Environment, Development and Sustainability
https://doi.org/10.1007/s10668-022-02646-3

Understanding the COVID-19 pandemic prevalence in Africa through optimal feature selection and clustering: evidence from a statistical perspective

Mohamed Lamine Sidibé[1] · Roland Yonaba[1] · Fowé Tazen[1] · Héla Karoui[1] · Ousmane Koanda[1] · Babacar Lèye[1] · Harinaivo Anderson Andrianisa[1] · Harouna Karambiri[1]

Received: 12 October 2021 / Accepted: 18 August 2022

Abstract
The COVID-19 pandemic, which outbroke in Wuhan (China) in December 2019, severely hit almost all sectors of activity in the world as a consequence of the restrictive measures imposed. Two years later, Africa still emerges as the least affected continent by the pandemic. This study analyzed COVID-19 prevalence across African countries through country-level variables prior to clustering. Using Spearman-rank correlation, multicollinearity analysis and univariate filtering, 9 country-level variables were identified from an initial set of 34 variables. These variables relate to socioeconomic status, population structure, healthcare system and environment and the climatic setting. A clustering of the 54 African countries is further carried out through the use of agglomerative hierarchical clustering (AHC) method, which generated 3 distinctive clusters. Cluster 1 (11 countries) is the most affected by COVID-19 (median of 63,508.6 confirmed cases and 946.5 deaths per million) and is composed of countries with the highest socioeconomic status. Cluster 2 (27 countries) is the least affected (median of 4473.7 confirmed cases and 81.2 deaths per million), and mainly features countries with the least socioeconomic features and international exposure. Cluster 3 (16 countries) is intermediate in terms of COVID-19 prevalence (median of 2569.3 confirmed cases and 35.7 deaths per million) and features countries the least urbanized and geographically close to the equator, with intermediate international exposure and socioeconomic features. These findings shed light on the main features of COVID-19 prevalence in Africa and might help refine effectively coping management strategies of the ongoing pandemic.

Keywords Africa · Cluster analysis · COVID-19 · Hierarchical clustering · Pandemic · Transmission factors

✉ Roland Yonaba
ousmane.yonaba@2ie-edu.org; roland.yonaba@gmail.com

[1] Laboratoire Eaux, Hydro-Systèmes Et Agriculture (LEHSA), Institut International d'Ingénierie de l'Eau Et de l'Environnement (2iE), 1 Rue de la Science, 01 BP 594, Ouagadougou 01, Burkina Faso

1 Introduction

Disease epidemics and even pandemics are nowadays becoming increasingly common occurrences (Madhav et al., 2017). In December 2019, the *2019-nCOV* acute respiratory disease (hereafter named '*COVID-19*' disease) emerged. This disease was later found to be caused by the *Sars-CoV-2* coronavirus, isolated for the first time in the province of Hubei (in China). The World Health Organization (WHO) declared the disease a pandemic two months later (Cucinotta & Vanelli, 2020). As of December 31, 2020, a year later, the global epidemiological situation indicated a cumulative total of 83,559,591 confirmed cases and 1,824,934 deaths, i.e., hence a global case fatality rate of 2.18% and a mortality rate of 234 deaths per million people (Dong et al., 2020).

A study published by the *London School of Hygiene & Tropical Medicine* (LSHTM) concluded that all African countries would have passed the 10,000-case mark for COVID-19 by early June 2020 (Pearson et al., 2020). Following this study, the vast majority of public health experts, including the WHO, had called on Africa to '*prepare for the worst*' (Nuwagira & Muzoora, 2020). Yet, the global observed case count reported for Africa remained largely 10 times lower than expected, suggesting that the African continent, as a whole, has remained largely unaffected: America, Europe and Asia reported, respectively, 43.4%, 28.5% and 24.8% of the global count of confirmed cases, while Africa only reported 3% (Dong et al., 2020). Another interesting point worth raising is that even at the scale of the African continent, the pandemic appears to be unevenly spread between its countries. South Africa, for example, has reported more than 38% of cases (Dong et al., 2020). Likewise, more than 82% of the confirmed cases come from 9 countries alone (Salyer et al., 2021).

The relatively low number of cases and deaths due to COVID-19 is thought to be largely attributed to the fact that forecast regarding the evolution of the pandemic in Africa has been made without regard to some specificities such as socio-demographic aspects (Zongo et al., 2020). African countries seem to be more resilient to COVID-19 because of the swift adoption of mitigation measures, the low rate of urbanization, the limited transport network and the youth of the population: in fact, the median age of the population lies between 31 to 42 years old for Europe, America, Oceania and Asia, as compared to 18 years old for Africa (Adams et al., 2021; Desjardins, 2019; Lulbadda et al., 2021). This might explain the low number of COVID-19 cases and deaths in Africa, since the case fatality rate of non-communicable diseases (such as cancer, cardiovascular accidents and diabetes), already known as comorbidities in the context of this pandemic, is unlikely with younger people (Lawal, 2021; Randazzo et al., 2020).

The role of climatic and environmental factors has also been highlighted in COVID-19-related studies. Temperature and humidity are the factors most often associated to COVID-19 (Kerr et al., 2021; Şahin, 2020). Wang et al. (2021), for instance, showed that a 1 °C rise in average temperature can be associated with a 3.1% decrease in the new cases of COVID-19 infections and a 1.2% decrease in related deaths. According to Baker et al. (2020), in the absence of effective control measures, stronger outbreaks are likely in wetter climates. Luo et al. (2020) found significant influence of absolute air temperature over transmission rates of COVID-19 in China. A significant correlation between geographical latitude and COVID-19-related deaths and confirmed cases has been reported in earlier studies (Braiman, 2020; Chen et al., 2021; Heneghan et al., 2020; Whittemore, 2020). Moreover, the concentration of fine particles in the air has been associated with a higher prevalence of COVID-19 (Rizvi et al., 2021; Zhu et al., 2020).

Based on the current findings, it appears that the spread of the COVID-19 is affected by various factors at different levels, but also that countries exhibit different vulnerabilities to the pandemic. To assess these various sensitivities, clustering has been carried in earlier studies to identify emerging risk profiles. Gilbert et al. (2020) used bottom-up hierarchical clustering to model transmission between Africa and China and identified three different clusters depending on the severity of the risk of exposure to COVID-19 (high, medium and low). Centroid-based method (*K-means*) clustering was used by Carrillo-Larco and Castillo-Cara (2020) at the global level using country-level variables, which helped identify 5 to 6 clusters of countries. Imtyaz et al. (2020) assessed the effectiveness of measures taken by countries to limit the spread of COVID-19 based on 5 clusters identified and concluded the positive association between the mortality rate and the proportion of people over 65 years of age. Sadeghi et al. (2021) used hierarchical clustering to rank and score 180 countries according to COVID-19 cases and fatality in 2020 and compare existing pandemic vulnerability prediction models and standard epidemiological scoring techniques. In Africa, the African Center for Strategic Studies (ACSS) identified 3 clusters of African countries by assessing their level of vulnerability through the rating of the 9 following socioeconomic factors: international exposure, healthcare system, urban density, urban population, population age, governmental transparency, press freedom, conflict and displacement (ACSS, 2020a). These clusters later served in establishing different risk profiles of exposure to COVID-19 (ACSS, 2020b). However, the lack of inclusion of climatic factors might constitute serious limitations for these results.

Despite the large number of publications related to COVID-19, very few focused on the African continent. This study aims at addressing this critical gap in the body of the available literature, through the evaluation of the relative importance of factors that might be associated with the spread of the COVID-19 pandemic within the African context, using country-level indicators. To the best of our knowledge, this is the first study addressing directly the African continent scale regarding COVID-19, considering confirmed cases and deaths reports. Moreover, it uses data acquired over two years (January 1, 2020 to March31, 2022), which is likely to support effectively in identifying long-term relevant conclusions regarding the spread of COVID-19 in Africa. In addition, this study is motivated by the lack of explicit assessment of the effect of variables related to the physical climate setting (temperature, rainfall, insolation, humidity, wind speed) and environment (air quality, environmental performance index) on COVID-19 in earlier studies, especially in the case of Africa. The objectives of this study are twofold: (i) to assess the potential factors explaining at best the COVID-19 prevalence in African countries (cumulative number of confirmed cases and deaths per million people); (ii) to identify clusters of African countries sharing similar prevalence profiles to the COVID-19 disease.

2 Material and methods

Figure 1 is the flowchart presenting the main steps of the methodology used in this study, which are: (i) the data preparation phase (including data collection, imputation of missing values and removal of potential redundant variables); (ii) the dataset optimization phase (consisting in filtering optimal variables explaining the maximum variance in the data); (iii) the clustering phase and the comparative analysis of the clusters. These phases are further described in detail in the following sections. The complete list of the 54 African countries considered in this study is presented in the Online Resource 2 (ESM2—Table S1).

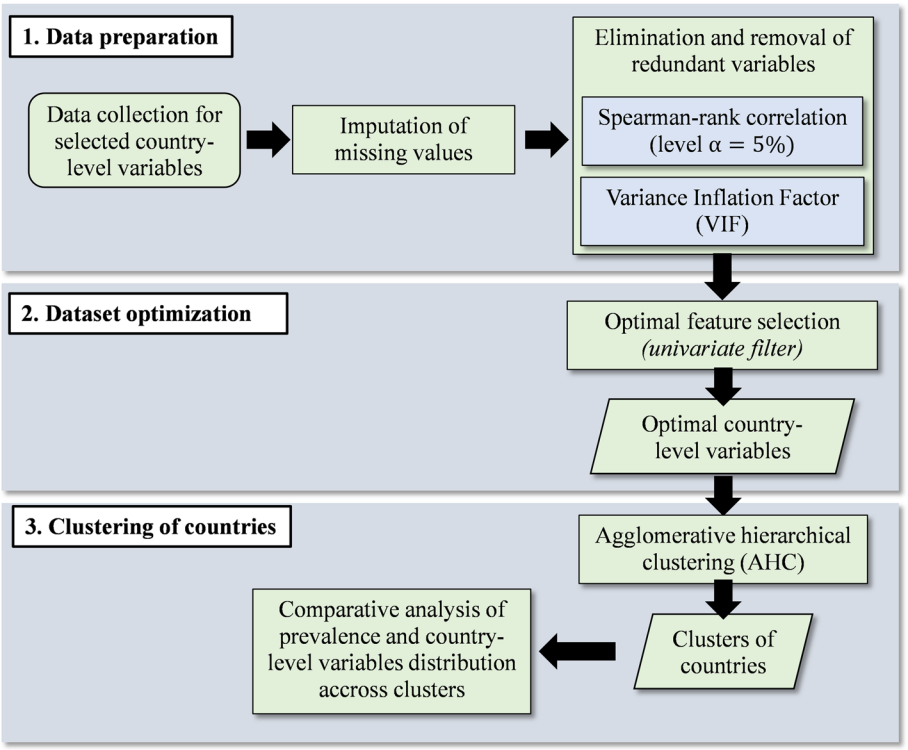

Fig. 1 Flowchart of the methodology used in this study

2.1 Data preparation

2.1.1 Selected country-level variables description

Based on a literature review, a set of factors previously associated with the spread of the COVID-19 pandemic has been identified and included in this study. These variables, presented in Table 1, are grouped into five categories: (i) international exposure and socioeconomic status; (ii) population structure; (iii) healthcare system and environment; (iv) disease prevalence and risk factors; and (v) climatic setting. The detailed dataset for all these variables is given in the Online Resource 2 (ESM2—Table S2).

