Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Scientific African 12 (2021) e00827

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Spread of COVID-19 in Zambia: An assessment of environmental and socioeconomic factors using a classification tree approach

Darius Phiri [a,*], Serajis Salekin [b], Vincent R. Nyirenda [c], Matamyo Simwanda [a], Manjula Ranagalage [d,e], Yuji Murayama [e]

[a] Department of Plant and Environmental Sciences, School of Natural Resources, Copperbelt University, Kitwe 10101, Zambia
[b] Scion, Titokorangi Drive (formerly Longmile Road), Private Bag 3020, Rotorua 3046, New Zealand
[c] Department of Zoology and Aquatic Sciences, School of Natural Resources, Copperbelt University, P.O. Box 21692, Kitwe 10101, Zambia
[d] Department of Environmental Management, Faculty of Social Sciences and Humanities, Rajarata University of Sri Lanka, Mihintale 50300, Sri Lanka
[e] Faculty of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba City, Ibaraki 305-8572, Japan

ARTICLE INFO

Article history:
Received 26 January 2021
Revised 27 May 2021
Accepted 27 June 2021

Editor: DR B Gyampoh

Keyword:
COVID-19
HIV/AIDS
Africa
GIS
Zambia
Classification Tree

ABSTRACT

The global pandemic emergent from SARS-COV-2 (COVID-19) has continued to cause both health and socio-economic challenges worldwide. However, there is limited information on the factors affecting the dynamics of COVID-19, especially in developing countries, including African countries. In this study, we have focused on understanding the association of COVID-19 cases with environmental and socioeconomic factors in Zambia - a sub-Saharan African country. We used Zambia's district-level COVID-19 data, covering 18 March 2020 (i.e., from first reported cases) to 17 July 2020. Geospatial approaches were used to organize, extract and establish the dataset, while a classification tree (CT) technique was employed to analyze the factors associated with the COVID-19 cases. The analyses were conducted in two stages: (1) the binary analysis of occurrences of COVID-19 (i.e., COVID-19 or No COVID-19), and (2) a risk level analysis which grouped the number of cases into four risk levels (high, moderate, low and very low). The results showed that the distribution of COVID-19 cases in Zambia was significantly influenced by the socioeconomic factors compared to environmental factors. More specifically, the binary model showed that distance to the airport, population density and distance to the town centres were the most combination influential factors, while the risk level analysis indicated that areas with high rates of human immuno-deficient virus (HIV) infection had relatively high chances of having many COVID-19 cases compared to areas with low HIV rates. The districts that are far from major urban establishments and that experience higher temperatures have lower chances of having COVID-19 cases. This study makes two major contributions towards the understanding of COVID-19 dynamics: (1) the methodology presented here can be effectively applied in other areas to understand the association of environmental and socioeconomic factors with COVID-19 cases, and (2), the findings from this study present the empirical

* Corresponding author.
E-mail address: dariusphiri@rocketmail.com (D. Phiri).

https://doi.org/10.1016/j.sciaf.2021.e00827

evidence of the relationship between COVID-19 cases and their associated environmental and socioeconomic factors. Further studies are needed to understand the relationship of this disease and the associated factors in different cultural settings, seasons and age groups, especially as the COVID-19 cases increase and spread in many countries.

Introduction

The novel Coronavirus disease 2019 (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) [20,47], has posed serious threats to global human health. This disease has also triggered widespread socioeconomic challenges around the world ([21]; Buheji et al., 2020; Nicola et al., 2020). The first confirmed case of COVID-19 was reported in Wuhan, China in late December 2019 [29,43,45]. By 31 December 2019, the World Health Organisation (WHO) received information about this epidemic [3], and WHO declared that COVID-19 was a public health emergency of international concern by 30 January 2020 [29,46]. Despite the different measures (e.g., mandatory lockdown, social distancing, masking) put in place across the world [14], influence of the virus remains high. As of 24 May 2021, there were more than 164 million confirmed cases of COVID-19; more than 3.4 million deaths and more than 1.4 million vaccine doses were administered [46].

The COVID-19 pandemic was confirmed to have spread in Africa on 14 February 2020, with the first confirmed case reported in Egypt [1,24,26]. Most of the cases that were identified in Africa in the initial stages of the pandemic were reported to have been imported from three locations; (1) Europe, (2) the United States of America (USA), and (3) China [1,24,26]. By mid-June, Africa had surpassed 200,000 cases, with 70% of these cases reported in South Africa, Egypt, Nigeria, Ghana and Algeria [1,24]. One of the major challenges in the management of and the fight against COVID-19 in many African countries, such as Zambia, is the poor health care systems [1,26]. Compared to other countries in the region such as South Africa, Botswana and Namibia, Zambia has a weaker health care system because of lack of modern facilities in form of infrastructure and equipment, and the low funding [1,23].

Zambia reported the first two cases of COVID-19 on 18 March 2020, which were imported from France [31,42]. By 25 March 2020, the number of confirmed cases had risen to 12. In July, the cases reached over 1,000 within 10 days reporting period, bringing the total number of cases to 4,481. Several studies have been done on COVID-19 in Zambia with specific emphases on the impact of COVID-19 on the education system [42], spreading [31], testing [23] and vulnerability [40].

Geospatial and statistical (i.e., geo-statistics) modelling offer an opportunity to assess the empirical relationships between different factors and challenges [30,35], such as other infectious diseases (e.g., Tuberculosis) [18]. This kind of assessments provide information on the most influential factors and how they relate with other factors [8,30]. For example, Mollalo et al. [29] used geo-statistical approaches to model the spread of COVID-19 in the USA.

To the best of our knowledge, the current study is the first study that has focused on assessing the relationships of COVID-19 cases with environmental and socioeconomic factors using a combination of a geospatial and a statistical approach for Zambia. Here we try to address the lack of information on the relationship between COVID-19 and socioeconomic factors, as well as environmental factors in Zambia. Therefore, this study has two primary objectives: (1) to understand the relationship between COVID-19 and the associated environmental and socio-economic factors, and (2) to explain the factors associated with different risk levels across Zambia.

Materials and methods

Study area

The study was conducted in Zambia, a sub-Saharan African country located between latitudes 8°S and 18°S and longitudes 22°E to 34°E, sharing its borders with eight countries (Malawi, Tanzania, Mozambique, Democratic Republic of Congo, Botswana, Angola, Namibia and Zimbabwe [34]. Elevation ranges between 2300 320 m above sea level, with the Mafinga Hills being the highest point, while the Zambezi river being the lowest level. Zambia is characterised by a sub-tropical climate with temperatures ranging from 7 to 37 °C; June and July are the coldest months, while September and October are the hottest months [33]. The country receives rainfall between October and April, which ranges from 800 mm (in the southern region) to 1500 (northern region) mm annually. The major vegetation type is the dry tropical forests called “Miombo” and Zambia is also rich in wildlife resources with one of the largest National Park in the world - the Kafue National Park [36].

