Clustering of the epidemic time trends: the case of COVID-19

1 Model

Suppose we observe a large number of time series $\mathcal{Y}_i = \{Y_{it} : 1 \leq t \leq T\}$ for $1 \leq i \leq n$ and each time series $\mathcal{Y}_i$ satisfies the following nonparametric regression equation

$$Y_{it} = m_i\left(\frac{t}{T}\right) + u_{it} \tag{1.1}$$

for $t \in \{1, \ldots, T\}$ with $m_i(\cdot)$ being an unknown smooth function defined on $[0, 1]$. As usual in nonparametric regression (see e.g. Robinson, 1989), we let the regression function m_i in model (1.1) depend on rescaled time t/T rather than on real time t. The assumptions for the error term u_{it} will be discussed later.

Suppose that all of the unknown functions $m_i(\cdot)$ can be divided into K classes in the following way:

- Each of the K classes of functions $\mathcal{F}_k$ are defined as
  $\mathcal{F}_k := \{f : [0, 1] \rightarrow R \,|\, f = c \cdot g_k(b \cdot u) \text{ with } c > 0, b \in [1, \bar{b}] \text{ and } g_k \text{ a density function}\}$.
  We assume that $\bar{b}$ is known beforehand. For the identification purposes, we also assume that the classes are distinct, i.e. $\mathcal{F}_k \cap \mathcal{F}_{k'} = $ for any $k \neq k'$.

- Suppose that $\{1, \ldots, n\} = \cup_{k=1}^{k} \mathcal{G}_k$ such that for any k we have

  $$m_i \in \mathcal{F}_k \text{ for all } i \in \mathcal{G}_k.$$

  In other words, for all $i \in \{1, \ldots, n\}$ we can write $m_i(u) = c \cdot g_k(b \cdot u)$ for some $k \in \{1, \ldots, K\}$, $c > 0$ and $b \in [1, \bar{b}]$.

We can regard c as the country-specific scaling parameter that accounts for the size of the country or population density. We introduce this additional parameter in order to be able to compare countries that differ substantially in terms of the population, i.e. Luxembourg and Russia. We can regard b as a time parameter that is responsible for the speed of the development of the pandemic. If we compare two countries i and j that belong to the same class k but have different time parameters, b_i and b_j respectively, and $b_i > b_j$, then we can say that country i experiences a more rapid development of the

pandemic relative to country j even though the general shapes of the regression functions m_i and m_j are the same.

In what follows, we present a method that allows researchers to discover the group structure of the time trends of new COVID-19 cases in different countries and to cluster the countries based on the shape of the respective regression functions.

2 Clustering procedure

Let i and j be two countries from our sample. In this subsection we construct a dissimilarity measure $\widehat{\Delta}_{ij}$ that estimates how far the functions m_i and m_j are from each other. This dissimilarity measure $\widehat{\Delta}_{ij}$ will serve as a distance measure between the functions m_i and m_j in our clustering algorithm later on.

Step 1

First, for each i we nonparametrically estimate $m_i(u)$ using Nadaraya-Watson estimation procedure with a rectangular kernel and a bandwidth window that covers 7 data points, i.e. 7 days. This choice of a bandwidth allows us to take care of possible weekly cycles in the data which are produced by delays in reporting new cases over the weekend. As a robustness check, we repeat our analysis for the multiples of 7 days, i.e. for bandwidth windows covering 14 and 21 days. We report the results of the robustness checks in the Supplementary Material.

Formally, the estimator $\hat{m}_i$ estimated at $u \in [0, 1]$ is defined as

$$\hat{m}_i(u) = \sum_{t=1}^{T} \frac{K_h(u - t/T)Y_{it}}{\sum_{s=1}^{T} K_h(u - s/T)}$$

with $K_h(u - t/T)$ being a rectangular kernel: $K_h(x) = \frac{1}{2}$ for $|x| \leq h$ and $K_h(x) = 0$ otherwise.

