Mathematical Biosciences and Engineering

https://www.aimspress.com/journal/mbe

MBE, 21(12): 7760–7782.
DOI: 10.3934/mbe.2024341
Received: 27 August 2024
Revised: 16 November 2024
Accepted: 22 November 2024
Published: 06 December 2024

Research article

Prescriptive temporal modeling approach using climate variables to forecast dengue incidence in Córdoba, Colombia

Ever Medina[1], **Myladis R Cogollo**[1] **and Gilberto González-Parra**[2,*]

[1] Departamento de Matematicas y Estadistica, Universidad de Cordoba, Monteria 230002, Colombia
[2] Department of Mathematics, New Mexico Tech, New Mexico 87801, USA

* **Correspondence:** Email: Gilberto.GonzalezParra@nmt.edu.

Abstract: We present a modeling strategy to forecast the incidence rate of dengue in the department of Córdoba, Colombia, thereby considering the effect of climate variables. A Seasonal Autoregressive Integrated Moving Average model with exogenous variables (SARIMAX) model is fitted under a cross-validation approach, and we examine the effect of the exogenous variables on the performance of the model. This study uses data of dengue cases, precipitation, and relative humidity reported from years 2007 to 2021. We consider three configurations of sizes training set-test set: 182-13, 189-6, and 192-3. The results support the theory of the relationship between precipitation, relative humidity, and dengue incidence rate. We find that the performance of the models improves when the time series models are previously adjusted for each of the exogenous variables, and their forecasts are used to determine the future values of the dengue incidence rate. Additionally, we find that the configurations 189-6 and 192-3 present the most consistent results with regard to the model's performance in the training and test data sets.

Keywords: Dengue; climate variables; time series; incidence rate; SARIMAX; exogenous variables

1. Introduction

Dengue is a viral disease transmitted by the *Aedes Aegypti* and *Albopictus* species of mosquitoes, that spreads rapidly around the world and is a major cause of acute morbidity in more than 125 countries. Dengue mainly affects countries with a tropical and/or subtropical climate in Central America, South America, and Southeast Asia. It is estimated that between 100 and 400 million people are infected with dengue each year and around 3.9 billion people live in countries where dengue is endemic [1, 2].

The prevalence of dengue has multiplied by 30 in the last 50 years, since its geographic distribution has continued to expand to new countries and within urban to rural areas. Dengue cases reported to

the World Health Organization (WHO) have considerably increased, from 505,430 in the year 2000 to 2.4 million in 2010. Furthermore, a total of 5.2 million cases were reported in 2019. With regard to deaths, the number of deaths went from 960 in the year 2000 to 4032 in the year 2015, with the highest number of victims being young people [1, 3].

In Colombia, 68.72% of its municipalities are considered endemic, which correspond to populations located at less than 2200 meters above sea level, and its main transmission vector is the *Aedes Aegypti* mosquito [4, 5]. In 2010, Colombia registered its highest number of dengue cases with 157,203 cases, of which 9777 corresponded to severe dengue and 217 deaths were reported.

It has been suggested that the effect of certain exogenous variables such as climate change, high temperatures, or the La Niña phenomenon have caused an increase in dengue cases throughout Colombia [6, 7]. Between 2010–2011, the Niña phenomenon left approximately 2.4 million homeless people and 900,000 affected persons in 1061 municipalities [8]. The department of C´ordoba stands out among the departments of Colombia affected by this phenomenon with a high incidence of dengue. The department of Córdoba is located in the Caribbean region of Colombia and is made up of 30 municipalities, of which 90% are characterized by the permanent presence of dengue and for being between 0–499 meters above sea level. When examining the number of cases of dengue, severe dengue, and deaths registered in the department of Córdoba by the National Institute of Health, during the period 2017–2021 (see Table 1), it is evident that the number of dengue cases reported in the year 2021 presented an increase of 328% in relation to the year 2017. The increase was 182% with regard to the cases of severe dengue. Additionally, it is observed that there have been deaths every year due to dengue. Regarding the fatality rate, it is found that during the 2017–2021 period, there were 18 fatal cases for every 100 reported cases of severe dengue. In the years 2018 and 2021, Córdoba was among the top 10 territorial entities with the most cases of dengue reported in the country. Therefore, it is important to gain further insight into the effects that climate variables have on the dengue incidence.

It is important to mention that the analysis of the incidence of dengue cases over time is an issue that has been addressed through the use of univariate models of time series, according to the characteristics of the data. When the current value of the series can be explained based on its own lags, the moving averages, and the autoregressive components, then the Autoregressive Integrated Moving Average (ARIMA) model can be used [9]. Additionally, if the time series presents seasonal fluctuations, then an ARIMA variation called the Seasonal Autoregressive Integrated Moving Averages model (SARIMA) can be used [10]. In addition, when the ARIMA model includes other time series as the input variables, the Autoregressive Integrated Moving Average model with exogenous variables (ARIMAX) can be used. Finally, if the ARIMAX model also considers seasonal fluctuations, then the Seasonal Autoregressive Integrated Moving Average model with exogenous variables (SARIMAX) can be implemented. Regarding the use of these aforementioned models to forecast dengue cases, it has been found that in the last 5 years, the ARIMA model continued to be the most frequently used (see for example [11–15]). For instance, in [16–19], the ARIMA model was used to investigate Dengue in different countries such as Brazil and Colombia. The SARIMA model is the second most used to analyze time series related to Dengue (see [9, 12, 16, 20, 21]). The next most popular models are the ARIMAX (see [15, 22]) and SARIMAX (see [9, 22]) models. Recently, in [23] the tuberculosis incidence was estimated using a SARIMAX-NNARX hybrid model by integrating meteorological factors. Moreover, SARIMAX models have been used to forecast gas production [24].

