American Journal of Theoretical and Applied Statistics
2023; 12(3): 37-42
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20231203.11
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

SciencePG
Science Publishing Group

Supervised Machine Learning and Bayesian Regression Kriging with Application to COVID-19 Incidences in Sub-Saharan Africa

Safari Godfrey Lyece[*]**, Samuel Mwalili, Joseph Kyalo Mung'atu**

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya

Email address:

godfreylyece738@gmail.com (Safari Godfrey Lyece), samuel.mwalili@gmail.com (Samuel Mwalili),
kmungatu@gmail.com (Joseph Kyalo Mung'atu)

[*]Corresponding author

To cite this article:

Safari Godfrey Lyece, Samuel Mwalili, Joseph Kyalo Mung'atu. Supervised Machine Learning and Bayesian Regression Kriging with Application to COVID-19 Incidences in SUB-Saharan Africa. *American Journal of Theoretical and Applied Statistics*. Vol. 12, No. 3, 2023, pp. 37-42. doi: 10.11648/j.ajtas.20231203.11

Received: April 19, 2023; **Accepted**: May 15, 2023; **Published**: May 24, 2023

Abstract: Since COVID-19 invasion of the World, human life has been affected greatly. Several studies have shown a positive correlation between COVID-19 infections and pre-existing conditions such as Diabetes, Cancer, Tuberculosis, and Hypertension. In this study, we would like to determine whether demographic variables have a contribution to the spread of COVID-19 infections. We will apply a machine language method to select the demographic variables which are impactful in the spread of COVID-19 cases in Sub-Saharan Africa. Then we shall determine the nature of COVID-19 cases patterns applying the K-Nearest Neighbor (KNN) in calculating the neighborhood weights between locations/countries. The weights would then be tested for significance to conclude whether the cases patterns are either random, sparsely or clustered. We would then perform simulations to estimate the social demographic/covariates/fixed effects parameters and the random effects parameters. The Bayesian Kriging would be applied to predict Covid-19 cases based on the estimated social demographical variables coefficients/parameters and the random effects parameters in unknown/new locations in Sub Saharan Africa with a known uncertainty. The results showed that Children aged (0-14) years living with HIV AIDS, Prevalence of HIV Total (percentage of population ages 15-49) and Access to electricity (as a percentage of the population) was estimated to contribute to the increase of COVID-19 cases. Prediction of the COVID-19 cases in unknown locations showed that most of the cases were predicted in the elevated locations/areas than in the lower/flatter locations. This could mean that high elevated areas are associated with lower temperatures which increases the spread of COVID-19 cases as opposed to lower/flatter areas which are associated with higher temperatures which reduces the spread of COVID-19 cases.

Keywords: K-Nearest Neighbor, Bayesian Kriging, COVID-19

1. Introduction

In December 2019 a novel virus known as Corona Virus was discovered in Wuhan in China. It is abbreviated as COVID-19 signifying the short name. This virus caused severe respiratory diseases including Pneumonia. The virus causing the infection has been referred to as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) by the World Health Organization (WHO). This infection is capable of spreading from one human being to the other which appears different from SARS-COV which is said to have originated from a bat and is not able to spread to human beings. Due to its fast spread, by April of 2020 the virus had already spread to all parts of the world as a result of human interaction. Due to its novelty, medical experts could not figure out its structure and component hence unable to explain its epidemiological behavior and which parts of the body it attacked most. This resulted to so many people contracting the virus and loss of lives including the medical workers. World Health Organization therefore, called for worldwide collaboration on February 24th 2020 after declaring the Virus a pandemic to have countries invest in research to learn the COVID-19 transmission and the mitigation measures before developing a treatment. According to John Hopkins University Corona Virus Resource Centre, the current number of Corona Virus Infection cases around the world stands at 608,709,777 with a total of 6,514,732 deaths as at September 12th 2022 with a lot cases and fatalities reported in the United States of America and Europe. Ongoing research has demonstrated that the virus mostly attacks the lungs resulting to difficulty in breathing and other respiratory problems. It has also been revealed that COVID-19 affection continues to affect people with pre-existing conditions such as cancer, diabetes, kidney failure, Tuberculosis, and Hypertension although it is not known to what extent. Because of the high number of cases of COVID-19 and deaths it is important to relate these to some socio-demographic variables to find out whether such variables have any statistical association with COVID-19 cases particularly in the Sub-Saharan Africa. We will also look at the spatial pattern spatial temporal of the COVID-19 cases across the Sub-Saharan Africa from April 1st 2020 to December 31st 2021. These patterns would enable us discover the COVID-19 spread patterns with time. We will apply Bayesian Kriging Regression method to predict COVID-19 cases with a given uncertainty particularly in locations/regions that COVID-19 cases have not been observed using the available COVID-19 cases.

