STEM Fellowship Big Data Challenge 2020

Combining Unsupervised and Supervised Learning Methods to Investigate Potential Country Development Factors Contributing to COVID-19 Progression

Feifei Li[1*], Mian Li[1], Jiaxiong Jiao[1], and Jun Leng[1]

[1]University of Toronto
[*]ff.li@mail.utoronto.ca

June 1, 2020

Abstract

To have a deeper understanding of what might cause the various levels of COVID-19 transmission figure R0 in different countries, we collected 144 countrys' GDP, population density, etc and their corresponding COVID-19 R0 value. By performing a Principal Component Analysis (PCA) and decision tree analysis, we were able to identify obesity rate and population as contributing factors correlated to COVID-19 transmissibility. Future research based on our preliminary results can focus on establishing the relationship between obesity and transmissibility of COVID-19 which may provide prospects in population health monitoring and disease control policy making.

Keywords
COVID-19, PCA, Basic reproductive number, Unsupervised learning, Supervised learning, Decision tree, Linear regression, GDP

1 Introduction

At the beginning of January 2020, a novel disease was identified in Wuhan, China and it soon started spreading across the world. Within 13 days, the disease was able to spread to another country and soon after, the outbreak escalated into a global pandemic (WHO, 2020)[6]. To date, there are more than six million confirmed cases of COVID-19 patients and over three hundred thousand people have died from this disease (JHU, 2020)[3]. With the ferocious COVID-19 outbreak, researchers around the world have come together to gain a thorough understanding of this disease in order to mitigate the current spread of COVID-19 virus and more importantly, to minimize the possibility of, if not prevent a similar epidemic in the future. With the current study in COVID-19 progression, we are able to identify some useful correlation to population characteristics which could be valuable information in contribution to disease control strategy, that is understanding how different factors may impact the current pandemic provides crucial insight for decision makers of governments in terms of how to prevent a similar catastrophe from happening again. In the current study, many population variables are taken into sensible consideration. Via principal component analysis (PCA) and decision tree analysis, we were able to narrow down from potential contributing population statistics including GDP, tourism, death from smoking, Obesity rate, and health expenditure to only two. It is hypothesized that at least one of these factors could potentially be correlated with the COVID-19 pandemic.

2 Materials & Methods

Data set:
Consist of information retrieved online regarding global GDP, Health Expenditure, Death from Smoking, Obesity Rate, Population Count and Tourism. We also obtained data regarding the Basic Reproduction number in order to perform PCA.

Basic Reproduction Number (R_0):
Quantitative measurement of transmissibility of a contagious disease that is the average new infection arising from a single infected case. When its value is less than one, it means the number of infected by this disease will decline. When larger than one, it means number of infected will go up. When its value is equal to 1, the number of infected will remain unchanged for that period. R_0 used in this analysis was calculated by the exponential growth (EG) method using the *R0* package(Obadia et al., 2012).[5] A mean serial interval **(SI)** of 3.96 days and its standard deviation of 4.75 days estimated by Du, *et al.* in China as of February 8, 2020 were used in estimating R_0.[11]

GDP:
Quantitative measurement of national GDP taken for 2018 in terms of US dollar.[1]

Government Health Expenditure:
Quantitative measurement of annual health expenditure for each nation for the year of 2017.[2]

Death from Smoking:
Quantitative measurement of people who died from smoking for the year of 2017.[9]

Obesity Rate:
Quantitative measurement of number of adults deemed as obese for the year of 2016.[7]

Population Count:
Quantitative measurement of a countries population for the year of 2019.[8]

Tourism:
Quantitative measurement for number of tourism arriving from each country during 2016.[10]

Principal Component Analysis (PCA):
Used to analyse the six variables mentioned above where they are used as data in constructing the plot. The data points were then grouped based on the R_0 level.

Decision Tree:
The response was a categorical variable that divided the sample, in this project, the countries, into 2 groups: Controlled, in which countries have R_0 less than 1, and Uncontrolled, where countries have R_0 higher than or equal to 1. All the 6 numerical variables are the potential predictors for the tree. 75% of the sample (countries) were selected from our data set for training, and 25% were used to validate the model.

