RESEARCH ARTICLE

Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach [version 1; peer review: 1 approved, 1 approved with reservations]

Rodrigo M. Carrillo-Larco [1-3], Manuel Castillo-Cara [4]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
[2]CRONICAS Centre of Excellence in Chronic Diseases, Universidad Peruana Cayetano Heredia, Lima, Peru
[3]Instituto de Investigación, Universidad Católica Los Ángeles de Chimbote, Chimbote, Peru
[4]Center of Information and Communication Technologies, Universidad Nacional de Ingeniería, Lima, Peru

v1 **First published:** 31 Mar 2020, **5**:56
https://doi.org/10.12688/wellcomeopenres.15819.1
Second version: 04 Jun 2020, **5**:56
https://doi.org/10.12688/wellcomeopenres.15819.2
Latest published: 15 Jun 2020, **5**:56
https://doi.org/10.12688/wellcomeopenres.15819.3

Abstract

Background: The COVID-19 pandemic has attracted the attention of researchers and clinicians whom have provided evidence about risk factors and clinical outcomes. Research on the COVID-19 pandemic benefiting from open-access data and machine learning algorithms is still scarce yet can produce relevant and pragmatic information. With country-level pre-COVID-19-pandemic variables, we aimed to cluster countries in groups with shared profiles of the COVID-19 pandemic.
Methods: Unsupervised machine learning algorithms (k-means) were used to define data-driven clusters of countries; the algorithm was informed by disease prevalence estimates, metrics of air pollution, socio-economic status and health system coverage. Using the one-way ANOVA test, we compared the clusters in terms of number of confirmed COVID-19 cases, number of deaths, case fatality rate and order in which the country reported the first case.
Results: The model to define the clusters was developed with 155 countries. The model with three principal component analysis parameters and five or six clusters showed the best ability to group countries in relevant sets. There was strong evidence that the model with five or six clusters could stratify countries according to the number of confirmed COVID-19 cases ($p<0.001$). However, the model could not stratify countries in terms of number of deaths or case fatality rate.
Conclusions: A simple data-driven approach using available global information before the COVID-19 pandemic, seemed able to classify countries in terms of the number of confirmed COVID-19 cases. The model was not able to stratify countries based on COVID-19 mortality data.

Keywords

COVID-19, pandemic, clustering, k-mean, unsupervised algorithms

Open Peer Review

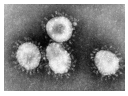
Reviewer Status ✔ ?

	Invited Reviewers	
	1	**2**
version 3 (revision) 15 Jun 2020		
version 2 (revision) 04 Jun 2020		? report
version 1 31 Mar 2020	✔ report	? report

1 **Alan E. Hubbard**, University of California, Berkeley, Berkeley, USA

2 **Nonie Alexander**, University College London, London, UK
Maria Pikoula, University College London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Coronavirus (COVID-19) collection.

Corresponding author: Rodrigo M. Carrillo-Larco (rcarrill@ic.ac.uk)

Author roles: Carrillo-Larco RM: Conceptualization, Data Curation, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Castillo-Cara M**: Conceptualization, Data Curation, Formal Analysis, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was funded by the Wellcome Trust. RMC-L has been supported by a Strategic Award, Wellcome Trust-Imperial College Centre for Global Health Research (100693), and Imperial College London Wellcome Trust Institutional Strategic Support Fund [Global Health Clinical Research Training Fellowship] (294834 ISSF ICL). RMC-L is supported by a Wellcome Trust International Training Fellowship (214185).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Carrillo-Larco RM and Castillo-Cara M. **Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2020, **5**:56 https://doi.org/10.12688/wellcomeopenres.15819.1

First published: 31 Mar 2020, **5**:56 https://doi.org/10.12688/wellcomeopenres.15819.1

Introduction

The ongoing COVID-19 pandemic has attracted the attention and interest of public health officers, practitioners, researchers and the general population. They all are working together to slow down the spread of the disease, thus reducing the number of severe cases and deaths. Their efforts have already produced relevant preliminary information on COVID-19 risk factors and the epidemiological profile of the disease[1–3], with plenty more information not published yet (e.g., academic pre-prints).

The available evidence—published and unpublished—has mostly focused on the individual level; that is, they have studied the patients, their characteristics, disease progression and outcomes. Little has been studied about large populations and geographic areas; in other words, ecological evidence and research addressing study units other than the patients are scarce, though can reveal relevant and pragmatic information. In this line, research with novel analytical approaches, such as machine learning algorithms, is also uncommon.

Research at the country level could reveal potentially modifiable associated factors that individual-level data is still unable to study because of the limited number of observations. Moreover, machine learning techniques informed by country-level variables can provide prediction and classification algorithms useful to understand how countries may behave during and after the COVID-19 pandemic. Consequently, we aimed to develop a simple unsupervised machine learning algorithm informed by country-level variables before the COVID-19 pandemic, that can classify countries regarding the number of confirmed COVID-19 cases and deaths. In so doing, we provide a preliminary framework to stratify countries with similar progression through the COVID-19 pandemic.

Methods

Data sources

We used different data sources to build a dataset with information on COVID-19, prevalence estimates of selected diseases, a socio-economic metric, an air pollution metric, and a metric of health system coverage (Table 1). The unit of analysis was a country. Variables and specific data sources are shown in Table 1. Except for the COVID-19 variables, the other variables were used in the clustering analysis. In other words, countries were clustered following unsupervised machine learning algorithms based on prevalence estimates of the selected diseases, socio-economic status, air pollution and health system coverage (Table 1).

