Research Article
Open Access
CrossMark

COVID-19 infection: a Mozambican case study

Abstract

In China, the country of COVID-19 origin, until February 23rd, 2020, more than 77000 cases of COVID-19 infection were reported, and 60% of confirmed cases were reported in the city of Wuhan. Mozambique declared a state of emergency in March 2020, different prevention measures were implemented to control and respond in a timely manner to the pandemic, including the early diagnosis of cases of the disease. The present work reports some details about a larger project with the main objective of computing models of analysis and visualization of COVID-19 data in Mozambique. The topic falls within the area of Statistics with the purpose of providing evidence that explains the stage of the country regarding the evolution of COVID-19 cases, (from the notification of the first case of COVID-19 in Mozambique on March 22nd, 2020, until May 31st, 2022) with the focus on the provinces of Maputo, Nampula, Cabo Delgado and Niassa. The work considered qualitative and quantitative data to allow decision-making in the health area on measures to prevent the pandemic and the trend of cases and deaths from the disease.

Keywords: COVID-19, risk analysis, multiple regression

Volume 13 Issue 1 - 2024

M. Filomena Teodoro,[1,2] Teresa A. Oliveira,[3,4] Francisco Arune[5]

[1]CINAV- Center for Naval Research, Portuguese Naval Academy, Military University Institute, Portuguese Navy, Portugal
[2]CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Portugal
[3]Department of Science and Technology, Universidade Aberta, 1269-001 Lisboa, Portugal
[4]CEAUL (Center of Statistics and Applications of the University of Lisbon), 1649-014 Lisboa, Portugal
[5]Department of Science and Technology, Universidade Aberta, 1269-001 Lisboa, Portugal

Correspondence: M. Filomena Teodoro, CINAV- Center for Naval Research, Portuguese Naval Academy, Military University Institute, Portuguese Navy, Portugal and CEMAT - Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Av. Rovisco Pais, n. 1, Portugal, Tel +351927348873,
Email mteodor64@gmail.com;maria.teodor@tecnico.ulisboa.pt

Received: January 26, 2024 | **Published:** February 15, 2024

Abbreviations: WHO, the World Health Organization; SARS-CoV-2, corona virus 2 that causes severe acute respiratory syndrome.

Introduction

Corona viruses are a family of viruses known since the mid-1960s. These viruses cause respiratory infections in humans and animals. Generally, coronavirus infections cause mild to moderate illnesses - the common cold.[1]

The WHO was informed on December 31st, 2019, about the existence of cases of pneumonia with unidentified origin in China, Hubei Province, specifically in the city of Wuhan.[2] In subsequent weeks, the etiological agent was identified as a new corona virus that was called SARS-CoV-2 and the disease caused by this agent called COVID-19. In the first months of 2020, cases of COVID-19 infection were identified in several countries around the world. On March 11st, 2020, the outbreak of this disease was declared a pandemic by the WHO.[3]

The outbreak of corona virus disease 2019 (COVID-19) in the world was first diagnosed in the city of Wuhan, Hubei province in China and has spread gallopingly to all provinces of China including 28 other countries in the world. In China, the country where the disease originated, on February 23rd, 2020, more than 77,000 cases of infection by the disease had been reported, and from the analysis carried out, 60% of confirmed cases were reported in the city of Wuhan.

To ensure control of the pandemic and reduce the spread of the corona virus, several countries including Mozambique have declared a state of emergency and different prevention measures are being implemented to control and provide a timely response to the pandemic, including early diagnosis of cases of the disease.

In Mozambique, the first reported case of COVID-19 infection was registered in the capital of the country, Maputo, on March 22nd, 2020, in a man, over 75 years of age, with Mozambican nationality, who returned from a trip to the United Kingdom in mid-March.

This is an imported case of infection with the new Coronavirus, diagnosed in people who entered the country from abroad (countries of high endemicity) and due to weak surveillance, lack of means of diagnosis and reduced mechanisms of protection, until then allowed the spread of the disease throughout the country with a greater incidence in the cities of Pemba, Nampula, Matola and Maputo City.

Following the accelerated spread of COVID-19 cases, the country recorded the first death from the disease on May 25th, 2020, in Nampula, in the northern province of Mozambique. It was a 13-year-old child, whose sample was collected on May 20th in that province.[4]

For a robust spatial presentation of data from the problem under study, in addition to different tools that allow data manipulation, it is essential to use GIS due to its ability to gather a large amount of conventional data of spatial expression, providing adequate structuring and integration for manipulating geographic information.[5,6]

The main objective is to build a model of analysis and visualization of COVID-19 data in Mozambique. We considered the purpose of providing evidence that explains the current stage (from the notification of the first case of COVID-19 in Mozambique (March 22nd, 2020) until May 31, 2022) of the country regarding the evolution of COVID-19 cases. We focused our study on the provinces of Maputo, Nampula, Cabo Delgado and Niassa. The work used qualitative and quantitative data to allow decision-making in the health area on measures to prevent the pandemic and the trend of cases and deaths from the disease.