In the literature, there are few studies related to time series modeling to forecast dengue cases in

Colombia. In [18] and [19], the authors examined the total number of dengue cases registered in Colombia using an ARIMA model. However, in these previous valuable works, they did not consider the effect of exogenous variables, nor did they perform a cross-validation procedure that allowed them to determine the appropriate forecast horizon. Moreover, only descriptive analyses of cases of dengue in the Córdoba department have been made. There are studies that have investigated the relationship between climate variables and diseases such as Leptopirosis, Respiratory syncytial virus (RSV), and others [25–27]. In particular, the correlation between climate variables, seasonality, and dengue cases have been investigated [22, 28–31]. Additionally, there are studies that have used time series related to dengue within the scientific literature [14, 32–35]. Some studies have used climate variables to predict the number of cases and fatalities associated with the West Nile virus, RSV, Zika, Chikungunya, and dengue [27, 36–39]. In particular, in [9], it was used to assess the effect of climate variables in Panama City. In [40], the weather variability in the spread of West Nile Virus in Texas was investigated. Finally, in [41], the effect of climate conditions on forecasting cases in Recife, Brazil, were studied in relation to three different viruses: Chikungunya, Dengue, and Zika.

In this article, we evaluate whether it is possible to apply an autoregressive model SARIMAX of moving averages with exogenous variables to forecast the short-term incidence rate of dengue per 100,000 inhabitants in the department of Córdoba (Colombia) using the historical information available in official sources, including some climate variables, such as precipitation and relative humidity. Furthermore, we propose the use of a prescriptive approach in the modeling strategy to determine the effect that the quality of the information of the exogenous variables has on the performance of the adjusted model. Three realistic scenarios are considered based on the available exogenous information and different forecast horizons. For each scenario and horizon, the SARIMAX model is adjusted, and finally the accuracy of the forecasts generated by each of them is evaluated.

The importance and originality of this work lies in the following:

1) The proposed modeling strategy allows us to know the quality of the information on the exogenous variables necessary to obtain a SARIMAX model with a good performance in terms of the accuracy of the forecasts of the dengue incidence rate.

2) To date, a time series model has not been proposed to forecast dengue cases reported in the department of Córdoba, thereby considering the effect of meteorological variables.

3) Due to the increase in reported cases of dengue in the department of Córdoba and the high human and economical cost, it is necessary to support health entities with a statistical tool that allows them to forecast potential new cases of dengue in the department. This tool should consider the effect of different measurable meteorological variables. This would contribute to the development of preventive public health policies.

2. Material and methods

The proposed methodology of this work incorporates the SARIMAX model and the Box-Jenkins methodology proposed in [10]. First, we will proceed with a description of the data set related to dengue. Then, we present the theory related to the SARIMAX model. Finally, we present the proposed methodology in detail.

Table 1. Number of dengue cases reported per year in the department of Córdoba, Colombia, between 2017–2021. The number of mild cases is in the second column, the number of severe cases is shown in the third column, and the total number of cases fatalities due to dengue that occurred each year is shown in the fourth column. The last column contains the position occupied by the department on a scale of 1–38, by ordering the 38 territorial entities of Colombia in descending order, according to the total number of dengue cases reported in a year. Source: Public Health Surveillance System (SIVIGILA), National Institute of Health, Colombia, 2017–2021.

Year	Dengue cases	Severe dengue cases	Deaths	National rank
2017	584	11	1	11
2018	3800	41	15	4
2019	4625	48	2	11
2020	1650	13	3	12
2021	2500	31	5	6
Total	13,159	144	26	

2.1. Dengue incidence data and climate variables in Córdoba

The information related to dengue cases was retrieved from the web portal of the National Public Health Surveillance System of Colombia (SIVIGILA), which contains the databases of dengue cases reported in Colombia from 2007 to 2021. The information reported for the department of Córdoba was filtered by week and epidemiological year. We consider the fact that an epidemiological year contains 13 epidemiological periods and each period is made up of four consecutive epidemiological weeks, for a total of 52 weeks per epidemiological year (with the exception of the years 2008, 2014, and 2020 that had 53 weeks; in fact, they had one epidemiological period with 5 consecutive weeks). Based on this information, the number of dengue cases per epidemiological period (denoted by *cases period*) was determined, which corresponds to the sum of the reported cases in 4 consecutive epidemiological weeks), such that for each of the fifteen years of the study, there are 13 records of dengue cases.

This data, together with the sizes of the populations estimated by the National Administrative Department of Statistics of Colombia (DANE) for the department of Córdoba since 2007 until 2021, are the basis for calculating the values of the response variable. When analyzing the relationship between the total number of reported dengue cases and the size of the population, the Pearson correlation coefficient was found, $\hat{\rho} = 0.49$ ($p-value = 2.825e-10$). This indicates that there is a positive and significant relationship between the total number of reported dengue cases and the population size, that is, the larger the size of the population, the greater the number of dengue cases are reported. This means that the study of the dynamics of the disease over time should not be solely studied by the number of cases. For this reason, the effect of population size is controlled by specifying the dengue incidence rate per 100,000 inhabitants in the department of Córdoba as a variable of interest, which is given by

the following:

$$\textbf{Incidence rate} = \frac{Cases\ period_i}{N_i} \times 100,000;\ i = 1, \ldots, 13 \tag{2.1}$$

where $Cases\ period_i$ corresponds to the total reported cases of dengue in the period i, and N_i is the population size of the department for the period i. In total, there are 195 values for the response variable and whose behavior is shown in Figure 1. The relative humidity (%) and precipitation (mm) reported in the department of Córdoba between 2007 to 2021 are considered as exogenous variables. The daily records of these variables were provided by the Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). To unify this information with the dengue incidence rate, it was necessary to gather the information of the exogenous variables by the epidemiological week and then add up the information by the epidemiological period. In the case of the relative humidity variable, its daily records are given as a percentage, so the value taken by an epidemiological week is the average of the daily records in that week. Therefore, the values registered in an epidemiological period are the average of the values of four consecutive epidemiological weeks.

Figure 2 shows the time series of the two climate variables. On the top side you can see the relative humidity records for each epidemiological period; on the bottom side, you can see the values corresponding to the precipitation. The behavior observed in the two exogenous variables reflects the well-defined seasonal pattern that they have. In the department of Córdoba, starting in the month of April (approximately in epidemiological period 5), the rainy season begins in most of the regions, and intense rain continues during the second and third quarters, that is, approximately until epidemiological period 10; the rest of the periods are characterized by a drought. In addition, the department of Córdoba is characterized by maintaining conditions of a high relative humidity throughout the year, ranging between 77 and 85%; the humidity is higher between the months of April and November (that is, between epidemiological periods 5 and 12) and lower in the other epidemiological periods. Additionally, the direct relationship between the relative humidity and precipitation is observed, that is, a greater amount or days with rain leads to an increase in the relative humidity, with the opposite case occurring in months of low rainfall.