1.1. Theoretical Background

Since COVID-19 invasion of the World, human living styles have been affected greatly. This involves wearing of masks, keeping of social distance and regular sanitizing and washing of hands to help reduce the spread of COVID-19 virus. However, in as much as there is evidence to show that pre-existing conditions such as Diabetes, Cancer, Tuberculosis, kidney failure and Hypertension increases the chances of COVID-19 infection and even death, other socio-demographic variables that are of importance have not been analyzed to find out if they can be linked to COVID-19 infections and deaths particularly in the Sub-Saharan African region. The data under consideration has a higher number of features some of which are collinear. So, considering these exploratory variables which greatly affect people in the Sub-Saharan African Region have not been analyzed to find out their strength of contribution toward COVID-19 infections. This would not only help in increasing the mitigating measures in curbing the spread of COVID-19 but also developing policies that would ensure that the spread of other diseases is also controlled through the construction of more health facilities. This study would utilize the LASSO machine language algorithms to select the impactful social-demographic variables that contribute to the spread of COVID-19 and then determine the COVID-19 patterns applying the spatial autocorrelation Index (Global Moran's 1) in the Sub-Saharan African countries. Finally, the study would use the estimated posterior distributions to predict the COVID-19 Cases in the unobserved locations in the Sub-Saharan Regions/locations using the built Bayesian Regression Kriging model.

1.2. Literature Review

Since the outbreak of COVID-19, several studies have been carried out ranging from the gene description of the virus, transmission, mutation, to association of the virus and various pre-existing conditions. Several researchers have applied correlation and regression methods to establish an association between COVID-19 infection cases and a number of social-economic variables. For example, (Cambaza and Viegas) [1] studied the correlations between Gross Domestic Product (GDP), the number of COVID-19 tests, and the number of confirmed COVID-19 cases in just thirteen African countries. Lin et al. [2] used data on the number of confirmed COVID-19 cases from thirty-nine well developed cities of China and modelled the effects of several socio-economic indicators. However, these studies failed to account for the possibility of spatial dependency in the COVID-19 prevalence. Therefore, to understand the impact of neighboring and the presumed influencing factors on COVID-19 prevalence in Africa, spatial statistics analysis would be very fundamental according to Fatima et al. [3]. Fatima et al. [3] observed that most studies that used spatial statistics analysis on COVID-19 prevalence have been primarily been carried out in Brazil, China, and USA. According to my knowledge this is the first study in Sub-Saharan Africa that does not only look at the association between the spatial dependency of COVID-19 cases and most impactful social demographic variables but also apply some improved geostatistical methods to predict the COVID-19 cases in the unobserved locations/regions with a specific uncertainty. Rahman et al. [4], acknowledged the importance of interpolation from individual observations to a larger geographical area for mapping purposes is a paramount importance for geotechnical engineers decision-making especially in data scarce regions. Therefore, to make the most of available data to inform the acquisition of new data in a cost-effective way requires new algorithms for analysis. Thompson et al. [5] noted that Kriging methods have been used to combine several geotechnical data sources to obtain accurate and reliable shear wave velocity in the uppermost 30m of soil V_{s30} maps, cite implification factor maps according to Thompson et al. [6]. Other necessary geotechnical parameters according to Marache et al. [7], or liquefication potential maps as reported by Pokhrel et al. [8]. As noted by Pilz and Spöck [9] in their literature, kriging interpolation in combination with the Bayesian approach is known as Bayesian Kriging as earlier noted by Omre [10] and Omre and Halvorsen [11] which is particularly used to study a region/location variation according to Chakraborty and Goto [12]. De Risi et al. [13] applied the Bayesian Kriging approach to improve the knowledge about the spatial Variation of the shear wave velocity in the uppermost 30m V_{s30} for the Kathmandu valley Cui et al. [14]. Unlike other forms of kriging which require the data to be stationary before prediction can be performed to reduce the prediction errors according to Machuca et al. [15] Bayesian Kriging does not require the data to be stationary. Respecting the traditional kriging approach, this approach provided a mechanism of quantifying and propagating the uncertainties on the parameters that define the spatial modeling of the shear wave velocity in the uppermost 30m V_{s30}. This approach would be applied in the health para dim in quantifying the uncertainties associated with the prediction of COVID-19 cases of unobserved locations/regions in Sub-Saharan Africa.

