




SPATIAL AND TEMPORAL ANALYSIS OF COVID-19 CASES IN PARAÍBA

ANÁLISE ESPACIAL E TEMPORAL PARA OS CASOS DE COVID-19 NA PARAÍBA

ANÁLISIS ESPACIAL Y TEMPORAL DE LOS CASOS DE COVID-19 EN PARAÍBA

 <https://doi.org/10.56238/isevmjv4n5-004>

Submission date: 08/01/2025

Publication date: 09/01/2025

Patrícia Silva Nascimento Barros¹

ABSTRACT

This article focuses on performing a spatial and temporal analysis of COVID-19 cases in Paraíba. Spatial analysis of area data is used in geoprocessing when the occurrence of the phenomenon under study is measured based on aggregated data by area, such as the number of COVID-19 cases per city. A time series is a set of ordered observations (in time). Time can be spatial, deep, or other variables. R software was used for the analyses. The data were obtained from the Paraíba State Department of Health. The results of the spatial analysis showed that the number of cities with more than 1,000 COVID-19 cases increased in 2021, and a slight decrease in the number of COVID-19 cases in Paraíba in 2022. The cities of João Pessoa, Campina Grande, and Patos had the highest number of cases each year. In the temporal analysis, the data needed to be transformed to a Normal distribution to apply the Box-Jenkins technique. The most suitable model was the ARMA(2,1) model, which obtained the lowest values for the selection criteria, the residuals also satisfied the conditions and the predicted values were within the confidence interval.

Keywords: Spatial Analysis. Temporal Analysis. Covid-19.

RESUMO

Este artigo se concentra em realizar uma análise espacial e temporal para os casos de Covid-19 na Paraíba. A análise espacial de dados de áreas é utilizada em geoprocessamento quando a ocorrência do fenômeno em estudo é mensurada a partir de dados agregados por área, como é o número de casos de Covid-19 por cidade. Uma série temporal é um conjunto de observações ordenadas (no tempo). Tempo pode ser: espaço, profundidade, ente outros. Utilizou-se o software R para fazer as análises. Os dados foram obtidos na Secretaria de Estado da Saúde da Paraíba. Com os resultados da análise espacial observou-se que no ano de 2021 aumentou o número de cidades com mais de 1000 casos de Covid-19 e observamos também que em 2022 houve uma pequena diminuição no número de casos de Covid-19 na Paraíba. Verificou-se que as cidades de João Pessoa, Campina Grande e Patos obtiveram maior número de casos em todos os anos. Na análise temporal foi preciso transformar os dados em distribuição Normal para aplicar a técnica de Box-Jenkis. O modelo mais adequado foi o modelo ARMA(2,1), que obteve os menores valores para os critérios de seleção, os resíduos também satisfaz as condições e os valores previstos ficaram dentro do intervalo de confiança.

¹ Dr. Universidade Federal da Paraíba. E-mail: patricia@dcx.ufpb.br

ORCID: <https://orcid.org/0000-0003-0681-2029> Lattes: <https://lattes.cnpq.br/8157392704703268>



Palavras-chave: Análise Espacial. Análise Temporal. Covid-19.

RESUMEN

Este artículo se centra en el análisis espacial y temporal de los casos de COVID-19 en Paraíba. El análisis espacial de datos de área se utiliza en el geoprocesamiento cuando la ocurrencia del fenómeno en estudio se mide con base en datos agregados por área, como el número de casos de COVID-19 por ciudad. Una serie temporal es un conjunto de observaciones ordenadas (en el tiempo). El tiempo puede ser espacial, profundo u otras variables. Se utilizó el software R para los análisis. Los datos se obtuvieron del Departamento de Salud del Estado de Paraíba. Los resultados del análisis espacial mostraron que el número de ciudades con más de 1000 casos de COVID-19 aumentó en 2021 y una ligera disminución en el número de casos de COVID-19 en Paraíba en 2022. Las ciudades de João Pessoa, Campina Grande y Patos tuvieron el mayor número de casos cada año. En el análisis temporal, los datos debieron transformarse a una distribución normal para aplicar la técnica de Box-Jenkins. El modelo más adecuado fue el modelo ARMA(2,1), que obtuvo los valores más bajos para los criterios de selección, los residuos también cumplieron las condiciones y los valores predichos estuvieron dentro del intervalo de confianza.