Step 2

Second, for a given value of $b \in [1, \bar{b}]$ and for a given pair of countries (i, j), consider the following statistic:

$$\delta_{ij}(b) = \frac{1}{1/b} \int_0^{1/b} \left(\frac{\hat{m}_i(b \cdot u)}{\int_0^{1/b} \hat{m}_i(b \cdot v)dv/(1/b)} - \frac{\hat{m}_j(u)}{\int_0^{1/b} \hat{m}_j(v)dv/(1/b)} \right)^2 du.$$

$\delta_{ij}(b)$ can be regarded as a measure of dissimilarity between $m_i(b \cdot u)$ and $m_j(u)$ on an interval $[0, 1/b]$ for a specific value of b. Note that generally speaking $\delta_{ij}(b) \neq \delta_{ji}(b)$ for

2

$i \neq j$.

Step 3

We now construct a dissimilarity measure between two countries i and j that does not depend on a specific choice of a time parameter b. In order to do so, we would like to aggregate $\delta_{ij}(b)$ for different values of b. If the functions m_i and m_j belong to the same class, then for some $b_0 \in [1, \bar{b}]$ we will have $\delta_{ij}(b_0)$ close to zero. Need to show that! Hence, we aggregate the measures $\delta_{ij}(b)$ by taking the infimum over all possible values of b:

$$\widehat{\Delta}_{ij} = \min\{ \inf_{b \in [1,\bar{b}]} \delta_{ij}(b), \inf_{b \in [1,\bar{b}]} \delta_{ji}(b) \}$$

Step 4

Based on $\widehat{\Delta}_{ij}$, we run a hierarchical agglomerative clustering (HAC) algorithm using complete linkage criterion. Detailed description of the algorithm and its properties are presented in Section 3.

3 Clustering algorithm

Let $S \subseteq \{1, \ldots, n\}$ and $S' \subseteq \{1, \ldots, n\}$ be two sets of time series from our sample. There are several ways to define a dissimilarity measure between S and S'. In our paper, we work with the complete linkage measure of dissimilarity defined as

$$\Delta(S, S') = \max_{i \in S, j \in S'} \widehat{\Delta}_{ij}.$$

Alternatively, we may use single or average linkage measure

This description I copied from your paper.

To partition the set of time series $\{1, \ldots, n\}$ into groups, we combine the dissimilarity measure Δ with a HAC algorithm which proceeds as follows:

Algorithm (HAC Algorithm).

Step 0 (Initialization): Let $\hat{G}_i^{[0]} = \{i\}$ denote the ith singleton cluster for $1 \leq i \leq n$ and define $\{\hat{G}_1^{[0]}, \ldots, \hat{G}_n^{[0]}\}$ to be the initial partition of time series into clusters.

Step r (Iteration): Let $\hat{G}_1^{[r-1]}, \ldots, \hat{G}_{n-(r-1)}^{[r-1]}\}$ be the $n - (r-1)$ clusters from the previous step. Determine the pair of clusters $\hat{G}_k^{[r-1]}$ and $\hat{G}_{k'}^{[r-1]}$ for which

$$\Delta(\hat{G}_k^{[r-1]}, \hat{G}_{k'}^{[r-1]}) = \min_{1 \leq l < l' \leq n-(r-1)} \Delta(\hat{G}_l^{[r-1]}, \hat{G}_{l'}^{[r-1]})$$

and merge them into a new cluster.

Iterating this procedure for $r = 1, \ldots, n-1$ yields a tree of nested partitions $\{\hat{G}_1^{[r]}, \ldots, \hat{G}_{n-r}^{[r]}\}$, which can be graphically represented by a dendrogram. Roughly speaking, the HAC algorithm merges the n singleton clusters $\hat{G}_i^{[0]} = \{i\}$ step by step until we end up with the cluster $\{1, \ldots, n\}$. In each step of the algorithm, the closest two clusters are merged, where the distance between clusters is measured in terms of the dissimilarity Δ We refer the reader to Ward (1963) for an early reference on HAC clustering and to Section 14.3.12 in Hastie et al. (2009) for an overview of hierarchical clustering methods.

