Machine Learning and the Epidemiological Spread of COVID-19

Ghislain Niyongabo (ghislainniyongabo@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Professor Bruce Basset
African Institute for Mathematical Sciences, South Africa
University of Cape Town, South Africa

14 May 2020

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Declared a pandemic on 11 March 2020 by the World Health Organization (WHO), the Novel Coronavirus 2019 (COVID-19) is having a significant impact on public health, economy, and social life, and has caused over 1.5 million infections and 108113 deaths by 10th April 2020. In this essay, we investigate the factors impacting the spread of COVID-19. Our main goal is to evaluate the impact of environmental factors - such as temperature, humidity, and wind speed on the infection rate using random regression forests. The lockdown dates, population density, Gross Domestic Product (GDP), were also used as independent variables. First, we made an analysis on a global dataset and compared it with the African countries. The results did not indicate evidence of the impact of the environmental factors on the infection rate. The most important feature in our model was the GDP. This could be either because developed countries were more exposed to the pandemic (as of April 2020) or because they were conducting many more tests than poorer countries. From this, the study suggested that frequent hygiene, public health interventions, and social distancing policies should be primarily considered to fight the pandemic.

Keywords: COVID-19, SARS-CoV-2, WHO, Environmental Factors, Lockdown Dates, Transmission Rate, Initial Travel Risk(ITR), Random Forest.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Ghislain Niyongabo, 14 May 2020

Contents

1. Introduction

The spread of the coronaviruses has been characterized by a high rate of infection. Novel to epidemiologist, COVID-19 is respiratory illness caused by SARS-CoV-2. There is a growing hypothesis predicting the decline of the virus during warm weather (Notari, 2020), hence suggesting that environmental factors such as temperature, humidity, wind speed, air pollution, and so on, could play a major influence in eradicating the disease. However, the lack of enough epidemiological knowledge on the SARS-CoV-2 virus emerges as the measure challenge in stopping this new virus.

Initially, a preliminary analysis has investigated the possible impact of the environmental factors on SARS using mathematical and statistical techniques (Luo et al., 2020). In this essay, we study the spread of COVID-19 using machine learning more specifically using the random forest method with a focus on the interaction between COVID-19 infections and climate change. We investigate whether weather conditions have affected the rate of infection caused by the disease in different regions. However, also measuring the infection rate of different countries can be risky and lead to inaccurate predictions in a machine learning language. First, we have no real infected cases numbers by country because some countries do fewer tests than others. Second, since social distancing policies have a positive effect on the severity of epidemics based on our experience with SARS in 2003, we take into account the country lockdown dates as predictor of the number of infected cases , and also include population density to measure the correlation between population and infection rate. Here, the applied approach consists in developing a supervised learning model which can predict the number of confirmed cases in various countries or regions using different variables and compare the importance of each variable in the final predictions.

1.1 Background of the Novel Coronavirus 2019 (COVID-19)

Coronaviruses are family of viruses, which are infectious and fatal. They can be transmitted from both animals to humans and humans to humans. They are of different categories such as common human coronaviruses, MERS and SARS. COVID-19 is a new coronavirus identified since December 2019 in Wuhan, China. Initially, coronaviruses were identified circulating in animals and they could be transmitted from animal to human resulting in a "zoonotic event". However, they can be spread from humans to human resulting in a "Spillover infection". Spillover event is an infection process whereby a pathogen is transmitted from one population and to another population resulting in an outbreak (Johnson et al., 2015).

Some of the most common symptoms of infection identified from infected persons include fever, cough, fatigue and respiratory symptoms such as breathing difficulties. Some patients may also include other symptoms such as nasal congestion, diarrhoea, vomiting, sore throat, headache and other patients may experience a continuous weakness. The infection rate of COVID-19 becomes high as some people get infected but do not test since they present mild symptoms. Another group of COVID-19 patients are asymptomatic, who continually transmit the virus to other people without presenting signs of infection.

The COVID-19 virus is mainly spread from person to person when an infected person sneezes or coughs by dropping droplets of saliva. An individual can easily be exposed to these droplets if he stands closer to the infected person, thereby contracting the virus. When infected droplets fly on the surface or on an object, they can last for hours, so transmission can also occur after shaking hands, being in contact with a contaminated object, and then touching the mouth, nose or eyes before washing the hands.

The COVID-19 infection rate has increased exponentially worldwide since the beginning of 2020, which led the WHO to declare it a pandemic on 11 March 2020. Some countries have been described as at high risk of contracting the epidemic from China based on the volume of passengers they receive from China over a period of three months, 15 days before the Lunar New Year and around 75 days after according to Lai et al. (2020) at the University of Southampton. The survey has predicted 30 countries at high risk of catching the epidemic based on the volume of passengers they receive from China over the period of approximately three months.

This enormous mobility was at the origin of a wide diffusion towards other regions outside China. However, environmental factors could also have affected the early outbreak. Environmental conditions such as temperature and humidity can have a huge impact on various respiratory viruses (Pica and Bouvier, 2012). Although they do not have much information on COVID-19, some researchers have identified similarities to SARS which affected 8000 cases in 26 countries and the influenza pandemic which has been limited by the summer seasons. The serial interval for COVID-19 is estimated to be 5-6 days, which is higher than the serial interval for influenza virus estimated at three days. (WHO, a).

SARS-CoV-2 can survive longer on some surfaces and can last up to 24 hours on cardboard and stays up to 2-3 days if dropped on stainless steel and plastic surfaces (NHI). An infected individual can release up to 3000 droplets in a single cough. The stability of these droplets in the air or on surfaces can be strongly influenced by weather conditions and this can affect the rate of transmission among the susceptible population.

According to SARS and MERS, the stability of coronaviruses depends on the environmental temperature to which they are exposed. On one hand, low temperatures and low humidity, the virus tends to live longer, thereby increasing the community transmission. On the other hand, exposing the virus to high temperatures and high humidities significantly shortens the lifespan of the virus (Chan et al., 2011a). The variability of weather conditions can influence both the stability of the pathogenicity of viruses and the human immune system. There is evidence that in the presence of strong antibodies in the human body, respiratory diseases can be eliminated from the host (Xu and Gao, 2004). The impact of the recovery rate can therefore be vital for the eradication of infectious diseases and, if it is high for COVID-19, the infection rate could decrease, leading to a decline in the disease. However, we must ensure that patients recovered from COVID-19 are no longer contagious and determine the time it takes for the virus to be completely eliminated from recovered individuals.

The effect of environmental factors on the stability of SARS-CoV-2 is not well understood, but for such an unknown virus, it is worth using all possible research to find a solution to this pandemic. Although temperature and relative humidity are presented as the main environmental factors affecting the transmission of SARS-CoV-2 (Wang et al., 2020), other variables can also be included such as air pollution and absolute humidity can influence the stability of any respiratory virus if it is placed in different conditions. It is therefore necessary for epidemiologists to understand in depth the transmission of SARS-CoV-2 and also to obtain precise information on the biological nature of the virus.

The WHO provided the first models of transmission of SARS-CoV-2, including direct human-to-human transmission when there is a physical contact between an infected person and a healthier person. However, due to the rapid spread of the virus, some epidemiologists predict that the virus can be transmitted by air and from the WHO report published on March 29, 2020, they recommended that people take important steps and precautions to avoid any surprises that may occur although there is no evidence of an airborne infection yet noticed in China or the world (WHO, b).

1.2 Random forest regression

In this section, we will provide a brief description of the random forest algorithm applied for regression tasks.

1.2.1 Basics of regression tree

Various approaches can be employed to solve regression problems. In the traditional method, the analysis can be achieved using a simple regression model that takes one variable X as the independent variable and use it to predict a variable Y as the target or dependent variable, e.g Predicting the salary based the level of education. In some cases the target can be predicted by a set of multiple variables where the features can be represented as a set $X = \{X_1, X_2, ..., X_n\}$, and in this scenario, we work with multiple regression model, e.g The same salary can be estimated using both the level of education and the type of program or course taken by an employee.

Let consider π to be a function of X such that

$$\pi(X) = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_n\beta_n, \tag{1.2.1}$$

where β_0 represents the intercept and β_1 is the slope, the linear function Y can be mathematically be expressed as follows:

$$Y = \pi(X) + \epsilon, \tag{1.2.2}$$

where Y is the dependent variable and ϵ is the error term.

This function works best for linear data, but it might work poorly when given non-linear data. In this situation, we need a model that can perform predictions regardless of the type of dataset, and decision trees are primarily accredited for this case because of their functionality of making predictions across multiple decisions.

The methodology of growing regression trees has been introduced by Breiman et al. (1984) under Classification and Regression Trees (CART) algorithm. The CART algorithm is a binary decision algorithm that construct trees by repeatedly splitting a node into two separate nodes. The process of growing the CART algorithm is done by choosing the best split for each predictor and the the split or node that minimizes the errors will be chosen at the root node. Ideally, a stopping criterion for the spitting process has to be determined depending on whether the problem is classification or regression.

