCA), this task, unlike all the previous tasks which were performed in R, was performed in the MATLAB extension "MultiBiplot".

III. Results and Discussion

A. Public Health Situation

*1) **Clustering**:* First, we performed partitioning clustering, in which we used two different methods to determine the optimal number of clusters to be used. We used the elbow method and the group dissimilarity techniques.

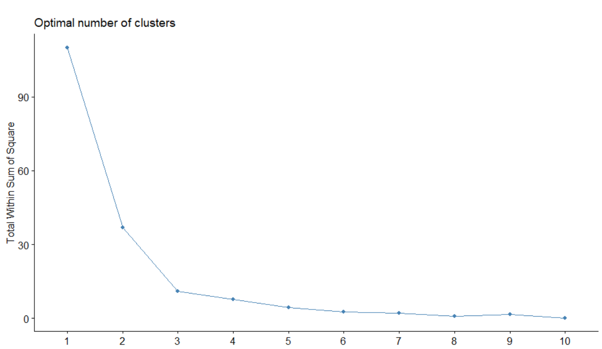

Fig. 1. Elbow method for selection of an optimal number of clusters

We notice that in both pictures, Fig. 1. and Fig. 2. that the function stabilizes at point three, thus it is our optimal

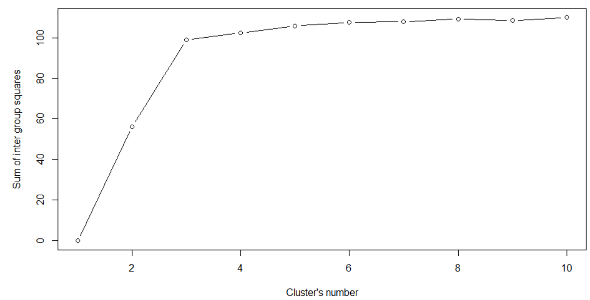

Fig. 2. Group dissimilarity for selection of an optimal number of clusters

number of clusters to use in our clustering process, k-means in this case.

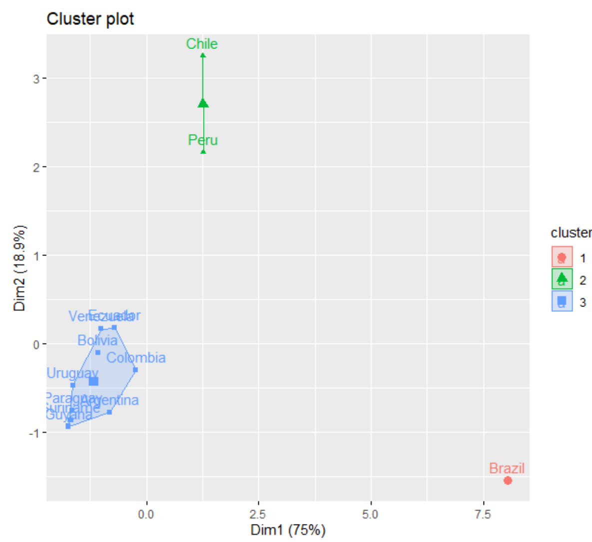

Fig. 3. K-means on public health data

Clusters	Country
Cluster 1	Brazil
Cluster 2	Chile, Peru
Cluster 3	Argentina, Bolivia, Colombia, Ecuador,Guyana, Paraguay, Suriname, Uruguay, Venezuela

TABLE I
Groups formed by K-means

In Fig.3. and Table I we can appreciate the clustering performed by the K-means algorithm, in which we can differentiate three groups, where it is noticeable that Brazil is forming a group alone, that is understandable due to the situation of Brazil being the second most affected country in the world by the pandemic. Peru and Chile are in another group, indicating that they are affected in the same proportion, and the most affected countries in the region after Brazil. In the third group we have the rest of the countries, which is reasonable knowing that Brazil, Peru and Chile are among the top ten countries in the world with the worst numbers of infected people. Probably, if we add one more group, we will have countries like Ecuador, Argentina or Colombia in it, and the rest in another group.