Based on the projection from the 2010 national census, Zambia has an estimated population of 18 million people, with Lusaka being the most populated district (over 2 million people) [9]. Of this population, 35% is found in urban areas, and

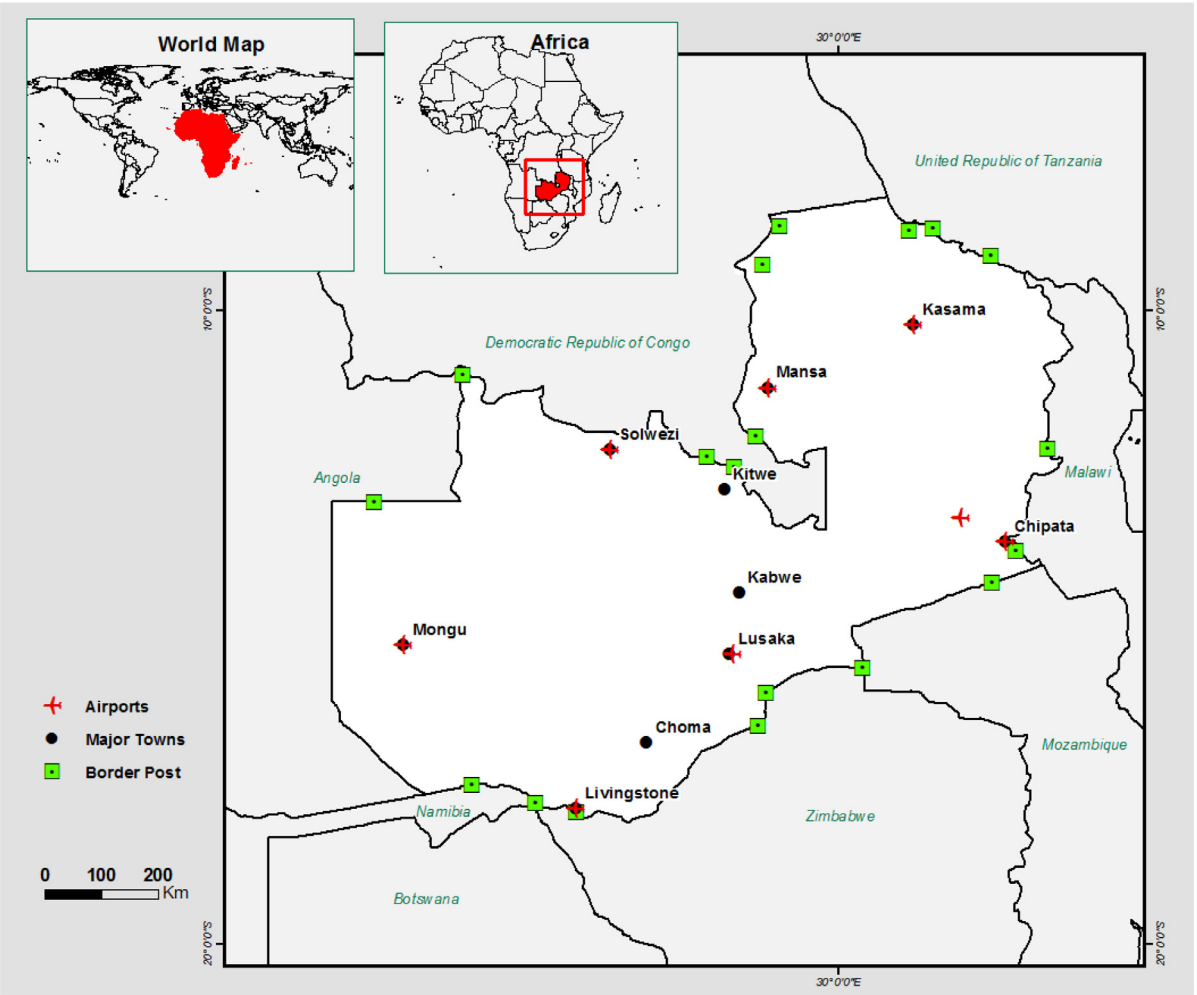

Fig. 1. The location of the study site - Zambia. The map also shows the location of the neighboring countries, airports, border posts and the major towns in Zambia.

65% are in rural areas. During the 2010 population census, 99% of the population were black Africans and 1% consisted of other racial groups, mostly from Europe and America [9]. In the last ten years, Zambia has also seen an increase in the population of immigrants of Asian origin, especially from China [9]. The major challenges in terms of diseases include seasonal outbreaks of waterborne diseases, such as cholera, and malaria which is caused by mosquitos [17,32].

Due to trade and tourist activities, Zambia has experienced migration involving people from all over the world, mainly through the international airports. Migration is also common through the land border posts, shared with the eight neighbouring countries (Fig. 1). The major economic activities in Zambia are copper mining and small-scale agriculture [34].

Zambia's population comprises more than 70 tribes, which are defined by distinct cultural practices, and these ethnic groups are located across the whole country in traditional structures [9]. The rural population depends on agriculture, while the people in urban areas depend on agricultural output from the rural areas for food. Thus, there is continuous migration by both rural and urban populations, driven by agricultural production [5,9].

Methods

Data

Response variables: COVID-19 cases per district. The major dataset used in this study is the reported cases of COVID-19 at district level (see Fig. 2). As a way of monitoring the COVID-19 disease, the Ministry of Health in Zambia presents daily public updates on the spread of COVID-19. We used the data reported on 17 July 2020 because this was when we started the analysis for this study. It is important to note that, as of 2018, Zambia was divided, administratively, into 10 provinces and 117 districts.

