Data Preprocessing Report

0.1 The COVerAGE-DB Project

The COVerAGE-DB project (https://osf.io/mpwjq/) contains 3 data files:

- the “Input” data file which consist in the accumulation of all the raw data communicated by governments
- the “Output_5” data file which is a projection of Covid-19 cases by group of 5 years
- the “Output_10” data file which is a projection of Covid-19 cases by group of 10 years

Given the necessity to distinguish finely between the age groups, the use of the “Output_5” data file appears as the best option. Indeed, contrary to the “Input” data file in which data can be given in different age brackets according to each countries’ reporting standard, the “Output_5” data file provides an homogeneous analysis of Covid-19 cases by age brackets of 5 years using spline approximations when the data for this age bracket is not available for a country.

0.2 The “Output_5” Dataset

The original data consist in 14089320 observations of 10 variables (117 distinct *country*, *region* within the country, an unique observation *code*, the *date* of the observation, the *gender* which can be male, female or both, the *age* bracket by 5 years from 0 to 100, a confirmation of the *age interval* for each bracket, the total number of *cases* so far, the total number of *deaths* and the total number of *tests* performed).

However, all the data are not available for each observation. For instance, out of the 14089320 observations, 1423172 are missing from the *cases* variable, 2661428 are missing from the *deaths* variable, and 12020690 are missing from the *tests* variable.

After having kept observation for whole countries, including both male and female cases and removing all artifact observations recorded before December 31st, 2019, date of the first official case reported in Wuhan, China, the data from 114 are then available.

Among the covid numbers reported, only the total number of cases is relevant for our analysis. However, after removing the missing values in the *cases* variable 107 countries are remaining.

0.3 Disparities Between Countries

Even if each of the remaining 107 countries have exploitable *cases* number reported, the data file is not homogeneous in term of the period covered and regularity in reported data (Figure 1).

Regarding the period covered by the data collected, it appears that some countries have started their data reporting very late in the pandemic and other have ended their reporting very early. In average, each country have a total of 252 days reported with a standard deviation of 241. However the most important problem is the continuity in the reporting of data. Some countries are missing days even weeks without any regularity that would explain this phenomenon. As a result, many more countries will be doped out from the pre-processing.