2.2. SARIMAX model

The SARIMA model is characterized by combining seasonal and non-seasonal components in a multiplicative model, and is denoted $SARIMA(p, d, q)(P, D, Q)[S]$, where p is the order of the autoregressive part, d is the number of differences necessary for the series to be stationary, q is the order of the part of moving average, P corresponds to the seasonal autoregressive order, D is the seasonal differentiation order, Q is the order of the seasonal moving average, and S is the periodicity of the data [42]. Equation (2.2) corresponds to the mathematical representation of the SARIMA model [10]:

$$\phi_p(B)\Phi_P(B^s)\nabla_s^D\nabla^d y_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{2.2}$$

where B corresponds to the backshift operator, $\phi_p(B) = 1 - \phi_1(B) - \cdots - \phi_p(B^p)$ is the regular autoregressive operator of order p, $\Phi_P(B^s) = 1 - \Phi_1(B^s) - \cdots - \Phi_P(B^{sP})$ is the seasonal autoregressive operator of order P, $\nabla_s^D = (1 - B^s)^D$ represents the seasonal difference ($D = 1$ if there is seasonality and $D = 0$ otherwise), $\nabla^d = (1 - B)^d$ represents the regular difference, $\theta_q(B) = 1 - \theta_1(B) - \cdots - \theta_q(B^q)$ is the seasonal moving average operator of order q, $\Theta_Q(B) = 1 - \Theta_1(B^s) - \cdots - \Theta_Q(B^{sQ})$ is the seasonal moving average operator of order Q, and a_t is a white noise process.

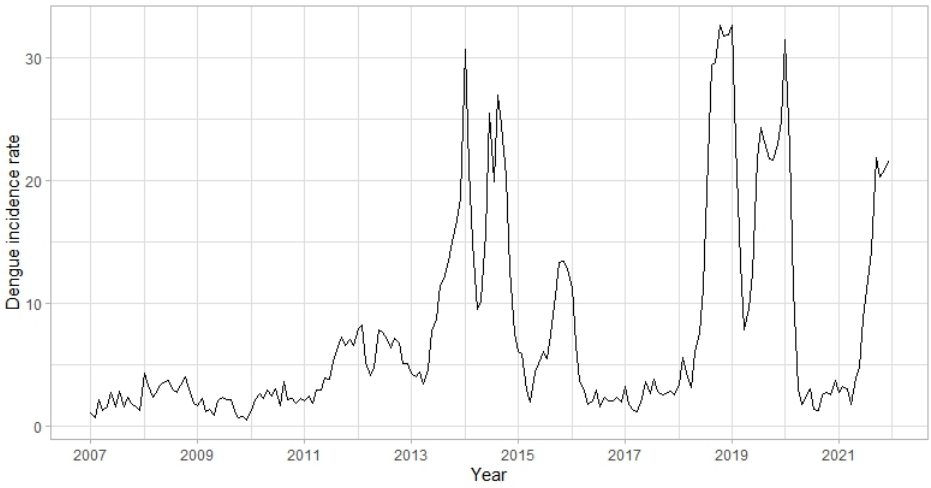


Figure 1. Time series of the dengue incidence rate per 100,000 inhabitants in the department of Córdoba, Colombia (from the epidemiological period 01/2007 to the epidemiological period 13/2021).

When exogenous variables are included in the SARIMA model, the $SARIMAX(p, d, q)(P, D, Q)_{[S]}$ model is obtained [43]:

$$\phi_p(B)\Phi_P(B^s)\nabla_s^D\nabla^d y_t = \theta_q(B)\Theta_Q(B^s)a_t + \beta\mathbf{X}_t \tag{2.3}$$

where β represents the vector of parameters of the exogenous variables $\mathbf{X}_t$, and the other terms are defined analogously to Eq (2.2).

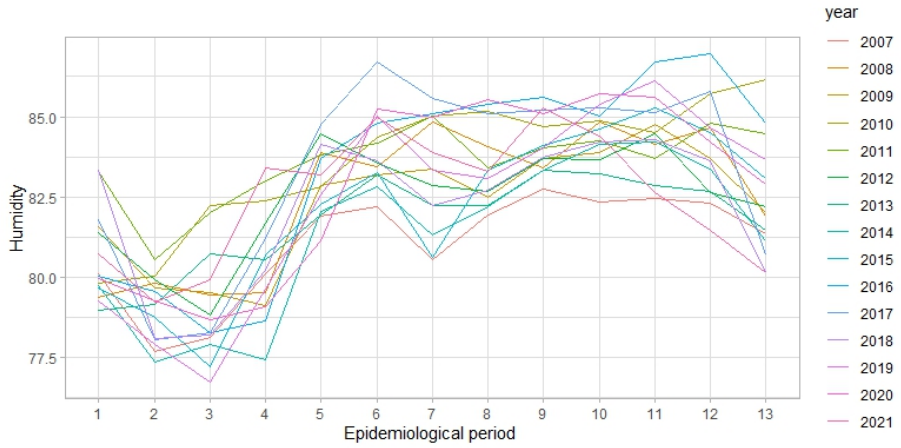
2.3. Methodology

In this section, we present the proposed methodology to forecast the dengue incidence rate per 100,000 inhabitants in the department of Córdoba for different epidemiological periods. The methodology is characterized by applying a rigorous statistical process under a cross-validation approach, as well as carrying out a prescriptive analysis of the performance of the forecasting model of the incidence rate of dengue according to the degree of knowledge of the values of the exogenous variables in the epidemiological periods, for which it is desired to forecast the incidence rate of dengue. We begin by explaining each stage of the proposed methodology.