The data was collected at the country level for the latest year available (2020 in most of the cases). Climatic setting variables were collected from MERRA-2 global reanalysis, which is a gridded and global model operating at the hourly/daily timestep at a spatial resolution of $0.625° \times 0.5°$ providing data since 1980 (Gelaro et al., 2017). For this study, the climate data was collected using NASA POWER Data Access Viewer (https://power.larc.nasa.gov/data-access-viewer/), using the R package *nasapower* (Sparks, 2021). The climate data was first collected as daily time series for the period January 2020 to March 2022 and later on averaged over the period for each country. Absolute humidity (AH) was calculated using an approximation of the Clausius–Clapeyron equation, presented in Eq. **(1)** (Iribarne & Godson, 1973):

Table 1 Country-level variables selected for this study

Category	Variables	Description	Sources
COVID-19 prevalence	conf_pm	Cumulative confirmed cases (as of 08/31/21)	Dong et al. (2020)
	death_pm	Cumulative confirmed deaths (as of 08/31/21)	Dong et al. (2020)
International exposure and socioeconomic status	arriv	International tourism, number of arrivals (thousands)	WorldBank (2021)
	hdi	Human development index (HDI)	UNDP (2020)
	gini	Gini index (metric for inequalities)	WorldBank (2021)
	gdp_cap	Gross domestic product per capita (GDP) ($US)	WorldBank (2021)
	alphab	Literacy rate (%)	WorldBank (2021)
Population structure	dens_pop	Population density (people/km^2)	WorldBank (2021)
	urb_pop	Urban population percentage (%)	WorldBank (2021)
	median_age	Median age of the population (years old)	WorldBank (2021)
	life_exp	Life expectancy (years old)	WorldBank (2021)
	p65yrs	Percentage of people aged over 65 years (%)	WorldBank (2021)
Healthcare system and environment	lack_hygien	Mortality rate due to lack of hygiene, unsafe water and sanitation (per 100,000 people)	WorldBank (2021)
	hous_fossf	Mortality rate due to air pollution from the use of household solid fuels (per 100,000 people)	Yale (2020)
	med_1000	Number of physicians (per 1000 people)	WorldBank (2021)
	pm25	Annual mean concentration of particulate matter of less than 2.5 microns of diameter (PM2.5) [µg/m3] in urban areas	WorldBank (2021)
	health_exp	Current health expenditure per capita ($US)	WorldBank (2021)
	epi	Environmental performance index	Yale (2020)
	immuniz_dtp1	Immunization coverage / DTP1 (%)	WHO (2020)
	immuniz_bcg	Immunization coverage / BCG (%)	WHO (2020)

Table 1 (continued)

Category	Variables	Description	Sources
Diseases prevalence and risk factors	prev_diab	Diabetes prevalence (number of people)	IHME (2020)
	prev_cvlds	Cardiovascular diseases prevalence (number of people concerned)	IHME (2020)
	prev_ch.resp	Chronic respiratory diseases prevalence (number of people concerned)	IHME (2020)
	prev_malaria	Malaria prevalence (number of people concerned)	IHME (2020)
	prev_nutdef	Malnutrition and nutritional deficiencies prevalence (number of people concerned)	IHME (2020)
	prev_respdtub	Respiratory infections and tuberculosis prevalence (number of people concerned)	IHME (2020)
	alcohol_cons	Total alcohol consumption per capita (liters)	WorldBank (2021)
Climatic setting	lat_abs	Absolute latitude (°)	Gelaro et al. (2017)
	ws2m_avg	Average daily wind speed (m/s)	Gelaro et al. (2017)
	rh2m_avg	Average daily relative humidity (%)	Gelaro et al. (2017)
	tmax_avg	Average daily maximum temperature (°C)	Gelaro et al. (2017)
	tmin_avg	Average daily minimum temperature (°C)	Gelaro et al. (2017)
	tmoy_avg	Average daily temperature (°C)	Gelaro et al. (2017)
	insol_avg	Average daily insolation ($MJ/m^2/j$)	Gelaro et al. (2017)
	tdew_avg	Average dew point temperature (°C)	Gelaro et al. (2017)
	ah_avg	Absolute air humidity (%) – *calculated* (Iribarne & Godson, 1973)	Gelaro et al. (2017)

$$AH = \frac{6.112 \times \frac{17.67 \times T}{e^{T+243.5}} \times RH \times 2.1674}{T + 273.15} \quad (1)$$

where AH is the absolute humidity, T is the average temperature (°C), RH is the air relative humidity (%), and e is the base of the natural logarithm.

These variables were selected because they have been previously associated with the COVID-19 pandemic. International exposure reflects the number of people entering a country through airports, which denotes an increased probability of welcoming confirmed cases of COVID-19 in the country (Moosa & Khatatbeh, 2020). Socioeconometric variables describe the level of development of countries and are closely related to healthcare system equipment and the effectiveness of policy management during health crises (Freed et al., 2020). Population structure has been associated with COVID-19 prevalence, especially to deaths (Medford & Trias-Llimós, 2020). Healthcare systems, environment and disease prevalence have also been identified as determinants of COVID-19 prevalence (Aydın et al., 2021; Carrillo-Larco & Castillo-Cara, 2020). Finally, the relationship between climate variables and COVID-19 has been a trending and active topic of research since the outbreak of the pandemic (Chen et al., 2021; Islam et al., 2021; Rahman et al., 2021; Singh et al., 2021; Zaitchik et al., 2020).

COVID-19 prevalence data used in this study (cumulative number of confirmed cases and deaths) includes cumulative cases and deaths since the outbreak of the pandemic until March 31, 2022 for all 54 African countries. The data was normalized by current countries population estimates (WorldBank, 2021) to enable the comparison of the pandemic prevalence across countries, as suggested by Goldstein and Lee (2020). The counts were later on translated into confirmed cases per million people (*conf_pm*) and deaths per million people (*death_pm*).

2.1.2 Dataset imputation

The data collected for all country-level variables listed in **Table 1** initially presented gaps for specific countries such as Eritrea (ERY), Equatorial Guinea (GNQ), Libya (LBY), Somalia (SOM), South Sudan (SSD) and Lesotho (LSO). Missing values were identified especially for international arrivals (*arriv*: 4 missing values out of 54), Gini index (*gini*: 2 missing values out of 54), number of physicians (*med_1000*: 1 missing value out of 54), current health expenditure per capita (*health_exp*: 1 missing value out of 54), environmental performance index (*epi*: 3 missing values out of 54), alcohol consumption (*alcohol_cons*: 1 missing value out of 54), exposure to air pollution from household fossil fuels (*hous_fossf*: 3 missing values out of 54). To form a fully complete initial dataset, the missing values were imputed using the R package *missforest* (Stekhoven, 2013), which implements a random iterative and nonparametric gap-filling approach based on random forests (Stekhoven & Buhlmann, 2012). The *missForest* algorithm has been used in previous COVID-19 research (Gangloff et al., 2021) and has proven to be more effective at gap-filling than other nonparametric approaches (Ramosaj & Pauly, 2019). In this study, the random forest model was trained on the matrix formed by the initially selected variables using 100 trees and 3 iterations, yielding a normalized root mean square error (NRMSE) = 0.1480. The detailed dataset for all variables is given in the Online Resource 2 (ESM2—Table S3).

2.1.3 Identification and removal of redundant factors

The initial dataset contained 34 variables which were assessed for potential redundancy and multicollinearity. In this perspective, the correlation matrix (Spearman's ρ nonparametric coefficient) was evaluated and a threshold of 0.90 was considered to eliminate country-level variables highly correlated ($\rho > 0.90$) with an already existing variable within the dataset, to lessen redundancy.

To further avoid potential issues related to multicollinearity and form a dataset of independent variables, the variance inflation factor (VIF) was evaluated for the remaining country-level variables. The VIF is a measure of multicollinearity in a set of multiple regression variables, and is defined as the ratio of the overall model variance to the variance of a model including a single independent variable (Akinwande et al., 2015). The VIF formula is defined as in Eq. **(2)**:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where VIF_i is the VIF for the i_{th} independent variable, R_i^2 is the unadjusted coefficient of determination for regressing the i_{th} independent variable on the remaining ones. In this study, only variables presenting a VIF value below 10 were retained, this threshold being commonly advised as a cutoff for high multicollinearity (Kutner et al., 2004)..

2.2 Optimal feature selection

The optimal feature selection is a dimensionality reduction method which helps in retaining a subset of relevant variables maximizing the variance of the original dataset, while minimizing the loss of information resulting from the removal of some of the original variables (Friedman, 1997). However, the procedure for optimal feature selection is likely to be affected by the presence of atypical observations, i.e., outliers. It is therefore critical to identify and remove these outliers from the dataset before looking for optimal variables (Acuña & Rodríguez, 2005).

In this study, outliers were identified using the multivariate Cook's D distance statistic (Cook & Weisberg, 1982). Cook's D-statistic is calculated by removing the i_{th} data observations from the model and recalculating a regression, hence summarizing how much all the values in the regression model change when the i_{th} observation is removed. The calculation of Cook's distance is defined by Eq. **(3)**:

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \times MSE} \tag{3}$$

where D_i is the Cook distance for the i_{th} observation (for $i = 1, \ldots, n$), $\widehat{Y}_j$ is the regression model response fitted on all observations, $\widehat{Y}_{j(i)}$ is the regression model response fitted on all but the i_{th} observation, p is the number of coefficients in the regression model, and MSE is the mean square error. The calculation of D-statistic was conducted through a linear regression model, using a cutoff of 4 times the standard deviation to flag outlier observations (Cook & Weisberg, 1982).

The identified outliers were temporarily set aside in order to avoid bias in the selection conducted later on during feature selection. The optimal feature selection was performed through a univariate filter using the R *caret* package function *sbf* for selection by filter (Kuhn, 2021).

Seventy-five percent of the data was used for training and 25% of the data was used for validation, using repeated tenfold cross-validation. This procedure for optimal feature selection was conducted separately on confirmed cases (*conf_pm*) and on deaths (*deaths_pm*) as response variables. The resulting dataset was finally normalized to bring all the variables to the same scale between 0 and 1, prior to clustering (Visalakshi & Suguna, 2009), and using *min–max* normalization.

2.3 Clustering of countries

Clustering aims at partitioning the whole of African countries into homogeneous groups called *clusters*. These clusters are obtained by maximizing the inertia between clusters and therefore minimizing the inertia within all clusters to obtain well-differentiated groups of observations. The different clustering methods includes hierarchical clustering, partitioning methods and machine learning-based methods. In this study, the bottom-up or agglomerative hierarchical clustering (AHC) was used as it does not require as an input a number of clusters, unlike partitioning methods such as *K-means*. Machine learning-based methods are also available and effective; however, they poorly compare to AHC in terms of ease of interpretation of their results. The AHC procedure used in this study provides the analyst a dendrogram, whose goodness can be assessed through the correlation between the cophenetic distances between observations (vertical *y-axis* on the dendrogram) and the original distances. The closer the value of this correlation coefficient to 1, the more reliable the classification presented through the dendrogram in terms of reflection of the data. Cophenetic distances above 0.5 are deemed to be acceptable (Kassambara, 2017).

In this study, the cophenetic correlation coefficient for various clustering schemes produced by combinations of various distance metrics (*Manhattan, Canberra, Minkowski and Euclidean*) and aggregation methods (*Average, Complete, Ward.D and Ward.D2*), hence a total of 16 combinations, was examined. These combinations were ranked out by decreasing values of cophenetic correlation coefficients. For each of these combinations, the optimal number of clusters to be produced was evaluated with the R package *NbClust* (Charrad et al., 2014), which uses an array of indices to select the appropriate number of clusters (Charrad et al., 2014; Milligan & Cooper, 1985). Using this number of clusters, the AHC is applied and the statistical differences in COVD-19 prevalence between clusters are assessed. The final combination of distance metric and aggregation method selected is the one producing significantly different clusters (in terms of COVID-19 prevalence), with the highest cophenetic correlation coefficient.

The significance of differences between COVID-19 clusters prevalence (confirmed cases and deaths per million) was further evaluated with the nonparametric Kruskal–Wallis test for multiple groups comparison (at level α=5%), associated with the post hoc nonparametric Mann–Whitney U test for pairwise comparison of group medians (at level $\alpha = 5\%$). Also, the significant differences between the distribution of variables used to form clusters were similarly assessed.

3 Results

3.1 COVID-19 situation in Africa

In this section, several aspects of the epidemiological situation of the pandemic are presented: the chronological onset of COVID-19 in Africa, the evolution of the cumulative number of cases and deaths and the spatial and temporal evolution of COVID-19 within the African continent.