These predictors were selected because they are closely related to the COVID-19 pandemic, both from a clinical and public health perspective. We chose two chronic non-communicable diseases (diabetes and chronic obstructive pulmonary disease [COPD]) and two infectious diseases (tuberculosis and HIV/IDS). Diabetes seems to be very frequent among COVID-19 patients[4]. Although hypertension had a higher frequency than respiratory diseases[4], we chose COPD because of the structural and pathophysiological pathways it can share with an acute respiratory disease such as COVID-19; the same logic would apply for tuberculosis. We chose HIV/AIDS because of the high potential of impaired immune response. We chose 2.5 particulate matter (particles of width <2.5 μm) as a metric of air pollution; 2.5 particulate matter has been related to severe acute respiratory syndrome[5]. Finally, we chose a metric of socio-economic status and health system coverage, which could impact on the probability of a persona to adopt preventive care and access to appropriate healthcare should it be necessary.

Table 1. Extracted data, variables and data sources.

Concept	Variables	Data source	Used for
COVID-19 prevalence	Country; number of confirmed cases (as of 23/03/2020); number of confirmed deaths (as of 23/03/2020); case fatality rate per 1,000 cases (as of 23/03/2020); order number at which the country experienced the first case (e.g., 1st country, 2nd country…)	COVID-19 global surveillance system by Johns Hopkins Univeristy[16,17]	Cluster evaluation
Disease prevalence	Age-standardized prevalence of diabetes, chronic obstructive disease [COPD], HIV/AIDS and tuberculosis (as of year 2017)	2017 Global Burden of Disease / Institute for Health Metrics, Washington University[18]	Clustering
Air quality metric	Concentration of 2.5 particulate matter by country	Global Health Observatory data repository, World Health Organization[19]	Clustering
Socio-economic metric	Gross domestic product per capita (as of year 2017)[a]	World Bank[20]	Clustering
Health system metric	Universal health coverage index of service coverage (as of year 2017)	Global Health Observatory data repository, World Health Organization[21]	Clustering

[a]When a country did not have data for 2017, we used the latest available; when a country did not have any data on this source, we used data as reported by a Google search (this was the case for four countries).

Data analysis – clustering

Predictors. The variables used to develop the clustering model had different values between them, thus each of them carries a different variance. Because of this characteristic, it is relevant to standardize these variables to set reliable clusters without losing information. Consequently, before running the unsupervised clustering algorithms, the predictors were treated with an orthogonal transformation and then with principal component analysis (PCA).

PCA. The PCA is a technique within the remit of unsupervised machine learning algorithms. PCA follows an orthogonal transformation, which turns correlated variables into an uncorrelated set of variables. The PCA aims to create a set of characteristics, or components, that represents the relevant information from the original group of variables[6,7]. The PCA seeks to reduce the number of predictors while maximizing the variance.

In this work, and to avoid losing information explained by the original predictors, we prespecified three PCA components; of the three components, the third one had an explained variance of 1. This method of obtaining 1 as an explained variance imply keeping 100% of the information explained by the original predictors. Moreover, these three components gave the most reliable clusters as reported in the results section. We used the PCA algorithm available in the Scikit-Learn library[8].

K-means. This technique seeks to group heterogenous elements into homogenous clusters. This approach is considered a paradigm in unsupervised machine learning, because it assigns the elements into clusters which were unknown at the beginning of the analysis[9]. A few authors have used this methodology in clinical and public health research[10–13].

There are different methods for unsupervised clustering depending on the data characteristics[14]. Given our data and aims, we chose a centroid-based algorithm: k-means. This approach works well when the clusters have similar size, similar densities and follow a globular shape.

Regarding the number of clusters that optimizes the function convergence to the centroids, we estimated a cost function which supported the choose of five and six clusters (Figure 1). This function cost, paired with the overall results (results section), suggested that five or six cluster was an adequate decision. We used the k-mean algorithm available in the Scikit-Learn library, with five and six clusters, 500 iterations, and a fast initiation of convergence with k-mean++[15].

Statistical analysis

The COVID-19 variables—number of confirmed cases, number of deaths, case fatality rate and order when the first case appeared—were compared across clusters with the one-way ANOVA tests. Within clusters, pairwise combinations were analysed with t-tests adjusted for multiple comparisons with the Bonferroni method. The statistical analysis was conducted with COVID-19 data until March 23rd, 2020. Analysis was performed in R (v3.6.1).

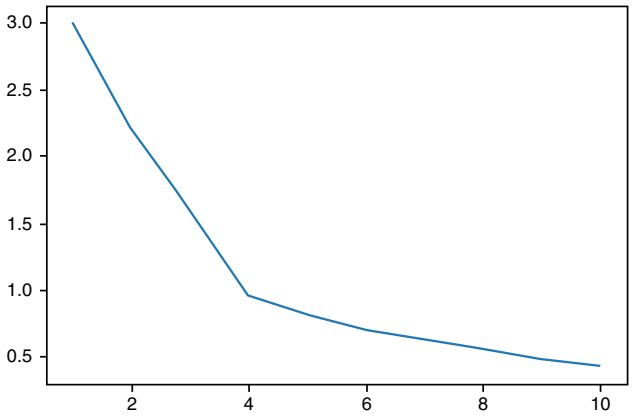


Figure 1. Cost function for the k-mean analysis.

Ethics

This work analysed open-access data and did not involve any human subjects. No approval by an IRB or ethics committee was sought.

Results

Data points

The clustering models were built with 155 countries and territories. Based on visual inspection of maps and boxplots, and on statistical parameters, the clustering models with three PCA components and five (Figure 2A) or six (Figure 2B) clusters performed the best to stratify countries according to COVID-19 variables (Figure 3; data available with the manuscript).

Clusters prediction

The one-way ANOVA test comparing the confirmed number of COVID-19 cases across the five and six clusters, strongly suggested there was a difference between groups ($p<0.001$). Regarding the model with five clusters, the strongest differences were between clusters 0 and 1, 0 and 4, 1 and 2, 2 and 3, as well as 2 and 4 (Figure 3, Table 2). Similarly, for the model with six clusters there were ten pairwise combinations with strong differences in the number of confirmed COVID-19 cases (Figure 3, Table 2).

The proposed clustering with five groups did not stratify well according to number of total deaths ($p=0.067$); adding one more cluster did not improve the prediction ($p=0.864$). None of the pairwise combinations revealed a strong difference (Figure 3, Table 2). Overall, the same findings applied to case fatality rate for five ($p=0.320$) and six ($p=0.373$) clusters, with no differences in pairwise comparisons (Figure 3, Table 2).