The outline of this work consists in six sections. Section 2 describes some preliminary details about the COVID-19 pandemic. In Section 3 are presented some possible data analysis models. Fourth Section details the multiple linear regression. The fifth Section corresponds to the empirical part of the work, where is described the used data and its

presented the estimated model by multiple linear regression approach. Also, it is described the model validation process and summarized the residual analysis. The manuscript ends with some discussion and some conclusions.

Preliminaries

The COVID-19 pandemic is a major public health threat globally. There were 332930 cases and 14510 deaths confirmed worldwide on March 23rd, 2020. Since initial identification in China, the spread of the disease has been rapid, with 182 of 202 countries reporting at least one case. Country experience to date has highlighted the intense pressure that the COVID-19 pandemic places on national healthcare systems, with demand for intensive care beds and mechanical ventilators.[4–7]

Several governments of countries hit by the COVID-19 pandemic have been introducing policies and strategies that aim to combat and control the new normal of the disease, based on a model of analyzing the behavior of pandemic cases with assumptions of social distancing, isolation, as well as monitoring of positive cases and their contacts.

However, it is not enough to analyze the problem of the COVID-19 pandemic by looking only at the attitudes and behavior of the population susceptible and exposed to the disease,[8,9] but it is also essential to analyze the problem under study from an environmental perspective that is, the analysis models must incorporate the region (physical space) and temperature, among others, as variables.[10]

Many uncertainties exist about the underlying determinants of the severity of COVID-19 infection. However, very clear risk factors for the disease include age, with older people more likely to require hospitalization and subsequently die as result of infection and underlying comorbidities, including hypertension, diabetes and coronary heart disease that exacerbate symptoms.[11]

The epidemiological consequences of the estimated mixture patterns are assessed by simulating the age distribution of infected individuals during the initial phase of an epidemic in a fully susceptible population. This distribution was obtained by computing the eigenvector of the matrix of the next generation associated with the contacts, considered for any possible age of the index case in the community, and assuming an age-independent transmission rate per contact, under the so-called "Hypothesis of social contact (Imperial College COVID-19 Response Team, 2020)".[6]

According to the same source, it is worth highlighting that the age distribution of cases during the initial phase of the epidemic does not depend either on the choice of the duration of the infectious period, nor on the considered value of the basic reproductive number.[12] This means that the timing and severity associated with an infection are generally disease-specific, although the impact of mixing patterns on the age distribution of cases is mainly caused by the type of contacts relevant to the infection, transmission of infections and socio-demographic structure of the considered population.[13] The work uses qualitative and quantitative data to allow decision-making in the health area on measures to prevent the pandemic and the trend of cases and deaths from the disease. Another interesting contribution about the COVID-19's epidemiological and demographic issues in Africa can be found.[14]

COVID-19 data analysis models

Mathematical modeling as well as statistical modeling seek to define models that describe any process (or Data can be analyzed in several ways. In detail, we can characterize them according to their temporal characteristics (static or dynamic), distribution, size, their nature (hydraulic, geological, accounting, linguistic, etc.), geographic location, among other forms.

Currently, there is a range of techniques that allow data modeling and visualization used to map information. In data modeling, before choosing the techniques to be used for data analysis, it is a priority to identify the appropriate mathematical model. However, the model must conveniently "translate" the relationship between the variables and, therefore, the real world and the respective data.

It is essential in statistical (scientific) modeling to eliminate possible noise (bias in data with abnormal characteristics when analyzed over time) through filtering, cleaning, and normalizing the database.

In problems where the objective is to analyze and study the relationship between variables (predictive and predictive), regression models are used. There are several statistical analysis methodologies[15] that allow explaining or describing the relationship between a variable of interest (response variable) and one or more variables (explanatory variables).

To understand and predict trends in COVID-19 cases, there are several data analysis models that can be applied. Below we present some of the most common and effective models that can be used to analyze COVID-19 data.

SIR and variant epidemiological models (SEIR, SIRD, etc.): These models are used to understand the spread of infectious diseases in a population. They divide the population into compartments (Susceptible, Exposed, Infected, Recovered) and use differential equations to model transmission rates, incubation, recovery, etc. These models are used to predict the evolution of the disease, understand the impact of control measures (such as isolation), and estimate the basic reproduction rate.[16–18]

Exponential and logistic growth models: These models are often used to predict the initial spread of a disease, such as COVID-19. They are based on an exponential growth rate that eventually reaches a limit (logistic model) as the population becomes susceptible.[19–21]

Time series models: these models analyze data over time to identify seasonal trends and patterns. They can be used to predict future numbers of cases, deaths or other indicators related to COVID-19 as well as to understand and predict patterns in temporal data, such as the evolution of a variable over time.[22,23]

Network models: these models consider the structure of the social network and contacts to understand how the disease spreads in interconnected communities. They can be useful for analyzing the effectiveness of social distancing measures.[24]

Risk analysis models: these models assess the risk of disease spread based on demographic, behavioral, geographic factors, among others.[25]

Spatial analysis models: using geoprocessing and GIS (Geographic Information Systems) techniques, these models help to understand how COVID-19 spreads geographically and how local factors can influence the spread.[26] Generally, spatial analysis of epidemiological events involves the study of patterns and processes in relation to their location in geographic space. There are several approaches and models in spatial analysis, and the specific equation will depend on the type of model being used to highlight the effect of the disease, for example, the spatial interpolation method denominated Kriging.