3.1.1 Chronological onset of COVID-19

Africa reported its first COVID-19 case in Egypt (EGY), on February 14, 2020. Neighboring countries such as Algeria (DZA), Tunisia (TUN) and Morocco (MAR) reported their first cases a few days later. It appears that the first countries to be affected are countries with higher international exposure (Online Resource 1, ESM1–Fig. S1). These countries are also those farthest from the equator, such as Tunisia (TUN), Egypt (EGY) and South Africa (ZAF). Less than 2 months after the 1st case was reported in Egypt, 52/54 countries (96.3%) had reported at least one confirmed case. For most countries, the date of the first reported death case follows quite closely the date of the first reported confirmed case, by an order of 2 to 125 days. Only Seychelles (SYC) has not reported a single death in 2020 despite a first confirmed case being reported on March 14, 2020.

3.1.2 Cumulative number of cases and deaths

Figure 2 shows the cumulated numbers of cases and deaths, along with the daily estimates in Africa during the early beginning of the pandemic up to March 31, 2022.

From February 14, 2020 to March 31, 2022, a total of 11,558,931 COVID-19-related confirmed cases and 251,953 deaths is reported in Africa (roughly 2.37% of the total number of worldwide cases and 4.10% of worldwide deaths counts). This yields an average of 8848 confirmed cases per million and 193 deaths per million in Africa for this period, compared to 67,977 cases per million and 809.1 deaths per million globally (Dong et al., 2020). These figures translate to a case fatality rate of 2.17% in Africa, twice higher than the global case fatality rate which is 1.19% (Dong et al., 2020). This shows that even though Africa is much less affected than the rest of the world, the COVID-19 lethality in Africa is more severe (Lawal, 2021).

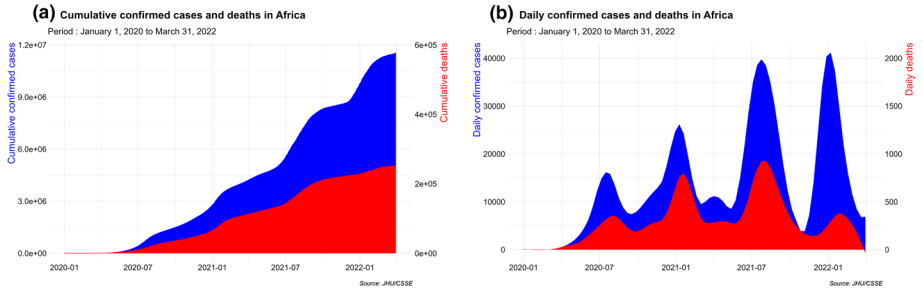

Fig. 2 COVID-19 prevalence evolution in Africa. **a** Cumulative confirmed cases and deaths. **b** Daily confirmed cases and deaths

As of March 31, 2022, the 10 COVID-19 hard-hit countries are: Seychelles (SYC: 414,044 cases per million, 1680 deaths per million), Cameroon (CMR: 217,378 cases per million, 3504 deaths per million), Mauritius (MUS: 166,186 cases per million, 765 deaths per million), Botswana (BWA: 132,624 cases per million, 1166 deaths per million), Tunisia (TUN: 88,577 cases per million, 2422 deaths per million), Libya (LBY: 74,026 cases per million, 947 deaths per million), South Africa (ZAF: 63,509 cases per million, 1708 deaths per million), Namibia (NAM: 63,197 cases per million, 1611 deaths per million), Swaziland (SWZ: 60,769 cases per million, 1214 deaths per million) and Morocco (MAR: 31,895 cases per millions, 440 deaths per million). In terms of raw confirmed cases and deaths count, the top 10 countries include South Africa (ZAF), Morocco (MAR), Tunisia (TUN), Egypt (EGY), Libya (LBY), Ethiopia (ETH), Kenya (KEN), Zambia (ZMB), Botswana (BWA), Algeria (DZA) and 74.5% of the confirmed cases counts and 80.1% of deaths come from these topping countries; yet, their cumulative population account for 33.8%of the continent population (Dong et al., 2020). Also, it is worth noting that most of these countries are those mostly located at the northernmost (or southernmost) parts of the continent, and features high standard of living and international exposure.

Over the study period, the African continent experienced four waves of increasing magnitude, both for new daily cases and deaths, as shown in daily estimates presented in Fig. 2b. The first wave peaked around July–September 2020, the second in January–February 2021, the third one in early August 2021 and the fourth one (of similar magnitude with the third one) occurred February 2022. Also, a strong periodicity of around 6 months is observed.

3.1.3 Spatial and temporal evolution of COVID-19

The spatial and temporal spread of the COVID-19 in Africa is presented on choropleth maps in Fig. 3 at different dates (September 30, 2020; March 31, 2020; September 30, 2021; March 31, 2022).

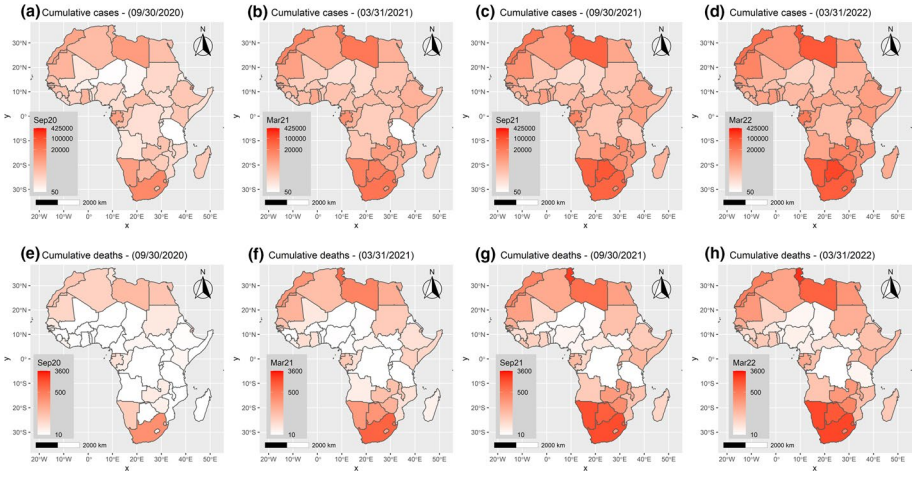

Fig. 3 Choropleth map showing the spatial and temporal spread of COVID-19 cumulative cases and deaths in Africa over the period January 2020 to March 2022. **a–d** Cumulative cases per million people. **e–h** Cumulative deaths per million people

From the beginning of the pandemic in Africa to March 31, 2022, it appears that the countries located at the extremities of the continent (north and south) are the most affected in terms of cumulative confirmed cases and deaths per million, while countries closer to the equator seem less affected, in terms of magnitude. The Southern Africa region especially is the most affected region on the continent in terms of prevalence. This region, which is home to only 13.52% of the people living in Africa (WorldBank, 2021), accounts for over 47.1% of cumulative confirmed cases and 49.8% of deaths. This is in sharp contrast to the findings of Heneghan et al. (2020) who stated that the pandemic had a higher prevalence in the Northern hemisphere of the continent. In contrast, the West Africa region, which is home to 29.67% of the African population, reported only 7.8% of cumulative confirmed cases and 5.2% of deaths. The North Africa region, with 18.69% of the continent's population, reported 31.1% of confirmed cases and 34.9% of deaths, while the East Africa region reported 11.0% of confirmed cases and 8.3% of deaths. The Central Africa region is the least affected by the pandemic, with only 3.0% of confirmed cases and 1.7% of deaths (Online Resource 1, ESM1–Fig. S2-S5).

3.2 Selection of optimal factors for COVID-19 prevalence analysis

3.2.1 Spearman's correlation rank analysis

Figure 4 shows Spearman's ρ correlation matrix for the complete dataset of 34 country-level variables initially selected and their association to the two response variables, which are the cumulative confirmed cases per million (*conf_pm*) and deaths per million (*death_pm*). The complete correlation matrix values and associated significance (p values) is given in the Online Resource 2 (ESM2—Tables S4 and S5).

The variables highly and positively correlated to COVID-19 prevalence data (respectively, cumulative confirmed cases and deaths) are Human Development Index (*hdi*: $\rho = 0.77, 0.73$), health expenditure (*hdi*: $\rho = 0.76, 0.77$), median age (*median_age:* $\rho = 0.74, 0.71$), literacy rate (*alphab:* $\rho = 0.72, 0.64$), number of physicians for 1000 people (*med_1000:*$\rho = 0.71, 0.70$), Gross Development Product per capita (*gdp_cap:*$\rho = 00.71, 0.64$) and mortality rate due to air pollution from the use of household solid fuels (*hous_fossf:* $\rho = 0.70, 0.72$). These results are in line with those of Gilbert et al. (2020) and ACSS (2020a) for variables reflecting the standard of living, and of Lulbadda et al. (2021) for the age of the population. Some variables are also found to be highly and negatively correlated to COVID-19 prevalence, including the mortality related to the lack of hygiene (*lack_hygien*: $\rho = -0.78, -0.79$), prevalence of malaria (*prev_malaria:*$\rho = -0.71, -0.79$), the lack of hygiene related mortality (*lack_hygien:*$\rho = -0.77, -0.70$) and prevalence of malnutrition and nutritional deficiencies (*prev_nutdef*: $\rho = -0.62, -0.53$). Such findings are in line with the recent work of Weiss et al. (2021).

Moderate association is found between COVID-19 prevalence (confirmed cases and deaths per million, respectively) and prevalence of nutritional deficiencies (*prev_nutdef:*$\rho = -0.62, -0.53$), immunization coverage with BCG (*immuniz_bcg:* $\rho = 0.57, 0.50$), life expectancy (*life_exp:* $\rho = 0.55, 0.56$), percentage of people aged over 65 years (*p65yrs:* $\rho = 0.54, 0.58$), environmental performance index (*epi:* $\rho = 0.54, 0.56$), prevalence of respiratory diseases and tuberculosis (*prev_respdtub:* $\rho = -0.51, -0.45$), prevalence of chronic respiratory diseases (*prev_ch.resp:* $\rho = -0.46, -0.37$), PM2.5 air pollution (*pm25:* $\rho = -0.45, -0.40$), daily annual maximum temperature (*tmax_avg:* $\rho = -0.44, -0.36$) and urban population (*urb_pop:* $\rho = 0.43, 0.42$).

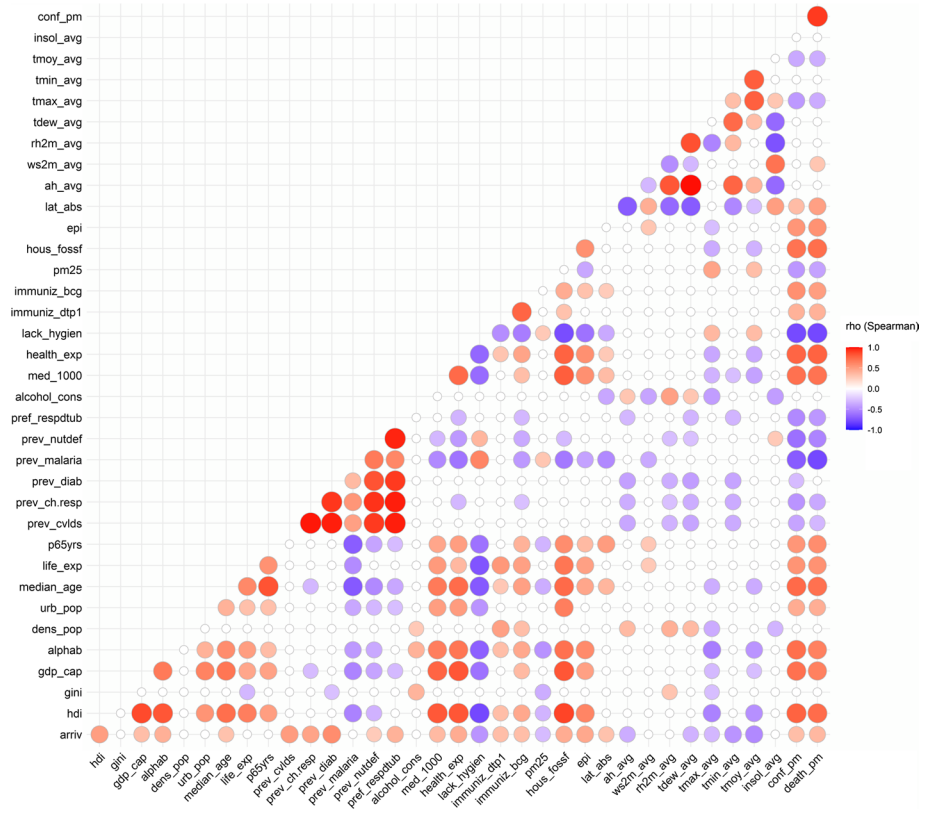

Fig. 4 Spearman's rho correlation coefficient between country-level variables and their association to COVID-19 prevalence in African countries. Blank values show nonsignificant correlation coefficients (at $\alpha = 5\%$ level)

Variables such as prevalence of nutritional deficiencies (*prev_nutdef*), prevalence of respiratory infections and tuberculosis (*prev_respdtub*), prevalence of cardiovascular diseases (*prev_cvlds*) and prevalence of diabetes (*prev_diab*) and prevalence of chronic and respiratory diseases (*prev_ch.resp*) were found to be highly correlated with each other ($\rho > 0.9$). As such the first four variables were removed from the dataset, only leaving out the *prev_ch.resp* variable, found to be lesser correlated with the remaining country-level variables in the entire dataset. To further avoid potential issues of collinearity, redundant variables in the dataset were screened through the calculation of the VIF index, presented in Table 2.