Regression tree is a structural algorithm in machine learning that uses a tree-based model to fit the data and make decisions. This algorithm is commonly used in data analysis to build predictive models in a situation where the target is a real number or a continuous variable. First, growing the regression tree is done by dividing the predictor space into N distinct partitions $(R_1, R_2, R_3, ..., R_k)$. Second, the predictions are performed by averaging the observations located in the same region R_k. The main purpose of this tree-building process is to find k that minimizes the following equation:

$$RSS = \sum_{k=1}^{N} \sum_{i \in R_k} (y_i - \hat{y}_{R_k})^2, \tag{1.2.3}$$

where RSS is the residual sum of squares, $\hat{y}_{R_k}$ is the mean value of the observations located in one particular region k and y_i represents the actual observations. The above computations are time-consuming

by doing the same measurements on all points. However, to minimize the algorithm running time, we introduce the Recursive Binary Splitting (RBS), which is generally used to measure the importance of the variables for a regression tree. Hence, the variable given high importance will start on top of the tree. The RBS can be examined by the following equation:

$$RBS = \sum_{i:x_i \in R_1(n,l)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(n,l)} (y_i - \hat{y}_{R_2})^2, \tag{1.2.4}$$

where $R_1(n,l)$ and $R_2(n,l)$ represents two regions around a point l minimizing the RSS, resulted from splitting the tree into two branches by the RBS, such that $R_1(n,l) = X|X_n < l$ and $R_2(n,l) = X|X_n > l$. Also, X_n is a predictor located in a given partition. The same procedure is recursively repeated, by evaluating at the same time the RSS at each split of the partition. There are a few criteria that we could consider, including limiting the maximum depth of the tree, and this criterion will be responsible for terminating or stopping the algorithm. Note that the predictions are made from the average values of all the observations located in each region. The regression tree algorithm is highly exposed to over-fitting but there are a lot of techniques to reduce the high variance, as explained in Section 1.2.3.

1.2.2 Regression tree architecture

Decision trees are generally implemented based on the same model known as "tree-model". However, regression trees differ from classification trees in the final predictions which are discrete values for classification trees whereas regression trees output continuous values impacting also on their differences in evaluation metrics.

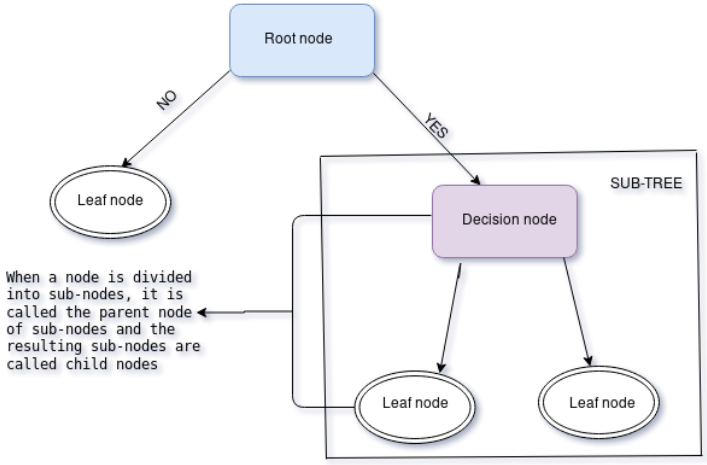


Figure 1.1: Architecture of the decision-tree model. The main decision node is denoted as the root node where the binary splitting is performed to give two regions depending on the selected conditions. At each internal node in the tree, one of the inputs is split and the result indicate the left-hand or right-hand of the sub-tree.

The basic terminologies of the regression tree are as follows:

1. Root node: Similarly to a tree, a tree-based model is also build on a root node which is another name to call a top decision node.

2. Decision node: A decision node arises after the splitting of a sub-node if the results are other sub-nodes. A decision node is also called a parent node.
3. Leaf node: Also called terminal node, it determines a node without any other split on it. Predictions are usually made on a leaf node and it is usually called a child node.
4. Sub-tree: a tree resulted from another node of a main tree. It is commonly called a branch.
5. Parent and Child node: A parent node is a node which is divided into sub-nodes and each sub-node is referred as the child node of that particular parent node.

1.2.3 Pruning decision trees

In machine learning, some algorithms learn too much from the training data which can result in inaccurate predictions.This phenomenon is technically called overfitting, which has a significant impact on the generalization capacity of an algorithm. In the decision tree, a technique called pruning is used to lower the size of decision trees by detaching certain sections, thereby reducing the complexity of the tree. The sections are mainly nodes that reduce the final accuracy of the model. Several metrics are used to measure this tree complexity and those include nodes, maximum leaves, tree depth, and also the number of attributes used in the model.

To reduce this complexity, a technique called "cost complexity pruning" can be applied. The main concept is to build a tree with the entire data which means that the tree will develop nodes containing a small number of instances and prune the tree until all the nodes with noisy information are removed.

The first step with cost complexity pruning is to compute the residual sum of squares for each tree (sub-trees included) and this is given by the Equation 1.2.4. The second step is to compute the score for each tree. The score is given by the sum of RSS and the tree complexity penalty as expressed in Equation 1.2.5

$$Score = \sum_{k=1}^{T} \sum_{i:x_i \in R_k} (y_i - \hat{y}_{R_k})^2 + \alpha |T|, \tag{1.2.5}$$

where α is the tuning parameter that we get using cross-validation and $|T|$ represents the number of leaves in each tree or sub-tree. During the cross-validation, the goal is to find α that minimizes the tree-score equation and therefore this tree will be selected as the model of our algorithm.

1.2.4 Random forest

Decision trees are machine learning algorithms very exposed to overfitting. Random forest is a machine learning method that reduces overfitting by combining multiple decision trees to make predictions (Breiman, 2001). In regression, the predictions will only be the average of the outcome from each tree. The main methodology used in random forest for regression is called "bagging". According to Breiman (2001), the average predictions obtained from the random forest will likely have smaller variance compared to a model implemented with one tree, and the variance reduction effect is proportional to the number of trees applied. Two main concepts are used which include randomization of the samples in a dataset and selection of a random subset of features.

Figure 1.2 shows the illustration of how the random forest is achieved by dividing the full dataset into multiple training data and assigning them to different trees randomly.

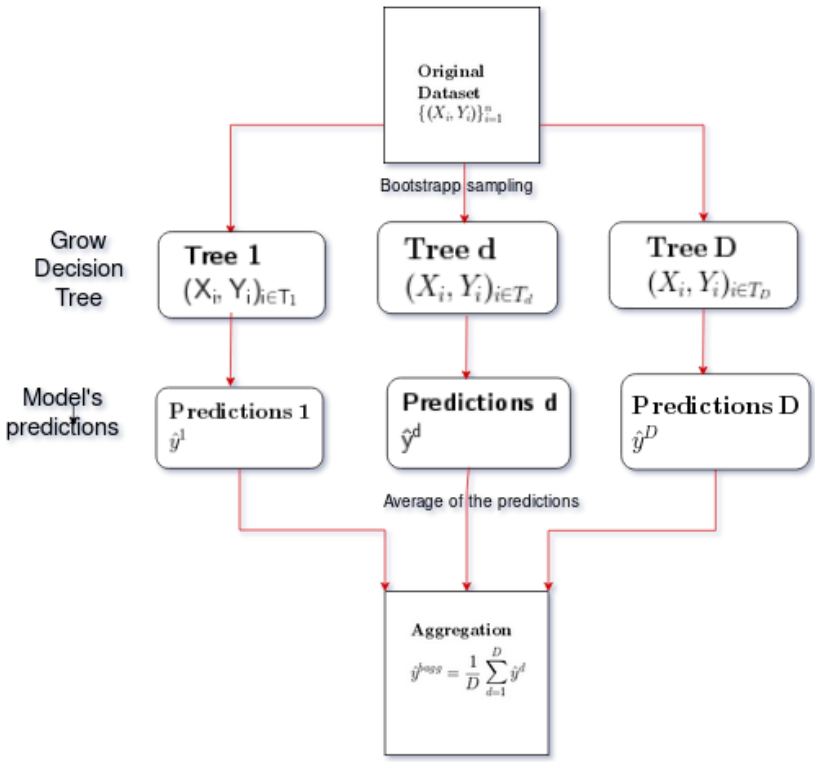


Figure 1.2: The algorithm is centered on bootstrap sampling. This architecture is employed for both regression and classification. The difference is based on the final predictions as classification applies voting to select the most common class.

Bagging: Prediction errors produced in a single tree are due to the type of data used in the training process, resulting in high variance. Bagging is an aggregation technique initiated by Breiman (1996) to reduce the variance with the main idea of dividing the data into many subsets.

For a given dataset D of a pair $\{(X_i, Y_i), ..., (X_n, Y_n)\}$, we can represent the dataset as a matrix D of n-dimension.

$$D = \begin{pmatrix} X_{11} & X_{21} & ..... & X_{i1} & Y_1 \\ X_{12} & X_{22} & ..... & X_{i2} & Y_2 \\ X_{13} & X_{23} & ..... & X_{i3} & Y_3 \\ . & . & ..... & . & . \\ . & . & ..... & . & . \\ X_{1n} & X_{2n} & ..... & X_{in} & Y_n \end{pmatrix}$$

Let us consider a set of trees $T = T_i, ...T_n$ and choose randomly a bootstrap sample D_i from the main dataset D. Using D_i, we will create several trees by assigning at each node random subsets of features as well as splitting on those features occurring to the randomization of samples.