Kriging is a spatial interpolation method that estimates values at unsampled locations based on spatial relationships between sampled points.[27] In a spatial regression model, the relationship between the dependent variable and the independent variables is adjusted taking spatial dependence into account.[27–29] Also, we can consider longitudinal models to include the spatial dependence.[30]

Regression models: Regression - is a statistical technique that seeks to identify relationships between variables, allowing the prediction or understanding of how a dependent variable is influenced by one or more independent variables.[31]

When analyzing COVID-19 data, regression models can be applied in several ways, as is the case in predicting cases, analyzing risk factors, evaluating interventions developed to control the pandemic, analyzing the impact of interventions and among other aspects that allow the understanding of the phenomenon under study.

Below it is characterized the context of each approach:

- **Case Forecasting:** Regression models can be used to predict the future number of cases, deaths or other indicators related to COVID-19 based on historical data. These forecasts can assist in allocating resources, planning public health measures and making decisions.
- **Analysis of risk factors:** Regression models can help identify which factors (such as age, sex, comorbidities, population density, social distancing measures) are associated with a greater risk of infection or complications from COVID-19. This is essential to protect more vulnerable groups.
- **Evaluation of interventions:** Regression models can be used to evaluate the effectiveness of control measures, such as lockdowns, use of masks or vaccination campaigns. They can help understand how these interventions affect the spread of the disease.
- **Impact analysis:** Regression models can be used to evaluate the impact of different scenarios or variables on disease behavior. This is especially useful for anticipating the impact of changes in healthcare policies.
- However, it is important to keep in mind that regression models have limitations. They assume linear relationships between variables, may be sensitive to outliers. Regression models still are a valuable tool for understanding COVID-19 patterns, allowing predictions to be made that enable evidence-based decision making.[32,33]

Multiple linear regression

Multiple linear regression aims to study the relationship between two or more explanatory variables that are presented in linear form, and a metric dependent variable. Multiple regression is appropriate when the investigator seeks to understand how multiple independent variables (factors) are related to a continuous dependent variable. For the particular case of this study, multiple regression can be used to analyze how age, sex, comorbidity and social distancing measures are related to the number of COVID-19 cases in the country and particles in the four provinces under analysis.

Assumptions of the multiple linear regression model:

1. The dependent variable is a linear function of a specific set of variables and the error.
2. The expected value of the error term is zero.
3. The error has a normal distribution and does not present auto-correlation or correlation with any explanatory variable.
4. Observations of explanatory variables can be considered fixed in repeated samples.
5. There is no exact linear relationship between the explanatory variables and there are more observations than explanatory variables.

In multiple linear regression it is assumed that there is a linear relationship between a variable Y (the dependent variable) and k independent variables, $X_j\,(j=1,...,k)$. Independent variables are also called explanatory variables or regressors, once they are used to explain the variation in Y. They are often also called predictor variables, due to their use to predict Y.

The conditions underlying multiple linear regression are analogous to simple linear regression:

The independent variables $X_j\,(j=1,...,k)$ are non-random (fixed);

For each set of $X_j\,(j=1,...,k)$ values there is a subpopulation of Y values. To construct confidence intervals and hypothesis tests, it must be assumed that these subpopulations follow a normal distribution;

The variances of the subpopulations of Y are equal;

The Y values are statistically independent. In other words, when extracting the sample, it is assumed that the Y values obtained for a given set of $X_j\,(j=1,...,k)$ values are independent of the Y values obtained for any other set of $X_j\,(j=1,...,k)$ values.

The multiple linear model is, in its most general form, expressed by equation (1):

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k+\varepsilon. \tag{1}$$

The errors $\varepsilon_1,...,\varepsilon_n$ are independent and identically distributed $(\ddot{u}\ )$ with n the sample size. The errors have null expected value and are homoscedastics:

$$E(\varepsilon_i)=0\,,\quad \ddot{u}\quad (\varepsilon_i)=\sigma^2 \tag{2}$$

With $i=1,2,...,n.$

The estimator of $\boldsymbol{Y}$ is given by $\hat{\boldsymbol{Y}}$. The estimate of $\boldsymbol{Y}$ is given by $\hat{\boldsymbol{y}}$ and is computed as displayed in formula (3):

$$\hat{y}_i=f(x_1,...,x_k)=\ddot{\beta}_0+\beta_1x_1+...+\beta_kx_k. \tag{3}$$

In a linear model, to each regressor is assigned a numerical weight, called a regression coefficient, regression slope, or simply regression weight, which determines how much that explanatory variable contributes to the estimate $\hat{\boldsymbol{y}}$. These regression weights are derived by an algorithm that produces a mathematical equation or model for y that best fits the data, using some type of best-definition criteria (usually minimal square errors).