2. Methods

The social demographic variables we considered in this study to assess their contribution toward COVID-19 infections in Sub-Saharan Africa are: HIV prevalence among people aged 15-49 years, infants who received their-dose of pneumococcal conjugate-based vaccine, Incidence of TB (per 100,000 people), HIV incidence rate among children aged 0-14 years, Incidence of HIV among individuals aged 15-24 years, Smoking prevalence (15 years and above), HIV incidence rate among adolescence aged 10-19 years, Diarrhea treatment (children under-5 years), Incidence of Malaria (per 1000 persons at risk), Percentage of population with access to basic drinking water, GDP per capita, Percentage of population with access to electricity, and Life expectancy at birth.

We selected the features which were associated with the spread of COVID-19 cases in Sub-Saharan African using the LASSO algorithm.

$$J(\beta_0, \beta_1, \beta_2, \dots, \beta_k) = \frac{1}{2M} \sum \left(\left(\beta_0 + \beta_1 X_{obs}^{i} \right) - Y_{obs} \right)^2 + \lambda \sum_{j=1}^{K} \left| \beta_j \right|$$

The selected features in relationship to COVID-19 are given in Table 1 ranked according to the strength of association.

The nature of COVID-19 pattern in Sub-Saharan African was determined applying the Global Moran I method. In this case, we formulated a hypothesis to determine whether or not there was no spatial correlation in the recorded COVID-19 cases. Table 2 shows the Global Moran I values along with their respective P-values. From the P-Values, we fail to reject the null hypothesis and conclude that the COVID-19 patterns are randomly distributed as shown in Figure 1.

In predicting the COVID-19 cases in unknown locations in Sub-Saharan African with a given uncertainty, we used Bayesian Regression Kriging. The Bayesian Regression Kriging (BRK) is a geostatistical Kriging model with a Gaussian assumption in the Bayesian Framework. Bayesian Framework considers model parameters as random variables. Therefore, (BRK) has the ability to quantify uncertainties that are associated with the model parameters. The BRK can be expressed as follows,

$$y = X^t \beta + W + e$$

Where y is the vector of COVID-19 cases while the X is the vector of the covariates/features. W is a vector of spatial random effects which follows a multivariate Gaussian distribution with zero-mean and covariance matrix Σ which shall be computed from the correlation function $C(s_1, s_2; \sigma^2, \phi) = \sigma^2 e^{-\phi|s_1 - s_2|}$. The error e is an Nx1 random error with zero mean and correlation matrix $\sum_w = \tau^2 I$ which is a diagonal matrix. The Bayesian Regression Kriging considers all its parameters as random variables and assigns prior distributions to them. We shall denote all the model parameters as $[\theta = \beta, \sigma^2, \phi, \tau^2]$. The β will represent the coefficients of the fixed effects. The random effects parameters are σ^2, ϕ and τ^2 representing the response spatial variation, correlation function between the spatial locations and variation in the measurement errors respectively. The BRK parameters were estimated through simulations. The estimates of the fixed effects coefficients are shown in Table 3 while the estimates of the random effects parameters are shown in Table 4.

Predicting the COVID-19 cases in new locations/unknown locations was performed using the estimated parameters. Figure 2, represents the image of prediction of the COVID-19 cases in a hundred new locations in Sub-Saharan Africa. Figure 3, represents the predicted cases in perspective. Figure 4, represents the contour lines which shows the nature of the predicted locations. Finally, Figure 5 shows the contour representation of the standard errors on the predicted cases in Sub-Saharan Africa.