A total of 16 variables have VIF values below 10 and were considered significant for further analysis. These variables refer to international exposure (*arriv*), socioeconomic status (*gini, alphab*), population structure (*dens_pop, urb_pop*), healthcare systems and environment (*pm25, med_1000, epi, lack_hygien, immuniz_dtp1, immuniz_bcg*), disease prevalence and risk factors (*alcohol_cons, prev_malaria, prev_ch.resp*), climate setting (*ws2m_avg, lat_abs*). The remaining variables show VIF values over 10, indicating high collinearity. Therefore, these latter variables were excluded.

Table 2 VIF values for all variables

N°	Variable	conf_pm	death_pm	N°	Variable	conf_pm	death_pm
1	alcohol_cons	3.11	3.11	16	prev_ch.resp	9.52	9.52
2	dens_pop	3.36	3.36	17	insol_avg	10.57	10.57
3	urb_pop	3.59	3.59	18	life_exp	11.48	11.48
4	pm25	3.66	3.66	19	p65yrs	14.63	14.63
5	gini	4.99	4.99	20	hous_fossf	17.20	17.20
6	arriv	6.10	6.10	21	hdi	18.74	18.74
7	med_1000	7.48	7.48	22	gdp_cap	32.48	32.48
8	prev_malaria	7.51	7.51	23	health_exp	32.54	32.54
9	ws2m_avg	7.75	7.75	24	median_age	35.61	35.61
10	epi	7.83	7.83	25	rh2m_avg	140.79	140.79
11	alphab	7.96	7.96	26	ah_avg	180.74	180.74
12	immuniz_bcg	8.46	8.46	27	tmax_avg	258.91	258.91
13	lack_hygien	8.82	8.82	28	tmin_avg	575.63	575.63
14	immuniz_dtp1	8.92	8.92	29	tmoy_avg	1349.89	1349.89
15	lat_abs	9.08	9.08				

3.2.2 Optimal factors subset selection

The Cook's distance *D*-statistic helped in flagging some countries as outliers for cumulative confirmed cases and deaths per million, especially 3 countries: Cabo Verde (CPV), Mauritius (MUS) and Seychelles (SYC). These countries present highest GDP per capita values (*gdp_cap*: 11,099.2 $US/capita and 17,448.3 $US/capita for MUS and SYC, respectively), highest population densities (*urban_pop*: MUS: 620.4 inhabitants/km^2 for MUS), highest health expenditure (*health_exp*: 653.3 $US/capita and 833.1 $US/capita for MUS and SYC, respectively). Also, these countries share the highest prevalence estimates (166,185–414,043 confirmed cases per million and 764–3,504 deaths per million). Since such atypical values are likely to affect the feature selection procedure (Online Resource 1, ESM1–Fig. S6), these countries were temporarily removed from the dataset, and later re-included in the set of countries.

Table 3 shows the optimal variables selected through the feature selection, ranked by order of decreasing importance, evaluated at different time points during the study period.

It appears that 8 to 9 optimal features stand out as the most important ones, both for confirmed cases and deaths. These optimal variables include mortality attributed to the lack of hygiene (*lack_hygien*), literacy rate (*alphab*), number of physicians per 1000 inhabitants (*med_1000*), coming as the most important ones. These variables are further followed by EPI (*epi*), air pollution with PM2.5 (*pm25*) and urban population (*urb_pop*). The lesser important one, with varying ranks of importance depending on the analysis period, are latitude (*lat_abs*), international tourism (*arriv*) and Gini index (*gini*).

From the above results, it appears that variables relating to the healthcare system and environment-related variables (*lack_hygien, med_1000, epi, pm25*), international exposure (*arriv*) and socioeconomic status (*alphab, gini*) are closely related to COVID-19 prevalence. The latter are followed by variables related to population structure (*urb_pop*) and to a lesser extent, climatic setting (*lat_abs*). These 9 variables were finally used for the clustering of countries.

Table 3 Optimal features explaining variability in COVID-19 prevalence across African countries

N°	Variable	Conf_pm (ρ)	Variable	Death_pm (ρ)	N°	Variable	conf_pm (ρ)	Variable	Death_pm (ρ)
Period: January 1, 2020 to September 30, 2020					Period: March 31, 2021 to September 30, 2021				
1	lack_hygien	−0.69 ***	lack_hygien	−0.62 ***	1	lack_hygien	−0.74 ***	lack_hygien	−0.73 ***
2	med_1000	0.62 ***	med_1000	0.56 ***	2	alphab	0.65 ***	med_1000	0.63 ***
3	urb_pop	0.58 ***	urb_pop	0.54 ***	3	med_1000	0.64 ***	alphab	0.55 ***
4	alphab	0.51 ***	alphab	0.41 **	4	epi	0.49 ***	epi	0.49 ***
5	epi	0.38 **	lat_abs	0.37 **	5	pm25	−0.44 **	lat_abs	0.49 ***
6	lat_abs	0.3 *	epi	0.33 *	6	urb_pop	0.43 **	pm25	−0.41 **
7	arriv	0.18	arriv	0.16	7	arriv	0.35 *	arriv	0.40 **
8	gini	0.15	gini	0.02	8	lat_abs	0.35 *	urb_pop	0.38 **
					9	gini	0.32 *	gini	0.19
Period: September 30, 2020 to March 31, 2021					Period: September 30, 2021 to March 31, 2022				
1	lack_hygien	−0.73 ***	lack_hygien	−0.68 ***	1	lack_hygien	−0.74 ***	lack_hygien	−0.76 ***
2	med_1000	0.68 ***	med_1000	0.63 ***	2	alphab	0.67 ***	med_1000	0.66 ***
3	alphab	0.62 ***	lat_abs	0.53 ***	3	med_1000	0.66 ***	alphab	0.58 ***
4	epi	0.51 ***	alphab	0.49 ***	4	epi	0.49 ***	epi	0.53 ***
5	urb_pop	0.46 ***	epi	0.47 ***	5	pm25	−0.45 **	lat_abs	0.51 ***
6	lat_abs	0.4 **	urb_pop	0.42 **	6	urb_pop	0.42 **	pm25	−0.41 **
7	pm25	−0.39 **	pm25	−0.35 *	7	arriv	0.37 **	urb_pop	0.39 **
8	arriv	0.33 *	arriv	0.34 *	8	lat_abs	0.37 **	arriv	0.39 **
9	gini	0.25	gini	0.14	9	gini	0.31 *	gini	0.16

ρ indicates Spearman's rank correlation coefficient. '***' indicates significance at the 0.001 level. '**' indicates significance at the 0.01 level. '*' indicates significance at the 0.05 level. Variables are ranked out by order of decreasing importance

3.3 Clustering of African countries

3.3.1 Creation of clusters

The examination of various clustering schemes (as presented in Sect. 0) resulted in an optimal set of 3 clusters, produced through the combination of *Canberra* distance and *Ward. D2* aggregation method. The cophenetic correlation associated is 0.585, therefore considered acceptable (Kassambara, 2017). Figure 5 shows the resulting dendrogram from the AHC clustering.

The associated map in Fig. 6 shows the spatial configuration of the clusters obtained. The detailed dataset presenting the clusters and their associated features and prevalence is presented in the Online Resource 2 (ESM2—Table S6).

Cluster 1 is composed of 11 countries, mostly located in the Northern Southern regions of Africa: Algeria (DZA), Botswana (BWA), Egypt (EGY), Libya (LBY), Mauritius (MUS), Morocco (MAR), Namibia (NAM), Seychelles (SYC), South Africa (ZAF), Tunisia (TUN) and Zambia (ZMB). Most of these countries feature a high socioeconomic status and large international exposure.

Cluster 2 is the largest and is composed of 27 countries: Angola (AGO), Burundi (BDI), Cabo Verde (CPV), Central African Republic (CAF), Comoros (COM), Congo Brazzaville (COG), Cote d'Ivoire (CIV), Djibouti (DJI), Equatorial Guinea (GNQ), Eritrea (ERI), Eswatini (SWZ), Ethiopia (ETH), Gabon (GAB), Ghana (GHA), Kenya (KEN), Lesotho (LSO), Madagascar (MDG), Malawi (MWO), Mozambique (MOZ), Rwanda (RWA), Sao Tome and Principe (STP), Somalia (SOM), South Sudan (SSD), Sudan (SDN), Tanzania (TZA), Uganda (UGA) and Zimbabwe (ZWE). Most of these countries feature a middle to low socioeconomic status.

Cluster 3 is composed of the 16 remaining countries: Benin (BEN), Burkina Faso (BFA), Cameroon (CMR), Chad (TCD), Congo Kinshasa (COD), Gambia (GMB), Guinea

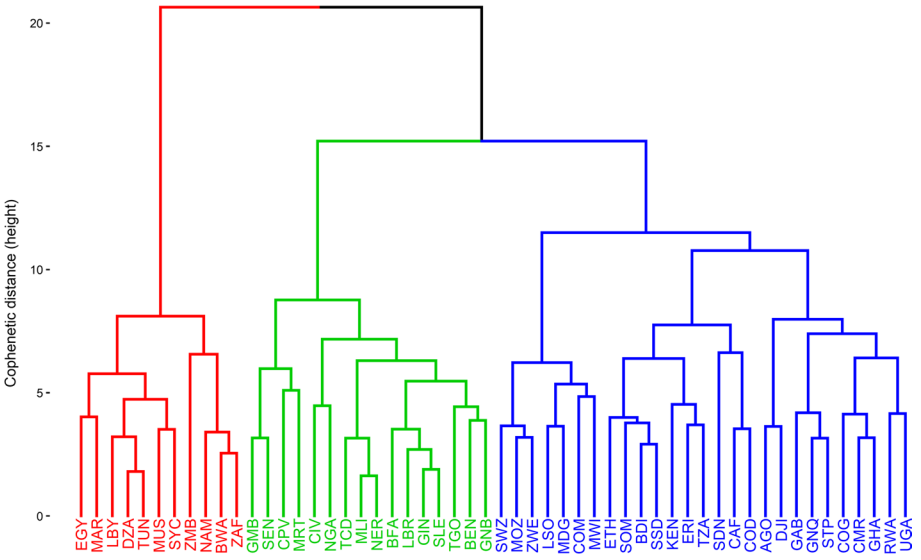

Fig. 5 Dendrogram of observations based on AHC using the optimal subset of 9 variables