Palabras clave: Análisis Espacial. Análisis Temporal. Covid-19.



1 INTRODUCTION

Covid-19 is an acute respiratory infection caused by the SARS-CoV-2 coronavirus, potentially serious, highly transmissible and globally distributed. SARS-CoV-2 is a betacoronavirus discovered in bronchoalveolar lavage samples obtained from patients with pneumonia of unknown cause in the city of Wuhan, Hubei province, China, in December 2019 [BRASIL, 2023].

In Brazil, there have been 703,719 deaths from the disease since the beginning of the pandemic in 2020. In total, the country has 37,656,050 confirmed diagnoses, also since 2020. The most recent wave was in February 2022, when the number recorded in one week exceeded 6,000 deaths [BRASIL, 2023]. The new coronavirus, which causes Covid-19, has killed 10,545 people in Paraíba, according to official data released by the State Department of Health (SES). On April 3, 2021, the disease reached its most critical point in Paraíba: in one day, the SES confirmed 60 deaths from Covid-19 [PARAÍBA, 2023].

This article has the general objective of making a spatial and temporal analysis of Covid-19 cases in Paraíba from 2020 to 2022. The spatial analysis of area data is used in geoprocessing when the occurrence of the phenomenon under study is measured from data aggregated by area, such as the number of cases per city. On the other hand, temporal analysis is used in Epidemiology, when it is intended to analyze the behavior of the epidemiological indices and patterns expected from some phenomenon over time, thus allowing the planning of actions and public policies [LATORRE, 2001]. Time series analysis is a statistical method that can be used for the planning of actions and public policies, as it allows forecasting future events based on past data.

2 MATERIALS AND METHODS

The data were obtained from the website of the Department of Health of the Government of Paraíba [PARAÍBA, 2023]. Monthly data on Covid-19 in Paraíba from March 2020 to December 2022 were used to apply the time series technique. For the spatial analysis, data by cities from 2020 to 2022 were used.

Geoprocessing is a set of technologies for the analysis of spatial data, among which the GeoFigureic Information System (GIS) technology stands out. A GIS allows you to capture, store, retrieve, manipulate, analyze and present data. [MORAES, 2003]. In order to verify the distribution of Covid-19 cases in Paraíba, spatial analysis of area

data was used. The spatial analysis of area data is used in geoprocessing when the occurrence of the phenomenon under study is measured from data aggregated by area, such as the number of cases per city.

The analysis tool used was clustering, which is a process of grouping geo-objects, based on the values of their variables. These clusters are formed by dividing the data of a given variable into intervals, hence the choropleth (colored) maps are obtained [MORAES, 2003]. The data were grouped by area (cities), so it is necessary to organize them for visualization in a type of map called cadastral map. The Cadastral Map is a map where each of its elements is a geoFigureical object, which has attributes and can be associated with various Figureic representations [SOUZA, 2003].

Time series analysis is a statistical method that can be used for the planning of actions and public policies, as it allows forecasting future events based on past data. However, it is necessary that the data have a Normal distribution, which can be verified through the Lilliefors test (which is a variation of the Kolmogorov Smirnov Adherence Test) [SIEGEL, 1975]. To perform the test, consider the following statistic:

$$D = \max |F_n(x) - F(x)| \quad (1)$$

Where:

$F(x)$ represents the cumulative distribution function that is to be tested; $F_n(x)$ represents the empirical cumulative distribution function of the data. This function is defined for the entire value of x , and for each x gives the proportion of elements in the sample less than or equal to x . This procedure tests the hypothesis of normality of the data, if the p-value is greater than 0.05, then the data have a Normal distribution.

The time series analysis technique used was the Box-Jenkins technique, whose main objective is to make predictions. This methodology allows you to predict future values based only on your present and past values. One of the most frequent assumptions that is made about a time series is that it is stationary, that is, the mean and variance do not vary over time [MORETTIN, 2006]. A widely used test to check the stationarity of a series is the unit root test. The test used in this study was the ADF (Augmented Dickey-Fuller) test, which tests the hypothesis of non-stationarity of the series. If the p-value is less than 0.05, then the hypothesis of non-stationarity of the series is rejected, so the series is stationary [DICKEY-FULLER, 1979].