4 Application

We now use our clustering procedure to analyze the outbreak patterns of the COVID-19 epidemic. We proceed in two steps. In Section 4.1, we assess the finite sample performance of our method by Monte-Carlo experiments. In Section 4.2, we apply the method to a sample of COVID-19 data for 104 different countries.

4.1 Simulation

4.2 Analysis of the COVID-19 data

4.2.1 Data

We analyze data from 104 countries. We chose only those countries that have a total number of not less than $20\,000$ cases of infection during the considered time period. For each country i, we observe a time series $\mathcal{Y}_i = \{Y_{it} : 1 \leq t \leq T\}$, where Y_{it} is the number of newly confirmed COVID-19 cases in country i on day t. The data are freely available on the homepage of the European Center for Disease Prevention and Control (`https://www.ecdc.europa.eu`) and were downloaded on 25 February 2021.[1] As already mentioned in the Introduction, we take the first Monday after reaching 100 confirmed cases in each country as the starting date $t = 1$. Beginning the time series of each country on the day when that country reached 100 confirmed cases is a common way of "normalizing" the data (see e.g. Cohen and Kupferschmidt, 2020). Additionally aligning the data by Monday allows to take care of possible weekly cycles in the data which are produced by delays in reporting new cases over the weekend. The time series length T is taken to be

---

[1] ECDC switched to a weekly reporting schedule for the COVID-19 situation on 17 December 2020. Hence, all daily updates have been discontinued from 14 December. The downloaded daily data set presents historical data until 14 December 2020.

the longest interval for which we have observations for all 104 countries. The resulting dataset thus consists of $n = 104$ time series, each with $T = 192$ observations. Some of the time series contain negative values which we replaced by 0. Overall, this resulted in 14 replacements.

We assume that the data Y_{it} of each country i in our sample follow the nonparametric trend model

$$Y_{it} = m_i\left(\frac{t}{T}\right) + u_{it},$$

which was introduced in equation (1.1). As in the theoretical part of the paper, we assume that all of the countries i can be partitioned into K classes $\mathcal{G}_1, \ldots, \mathcal{G}_K$ such that for each $1 \leq k \leq K$ and for all $i \in \mathcal{G}_k$, $m_i(u) = c \cdot g_k(b \cdot u)$ for some $c > 0$ and $b \in [1, \bar{b}]$. We thus suppose that the general shape of the time trends m_i are the same (or at least very similar) in all countries i in a given group $\mathcal{G}_k$, but the scale and time parameters may vary across countries from the same group. We take $\bar{b} = 2$ which is more than enough for our purposes. Note that $\bar{b}$ accounts for the largest value of the time parameter, and allowing for some countries to be twice as fast as some others is already a very generous assumption.

Throughout the section, we set implement the clustering procedure in exactly the same way as in the simulation study of Section 4.1. In particular, we use a rectangular kernel K_h to compute the local linear smoothers $\hat{m}_i$ and consider the bandwidth $h = 3.5/T$, which corresponds to an effective sample size of 7 days of data. Moreover, we assume that the number of classes K is equal to 7.

Furthermore, the values $\delta_{ij}(b)$ and $\delta_{ji}(b)$ are calculated not for all $b \in [1, \bar{b}]$, but on a grid that consists of the values $\{1, 1.01, 1.02, \ldots, \bar{b}\}$. Since this grid is finite, in what follows we will be working with $\min_{b \in [1,\bar{b}]} \delta_{ij}(b)$ and $\min_{b \in [1,\bar{b}]} \delta_{ji}(b)$ instead of $\inf_{b \in [1,\bar{b}]} \delta_{ij}(b)$ and $\inf_{b \in [1,\bar{b}]} \delta_{ji}(b)$.

The results of our algorithm are presented in Figures 1 - 3. First, Figure 1 presents the corresponding dendrogram of a HAC. Different colours of the countries correspond to the different classes they belong to (by our clustering algorithm). Then, Figures 2a - 2g present the trend functions of the countries that belong to the group $\mathcal{G}_1, \ldots, \mathcal{G}_6$ respectively. For comparability reasons, instead of plotting $\hat{m}_i(u)$, we opted for plotting $\hat{m}_i(\hat{b}_i \cdot u)/\hat{c}_i$ where $\hat{b}_i$ and $\hat{c}_i$ are computed according to (4.1) and (4.2) in the algorithm described below.