The theory of linear statistical modeling presents limitations. It assumes normality, but the distribution of the response variable can be non-Gaussian. To solve this problem (non-normal distribution) the generalized linear model cam be used.

The multiple linear model[34] is a particular case of generalized linear models (GLMs). GLMs are an extension of the linear model.[35] where the distribution of the response variable does not have to be normal, but rather another distribution from the exponential family (and the function that relates the expected value and the vector of explanatory variables can be any differentiable function for a probability density function written in general form given by equation 4:

$$f(y\mid\theta;\varphi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\varphi)}+c(y;\varphi)\right\}, \tag{4}$$

Citation: Teodoro MF, Oliveira TA, Arune F. COVID-19 infection: a Mozambican case study. *Biom Biostat Int J.* 2024;13(1):7–14.
DOI: 10.15406/bbij.2024.13.00408

where

θ - is the location parameter,

φ - is the dispersion parameter,

$a(.)$, $b(.)$ and $c(.)$ - are known real functions.

The Explanatory Power of the Regression Model is given by R^2.

The fraction of the sample variance of Y_i that is explained (predicted) by the regressions is designated by R^2 Likewise, R^2 considers "the proportion of the variation in the sample of the dependent variable explained by the explanatory variables". R^2 is used as a measure of the degree of adjustment".

The explanation (determination) coefficient is obtained by formula (5):

$$R^2 = \frac{SSR}{SST} \quad (5)$$

The adjustment or explanation coefficient can vary between 0 and 1 (0 to 100%), however it is practically impossible to obtain an explanation coefficient equal to 1, since it is unlikely that all points will fall on a straight line. Notice that when the explanation coefficient is 1, there will be no residuals for each observation in the sample under study. However, if the explanatory variables are not adequate to explain the behavior of y, the explanation coefficient will be close to 0. Therefore, as regression errors are minimized, the explanation coefficient will increase.[36–38]

Empirical application

Data

In the present study, to allow the analysis and comparison of the epidemiological profile of COVID 19 cases in Mozambique, four of the eleven provinces of the country were considered, namely Niassa, Cabo-Delgado, Nampula and Maputo. The inclusion and exclusion criteria for other provinces was based on the reasoning of the epidemiological profile with community transmission and the evolution of COVID-19 cases. The situation in Cabo-Delgado may present a different pattern due to insurgencies.

The four selected provinces presented a higher number of cases and deaths from the disease and a profile with an increasing trend until a certain period and shortly after the change in the profile presenting a curve of deceleration of cases. The climatic conditions of each province allowed a succinct analysis of the problem under study. This work was based on data reported from the notification of the first case of COVID-19 in Mozambique (March 22nd, 2020) until May 31st, 2022, considering this period as a cohort period for the study.

With the registration of a new wave of the disease reported in mid-January 2021, the provinces in question registered an increase in new COVID-19 infections and a galloping increase in cases and deaths in the first two months of 2021, taking a new epidemiological profile regarding records of new cases of people infected by COVID-19.

In these terms, the study was based on the analysis of data from the COVID-19 pandemic from the registration of the first case of the disease (2020) until April 2021 (the period in which pandemic vaccination began).

The considered database was compiled and standardized. It was built based on data obtained by the Provincial Health Services of the four provinces under study, with the objective of identifying and selecting analysis models and visualization mechanisms. This dataset covers a two-year period, from 2020 to 2022, and offers a comprehensive view of several variables related to population health.

The database includes demographic information, health status and other factors relevant to understanding the overall COVID-19 picture. The available variables cover a wide range of topics, from age and gender to location.

This database serves as a valuable source of information for researchers, health professionals and academics interested in exploring and understanding the various aspects related to population health in the context of the pandemic during the period considered. Before we proceed to the modeling step we have performed an exploratory multivariate data analysis.[39–41]

The model

For the particular case of this manuscript, the influence of certain variables that contributed to explaining the variability of COVID-19 cases in the country and, in particular, in the considered four provinces for the analysis of the pandemic, was investigated. Multiple regression provides diagnostics that can determine whether such effects exist based on empirical or theoretical argument. Indications of a high degree of inter-relationships (multicollinearity) between the independent variables may suggest the use of multiple scales.

A statistical relationship is characterized by two elements:

- When multiple observations are collected, more than one value of the dependent measure will generally be observed for any value of an independent variable.
- Based on the use of a random sample, the error in predicting the dependent variable is also considered random, and for a given independent variable, we can only expect to estimate the mean value of the dependent variable associated with it.

However, before starting with the analysis of the dependent variables that influenced the variation in the number of cases of COVID-19 infection, it is worth mentioning that the provinces taken as a sample showed community transmission over time, which provided an acceleration in the increase in number of COVID-19 cases. In addition to endogenous factors (household income, the insurgency in Cabo Delgado – armed attacks, non-compliance with safety standards – social distancing and mandatory collection) socio-economic and political factors negatively influenced the epidemiological profile of COVID-19.