The models used to describe time series are stochastic processes, that is, ordered processes (in time) controlled by probabilistic laws. Box-Jenkins models are used to model stationary series. They encompass the following models: Autoregressive (AR), Moving Averages (MA) and Autoregressive and Moving Averages (ARMA) [MORETTIN, 2006]. In stationary models, the observations are independent, that is, they are not affected by the change of a time origin, they do not present a trend or seasonality [MORETTIN, 2006]. To test seasonality, the Kruskal-Wallis test is used: if the p value is greater than 0.05, the series does not have seasonality. The Cox Stuart test was used to verify if the series has a trend, if the p value is greater than 0.05, we conclude at the level of 5% of significance that the series does not have a trend.

An autoregressive model is a univariate time-series model in which the random variable of interest is described only by its past values and random error. An autoregressive model of order p is represented by $AR(p)$, and will be the weighted sum of the past p values of the variable, in addition to white noise (random error) [MORETTIN, 2006]. A moving average model results from the linear combination of random errors (white noise) that occurred in the current period and in past periods. A moving averages model of order q involves lagged values and is indicated by $MA(q)$. An ARMA model is obtained by combining the autoregressive and moving average components, i.e., Z_t is described by its past values and by current and past random errors. The generic specification of a ARMA model admits an autoregressive component of order p and a component of moving averages of order q $ARMA(p,q)$ [MORETTIN, 2006].

Steps of the Box-Jenkins methodology

This method makes use of an iterative approach, which has three main steps, starting from a stationary time series:

- i. Identification: In this first step, the appropriate values of the parameters p , d and q are identified through the analysis of autocorrelation and partial autocorrelation, evaluating the correlations in lags of k periods;

Table 1

FAC and FACP Behavior of an ARMA Process(p , q)

Model	Autocorrelation (FAC)	Partial autocorrelation (FACP)
$AR(p)$	Exponential decay	abrupt drop to zero when $k > p$
$MA(q)$	$\rho(k = 0)$ if $k > q$	Exponential decay
$WEAPON(p,q)$	Exponential decay from q	Exponential decay from p

- ii. Model diagnosis: In this step, through computational resources, the autoregressive parameters and moving averages are estimated and tested to obtain coefficients that best represent the selected model. This step also consists of evaluating the model's compliance with the Akaike Information Criterion (AIC), which should be as low as possible. The most commonly used definition is:

$$AIC = -2 \log(\text{verossimilhança maximizada}) + 2m \quad (2)$$

Where:

m is the number of estimated parameters (in ARMA(p,q) models, $m = p + q + 1$) [AKAIKE, 1974].

Some measures of interest based on actual and predicted values can also be calculated to help choose the best model, among them we can mention:

- a) Criterion 1: Total Error (TE)

$$C1 = \sum_{j=t+1}^{t+k} y_j - \sum_{j=t+1}^{t+k} \hat{y}_j \quad (3)$$

- b) Criterion 2: Mean Percent Error (MPE)

$$C2 = \left(\frac{\sum_{j=t+1}^{t+k} y_j - \sum_{j=t+1}^{t+k} \hat{y}_j}{\sum_{j=t+1}^{t+k} y_j} \right) * 100 \quad (4)$$

- c) Criterion 3: Mean Squared Error (MSE)

$$C3 = \frac{1}{h} \sum_{j=t+1}^{t+k} (y_j - \hat{y}_j)^2 \quad (5)$$

- d) Criterion 4: Mean Absolute Error (MAE)

$$C4 = \frac{1}{h} \left(\sum_{j=t+1}^{t+k} |y_j - \hat{y}_j| \right) \quad (6)$$

It is also checked whether the estimated residuals can be seen as white noise.



The verification of the assumption of constant variance is done through the residual Autocorrelation Function (FAC) and the Ljung and Box (1978) test. To assume normality, the analysis of the histogram of the residuals is used and the Lilliefors test is applied to the residuals of the series. If the model is not suitable, you should return to the first step and look for a new model. When a suitable model is found, it is time to move on to the next step.

iii. Forecast: Once the one that proved to be the most adequate has been found among the estimated models, the last and most important stage of the Box-Jenkins methodology is reached, which consists of making predictions for the series at moments of time after the sampled. The optimal predictor "l steps ahead", represented by is the one that minimizes the mean square error of the forecast: $\hat{y}_n(l)$

$$E[y_{n+1} - \hat{y}_n(l)]^2 = E[e_n^2(l)] \quad (7)$$

Where:

$e_n(l)$ is the prediction error l steps ahead of n.