Step 1

First, define the following quantities:

$$
b_{ij} = \begin{cases} \arg\min_{b \in [1, \bar{b}]} \delta_{ij}(b), & \text{if } \min_{b \in [1, \bar{b}]} \delta_{ij}(b) \leq \min_{b \in [1, \bar{b}]} \delta_{ji}(b), \\ 1, & \text{otherwise;} \end{cases}
$$

$$
\text{and } b_{ji} = \begin{cases} \arg\min_{b \in [1, \bar{b}]} \delta_{ji}(b), & \text{if } \min_{b \in [1, \bar{b}]} \delta_{ji}(b) < \min_{b \in [1, \bar{b}]} \delta_{ij}(b), \\ 1, & \text{otherwise.} \end{cases}
$$

The idea behind these values is that our algorithm estimates the dissimilarity measure between countries i and j to be the smallest when their respective time parameters are equal to b_{ij} and b_{ji}. These quantities are calculated for all pairs i and j irrespective of countries i and j being in the same class. However, if the countries i and j are indeed in the same class k, we have that $m_i(u) = c_i \cdot g_k(b_i \cdot u)$ and $m_j(u) = c_j \cdot g_k(b_j \cdot u)$. Hence,

$$
m_i\left(\frac{u}{b_i}\right) / c_i = g_k(u) = m_j\left(\frac{u}{b_j}\right) / c_j,
$$

and thus,

$$
m_i(u)/c_i = m_j\left(\frac{b_i}{b_j} u\right) / c_j,
$$

or, alternatively,

$$
m_i\left(\frac{b_j}{b_i} u\right) / c_i = m_j(u)/c_j
$$

If $b_i \geq b_j$, then b_{ji} can be regarded as an approximation of b_i/b_j (and b_{ij} should be equal to 1), and if $b_i < b_j$, then b_{ij} can be regarded as an approximation of b_j/b_i (and b_{ji} should be equal to 1). In any case, $m_i(b_{ij}u)/c_i \approx m_j(b_{ji}u)/c_j$.

Step 2

Using the quantities from Step 1, we pick a representative time series i_k for each cluster k. If the cluster k consists only of one element, then the representative time series should be the only member of the cluster. This is, for example, the case for clusters 6 and 7 (see Figures 2f and 2g). However, if the number of elements in a cluster is greater than 1, the procedure of picking a representative country is slightly more compicated. The representative time series should be "the slowest" out of two countries in most of the the
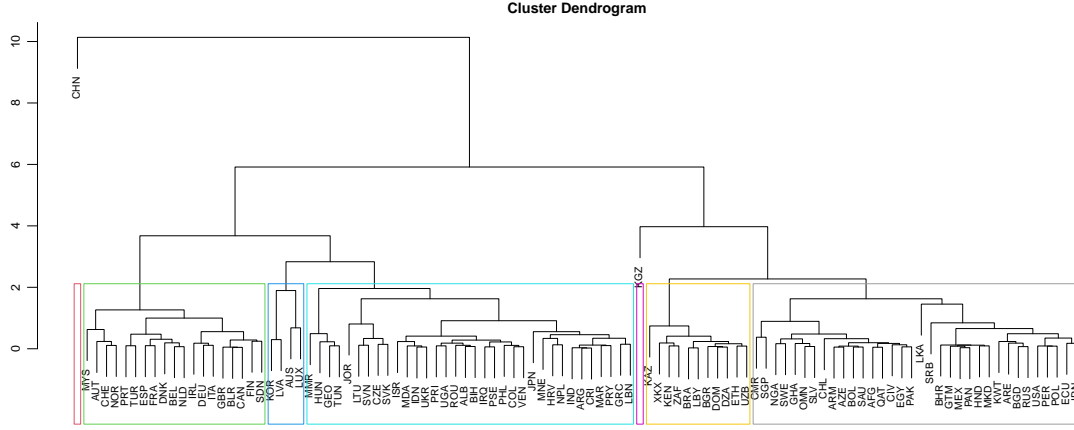
Figure 1: Results of HAC.