To analyze the epidemiological profile of COVID-19 in Mozambique and in the provinces under analysis, it is important to look at the multiple factors that influenced the transmissibility behavior of the disease in different geographical areas of the Indian Ocean region (Mozambique). It was followed the Multiple Linear Regression approach to allow the determination of the value of response variable Y based on multiple influencing variables (predictors) considering the expression given by formula (1).

To understand the variability in the number of COVID-19 cases in Mozambique and in particular in the provinces of Niassa, Cabo Delgado, Nampula and Maputo Province, variables that influence the number of cases and which condition the epidemiological profile of the pandemic were taken as following:

- Age of the susceptible population (Child, Young, Adult or Elderly) ,
- Gender of susceptible population (Male, Female),
- Province - Geographic area (Urban or Rural).

Deplorable living conditions, weak community involvement in health activities, climatic/environmental factors (temperatures), socio-economic factors (poverty) and non-compliance with prevention standards (social distancing and mandatory collection) can be analyzed in isolation as well as a part of factors that influence the disease transmission chain.

To analyze the correlation between the variables (influencing factors) under study, a statistical measure was taken that shows the degree of relationship or association between the variables, one is denominated dependent or explained, the other ones are called independent (or explanatory or predictor variables). It was considered the following notation:

Y - COVID-19 cases as dependent or explained variable,

X_1 - Geographic area as independent or explanatory variable,

X_2 – Gender as independent or explanatory variable,

X_3 – Age as independent or explanatory variable,

without ruling out the possibility of other external factors of influence, such as social distancing.

From the analysis of the variables, one mathematical model that contribute to explain the epidemiological profile of the COVID-19 pandemic in Mozambique, followed the theoretical formula (1). The estimated model is given in equation (6):

$$\hat{y} = 474.32 - 409.53 x_1 - \quad - \quad (6)$$

The inclusion of the variable "geographical area – Province" in the model was to seek an explanation about the variation in the number of COVID-19 cases in the country and in the provinces under analysis. This variable inclusion translate the concentration of mass population in the areas urban areas or poor access to real-time prevention resources, as well as non-compliance with pandemic control policies.

Residual analysis, whether with residual plots or statistical tests, provides a simple but powerful set of analytical tools for examining the adequacy of our regression model. However, too often these analyzes are not done and violations of assumptions are left intact. Thus, users of the results are not aware of the potential inaccuracies that may be present, which range from inadequate tests of the significance of coefficients (showing significance where none exists or the opposite) to biased and inaccurate predictions of the dependent variable.

The model adjustment of the original series, the residuals of the observations show a determination about the trend of COVID-19 cases in Mozambique.

The Figure 1 allows to analyze linearity, normality of residuals, homoscedasticity. The top left graphic of Figure 1 presents the residual stage of adjusted values to analyze the linearity of the residuals of the linear regression model of the driving factors of the problem under study. However, the model residuals show some strange observations.

The top right graphic of Figure 1 displays standardized residues versus theoretical gaussian quantiles. The normality of residuals is not rejected, the concentration of points shows residuals are mainly close to the line.

Regarding the analysis of homoscedasticity, in the bottom left graphic of Figure 1, the residuals variability seems close to a constant (exception to the strange observations) which it's in agreement with the variance of the error terms (Ɛ) be constant in the range of values of the independent variable. The residuals do not present a triangular pattern around the regression line, which allows us to state that the residuals do not present heteroscedasticity.