We can also represent this process of bootstrap sampling in a matrix form.

$$D_1 = \begin{pmatrix} X_{15} & X_{25} & Y_{15} \\ X_{16} & X_{26} & Y_6 \\ X_{17} & X_{27} & Y_7 \\ X_{18} & X_{28} & Y_8 \end{pmatrix} D_i = \begin{pmatrix} X_{42} & X_{22} & Y_2 \\ X_{43} & X_{23} & Y_3 \\ X_{44} & X_{24} & Y_4 \\ X_{45} & X_{25} & Y_5 \end{pmatrix} ..... \quad D_n = \begin{pmatrix} X_{72} & X_{83} & Y_3 \\ X_{73} & X_{84} & Y_4 \\ X_{74} & X_{85} & Y_5 \\ X_{75} & X_{86} & Y_6 \end{pmatrix}$$

Recalling D as the sequence of the bootstrapped subsets, the forest is built by selecting a bootstrap sample D_i denoted as the i^{th} bootstrap. A random subset of features b is drawn at each tree split and then a decision-tree is developed. The algorithm is adjusted at each node of the tree, such that instead of using all possible feature-splits, we select randomly $b \subseteq F$, where F is the set of features. The node then splits on the best feature in b rather than F. In reality, the optimal number of bootstrap samples depends highly on the available data.

Finally, the predictions are the averages from the various trees trained in the forest as expressed mathematically in the Equation 1.2.6:

$$\hat{y}^{bagg} = \frac{1}{D} \sum_{i=1}^{D} \hat{y}^{(i)} \tag{1.2.6}$$

where $\hat{y}^{bagg}$ represents the average or mean of the predictions from the given number of constructed trees D and $\hat{y}^{(i)}$ represents the prediction from each individual tree.

This process of randomly selecting bootstrap samples reduces similarity between trees, thus when averaging the final predictions, the model will likely to reduce the variance. The prediction errors in random forest are estimated by computing the Out-of-bag error which is generated from the process of bootstrap sampling with replacement when growing the trees.

Feature importance: In the presence of multiple independent variables, all variables may not have an equal influence in a machine learning algorithm and this includes measuring the importance of each feature, thus getting information on which feature does better predictions. The random forest feature importance is computed using straightforward techniques: mean decrease impurity (MDI) and mean decrease accuracy (MDA). For a given a set of data $D_n = \{(X_i, Y_i)\}_{i=1}^{n}$, the random forest records the out-of-bag error for each data points during the fitting process. On one hand, the importance of the feature is measured according to the optimal criterion (node impurity) chosen typically variance for regression. This measurement is referred to as the MDI which is simply achieved by computing how much each feature reduces the impurity in a tree. For an ensemble, the impurity reduction from each feature can be averaged and the features are classified according to this measure. The other technique used in the ranking of the features in a random forest is MDA which directly measures the influence of each variable on the final predictions of the model. The process is done by permuting the values of the feature and then measuring how the accuracy of the model decreases. Clearly, the less important variable should not have a huge impact on the accuracy of the model, whereas if the variable is very important, its permutation should imperatively have more impact on the final predictions of the model.

Hyperparameter Tuning: The main concept behind random forest is combining the predictions from multiple decision trees to find the mean of the final output for regression. However, the accuracy of the model could also be affected by the number of trees deployed in the forest. Bagging of the dataset is also done based on the number of trees to define the number of subsets created from the original

dataset. To address this challenge we need to optimize the model through hyperparameter Tuning. The random forest algorithm depends on various hyperparameters and apart from the number of trees, we must also take into account the number of entities assigned to each tree and also determine the maximum depth of each tree which provides the optimal architecture. The process of identifying the right parameters to use requires an experimental study that is done by building several algorithms and combining different hyperparameters to select those which offer better performance.

There are several strategies used in hyperparameter tuning for the random forest algorithm. The most popular include grid search and random search, others include F-Race, generally simulated annealing, sequential model-based optimization (Probst et al., 2019), and bayesian optimization. It is then important to choose the right hyperparameter optimization metric and this depends highly on which parameters to be tuned to reduce the errors in the predictions. The parameters used in a random forest also depend on whether we are in the case of regression or classification and tuning them requires a brief understanding of their usefulness in the algorithm.

1.3 The purpose of the essay

The purpose of this essay is to investigate the epidemiological spread of COVID-19 using machine learning. In this essay, we will focus on the factors that might influence the transmission of the virus more specifically the main contribution here is to investigate if the environmental factors have an impact on the spreading of the virus. This will help governments and policymakers to take measures to fight against this disease. A random regression forest will be developed to assess the variability of COVID-19 infected persons from different countries to weather conditions.

1.4 The outline of the essay

The essay is subdivided into four main chapters. Chapter 2 provides details on the methodology used in developing the model. Initially, the source of the data was given describing the type of data as well as to give a better understanding of the problem. This was followed by data analysis and processing to understand the underlying patterns of the data. It allowed us to get a better hypothesis and assumptions that were used in the modelling process especially for the interpretation of the results. The model was also explained in this chapter summarizing the steps followed in building the random forest regression and Two main analyses have been done. This chapter provided also details on the metrics used in evaluating the performances of the three models.

Chapter 3 provided all the results we got in the analyses done and followed by the discussion that leads us to conclusions later in the fourth chapter.

Chapter 4 contained the conclusions commenting on the findings we got from the modelling process. This chapter also included the different recommendations made and the future work that might be done to better study this new global pandemic especially on the factors that influence the transmission of the SARS-CoV-2 virus.

2. Methodology

In this chapter we discuss the methods used to make the analysis. First, it provides the source of the data used in the study. Second, we talked about the exploratory analysis done of the data collected. Third, we describe the machine learning model used which presents the steps and how the data was analysed using the model. At last, the chapter shows the metrics used in evaluating the performance of our random forest model.

2.1 Data collection

The study uses data obtained from various sources. These include the epidemiological data of COVID-19, the weather conditions, and human mobility data interpreted as the Initial Travel Risk(ITR) for this study. We will discuss this in further details in the following subsections. These three sources have been combined in one table whereby each high risked country was associated with its respective risk rate, and average February temperature.

2.1.1 COVID-19 report data

The data was collected from the COVID-19 Data Github Repository provided by Johns Hopkins CSSE [1]. The source provides three main time series datasets. First, we have a dataset for daily reports of confirmed cases that contain information about the city, country/region, location coordinates and a time series of 01-22-2020 containing the cases reported daily in each city. Second, we have daily report of recovered cases which contains the same information except that a time series of dates contain information related to daily reported recovered cases from each city. Lastly, we have the dataset storing the recorded cases where for each data recorded cases are cumulative to the previous date. For example if five cases are reported then they will be added to the previous numbers. Despite being mainly reported as a time series with each daily report as a column, the data has been reproduced in one table for all future days starting from 01/22/2020 in the following format.

- Observation Date: The collection date of each observation.
- Province/State: Province or state
- Country/Region: Country or region
- Confirmed: The total number of confirmed cases

2.1.2 ITR data

This data was obtained from the preliminary risk analysis of the COVID-19 spread in China and beyond, carried out by Lai et al. (2020). Considered the highest human mobility in the world, the analysis used the rate of travelers during the 15 days preceding Lunar New Year's Day and around 75 days after. Mainly, the analysis provides the ITR for the other cities inside China based on the passenger volume of Wuhan identifying 18 high-risk cities. Second, the ITR for countries beyond China was also calculated by ranking the top 30 countries or regions that receive the highest volume of passengers from 18 high-risk cities in China. This data is constructed with three main columns, including Country, Volume for the number of travelers and the percentage risk.

[1] https://github.com/CSSEGISandData/COVID-19s

Table 2.1: The top 30 countries or regions at high risk of contracting COVID-19 due to the volume of passengers they receive from China a few days preceding Luna New Year's Day and more than 75 days after (Lai et al., 2020).

Rank	Country	Volume	Risk
1	Thailand	2031.9	15.03
2	Japan	1563.3	11.57
3	Hong Kong SAR, China	1001.7	7.41
4	Taiwan, China	979.7	7.25
5	South Korea	936.6	6.93
6	United States	773.3	5.72
7	Malaysia	634.3	4.69
8	Singapore	568.1	4.20
9	Viet Nam	468.4	3.47
10	Australia	455.6	3.37
11	Indonesia	412.5	3.05
12	Cambodia	262.9	1.95
13	Macao SAR, China	260.4	1.93
14	Philippines	250.3	1.85
15	Germany	234.9	1.74
16	Canada	208.5	1.54
17	United Kingdom	190.7	1.41
18	United Arab Emirates	162.3	1.20
19	Italy	152.9	1.13
20	Russia	151.3	1.12
21	France	137.9	1.02
22	New Zeland	120.7	0.89
23	India	106.7	0.79
24	Spain	105.8	0.78
25	Turkey	66.5	0.49
26	Egypt	57.5	0.43
27	Sri Lanka	55.7	0.41
28	Maldives	50.7	0.37
29	Netherlands	44.9	0.33
30	Myanmar	43.3	0.32

2.1.3 Environmental data

The environmental data was obtained from a time series dataset collected from Github [2]. Our goal is to investigate the severity of the outbreak from 22 January 2020 up to the lockdown dates. As stated in Chapter 1, the analysis included other variables to support the results such as lockdown dates collected

[2] https://github.com/chrisfinlay/covid19/tree/master/data

from Wikipedia [3], population density collected from Kaggle [4] and the GDP obtained from Wikipedia [5].