Then, we did a hierarchical clustering to confirm the results obtained with partitioning. In Fig. 4. we notice that Venezuela could be separated from the rest of the countries in the third group we just mentioned, however, Venezuela is there just because the data provided by the government on tests which is not very trustworthy. On the other hand, in Fig. 5. we can confirm the results obtained by partitioning, with three groups with the same individual in each group.

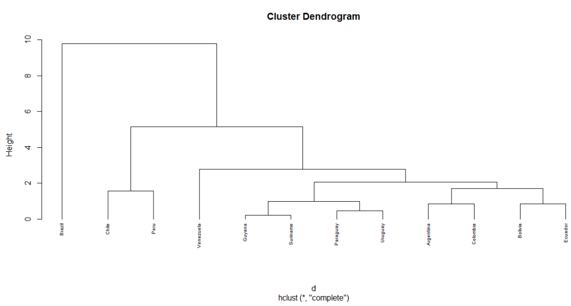

Fig. 4. Hierarchical clustering

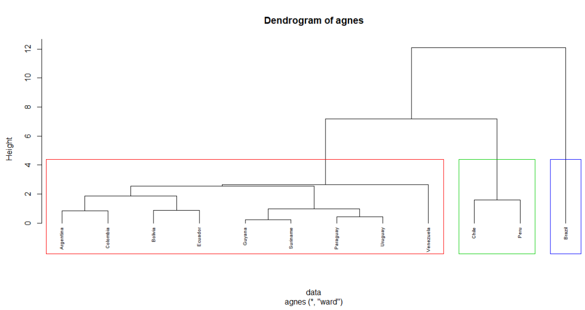

Fig. 5. Cut Hierarchical clustering

*2) **PCA and Biplot**:* To explain all the variability of the data-set related to public health outcome of the countries, we just need two dimensions or components extracted from PCA, those two dimensions explain almost 94% of data variability, more than enough. That can be seen in Fig. 6.

```
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  7.503165e+00           7.503165e+01                          75.03165
comp 2  1.893258e+00           1.893258e+01                          93.96423
comp 3  4.487209e-01           4.487209e+00                          98.45144
comp 4  1.048777e-01           1.048777e+00                          99.50022
comp 5  3.801222e-02           3.801222e-01                          99.88034
comp 6  9.276344e-03           9.276344e-02                          99.97310
comp 7  2.305205e-03           2.305205e-02                          99.99615
comp 8  2.991714e-04           2.991714e-03                          99.99914
comp 9  8.554468e-05           8.554468e-04                         100.00000
comp 10 1.343495e-32           1.343495e-31                         100.00000
```

Fig. 6. Dimension's cumulative percentage of variance

Fig. 7. and Fig. 8. obtained from a Principal Component Analysis help us to explain how the individual behave and also to explain what we see in the previous clustering and in the following biplot. In Fig. 8. we can notice that variables which contribute the most to the first two dimensions are "active_cases", "total_cases", "deaths", among others. We also notice in Fig. 7. that the individual which contributes the most to the variability of data is by far Brazil, and that is why this individual forms a group alone and is that far away from the rest in the biplots, second and third best contributors are Chile and Peru, as expected.