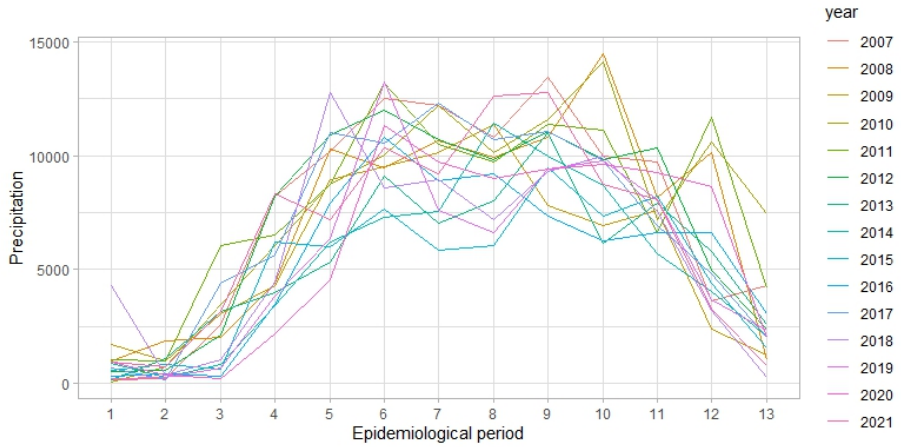
2.3.1. Preliminary analysis

Initially, both the time series of the dengue incidence rate and the series of exogenous variables are cleaned. The outliers are smoothed by linear interpolation using the R software function *tsclean()* [44]. Then, the stationarity of the dengue incidence rate is determined using the Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests [45–47]. In addition, the degree of relationship between the dengue incidence rate and exogenous variables is quantified using the cross-correlation technique.

Before proceeding with the adjustment and selection of the model, the smoothed values of the dengue incidence rate, relative humidity, and precipitation are normalized in the interval [0, 1], using



(a) Relative humidity.



(b) Precipitation.

Figure 2. Time series of the exogenous variables.

the linear scaling technique [48], that is

$$y_{Norm} = \frac{y - y_{min}}{y_{max} - y_{min}},$$

where y is a generic notation for the value of a variable and y_{min} and y_{max} are the minimum and maximum values reported for that variable, respectively.

2.3.2. Cross-validation

The degree of temporal dependence presented by the observations of a time series makes it impossible to directly apply the standard cross-validation approach because an alteration of the temporal order of the observations can lead to obtaining unrealistic predictions [49]. For this reason, a modification of the cross-validation technique, called the out-of-sample (OSS) method, is applied to the normalized time series of the dengue incidence rate, which consists of dividing the data set $y_1, \ldots, y_n$ in two parts: (i) the n_{train} initial data $y_1, \ldots, y_{n_{train}}$, called training set, is used to identify and adjust the SARIMAX model; and (ii) the n_{test} subsequent observations $y_{n_{train}+1}, \ldots, y_n$, called test set, is used to evaluate the accuracy of the model to predict new data.

The goal is to forecast H future values $y_{n+h}; h = 1, \ldots, H$, where H is the forecast horizon. For this, the test set size n_{test} to consider is as large as the defined prediction horizon (i.e., $n_{test} = H$) [50]. Furthermore, different combinations of training and test set sizes, namely $n_{train} - n_{test}$, are adopted in this study to evaluate a moving forecast origin determined by the forecast horizon values H; thus, this avoids the potential for bias that arises from arbitrarily selecting a single training and testing set [51]. To establish the values of H, it is considered that the SARIMAX models are suitable to obtain short-range forecasts [50]. In total, three configurations $n_{train} - n_{test}$ are established: (i) The period of the time series is taken as the value of H; (ii) the size of the test set is reduced to half the period value; and (iii) H is reduced to a quarter of the period value of the series.

2.3.3. Identification and fitting of the model.

The identification of the order of an initial model *SARIMAX* is performed using the normalized values of the dengue incidence rate that belong to the training data set, and by applying the strategy suggested by [50]:

- **Step 1:** The functions *ndiffs* and *nsdiffs* from the *forecast* package of the R software [44] are used to determine the number of regular differences (d) and the number of seasonal differences (D) so that the time series of the response variable is stationary.

- **Step 2:** The graphs of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the response variable are made and examined:
  - Use the ACF graph to determine the values of q and Q: the number of consecutive significant lags corresponds to the value of q , and the number of consecutive significant lags each S indicates the order Q.
  - Use the PACF chart to determine the value of p and P: take p as the number of significant consecutive lags and P as the number of consecutive significant lags significant every S lags.

Once the possible order of the SARIMAX model has been identified, we proceed to adjust it using the *Arima()* function of the *forecast* [44] package.

2.3.4. Performance evaluation in the training set

A performance evaluation is used to establish the effectiveness of the model in forecasting future values. One way to measure the efficiency is through computing the difference between the N actual values y_i and the predicted values $\hat{y}_i$, using various metrics. Among these metrics, the following stand out [52]:

- Mean error (ME),

$$ME = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i). \tag{2.4}$$

- Root Mean Square Error (RMSE),

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{2.5}$$

- Mean Average Error (MAE),

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{2.6}$$

- Mean absolute percentage error (MAPE),

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%. \tag{2.7}$$

- Akaike Information Criterion (AIC),

$$AIC \approx ln(\hat{\sigma}_a^2) + r\frac{2}{n} \tag{2.8}$$

- Bayesian Information Criterion (BIC),

$$BIC = ln(\hat{\sigma}_a^2) + r\frac{ln(n)}{n} \tag{2.9}$$

where $\hat{\sigma}_a^2$ is the maximum likelihood estimate of σ_a^2 (variance of model's residuals) and $r = p + q + 1$ is the number of estimated parameters of the model.

Before calculating the metrics mentioned above, it is important to denormalize the forecasts obtained for the dengue incidence rate using the min-max denormalization technique described in [48]. The denormalized forecasts are used together with the clean dengue incidence rate data to evaluate the performance of the model. In this stage, the diagnosis of the model is also examined at a significance level of 5%. The independence and normality of the residuals are determined using the Ljung-Box test and the Anderson-Darling test, respectively [10, 53]. Likewise, it is verified that the residuals are centered at zero, by means of a T test.