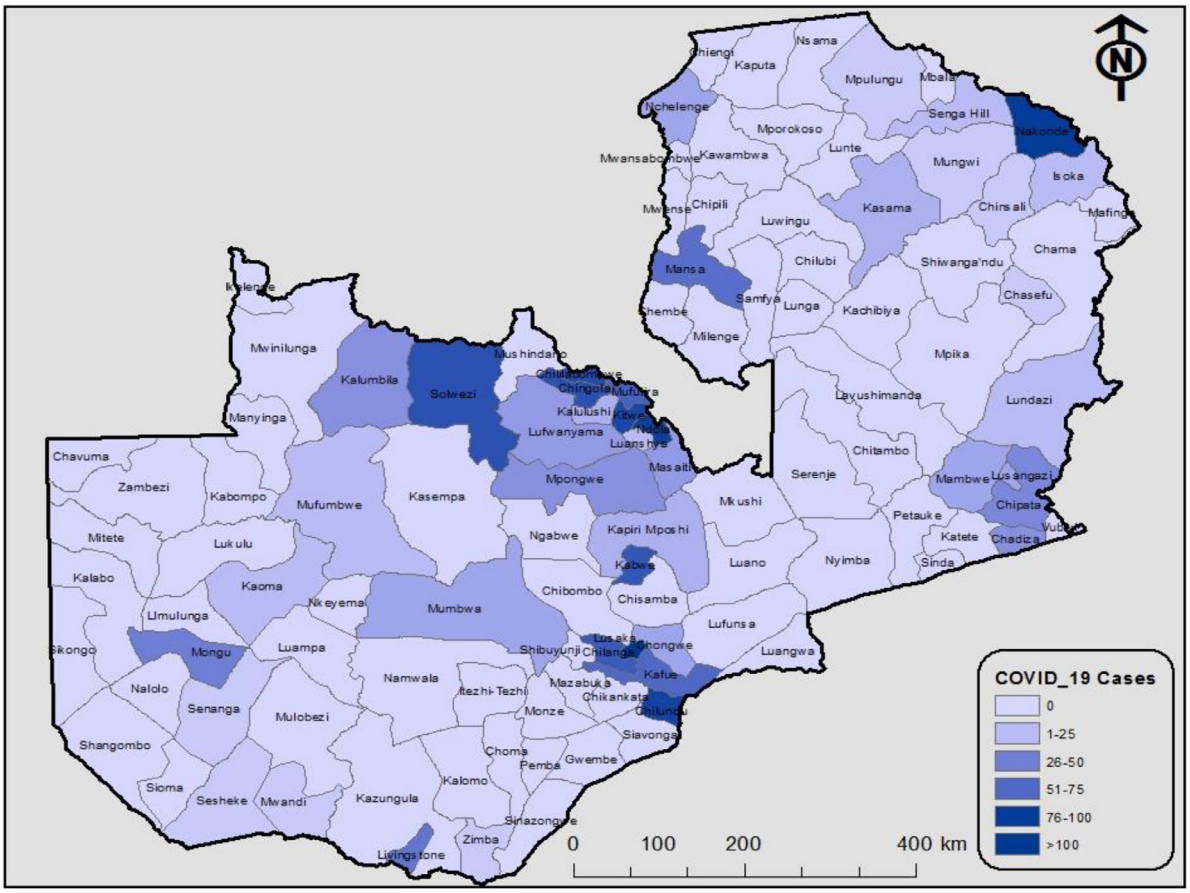

Fig. 2. The distribution of COVID-19 cases in Zambia at district level as of 17 July 2020 (Modified from the Ministry of Health daily reports).

Table 1
The response variables used for both the binary and risk level Classification Tree models for assessing COVID-19 cases in Zambia.

Model	Level	Range of case	Number of Districts
Binary	Presence of COVID-19	≥1	68
	Absence of COVID-19	0	44
Risk levels	High	>100	9
	Moderate	50–100	21
	Low	5–50	24
	Very low	<5	12

As of 17 July 2020, Zambia had 2,810 cases, 1,450 recoveries, and 102 deaths. Then, on 28 July 2020, Zambia had recorded 5,002 cases with 137 deaths (39 COVID-19 and 98 COVID-19 related) and 3,195 recoveries. The general spatial distribution of COVID-19 in Zambia is that the areas with high population density and border towns have many cases of COVID-19. Lusaka, the capital district, was the worst affected, with 1,232 cases, followed by Nakonde (645 cases), a border town in Muchinga Province in the northern part of the country. The Ministry of Health highlighted the sudden increase in the number of cases during the last two weeks of July [23].

Here, the response variables were categorized into two groups, depending on the models developed: (1) the binary, and (2) the risk levels. The binary responses included presence or absence of COVID-19 cases, while the risk level responses had four levels: high, moderate, low, or very low (Table 1). The districts which had reported at least one case of COVID-19 were categorised as “COVID-19 presence”, while those areas were the disease was yet to be reported were considered to be “COVID-19 absent”.

Explanatory variables. Potential factors that influence the spread of COVID-19 were divided into two groups: (1) environmental factors, and (2) socioeconomic factors. The environmental factors include topographic and climatic factors, while so-

cioeconomic factors included demographics, proximity to different commercial and urban establishments, such as airports, town centres, accessibility and economic factors (Table 2).

These factors are linked to COVID-19 in different ways; for example, environmental factors such as elevation, slope, aspects influence how habitable an area is. Hence, dictate the population and social economic activities (e.g. migration), while climatic conditions such as the temperature, humidity and rainfall have been closely associated with the spread of COVID-19 [3]. This is so because COVID-19 has been reported to quickly spread during the cold season. Socioeconomic activities determine many aspects of a population such as movements, access to facility and lifestyle. These aspects are some of the issues which have been considered by many countries for COVID-19 mitigation such as reducing people's movements through lockdowns.

So far, many studies [3,6,27] have associated the distribution of COVID-19 with climatic factors throughout the world, especially temperature. Menebo [27] assessed the correlation of COVID-19 cases with temperature and precipitation in the USA and reported a strong pattern. Precipitation is also an important factor that has the potential to influence the distribution of COVID-19 cases because precipitation is also associated with the temperature of a location [6]. Solar radiance was also considered in the analysis because it is a measure of the solar energy received in an area [13]. In our study, the climatic factors (i.e., precipitation, temperature and solar radiance) were based on simulated monthly values provided under the WorldClim dataset [12] (see Table 2).

Topographic factors, such as altitude, slope, and aspect, influence the environmental conditions in different ways, including influencing the local climate. The topographic factors (i.e., slope, altitude and aspect) were processed from the digital elevation models (DEM) acquired from the United States Geological Survey (USGS) website (https://earthexplorer.usgs.gov/). Both the climatic factors and topographic factors were prepared in a raster format, and the values were extracted using a "Zonal Statistics" tool in ArcMap 10.7 (ESRI, Redland, CA, USA) (Table 2).

Proximity factors are those factors associated with distances from features or locations (e.g. towns) [41]. We considered factors such as distance to the mines, towns and border posts. These proximity factors determine different economic activities that take place in a given area. For example, it is more likely that those areas near towns will have high employment and good road infrastructures. All these factors coerce more inhabitants to migrate to those areas hence increasing the chances of spreading COVID-19 [21] (see Table 2).

Accessibility determines the transportation infrastructure of an area and accessibility influences people's mobility. Here we included distance to road networks, railways, and airports [19,22]. Given that the first COVID-19 cases were reported to have been imported into Zambia, distance to the international and provincial airports was an important factor. Both the accessibility and proximity factors were derived based on the Euclidean distance [10], which was implemented in ArcMap 10.7 (ESRI, Redland, CA, USA) (see Table 2).

Demography has an impact on the distribution of diseases because it determines attributes such as population density, birth rates, and the total number of people in an area [39]. Here, population density, total population and population growth rates were considered as potential factors. Since Zambia has also been affected by the Human Immuno-deficiency Virus (HIV) / Acquired Immuno-Deficiency Syndrome (AIDS), the number of people with HIV/AIDS was also considered as one of underlying factors for the contraction of COVID-19 [4,11] (see Table 2).