2.2 Data preparation and analysis

2.2.1 Exploratory analysis

The data was analyzed from different angles, but before considering the factors that may be affecting the spread of the COVID-19, we undertook the empirical analysis using the data collected to get a clear understanding of the global situation. We proceeded by grouping all the confirmed cases by the observation day and we visualized the data using a curve line (see Figure 2.1). Our data show that the number of confirmed cases increases exponentially. At this instant, we grouped the cases by Country/Region and observation date (see Figure 2.2), to get the cumulative number of confirmed cases. This showed 185 countries that have already confirmed the existence of the virus in their territory as per the report of 10 April 2020.

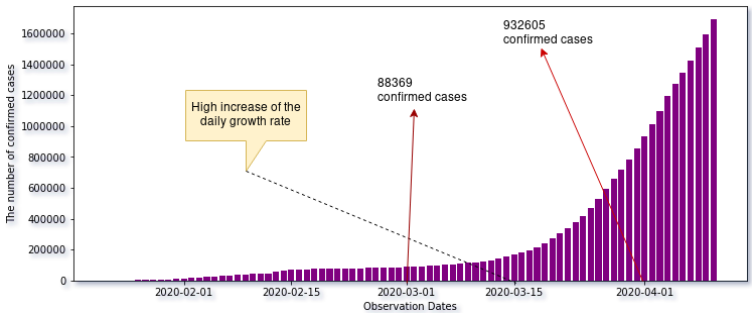


Figure 2.1: Analysis of the global situation for the confirmed cases from 2020-01-22 up to 2020-04-10.

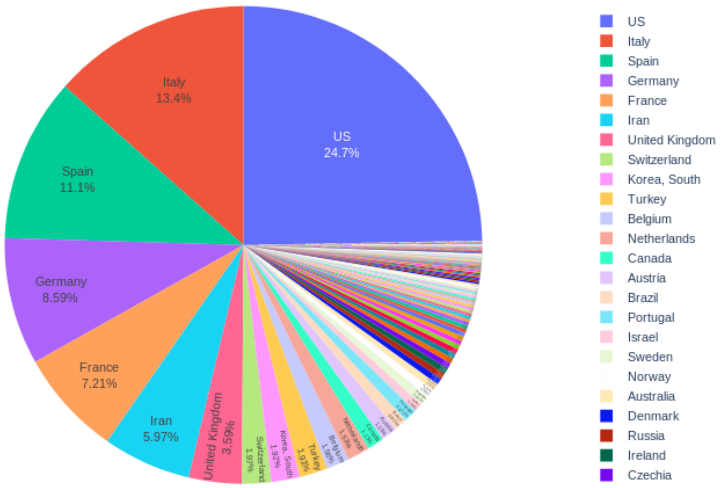
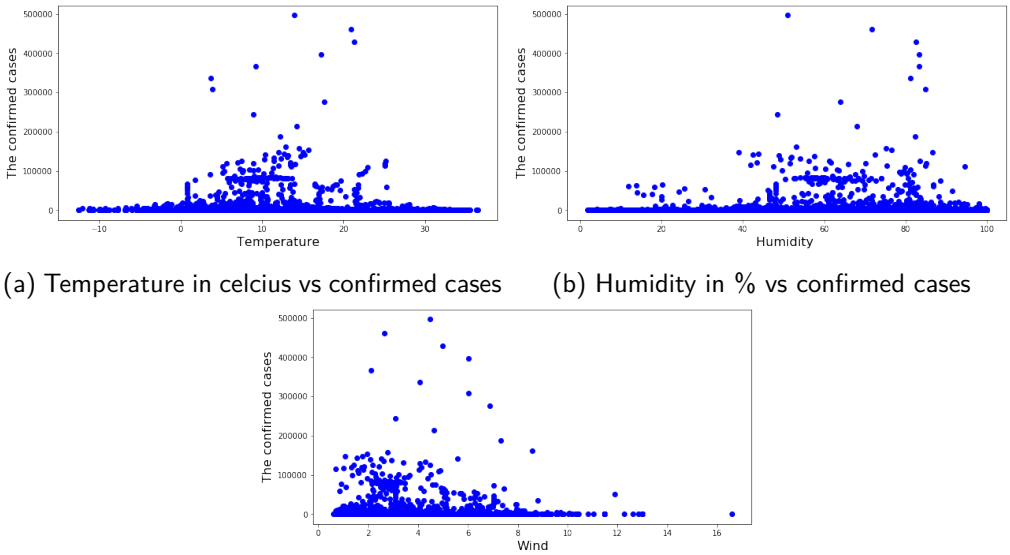


Figure 2.2: Distribution of the number of confirmed case in the period of 2020-01-22 up to 2020-04-10 for the global countries out of China. 24.7% of patients are from the United States.

[3] https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density
[4] https://www.kaggle.com/jcyzag/covid19-lockdown-dates-by-country#countryLockdowndatesJHUMatch.csv
[5] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)



(a) Temperature in celcius vs confirmed cases (b) Humidity in % vs confirmed cases

(c) Wind in m/s vs confirmed cases

Figure 2.3: Data visualization of the number of confirmed cases as a function of the variation of the weather conditions where each points represents daily cumulative confirmed cases for each country.

Figure 2.3 show the distribution of the daily confirmed cases in the variation of weather conditions (temperature, humidity, and wind speed). The findings do not give evidence of a linear relationship between the independent and the target. However, we can see that the two most affected countries have a wind speed between 1 and 5 m/s. Interestingly, the data shows that countries with wind speeds varying from 10 to 18 m/s have fewer confirmed daily cumulative cases than countries with lower wind speeds. Despite showing that the highest number of confirmed cases have been observed in temperatures ranging from 0° to 25°, there is also no clear evidence of linear association in the data. As shown in Figure 2.3; (a), when the temperature is less than 0° or above 25°, the COVID-19 confirmed cases are likely to be reduced. The daily recorded humidity conditions do not also show strong linearity although it is visible that some days with high humidity, the numbers of confirmed cases are very high (see Figure 2.3; (b)).

Interested in the countries that have been categorized as the top 30 high risked countries based on the assumptions that those countries might be the most affected regions due to the volume of travelers they receive from the mainland China, we analyzed the severity of the outbreak within those countries. A machine learning algorithm is more likely to suffer from overfitting and therefore if a model is trained with less amount of training data, it will be unable to learn enough which might lead to the lack of generalization, hence, resulting in inaccurate predictions when given the testing data. To strengthen the analysis, we increased the number of samples by adding other countries that present a severe situation of the disease as it was in early April, to get enough observations and their ITR were set to 0% as shown in Table 2.2.

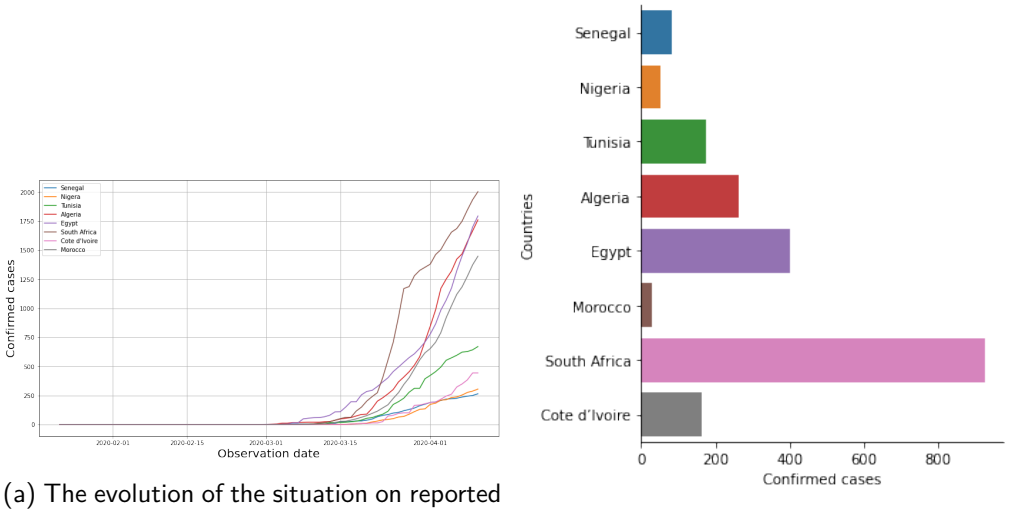
Table 2.2: The 52 countries selected for the global analysis sample dataset