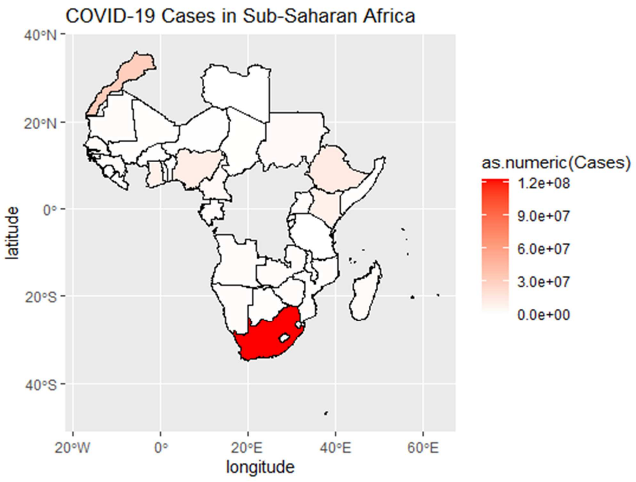

Figure 1. *COVID-19 Cases Patterns in countries in Sub-Saharan Region.*

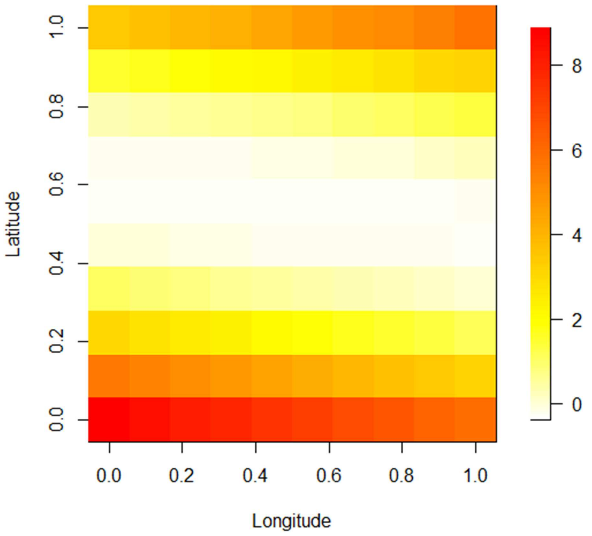

Figure 2. *The image representation of the predicted cases in Sub-Saharan Africa.*

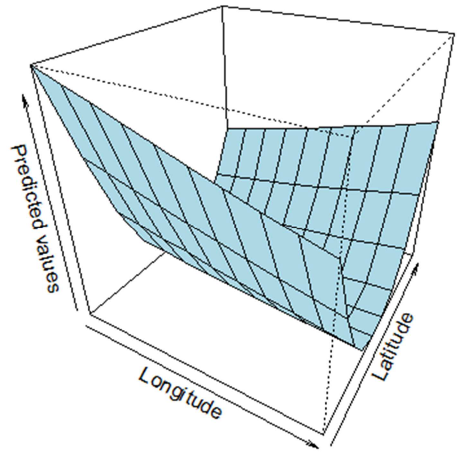

Figure 3. *The perspective representation of the predicted cases in Sub-Saharan Africa.*

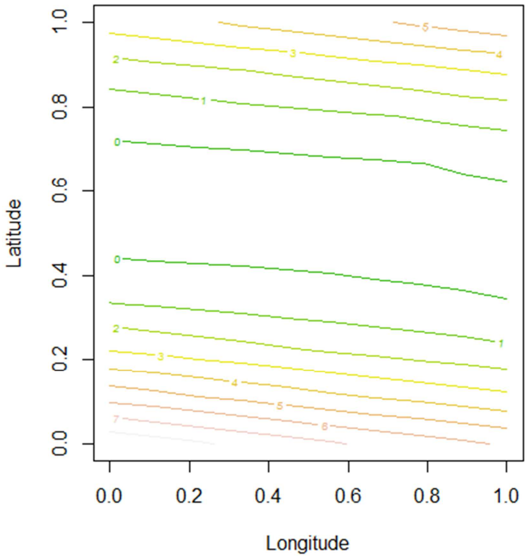

Figure 4. *The contour representation of the predicted cases in Sub-Saharan Africa.*