In order to apply the previous tests, it is desirable that some assumptions are satisfied. However, in [54], the authors demonstrated that linear time series models are robust to the violation of the normality assumption when it comes to hypothesis testing and an estimation of the parameters, as long as the outliers are treated correctly. Given that an adequate treatment is given to the outliers in the preliminary analysis of the series and in the methodological process proposed in this article, then we give priority to the compliance of independence and zero mean assumptions in the diagnosis of the residuals.

A model is considered valid if it presents the lowest value in most of the performance metrics, and satisfies the greatest number of assumptions. However, if the model does not present a good performance, then we proceed to adjust it again by sequentially increasing the values of p, q, P, and Q, thereby leaving the initial values of d and D until there is an improvement in the performance of the model. Finally, a set of candidate models is selected and these models are evaluated with the test data.

2.3.5. Forecasting and evaluation of performance with the test set

At this stage, three possible scenarios are considered by taking the availability the values of the exogenous variables during the study period into account. Here, the main idea is to foresee what impact they would have on the performance of the models and on decision making. Specifically, we predict the dengue incidence rate in the test set for each of the following scenarios:

- **Scenario 1:** When the values of the exogenous variables are known in advance.
- **Scenario 2:** When the values of the exogenous variables are not known, and it is decided to forecast these using time series models as a stage prior to modeling the response variable.
- **Scenario 3:** When the values of the exogenous variables are not known, and it is decided to use the average of the historical data. For this, the averages obtained for the exogenous variable in each epidemiological period should be compared. If there is no significant difference between them, then the average of the historical data of the series is taken as a representative value. If there is a significant difference between the averages, then the averages of the periods to be forecast must be considered as the values of the exogenous variables.

For each scenario, the test data is forecast using the candidate models, and performance metrics are calculated as described in Subsection 2.3.4.

2.3.6. Model selection

Finally, for each scenario, the best model is chosen as the one that has presented the best values in the performance metrics in both the training and test sets. In addition, the consistency of the results obtained in the metrics for the two data sets is taken into account.

3. Results

The results obtained by applying the methodology described in Section 2 are presented below.

3.1. Preliminary analysis of the incidence rate of dengue in Córdoba, Colombia

After cleaning the series of the dengue incidence rate, we obtained a series with smoother peaks. The Phillips-Perron (PP) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) stationarity tests show that it is not stationary in the mean. Using the *ndiffs* function, we obtained that it is enough to differentiate the time series once to make it stationary. This result is verified by applying the PP and KPSS tests, at a significance level of 5%, to the differentiated series. We obtained the p-values of 0.01 and 0.1, respectively. Subsequently, the existence of a linear relationship between the dengue incidence rate, the relative humidity, and the precipitation variables was determined using the cross-correlation coefficient and its respective significance at 5%. Table 2 shows the results of the cross-correlation coefficients between the variable of interest and the exogenous variables, where significant lags are in bold.

It is observed that the cross correlation between each of the exogenous variables and the incidence rate has a high and significant coefficient at a lag of 0, which means that the dengue incidence rate per period either increases or decreases simultaneously with the precipitation and humidity. Additionally, it is obtained that lags 1, 12, and 13 of the cross correlation of each of the exogenous variables and the incidence rate are positive and significant. This means that when the relative humidity or precipitation increases in an epidemiological period, the incidence rate of dengue will increase one, twelve, or thirteen periods later. For example, if the relative humidity increases in the fourth period, then the incidence rate of dengue is expected to increase in the fifth period. This is contrary to what happens in lags 6 to 8, where the incidence rate will decrease by 6 or 8 more periods.

The significant cross-correlations identified in this study support the fact that the increase in dengue cases is associated with rainy periods. A seasonal behavior is identified, where the persistence of

dengue cases occurs approximately between periods 5 and 10, which correspond to the months of rainy seasons and a high relative humidity in the department of Córdoba. In addition, a sustained increase is highlighted in the number of cases in weeks 10 and 11, that is, mid-September and the entire month of October. (see Figure 3).

Table 2. Cross-correlation coefficients between the dengue incidence rate and the climate variables: Relative humidity and precipitation. Values in bold indicate that the estimated coefficient is statistically significant at the 5% level.

Lag	Humidity	Precipitation
0	**0.306**	**0.300**
1	**0.224**	**0.237**
2	0.084	0.111
3	0.061	0.006
4	0.072	-0.040
5	-0.093	-0.192
6	**-0.212**	**-0.250**
7	**-0.308**	**-0.259**
8	**-0.345**	**-0.200**
9	-0.296	-0.068
10	-0.172	0.031
11	0.030	0.163
12	**0.165**	**0.223**
13	**0.228**	**0.199**

3.2. Cross-validation

To determine the configuration of the two sets (training and test), we consider the fact that an epidemiological year contains 13 epidemiological periods. Therefore, the maximum number of epidemiological periods will initially be considered as the forecasting horizon, that is, $H = 13$. Additionally, in order to examine the performance of the model in low range time horizons, this value was halved, alongside considering the horizons $H = 6$ and $H = 3$ per epidemiological methods. This lead us to the evaluation of three configurations of sizes for the training set-test set ($n_{train} - n_{test}$): 182-13, 189-6, and 192-3.

3.3. SARIMAX model identification and fitting

For the three training sets size $n_{train} = \{182, 189, 192\}$, it is found that the normalized dengue incidence rate time series become stationary when they are differentiated regularly once. Furthermore, it is not necessary to make differences for the seasonal part. Therefore, we assume the orders $d = 1$ and