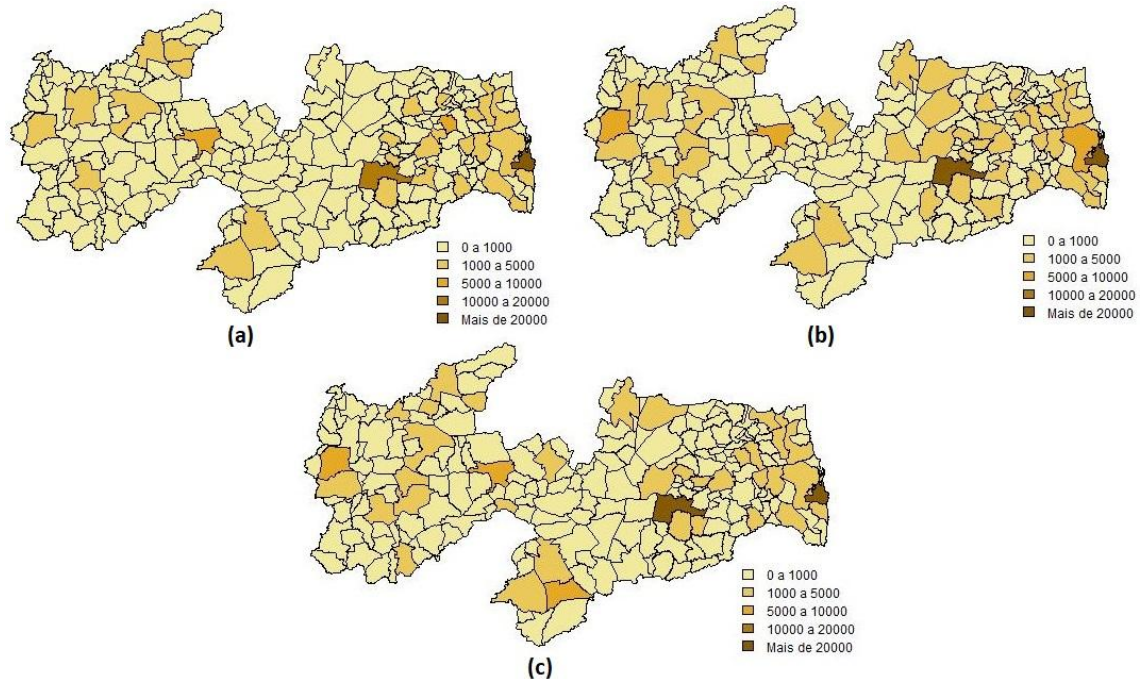
3 FINDINGS

3.1 RESULTS DESCRIPTIVE AND SPATIAL ANALYSIS

The groupings carried out refer to the number of Covid-19 cases in Paraíba, by city. In this way, we will verify the cities with the highest number of cases in each year studied. The results of the groupings were as follows:

Figure 1

Spatial distribution of the number of Covid-19 cases in Paraíba in the years 2020 (a), 2021 (b) and 2022 (c)



In Figure 1 we can see the distribution of the number of Covid-19 cases in Paraíba in the years 2020 (a), 2021 (b) and 2022 (c). It can be seen that in 2021 the number of cities with more than 1000 cases of Covid-19 increased. We also observed that the cities with more than 5000 cases of Covid-19 in 2020 were: João Pessoa (49314), Campina Grande (17540), Patos (7644), Guarabira (5651). In 2021, the cities with more than 5000 cases were: João Pessoa (63214), Campina Grande (30627), Cajazeiras (7996), Patos (7987), Bayeux (5642), Cabedelo (5430) and Santa Rita (5346). In 2022, the cities with more than 5000 cases were: João Pessoa (65016), Campina Grande ((21386), Patos (7947), Cajazeiras (7298) and Camalaú (5001). We can also observe that in 2022 there was a small decrease in the number of Covid-19 cases in Paraíba.