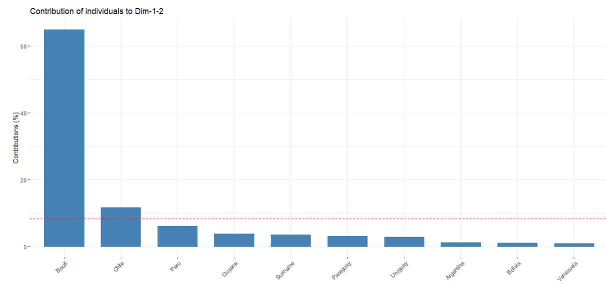

Fig. 7. Individual's contribution to 1 and 2 dimensions

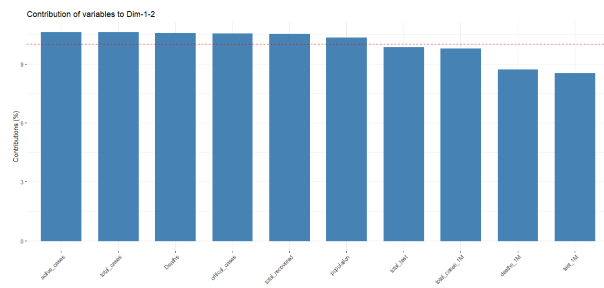

Fig. 8. Variable's contribution to 1 and 2 dimensions

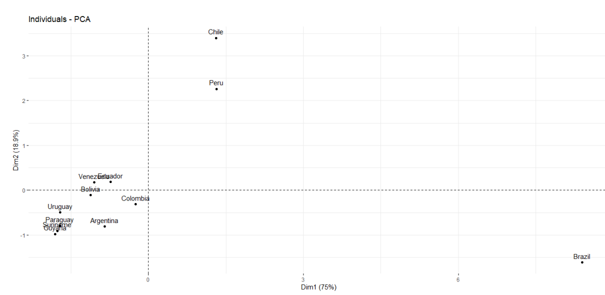

Fig. 9. Individual's Biplot

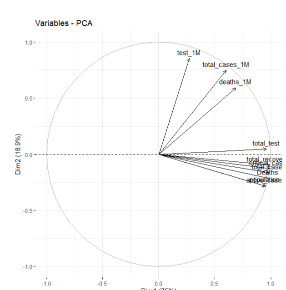

Fig. 10. Variable's Biplot

It is noticeable, from Fig. 11. that countries like Uruguay, Paraguay, Suriname or Guyana were the least impacted countries of the region in terms of public health outcome, although they were classified in the same group with other countries like Ecuador or Colombia, those countries have reasonable numbers in terms of total cases or total deaths, but that is due to the lack of testing, they are inversely correlated. Also, although Brazil is by far the country with most cases in the region, Chile and Peru are more affected in terms of infections and deaths per million, in terms of test per million Chile and Peru also have better performances than Brazil, although they have performed a higher number of total tests.

It is important to mention, as we can see in the biplot, that total tests and deaths have a high correlations, that might be due to the fact that the more test you take, the more identified deaths you have.

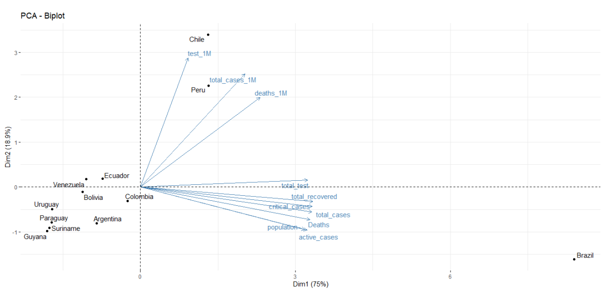

Fig. 11. Individual and variable's Biplot

B. Economic Situation

*1) **Clustering**:* We carried out the same steps to analyze the economic outcome thorough the pandemic for the countries in South America, then for clustering we first determine the optimal numbers of clusters to use, with the same methods, elbow and group dissimilarity. The variables added to the economic study were the GDP variance in 2019, with respect to 2018, the same with 2020, and 2021. The same for unemployment rates.

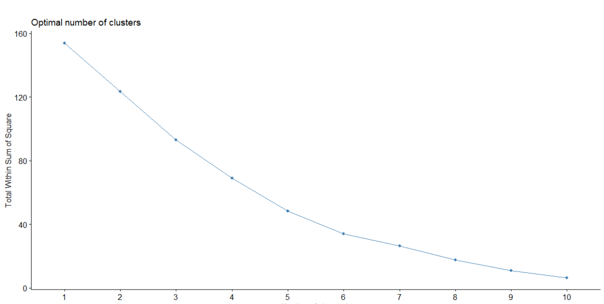

Fig. 12. Elbow method for selection of an optimal number of clusters

From Fig. 12. and Fig. 13. we can see that, for the same reasons mentioned before, the optimal number of clusters is five, and with that number of clusters we performed partitioning.

We realize, from Fig. 14. that when economic takes part, Venezuela is alone, as expected, but we also see Brazil and Chile forming groups alone, although Brazil could be in a group alone because of its public health situation, we can not actually see a reason of Chile being in a group alone.