Country	Risk	Confirmed_cases	Lockdown_date	Temperature	Humidity	Wind	Population density	BMI	GDP
Algeria	0	264	24/03/2020	22.38	10.65	5.76	18.41	26.2	180689
Argentina	0	128	20/03/2020	26.48	45.29	7.02	41	27.7	518475
Australia	3.37	2044	24/03/2020	18.73	69.17	3.72	3	27.2	1432195
Austria	0	1332	17/03/2020	-3.59	62.57	1.77	106	25.4	455737
Belgium	0	559	13/03/2020	5.8	53.1	5.13	376	25.5	531767
Cambodia	1.95	93	28/03/2020	30.24	60.29	3.67	90	21.9	24572
Canada	0	657	18/03/2020	-0.4	72.78	3.96	4	27.2	1712510
Colombia	0	470	25/03/2020	14.04	93.42	1.59	41	25.9	330228
Czech Republic	0	258	16/03/2020	-1.6	55.58	3.51	135	26.9	245226
Egypt	0.43	402	24/03/2020	21.77	19.61	4.17	100	29.2	250895
Finland	0	1041	27/03/2020	0.19	63.19	3.98	19	25.9	273961
France	1.02	6683	16/03/2020	21.8	81.79	6.1	14	25.3	2777535
Germany	1.74	9257	17/03/2020	2.1	54.9	3.2	233	26.3	3996759
India	0.79	536	24/03/2020	29.23	27.78	1.83	414	21.9	2726323
Indonesia	0	117	15/03/2020	26.25	86.76	0.8	141	22.9	1042173
Iran	0	13938	15/03/2020	9.98	35.17	4.25	51	26.2	454013
Ireland	0	43	12/03/2020	6.16	77.28	4.91	70	27.5	382487
Israel	0	100	12/03/2020	14.2	57.5	2.33	416	26.3	369690
Italy	1.13	7375	08/03/2020	6.27	53.99	5.43	200	26	2073902
Japan	11.57	1468	27/02/2020	3.03	65.04	1.82	200	22.6	4970916
Lebanon	0	110	16/03/2020	8.87	62.07	1.64	333	27.8	56639
Lithuania	0	17	16/03/2020	-0.87	49.82	2.19	43	26.6	53251
Luxembourg	0	77	16/03/2020	3.03	51.04	5.33	237	26.5	69488
Malaysia	4.69	1183	21/03/2020	24.07	97.37	0.83	99	25.3	354348
Maldives	0	8	11/03/2020	28.89	76.54	3.66	1258	25.1	5272
Mexico	0	26	14/03/2020	21.43	35.75	2.34	64	28.1	1223809
Morocco	0	29	16/03/2020	11.53	67.56	2.15	80	25.6	118495
Netherlands	0.33	1416	16/03/2020	22.24	73.16	7.69	420	25.4	913658
New Zealand	0.89	8	25/03/2020	16.14	91.14	7.89	18	27.9	205025
Norway	0	702	12/03/2020	-2.83	62.46	2.87	17	26	434751
Panama	0	443	25/03/2020	20.41	94.15	4.49	56	27.1	65055
Peru	0	11	10/03/2020	23.45	91.58	0.95	25	26.3	222238
Philippines	1.85	142	16/03/2020	26.76	83.96	5.74	362	23.2	330910
Poland	0	49	12/03/2020	0.67	48.8	4.53	123	26.4	585783
Portugal	0	448	18/03/2020	12.78	73.77	2.87	112	26.2	237979
Qatar	0	262	15/03/2020	18.89	45.08	5.3	237	29.2	192009
Romania	0	906	25/03/2020	0.07	74.69	5.58	81	25.3	239553
Russia	1.12	1036	27/03/2020	-8.75	68.6	4.5	9	26.5	1657554
Saudi Arabia	0	103	15/03/2020	15.52	15.2	2	16	28.5	782483
Serbia	0	48	17/03/2020	1.12	48.09	3.9	89	25.8	50508
South Africa	0	927	26/03/2020	24.08	43.02	4.11	48	27.3	368288
South Korea	4.69	9661	23/03/2020	8.1	49.72	2.22	517	23.9	1619424
Spain	0.78	80110	29/03/2020	10.8	75.55	4.43	93	26.7	1426189
Sri Lanka	0.41	77	21/03/2020	22.7	71.43	1.6	332	23	88901
Sweden	0	3700	29/03/2020	-0.1	61.41	3.02	23	25.8	551032
Switzerland	0	3028	18/03/2020	-0.45	79.29	1.44	208	25.3	705501
Taiwan	7.25	59	15/03/2020	16.16	81.7	1.41	652	0	574905
Thailand	15.03	75	13/03/2020	30.97	58.97	3.79	130	24.1	504993
Turkey	0	1529	23/03/2020	5.3	60.93	1.96	106	27.8	766509
Ukraine	0	14	17/03/2020	1.58	51.97	6.33	69	26	130832
United Kingdom	1.41	6726	23/03/2020	16.93	79.76	6.82	274	27.3	2825208
Vietnam	3.47	113	22/03/2020	20.59	89.28	1.21	290	21.6	244948

Preparing data is one of the most indispensable tasks in machine learning, especially when we have multiple sources of information. Once the sample data has been selected, we assigned the ITR (expressed in percentage), population density, the lockdown date for each country, and construct a sample dataset made of 52 countries as shown in the Table 2.2 above. The dataset collected from Johns Hopkins CSSE provides daily reports starting from the 22 January 2020. We, therefore, denoted this date as the

starting date column and we computed the number of days between this initial date to the lockdown date of each country to obtain total days. As we measure the cumulative number of confirmed cases on the exact day of the first lockdowns, we then made predictions regarding the number of days it took for governments to take rigorous social distancing policies. To illustrate the impact of weather conditions on the infection rate, we also calculated the average weather conditions for each country based on the lockdown dates. Recall that some regions adopted partial lockdowns and were involving a ban on some activities and cities. However, this did not have a significant impact on the analysis since we study the impact of the independent variables before the lockdowns measures.

Further analyses were carried out with African countries. Typically, our hypothesis is based on the fact that the weather conditions would impact the stability of the SARS-CoV-2 virus, thus summer season characterized by a rise in temperature and humidity would reduce the spread when the virus is exposed to the sunlight and heat. African countries are generally warm and we would expect SARS-CoV-2 to be less contagious. Consequently, we proceeded by analyzing the impact of the outbreak in the continent using eight affected African countries till their lockdown dates and it is shown in the Figure 2.4 below.



(a) The evolution of the situation on reported cases in function of Time

(b) Cases at the time of lockdown

Figure 2.4: Data visualization of the number of confirmed cases in Africa over time (a), and the cumulative number of confirmed cases per country (b)

2.2.2 Association analysis

Additionally, we performed further data analysis by computing the correlation to understand the behavior of each variable with respect to the change of the other variables. Pearson's correlation was used as the statistical measure to get the association of the independent variables and the dependent variable.

$$r(xy) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{2.2.1}$$

where $r(x, y)$ is the correlation of the two variables x and y. x_i and y_i are the observations. $\bar{x}$ and $\bar{y}$ are the respective means of the observations of the two variables.

First of all, we examined the relationship between the number of confirmed cases from a global perspective. Then, we also studied the impact of the variability of the weather conditions and the ITR on the severity of the SARS-CoV-2. We also measured the association between the other four variables such as population density, GDP, closing dates and cumulative confirmed cases. A negative value of the correlation between two different variables indicates an opposite behavior while a positive correlation shows that the one variable increase as the other rises and vice-versa.

To strengthen the association test results, we calculated the p-value (α) which is a statistical measure showing the probability value indicating whether the obtained results occurred by chance alone. The designated threshold of significance is standard with $\alpha = 0.05$. In this study, we used a Python module (statsmodels) that compute many different statistical measures and the p-value included. If the p-value is less than 0.05, this shows that the results are statistically significant and and they may unlikely to be occurred by chance itself. In contrast if the p-value is greater than 0.05, then the test is not statistically significant which indicates weak evidence against the null hypothesis. The p-value assumes two hypotheses (null and alternative). In this study, the null hypothesis says that there is no statistical significance between the independent variables and the dependent variable. The alternative hypothesis indicates that the independent variables studied in this study have a significant impact on the dependent variable.

2.2.3 Data pre-processing

Generally, a machine learning model is built with two main types of variables. The input is called the independent variable or technically known as a feature while the output is known as the dependent variable or target in the machine learning language. The target is the numbers of confirmed cases so it is a continuous variable that indicates the existence of the regression task. We have also performed the scaling of the number of confirmed cases using a logarithmic function of base 10. We divided the dataset into features X and the target Y the latter was the variable to be predicted, in this case, the "Confirmed cases" column in Table 2.2. After this step, the variables X and Y were adapted for the model into a NumPy array. The main goal of the pre-processing step is to convert the data from pandas to vector form.

One of the most frequent scenarios with Machine Learning algorithms is that they can perform well in the simulation but when given real data, the accuracy can reduce dramatically resulting in the lack of generalization. One solution to this problem is to divide the data into training and test subsets. This enables the model to be trained and through the cross-validation process, a good model can be chosen to be used in the predictions using a separate dataset. If the model performs better in the training process and provides accurate predictions when given new data, this may justify the efficiency of the model.

In the next step after separating the attributes from the labels, we split 75% of the data for the training subset while 25% of the data for the testing subset. After dividing the data into training and testing subsets, we implement the machine learning model. As discussed earlier in chapter 1, we have chosen to use the random forest algorithm. Generally, the idea is to have a method that can make predictions using multiple decisions and the best solution is to use a tree-based model. The target of our model is continuous which implies that the suitable method is a regression tree. Despite, the numerous advantages presented in decision trees algorithms such as less effort for data preparation or less impact of missing values in the final predictions, these algorithms present huge flaws in predictions. They often take time to train the model since a tree may be developed with a lot of the number of leaves. Thus, these computations involve complex calculations and a little change in the data can have an impact

on the entire structure of the tree. To address these challenges, we decided to use a random forest which easily handles the bias-variance problem using **bootstrap aggregation** as briefly described in the Subsection 1.2.4.