Linear Regression:
After identifying possible correlating features through the classification tree, two linear regressions were performed using the variables, which were to be compared against R_0.

3 Results

For the first round of PCA, the result is shown through Figure 1. We used six country indicators as variables to construct the data and grouped the countries based on the R_0 level. As shown in Figure 1, all the data points are clustered around the top left corner with a few outliers. Figure 2 presents the loading score of each principal component. This represents how much each original factor weighs in the new component with PC1 explaining the most amount of variation.

The variables ***population*** and ***obesity*** were selected by the decision tree, in which all the leaves have plausible error rates: 11.8%,25.0%, 6.0%, from left to right. The accuracy of the prediction model, validated with the test set, was evaluated to 83.3%. The sensitivity and specificity were 0.5714 and 0.8966, respectively. The ROC curve (Figure 4) presents the sensitivity against the false positive rate for various classifiers. Another round of PCA was performed using the 2 variables selected by the tree (Figure 5.); there was still no obvious separation between groups of different R_0 levels. Two linear regressions performed using *population* and *obesity* as predictors respectively, R_0 as the target variable for both. The R^2 for the linear regression of R_0 against population is 0.0017, and the *p-value* for this regression is 0.6216. The R^2 for the linear regression of R_0 against obesity is 0.0014, and the *p-value* 0.6558.

4 Discussion

To gauge the severity of the outbreak in affected nations, we selected *basic reproduction numbers* (R_0) as our epidemiology indicator. R_0 is affected by various factors including human-human interaction, transmission mechanisms, seasonality.(Delamater et al., 2019) R_0 can be used to interpret the growth dynamics of a contagious disease, which reflects the effectiveness of disease control policy in the region of effect. In the current COVID-19 setting, R_0 is one of the apex measurements, logically illustrating the progressional and consequential aspects of the pandemic. However, there are various algorithms to estimate R_0 such as the AR

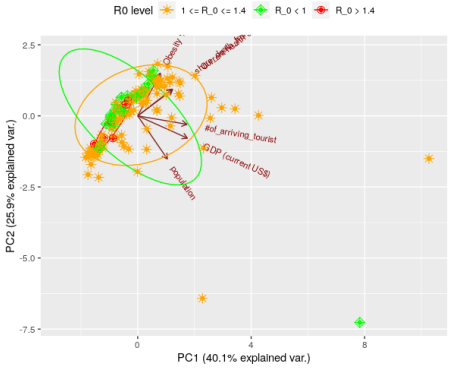

Figure 1: **PCA plot for all six variables:** a cluster of data points can be seen, but little to none separation of groups of different R_0 levels is present.

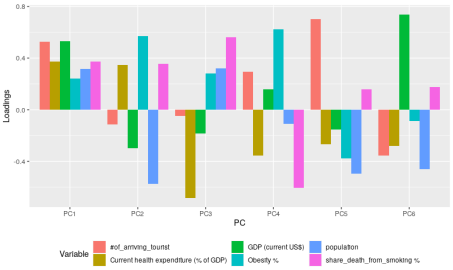

Figure 2: **Loading plot:** Loading scores for each component of the PCA on all the six variables.

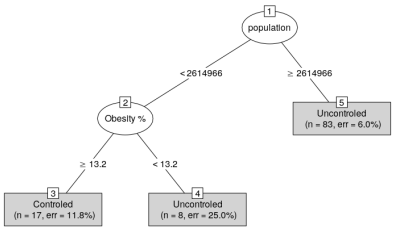

Figure 3: **Split of the decision tree:** Obesity and population were selected as a result of minimizing the impurities of terminal nodes.