Economic activities taking place in a location have an influence on the population structure and available facilities [2,35]. These facilities have both direct and indirect effects on people's health and regional health systems, which can be the factors influencing the spread of other diseases. Generally, diseases like cholera spread fast in districts with high economic activity due to the high population densities, especially when the sanitation conditions are poor [25,37]. Here the selected economic factors include the status of a district (urban or rural), crop yield and employment rates (Table 2).

Geospatial data processing

The geospatial data processing was done in ArcMap 10.7 (ESRI, Redland, California). The major processing included calculating Euclidean distance using the Spatial Analyst tool (Euclidean Distance) of features of interests including distance to roads, railway, towns and mines (Fig. 3). Other factors, such as population density, income, and crop yield were converted from shape to raster format. Topographic features, such as slope, elevation above sea level, and aspect were derived from the Shuttle Radar Topographic Mission (SRTM) digital elevation model (DEM). The DEM was downloaded from United States Geological Survey (USGS) website (http://glovis.usgs.gov). All the datasets were then resized using the country boundary for Zambia by employing the "clipping" tool in ArcMap and then projected to a common coordinated system World Geodetic System 84 (WGS 84) - Zone 35 south. To extract the data, we used both the zonal statistics to find the averages for districts and we also employed the central points of the districts by using a tool called "Extract values to points" under Spatial analyst.

Statistical modelling - classification tree

There are many different approaches for modelling disease incidences, and these methods are commonly based on a combination of geospatial and statistical techniques [,30,31]. Phiri et al. [35] indicated that the most common modelling techniques include regression analysis, logistic regression with decision or classification trees (CT). CT models, which employ a recursive partitioning principle, have an advantage over other methods because they are non-parametric and can handle both numerical and categorical data [15,30]. As an initial step during data analysis, the all the variables were check for their correlation with other explanatory variables and the independent variables (Supplementary Material Figure 1S and 2S).

Table 2
The description of socioeconomic and environmental factors used in this study with their spatial and temporal resolutions, and sources.

Category	Factors and units	Range	Spatial resolution	Temporal resolution	Sources
Topographic	Elevation (m)	325–2296	30 m	_	USGS
	Slope (°)	0–57.42	30 m	_	USGS
	Aspect (°)	-1–359.90	30 m	_	USGS
Climatic	Total annual precipitation (mm)	590–1503	1 km	1970–2000	WorldClim
	Solar radiance (w m^{-2})	15763–20511	1 km	1970–2000	WorldClim
	Maximum temperature (°C)	19.78–33.63	1 km	1970–2000	WorldClim
	Minimum temperature (°C)	8.60–21.57	1 km	1970–2000	WorldClim
	Mean temperature (°C)	14.30–26.30	1 km	1970–2000	WorldClim
Social	District status (urban, rural)	-	District level		Central Statistics Office of Zambia (CSO)
	Total population (count)	25, 294–1,701,640	District level	1969–2010	CSO
	Human immunodeficiency virus (HIV) rate		District level	2010	CSO
	Population density (persons km^{-2})	2.70–4,841.60	District level	2010	CSO
	Crop yield (tonnes)	451.93–67,600.78	District level	1990–2010	CSO
	Population change (count)	17,369–1,359, 354	District level	1990–2010	CSO
Proximity	Euclidean distance to active mine centres (km)	0–280.35	30 m	_	Ministry of Mines for Zambia
	Euclidean distance to waterbody edges (km)	0–108.62	30 m	_	Forest Department (FD)
	Euclidean distance to town centres (km)	0–82.56	30 m	_	FD
Accessibility	Euclidean distance to road (km)	0–104.25	30 m	_	Road Development Agency of Zambia (RDA)
	Euclidean distance to railway (km)	0–108.67	30 m	_	FD
	Euclidean distance to rivers (km)	0–128.62	30 m	_	FD
	Euclidean distance to border towns (km)	0–355	30 m	-	Ministry of Tourism
	Distance to airports (km)	0–305	30 m		Ministry of Tourism

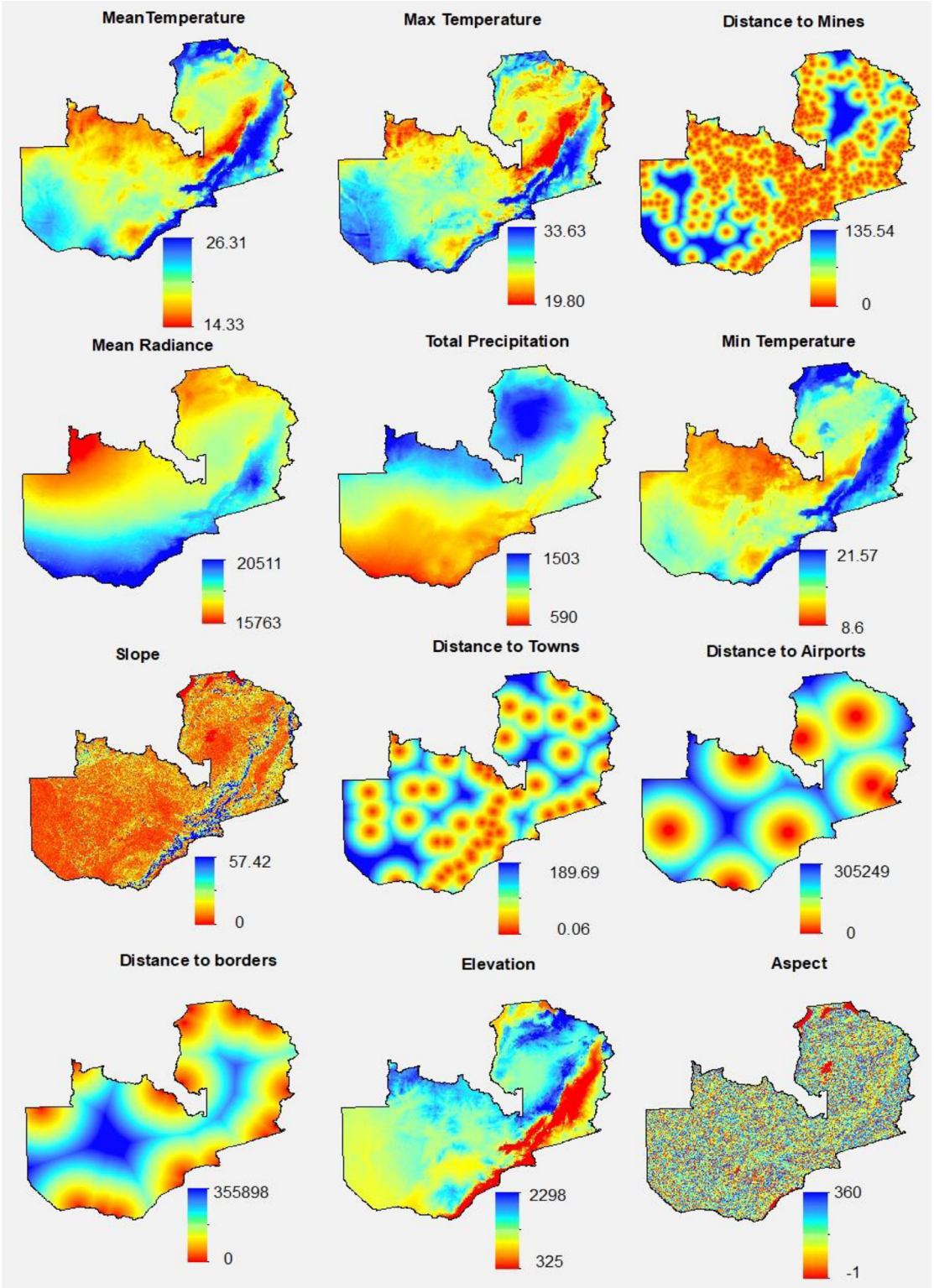

Fig. 3. Graphic representation of selected factors used in this study. Note that distance to the airport, borders, towns and mines were calculated using the Euclidean distance approach.