There was strong difference among cluster regarding the order at which each country had the first confirmed case, regardless of the number of clusters ($p<0.001$). For the model

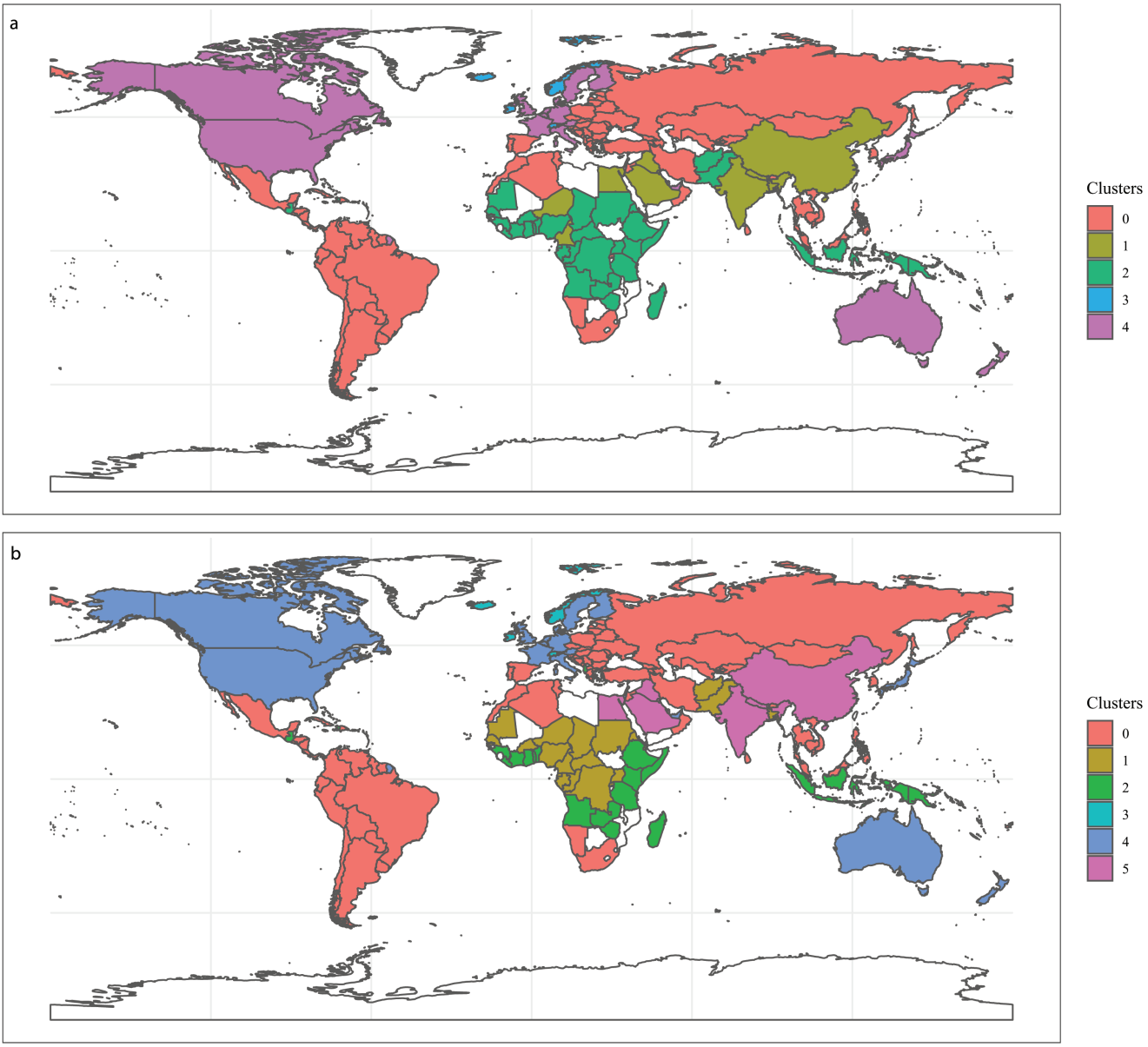


Figure 2. World map showing countries coloured as per the model with five (**A**) and six (**B**) clusters.

with five clusters, there were strong pairwise differences in all but four pairs (Figure 3, Table 2). In a similar line, eight of the pairwise combinations in the model with six clusters revealed a strong difference (Figure 3, Table 2)

Discussion

Main results

Based on open-access variables at the country level, along with unsupervised machine learning algorithms (k-means), we developed a clustering model that can classify countries well regarding the number of confirmed COVID-19 cases. However, the model did not stratify countries well according to the number of deaths or case fatality rate.

The clustering model we proposed has potential applications. First, for each cluster we report a median and a range of number of confirmed COVID-19 cases. Although still early and deserving of further scrutiny as the outbreak progresses, the results could suggest that the number of cases in one country in one cluster will be within the proposed range for that cluster, unless one country performs below the expectation (i.e., exceeds the proposed range).