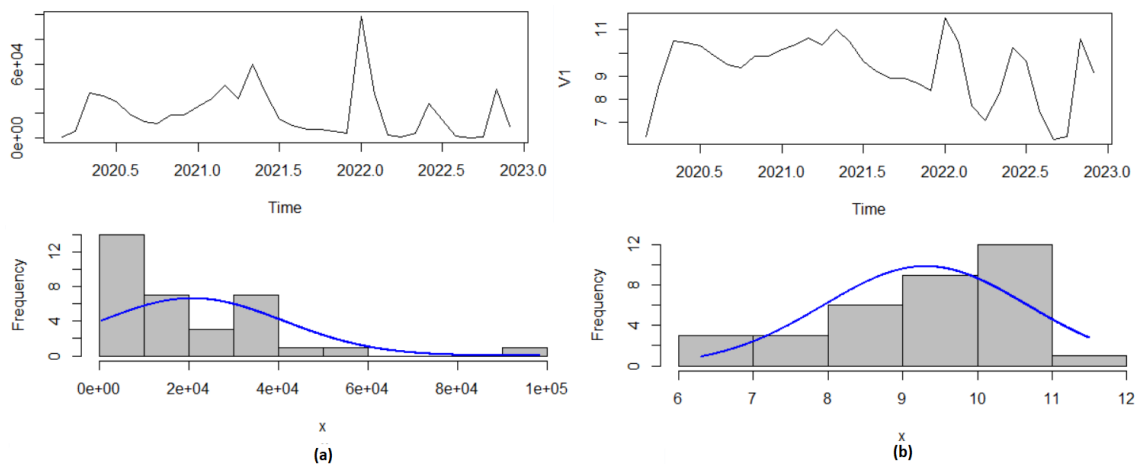
3.2 TIME SERIES RESULTS

For modeling through the Box & Jenkins methodology, the series needs to have a Normal distribution. For this, the Lilliefors normality test was performed obtaining a p-value of 0.0244, so the series does not have a normal distribution (Figure 2a), it is necessary to transform the series into a Normal distribution before proceeding with the

adjustment of time series. The transformation was performed through the logarithm obtaining a p-value of 0.2141, as can be seen in Figure 2b.

Figure 2

Time series and histogram of Covid-19 cases in Paraíba from March 2020 to April 2023



It is observed in Figure 3, of the series of Covid-19 cases and through the ADF (Augmented Dickey-Fuller) unit root test with a p-value of 0.01, that the series is stationary. Figure 4 shows that two autocorrelations exceed the significance threshold and the others are relatively small. In Figure 4, it is observed that only a partial autocorrelation exceeds the significance limit. Thus, one can start with an ARMA model(2,1).

Figure 3

Time series of Covid-19 cases in Paraíba from March 2020 to April 2023

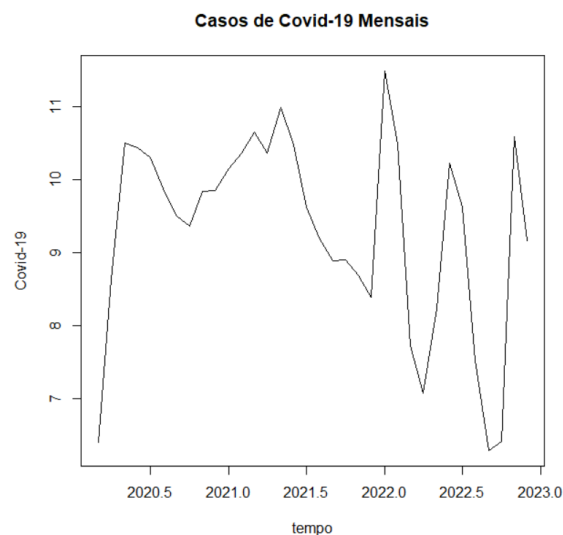


Figure 4

Autocorrelation function and partial autocorrelation of the series of Covid-19 cases in Paraíba