Table 3
Confusion matrices for accuracy assessment of binary model (COVID-19/NO COVID-19)[1].

	Reference			
Prediction	COVID-19	No COVID-19	Total	UA (%)
COVID 19	15	0	15	100
No COVID-19	2	5	7	71
Total	17	5	22	
PA (%)	88	100	OA = 91%	

Table 4
Confusion matrices for accuracy assessment of the four risk levels[2].

	Reference					
Prediction	High	Moderate	Low	Very low	Total	UA (%)
High	3	1	0	0	4	75
Moderate	0	6	1	0	7	86
Low	1	1	8	0	10	80
Very low	0	0	0	1	1	100
Total	4	8	9	1	22	
PA (%)	75	75	89	100	OA = 82%	

The decision tree was implemented in R statistical software environment [38] using the "rpart" package [44], and the graphic outputs were produced using the "rpart.plot" package [28]. To minimise the production of large decision trees (over-fitting), pruning was applied by setting the complex parameter (cp) to a minimum cross-validation error.

Validation

The whole dataset was divided into two: (1) training, and (2) validation datasets, using a 70:30 ratio [15,35] for both the binary and the risk level models. To assess the accuracy of the CT models, a cross-validation approach was used. Measures of accuracy, such as overall accuracy, user's and producer's accuracy were used to assess the model accuracies [7].

Results

In this study, two classification tree models were developed. The first one had a binary approach; the response variables were the presence of COVID-19 or its absence (No COVID-19). In the second model, four risk levels were established: high, moderate, low, and very low.

Binary model: Covid-19 / No Covid-19 cases

Fig. 3 presents the binary CT model, which focused on understanding the factors associated with districts where there is COVID-19 or No COVID-19. An assessment of the accuracy of this model showed that the CT model predicted the presence and absence of COVID-19 cases with an overall accuracy of 90.6%, together with the user's and producer's accuracy ranging from 70.8% to 100% (Table 3).

Of the seven terminal nodes, four predicted the presence of COVID-19 and three predicted the absence of the disease. The CT models showed that, as of 17 July 2020, 71% of the districts were likely to have no COVID-19, compared to 29% that had COVID-19 cases. The CT model had six factors, which included population density, total precipitation, distance to the airport, distance to the mines, distance to town and distance to the border.

Of all these factors included in the analysis, distance to the airport was identified as the most influential factor. If a district is within 47 km of the airport, it is likely to report a case of COVID-19, compared to those districts that were beyond this threshold. Most of the districts that had no COVID-19 are located far from the airports (>47 km), have a low population density (<44 people/km^2), are far from the mines (>41 km), and have a total precipitation of over 677 mm per annum.

Factors associated with different risk levels

Fig. 4 shows the CT model for the four COVID-19 risk levels. The model showed an overall accuracy of 81.5%, with a user's and producer's accuracy ranging from 74.9 % to 100% (Table 4). The low accuracy indicated a high percentage of errors from these classes. The CT model here had nine terminal nodes: very low = 1 node, low = 4 nodes, with moderate and high accuracy both having two nodes each. Of all the districts that had reported having COVID-19, 46% were classified as low risk, while 14% had high risk. The CT model had six factors, which included HIV cases, population density, mean temperature, distance to the mines, towns, and border posts.