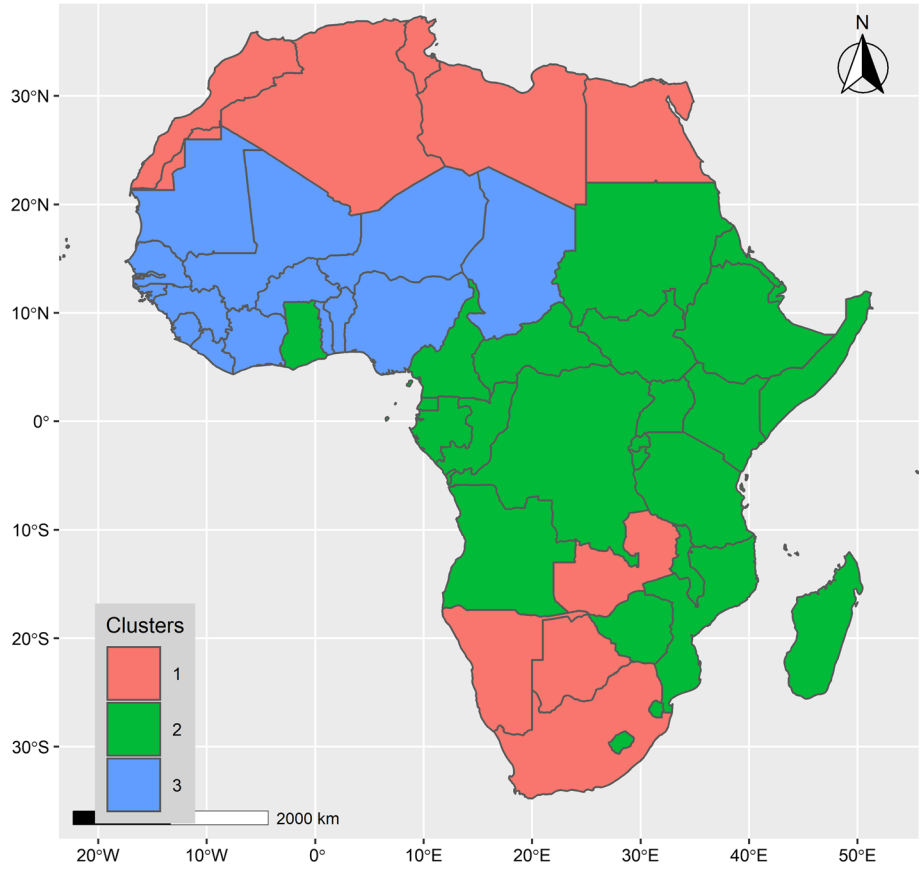

Fig. 6 Map of the 3 clusters identified in this study

(GIN), Guinea-Bissau (GNB), Liberia (LBR), Mali (MLI), Mauritania (MRT), Niger (NER), Nigeria (NGA), Senegal (SEN), Sierra Leone (SLE) and Togo (TOG). These countries feature an intermediate socioeconomic status, between countries of Cluster 1 and Cluster 2.

The map of clusters shows a substantial spatial differentiation. With a few exceptions, the countries in Cluster 1 (11 countries) are located at the Northern and Southern poles of the continent. The majority of countries in Cluster 2 (27 countries) are located in Central and Eastern regions of the continent. Finally, Cluster 3 (16 countries) is essentially composed of countries located in the western part of the continent.

3.3.2 Statistical analysis of clusters

The Cluster 1 is by far the largest affected cluster (median of 63,508.6 confirmed cases per million and 946.5 deaths per million), followed by Cluster 2 (median of 4473.7 confirmed cases per million and 81.2 deaths per million) and Cluster 3 (median of 2569.3 confirmed cases per million and 35.7 deaths per million). Clusters 2 and 3 share similar orders of magnitude in terms of COVID-19 prevalence.

Table 4 presents the statistical description of the 3 clusters. Cluster 1 is by far the largest affected cluster (median of 63,508.6 confirmed cases per million and 946.5 deaths per million), followed by Cluster 2 (median of 4473.7 confirmed cases per million and 81.2 deaths per million) and Cluster 3 (median of 2569.3 confirmed cases per million and 35.7 deaths per million). Clusters 2 and 3 share similar orders of magnitude in terms of COVID-19 prevalence.

Figure 7 compares the COVID-19 prevalence between the 3 clusters. The cumulative number of confirmed cases per million (*conf_pm*) and deaths per million (*death_pm*), respectively, in Fig. 7a and Fig. 7b shows significant differences for groups between Clusters 1–2 and Clusters 1–3 (p *values* < 0.01). Likewise, the mortality rate (shown in Fig. 7d) is found to be significantly different in Clusters 1–2 and 1–3 pairs (p *values* < 0.05). The case fatality rate, however, shown in Fig. 7c, remains similar across the 3 clusters (p *values: 0.648–1.000*).

Figure 8 compares the distribution of the 10 features used to form the 3 clusters. Significant differences (at $\alpha = 5\%$ level) are observed between the 3 pairs of clusters for the literacy rate (*alphab*) and environmental performance index (*epi*). Variables such as

Table 4 Statistical description of the 3 clusters of countries

Clusters	1	2	3	1	2	3	1	2	3	1	2	3
	lack_hygien			alphab			med_1000			epi		
Min	0.2	11.4	4.1	71.2	34.5	22.3	0.2	0.0	0.0	34.7	26.5	22.6
Q1	0.8	26.1	37.8	80.2	61.4	39.8	0.5	0.1	0.1	41.4	30.2	26.6
Q2	1.9	38.4	45.9	86.7	76.5	46.4	1.2	0.1	0.1	43.3	33.8	29.3
Avg	7.9	39.8	50.5	84.6	70.6	47.6	1.2	0.2	0.2	43.9	33.2	29.0
Q3	12.8	49.8	69.1	89.2	79.7	52.3	1.9	0.2	0.1	45.0	36.0	30.7
Max	34.9	86.6	101.0	95.9	94.4	86.8	2.5	0.7	0.8	58.2	45.8	38.3
Clusters	1	2	3	1	2	3	1	2	3	1	2	3
	pm25			urb_pop			arriv (in thousands)			lat_abs		
Min	10.5	15.3	42.2	40.8	13.4	16.5	428.0	33.4	30.0	4.7	0.0	6.4
Q1	24.6	26.1	54.0	47.6	25.9	40.8	1342.0	258.5	96.0	21.3	2.7	9.0
Q2	28.4	37.0	57.7	63.0	36.5	45.7	1830.0	812.0	277.0	26.3	6.9	12.0
Avg	31.6	35.1	59.5	59.9	42.0	44.4	5462.3	817.4	792.2	23.7	9.2	12.5
Q3	34.2	42.6	62.6	69.7	56.8	51.3	11,227.5	1184.0	787.5	29.3	13.1	15.7
Max	72.3	65.4	93.2	80.4	89.7	66.2	14,797.0	2294.0	5265.0	33.9	29.6	21.0
Clusters	1	2	3	1	2	3	1	2	3			
	gini			Confirmed cases (conf_pm)			Deaths (death_pm)					
Min	27.6	34.2	32.6	5033.1	583.0	377.6	159.7	3.3	12.0			
Q1	31.9	40.8	34.9	24,817.5	2577.6	1209.8	341.8	38.5	15.9			
Q2	36.8	43.7	35.8	63,508.6	4473.7	2569.3	946.5	81.2	35.7			
Avg	42.2	44.6	38.8	96,636.4	9227.1	16,575.5	1033.1	149.4	274.5			
Q3	55.2	47.9	42.6	110,600.6	11,036.5	4704.3	1645.5	140.2	96.5			
Max	63.0	56.3	50.7	414,043.5	60,769.3	217,378.4	2421.9	1214.1	3504.1			

'Min' is the minimum value, 'Q1' the first quartile, 'Q2' the second quartile, i.e., the median, 'Avg' is the average, 'Q3' is the third quartile, 'Max' is the maximum value

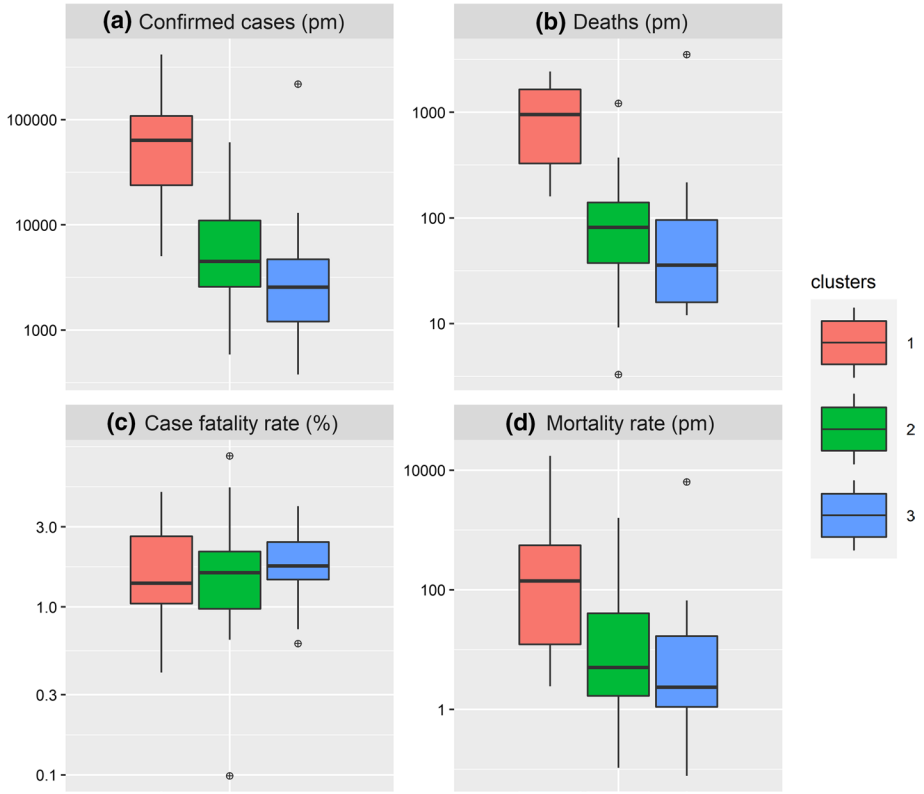

Fig. 7 Box-plot comparison of COVID-19 prevalence across clusters. **a** Cumulative cases per million (conf_pm). **b** Cumulative deaths per million (death_pm). **c** Case fatality rate (%), calculated as the cumulative number of deaths out of the cumulative number of confirmed cases. **d** Mortality (per million people), calculated as the cumulative number of deaths out of population estimates (WorldBank, 2021). The vertical axis was transformed to log10 scale for easier visual cross-comparison of clusters

international arrivals (arriv), urban population (urb_pop), number of physicians for 1000 inhabitants (*med_1000*), mortality attributed to the lack of hygiene (*lack_hygien*) and absolute latitude (lat_abs) present significant differences only for Clusters 1–2 and Clusters 1–3 pairs. Differences in Gini index distribution (*gini*) are found to be significant only between Clusters 2 and 3 (p value = 0.011), while air pollution due to PM2.5 particles (*pm2.5*) shows significant differences between Clusters 1–3 and Clusters 2–3 pairs.

Cluster 1 is by far the hard-hit cluster by COVID-19 with a median of 63,508.6 confirmed cases per million and 946.5 deaths per million. The countries in this cluster have the lowest mortality related to the lack of hygiene (median of 1.9%) and air pollution due to PM2.5 (median of 28.4), the highest literacy rate (median of 86.7%) and EPI score (median of 43.3). These countries are the most urbanized (median of 63.0%) and are located the farthest from the equator (median absolute latitude of 26.3°). International exposure, with the annual number of tourist arrivals is the highest for this cluster (median of 1,830,000 tourists).

Cluster 3 is the least affected by COVID-19 with a median of 2569.3 confirmed cases per million and 35.7 deaths per million. Interestingly, the countries in this cluster have

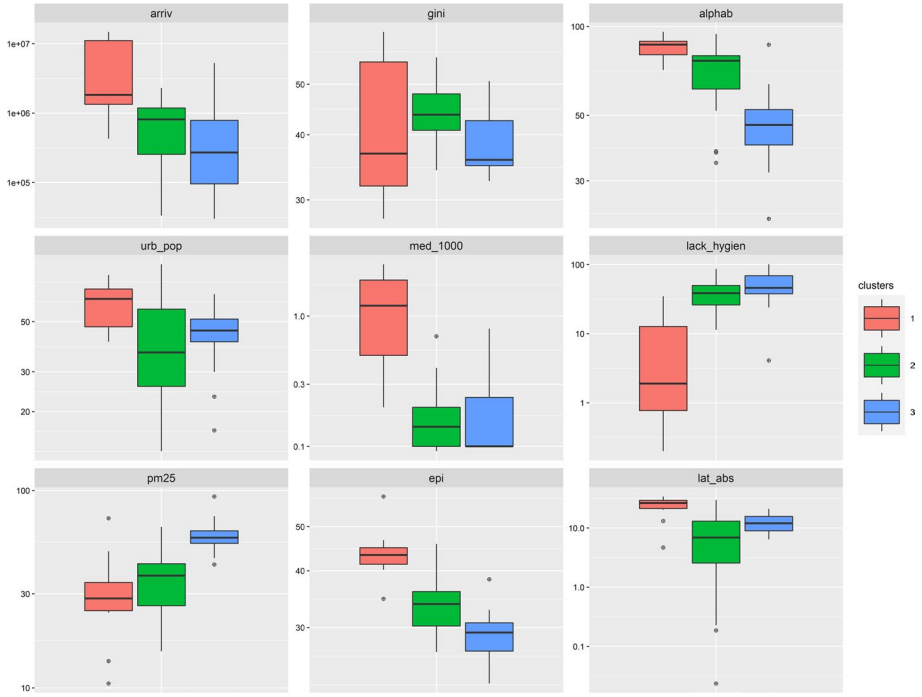

Fig. 8 Cluster comparison by variables. The vertical axis was transformed to log10 scale to enable visual cross-comparison across clusters

the highest mortality related to the lack of hygiene (median of 45.9%), the lowest literacy rate (median of 46.4%), EPI score (median of 29.3) and international exposure (median of 277,000 arrivals). Also, air pollution due to PM2.5 particles in countries of Cluster 3 is the highest (median of 57.7).