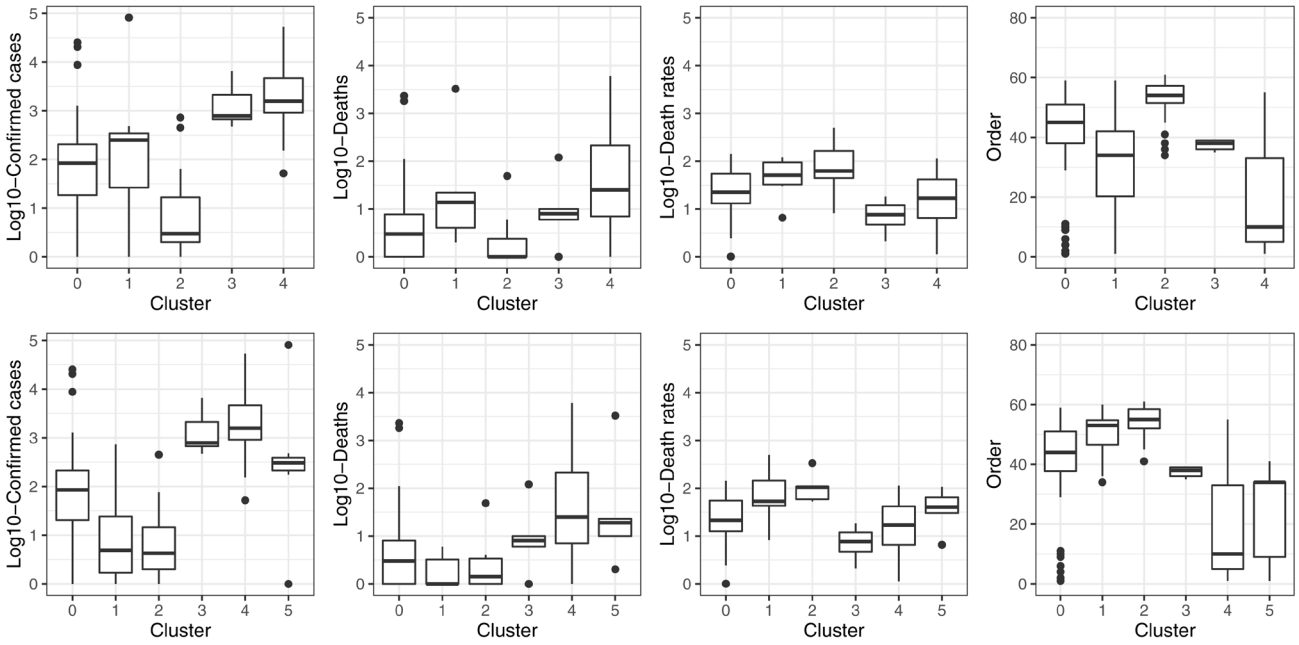


Figure 3. Boxplots showing the distribution of COVID-19 pandemic variables across clusters.

Table 2. Pairwise combinations between clusters according to COVID-19 variables (as of March 23rd, 2020).

	Number of confirmed cases					Number of confirmed cases				
Clusters	**0**	**1**	**2**	**3**	**Clusters**	**0**	**1**	**2**	**3**	**4**
1	1.000				**1**	<0.001				
2	<0.001	<0.001			**2**	<0.001	1.000			
3	0.023	0.300	<0.001		**3**	0.034	<0.001	<0.001		
4	<0.001	0.003	<0.001	1.000	**4**	<0.001	<0.001	<0.001	1.000	
					5	0.771	<0.001	<0.001	1.000	0.270
	Number of deaths					**Number of deaths**				
Clusters	**0**	**1**	**2**	**3**	**Clusters**	**0**	**1**	**2**	**3**	**4**
1	1.000				**1**	1.000				
2	1.000	1.000			**2**	1.000	1.000			
3	1.000	1.000	1.000		**3**	1.000	1.000	1.000		
4	0.110	1.000	0.096	1.000	**4**	0.180	0.320	0.290	1.000	
					5	1.000	1.000	1.000	1.000	1.000
	Case fatality rate per 1,000 cases					**Case fatality rate per 1,000 cases**				
Clusters	**0**	**1**	**2**	**3**	**Clusters**	**0**	**1**	**2**	**3**	**4**
1	1.000				**1**	0.460				
2	0.430	1.000			**2**	1.000	1.000			
3	1.000	1.000	1.000		**3**	1.000	1.000	1.000		
4	1.000	1.000	1.000	1.000	**4**	1.000	1.000	1.000	1.000	
					5	1.000	1.000	1.000	1.000	1.000
	Order					**Order**				
Clusters	**0**	**1**	**2**	**3**	**Clusters**	**0**	**1**	**2**	**3**	**4**
1	0.123				**1**	0.064				
2	<0.001	<0.001			**2**	<0.002	1.000			
3	1.000	1.000	0.198		**3**	1.000	0.649	0.169		
4	<0.001	0.040	<0.001	0.025	**4**	<0.001	<0.001	<0.001	0.007	
					5	0.004	<0.001	<0.001	1.000	0.856

Cells in red show not significant results (p>0.05); cells in yellow show significant results (p<0.05 & p>0.001); cells in green show strong significant results (p<0.001)

Unless there are substantial changes in the predictors used to define the clusters, these could signal countries that are particularly vulnerable or resilient for future respiratory outbreaks of this kind. Future research in a similar situation can test whether the proposed clusters also stratify countries well regarding the number of cases. Alternatively, the model could be tested with data of old respiratory pandemics to assess if it would have classified countries well.

Overall, considering the limitations of this work, the stage of the ongoing COVID-19 pandemic, and the general knowledge about this disease and its epidemiological profile, we provided a preliminary clustering model that could be useful to understand similarities and differences across countries, and how they may be affected by the ongoing pandemic.

Results in context

We are unaware of other studies that have aimed to classify countries based on simple open-access variables, and that can stratify the countries based on the number of COVID-19 cases. Most of the previous research using unsupervised machine learning clustering algorithms on health research has focused on individuals and diseases[10–13]. This work complements the available evidence at the individual level with preliminary information on clusters at the country level, with potential relevant applications in the current COVID-19 pandemic. Nevertheless, future research should verify the accuracy and stability of our findings, so that they can be applied for this and future similar scenarios.

Strengths and limitations

We proposed a simple algorithm to classify countries regarding the number of confirmed COVID-19 cases. In that sense, this model and others can be easily applied and developed. However, there are limitations to acknowledge. First, one could argue that there were few predictors to define the clusters. However, these were relevant variables that are freely available for research and analysis. Moreover, finding reliable, consistent and comparable information for all -or most- countries in the world may be challenging. This calls to researchers and international organizations to produce more information at the country level following similar methods that will allow global comparisons and analysis. Second, we did not find any strong evidence for the total number of deaths or case fatality rate. This could be because there are, fortunately, still very few deaths in most countries precluding strong comparisons. Our model can be tested again in the future, when the outbreak ends and there would be potentially more deaths, to assess whether the performance on this outcome improves. Third, we based our analysis on the confirmed number of cases and deaths. It is expected that this number may not reflect the actual number of people with the disease. In other words, it is more likely that there are more COVID-19 cases that have not been diagnosed or confirmed. This could be a limitation if we had aimed to predict the exact number of sick people, in which case we should have somehow accounted for the under-reporting.