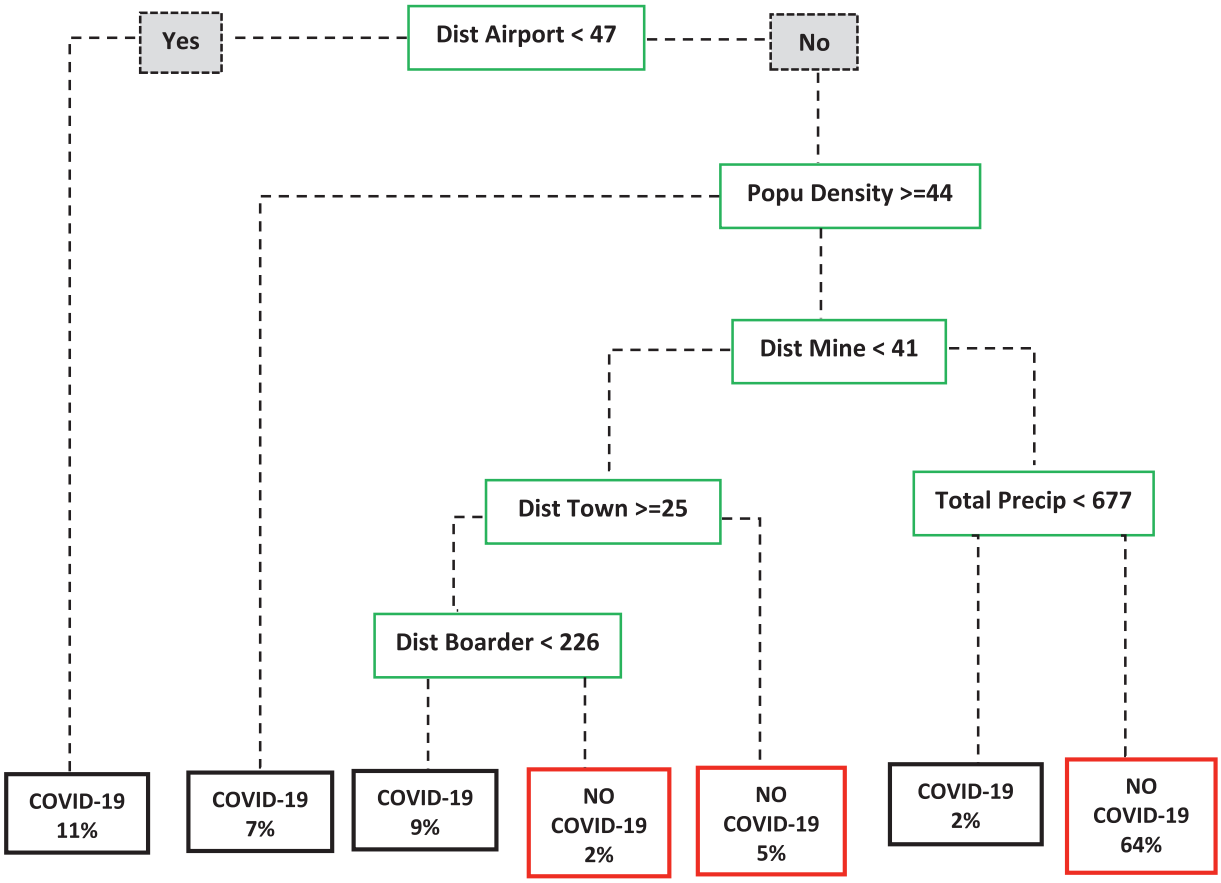

Fig. 4. The CT model showing the relationship between COVID-19 cases and the environmental and socioeconomic factors. The dark grey shade represents the districts where COVID-19 was reported, while the light grey shade represents the districts without COVID-19. Dist Airport refers to distance to the nearest airport, Popu Density refers to population density, Dist Mine refers to distance to the mines, Dist Town is distance to the nearest town, Total Precip is total precipitation, while Dist Boarder refers to distance to the boarder.

The number of HIV cases per district was the most influential factor, followed by distance to the mines and towns. The high number of COVID-19 cases corresponded with high numbers of HIV cases, while very low numbers of COVID-19 cases occurred in districts with low numbers of HIV cases, in locations far from the mines, and in areas having high mean temperatures (>22 °C). The CT model also showed that if an area is within 57 km of a border post, it has a high chance of having COVD-19 cases.

Discussion

The findings from this study show that the factors influencing the presence or absence of the disease and the risk levels of the disease are different. All the models had high accuracy; however, the models that were assessing the risk levels had relatively lower accuracy. This low accuracy could have been because the districts with high and very numbers of COVID-19 cases were fewer compared to those in the moderate class; hence the percentage errors from these classes were higher. The binary CT model, which focuses on assessing the factors associated with the presence or absence of COVID-19, showed that the major factors influencing the chances of contracting COVID-19 in a given district, in order of importance, included distance to the airport, population density, and distance to the towns. On the other hand, the CT model for the risk levels of the diseases showed that districts with higher numbers of HIV-positive cases had a high chance of having high levels of COVID-19 cases. Besides, other factors that had more influence in determining the disease risk levels include distance to the mines, towns and borders.

This study clearly indicated that COVID-19 cases were high in urban districts compared to rural districts in Zambia. This may be due to the association of factors influencing the occurrence and risk levels with urban areas. For example, the cases were high in areas close to airports, those with high population densities, and those that are close to urban centres. These findings are supported by Rocklöv and Sjödin [39], who indicated that areas with high population density are a catalyst for the spread of COVID-19. Generally, the findings from this study agree with other studies [21,29,39] that socioeconomic

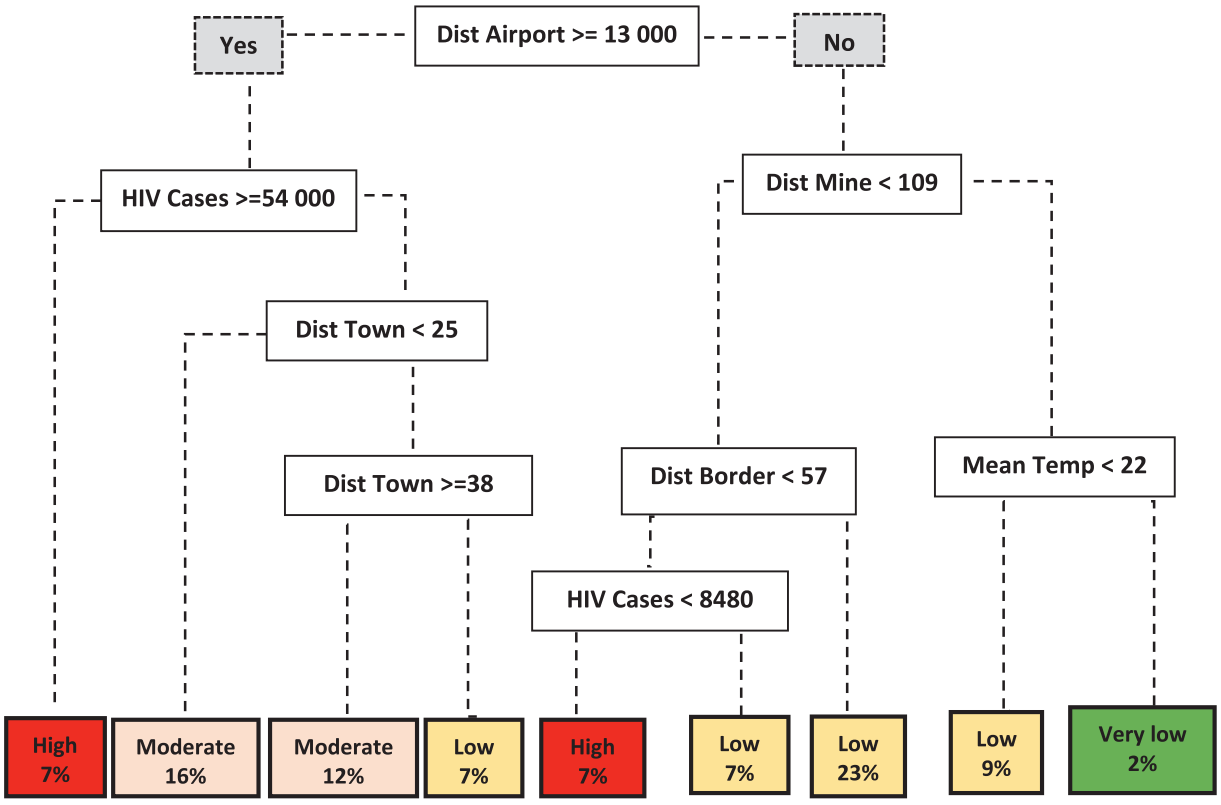

Fig. 5. The CT model showing the relationship between four COVID-19 risk levels and the environmental and socioeconomic factors. Where Dist Airport refers to distance to the nearest airport, HIV cases refers to the number of people living with HIV, Dist Mine refers to distance to the mines, Dist Town refers to distance to the nearest towns, Dist Border is distance to the border, while Mean Temp is Mean Temperature.

factors, which include proximity to towns, airports and borders, population density, and HIV infections, have more influence on the occurrence and spread of the disease than environmental factors.