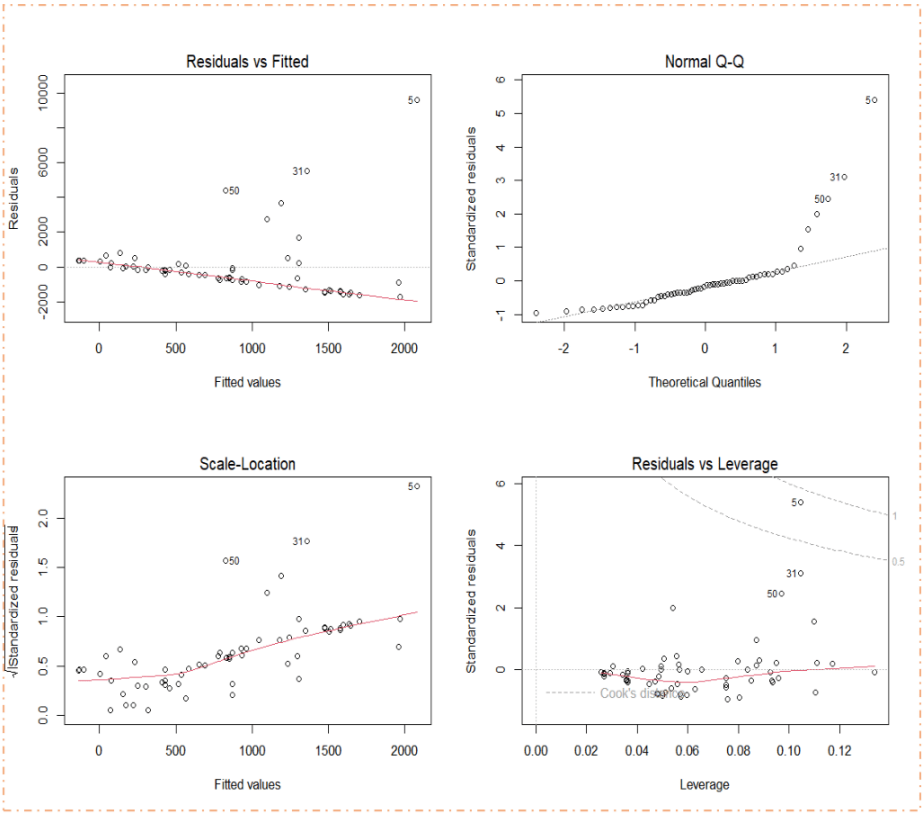

Figure 1 Graphical analysis of linear model assumptions.

Citation: Teodoro MF, Oliveira TA, Arune F. COVID-19 infection: a Mozambican case study. *Biom Biostat Int J.* 2024;13(1):7–14.
DOI: 10.15406/bbij.2024.13.00408

Regarding the analysis of data, we could identify and analyze strange observations.

Also, we can observe that at the bottom right graph of Figure 1, we find the leverage associated to standardized residuals. It is a diagnostic tool for influent observations. Notice that leverage refers to the extension to which the coefficients in the regression model would change if a particular observation is removed from the dataset. We can observe only one influent observation closest to the border of Cook's distance (n. 5), falling outside of the dashed line. It means that it is necessary to remove it and estimate again the model.

A key issue in interpreting the regression statistical variable is the correlation between the independent variables. This is a data problem, not a model specification problem. The ideal situation for a researcher would be to have several independent variables that are highly correlated with the dependent variable, but with little correlation between themselves. To perform the diagnostic of multicollinearity it was evaluated the tolerance value VIF parameter, which directly expresses the degree of multicollinearity with an impact on the estimation process. When the standard error increases, the confidence intervals around the estimated coefficients become larger, making it more difficult to demonstrate that the coefficient is significantly different from zero.

The VIF values show that there is no strong association between the independent variables (Province, Gender and Age): are equal to 1.0. This fact allow us to claim that there is no multicollinearity between the variables (factors associated with the behavior of COVID 19 cases). The same result is confirmed in Figure 2.

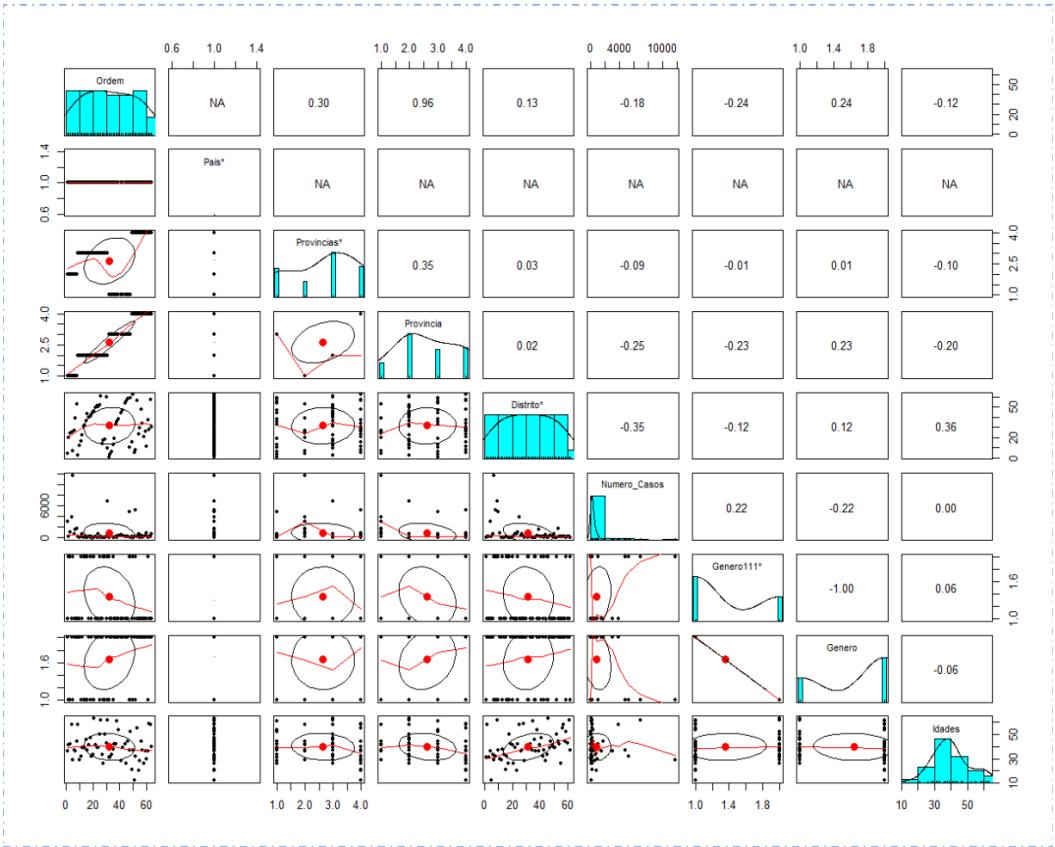

Figure 2 Analysis of multicollinearity through visualization.