Cluster 2 is intermediate between Cluster 1 and 3 in terms of COVID-19 prevalence, as shown by the median values for confirmed cases (4473.7 cases per million) and deaths (81.2 deaths per million). Urban population densities are the lowest in this cluster (median of 36.5%). Also, these countries are geographically close to the equator (median absolute latitude of 6.9°). International exposure is intermediate for this cluster (median of 812,000 tourists).

Figure 9 shows the linear association between the natural logarithm of cumulative confirmed cases and deaths per million opposed to the absolute latitude.

The coefficients of determination (R^2) for this linear association are, respectively, of 0.098 (*p value = 0.063, not significant*) and 0.198 (*p value = 0.005, significant*) for cumulative confirmed cases and deaths per million. It shows that to some extent, the farther from the equator a country is located, the more deaths are to be expected to COVID-19, with a semi-elasticity of around 7.9% increase in deaths cases per million by one degree of absolute latitude increase. Similar observation have been reported in previous studies, which suggested that the higher sunlight and heat (near equator) is likely to hinder the spread of the COVID-19 (Braiman, 2020; Chen et al., 2021; Whittemore, 2020).

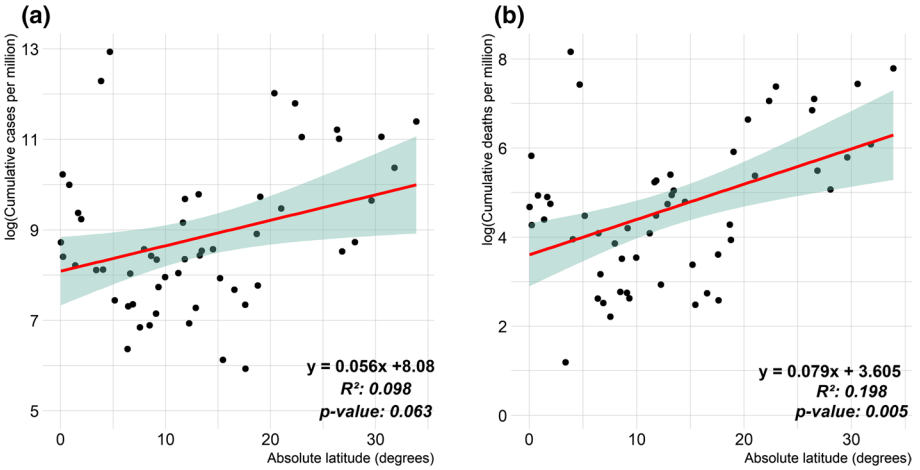

Fig. 9 Scatterplots of natural logarithm (log) of COVID-19 cases and deaths per million people opposed to absolute latitude (in degrees) for African countries. **a** COVID-19 cumulated confirmed cases ($R^2 = 0.063$, p value $= 0.063 > 0.05$). **b** COVID-19 cumulated deaths ($R^2 = 0.198$, p value $= 0.005 < 0.05$)

4 Discussion

4.1 On transmission factors and COVID-19 clustering

At a global level, the spread of the pandemic indicates that developed countries (such as USA, Italy, England, France, China and Russia) are also the most affected by the pandemic. A positive correlation between the high socioeconomic status, standard of living and COVID-19 prevalence has been reported in earlier studies (Dong et al., 2020), which is similar, to some extent, to the findings in this study: Cluster 1 in this study, for example, is the most affected by the pandemic and is also the one concentrating the leading countries in Africa, in terms of socioeconomic features (Cash & Patel, 2020). On the other hand, it appears that the prevalence of chronic respiratory infections and diseases is not relevant to COVID-19 prevalence in the context of Africa, which calls into question some of the previous studies (Carrillo-Larco & Castillo-Cara, 2020; Renzaho, 2020). According to Bigna and Noubiap (2019), there is a rising concern regarding the recent increase of non-communicable diseases (cardiovascular and respiratory diseases, cancers, diabetes) in Sub-Saharan Africa, mostly attributed to rapid urbanization and increased risk factors such as unhealthy diets, reduced physical activity, hypertension, obesity and air pollution (Kraef et al., 2020).

The countries which reported the first COVID-19 cases and deaths in this study are those found to have a higher international exposure mostly through tourism. This is in line with a recent large-scale genomic analysis, which specifically revealed that COVID-19 in most African countries was triggered by importations, predominantly from Europe. Yet, this spread slowed down following the early introduction of international travel restrictions. Furthermore, ongoing transmission and increasing mobility led to the emergence and spread of many variants within the continent (Wilkinson et al., 2021; Zongo et al., 2020).

Some variables related to the structure of the population (life expectancy, urban population) in this study also best explain the spread of the pandemic in Africa. The highest

values of life expectancy obtained for Cluster 1 appear mostly as a typical trait of topping countries, for which it is expected to be high because of the better standard of life, life amenities, healthcare facilities and management systems found in such countries. Similarly, for such countries, the rate of urbanization is expected to be high, which increases the risk of COVID-19 transmission (Carrillo-Larco & Castillo-Cara, 2020; Rizvi et al., 2021).

Zhu et al. (2020) and Rizvi et al. (2021) established that air quality is a positive predictor of the COVID-19 confirmed cases. This is supported by the findings in this study, as shown by the prominent variables highlighted in this study such as mortality related to air pollution due to PM2.5 and also the EPI score. COVID-19 and air pollution is already known to be a hazardous association. Recently emerging evidence suggests that exposure to air pollution worsens the severity of COVID-19 on human health (Bourdrel et al., 2021).

Regarding the climatic setting, only latitude was found to be effective at explaining COVID-19 prevalence, with a higher and significant association to COVID-19 deaths. Meo et al. (2020) showed that an increase in relative humidity and temperature is associated with a decrease in the number of daily cases and deaths due to COVID-19 in Africa. Other studies highlighted association between COVID-19 and various climate parameters, such as rainfall, wind speed and surface pressure (Bashir et al., 2020; Bilal et al., 2021; Hossain et al., 2021; Raza et al., 2021; Rendana, 2020; Ward et al., 2020). Interestingly, insolation has been reported as a negative predictor of COVID-19 prevalence, which in turn might explain why countries located farther from the equator tend to report more confirmed cases and especially deaths (Braiman, 2020; Chen et al., 2021; Whittemore, 2020). This latter finding is in line with our results. However, an in-depth assessment of the clear connection between climate and the current pandemic is yet to be carried. (Wang and Crameri 2014; Lone & Ahmad, 2020).

On the overall, little previous work has examined factors associated with the COVID-19 pandemic within the context of Africa. In this research, 3 clusters of countries are identified. In comparison, ACSS (2020b) conducted a clustering in Africa and identified 7 country profiles. Yet, our approach presents a significant difference as it tries to relate the variability in COVID-19 prevalence (cases and deaths) across countries through country-level variables, later used to form clusters. Moreover, environment-related variables are considered here, unlike ACSS (2020b). Other clustering-related research work conducted outside of Africa or at the global level concluded that countries with similar socioeconomic profiles fall within the same cluster (Carrillo-Larco & Castillo-Cara, 2020; Freed et al., 2020; Zarikas et al., 2020). It is, at a first glance, surprising to note that countries considered to be '*developed*' from the viewpoint of socioeconomic status or standard of living are the most severely affected by the COVID-19 pandemic. However, Freed et al. (2020) discusses the clear distinction to be made between the level of socioeconomic achievement for a country on the one side, and on the other side, the preparedness of healthcare systems as well as the willingness of populations to cope with restrictive measures promoted by authorities. These features are decisive to achieve a swift and effective response to the ongoing health crisis (Sadeghi et al., 2021; Zhang et al., 2020).

4.2 On the lack of hygiene and COVID-19 transmission

Handwashing is considered to be one of the most effective ways to prevent the transmission of diseases, including COVID-19. In this study, the mortality attributed to the lack of hygiene (*lack_hygien*) is found to be significant for both confirmed cases and deaths. Similarly, it was found to be determinant at separating optimally the 3 clusters

found. The *lack_hygien* variable is a negative predictor of the pandemic (conf_pm: ρ = -0.74; death_pm: ρ = -0.76). The lack of sanitation associated with poor hygiene practices is already deemed to be responsible for the higher communicable disease burdens, especially for developing countries (James et al., 2018). It is, therefore, reasonable to expect that better hygiene standards, safe sanitation and safe drinking water are likely to be negatively correlated with COVID-19 cases and deaths. Interestingly, the findings in our study suggest quite the opposite. In our understanding, this should not be perceived as a causation, but rather as a typical trait of the clusters formed instead. An explanatory hypothesis can be found in the possibility of '*immune training*,' as suggested by Chatterjee et al. (2020). In fact, the African context is a unique case where previous infectious diseases such as HIV, tuberculosis and malaria as well as infections are highly prevalent and are known to influence immune function, which might also, in turn, affect the immune response to COVID-19 (Adams et al., 2021; Tessema & Nkengasong, 2021). Also, along these lines, a lower prevalence of COVID-19 in malaria-endemic areas has been reported, although the reasons are yet to be further investigated (Anjorin et al., 2021; Iesa et al., 2020).

4.3 Implications for policies and decision making

Since the onset of the pandemic, Africa has experienced four waves in daily new cases, which seems to display a strong periodicity of approximately 6 months each. Such finding has direct implications for management policies, as it suggests that barrier measures, social distancing and eventually specific measures should be undertaken prior to the occurrence of such predictable peak periods (especially in the months of June-July and December-January).

Besides this, coupling some strategies and preventive measures might help in a strong mitigation of the spread of the pandemic on the African continent. For example, authorities should focus on good governance regarding health directives and open communication, to foster willingness of population to adopt mitigation measures, and also encourage them to get vaccinated. Providing financial support to vulnerable sectors of activities and populations might help in this regard, considering the limited resources of many African countries (James et al., 2018; Sadeghi et al., 2021).

Regarding the governance of the health sector, lack of knowledge is still hindering our understanding of the pandemic. Also, the emergence of new variants more virulent in younger populations is likely, which might lead in turn to reconsideration of Africa's susceptibility to the COVID-19 pandemic. As such, studies to assess risk factors including detailed cohort studies with appropriate controls are needed (Adams et al., 2021).

Regarding environment and sustainability aspects, the current figures of the COVID-19 pandemic can be perceived late lesson from an early warning. Human-induced environmental degradation increases the risk of pandemics through the complex interplay between ecosystem disturbance, urbanization, international travel and climate change. Therefore, a transition to a sustainable society and economy appears necessary to protect human health. As such, decision makers (at the institutional level) and societies (at the individual and community level) should start thinking about what to differently to move forward more sustainable practices (EEA, 2020; Harremoës et al., 2001).

4.4 Limitations of this study

It should be fully acknowledged that our study is fraught with a few limitations: first, since it is mostly based on statistical analysis, it might help highlighting evidence on a macroscopic scale. However, our framework could be less performant at explaining individual and specific variations among observations, which are typically masked. Moreover, since any model is as good as the data used, the limitations related to the supporting data used in this study should be considered. The reporting of COVID-19 data is subjected to different strategies depending on countries and might not be entirely accurate, nor up to date, depending on either technical limitations or communication strategies. Similarly, it is well known that only positive tests results are considered as confirmed cases: therefore, the less testing is done, the less confirmed cases are detected, which might not reflect accurately the actual state of the pandemic. On several occasions, the reliability of the tests has been questioned (Danilova, 2020). These issues and the subsequent uncertainty around the COVID-19 prevalence estimates should be considered as they might distort our understanding of the spread of the pandemic across different countries.