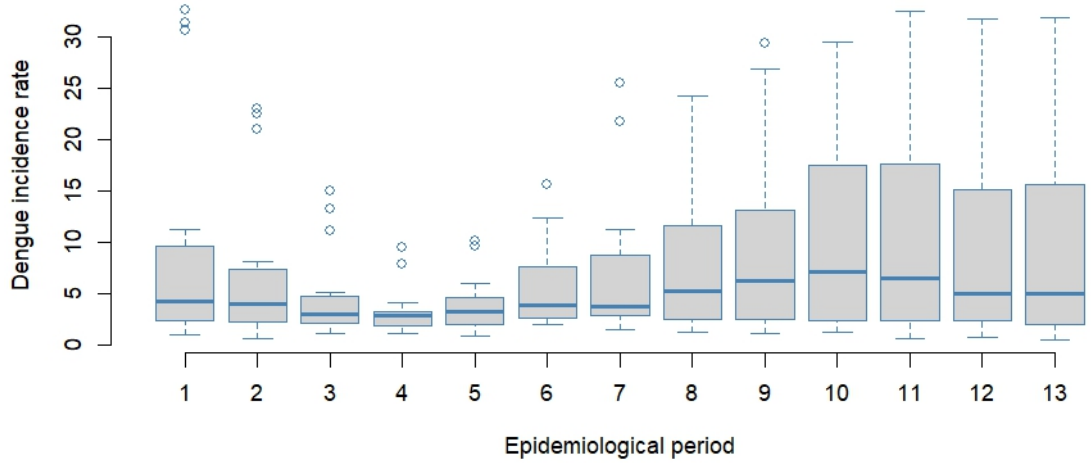


Figure 3. Box plot of the dengue incidence rate per 100,000 inhabitants in the department of Córdoba, Colombia by epidemiological period.

$D = 0$ for the SARIMAX model. In addition, we have that $S = 13$, which corresponds to the number of epidemiological periods per year.

The ACF and PACF graphs of the first training data set is shown in Figure 4. All the three configurations share the same pattern in the ACF and PACF. This entails that the same initial model is selected for each configuration, as follows: a first significant lag in the PACF, which indicates that $p = 1$ in the non-seasonal part, and no significant lag in the seasonal part, which indicates that $P = 0$. With regard to the ACF, a first significant lag is found, which leads to the value of $q = 1$ and the 13th significant lag for a seasonal MA(1) (i.e., $Q = 1$). Therefore, the initial model in all three configurations is a $SARIMAX(1, 1, 1)(0, 0, 1)_{[13]}$.

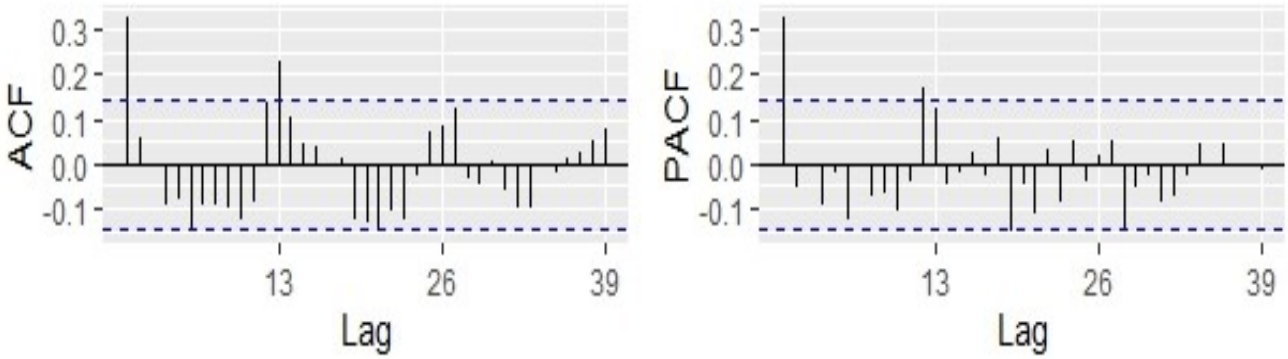


Figure 4. ACF and PACF of the differentiated dengue incidence rate ($d = 1$), for the 182-13 configuration.

3.4. Performance evaluation with the training set

Starting from the initial model determined in the previous stage, for each of the three training sets, different models were fitted by gradually varying the values of p, q, P, and Q until no significant gain was obtained for the performance metrics. Table 3 contains the fitted models with their respective performance evaluations and the p-value associated with the Ljung-Box test for the first configuration 182-13. It is observed that of the six fitted models, the SARIMAX $(1, 1, 1)(0, 0, 1)_{[13]}$ model is the one with the largest number of good results for the proposed metrics. In addition, it is found that the residuals of all the models satisfy the assumption of independence, since we obtain a p-value >0.05 in the Ljung-Box test, and they are centered at zero (since in all the T tests a p-value<0.05 was obtained). Although no set of residuals is found to satisfy the normality assumption, the models can still be used to forecast, as established in [54]. On the other hand, the adjusted models for the 189-6 configuration are presented in Table 4. It is observed that for this training set, the model SARIMAX $(4, 1, 3)(1, 0, 2)_{[13]}$ is the one that presents the best results for the proposed metrics. Similar to the previous configuration, it is verified that all the residuals of the models satisfy only the assumptions of independence and a zero mean. Finally, for the last configuration, most of the models are different from those adjusted in the previous configurations (see Table 5). In this case, the model SARIMAX $(4, 1, 3)(1, 0, 3)_{[13]}$ is the one that has the largest number of lowest values for the proposed metrics. In this case, we found that the residuals of the models satisfy the diagnostic tests of independence and center at zero.

Table 3. Performance statistics of the fitted models for the training data set of the 182-13 configuration, and the corresponding p-value of the Ljung-Box test applied to the residuals of each model. The values in bold correspond to the best results obtained for each performance metric and the highest p-value obtained in the independence test.

Model	ME	RMSE	MAE	MAPE	AIC	BIC	Ljung-Box
SARIMAX$(1, 1, 1)(0, 0, 1)_{[13]}$	**0**	1.926	**1.263**	29.937	764.682	**783.873**	**0.589**
SARIMAX$(2, 1, 1)(1, 0, 1)_{[13]}$	-0.0023	1.924	1.276	30.354	768.183	793.771	0.422
SARIMAX$(3, 1, 2)(0, 0, 2)_{[13]}$	-0.0059	1.886	1.279	30.366	766.405	798.390	0.513
SARIMAX$(4, 1, 3)(1, 0, 2)_{[13]}$	0.139	1.842	1.271	**29.621**	766.037	807.617	0.421
SARIMAX$(3, 1, 4)(1, 0, 2)_{[13]}$	0.159	1.873	1.286	29.915	770.597	812.178	0.309
SARIMAX$(4, 1, 3)(0, 0, 2)_{[13]}$	0.143	**1.827**	1.268	30.196	**762.946**	801.328	0.246

3.5. Forecasting and evaluation of performance with the test set

The models selected in the previous stage for each configuration are used to predict values of the dengue incidence rate and for the respective size of the test set. The values of the exogenous variables were included in the SARIMAX models to calculate the forecasts of the test sets, under the three scenarios previously defined in the methodology.