Discussion and conclusion

This work results from a study about Covid – 19 infection in Mozambique. In the present manuscript we revisit some background about COVID-19 pandemic. We also got a model by GLM approach[38] that provided the distribution of COVID-19 cases in different age, gender and geographical area groups of the infected population. It is useful to understand how the pandemic affects different groups of people by age, gender and local which will contribute to formulate health policies aiming to control the spread of the pandemic.

In general, the data shows that youth and adults up to 39 years of age are the age group most infected by the corona virus in both social strata (male and female). Due to the characteristics of the country's population (smaller elderly population), the male population aged between 25 to 39 years old is the one more exposed to the risk factors of contracting COVID-19 in the 4 provinces under analysis. This issue can be justified by the mobility of this layer in search of living conditions without sparing efforts under the motto "stay at home", one of the methods used to control the spread of the disease.

The distribution of COVID-19 cases in the province of Cabo Delgado evidenced that the population in the age group between 25 and 34 years of age is the most infected by COVID-19, with a percentage around 33% of the total cases reported by the province, with emphasis on males, with a percentage of around 20%.

The distribution by age and gender of COVID-19 cases in the province of Niassa is not similar to Cabo Delgado case. The population considered vulnerable to COVID-19 was mostly female,

Citation: Teodoro MF, Oliveira TA, Arune F. COVID-19 infection: a Mozambican case study. *Biom Biostat Int J.* 2024;13(1):7–14.
DOI: 10.15406/bbij.2024.13.00408

aged between 20 and 34 years old. Between the 44% of the young population that became infected with the corona virus, around 22% were women.

The age characteristic of Maputo province presents a high-risk population concentration for the age groups from 20 to 44 years. The distribution of COVID-19 cases in this province does not show greater variability between females and males, despite the dominance of cases in the 40 to 44 age group being female. It is notable that the male population aged between 25 and 34 years old is the one that smoothes the curve of cases of the disease in the young population most infected by COVID-19 with a percentage of 14%.

The distribution of COVID-19 transmission in Nampula province shows a contamination situation similar to that of other geographic areas (Cabo Delgado, Niassa and Maputo) in which the population most affected is in the age group between 20 and 44 years old and with a greater weight for male individuals.

It should be noted that the population in the age group between 25 and 34 years is the most vulnerable and infected by COVID-19, with a percentage of around 30% of the total cases reported in the province, with emphasis on males with a percentage of around 15%. As extension of this manuscript we will present another models using some of the techniques described in section 3.

Also, we will evidence the comprehensive and accessible visualization of data and/or complex information, allowing people to understand patterns, trends and information relationships intuitively; simplifying complex data and/or information and making communication more effective; data visualization allows informed decision-making, that is, people make decisions based on evidence, as it is possible to highlight data patterns as well as anomalies and unusual points (normal patterns) that allow for accurate analysis of data.[5,29,42–47]

Acknowledgments

This work was supported by Portuguese funds through the Center of Naval Research (CINAV), Portuguese Naval Academy, Portugal and The Portuguese Foundation for Science and Technology (FCT), through the Center for Computational and Stochastic Mathematics (CEMAT), University of Lisbon, Portugal, project UID/Multi/04621/2019 and this research was partially funded by FCT - Fundação para a Ciência e a Tecnologia under the project - UIBD/00006/2020.

Conflicts of interest

The authors declare there is not any conflict of interest.

Funding

None.