Conclusions

Using readily available variables we developed an unsupervised machine learning algorithm that can stratify countries based on the number of COVID-19 confirmed and reported cases. This preliminary work provides a timely algorithm that could help identify countries more vulnerary or resistant to the ongoing pandemic.

Data availability

Source data

The source data for this study are described in Table 1.

Extended data

Figshare: Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. https://doi.org/10.6084/m9.figshare.12030363.v1[22].

This project contains the following extended data:

- Datasets.zip (containing the pooled data used in this analysis).
- Codes.zip (containing codes used in the analysis to develop the cluster and to assess its performance).

Author contributions

RMC-L conceived the idea with support of MC-C. RMC-L pooled the data. MC-C conducted the clustering analysis. RMC-L conducted the statistical analysis. RMC-L drafted the manuscript with input from MC-C. Both authors approved the submitted version.

References

1. Chan JF, Yuan S, Kok KH, *et al.*: **A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.** *Lancet.* 2020; **395**(10223): 514–23.
PubMed Abstract | Publisher Full Text

2. Chen N, Zhou M, Dong X, *et al.*: **Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study.** *Lancet.* 2020; **395**(10223): 507–13.
PubMed Abstract | Publisher Full Text

3. Huang C, Wang Y, Li X, *et al.*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.** *Lancet.* 2020; **395**(10223): 497–506.
PubMed Abstract | Publisher Full Text

4. Yang J, Zheng Y, Gou X, *et al.*: **Prevalence of comorbidities in the novel Wuhan**

coronavirus (COVID-19) infection: a systematic review and meta-analysis. *Int J Infect Dis.* 2020; pii: S1201-9712(20)30136-3.
PubMed Abstract | Publisher Full Text

5. Cui Y, Zhang ZF, Froines J, *et al.*: **Air pollution and case fatality of SARS in the People's Republic of China: an ecologic study.** *Environ Health.* 2003; **2**(1): 15.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Yang MS, Wu KL: **Unsupervised possibilistic clustering.** *J Pattern Recogn.* 2006; **39**: 5–21.
Publisher Full Text

7. Rodríguez-Sotelo JL, Delgado-Trejos E, Peluffo-Ordóñez D, *et al.*: **Weighted-PCA for unsupervised classification of cardiac arrhythmias.** *Conf Proc IEEE Eng Med Biol Soc.* 2010; **2010**: 1906–9.
PubMed Abstract | Publisher Full Text

8. Scikit learn: **sklearn.decomposition.PCA.**
Reference Source

9. Figueiredo MAT, Jain AK: **Unsupervised learning of finite mixture models.** *IEEE Trans Pattern Anal Mach Intel.* 2002; **24**(3): 381–96.
Publisher Full Text

10. Ahlqvist E, Storm P, Käräjämäki A, *et al.*: **Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables.** *Lancet Diabetes Endocrinol.* (2213-8595 (Electronic)). 2018; **6**(5): 361–369.
PubMed Abstract | Publisher Full Text

11. Carruthers SP, Gurvich CT, Meyer D, *et al.*: **Exploring Heterogeneity on the Wisconsin Card Sorting Test in Schizophrenia Spectrum Disorders: A Cluster Analytical Investigation.** *J Int Neuropsychol Soc.* (1469-7661 (Electronic)). 2019; **25**(7): 750–760.
PubMed Abstract | Publisher Full Text

12. Pikoula MA, Quint JK, Nissen F, *et al.*: **Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records.** *BMC Med Inform Decis Mak.* (1472-6947 (Electronic)). 2019; **19**(1): 86.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Sugihara G, Oishi N, Son S, *et al.*: **Distinct Patterns of Cerebral Cortical Thinning in Schizophrenia: A Neuroimaging Data-Driven Approach.** *Schizophr Bull.* (1745-1701 (Electronic)). 2017; **43**(4): 900906.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Fisher DH, Pazzani MJ, Langley P: **Concept Formation: Knowledge and Experience in Unsupervised Learning.** Elsevier Science; 2014.
Reference Source

15. Scikit learn: **sklearn.cluster.KMeans.**
Reference Source

16. **Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE.**
Reference Source

17. Dong E, Du H, Gardner L: **An interactive web-based dashboard to track COVID-19 in real time.** *Lancet Infect Dis.* 2020; pii: S1473-3099(20)30120-1.
PubMed Abstract | Publisher Full Text

18. Global Burden of Disease Collaborative Network: **Global Burden of Disease Study 2017 (GBD 2017) Results.** Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018.
Reference Source

19. World Health Organization: **Global Health Observatory data repository**.
Reference Source

20. **The World Bank. Data.**
Reference Source

21. World Health Organization: **Global Health Observatory data repository.**
Reference Source

22. Carrillo Larco R: **Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach.** *figshare.* Dataset, 2020.
http://www.doi.org/10.6084/m9.figshare.12030363.v1

Open Peer Review

Current Peer Review Status: ✔ ?

Version 1

Reviewer Report 27 May 2020

https://doi.org/10.21956/wellcomeopenres.17350.r38663

?

Nonie Alexander
Institute of Health Informatics, University College London, London, UK
Maria Pikoula
Institute of Health Informatics, University College London, London, UK

Carillo-Larco *et al*. used freely available data sources to perform a country-level cluster analysis of COVID-19 related variables. The resulting clusters were validated against outcomes related to mortality and confirmed cases. A statistically significant difference was observed between clusters with regards to number of confirmed cases. There was no correlation between cluster membership and mortality outcomes.