Finally, this study also focused on the African continent to explain the variability in COVID-19 prevalence across countries through country-level related variables. Such focus might bring a loss of generalization of our findings to other contexts outside Africa, or to urban areas. However, the framework of methodology could still be applied in such cases, with consideration of new potential variables more related to such contexts. Also, the clusters identified might help in COVID-19 modeling studies, since better performance might be achieved through the tuning models according to each cluster.

The findings in this study open new avenues for research regarding COVID-19 prevalence in Africa. Future studies should consider forecasting COVID-19 confirmed cases through time series modeling (Takele, 2020). Such modeling efforts might critically help in handling effectively the pandemic, but could also be useful in understanding spatial patterns of evolution of the pandemic and in assessing the effectiveness of mitigation of restrictive measures (Likassa et al., 2021). Also, future studies should consider assessing the potential effects of the pandemic on critical sectors to which most African countries are dependent, such as agricultural trade (Dugué et al., 2021; Lèye et al., 2021).

5 Conclusion

The current COVID-19 pandemic took the world by surprise early in 2020. The African continent, which turned out to be the least affected, has outwitted even the most sophisticated prognosticators. In this study, a set of 9 country-level descriptors have been identified as the optimal ones at explaining the variability of cumulative confirmed cases (per million) and cumulative deaths (per million) figures across the 54 African countries. The variables relating to the healthcare system and environment, international exposure and socioeconomic status are found to be closely related to COVID-19 prevalence, followed by variables relating to population structure. To a lesser extent, climate (through the geographical distance to the equator) might explain the current pandemic figures, more specifically in terms of cumulative deaths per million. A negative predictor, which is also the most significant variable, is the mortality related to lack of water, hygiene and sanitation.

Based on these optimal features, the African continent is partitioned into 3 epidemiological clusters using the AHC method. Cluster 1 is composed of 11 countries mainly located mainly at the northernmost and southernmost parts of the continent, and characterized by the highest median values for confirmed cases, deaths of COVID-19. It also has the highest socioeconomic and standard of living features. Conversely, the median value for mortality related to lack of water, hygiene and sanitation is the lowest for this cluster. Cluster 2 (27 countries) is the one the most spared by the current pandemic. It also has the lowest standard of living and is the part of the continent where mortality due to lack of water, hygiene and sanitation seems to be the highest. Cluster 3 (1 countries) is intermediate between Cluster 1 and 2 in terms of COVID-19 prevalence and mostly features countries with likewise similar socioeconomic features. Overall, in Africa, as in the rest of the world, richer or topping countries seem the most affected by this pandemic, as regard to reported statistics.

Some limitations of this study include the reliability of cases and deaths data reports, which might not be accurately reported on time. Also, it is of utmost importance to keep in mind that these reports are also limited by the testing policies applied by the different countries: the less testing is carried, the less active cases or deaths are reported. However, despite these limitations, the clustering produced in this study shed light on the nature and the exposure level to COVID-19 in African countries and might help fostering informed and strategic interventions by public authorities and decision makers for the current COVID-19 crisis and further epidemics or pandemics.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10668-022-02646-3.

Acknowledgements The authors are grateful to authors and maintainers of the COVID-19 Data Repository through the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19) and the NASA POWER global meteorology database (https://power.larc.nasa.gov/). The authors also thank M. Adam Sparks, author and maintainer of the R package *nasapower* (https://cran.r-project.org/web/packages/nasapower/), who swiftly provided critical help for collecting meteorological data through the use of his R client.

Code and data availability The code and data supporting the findings of this study are openly available through a public repository at https://github.com/Yonaba/covid19_clustering_africa.

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

ACSS. (2020b). Africa's Varied COVID Landscapes. *Africa Center for Strategic Studies*. https://africacenter.org/spotlight/africa-varied-covid-landscapes/. Accessed 14 September 2021

ACSS. (2020a). Mapping COVID-19 Risk Factors. *Africa Center for Strategic Studies*. https://africacenter.org/spotlight/mapping-risk-factors-spread-covid-19-africa/. Accessed 14 September 2021

Acuña, E., & Rodríguez, C. (2005). An empirical study of the effect of outliers on the misclassification error rate. *Submitted to Transactions on Knowledge and Data Engineering*.

Adams, J., MacKenzie, M. J., Amegah, A. K., Ezeh, A., Gadanya, M. A., Omigbodun, A., et al. (2021). The conundrum of low COVID-19 mortality burden in sub-Saharan Africa: Myth or reality? *Global Health: Science and Practice, 9*(3), 433–443. https://doi.org/10.9745/GHSP-D-21-00172

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics, 05*(07), 754–767. https://doi.org/10.4236/ojs.2015.57075

Anjorin, A. A., Abioye, A. I., Asowata, O. E., Soipe, A., Kazeem, M. I., Adesanya, I. O., et al. (2021). Comorbidities and the COVID-19 pandemic dynamics in Africa. *Tropical Medicine and International Health, 26*(1), 2–13. https://doi.org/10.1111/tmi.13504

Aydın, S., Nakiyingi, B. A., Esmen, C., Güneysu, S., & Ejjada, M. (2021). Environmental impact of coronavirus (COVID-19) from Turkish perceptive. *Environment, Development and Sustainability, 23*(5), 7573–7580. https://doi.org/10.1007/s10668-020-00933-5

Baker, R. E., Yang, W., Vecchi, G. A., Metcalf, C. J. E., & Grenfell, B. T. (2020). Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. *Science, 369*(6501), 315–319. https://doi.org/10.1126/science.abc2535

Bashir, M. F., Ma, B., Bilal, Komal, B., Bashir, M. A., Tan, D., & Bashir, M. (2020). Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Science of The Total Environment*, *728*, 13883. https://doi.org/10.1016/j.scitotenv.2020.138835

Bigna, J. J., & Noubiap, J. J. (2019). The rising burden of non-communicable diseases in sub-Saharan Africa. *The Lancet Global Health, 7*(10), e1295–e1296. https://doi.org/10.1016/S2214-109X(19)30370-5

Bashir, M. F., Shahzad, K., Komal, B., Bashir, M. A., Bashir, M., et al. (2021). Environmental quality, climate indicators, and COVID-19 pandemic: insights from top 10 most affected states of the USA. *Environmental Science and Pollution Research*, *28*(25), 32856–32865.https://doi.org/10.1007/s11356-021-12646-x

Bourdrel, T., Annesi-Maesano, I., Alahmad, B., Maesano, C. N., & Bind, M.-A. (2021). The impact of outdoor air pollution on COVID-19: A review of evidence from in vitro, animal, and human studies. *European Respiratory Review, 30*(159), 200242. https://doi.org/10.1183/16000617.0242-2020

Braiman, M. (2020). Latitude dependence of the COVID-19 mortality rate—A possible relationship to Vitamin D deficiency? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3561958

Carrillo-Larco, R. M., & Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. *Wellcome Open Research*, *5*, 56. https://doi.org/10.12688/wellcomeopenres.15819.3

Cash, R., & Patel, V. (2020). Has COVID-19 subverted global health? *The Lancet, 395*(10238), 1687–1688. https://doi.org/10.1016/S0140-6736(20)31089-8

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*(6), 1–36.

Chatterjee, B., Karandikar, R. L., & Mande, S. C. (2020). Paradoxical case fatality rate dichotomy of Covid-19 among rich and poor nations points to the "hygiene hypothesis" (preprint). *Epidemiology*. https://doi.org/10.1101/2020.07.31.20165696

Chen, S., Prettner, K., Kuhn, M., Geldsetzer, P., Wang, C., Bärnighausen, T., & Bloom, D. E. (2021). Climate and the spread of COVID-19. *Scientific Reports, 11*(1), 9042. https://doi.org/10.1038/s41598-021-87692-z

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall.

Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Bio-Medica: Atenei Parmensis*, *91*(1), 157–160. https://doi.org/10.23750/abm.v91i1.9397

Danilova, I. (2020). Morbidity and mortality from COVID-19. The problem of data comparability. *Demograficheskoe obozrenie*, 6–26.

Desjardins, J. (2019, February 15). Mapped: The Median Age of the Population on Every Continent. *Visual Capitalist*. https://www.visualcapitalist.com/mapped-the-median-age-of-every-continent/. Accessed 9 September 2021

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases, 20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Dugué, P., Kohio, E. N., & Tiemtoré, J. (2021). L'agriculture burkinabè face à la crise de la Covid-19: Cas des régions du Yatenga et des Hauts-Bassins. *Cahiers Agricultures, 30*, 16. https://doi.org/10.1051/cagri/2021002

EEA. (2020). Healthy Environment, Healthy Lives: How the Environment Influences Health and Well-Being in Europe. Publications Office of the European Union Luxembourg.

Freed, J. S., Kwon, S. Y., El Jacobs, H., Gottlieb, M., & Roth, R. (2020). Which country is truly developed? COVID-19 has answered the question. *Annals of Global Health, 86*(1), 51. https://doi.org/10.5334/aogh.2894

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery, 1*(1), 55–77. https://doi.org/10.1023/A:1009778005914

Gangloff, C., Rafi, S., Bouzillé, G., Soulat, L., & Cuggia, M. (2021). Machine learning is the key to diagnose COVID-19: A proof-of-concept study. *Scientific Reports, 11*(1), 7166. https://doi.org/10.1038/s41598-021-86735-9

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate, 30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gilbert, M., Pullano, G., Pinotti, F., Valdano, E., Poletto, C., Boëlle, P.-Y., et al. (2020). Preparedness and vulnerability of African countries against importations of COVID-19: A modelling study. *The Lancet, 395*(10227), 871–877. https://doi.org/10.1016/S0140-6736(20)30411-6

Goldstein, J. R., & Lee, R. D. (2020). Demographic perspectives on the mortality of COVID-19 and other epidemics. *Proceedings of the National Academy of Sciences, 117*(36), 22035–22041. https://doi.org/10.1073/pnas.2006392117

Harremoës, P., Gee, D., MacGarvin, M., Stirling, A., Keys, J., Wynne, B., & Vaz, S. G. (2001). *Late lessons from early warnings: the precautionary principle 1896–2000*. Citeseer.

Heneghan, C., Aronson, J., Hobbs, H., & Mahtani, K. (2020, March 16). Rapidly managing pneumonia in older people during a pandemic. *The Centre for Evidence-Based Medicine*. https://www.cebm.net/covid-19/rapidly-managing-pneumonia-in-older-people-during-a-pandemic/. Accessed 14 September 2021

Hossain, Md. S., Ahmed, S., & Uddin, Md. J. (2021). Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model. *Science of the Total Environment, 761*, 143315. https://doi.org/10.1016/j.scitotenv.2020.143315

Iesa, M. A. M., Osman, M. E. M., Hassan, M. A., Dirar, A. I. A., Abuzeid, N., Mancuso, J. J., et al. (2020). SARS-CoV-2 and Plasmodium falciparum common immunodominant regions may explain low COVID-19 incidence in the malaria-endemic belt. *New Microbes and New Infections, 38*, 100817. https://doi.org/10.1016/j.nmni.2020.100817

IHME. (2020). Global Health Data Exchange | GHDx. *Global Health Data Exchange*. http://ghdx.healthdata.org/. Accessed 14 September 2021

Imtyaz, A., Haleem, A., & Javaid, M. (2020). Analysing governmental response to the COVID-19 pandemic. *Journal of Oral Biology and Craniofacial Research, 10*(4), 504–513. https://doi.org/10.1016/j.jobcr.2020.08.005

Iribarne, J. V., & Godson, W. L. (1973). *Atmospheric thermodynamics*. Dordrecht; Boston: Reidel. http://books.google.com/books?id=MoYuAAAAIAAJ. Accessed 18 September 2021

Islam, A. RMd. T., Hasanuzzaman, Md., Azad, Md. A. K., Salam, R., Toshi, F. Z., Khan, Md. S. I., et al. (2021). Effect of meteorological factors on COVID-19 cases in Bangladesh. *Environment, Development and Sustainability, 23*(6), 9139–9162. https://doi.org/10.1007/s10668-020-01016-1

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet, 392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Kassambara, A. (2017). *Practical guide to cluster analysis in R: unsupervised machine learning* (Edition 1.). Erscheinungsort nicht ermittelbar: STHDA.