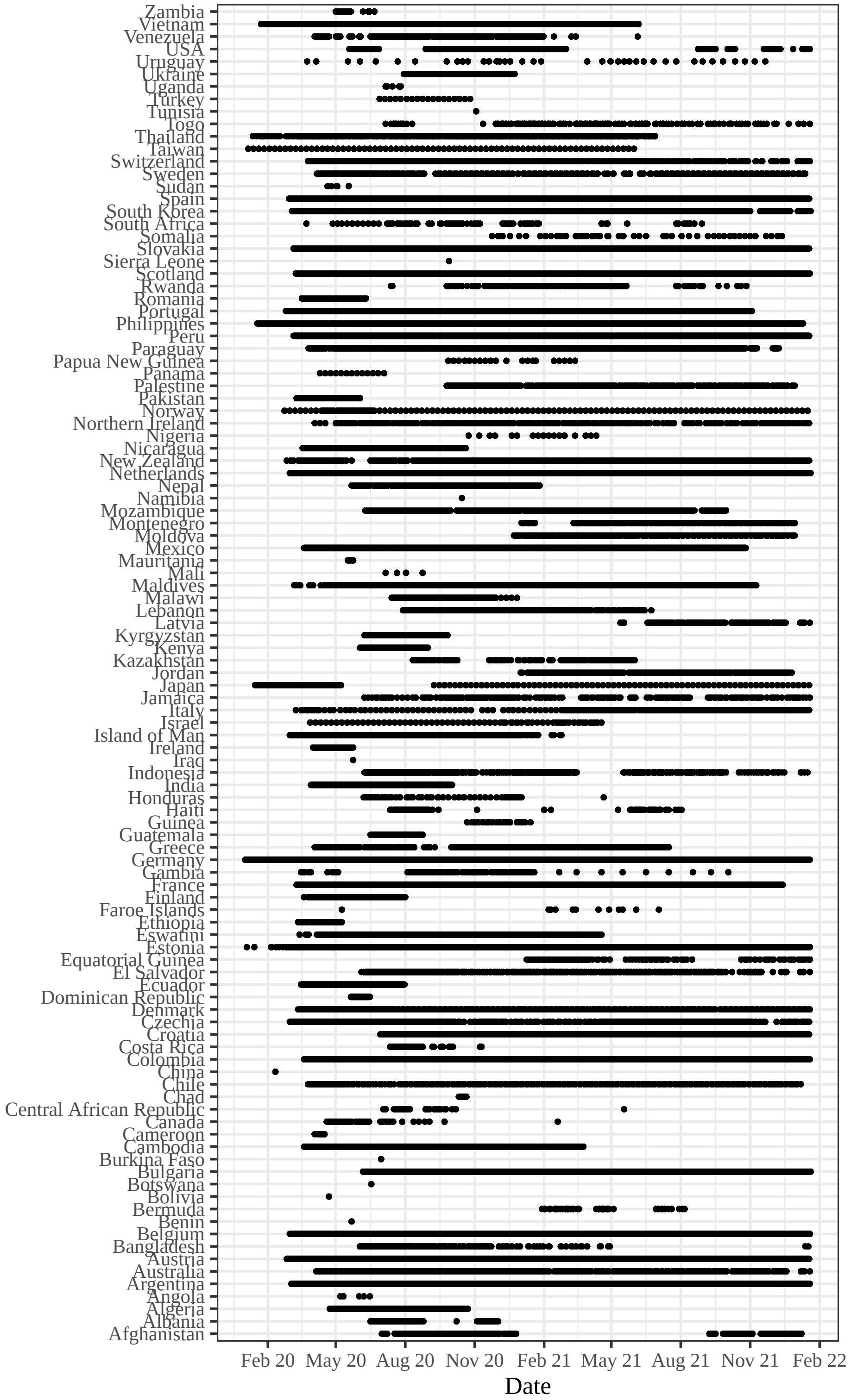

Figure 1: Disparities between countries regarding the reporting of Covid-19 cases. Each dots is a valid observation.

0.4 Changes in Covid-19 Cases After School Closures

Despite the data of the remaining 107 countries are not homogeneous, none of them has been dropped so far for this reason. However, the following steps in the data pre-processing will lead to significantly reduce the amount of data analysed.

First, the data reported are the total number of Covid-19 cases from the start to the pandemic to a specific date, also called total cases, but the analysis needs to be done using the daily number of cases at a specific date. The *daily number of cases* at a specific date n_x is calculated with the difference between the total cases at a date t_x and the total cases at a date t_{x-1} (i.e., lag 1). If t_x or t_{x-1} is missing then n_x will be missing as well. This brings the amount of countries to 88.

Second, our aim is to evaluate the change in daily cases after school closure. Therefore, countries time series are cut in multiple Target Period from the first day of the school closure to 28 days (4 weeks) after the closure whether the schools have reopen or not. However, to be taken into account, a Target Period should include a closure for at least 21 days after the closure. As a result, 50 countries have at least one of the research Target Period.

Finally, due to the non-linear nature of spline approximation to calculate the daily cases, the difference between t_x and t_{x-1} can lead to negative values. However, because negative daily cases are not plausible, if a target period contains a negative value then the entire period is removed. In addition, ensure the quality of this analysis, Target Periods with missing values are also removed.

As a result, the exploitable final number of countries is 22, each having a different amount of target period (Table 1).

Table 1: List of remaining countries and the corresponding number of target periods that can be analysed from each country.

Countries	Nb of Target Periods
Argentina	4
Austria	3
Belgium	7
Bulgaria	3
Cambodia	1
Colombia	2
Croatia	4
Czechia	1
Estonia	5
France	6
Germany	5
Greece	3
Maldives	4
Mexico	2
Netherlands	7
Paraguay	1
Peru	4
Philippines	1
Portugal	5
Slovakia	6
Spain	2
Vietnam	2