The study design and results are, for the most part, clearly presented, and the article is well-written. However, information is lacking with regards to both methodological aspects as well as the presented findings of the study. Most importantly, it is not clear what question the study is trying to answer.

Is the study design appropriate and is the work technically sound?

1) In terms of the appropriateness of the study, besides the lack of similar studies in the literature, no further justification is given as to why this study design was selected. If the purpose of the study is to allow for prediction of COVID-19 outcomes, a predictive model might have been more appropriate. The rationale for selecting cluster analysis is not sufficiently explained. Furthermore, the selection of input variables seems to be based on their availability rather than evidence from the literature that would make them suitable candidates for inclusion. The authors mention in the discussion that these variables are “relevant”, however this claim is not substantiated.

Are sufficient details of methods and analysis provided to allow replication by others?

2) The paragraph explaining the selection of principal components should be re-written as it is ambiguous whether the retention of three PCA components was pre-specified or whether keeping 100% of the explained variance was the original target. It is my understanding that four variables were used as input in the PCA and the first three components were selected, and that the three together explain 100% of the variance. It makes no sense for solely the third component to explain 100% of the variance, especially

given that the output of PCA lists components in descending order of % explained variance.

3) Related to the comment above, It is mentioned that “three components gave the most reliable clusters”. By which metric was reliability assessed? If this is to do with cluster stability, typically this entails re-sampling the data and verifying cluster stability with regards to the cluster characteristics using a metric such as the Jaccard coefficient[1].

4) The following sentence in the section labelled k-means needs rephrasing: “Regarding the number of clusters that optimises the function convergence to the centroids, we estimated a cost function which supported the choose of five and six clusters”. At the moment it is not clear which cost-function is being referred to and what is meant by estimating a cost function. I suspect the authors are referring to the standard k-means cost function, the sum of squared distances from each point’s cluster centre.

5) It is not clear how the choice of 5 or 6 clusters was made. According to the elbow plot in Figure 1, the elbow point is at 4 clusters. It is also unclear how the clustering results were used for the purpose of selecting k “based on visual inspection of maps and boxplots”. The maps in Figure 2 are fairly similar between the 5- and 6-cluster solutions and the boxplots in Figure 3 also suggest that clusters 0, 3 and 4 remain the same with some countries in clusters 1, 2 of the 5-cluster solution redistributed between them and with the additional cluster 5 in the 6-cluster solution.

6) There are more reliable metrics to aid with cluster selection, including the silhouette coefficient[2], and the GAP statistic[3]. The elbow plot is simply a heuristic. The authors should at least explain their choice of method.

If applicable, is the statistical analysis and its interpretation appropriate?

7) Although appropriate, the statistical analysis lacks further interpretation. The usefulness of the model could be illustrated by evaluating the predictive value of cluster labels to answer the question “Are the labels more predictive than individual variables?”

8) The resulting clusters are difficult to interpret without a summary table of cluster characteristics in terms of the 4 input variables used in the analysis.

Are the conclusions drawn adequately supported by the results?

9) No specific conclusions are drawn in the discussion. What are the cluster characteristics and how are they associated with confirmed COVID-19 cases? Are the results expected, surprising? There is little discussion on the characteristics, whether present or absent in the model, that would drive the countries to cluster together with regards to the number of reported cases. A few example points for discussion are listed below.

10) It appears from the map distribution that the clusters loosely correlate with GDP - although without a summary table confirming this is hard to tell for certain. I am not an epidemiologist and neither is NA, therefore it is not our area to comment, but countries with higher GDP are more likely to perform more tests, and are thus more likely to have a higher number of cases.

11) Additionally, some countries are more connected than others (e.g. because of air travel), and the spread of COVID-19 is not uniform across the world (e.g. countries that are closer to China reported cases earlier) and therefore, different countries are at different stages of the pandemic. It would make

more sense to separately cluster countries with similar exposure to the virus as well as comparable reporting standards.

Minor edits:
1. Figure 1 needs axis labels.

References

1. Rousseeuw P: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987; **20**: 53-65 Publisher Full Text
2. Fletcher S, Islam M: Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems*. 2018; **22**. Publisher Full Text
3. Tibshirani R, Walther G, Hastie T: Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2001; **63** (2): 411-423 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cluster analysis; phenotype discovery; airways disease; health informatics.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 29 May 2020

Rodrigo M Carrillo-Larco, Imperial College London, London, UK

Reviewer #2

Q1. The study design and results are, for the most part, clearly presented, and the article

is well-written. However, information is lacking with regards to both methodological aspects as well as the presented findings of the study. Most importantly, it is not clear what question the study is trying to answer.
A1. We appreciate the comprehensive evaluation; the comments will most certainly improve our work. We have included more details about the methodology (please refer to answers 4, 5 and 6); moreover, we have further elaborated on the results and discussion (please refer to answers 7, 8, 9, 10 and 11).

More than pursuing a specific research question, we aimed to develop a classification model that, benefiting from simple and available ecological variables, could cluster countries according to COVID-related outcomes (number of cases and deaths). If anything, our research question would be: *can country characteristics before the COVID-19 pandemic be useful to cluster countries according to COVID-19 number of cases and deaths?* We have modified the last paragraph of the introduction to include this question.

Q2. In terms of the appropriateness of the study, besides the lack of similar studies in the literature, no further justification is given as to why this study design was selected. If the purpose of the study is to allow for prediction of COVID-19 outcomes, a predictive model might have been more appropriate. The rationale for selecting cluster analysis is not sufficiently explained. Furthermore, the selection of input variables seems to be based on their availability rather than evidence from the literature that would make them suitable candidates for inclusion. The authors mention in the discussion that these variables are "relevant", however this claim is not substantiated.
A2. We agree that lack of evidence is not a strong justification, and we acknowledge we were not clear on our motivations. These have been further elaborated in the last paragraph of the introduction; these lines read: *Therefore, classification algorithms can reveal patterns to identify countries where the pandemic may have a similar effect. Countries could use this information to prevent worse-case scenarios given the cluster to which they belong. Global and regional organizations could use country clusters to organize similar aid to countries in the same cluster while prioritizing clusters likely to experience the worse outcomes.*

We certainly included variables that were readily available. However, we also chose variables that were closely related to the COVID-19 pandemic. The rationale behind our variable selection was explained in the paragraph immediately before the "Data analysis–clustering" sub-heading. In these lines, we elaborated on why we chose the selected variables, what their relationship may be with COVID-19, and why we did not choose other variables that could have been available as well. References were included to support our statements.