Temperature was one of the major factors that determined whether an area will have a low number of cases of COVID-19 or not. Our finding was that districts with a temperature above 22 °C had very low numbers of cases of COVID-19. Since temperature was highly correlated to other climatic factors such as solar irradiance and rainfall; this was an indication that even humidity would have similar effects even when it was not included as one of the explanatory factors due to limited access to datasets. This indicates that areas with low temperatures are likely to experience increasing numbers of COVID-19 cases [13,16,27]. These findings are similar to Bashir et al. [3], who tested the correlation of COVID-19 cases with climatic factors (i.e., temperature and precipitation) and reported a strong correlation between the two in New York, USA. In our study, rural districts with high precipitation (> 677 mm/annum) had less chances of promoting the spread of COVID-19 cases, perhaps because these areas have low population density and experience high temperatures (> 22 °C) during the rainy season. Another possible reason would be that people are usually housebound during the rainy season, especially in areas with high rainfall, where there is limited accessibility.

This study makes two significant contributions to the understanding of the factors relating to the spread of COVID-19, which include: (1) the approach used, which combines a geospatial approach with statistical methods (See Morgenroth et al. [30]; Guo et al. [15]; Phiri et al. [35]), and (2) the results, which provide empirical evidence on the factors associated with the spread of COVID-19. The combined geo-statistical approach used in this study is simple and easy to interpret, especially the CT models, which produce clear decision tree graphics. This approach can be used with different datasets because the results can be easily replicated. The results from this study present empirical evidence, and hence reduce the increasing speculation surrounding the factors relating to the spread of COVID-19 cases in Zambia and other countries in sub-Saharan Africa, as well as in developing countries around the world.

The findings from this study need to be interpreted by taking into consideration the limitations of the study. Firstly, the data used was accessed at the district level, and hence some of the details might not have been captured because they need small mapping units. Furthermore, it was not possible for the data to include human behavior attributes, age group and socioeconomic situation because of limited access to detailed information on COVID-19 patients. Secondly, the numbers of COVID-19 cases have continued to rise, especially during the cold season (July-August), and this is likely to affect the patterns and the distribution of the COVID-19 cases. As such, these results might not be replicated, or might vary if datasets

for later dates are used, yet they remain relevant to use in controlling the surge of COVID-19 cases in similar situations. Thirdly, the factors considered in this study do not represent all the potential factors that can be considered. Finally, due to the limitation in testing facilities, and to the logistics in fighting the COVID-19 pandemic, all the districts did not have the same testing facilities and opportunities. Testing was mainly focused on contact tracing from the outbreaks in the districts. The challenges in accessing most of the factors posed a limitation on our study, and hence we suggest that future studies should examine other factors beyond those that have been presented here. As another avenue for further studies, a detailed understanding is needed of the relationship between these factors and other factors, such as different age groups in different seasons. We also recommend a detailed time-series study based on normalised data to understand the pattern between daily COVID-19 cases and daily climatic factors.

Conclusions

This study aimed at assessing the relationship between the distribution of COVID-19 cases and socioeconomic as well as environmental factors in Zambia. Our findings indicated that the distribution of COVID-19 cases at the district level was mainly associated with socioeconomic factors, such as population density, HIV rates and proximity to airports, country borders and towns. Although environmental factors, such as temperature and precipitation did not appear as the most influential factors, high temperature, as a factor, was one of the factors which was associated with areas with low cases of COVID-19 in Zambia. The findings from this study may be used as a preliminary guide in understanding the distribution of COVID-19 in sub-Saharan Africa as well as in other developing countries in the world. Further studies could focus on the relationships of these factors in different seasons, within different age groups and cultural settings.

Fig. 5

Declaration of Competing Interest

The authors declare that they have no financial or non-financial competing interests.

Funding sources

This research was funded by Japan Society for the Promotion of Science under Grant-in-Aid for Scientific Research - JSPS grant 21K01027.

References

[1] Anjorin, A. A. J. A. P. J. o. T. M. (2020). The coronavirus disease 2019 (COVID-19) pandemic: a review and an update on cases in Africa. 13(5), 199.
[2] P.-A. Balland, C. Jara-Figueroa, S.G. Petralia, M.P. Steijn, D.L. Rigby, C.A. Hidalgo, Complex Econ. Activ. Concent. Large Cities 4 (3) (2020) 248–254.
[3] M.F. Bashir, B. Ma, Komal Bilal, Bashir B., A. M., D. Tan, M Bashir, Correlation between climate indicators and COVID-19 pandemic in New York, USA, Sci. Total Environ. 728 (2020) 138835, doi:10.1016/j.scitotenv.2020.138835.
[4] J.L. Blanco, J. Ambrosioni, F. Garcia, E. Martínez, A. Soriano, J. Mallolas, J.H. Miro, COVID-19 in Patients with HIV: Clin. Case Ser. 7 (5) (2020) e314–e316.
[5] J. Chamberlin, T. Jayne, N.J.J.A.E. Sitko, Rural in-Migr. Agricult. Dev.: Evid. Zambia 51 (4) (2020) 491–504.
[6] Cong, R.-G., & Brady, M. J. T. S. W. J. (2012). The interdependence between rainfall and temperature: copula analyses. *2012*.
[7] R.G. Congalton, K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (Vol. 2nd), CRC Press/Taylor & Francis, Boca Raton, 2009.
[8] N.C. Coops, R.H. Waring, T.A. Schroeder, Combining a generic process-based productivity model and a statistical classification method to predict the presence and absence of tree species in the Pacific Northwest, U.S.A, Ecol. Modell. 220 (15) (2009) 1787–1796 http://dx.doi.org/, doi:10.1016/j.ecolmodel.2009.04.029.
[9] CSOZambia 2010 Census of Population and Housing, GRZ, Lusaka, Zambia, 2010 Retrieved from Lusaka, Zambia.
[10] P.-E. Danielsson, Euclidean distance mapping, Comput. Graph. Image process. 14 (3) (1980) 227–248.
[11] Del Amo, J., Polo, R., Moreno, S., Díaz, A., Martínez, E., Arribas, J. R., ... Hernán, M. A. J. A. o. i. m. (2020). Incidence and severity of COVID-19 in HIV-positive persons receiving antiretroviral therapy: a cohort study.
[12] S.E. Fick, R.J. Hijmans, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int. J. Climatol. 37 (12) (2017) 4302–4315, doi:10.1002/joc.5086.
[13] Guasp, M., Laredo, C., & Urra, X. J. C. I. D. (2020). Higher solar irradiance is associated with a lower incidence of COVID-19.
[14] Guner, H., Rahmet, Hasanoğlu, I., & Aktaş, F. J. T. J. o. m. s. (2020). COVID-19: Prevention and control measures in community. 50(SI-1), 571–577.
[15] T. Guo, J. Morgenroth, T. Conway, Redeveloping the urban forest: the effect of redevelopment and property-scale variables on tree removal and retention, Urban Forest. Urban Green. 35 (2018) 192–201.
[16] Z. Huang, J. Huang, Q. Gu, P. Du, H. Liang, Q. Dong, Optimal temperature zone for the dispersal of COVID-19, Sci. Total Environ. 736 (2020) 139487, doi:10.1016/j.scitotenv.2020.139487.
[17] Jumbam, D. T., Stevenson, J. C., Matoba, J., Grieco, J. P., Ahern, L. N., Hamainza, B., ... Munachoonga, P. J. B. p. h. (2020). Knowledge, attitudes and practices assessment of malaria interventions in rural Zambia. 20(1), 216.
[18] Kalinda, C., Chimbari, M. J., Grant, W. E., Wang, H.-H., Odhiambo, J. N., & Mukaratirwa, S. J. P. n. t. d. (2018). Simulation of population dynamics of Bulinus globosus: Effects of environmental temperature on production of Schistosoma haematobium cercariae. 12(8), e0006651.
[19] J. Kleemann, G. Baysal, H.N.N. Bulley, C. Fürst, Assessing driving forces of land use and land cover change by a mixed-method approach in north-eastern Ghana, West Africa, J. Environ. Manage. 196 (2017) 411–442, doi:10.1016/j.jenvman.2017.01.053.
[20] Kontis, V., Bennett, J. E., Rashid, T., Parks, R. M., Pearson-Stuttard, J., Guillot, M., ... Corsetti, G. J. N. m. (2020). Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. 1–10.
[21] M.U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D.M. Pigott, ..., W.P.J.S. Hanage, The effect of human mobility and control measures on the COVID-19 epidemic in China 368 (6490) (2020) 493–497.
[22] R. Kumar, S. Nandy, R. Agarwal, S.P.S. Kushwaha, Forest cover dynamics analysis and prediction modeling using logistic regression model, Ecol. Indic. 45 (2014) 444–455, doi:10.1016/j.ecolind.2014.05.003.