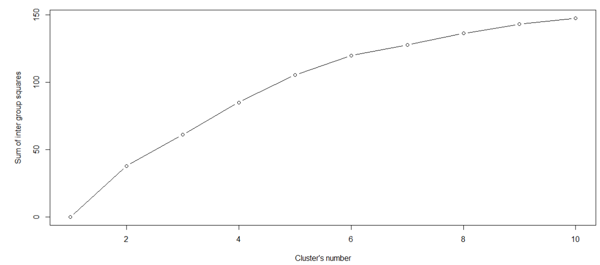

Fig. 13. Group dissimilarity for selection of an optimal number of clusters

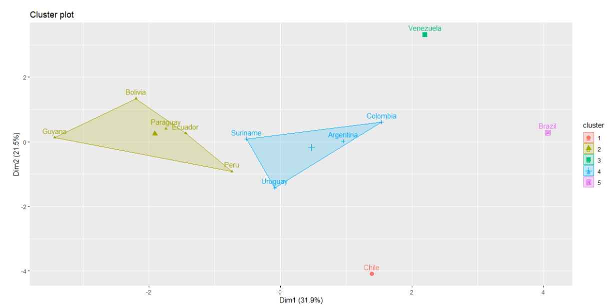

Fig. 14. K-means on public health and economic data

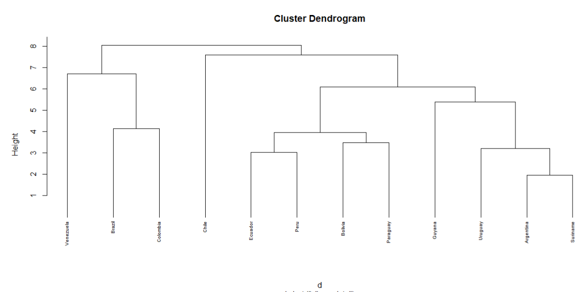

Fig. 15. Hierarchical clustering on public health and economic data

We can start explaining why Chile is alone in a group by looking at Fig. 16. from hierarchical clustering which throws slightly different results from the ones obtained in partitioning. That "incongruence" is due to the quality of representation of the data in the first two dimensions, in our partitioning plot, we see that data is plotted just over two dimensions which explain just 53.2% of the variability of the data, that is low, and that is why hierarchical clustering and partitioning show those differences in the results. We could say that hierarchical is more accurate and take some assumptions from it.

In the dendogram from Fig. 16., we can differentiate five groups, we explained why Brazil and Venezuela could be in groups alone. Then, Bolivia, Ecuador and Peru are in another group, those countries could be recognized as the ones with the worst public health outcome after Brazil, and with the worst economic outcome after Venezuela, but combining those

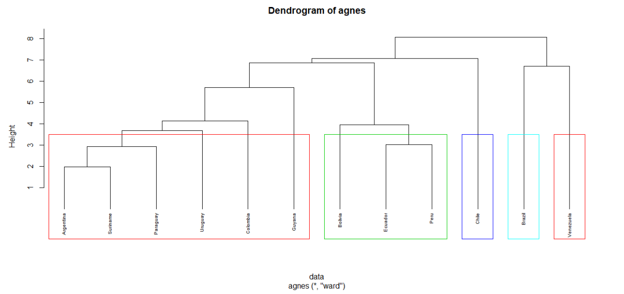

Fig. 16. Cut Hierarchical clustering on public health and economic data

factors and treating them as an only one. We can notice in another group countries like Argentina, Suriname, Paraguay, Uruguay, Colombia and Guyana, which can be labeled as the ones less affected in both ways, public health and economic, however, that is not equivalent to say that they are not affected at all, they are, but in a smaller proportion than the ones mentioned before. The case of Chile could be explained by analyzing the contributions in PCA.