2.3 Modelling process

Before implementing the final model, we performed the hyperparameter tuning. The features and the target variable values of the training dataset were processed to the method created for the default random forest regression model using a sklearn package RandomForestRegressor used for ensemble learning. In section 1.2.4 we defined the role played by hyperparameter tuning in machine learning. Random forest is implemented using several hyperparameters but in this model, we focused on optimizing the most important based on their influence on improving the predictions. Also, the method used for tuning the hyperparameters was the grid search cross-validation method from the sklearn library to determine the optimal parameters to fit the random forest algorithm as follows:

1. **n_estimators:** This defines the number of trees to be deployed in the forest.
2. **max_depth:** The maximum depth of the tree.
3. **max_feature:** The maximum number of features to be considered for splitting a node.
4. **min_samples_leaf:** The minimum number of data points authorized in a terminal node.

According to the result computed from the Grid search algorithm trained on 10-fold cross validation, the model preformed better with 49 trees. Hence, we observed that the accuracy of the random forest varies in function of the number of trees deployed in the forest. The maximum number of depth was investigated and 3 was specified as the optimal number for my model. We therefore set the stopping criteria to be no fewer than 6 countries in the terminal node. The features are randomly selected by taking the square roots of all the features and ideally, will take 2 features per tree since we have 7 features. A new model was then built using the aforementioned parameters. The features were processed in the model where they followed the method of bagging. After training the model using the 49 trees, we obtained the final performance by averaging the model performance or prediction errors from each tree using evaluation metrics as explained in section 2.4.

Algorithm : Random Forest pseudo code

1: Input: Initialize $D = \{D_1, ..., D_n\}$ as the training dataset with n bootstrap samples and
2. the model's hyperparameters
3. Output: Set $T = \{1, ..., T_n\}$ as the n trees
4: For each i in T:
5: select the bootstrap sample D_i from D with replacement:
6: Construct a Regressor T_i on D_i
7: i++
8: Results: average the predictions

2.4 Evaluating the model

After fitting the model, we examined the performance of the model by estimating how good the model will perform when given new data. In this case of a continuous target, we used two main metrics which are the Mean Absolute Error (MAE) and Mean squared error (MSE).

1. The MAE return the mean of the magnitude of errors between the actual values and the predictions. It is expressed mathematically by the following expression:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2.4.1}$$

   where y_i represents the actual observations and the $\hat{y_i}$ represents the predicted values.

2. The MSE returns the average of squared differences between predicted and actual observations by measuring how far away the average prediction is from the actual value. It can be expressed mathematically as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{2.4.2}$$

The smaller the values are for the above metrics, the better the model performs since this is an indication that the predictions are closed to the line of best fit. Ideally, for the MAE and MSE are all oriented negatively ranging from 0 to ∞, the best value should be closer to zero. Recall that in random forest, the predicted values $\hat{y}_i$ are widely designated by $\widehat{\bar{y}}_{i,OOB}$ to express the average prediction for the ith observation from the forest.

3. Results and discussions

3.1 Exploratory analysis findings

After grouping the number of confirmed cases by the observation date, we visualized the global situation of the impact of the disease from the 22 January 2020 to 10 April 2020. The result from Figure 2.1 shows that a few days before with the Lunar new year' day on 25^{th} January 2020, there is a smooth increase and from the 1^{st} February 2020, we then remarked a rise of infection rates indicating the human mobility as the influential factor for the transmission of the disease. The number of global infected cases was 12038 with more than 98% of the reported cases originating from China as for February 2020.

Around the 1st March 2020, the situation was seemingly constant but from 15 March 2020, we observed an increase of infections (see Figure 2.1). A month later, exactly on 1 April 2020, the numbers have been subjected to another rise, defining a larger exponential growth rate of the COVID-19 with a total global report of 932605 cases of SARS-CoV-2 infections on the 1st April 2020 (see Figure 2.1). This defines more than 10 times of infections compared to the report of 88369 cases recorded in the previous month. In fact, this can be explained by the fact that countries adopted a high rate of COVID-19 testing resulting in an increase of reported cases.

The figure shows the cumulative number of cases for each country classified in the top 30 ranked high-risk countries (see Table 2.1) for getting infected by the COVID-19. Ranked 19^{th} in the high-risk countries ranking, Italy has confirmed many COVID-19 cases up to the lockdown date with 9172 followed by South Korea ranked 5^{th} and with 7478 confirmed cases. This pod is followed by France, Germany, and Spain with respectively 1209, 1176, and 1073 confirmed COVID-19 cases. Should we conclude that human mobility has been the spreading factor for COVID-19? Saying so would raise questions for countries like Taiwan and Thailand. These two countries recorded only 45 and 46 infected cases while they both had higher ITR.

The seasonality of other respiratory viruses has also been a case study by other researchers, and some results show that seasonality is not always the case for all respiratory viruses. (Hazlett et al., 1988). One notable example is the Human Adenovirus (HAdv) which has become endemic in Taiwan. Existing in Kenya, the human adenovirus appears from November to February, which represents the warm period of the region (Hazlett et al., 1988) .

We then investigated the situation of the outbreak in Africa. Recall that African countries are generally warm which might tell us that the virus should not be harmful based on the weather conditions. From the data collected from 22 January 2020 till the lockdown dates in March, we selected 8 countries that reported some of the highest numbers early April, and our findings did not reveal clear correlations between the weather conditions and the number of confirmed cases. Located north of the Equator, Algeria, Egypt, Morocco and Tunisia are located in the Northern Hemisphere. Algeria counted 264 on the 24 March 2020, Tunisia with 173, Morocco with 29, and Egypt with 402 confirmed cases on their lockdown dates. Surprisingly, South Africa which is located in the southern hemisphere, recorded 927 on 26 March 2020 (South Africa lockdown date) whereas other countries such as Nigeria, Côte d'Ivoire, and Senegal recorded fewer cases although they banned large gatherings of people earlier to prevent the pandemic to be more contagious. Looking at the ITR, Egypt was also ranked as the highest-ranked country in Africa based on the preliminary risk analysis investigated by Lai et al. (2020). However, as reported in Figure 2.4, as the time increases, we do not see any impact of the ITR on the number of cases, and in fact, South Africa did not feature among the top 30 high risked countries but looks to be the most affected country in Africa. On the other hand, we used the averages weather conditions from

the 22 January 2020 to the specific dates of lockdown for each country and collected the cumulative number of cases. A country like Morocco located in the northern hemisphere with respectively low temperatures, the country has registered fewer cases of COVID-19 more than the other countries which are generally warmer. Would we say that the lockdown date set on 16/03/2020 has impacted these numbers compared to countries like South Africa which entered in confinement from a little bit later? We would doubt this, because as reported in Figure 2.4, we observed that the numbers increased dramatically after this date. This may be an increase in the number of tests. To sum up, from this analysis, we could not get a clear indication of both the impact of the ITR and weather conditions on the number of confirmed cases either globally or in Africa.

To investigate our study statistically, we used a common technique used to measure the relationship between two variables. The correlation between the independent variables and the dependent variable was measured. The correlation between the risk of getting infected and the number of confirmed cases returned a positive value which implies that they vary in the same direction. The correlation between the temperature returned a negative which shows that they vary in the opposite direction, thus, when the temperature increase, the numbers are more likely to decrease. However, these correlations tend to be weak, 0.066 and -0.249 for the ITR and temperature.

Table 3.1: The Pearson's correlation coefficients for the seven independent variables computed from the 52 countries used in the sample dataset of the analysis (see Table 2.2).

Feature	Correlation coefficient	p-value
ITR	0.066	0.35
Temperature	-0.249	0.178
Humidity	0.051	0.352
Wind speed	0.047	0.799
Lockdown date	-0.246	0.11
Population density	-0.132	0.738
GDP	0.530	0.0001

The value of the correlation between the temperature and the confirmed cases is -0.249 which shows a negative linear relationship however it is very weak to confirm the impact of the temperature on the variation of the number of cases. For the variable ITR, the correlation was 0.066 very weak although positive. Despite being also positive, the correlation between humidity and confirmed cases was very weak. Both the lockdown date and the population density have returned negative values and also weak. The above results computed from the initial analysis were very strange because none of our independent variables showed us strong relationships as seen in Table 3.1 and there was no significant effect shown by the weather temperatures. However, it was quite unexpected to get a negative correlation coefficient for the population density since a very dense country should imply a high number of cases. Countries like Maldives and Taiwan with 1258 and 652 respectively, have confirmed few COVID-19 cases as compared to less dense countries such as Italy, Russia, and even African countries such as South Africa.

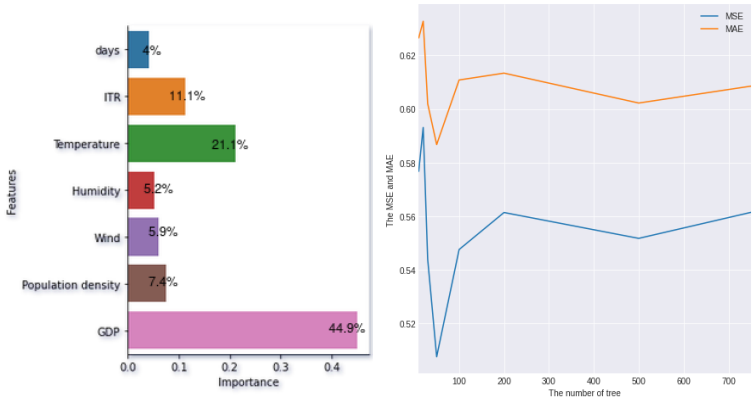
To a very large extent, the GDP show a better correlation coefficient as compared to other independent variables registering 0.530 respectively. As shown in Figure 2.2, we observe that most of the developed countries have the highest number of confirmed cases as for April data. For example, the United States (US) was the most affected country in early April having 24.7% of the worldwide confirmed cases. Also, it is important to point out that countries with higher GDP have conducted COVID-19 tests more than developing countries impacting the numbers reported. However, it is not usual for developed countries

to experience a worse situation because they have improved health facilities and advanced public health interventions.