Q3. The paragraph explaining the selection of principal components should be re-written as it is ambiguous whether the retention of three PCA components was pre-specified or whether keeping 100% of the explained variance was the original target. It is my understanding that four variables were used as input in the PCA and the first three components were selected, and that the three together explain 100% of the variance. It makes no sense for solely the third component to explain 100% of the variance, especially given that the output of PCA lists components in descending order of % explained variance.
A3. We apologise for the misunderstanding, as it was the consequence of a miscommunication. A priori, we decided on three PCA variables. We included eight input variables (please refer to answer 8) and applied the PCA. As you inferred correctly, these three PCA variables retained or

explained 100% of the variance. As you correctly pinpointed, it made no sense for solely the third component to explain 100% of the variance. We have modified the text in the “PCA” sub-heading to better reflect this procedure: *In this work, and to avoid losing information explained by the original eight predictors, we prespecified three PCA components; the three PCA components retained a variance of 1. This method of obtaining 100% as an explained variance imply keeping 100% of the information explained by the original eight predictors.*

Q4. The following sentence in the section labelled k-means needs rephrasing: “Regarding the number of clusters that optimises the function convergence to the centroids, we estimated a cost function which supported the choose of five and six clusters”. At the moment it is not clear which cost-function is being referred to and what is meant by estimating a cost function. I suspect the authors are referring to the standard k-means cost function, the sum of squared distances from each point’s cluster centre.
A4. We referred to the “elbow” plot (Figure 1). We have rephrased this sentence to make it clearer, that we were talking about the “elbow” plot in figure 1. Please, refer to answers 5 and 6 for details about other modifications made regarding the analysis and cluster selection.

Q5. It is not clear how the choice of 5 or 6 clusters was made. According to the elbow plot in Figure 1, the elbow point is at 4 clusters. It is also unclear how the clustering results were used for the purpose of selecting k “based on visual inspection of maps and boxplots”. The maps in Figure 2 are fairly similar between the 5- and 6-cluster solutions and the boxplots in Figure 3 also suggest that clusters 0, 3 and 4 remain the same with some countries in clusters 1, 2 of the 5-cluster solution redistributed between them and with the additional cluster 5 in the 6-cluster solution.
A5. Selection of 5 and 6 clusters was informed, mostly, by epidemiological knowledge about the countries, and how these were clustered. We did not choose 4 clusters, as the elbow plot would have suggested, because some countries were clustered with others they have little in common, epidemiologically speaking. This is what we meant by “visual inspection of maps and boxplots”. Mostly maps, though we also checked the boxplots. We have included a few lines the methodology section (“K-means” sub-heading) to explain our rationale: *…That is, five and six cluster classified countries in groups with shared socio-demographic and epidemiological profiles. Although five and six clusters provided similar groups, six clusters classified central Africa with greater detail, which could be useful for these countries and regional organizations. Overall, the function cost (elbow plot, Figure 1), paired with the overall results (boxplots and maps), suggested that five or six clusters were a sensitive decision.*

The maps with 5 or 6 clusters look similar. However, the map with 6 clusters classified countries in central Africa with greater detail. Although in the same sub-region, socio-economic and epidemiological differences provide unique features to these countries, that a 6-cluster model can identify. We have also included this argument in the new lines (please, refer to the text in italic in the previous paragraph).

Please, for further arguments about the choice of 5 and 6 clusters, referrer to answer 6.

Q6. There are more reliable metrics to aid with cluster selection, including the silhouette coefficient[2], and the GAP statistic[3]. The elbow plot is simply a heuristic. The authors should at least explain their choice of method.
A6. We did not follow any of these methods because of the limited number of observations available; that is, the number of countries (analysis units) studied. Given the reduced number of

observations, the elbow function would be fairly similar for the number of clusters close to the "elbow". At this stage, it is advisable to subjectively assess which clusters gives the best information or correlates better with expert knowledge,[1],[2] rather than relying only on performance metrics. As requested, we have further elaborated on the rationale for the choice of method: *When there is a limited number of observations, as it is arguably in this analysis, the number of clusters around the "elbow" function (Figure 1) provides similar information. At this point, it may be advisable to select the number of clusters which relates better to expert knowledge. Therefore, we used visual inspection of maps and plots to decide on the number of clusters that provide the best results, grouping countries in consistent clusters with a similar background.*

In addition, to further elaborate on our current choice of method, for clarity, transparency and consistency, we have conducted further analysis. First, the dendrogram with Euclidean distances showed the 5 clusters was the optimum number; this agrees with our current choice. The Silhouette analysis showed the metrics summarised in the table below. These show that the largest metrics (>40%) were retrieved for 3, 4, 5 and 6 clusters (please, see rows highlighted in green). After visual inspection of the maps with 3 and 4 clusters, we agreed that these did not classify or stratify countries well. In order words, there were countries in one cluster that may not have strong similarities (at least in epidemiological or socio-demographic terms). Consequently, 5 and 6 clusters appeared to be better options; again, the average silhouette score agreed with our original choice.