*2) **PCA and Biplot**:* In opposite with the data about public health, the dataset used this time will need more dimensions to explain more variability. Dimensions one to four can explain about 85% of the variability of the data, which is enough, so to successfully analyze this dataset, we should take four dimensions.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.46047714	31.86055102	31.86055
comp 2	3.00687959	21.47771135	53.33826
comp 3	2.39878845	17.13420325	70.47247
comp 4	2.10042334	15.00302389	85.47549
comp 5	0.78164738	5.58319555	91.05869
comp 6	0.52249299	3.73209278	94.79078
comp 7	0.34874214	2.49101528	97.28179
comp 8	0.16234724	1.15962317	98.44142
comp 9	0.11910152	0.85072515	99.29214
comp 10	0.09191081	0.65650576	99.94865
comp 11	0.00718939	0.05135279	100.00000

Fig. 17. Dimension's cumulative percentage of variance

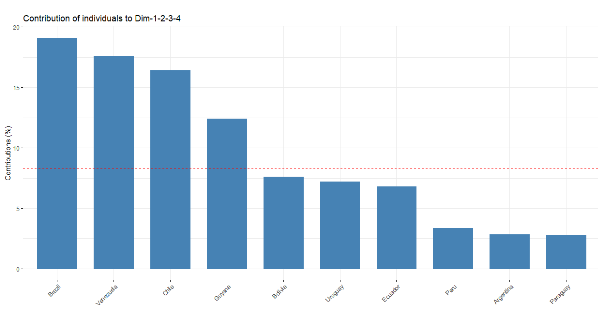

Fig. 18. Individual's contribution to 1 and 2 dimensions

In Fig. 18. we can see that the best contributors to the four dimensions we are analyzing are Brazil, Venezuela and Chile, and that is why they appear alone when doing clustering. However, it is not still very clear why Chile is alone exactly, that can be answered by looking at Fig. 19. and Fig. 20. In the second one, we can see that Chile is a great contributor of the second dimension, knowing that and looking at Fig. 19.,

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
total_cases_per_million	2.601887661	15.569211606	13.2459015	0.006518038	9.37555094
total_deaths_per_million	1.702560534	7.674904351	26.3604728	0.654335710	0.01691510
total_tests_per_thousand	5.131378723	15.863613373	0.1860695	6.309352695	7.78277181
stringency_index	2.320745764	11.383378993	5.0726534	14.936263965	0.68360352
population	11.898874863	0.248163193	2.4787724	15.232292186	0.01342484
population_density	1.689458594	0.520730619	10.5596249	5.914771181	59.65191905
gdp_per_capita	8.971619002	8.130936383	5.2031376	8.174703610	0.03619918
extreme_poverty	0.008777078	11.998684938	17.9849677	0.141531445	7.89656681
gdp2019	3.189976706	10.773785758	0.5827886	18.118435801	3.84470899
gdp2020	7.407166768	0.203230888	2.6500251	10.016891826	2.44090509
gdp2021	4.250169167	16.821398936	0.1706715	11.406426852	4.64357435
unemRt2019	13.036000026	0.005715468	11.8209661	4.907403104	0.32479080
unemRt2020	18.662088955	0.802554956	2.2859818	1.309830745	0.74741193
unemRt2021	19.129296157	0.003690538	1.3979671	2.871242842	2.54165760

Fig. 19. Variable's contribution to each dimension

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Argentina	1.85517497	6.249364e-04	9.609963043	1.265705e+00	1.62716242
Bolivia	9.82906561	5.406771e+00	12.848866294	5.634081e-04	21.05257459
Brazil	33.68839766	2.421238e-01	3.612887123	3.272344e+01	0.43590818
Chile	3.91829916	5.059297e+01	2.325598193	1.009598e+01	10.82537483
Colombia	4.82289292	1.111398e+00	0.005524259	1.833390e+00	10.52172170
Ecuador	4.22459492	2.121838e-01	22.478642716	3.882972e+00	40.67277814
Guyana	24.00162420	5.075237e-02	5.998878986	1.283053e+01	2.75259188
Paraguay	6.16182516	4.888293e-01	1.859012681	2.873009e-02	0.32823097
Peru	1.10873436	2.550943e+00	11.313260466	1.643325e-01	0.01196416
Suriname	0.53591944	2.024174e-02	5.627271607	9.684645e-01	1.33030001
Uruguay	0.01475642	6.127628e+00	23.494482909	5.425327e+00	5.19976150
Venezuela	9.83871519	3.319554e+01	0.825611725	3.078056e+01	5.24163162

Fig. 20. Individual's contribution to each dimension

we see that among the best contributors of dimension 2 are total cases per million, total tests per thousand, GDP 2021, and among others which are very important for Chile. So, we can name dimension 2 as the responsible for making it stay in a group alone.