pairwise comparisons, or formally,

$$i_k = \arg \max_{i \in \widehat{\mathcal{G}}_k} \#\{j \in \widehat{\mathcal{G}}_k, \, j \neq i : b_{ij} = 1\}.$$

The representative time series i_k will be plotted on the whole rescaled time interval $[0, 1]$, whereas all of the other time series $j \in \widehat{\mathcal{G}}_k$, $j \neq i_k$ due to the higher or equal time parameter (and thus faster speed of the development of the pandemic) will be plotted on the (possibly smaller) time interval $\left[0, 1/\hat{b}_j\right]$. The representative time series in Figures 2a - 2g are shown in red, all of the other representatives in the respective cluster are shown in black.

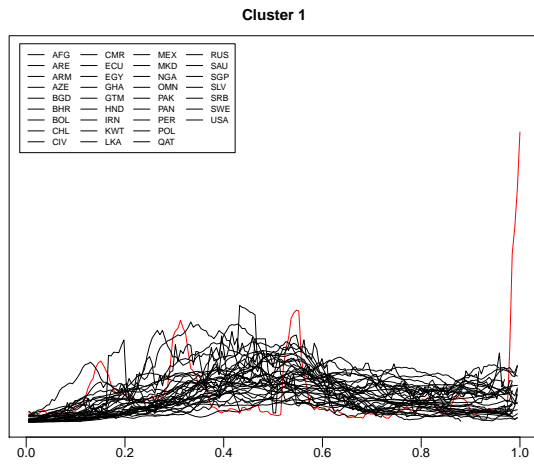The rule for determining the quantities $\hat{b}_j$ is as follows:

$$\hat{b}_j = \begin{cases} 1, & \text{if } j = i_k \text{ (representative country)}, \\ \min\left\{1, \frac{b_{ji_k}}{b_{i_k j}}\right\}, & \text{otherwise.} \end{cases} \tag{4.1}$$

Note that by construction $\hat{b}_j \geq 1$, hence, $\left[0, 1/\hat{b}_j\right] \subseteq [0, 1]$ for all j.
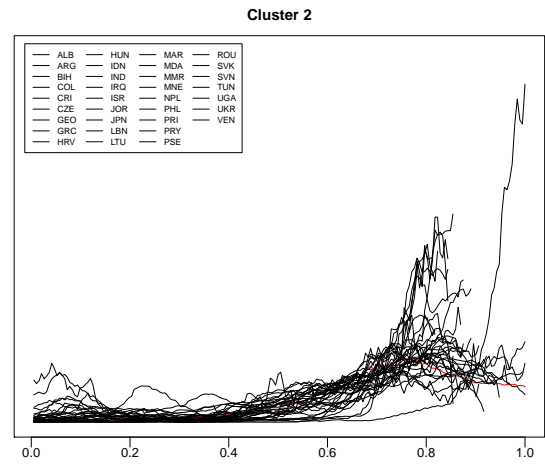
Step 3