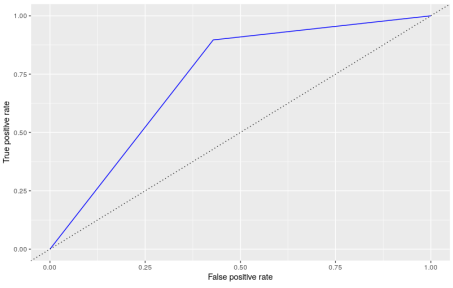

Figure 4: **ROC curve:** True positive rates against false positive rates for various cut points. The curve not being a straight diagonal line suggests its validity as a prediction model.

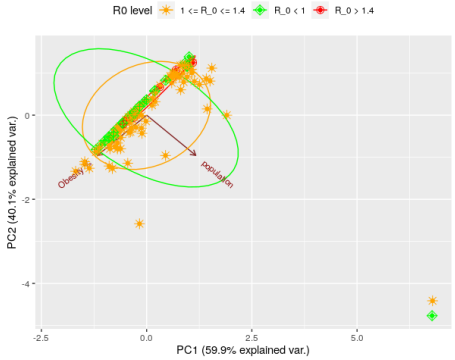

Figure 5: **PCA plot for the selected variables:** The cluster of different R_0 groups persists.

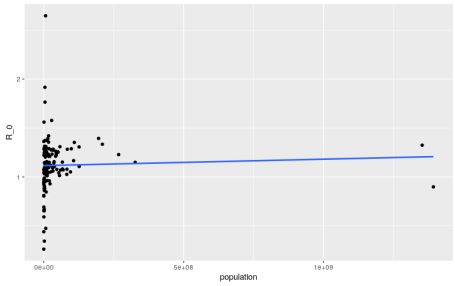

Figure 6: **Linear regression of Population and R_0:** Data points cluster and no obvious correlation could be observed.

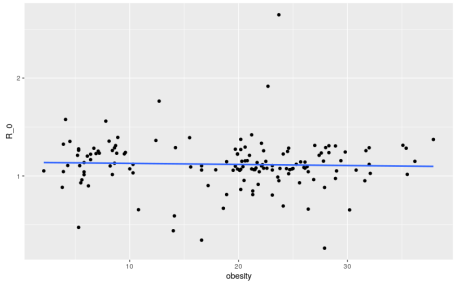

Figure 7: **Linear regression of Obesity and R_0:** Data points spread and no obvious correlation could be observed.

(Attack Rate) method, ML (Maximum Likelihood) method, and Sequential Bayesian method (SB). Given the limited epidemiological statistical information publicly available, we did not have data that any of these algorithms entails, and hence applied the EG method. Although we obtained sound estimates of R_0, a few of them might overestimate the true level of COVID-19 in countries where exponential growth of the virus never occurred due to the nature of the algorithm that requires periods of exponential growth of confirmed cases as one of its input.

From figure 1, it is evident that there are no clear clusters that separates certain countries from others. The same can be said to figure 2. In the plot of both analyses, all the countries clustered around the top left corner and the factors proposed in the method sections does not seem to help with distinguishing countries' degree of impact from the global pandemic. This also means that the level of GDP, health expenditure, tourism smoking and obesity rate, and population all fails to predict how a country would be impacted by the COVID-19 pandemic. Since these factors failed to produce a discrete cluster on the 2D plot, it would suggest that either none of them are useful in predicting the COVID-19 outcomes, or that these variables were not linear. However, From figure 1, we can also tell that Principal component 1 accounts for 40% of the variable. In this principle component, the number of tourism and the national GDP are the most heavily weighted variable. This would suggest that the economic aspect of a country may be the leading factor that plays a role in affecting the pandemic response. The second principal component accounts for 25% of the variable and it is composed of mainly obesity rates and population. Although obvious, the second factor that may play a major role in affecting pandemic response may be the national population and their health condition.