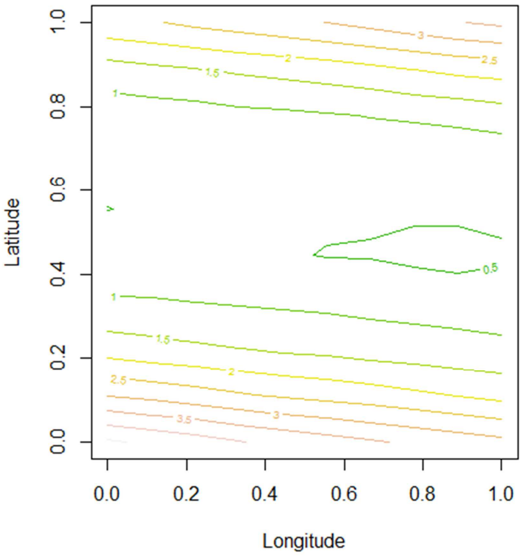

Figure 5. *The contour representation of the standard errors on the predicted cases in Sub-Saharan Africa.*

Table 1. *Lasso Regression Coefficients with an RMSE of 0.2035248.*

Name of Social Economic Variable	Coefficients
Intercept	-0.0686555
Prevalence of HIV total (percentage of population ages 15-49)	0.1006504
Children aged (0-14) years living with HIV AIDS	0.1291773
Incidence of HIV among Males Aged 15-24 Years	0
Incidence of Tuberculosis (per 100000 people)	0
Incidence of malaria (per 1000 population at risk)	0
People using at least basic drinking water services (percentage of population)	0
GDP per capita (current US Dollars)	0
Access to electricity (percentage of population)	0.2160687
Life expectancy at birth total (years)	0

Table 2. *The Moran I statistics with its associated K-Value and P-Value.*

K Values	The Maran I statistics	P values
3	-0.1260367	0.5501485
5	-0.5112407	0.6954087
7	-0.2163548	0.5856444
9	-0.2149674	0.5851036
11	-0.1743802	0.5692167
12	-0.1511768	0.5600819
13	-0.106398	0.5423667
14	-0.1413802	0.5562152
15	-0.1489331	0.5591968
17	-0.2852908	0.6122893

Table 3. *The posterior estimates of the covariate's coefficients.*

Betas	Mean	Median	Mode
β_0	-0.2149	-0.1450	-0.0153
β_1	-0.0472	-0.0410	-0.0409
β_2	-0.0790	-0.0712	-0.0655
β_3	0.00081	0.00075	0.00073
β_4	0.00346	0.00326	0.00299
β_5	0.002724	0.00262	0.00253

Table 4. *The posterior estimates of the random effect parameters.*

Parameters	Mean	Median	Mode
σ^2	4.1218	1.3834	0.8640
ϕ	87.2827	88.6344	3.5454
ν^2	0.4444	0.5000	0.0000

3. Discussion

This study applied LASSO regression method to identify a number of features/covariates which contribute to the spread of COVID-19 cases in countries in Sub-Saharan African region. It was found that three of the features put into consideration contributed to the spread of the virus. These social demographical variables are Access to electricity (percentage of population), Children aged (0-14) years living with HIV AIDS and Children aged (0-14) years living with HIV AIDS.

Access to electricity was found to be the most impactful in the spread of COVID-19 cases contributing to 20 percent of the cases. This could mean that access to electricity results to many economic activities taking place in an area/location such as manufacturing, construction etc. These economic activities would result to a frequent human interaction which will increase the spread of the virus. The analysis has also revealed that an increase in one unit of Children aged (0-14) years living with HIV-AIDS slightly contributed to the spread of COVID-19 cases accounting to 12.91 percent of the cases when holding the other two covariates constant. This could mean that children of this age interact with people including teachers and parents because they are school going children. Finally, the LASSO algorithm has also revealed that of the social demographic variables considered, an increase of one unit of prevalence of HIV total as a percentage of the population aged 15-49 had a least impact to the spread of COVID-19 cases accounting to 10.07 percent of the cases holding all the other two variables' constant.