Now we can calculate $\hat{c}_i$. As in the theoretical part of the paper, $\hat{c}_i$ can be regarded as a normalization parameter that accounts for the scale of the pandemic in country i:
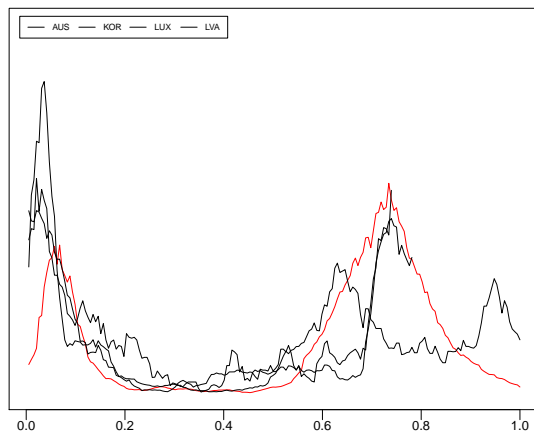
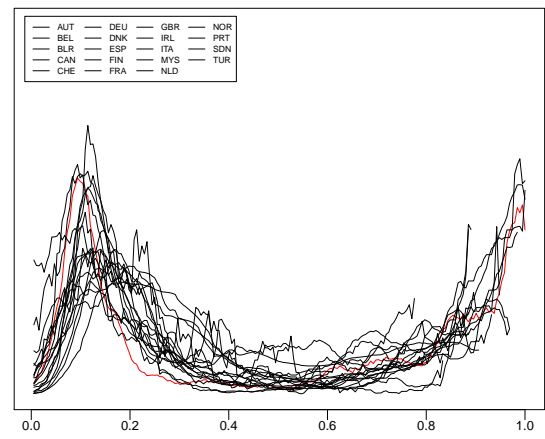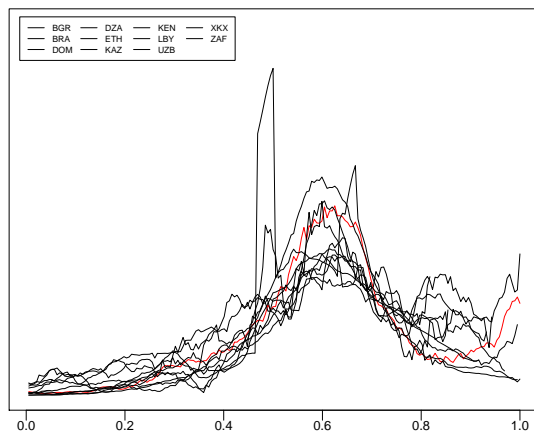$$\hat{c}_i = \frac{\int_0^{1/\hat{b}_i} \hat{m}_i(\hat{b}_i v) dv}{1/\hat{b}_i}. \tag{4.2}$$

7

Cluster 1

AFG	CMR	MEX	RUS
ARE	ECU	MKD	SAU
ARM	EGY	NGA	SGP
AZE	GHA	OMN	SLV
BGD	GTM	PAK	SRB
BHR	HND	PAN	SWE
BOL	IRN	PER	USA
CHL	KWT	POL	
CIV	LKA	QAT	

Cluster 2

ALB	HUN	MAR	ROU
ARG	IDN	MDA	SVK
BIH	IND	MMR	SVN
COL	IRQ	MNE	TUN
CRI	ISR	NPL	UGA
CZE	JOR	PHL	UKR
GEO	JPN	PRI	VEN
GRC	LBN	PRY	
HRV	LTU	PSE	

(a)

(b)

Cluster 3

AUS	KOR	LUX	LVA

Cluster 4

AUT	DEU	GBR	NOR
BEL	DNK	IRL	PRT
BLR	ESP	ITA	SDN
CAN	FIN	MYS	TUR
CHE	FRA	NLD	

(c)

(d)

Cluster 5

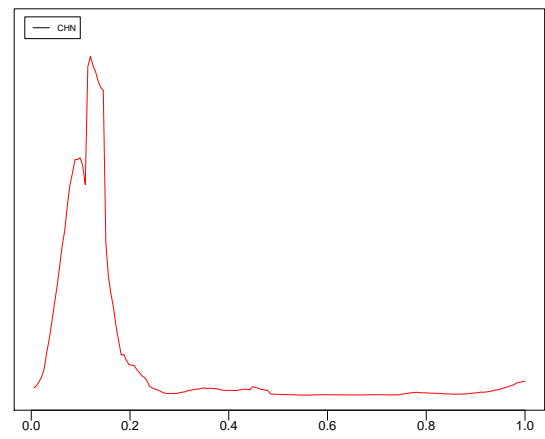| BGR	DZA	KEN	XKX
BRA	ETH	LBY	ZAF
DOM	KAZ	UZB	