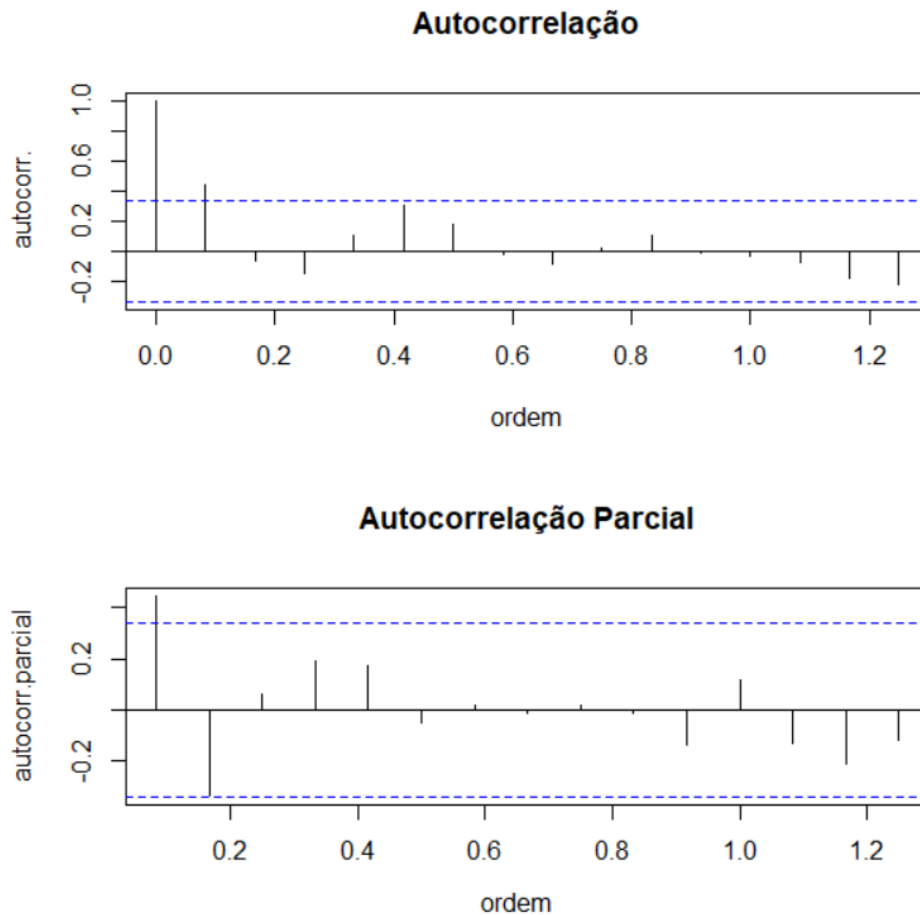
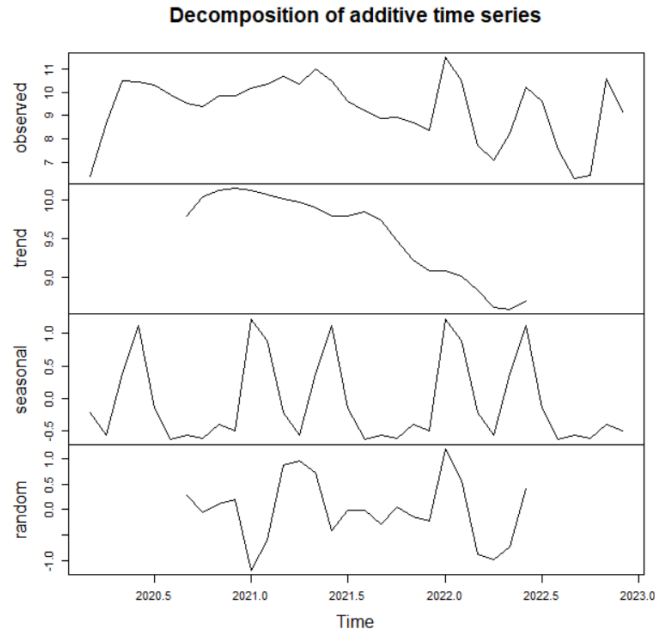


Figure 5 shows the breakdown of the series, where it is observed that it does not present a trend or seasonality. By the Kruskal-Wallis test we obtained a p-value of 0.4676, so the p value was greater than 0.05. At the level of 5% of significance, we have statistical evidence that the series does not have seasonality. The Cox Stuart test was used to verify whether the series has a trend, as the p value was greater than 0.05 (0.3323), we concluded at the level of 5% of significance that the series does not have a trend.

Figure 5

Decomposition of the series of Covid-19 cases in Paraíba into three components: seasonality, trend and residue



The models that obtained the lowest AIC were ARMA(2,1) with AIC = 81.76 and ARMA(3,1) with AIC = 83.62. Table 2 shows the values of the criteria to select the best model, so we chose the ARMA model(2,1) that has the lowest values of the AIC (81.76), C3 (4.16) and C4 (1.83) criteria.