References

1. National Institute of Health – INS. Análise da situação epidemiológica SARS-CoV-2/COVID-19. 2020.
2. World Health organization (WHO). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). 2020.
3. World Health organization (WHO). Coronavirus disease 2019 (COVID-19) Situation Report – 57. (2020a).
4. Ministério da Saúde Moçambicano COVID-19 em Moçambique Relatório do 1º Ano. 2021.
5. Gonçalves AC, Sousa AMO. Geographic Information Technologies. Universidade de **Évora**. 2017.
6. Ribeiro DM. Visualização de dados: mapas e cartografias do ciberespaço. Rio de Janeiro: ISBN. Imperial College COVID-19 Response Team. 2009.
7. Infarmed. SNS - Sistema Nacional de Saúde. 2020.
8. Silva CA, Carvalheira FM, Paulino JC. COVID 19 e os constrangimentos do mundo do trabalho dos profissionais de saúde: Contributos para a revisão da literature. 2021.
9. Manjate JLS, Chavane FS, Martins HR, et al. Knowledge, Attitudes and Practices of Mozambican Public Employees in relation to the Prevention of COVID-19 *Revista Produção e Desenvolvimento*. 2020;6:446.
10. Malheiro DR, dos Santos FAV, Araruna AOR, et al. *Perspetivas Socioambientais Sobre a COVID-19: Olhares interdisciplinares em ambiente e saúde*. Quipá Editora. 2021.
11. Campos MR, Schramm JAM, Emmerick ICM, Rodrigues JM, et al. Burden of disease from COVID-19 and its acute and chronic complications: reflections on measurement (DALYs) and prospects for the Brazilian Unified National Health System. *Cadernos de Saúde Publica.* 2020;36(11).
12. Diekmann O, Heesterbeek JAP, Metz JAJ. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J Math Biol*. 1990;28(4):365–382.
13. OMR. Caracterização das condições socioeconómicas dos deslocados internos no norte de moçambique ao longo do ano de 2021. 2022.
14. Martins HFB, Hansine R. Epidemiological and demographic analysis of COVID-19 in Africa. *Anais do Instituto de Higiene e Medicina Tropical (IHMT).* 2020;19:1–37.
15. Brauer F, Castillo-Chávez C. Mathematical models in population biology and epidemiology. Springer. 2012.
16. Anderson RM, May RM. *Infectious diseases of humans: dynamics and control*. Oxford University Press. 1992.
17. Gomes SCP, Rocha CR, Oliveira IO. Modelagem Dinâmica Aplicada à COVID-19. 2020.
18. Gomes SCP, Rocha CR, Oliveira IO. Dynamic modeling of COVID-19 applied to some Brazilian cities. *Revista Thema*. 2020a;18:1–25.
19. Malthus TR. *An essay on the principle of population*. J Johnson. 1798.
20. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. John Wiley & Sons; 2013.
21. Kleinbaum DG. *Logistic regression: A self-learning text*. Springer; 2010.
22. Durbin J, Koopman SJ. *Time series analysis by state space methods*. Oxford University Press; 2012.
23. Montgomery DC, Jennings CL, Kulahci M. *Introduction to time series analysis and forecasting*. John Wiley & Sons; 2008.
24. Newman MEJ. *Networks: An Introduction*. Oxford University Press; 2010.
25. Aven T, Vinnem JE. *Risk management with applications from the offshore petroleum industry*. Springer; 2007.
26. Diggle PJ, Ribeiro Jr PJ. *Model-based geostatistics*. Springer; 2007.
27. Cressie N. *Statistics for spatial data*. John Wiley & Sons; 1993.
28. Haining R. *Spatial data analysis: theory and practice*. Cambridge University Press; 2003.
29. Rosa A. Dados espaciais disponibilizados pelo sistema Geobases-Es e II. 2014.

30. Petersen MS, Kristiansen MF, Hanusson KD, et al. Long COVID in the Faroe Islands: a longitudinal study among non-hospitalized patients. *Clin Infect Dis*. 2021;73(11):E4058–E4063.

31. Cristina SR. *Linear regression model and its applications.* Monography. Universidade da Beira Interior Ciências. 2012.

32. Montgomery DC, Peck EA, Vining GG. *Introduction to linear regression analysis*. Wiley; 2012.

33. Chein F. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. In Coleção Metodologias de Pesquisa. 2019.

34. Yan X, Gang Su X. Linear regression analysis: theory and computing. World Scientific. 2009.

35. Watson, G. S. (1986). Generalized Linear Models (P. Mccullagh and J. A. Nelder). In SIAM Review (Vol. 28, Issue 1). https://doi.org/10.1137/1028043 (accessed January 18, 2024).

36. Cordeiro GM, Demétrio C. *Modelos Lineares Generalizados e Extensões*. 2010.

37. Campos MC. *Modelos de Regressão: uma aplicação em Medicina Dentária*. Master Thesis, Universidade Aberta, Lisboa. 2013.

38. Alvarenga AM. Tavares. *Modelos lineares generalizados: aplicação a dados de acidentes rodoviários Dissertação Mestrado em Gestão de Informação Especialização em Gestão e Análise de Dados*. Master Thesis, Universidade de Lisboa. 2015.

39. Fernandez PJ, Yohai V. *Introdução à Análise Exploratória de Dados Multivariados*. 2014.

40. Federighi E, Chagas B. Módulo 5 - Análise Multivariada no SPSS Análise Multivariada no SPSS. 2017.

41. Agresti A. *Categorical data analysis*. John Wiley & Sons; 2013.

42. Carvalho SM, Marcos AF. Visualização da Informação. Report. Centro de Computação Gráfica (CCG), Universidade do Minho; 2009.

43. Ricardo A, Grégio A, Pereira De Carvalho Filho B, et al. Capítulo 5 Técnicas de Visualização de Dados aplicadas à Segurança da Informação. 2009.

44. Santos M. A Visualização de dados na Teoria da Comunicação. Master Thesis, Universidade Federal de Juiz de For a. 2013.

45. Ferraz DR. Princípios de Visualização de dados aplicados no software de gestão financeira binfolio. Monography, NovaIMS-Universidade Nova de Lisboa. 2019.

46. Friendly M. *Visualizing categorical data*. 2005.

47. Dalla Corte AP, Silva C, Sanquetta R, et al. *Explorando o QGIS 3.X*. 2020.