Cluster 6

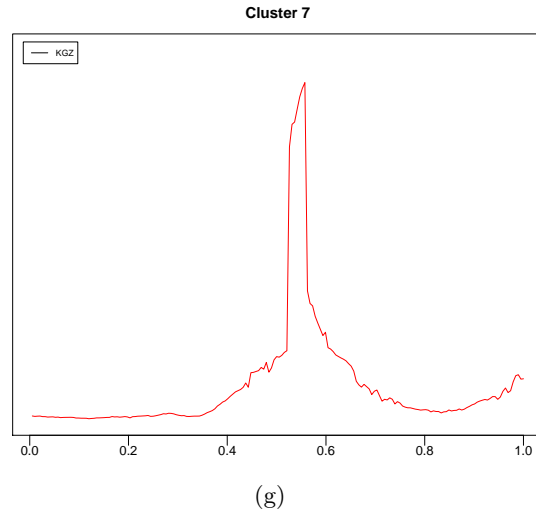| CHN |
|-----|

(e)

(f)

8

(g)

Figure 2: Clusters produced by the algorithm. Each panel presents appropriately scaled curve estimates $\widehat{m}_i$ that belong to a particular cluster.
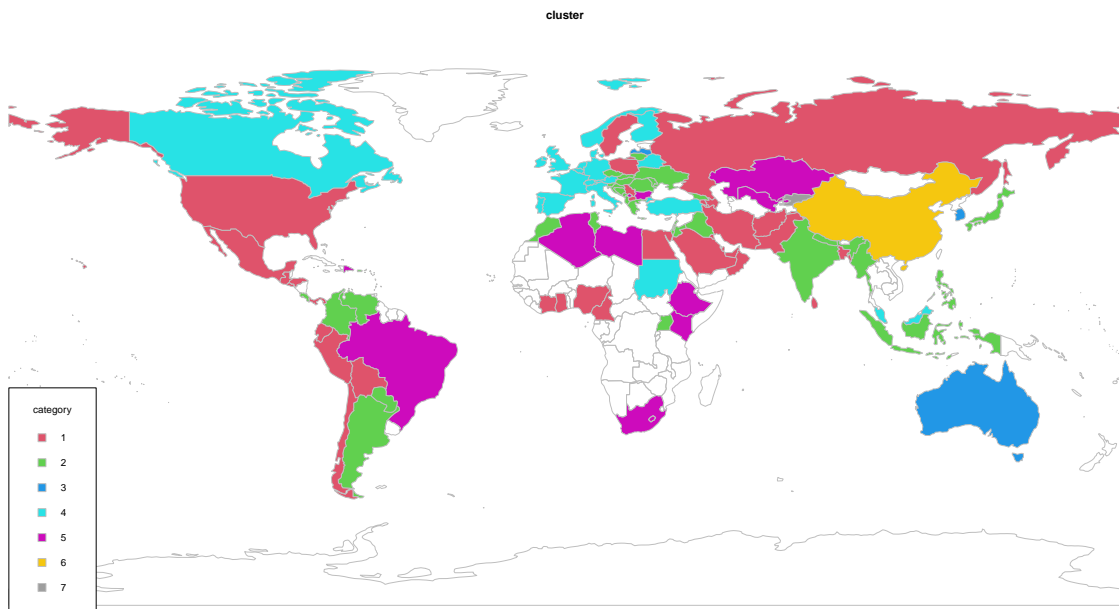

Figure 3: Results of HAC on a map: each country is coloured according to the group it belongs to.

References

COHEN, J. and KUPFERSCHMIDT, K. (2020). Countries test tactics in 'war' against COVID-19. *Science*, **367** 1287–1288.

HALE, T., PETHERICK, A., PHILLIPS, T. and WEBSTER, S. (2020a). Variation in government responses to COVID-19. *Blavatnik school of government working paper*, **31**.

HALE, T., WEBSTER, S., PETHERICK, A., PHILLIPS, T. and KIRA, B. (2020b). Oxford COVID-19 government response tracker. Blavatnik school of government. http://www.bsg.ox.ac.uk/covidtracker.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer.

ROBINSON, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical Analysis and Forecasting of Economic Structural Change* (P. Hackl, ed.). Springer, 253–264.

WARD, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, **58** 236–244.