Table 2

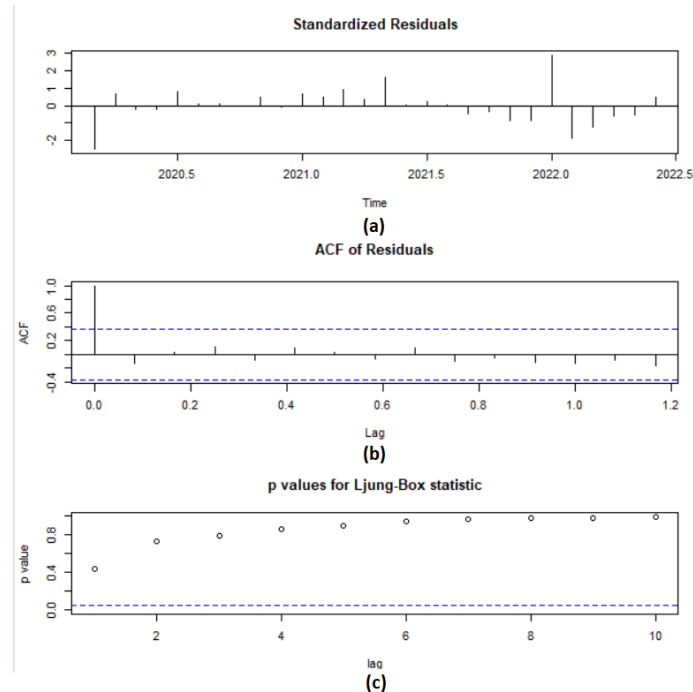
Values of the AIC, C1, C2, C3 and C4 criteria for forecasts of the series of Covid-19 cases in Paraíba

Model	AIC	C1	C2	C3	C4
GUN(2.1)	81,76	-8,59	-17,31	4,16	1,83
WEAPON(2.2)	83,67	-8,77	-17,67	4,48	1,89
WEAPON(3.1)	83,62	-8,66	-17,46	4,35	1,87

In the analysis of the residuals, the ARMA(2,1) model fulfilled the conditions. In Figure 6a are the standardized residuals in which it is verified that they are within the specified limits. It can be seen in Figure 6b that no autocorrelation is outside the confidence interval for both models. Figure 6c shows that all p-values are above the dotted line (0.05), indicating that the autocorrelations of the residuals are statistically equal to zero.

Figure 6

Standardized residuals (a), residual autocorrelation function (b) and p-values of the Ljung-Box statistic of the series of Covid-19 cases in Paraíba



According to the histogram (Figure 7a) and the Qqplot (Figure 7b), it can be seen that the residuals have an approximately normal distribution. This fact can be confirmed with the Lilliefors normality test, with a p-value of 0.68, it can be concluded that the residues are distributed in an approximately normal way.

Figure 7

Histogram and qq Plot of the residuals of the series of Covid-19 cases in Paraíba