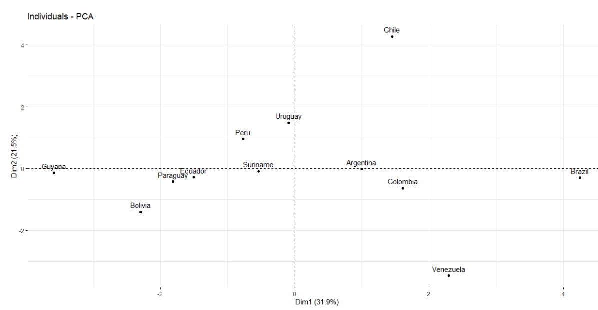

Fig. 21. Individual's Biplot

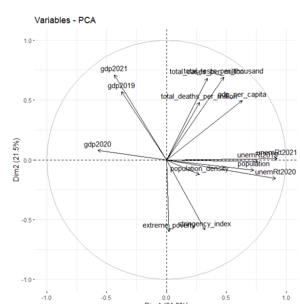

Fig. 22. Variable's Biplot

By looking at our biplots, specially at Fig. 23. we can lead to many conclusions, such as the inverse correlation, as expected, between stringency index and the GDP for each year, that is because the more strict are the lock-downs, the more severe is the damage to economy. Also something more or less obvious is the inverse correlation between unemployment rates and

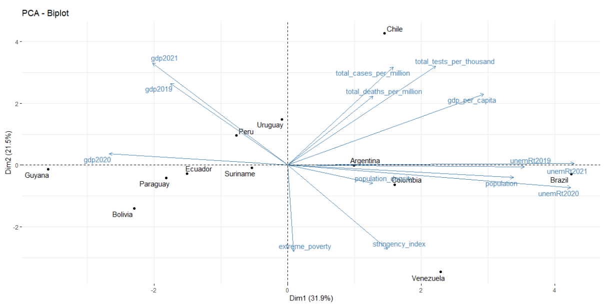

Fig. 23. Individual and variable's Biplot

the GDP's. Besides, we see that extreme poverty is directly correlated with stringency index, which makes sense.

A very remarkable fact is that variables like the cases per million, tests per thousand or deaths per million are not correlated with the GDP's, and although they are correlated with unemployment rates, those correlations are almost nonexistent.

For the variables related to public health, we can replicate the results from the previous analysis from the public health only data-set. For the economic variables, we can remark Venezuela, Brazil, Argentina and Colombia will be the countries with the slowest economic recuperation in 2021 after the pandemic, however, countries like Chile, Suriname, Peru or Uruguay will recover faster than the others, with the last two mentioned as the fastest.

In the same way as contrasting countries with their economic recoveries, we can contrast them with their unemployment rates by 2021, being the highest the ones from Brazil, Venezuela, Argentina and Colombia. And the ones with faster recovery are Chile, Suriname, Peru and Uruguay.

IV. Conclusions

Stringency index should be a variable which has inverse correlation with the total of cases and total of deaths, however, in our region that is not the case, and it has no correlation with that kind of variables, and also those variables did not actually have an appreciable correlation with GDP's. Knowing that, we can lead to the conclusion that lock-downs (stringency index) have not had a real effect in the public health outcome in our region, that might be because of the necessity of people to perceive some income through informal economy which is very recurrent in our region. Nonetheless, and unfortunately, the lock-downs actually had a severe effect in unemployment rates and growth of GDP, leading our region to an economic recession.

Although lock-downs seemed to work in countries like China, or in Europe, it does not seem to work in South America, just because our conditions are different, and there are not a lot of formal employment places, and people just need to perceive incomes in order to survive. We need to look further in the contingency plans applied in South America and look for alternatives, or the situation could just become worse each day.

References

[1] Imf.org. 2017. IMF Data. [online] Available at: https://www.imf.org/external/datamapper [Accessed 14 July 2020].

[2] Imf.org. 2017. Worldometers Data. [online] Available at: https://www.worldometers.info/coronavirus [Accessed 14 July 2020].

[3] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]