We chose to use principal component analysis (PCA) for a number of reasons. For one, because we are handling a big variety of data including several variables, a simple correlation graph would not be enough: we would need to create many separate graphs in order to determine the correlation between all the variables. Due to this condition, the usefulness of variables needs to be measured. PCA as a dimensional reduction method, was considered to be suited for the task of feature selection, since it can handle multiple variables, reducing it down to a single 2D plot. Furthermore, PCA can present the data in a clearer way that removes all the duplicating and correlated information. However, there are still limitations with PCA. The most obvious one is that the variables are converted to principal components which are a linear combination of all the original variables and hence less interpretable. Furthermore, by its SVD-based nature, PCA does not handle data with high variability well. Because the principal components are generated from a linear summation of all the standardized data, variables with high variance could bias the principal components. Thus, our data may introduce bias through the creation of PCA modelling since our variables are very different from each other both in types and variance. In addition, since the data we were able to collect online is limited and potentially biased, this may further impact the result of our principal component analysis.

Because the PCA plot did not generate much useful information, we decided to employ Decision tree learning. Unlike PCA, which is a SVD-based dimensional reduction method that assumes the subspace of useful data is linear, no independence between variables or functional relationships are required for decision trees. Indeed, *population* and *obesity* were discovered to correlate to the COVID-19 control level. Even though the accuracy of this model is plausible, the sensitivity is rather low. This could be attributed to the nature of the decision tree, since the algorithm looks for the split that minimizes the impurities in the leaves, and the strict ranking criteria of the algorithm could mask some useful features potentially also correlated. A better result could be yielded if random forests were applied. The result of PCA run with the 2 selected variables is also discouraging; groups of different R_0 levels still cluster. The linear regressions performed afterwards support the fact that these 2 variables are non-linear. R^2 for both linear regressions that are near zero, and the p-values as high as more than 0.6 fail to reject the null hypothesis that there are no linear correlations between the R_0 and the 2 variables. Nevertheless, as the result of the feature selection by the decision tree suggests, there are still underlying interactions between R_0 and these variables that are yet to be discovered. If we were to use more advanced dimensional reduction methods that require no linearities, possibly PCA that are extended to nonlinear assumptions such as kneral-PCA or sparse-PCA.

5 Conclusions

Through analyzing the data collected online, we were unable to find a linear correlation between the variables proposed and the transmissibility of COVID-19. However, through further analysis, we were able to identify two factors

that may relate to COVID-19 which are obesity rate and population. Although they are not linearly correlated, further research could take a deeper examination regarding how these factors relate to the disease. Our current study suggests that obesity rate and population correlates to how well a disease transmits through a specific population. Unsurprisingly, population number does relate to how well it transmits but more interestingly, we found obesity to correlate to how well the disease transmit as well. This brings out many interesting implications both for future direction and for general insight. It is not a surprise that obesity is correlated and would very likely trigger other health issues like heart disease and so on but this research suggests that obesity could also in some way help the disease transmit. We have found similarity in previous research in influenza A that obesity indeed impacts the risk of pathogenesis of lung infection, in this case, coronavirus [4]. Although this does not mean obesity caused greater transmissibility, it does provide direction as to a potential third variable which could affect both factors.

Like previously suggested, future research should focus on identifying the exact relationship between obesity and transmissibility. Another approach is to broaden the country variables to include more aspects such as their stance on the political spectrum and others. Through doing so, we will be able to gain a better, more accurate understanding of these types of disease and ideally make sure it does not happen again.

References

[1] World Bank. Gdp data.

[2] World Bank. Health expenditure by countries.

[3] JHU. Covid-19 dashboard.

[4] Stacey Schultz-Cherry Rebekah Honce. Impact of obesity on influenza a virus pathogenesis, immune response, and evolution. *frontiers in Immunology*, 10, May 2019.

[5] Pierre-Yves Boëlle Thomas Obadia, Romana Haneef. The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, 12, December 2012.

[6] WHO. Timeline.

[7] Our world in data. Obesity rate.

[8] Our world in data. Population.

[9] Our world in data. Shared death from smoking.

[10] Our world in data. Tourism.

[11] Ye Wu1 Lin Wang Benjamin J. Cowling Zhanwei Du1, Xiaoke Xu1 and Lauren Ancel Meyers. Serial interval of covid-19 among publicly reported confirmed cases. *Emerging infectious diseases*, 26, June 2020.