Kerr, G. H., Badr, H. S., Gardner, L. M., Perez-Saez, J., & Zaitchik, B. F. (2021). Associations between meteorology and COVID-19 in early studies: Inconsistencies, uncertainties, and recommendations. *One Health, 12*, 100225. https://doi.org/10.1016/j.onehlt.2021.100225

Kraef, C., Juma, P., Kallestrup, P., Mucumbitsi, J., Ramaiya, K., & Yonga, G. (2020). The COVID-19 Pandemic and Non-communicable diseases—A wake-up call for primary health care system strengthening in Sub-Saharan Africa. *Journal of Primary Care and Community Health, 11*, 215013272094694. https://doi.org/10.1177/2150132720946948

Kuhn, M. (2021). *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models*. McGraw-Hill.

Lawal, Y. (2021). Africa's low COVID-19 mortality rate: A paradox? *International Journal of Infectious Diseases, 102*, 118–122. https://doi.org/10.1016/j.ijid.2020.10.038

Lèye, B., Zouré, C. O., Yonaba, R., & Karambiri, H. (2021). Water Resources in the Sahel and Adaptation of Agriculture to Climate Change: Burkina Faso. In S. Diop, P. Scheren, & A. Niang (Eds.), *Climate Change and Water Resources in Africa* (pp. 309–331). Springer. http://link.springer.com/https://doi.org/10.1007/978-3-030-61225-2_14

Likassa, H. T., Xain, W., Tang, X., & Gobebo, G. (2021). Predictive models on COVID 19: What Africans should do? *Infectious Disease Modelling, 6*, 302–312. https://doi.org/10.1016/j.idm.2020.10.015

Lone, S. A., & Ahmad, A. (2020). COVID-19 pandemic – an African perspective. *Emerging Microbes and Infections, 9*(1), 1300–1308. https://doi.org/10.1080/22221751.2020.1775132

Lulbadda, K. T., Kobbekaduwa, D., & Guruge, M. L. (2021). The impact of temperature, population size and median age on COVID-19 (SARS-CoV-2) outbreak. *Clinical Epidemiology and Global Health, 9*, 231–236. https://doi.org/10.1016/j.cegh.2020.09.004

Luo, W., Majumder, M. S., Liu, D., Poirier, C., Mandl, K. D., Lipsitch, M., & Santillana, M. (2020). The role of absolute humidity on transmission rates of the COVID-19 outbreak (preprint). *Epidemiology*. https://doi.org/10.1101/2020.02.12.20022467

Madhav, N., Oppenheim, B., Gallivan, M., Mulembakani, P., Rubin, E., & Wolfe, N. (2017). *Pandemics: Risks, Impacts, and Mitigation* (3rd ed.). The International Bank for Reconstruction and Development / The World Bank, Washington (DC). http://europepmc.org/books/NBK525302

Medford, A., & Trias-Llimós, S. (2020). Population age structure only partially explains the large number of COVID-19 deaths at the oldest ages. *Demographic Research, 43*, 533–544.

Meo, S., Abukhalaf, A., Alomar, A., Aljudi, T., Bajri, H., Sami, W., et al. (2020). Impact of weather conditions on incidence and mortality of COVID-19 pandemic in Africa. *European Review for Medical and Pharmacological Sciences, 24*(18), 9753–9759.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159–179. https://doi.org/10.1007/BF02294245

Moosa, I. A., & Khatatbeh, I. N. (2020). International tourist arrivals as a determinant of the severity of Covid-19: International cross-sectional evidence. *Journal of Policy Research in Tourism, Leisure and Events*. https://doi.org/10.1080/19407963.2020.1859519

Nuwagira, E., & Muzoora, C. (2020). Is Sub-Saharan Africa prepared for COVID-19? *Tropical Medicine and Health, 48*(1), 18. https://doi.org/10.1186/s41182-020-00206-x

Pearson, C. A., Van Schalkwyk, C., Foss, A. M., O'Reilly, K. M., SACEMA Modelling and Analysis Response Team, CMMID COVID-19 working group, & Pulliam, J. R. (2020). Projected early spread of COVID-19 in Africa through 1 June 2020. *Eurosurveillance*, *25*(18). https://doi.org/10.2807/1560-7917.ES.2020.25.18.2000543

Rahman, M., Islam, M., Shimanto, M. H., Ferdous, J., Rahman, A.A.-N.S., Sagor, P. S., & Chowdhury, T. (2021). A global analysis on the effect of temperature, socio-economic and environmental factors on the spread and mortality rate of the COVID-19 pandemic. *Environment, Development and Sustainability, 23*(6), 9352–9366. https://doi.org/10.1007/s10668-020-01028-x

Ramosaj, B., & Pauly, M. (2019). Predicting missing values: A comparative study on non-parametric approaches for imputation. *Computational Statistics, 34*(4), 1741–1764. https://doi.org/10.1007/s00180-019-00900-3

Randazzo, W., Cuevas-Ferrando, E., Sanjuán, R., Domingo-Calap, P., & Sánchez, G. (2020). Metropolitan wastewater analysis for COVID-19 epidemiological surveillance. *International Journal of Hygiene and Environmental Health, 230*, 113621. https://doi.org/10.1016/j.ijheh.2020.113621

Raza, A., Khan, M. T. I., Ali, Q., Hussain, T., & Narjis, S. (2021). Association between meteorological indicators and COVID-19 pandemic in Pakistan. *Environmental Science and Pollution Research, 28*(30), 40378–40393. https://doi.org/10.1007/s11356-020-11203-2

Rendana, M. (2020). Impact of the wind conditions on COVID-19 pandemic: A new insight for direction of the spread of the virus. *Urban Climate, 34*, 100680. https://doi.org/10.1016/j.uclim.2020.100680

Renzaho, A. (2020). The need for the right socio-economic and cultural fit in the COVID-19 response in Sub-Saharan Africa: Examining demographic, economic political, health, and socio-cultural differentials in COVID-19 morbidity and mortality. *International Journal of Environmental Research and Public Health, 17*(10), 3445. https://doi.org/10.3390/ijerph17103445

Rizvi, S. A., Umair, M., & Cheema, M. A. (2021). *Clustering of Countries for COVID-19 Cases based on Disease Prevalence, Health Systems and Environmental Indicators* (preprint). Epidemiology. https://doi.org/10.1101/2021.02.15.21251762

Sadeghi, B., Cheung, R. C. Y., & Hanbury, M. (2021). Using hierarchical clustering analysis to evaluate COVID-19 pandemic preparedness and performance in 180 countries in 2020. *British Medical Journal Open, 11*(11), e049844. https://doi.org/10.1136/bmjopen-2021-049844

Şahin, M. (2020). Impact of weather on COVID-19 pandemic in Turkey. *Science of the Total Environment, 728*, 138810. https://doi.org/10.1016/j.scitotenv.2020.138810

Salyer, S. J., Maeda, J., Sembuche, S., Kebede, Y., Tshangela, A., Moussif, M., et al. (2021). The first and second waves of the COVID-19 pandemic in Africa: A cross-sectional study. *The Lancet, 397*(10281), 1265–1275. https://doi.org/10.1016/S0140-6736(21)00632-2

Singh, O., Bhardwaj, P., & Kumar, D. (2021). Association between climatic variables and COVID-19 pandemic in National Capital Territory of Delhi, India. *Environment, Development and Sustainability, 23*(6), 9514–9528. https://doi.org/10.1007/s10668-020-01003-6

Sparks, A. (2021). *nasapower: NASA-POWER Data from R*. https://doi.org/10.5281/zenodo.1040727

Stekhoven, Daniel J. (2013). *missForest: Nonparametric Missing Value Imputation using Random Forest*. https://cran.r-project.org/web/packages/missForest/index.html

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Takele, R. (2020). Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries. *Infectious Disease Modelling, 5*, 598–607. https://doi.org/10.1016/j.idm.2020.08.005

Tessema, S. K., & Nkengasong, J. N. (2021). Understanding COVID-19 in Africa. *Nature Reviews Immunology, 21*(8), 469–470. https://doi.org/10.1038/s41577-021-00579-y

UNDP. (2020). *UNDP Annual Report 2020*. UNDP. https://annualreport.undp.org/assets/UNDP-Annual-Report-2020-fr.pdf

Visalakshi, N. K., & Suguna, J. (2009). K-means clustering using Max-min distance measure. In *NAFIPS 2009 - 2009 Annual Meeting of the North American Fuzzy Information Processing Society* (pp. 1–6). Presented at the NAFIPS 2009 - 2009 Annual Meeting of the North American Fuzzy Information Processing Society, Cincinnati, OH, USA: IEEE. https://doi.org/10.1109/NAFIPS.2009.5156398

Wang, & Crameri, G. (2014). Emerging zoonotic viral diseases: -EN- -FR- Les maladies zoonotiques virales émergentes -ES- Enfermedades zoonóticas emergentes de origen vírico. *Revue Scientifique et Technique de l'OIE*, *33*(2), 569–581. https://doi.org/10.20506/rst.33.2.2311

Wang, T., & K., Feng, K., Lin, X., Lv, W., Chen, K., & Wang, F. (2021). Impact of temperature and relative humidity on the transmission of COVID-19: A modelling study in China and the United States. *British Medical Journal Open, 11*(2), e043863. https://doi.org/10.1136/bmjopen-2020-043863

Ward, M. P., Xiao, S., & Zhang, Z. (2020). The role of climate during the COVID-19 epidemic in New South Wales, Australia. *Transboundary and Emerging Diseases, 67*(6), 2313–2317. https://doi.org/10.1111/tbed.13631

Weiss, D. J., Bertozzi-Villa, A., Rumisha, S. F., Amratia, P., Arambepola, R., Battle, K. E., et al. (2021). Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: A geospatial modelling analysis. *The Lancet Infectious Diseases, 21*(1), 59–69. https://doi.org/10.1016/S1473-3099(20)30700-3

Whittemore, P. B. (2020). COVID-19 fatalities, latitude, sunlight, and vitamin D. *American Journal of Infection Control, 48*(9), 1042–1044. https://doi.org/10.1016/j.ajic.2020.06.193

WHO. (2020). WHO Immunization Data portal. *WHO Immunization Data portal*. https://immunizationdata.who.int/. Accessed 14 September 2021

Wilkinson, E., Giovanetti, M., Tegally, H., San, J. E., Lessels, R., Cuadros, D., et al. (2021). A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa (preprint). *Genetic and Genomic Medicine*. https://doi.org/10.1101/2021.05.12.21257080

WorldBank. (2021). DataBank | The World Bank. *The World Bank Data Bank*. https://databank.worldbank.org/home. Accessed 14 September 2021

Yale. (2020). Environmental Performance Index. *EPI | Environmental Performance Index*. https://epi.yale.edu/epi-results/2020/component/epi. Accessed 14 September 2021

Zaitchik, B. F., Sweijd, N., Shumake-Guillemot, J., Morse, A., Gordon, C., Marty, A., et al. (2020). A framework for research linking weather, climate and COVID-19. *Nature Communications, 11*(1), 5730. https://doi.org/10.1038/s41467-020-19546-7

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief, 31*, 105787. https://doi.org/10.1016/j.dib.2020.105787

Zhang, S., Wang, Z., Chang, R., Wang, H., Xu, C., Yu, X., et al. (2020). COVID-19 containment: China provides important lessons for global response. *Frontiers of Medicine, 14*(2), 215–219. https://doi.org/10.1007/s11684-020-0766-9

Zhu, Y., Xie, J., Huang, F., & Cao, L. (2020). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. *Science of the Total Environment, 727*, 138704. https://doi.org/10.1016/j.scitotenv.2020.138704

Zongo, P., Zorom, M., Mophou, G., Dorville, R., & Beaumont, C. (2020). A model of COVID-19 transmission to understand the effectiveness of the containment measures: Application to data from France. *Epidemiology and Infection, 148*, e221. https://doi.org/10.1017/S0950268820002162