The significance of the association between the independent variables and the dependent variable was again measured using the p-values, the latter were calculated using the Pearson's correlation coefficient. Our null hypothesis is that each of the seven independent variables (temperature, humidity, wind speed, ITR, population density, lockdown dates and GDP) has no association or linear relationship with the the transmission rate of COVID-19 (measured as the the confirmed cases), that is, the correlation coefficient is zero.As shown in Table 3.1, the p-values for all the independent variables, except GDP, are greater than 0.05. In practice, we can decide on an appropriate threshold value based on our study (α), but we chose to use the standard $\alpha = 0.05$ as a convention.The p-value for temperature was 0.178 and this indicates that there is no statistically significant relationship between the temperature and the number of confirmed cases. The p-values corresponding to the other environmental factors such as humidity, wind speed are allgreater than 0.05. Population density, ITR and lockdown dates did not show a significant relationship with the number of confirmed cases based on their respective p-values. Nevertheless, the GDP had a very small p-value with $p < 0.05$, which showed solid evidence against our null hypothesis, thus indicating that the variable has a significant association with confirmed cases globally

3.2 Model results

Result of the global analysis: The results for the random forest show that the GDP did better on the final predictions of the model, scoring around 44.9% of the predictions. The temperature was the second variable to influence the predictions on the confirmed cases with 21.1%. The variable humidity scored 5.2%. The variable ITR was the third to perform better with 11.1% and the population density contributed 7.4%. The wind speed contributed 5.9%. At last, the variable lockdown date impacted 4% in the final predictions of the random forest model. The results show that the GDP performs again better than the other variables. The MAE and the MSE were 0.58 and 0.50 respectively trained in a forest of 49 trees.



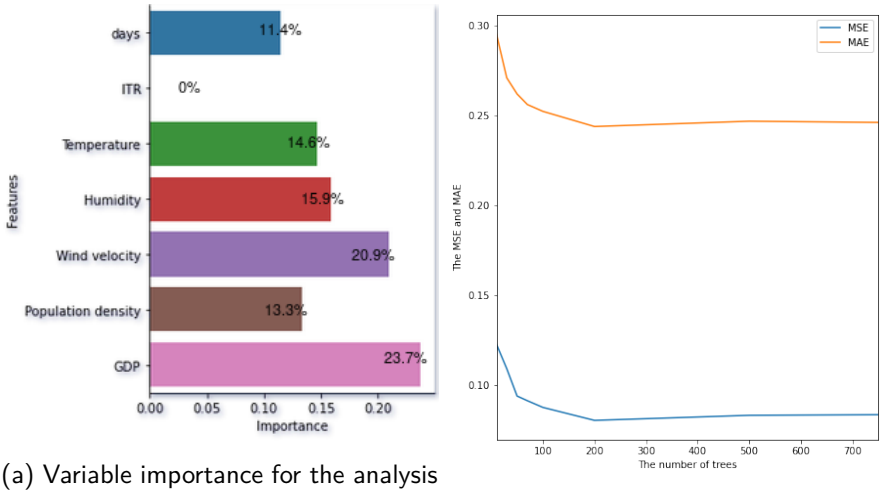
(a) The variable importance in the global predictions

(b) Random forest prediction errors

Figure 3.1: The plot (a) shows the feature importance with the GDP being more informative than the other variables. The generalized errors over 750 trees are also displayed in plot (b).

Analysis of the situation in Africa: The results for the random forest show that the temperature was not the best performer on the final predictions of the model, scoring around 14.6% of the predictions.

The variable humidity scored 15.9%, and the wind speed was 20.9%. The population density returned 13.3% and the GDP of the selected countries scored 23.7% in the random forest model. Finally, the variable lockdown date had an impact of 11.4% while the ITR showed no contribution in the final predictions of the model.



(a) Variable importance for the analysis of Africa

(b) The prediction errors

Figure 3.2: Results of the random forest model: The importance of the characteristic is found in plot (a) with again higher GDP while the generalized errors on 750 trees are also displayed in plot (b) with the optimal errors obtained in a model of 200 trees.

The result of the model trained on the entire African dataset was so considerable with the GDP performing again better than the other variables. The MAE and the MSE were respectively 0.24 and 0.08 obtained with 200 trees (see Figure 3.2b; (b)). Compared to the result of the global analysis model, our random forest has been improved although we have noticed a big impact of the datasets on the final results. The temperature was not important: neither tested on the global dataset nor tested on the African countries analysis. Additionally, the remaining environmental factors e.g humidity and wind speed performed poorly in the models. The results also highlighted that the GDP is more informative than the remaining variables for the two analysis. The ranking of these features was achieved by computing the average increase in node purity, measured by the decrease in sum of squared errors when X_i is chosen for split. For each feature, the sum of squared errors decrease across each tree in the forest is accumulated each time that feature is selected to split a node. The result is divided by the number of trees in the ensemble to give the average, so the feature that minimizes this value will be ranked as the most important feature of the model. The variables can also be ranked by counting the total number of trees in which X_i is used to split the root node because the root node is selected as a variable that minimizes the sum of squared errors when growing a regression tree.

3.3 Discussion

Understanding the variability of SARS-CoV-2 in different weather conditions is a key point in determining the seasonality of the Novel Coronavirus 2019 and the main goal of this analysis is to investigate whether the exponential curve can be flattened by the environmental factors. Initially, the migration of the virus from Wuhan, China to the other China cities and other regions out of China has been mainly caused

by human mobility during the period of the Luna New Year which is characterized by a huge volume of travelers. The countries that normally receive a high volume of airlines from China were more exposed to the outbreak and, in early February, regions ranked with higher ITR recorded more cases than other regions (Lai et al., 2020).

The incubation period of COVID-19 is estimated to vary from 2-14 days, but some studies have shown the existence of cases that did not show any symptom (Bai et al., 2020), and this made it hard to get an exact the basic reproduction number (R_0). However, if getting exact number of confirmed cases numbers becomes hard, then it looks good to take an alternative by evaluate the Case Fatality Rate (CFR) which was explained later in this section. However, the global infection rate has been growing exponentially since February and during March, we observed a rapid increase announcing deadly disruptions in various global domains such as, social, health, political, and economic. Globally, the total number of cases was 1657929 infections on 10 April 2020.

The transmission from human to human has been mainly characterized by direct contact between an infected person and another person. Indirect transmission is also observed where an individual by getting in contact with an object infected by the virus. Besides the normal spreading route (infected droplets and touching a contaminated surface), Airborne transmission is being studied although it has not been confirmed. Despite not having enough epidemiological data for SARS-CoV-2, the lifespan of the virus on the different environment has already been identified and the possible role of environmental factors in the stability of this virus needs to be addressed.

Conversely, we have observed that countries with very lower weather conditions especially in temperature have recorded fewer cases than countries with median conditions. Austria, Canada, and the Czech Republic are among countries that had a lower average temperature for the considered period (from 22 January 2020 to lockdown date) with, -3.59,-0.4, and -1.6 degrees, respectively. This could be explained by the fact that these countries do not have a higher volume of travelers coming from the 18 cities at high ITR of from China during this period.

The result from the random forest model showed that temperature was more important than the ITR in the predictions, scoring more than 20% of the model performance. These results revealed a strong influence of the weather conditions on the stability of the virus resulting in high infections when the virus is exposed to lower temperatures. This analysis could imply that the southern hemisphere may be prepared for the worst situation during its winter season, while the decline for the outbreak could be announced in the coming days for the Northern Hemisphere by entering its warm season. However, these conclusions could be wrong when we have to admit the fact that countries have been testing differently with some denying the existence of the pandemic. The number of tests can determine the numbers of some countries for example from the data we collected on 10 April 2020, we observed that some countries the pandemic was seen to be more deadly than the countries that were affected in the early days of the outbreak. Even though the testing protocols seem to be important in the COVID-19 reports, they do not affect the transmission rate. It is also interesting to consider the capability of some countries in dealing with pandemics. The study has also included factors such as the population density as social distancing is regarded as the remedy of the COVID-19 as no vaccine or treatment is available yet, as well as the GDP.

Importantly, the GDP was the most important feature in the random forest model scoring more than 40% in the final predictions but with the coefficient of the Pearson's correlation being positive, this suggests us that as the higher the GDP is, the higher the number of cases. On one hand, these facts make sense since countries have a high probability of having travelers from China or from any other country that might have contacted the pandemic earlier. On the other hand, having a high GDP implies

having better prepared medical systems and adequate public health interventions.