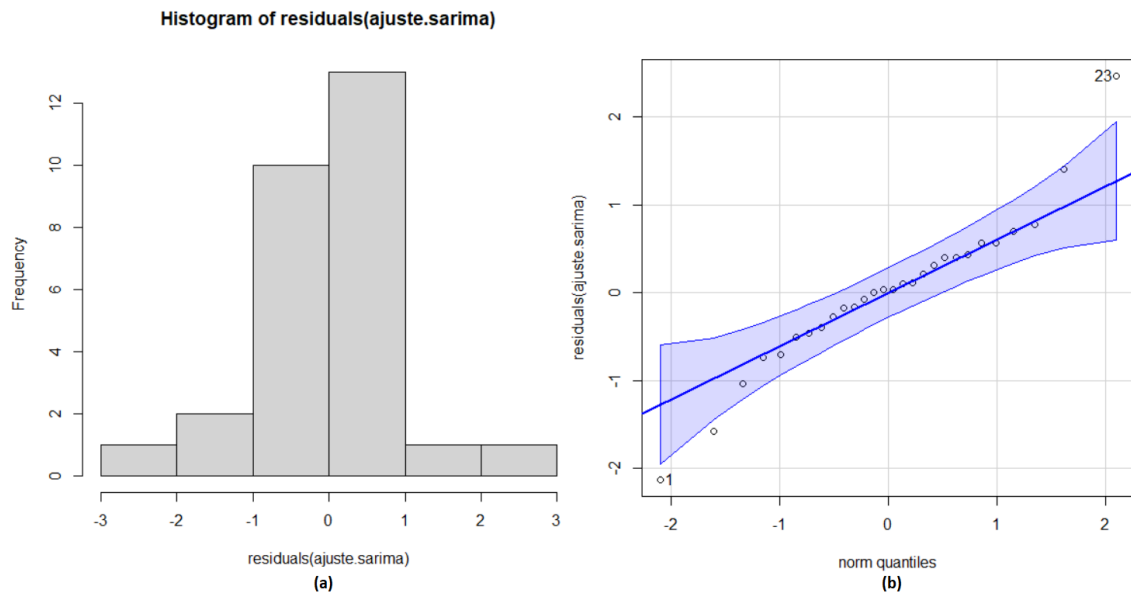


Table 3 shows the estimated coefficients for the chosen ARMA model(2,1), and it is verified that the estimated coefficients are significant. Table 3 presents a comparison between the actual and predicted values for the model and it is verified that the predicted values are within the specified limits.

Table 3

Estimated coefficients for the ARMA(2,1) model of the series of Covid-19 cases in Paraíba

Coefficient	Esteemed	Standard Error
Ar1	0,6662	0,3628
Ar2	-0,5167	0,2754
Ma1	0,2557	0,3975
Intercept	9,5449	0,2431

Table 4

Comparison of the actual values and predicted values of the ARMA(2,1) model of the series of Covid-19 cases in Paraíba

Months	Real Values	Expected Values	Lower Limit *	Upper Limit *
July 2022	9,63	10,78	9,11	12,45
August 2022	7,52	10,01	7,74	12,28
September 2022	6,29	9,22	6,94	11,50
October 2022	6,41	9,09	6,70	11,47
November 2022	10,60	9,41	6,97	11,85



December 2022	9,16	9,69	7,25	12,13
---------------	------	------	------	-------

*With 95% confidence

4 CONCLUSIONS

This article focused on performing a spatial and temporal analysis for Covid-19 cases in Paraíba. R software was used to perform the analyses. The data were obtained from the Paraíba State Department of Health. For the results of the spatial analysis, it was observed that in 2021 there was an increase in the number of cities with more than 1000 cases of Covid-19 and we also observed that in 2022 there was a small decrease in the number of cases of Covid-19 in Paraíba. It was found that the cities of João Pessoa, Campina Grande and Patos had the highest number of cases in all years. The city of Cajazeiras had the highest number of cases in 2021 and 2022. In the temporal analysis, it was necessary to transform the data into Normal distribution to apply the Box-Jenkins technique. The most appropriate model was the ARMA model(2,1), which obtained the lowest values for the selection criteria. In the analysis of the residuals, the ARMA model(2,1) fulfilled all the conditions. The estimated coefficients for the chosen ARMA model(2,1) are significant. Comparing the actual and predicted values for the model and it was found that the predicted values are within the specified limits.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Brasil, Ministério da Saúde. (2023). Coronavírus. <https://www.gov.br/saude/pt-br/coronavirus>
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), 427–431. <https://doi.org/10.1080/01621459.1979.10482531>
- Ehlers, R. S. (2009). Análise de séries temporais. <http://www.icmc.usp.br/ehlers/stemp/stemp.pdf>
- Latorre, M. R. D. O., & Cardoso, M. R. A. (2001). Análise de séries temporais em epidemiologia: Uma introdução sobre os aspectos metodológicos. *Revista Brasileira de Epidemiologia*, 4(3), 147–155.
- Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303. <https://www.jstor.org/stable/2335207>



Moraes, R. M., & Souza, I. C. A. (2003). Utilização de sistemas de informação geográfica na análise espacial de dados de saúde pública na Paraíba entre os anos de 1998 e 2001 [Relatório de PIBIC]. Universidade Federal da Paraíba.

Morettin, P. A., & Tolo, C. M. de C. (2006). Análise de séries temporais (2nd ed.). Blucher.

Paraíba, Secretaria de Saúde. (2023). Coronavírus. <https://paraiba.pb.gov.br/diretas/saude/coronavirus>

Siegel, S. (1975). Estatística não paramétrica para as ciências do comportamento. McGraw-Hill do Brasil.

Star, J., & Estes, J. (1990). Geographic information systems: An introduction. Prentice-Hall.