From the analysis, it was also revealed that COVID-19 cases in Sub-Saharan African were randomly spread as opposed to being concentrated in certain parts or sparsely distributed. This shows that majority of the countries in Sub-Saharan Africa had their own reported cases based on the internal community spread/infections which was not affected by the interaction of communities outside each country.

Prediction of the cases in unknown/ new locations was performed based on the covariates applying the Bayesian Kriging. Under the Bayesian frame work we estimated the covariates parameters and the prediction of the cases was done in a hundred new locations. It was interesting to note that even movement of people in addition to the three social demographic variables also had an effect on the spread of cases. Movement towards the East either from Western or Eastern sides particularly towards the point 0.5°E by one unit in Sub-Saharan Africa reduces the COVID-19 cases by 4.7 percent while movement towards the Equator 0°N either from the Northern or Southern sides in Sub-Saharan Africa also reduces the cases by roughly 7.9 percent. Additionally, based on the Bayesian frame work where parameters are considered random variables, requires that the model parameters are assigned prior information which would be updated by the data information, showed a slightly difference in the magnitude of the three social demographic variables contribution to COVID-19 cases when factoring other effects not explained by the model known as the random effects. For example, a unit change in Children aged (0-14) years living with HIV AIDS, Prevalence of HIV Total (percentage of population ages 15-49) and Access to electricity (as a percentage of the population) was estimated to increase the COVID-19 cases by 0.08 percent, 0.35 percent and 0.28 percent respectively. This implies that when all factors are included in analyzing factors associated with the spread of COVID-19 cases such as measurement errors, variations in measurement errors and the distance between which the measurements are taken, will definitely give a more accurate analysis than without. With the accurate estimates of the parameters of the covariates, one can make the predictions of cases in the unknown locations with minimal standard errors.

Finally, the Bayesian kriging used the estimated covariates parameters to make predictions in the unknown locations in Sub-Saharan African Countries. The prediction showed that most of the cases were predicted in the elevated locations/areas than in the lower/flatter locations. This could mean that high elevated areas are associated with lower temperatures which increases the spread of COVID-19 cases as opposed to lower/flatter areas which are associated with higher temperatures which reduces the spread of COVID-19 cases.

4. Conclusion

In conclusion, the research has shown that a number of social demographic variables also have an impact on the spread of COVID-19 Cases in Sub-Saharan Africa such as Children aged (0-14) years living with HIV AIDS, Prevalence of HIV Total (percentage of population ages 15-49) and Access to electricity (as a percentage of the population) with a contribution of 12.91 percent, 10.09 percent and 20 percent respectively to the COVID-19 cases. Therefore, setting policy decision to manage these factors is fundamental to help reduce the spread of COVID-19 Cases in Sub-Saharan Africa. It is also evident that the direction of movement in Sub-Saharan Africa also has an impact in the spread of COVID-19 cases. For example, movement towards the East either from Western or Eastern sides particularly towards the point 0.5°E by one unit in Sub-Saharan Africa reduces the COVID-19 cases by 4.7 percent while movement towards the Equator 0°N either from the Northern or Southern sides in Sub-Saharan Africa also reduces the cases by roughly 7.9 percent. Therefore, more COVID-19 cases were predicted on the elevated/upper locations and less COVID-19 cases are predicted on the lower locations in Sub Saharan Africa. This implies that individuals residing in elevated locations in Sub-Saharan Africa are required to be more cautious and observe all COVID-19 protocol measures to reduce the spread of the virus. However, those in the lower locations where the prediction shows lower cases of COVID-19 Cases still need to be cautious and follow all the guidelines provided by the health experts so that the spread of the virus does not increase.

Based on this research, I recommend that we can use more sophiscated machine learning models to select impactful features/covariates with an accurate quantification believed to be related to some output/response variables. For this case, when it comes to social demographical variables selections, I would recommend Deep Learning models such as CNN or RNN which are believed to have higher interpretability and prediction power. When it comes to determine the nature of a disease spread in regions such as Sub-Saharan African, I would recommend other alternatives such as considering the use of Local Moran I to measure how similar locations are to their neighbors. Optimal Posterior distribution estimation of model parameters can always be done through simulations. With super computers one can run large simulations of even to a million using a number of Markov Chain Monte Carlo packages such as STAN. Finally, when it comes to prediction of cases in Sub-Saharan Africa Countries based on a number of factors, one can use the co-kriging approach to not only get prediction of one output (Cases) but also other outputs such as vaccinates administration or nature of the hospital facilities.

