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: awaiting peer review]

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

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

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
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.

Copyright: © 2020 Carrillo-Larco RM and Castillo-Cara M. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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: awaiting peer review]

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\(^1\), 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/AIDS). Diabetes seems to be very frequent among COVID-19 patients\(^1\). Although hypertension had a higher frequency than respiratory diseases\(^2\), 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\(^3\). 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.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Variables</th>
<th>Data source</th>
<th>Used for</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19 prevalence</td>
<td>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 ...)</td>
<td>COVID-19 global surveillance system by Johns Hopkins University(^4)</td>
<td>Cluster evaluation</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Age-standardized prevalence of diabetes, chronic obstructive disease [COPD], HIV/AIDS and tuberculosis (as of year 2017)</td>
<td>2017 Global Burden of Disease / Institute for Health Metrics, Washington University(^19)</td>
<td>Clustering</td>
</tr>
<tr>
<td>Air quality metric</td>
<td>Concentration of 2.5 particulate matter by country</td>
<td>Global Health Observatory data repository, World Health Organization(^11)</td>
<td>Clustering</td>
</tr>
<tr>
<td>Socio-economic metric</td>
<td>Gross domestic product per capita (as of year 2017)(^5)</td>
<td>World Bank(^20)</td>
<td>Clustering</td>
</tr>
<tr>
<td>Health system metric</td>
<td>Universal health coverage index of service coverage (as of year 2017)</td>
<td>Global Health Observatory data repository, World Health Organization(^11)</td>
<td>Clustering</td>
</tr>
</tbody>
</table>

\(^{1}\) 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. 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 implies 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.

K-means. This technique seeks to group heterogeneous elements into homogeneous 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. A few authors have used this methodology in clinical and public health research.

There are different methods for unsupervised clustering depending on the data characteristics. 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++.

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).

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.

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).

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Number of confirmed cases</th>
<th>Clusters</th>
<th>Number of confirmed cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.023</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>5</td>
<td>0.771</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Number of deaths</th>
<th>Clusters</th>
<th>Number of deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.110</td>
<td>1.000</td>
<td>0.086</td>
</tr>
<tr>
<td>5</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

| Clusters | Case fatality rate per 1,000 cases | Clusters | Case fatality rate per 1,000 cases |
|----------|----------------|----------|----------------|----------------|
|          | 0  | 1            | 2  | 3  | 4  | 0  | 1            | 2  | 3  | 4  |
| 1        | 1.000 | 0.140       |    |    |    | 1    | 0.460 |    |    |    |
| 2        | 0.430 | 1.000       |    |    |    | 2    | 1.000 | 1.000 |    |    |
| 3        | 1.000 | 1.000       | 1.000 | 1.000 |    | 3    | 1.000 | 1.000 | 1.000 | 1.000 |
| 4        | 1.000 | 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 | 5    | 1.000 | 1.000 | 1.000 | 1.000 |

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Order</th>
<th>Clusters</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0.123</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>1.000</td>
<td>0.198</td>
</tr>
<tr>
<td>4</td>
<td>&lt;0.001</td>
<td>0.040</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>5</td>
<td>0.004</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

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\(^{8,11}\). 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 vulnerable 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\(^2\).

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).

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

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


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