The severity of COVID-19 observed in developed countries raises concerns compared to the developing countries. The high population densities in the main cities, lack of hygiene education, and inadequate healthcare facilities, the COVID-19 pandemic could be exposed to a high Case Fatality Rate(CFR). Although we can also associate with the fact that most of the developing countries e.g African countries have weather conditions that could limit the rate of respiratory virus transmission, the findings do not show a direct correlation between the infections in Africa and they are more likely to be influenced by the performed tests on the continent.

The results of our study could also be biased by the fact that we do not have the exact numbers. In fact, determining the attack rate of COVID-19 is blurred since the pandemic has been characterized by asymptomatic individuals who can also transmit the virus without being counted as cases. The incubation of period of COVID-19 is estimated to vary from 2-14 days, but some studies have shown the existence of cases that did not show any symptom (Bai et al., 2020), and this made it hard to get an exact reproduction number. However, if getting exact confirmed cases become hard, then evaluating the number of deaths since these numbers are hard to be hidden might indicate an better estimation of the exact infected cases.

The CFR which is the the proportion of deaths from all individuals diagnosed positive over a certain period of time, can be measured to provide statistics of the number of deaths. We selected 20 countries with the highest COVID-19 infections and we computed the CFR for three months to see the variation of the ratio. In the United Kingdom, the CFR was lower in the early stages of the outbreak (1,56% for cases up to 10 March) and one month later, there was a high increase up to 12% for the confirmed cases up to 10 April 2020. Despite being the most affected country with around 496535 COVID-19 confirmed cases, the United States recorded a CFR of just 3% in early April. In Italy, the outbreak affected around 10149 cases and the CFR was 6% in the data of 10 March, but an increase of confirmed cases up to 147577, was seen with a double CFR with 12% at the beginning of April.

The main advantage of the CFR is the ability to determine the number of confirmed cases when its value is constant by multiplying the number of deaths and the CFR. In general, the data show also that the CFR from all the countries was not constant which become hard to estimate the number of confirmed cases using the CFR. Given that the CFR is given by the total deaths over the total confirmed cases, the value is more likely to be also influenced by the number of tests performed.

We made other investigations to compare the number of conducted tests and the weather conditions on the infection rate of the COVID-19. We approached this analysis by selecting 88 countries with the highest total tests per 1 million population as it was on 10 April 2020 [1]. We introduced again the random forest by implementing a model that predicts the total cases of infected persons using the temperature, humidity, wind speed as well as the total tests per million performed by country. The results were quite interesting with the total tests per million outscoring the other variables in the model with a contribution of more than 60% in the final predictions. Clearly, weather conditions can help respiratory viruses spread faster. However, this requires a strong epidemiological investigation on the SARS-CoV-2 virus to get accurate information on the conditions that may inactivate rapidly the virus, hence resulting in a slow contagious rate. When inspecting the number of confirmed cases, it appears that the severity of the COVID-19 pandemic is more likely to be determined by the number of tests performed.

[1] https://github.com/owid/covid-19-data

4. Conclusions and future work

The influence of various environmental factors on the transmission of the viral respiratory virus cannot be ignored since several studies have shown evidence on the seasonality of some infectious diseases. However, the lack of epidemiological data of SARS-CoV-2 leads to many challenges as medical experts research on how to contain the COVID-19. In this essay, we investigated the weather conditions that impact the spreading of the pandemic. In our approach, we implemented random forest regression to evaluate the performance of the environmental factors in predicting the number of confirmed infected cases. To address the problem, we initially selected the highest risk countries (see Table 2.1) to be infected by the outbreak because of the volume of travellers they receive from China from the end of January to around March and we have added additional affected countries to increase observations.We proceeded by adding more predictors such as the ITR, population density, GDP and the days taken by each country to impose lockdown. We implemented models for the global situation and then for African countries. Finally, the models were compared using performance metrics by analyzing the prediction errors of each model and as well by measuring the contribution of each independent variable in the final predictions.

We also observed that the results of the model depend on the dataset. The result of the models did not show clear evidence of a weaken transmission of the virus in warm conditions. The variable GDP outscored other variables, which revealed that more developed countries are the most affected by the pandemic. Therefore, we cannot expect the outbreak to go away because of the change in temperature alone. Important public health measures, social distancing policies, and rigorous hygiene have to be maintained in all countries help limiting the rate of transmission.

The study presented some limitations which did not allow to fully conclude the impact of environmental conditions on the spread. In some countries, most of the variables such as temperature, humidity, wind speed, population density, and ITR may differ between cities. For future work, a good study should also investigate the impact of environmental factors by considering cities.

Acknowledgements

I wish to express my deepest acknowledgments to the Lord for the grace that brought me in the AIMS community. I would like to recognize all the people whose assistance was a milestone in the completion of this essay. I would like to thank Professor Bruce Basset for his supervision, support, invaluable advice and, above all, the confidence he has created in me. I want to acknowledge AIMS and it's funders for supporting this work, as well as thank the AIMS lecturers, tutors, administration who provided me with relevant training which opened up many future opportunities for me. I would also like to give special thanks to my family, friends, my tutor Samar El Sheikh, AIMS Students, and AIMS family for the undeserved love and support you provided me throughout my stay at AIMS South Africa.

References

Adhikari, S. P., Meng, S., Wu, Y.-J., Mao, Y.-P., Ye, R.-X., Wang, Q.-Z., Sun, C., Sylvia, S., Rozelle, S., Raat, H., et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (covid-19) during the early outbreak period: a scoping review. *Infectious diseases of poverty*, 9(1):1–12, 2020.

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet*, 395(10228):931–934, 2020.

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.-Y., Chen, L., and Wang, M. Presumed asymptomatic carrier transmission of covid-19. *Jama*, 323(14):1406–1407, 2020.

Breiman, L. Bagging predictors. *Machine Learning*, 24(2):123–140, aug 1996. doi: 10.1007/bf00058655. URL https://doi.org/10.1007%2Fbf00058655.

Breiman, L. *Machine Learning*, 45(3):261–277, 2001. doi: 10.1023/a:1017934522171. URL https://doi.org/10.1023%2Fa%3A1017934522171.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. Classification and regression trees, wadsworth statistics. *Probability Series, Belmont, California: Wadsworth*, 1984.

Chan, K., Peiris, J., Lam, S., Poon, L., Yuen, K., and Seto, W. The effects of temperature and relative humidity on the viability of the sars coronavirus. *Advances in virology*, 2011, 2011a.

Chan, K.-H., Peiris, J. S., Lam, S., Poon, L., ky, Y., and Seto, W. H. The effects of temperature and relative humidity on the viability of the sars coronavirus. *Advances in virology*, 2011:734690, 10 2011b. doi: 10.1155/2011/734690.

Chen, B., Liang, H., Yuan, X., Hu, Y., Xu, M., Zhao, Y., Zhang, B., Tian, F., and Zhu, X. Roles of meteorological conditions in covid-19 transmission on a worldwide scale. *medRxiv*, 2020.

Hazlett, D., Bell, T., Tukei, P., Ademba, G., Ochieng, W., Magana, J., Gathara, G., Wafula, E., Pamba, A., Ndinya-Achola, J., et al. Viral etiology and epidemiology of acute respiratory infections in children in nairobi, kenya. *The American journal of tropical medicine and hygiene*, 39(6):632–640, 1988.

Johnson, C. K., Hitchens, P. L., Evans, T. S., Goldstein, T., Thomas, K., Clements, A., Joly, D. O., Wolfe, N. D., Daszak, P., Karesh, W. B., et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific reports*, 5:14830, 2015.

Lai, S., Bogoch, I. I., Watts, A., Khan, K., Li, Z., and Tatem, A. Preliminary risk analysis of 2019 Novel Coronavirus spread within and beyond china. WorldPop, 2020.

Luo, W., Majumder, M., Liu, D., Poirier, C., Mandl, K., Lipsitch, M., and Santillana, M. The role of absolute humidity on transmission rates of the COVID-19 outbreak. medRxiv. 2020.

NHI. Study suggests new coronavirus may remain on surfaces for days | [national institutes of health (nih)]. https://www.nih.gov/news-events/nih-research-matters/study-suggests-new-coronavirus-may-remain-surfaces-days. (Accessed on 14/04/2020).

Notari, A. Temperature dependence of covid-19 transmission. *arXiv preprint arXiv:2003.12417*, 2020.

Pica, N. and Bouvier, N. M. Environmental factors affecting the transmission of respiratory viruses. *Current opinion in virology*, 2(1):90–95, 2012.

Probst, P., Wright, M. N., and Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), jan 2019. doi: 10.1002/widm.1301. URL https://doi.org/10.1002%2Fwidm.1301.

Wang, J., Tang, K., Feng, K., and Lv, W. High temperature and high humidity reduce the transmission of COVID-19. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3551767. URL https://doi.org/10.2139%2Fssrn.3551767.

WHO. Coronavirus disease 2019 (COVID-19) situation report-46. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4, a. (Accessed on 04/25/2020).

WHO. Modes of transmission of virus causing covid-19: implications for ipc precaution recommendations. https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations, b. (Accessed on 04/13/2020).

WHO. WHO on the Novel Coronaviruses (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses, c. (Accessed on 05/07/2020).

Xu, X. and Gao, X.-M. Immunological responses against sars-coronavirus infection in humans. *Cell Mol Immunol*, 1(2):119–122, 2004.