Number of clusters = Average silhouette score
2 = 0.388107
3 = 0.433095
4 = 0.477838
5 = 0.444210
6 = 0.415063
7 = 0.382376
8 = 0.354897
9 = 0.362776
10 = 0.365564

We have included the following paragraph in the "K-means" sub-heading: *Post-hoc analysis suggested we made a sensible choice when selecting 5 and 6 clusters. A dendrogram with Euclidean distances showed that 5 clusters were the optimum number. Similarly, the Silhouette analysis revealed the largest average Silhouette score for 3 (0.43), 4 (0.48), 5 (0.44), and 6 (0.42) clusters; all other options from 1 to 10 clusters were below 0.40. As explained above, the visual inspection of maps suggested that 3 or 4 clusters did not provide a good classification. That is, countries with no strong similarities were clustered. Overall, our choice of 5 and 6 clusters was supported by the analysed metrics (dendrogram and Silhouette).*

Q7. Although appropriate, the statistical analysis lacks further interpretation. The usefulness of the model could be illustrated by evaluating the predictive value of cluster labels to answer the question "Are the labels more predictive than individual variables?"
A7. We have further discussed (interpreted) about the relationship between the input variables, the cluster configuration, and how these relate to the outcomes. Please, refer to answers 9 and 10 for further details on the new text.

Although interesting, the proposed research question is beyond the aims of this work. The

research question and justification have been further elaborated (please refer to answers 1 and 2). Arguably, any cluster may predict better than individual variables. That is a strong argument in favour or risk prediction models, above and beyond risk/prognostic factors alone.

Q8. The resulting clusters are difficult to interpret without a summary table of cluster characteristics in terms of the 4 input variables used in the analysis.
A8. We have included a table showing the median and interquartile range of the eight input variables across clusters (Table 1).

There were eight input variables (Table 1); disease prevalence included 4 diseases. That is, four prevalence estimates hence the four variables in addition to air quality, GDP, universal health coverage index and proportion of male subjects in the country (eight input variables in total). We have included the following lines under the "Data sources" sub-heading to avoid confusions: …*that is, we used eight input variables for the cluster analysis: four diseases, air quality, gross domestic product per-capita, an universal health coverage index and the proportion of men in the country (Table 1).*

Q9. No specific conclusions are drawn in the discussion. What are the cluster characteristics and how are they associated with confirmed COVID-19 cases? Are the results expected, surprising? There is little discussion on the characteristics, whether present or absent in the model, that would drive the countries to cluster together with regards to the number of reported cases. A few example points for discussion are listed below.
A9. We have further discussed on the cluster characteristics (input variables) and how these may explain the clusters configuration in relation to COVID-19 outcomes. These lines in the discussion section read ("Results in context" sub-heading): *The input variables could potentially explain the clusters configuration. For example, cluster number four had the largest number of confirmed cases. This cluster also had the best universal health coverage index. It could be argued that such a strong health system is capable of performing tests to large populations, hence a large number of diagnosed cases. Conversely, cluster number two appeared to have the worst death rates; this cluster also had the largest tuberculosis prevalence as well as the smallest gross domestic product per capita and universal health coverage index. These epidemiological –large burden tuberculosis – and socio-demographic profiles could explain the high death rates.*

Q10. It appears from the map distribution that the clusters loosely correlate with GDP - although without a summary table confirming this is hard to tell for certain. I am not an epidemiologist and neither is NA, therefore it is not our area to comment, but countries with higher GDP are more likely to perform more tests, and are thus more likely to have a higher number of cases.
A10. We have further discussed how GDP, as an input variable in the clusters configuration, may relate to how the clusters reveal COVID-19 outcomes. Please, refer to the previous answer for details about the new text.

Q11. Additionally, some countries are more connected than others (e.g. because of air travel), and the spread of COVID-19 is not uniform across the world (e.g. countries that are closer to China reported cases earlier) and therefore, different countries are at different stages of the pandemic. It would make more sense to separately cluster countries with similar exposure to the virus as well as comparable reporting standards.
A11. It would difficult to separately cluster countries with similar exposure to the virus; it would be

more difficult to identify a threshold to define “similar exposure to the virus”. This approach will make the clustering more complex, which we tried to avoid by selecting variables readily available yet closely correlated to COVID-19 (please refer to answer 2). In this line, comparable reporting standards are not a static measure. Countries have improved their reporting standards at different paces and through different means during the pandemic. Finally, both the exposure to the virus and reporting standards are characteristics of the pandemic. However, our aim was to use pre-pandemic characteristics.

We have further discussed the relevance of flights or connections. Please, refer to the discussion section for the new text (“Results in context” sub-heading): *The cluster configuration herein presented did not seem to group countries closer to China, where the pandemic started. In other words, countries with the first imported cases did not cluster together. This could mean that the selected input variables do not correlate well with, for example, travel frequency or population movement from China to nearby countries. Alternatively, this unexpected finding could suggest that the selected input variables are more relevant than proximity or connections between countries.*

Q12. Figure 1 needs axis labels.
A12. We are providing a new figure with axis labels.

[1] Murugan Anandarajan, Chelsey Hill, Thomas Nolan. Practical Text Analytics: Maximizing the Value of Text Data. Chapter 7.5.1.

[2] Chia-Hui Chang, Zhi-Kai Ding, Categorical data visualization and clustering using subjective factors, Data & Knowledge Engineering, Volume 53,
Issue 3, 2005, Pages 243-262, ISSN 0169-023X.

Competing Interests: No competing interests were disclosed.

Reviewer Report 05 May 2020

https://doi.org/10.21956/wellcomeopenres.17350.r38301

Alan E. Hubbard
Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA

The main purpose of this article is to promote the use of clustering methods (k-means specifically) for aggregating regions into a smaller set of coherent clusters based on regional level data (non-Covid) that can be compared with Covid disease outcomes (case counts, deaths, order). The methodology is straightforward, reduces the dimension of the problem from a number of distinct regions to clusters of "like" regions and then if it also correlated with Covid outcomes, could be used to both simplify the

presentation of the results and possibly provide insights as to differences in the evolution of the epidemic in different regions. There are other unsupervised methods one could use as well as supervised methods (e.g., classification and regression trees). The results of this clustering exercise depend on the relevance of the data used to cluster on the dynamics of Covid, and thus as the understanding evolves, there are other sources that should be considered (e.g., mobility data).

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine learning, causal inference, epidemiology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.