[23] Lombe, D., Phiri, M., & Msadabwe, S. J. e. (2020). Negative impact of the COVID-19 pandemic on the management of cervical cancer patients in Zambia. 14.
[24] Lone, S. A., Ahmad, A. J. E. m., & infections. (2020). COVID-19 pandemic–an African perspective. 9(1), 1300–1308.
[25] M.E. Lucas, M. Jeuland, J. Deen, N. Lazaro, M. MacMahon, A. Nyamete, ..., F.F. Songane, Private demand for cholera vaccines in Beira, Mozambique, Vaccine 25 (14) (2007) 2599–2609.
[26] Maeda, J. M., & Nkengasong, J. N. J. S. (2021). The puzzle of the COVID-19 pandemic in Africa. 371(6524), 27–28.
[27] M.M. Menebo, Temperature and precipitation associate with Covid-19 new daily cases: a correlation study between weather and Covid-19 pandemic in Oslo, Norway, Sci. Total Environ. 737 (2020) 139659, doi:10.1016/j.scitotenv.2020.139659.
[28] S. Milborrow, rpart. plot: Plot rpart Models. an enhanced version of plot. rpart, R package version 1 (3) (2015).
[29] Mollalo, A., Vahedi, B., & Rivera, K. M. J. S. o. T. T. E. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. 138884.
[30] J. Morgenroth, J. O'Neil-Dunne, L.A. Apiolaza, Redevelopment and the urban forest: a study of tree removal and retention during demolition activities, Appl. Geogr. 82 (2017) 1–10, doi:10.1016/j.apgeog.2017.02.011.
[31] Mulenga, E. M. J. A. (2020). Spread of COVID-19 pandemic in Zambia: a mathematical model. 4(2), ep20019.
[32] Mwaba, J., Debes, A. K., Shea, P., Mukonka, V., Chewe, O., Chisenga, C., ... Chilengi, R. J. P. n. t. d. (2020). Identification of cholera hotspots in Zambia: a spatiotemporal analysis of cholera data from 2008 to 2017. 14(4), e0008227.
[33] Neubert, S., Kömm, M., Krumsiek, A., Schulte, A., & Tatge, N. (2011). *Agricultural development in a changing climate in Zambia: increasing resilience to climate change and economic shocks in crop production*: Studies.
[34] D. Phiri, J. Morgenroth, C. Xu, Four decades of land cover and forest connectivity study in Zambia—An object-based image analysis approach, Int. J. Appl. Earth Obs. Geoinf. 79 (2019) 97–109, doi:10.1016/j.jag.2019.03.001.
[35] D. Phiri, J. Morgenroth, C. Xu, Long-term land cover change in Zambia: an assessment of driving factors, Sci. Total Environ. (2019) 134206, doi:10.1016/j.scitotenv.2019.134206.
[36] D. Phiri, E. Phiri, R. Kasubika, D. Zulu, C. Lwali, The implication of using a fixed form factor in areas under different rainfall and soil conditions for Pinus kesiya in Zambia, South. Forests: J. Forest Sci. 78 (1) (2016) 35–39, doi:10.2989/20702620.2015.1108614.
[37] D. Phiri, M. Simwanda, V. Nyirenda, Mapping the impacts of cyclone Idai in Mozambique Using Sentinel-2 and OBIA approach, South Afr. J. Geogr. (2020), doi:10.1080/03736245.2020.1740104.
[38] R Core TeamR: A Language and Environment for Statistical Computing, R Foundation forStatistical Computing, Vienna, Austria, 2017 URL https://www.R-project.org/ .
[39] Rocklöv, J., & Sjödin, H. J. J. o. t. m. (2020). High population densities catalyse the spread of COVID-19. 27(3), taaa038.
[40] Rosa, W. E., Gray, T. F., Chow, K., Davidson, P. M., Dionne-Odom, J. N., Karanja, V., ... Nursing, P. (2020). Recommendations to leverage the palliative nursing role during COVID-19 and future public health crises. 22(4), 260–269.
[41] B. Shu, H. Zhang, Y. Li, Y. Qu, L. Chen, Spatiotemporal variation analysis of driving forces of urban land spatial expansion using logistic regression: a case study of port towns in Taicang City, China, Habit. Int. 43 (2014) 181–190.
[42] Sintema, E. J. J. E. J. o. M., Science, & Education, T. (2020). Effect of COVID-19 on the performance of grade 12 students: Implications for STEM education. 16(7), em1851.
[43] Spiteri, G., Fielding, J., Diercke, M., Campese, C., Enouf, V., Gaymard, A., ... Riutort, A. N. J. E. (2020). First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. 25(9), 2000178.
[44] Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines. In: Technical Report 61. URL http://www.mayo.edu/hsr/techrpt/61.pdf.
[45] Tian, H., Liu, Y., Li, Y., Wu, C.-H., Chen, B., Kraemer, M. U., ... Yang, Q. J. S. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. 368(6491), 638–642.
[46] WHO. (2021). Coronavirus (COVID-19) *Dashboard: Global Situation*. https://covid19.who.int/.
[47] Wu, J., Mamas, M. A., Mohamed, M. O., Kwok, C. S., Roebuck, C., Humberstone, B., ... Gale, C. P. (2021). Place and causes of acute cardiovascular mortality during the COVID-19 pandemic. 107(2), 113–119. doi:10.1136/heartjnl-2020-317912%JHeart