Funding

This study did not have any funding from any source. All the costs were incurred by the corresponding author.

Acknowledgements

I would like to thank Mr. Gregory Michael Mbishi for his great advice and insightful encouragement to undertake this research which has produced great results which when considered in policy making would results to informed decision making towards reducing or curbing the spread of COVID-19 cases in Sub-Saharan Africa.

I also thank John Hopkins University and Demographic Health Survey (DHS) for providing us with the COVID-19 Cases dataset and social demographic variables dataset respectively which were used in the study.

References

[1] E. M. a. V. G. C. Cambaza, "COVID-19: Correlation between gross domestic product, number of tests, and confirmed cases in 13 African countries," Journal of Public Health and Epidemiology, vol. 13, no. 1, pp. 14-19, 2021.

[2] Y. a. z. P. a. C. T. Lin, "Association between social economic factors and COVID-19 outbreak in the 39 well developed cities of China," Frontiers in Public Health, vol. 8, p. 546637, 2020.

[3] M. a. O. K. J. a. W. W. a. A. S. a. G. O. Fatima, "Geospatial Analysis of COVID-19: Ascoping review," International Journal of Environment Research and Public Health, vol. 18, no. 5, p. 2336, 2021.

[4] M. a. S. S. a. K. A. a. O. Rahman, "Geology and Topology based VS30 map for Sylhet City of Bangladesh," Bulletin of Engineering Geology and the Environment, vol. 78, no. 5, pp. 3069-3083, 2019.

[5] E. a. W. D. J. a. W. C. Thompson, "A VS30 Map for California with Geologic and Topographic Contraints," Bulletin of Seismological society of America, vol. 104, no. 5, pp. 2313-2321, 2014.

[6] L. G. a. K. R. E. a. T. Y. a. T. Thompson Eric M and Baise, "A geostatistical approach to mapping site response spectral amplifications," Engineering geology, vol. 114, no. 3-4, pp. 330-342, 2010.

[7] A. a. B. D. a. P. C. a. T. P. Marache, "Geotechnical modelling at the city scale using statistical and geostatistical tools: The Pessac case (France)," Engineering Geology, vol. 107, no. 3-4, pp. 67-76, 2009.

[8] R. M. a. K. J. a. T. S. Pokhrel, "A kriging method of interpolation used to map liquefaction potential over alluvial ground," Engineering geology, vol. 152, no. 1, pp. 26-37, 2013.

[9] P. a. Gunter, "Why do we need and how should we implement Bayesian kriging methods," Stochastic Environmental Research and Risk Assessment, vol. 22, no. 5, pp. 621-632, 2008.

[10] H. Omre, "Bayesian kriging Merging observations and qualified guesses in kriging," Mathematical Geology, vol. 19, no. 1, pp. 25-39, 1987.

[11] H. a. H. K. B. Omre, "The Bayesian bridge between simple and universal kriging," Mathematical Geology, vol. 21, no. 7, pp. 767-786, 1989.

[12] A. a. G. H. Chakraborty, "A Bayesian model reflecting uncertainties on map resolutions with application to the study of site response variation," Geophysical Journal International, vol. 214, no. 3, pp. 2264-2276, 2018.

[13] R. a. D. L. F. a. G. C. E. a. P. R. M. a. V. P. J. De Risi, "The SAFER geodatabase for the Kathmandu valley: Bayesian kriging for data-scarce regions," Earthquake Spectra, vol. 37, no. 2, pp. 1108-1126, 2021.

[14] H. a. S. A. a. M. D. E. Cui, "Extension of spatial information, Bayesian kriging and updating of prior variogram parameters," Environmetrics, vol. 6, no. 4, pp. 373-384, 1995.

[15] D. F. a. C. V. D. Machuka-Mory, “Non-Stationary geostatistical modeling based on distance weighted statistics and distribution,” Mathematical Geosciences, pp. 31-48, 2013.