- **Scenario 1:** The normalized values of precipitation and relative humidity corresponding to the last $n_{test} = \{3, 6, 13\}$ epidemiological periods of the database are considered.
- **Scenario 2:** Using their respective normalized training data, a SARIMA model is fitted for each exogenous variable, and the $n_{test} = \{3, 6, 13\}$ values are predicted for the precipitation

Table 4. Performance statistics of the fitted models for the training data set of the 189-6 configuration, and the corresponding p-value of the Ljung-Box test applied to the residuals of each model. The values in bold correspond to the best results obtained for each performance metric and the highest p-value obtained in the independence test.

Model	ME	RMSE	MAE	MAPE	AIC	BIC	Ljung-Box
SARIMAX$(1,1,1)(0,0,1)_{[13]}$	0.0378	1.922	1.261	29.782	792.685	**812.103**	**0.603**
SARIMAX$(1,1,1)(0,0,3)_{[13]}$	0.0324	1.917	1.265	29.999	795.921	821.813	0.547
SARIMAX$(3,1,2)(0,0,2)_{[13]}$	**0.0314**	1.901	1.254	30.122	796.909	829.273	0.482
SARIMAX$(4,1,3)(1,0,2)_{[13]}$	0.158	1.824	**1.245**	**29.362**	**790.496**	832.570	0.362
SARIMAX$(3,1,4)(1,0,2)_{[13]}$	0.149	1.830	1.260	29.553	791.430	833.503	0.418
SARIMAX$(4,1,3)(1,0,3)_{[13]}$	0.153	**1.802**	1.259	29.504	793.658	838.968	0.294

Table 5. Performance statistics of the fitted models for the training data set of the 192-3 configuration, and the corresponding p-value of the Ljung-Box test applied to the residuals of each model. The values in bold correspond to the best results obtained for each performance metric and the highest p-value obtained in the independence test.

Model	ME	RMSE	MAE	MAPE	AIC	BIC	Ljung-Box
SARIMAX$(1,1,1)(0,0,1)_{[13]}$	0.0847	1.966	1.304	30.188	813.887	**833.400**	0.428
SARIMAX$(1,1,1)(1,0,1)_{[13]}$	**0.0641**	1.930	1.290	29.711	**811.207**	833.973	0.331
SARIMAX$(1,1,1)(1,0,2)_{[13]}$	0.0668	1.926	1.290	29.771	812.403	838.421	0.384
SARIMAX$(1,1,1)(0,0,3)_{[13]}$	0.0799	1.961	1.310	30.463	817.152	843.171	0.351
SARIMAX$(1,1,1)(1,0,3)_{[13]}$	0.0667	1.904	1.269	29.147	811.564	840.835	**0.445**
SARIMAX$(4,1,3)(1,0,1)_{[13]}$	0.159	**1.855**	**1.263**	**28.869**	812.282	857.814	0.173

and relative humidity. Thus, the fitted models were SARIMA $(1,0,3)(1,1,2)_{[13]}$ and SARIMA $(0,0,1)(2,1,2)_{[13]}$ for the humidity and precipitation variables, respectively.

- **Scenario 3:** Figure 2 shows that the precipitation and relative humidity are not constant in the 13 periods. Therefore, it is necessary to use the training data of the exogenous variables to determine the average precipitation and the average relative humidity of each of the epidemiological periods considered in the test data.

Once the forecasts obtained for each combination of test set and scenario have been denormalized, the performance metrics proposed in Subsection 2.3.4 are calculated.

3.6. Model selection

The models selected in the previous stage for each configuration are used to forecast their respective forecast horizons under the considerations of each scenario. At the end, nine models are adjusted (3 scenarios for each of the 3 configurations). Table 6 contains the selected SARIMAX models for each combination of test set size $H = \{3, 6, 13\}$ (which is equivalent to the forecast horizon) and scenario related to the exogenous variables in the model, as well as the values of the performance metrics

obtained for each test data set of the dengue incidence rate. The best model is chosen as the one that presents the lowest in the performance metrics.

Table 6 shows that the order of the SARIMAX model varies according to the forecast horizon (H). In general, it is observed that the values of the performance metrics improve as the horizon decreases (as it is common in this type of studies). The obtained forecasts are more accurate for three epidemiological periods, which is equivalent to 12 weeks. On the other hand, regardless of the test set, the use of predicted values for the precipitation and relative humidity variables improves the performance of the SARIMAX model in the test set with respect to the other two scenarios considered. This is a relevant result of this article.

Based on the results, the best model is the SARIMAX$(4, 1, 3)(1, 0, 1)_{[13]}$ under the 192-3 configuration and Scenario 2, where its mathematical formulation can be obtained from Eq (2.3) as follows:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)(1-\Phi_1 B^{13})z_t=(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)(1+\Theta_1 B^{13})a_t+\beta_1 X_{Norm_{1t}}+\beta_2 X_{Norm_{2t}}, \tag{3.1}$$

where $z_t = y_{Norm_t} - y_{Norm_{t-1}}$ denotes the normalized differentiated series regularized once, and $X_{Norm_{1t}}$ and $X_{Norm_{2t}}$ denote the normalized values of the relative humidity and precipitation at time t, respectively. The general expression to obtain the forecasting of the normalized dengue incidence rate (y_{Norm_t}), is obtained by the product of the autoregressive polynomials and the moving averages in Eq (3.1), thereby applying the lag operator and solving for the variable y_{Norm_t}. Thus, one obtains the following adjusted model:

$$\begin{aligned}\hat{y}_{Norm_t} =& \hat{\phi}_1 z_{t-1} + \hat{\phi}_2 z_{t-2} + \hat{\phi}_3 z_{t-3} + \hat{\phi}_4 z_{t-4} + \hat{\Phi}_1 z_{t-13} - \hat{\phi}_1\hat{\Phi}_1 z_{t-14} - \hat{\phi}_2\hat{\Phi}_1 z_{t-15} - \hat{\phi}_3\hat{\Phi}_1 z_{t-16} - \hat{\phi}_4\hat{\Phi}_1 z_{t-17} \\ &+ \hat{\theta}_1 e_{t-1} + \hat{\theta}_2 e_{t-2} + \hat{\theta}_3 e_{t-3} + \hat{\Theta}_1 e_{t-13} + \hat{\theta}_1\hat{\Theta}_1 e_{t-14} + \hat{\theta}_2\hat{\Theta}_1 e_{t-15} + \hat{\theta}_3\hat{\Theta}_1 e_{t-16} \\ &+ \hat{\beta}_1 X_{Norm_{1t}} + \hat{\beta}_2 X_{Norm_{2t}} + y_{Norm_{t-1}}\end{aligned} \tag{3.2}$$

where e_t denotes the $i-th$ residual and the parameter estimates are shown in Table 7 (which are all significant at the 5% level). It is important to keep in mind that to obtain the forecasts in the SARIMAX models, it is assumed that $e_{T+h} = 0; h = 1, \ldots; H$ [50]. In addition, since the variable was normalized, then it is necessary to apply the following equation to obtain the estimate of the dengue incidence rate ($\hat{y}_t$):

$$\hat{y}_t = \hat{y}_{Norm_t} * (y_{max} - y_{min}) + y_{min}.$$

In order to illustrate the use of the model, suppose that it was trained with the first 192 observations, and it is of interest to forecast the number of dengue cases at time $t = 193$ (which corresponds to period 11 of the year 2021). To do this, the normalized series of the incidence rate is first differentiated once; then, from the resulting series, one selects the following positions: $z_{t-1} = z_{192} = 0.2353, z_{t-2} = z_{191} = 0.07846, z_{t-3} = z_{190} = 0.09515, z_{t-4} = z_{189} = 0.1235, z_{t-13} = z_{180} = 0.006756, z_{t-14} = z_{179} = 0.04054, z_{t-15} = z_{178} = -0.003378, z_{t-16} = z_{177} = -0.05405$, and $z_{t-17} = z_{176} = 0.02027$. In addition, suppose that one obtains the following residuals in the estimation process of the model: $e_{t-1} = e_{192} = 0.1172, e_{191} = -0.6105, e_{190} = -0.2910, e_{180} = -0.3557, e_{179} = -0.3447, e_{178} = -0.3982$, and $e_{177} = -0.4284$. Moreover, it is known that $y_{Norm_{192}} = 0.6657$, and also assume that the normalized real humidity and precipitation have been forecast for $t = 193$ as follows: $X_{Norm_{1t}} = 0.5776$ y $X_{Norm_{21t}} = 0.4797$, respectively.

Table 6. Performance statistics of the selected model for each of the test data sets $h = \{3, 6, 13\}$, under the conditions of each of the three prescriptive scenarios considered.

H	Scenario	Model	ME	RMSE	MAE	MAPE	AIC	BIC
13	1	SARIMAX$(1,1,1)(0,0,1)_{[13]}$	7.246	10.489	7.832	58.752	122.000	128.780
	2	SARIMAX$(1,1,1)(0,0,1)_{[13]}$	7.416	10.521	7.816	56.677	124.080	131.424
	3	SARIMAX$(1,1,1)(0,0,1)_{[13]}$	7.460	10.560	7.849	56.817	124.175	131.520
6	1	SARIMAX$(4,1,3)(1,0,2)_{[13]}$	8.021	8.667	8.021	41.168	54.941	53.692
	2	SARIMAX$(4,1,3)(1,0,2)_{[13]}$	7.827	8.561	7.827	39.776	54.794	53.545
	3	SARIMAX$(4,1,3)(1,0,2)_{[13]}$	7.894	8.596	7.894	40.267	54.843	53.593
3	1	SARIMAX$(4,1,3)(1,0,1)_{[13]}$	-2.285	2.395	2.285	11.023	41.754	29.135
	2	SARIMAX$(4,1,3)(1,0,1)_{[13]}$	-2.274	2.386	2.274	10.972	41.732	29.113
	3	SARIMAX$(4,1,3)(1,0,1)_{[13]}$	-2.372	2.496	2.372	11.448	42.002	29.382

Table 7. Estimation of the parameters of the best model SARIMAX$(4,1,3)(1,0,1)_{[13]}$, with their respective standard errors.

Parameter	Estimation	s.e.
ar1(ϕ_1)	-0.896	0.0790
ar2(ϕ_2)	-0.249	0.0793
ar3(ϕ_3)	-0.433	0.0783
ar4(ϕ_4)	0.742	0.0764
ma1(θ_1)	0.443	0.0431
ma2(θ_2)	0.365	0.0257
ma3(θ_3)	0.172	0.0405
sar1(Φ_1)	-0.644	0.0636
sma1(Θ_1)	0.643	0.0504
Relative humidity (β_1)	0.695	0.0104
Precipitation(β_2)	0.980	0.0801

Then, by replacing these values in Eq (3.2), one obtains that $\hat{y}_{Norm_{193}} = 0.6663$. Thus, we can compute the expected incidence rate of dengue in Córdoba per 100,000 inhabitants for the 11th period of the year 2021:

$$\hat{y}_t = 0.6663 * (32624.82 - 494.32) + 494.32 = 21903.2338.$$

Notice that once the forecast of the dengue incidence rate is computed, the estimate of dengue cases can be obtained. It is enough to solve for the variable *Cases period* in Eq (2.1), and replace the

respective values of the population size and the predicted rate. For our particular example, it is known that the size of the estimated population is $N_{193} = 1,864,336$. Thus, the total number of dengue cases reported for the 11th period of the year 2021 is given by the following

$$NCasos_{193} = 21903.2338 \times 1,864,336/100,000 \approx 408.$$

Additionally, the prediction intervals are obtained following the Bootstrap scheme proposed by [50]. In Figure 5, these intervals can be seen for periods 11, 12, and 13 of the year 2021, under Scenario 2, and the adjusted model given by Eq (3.2).

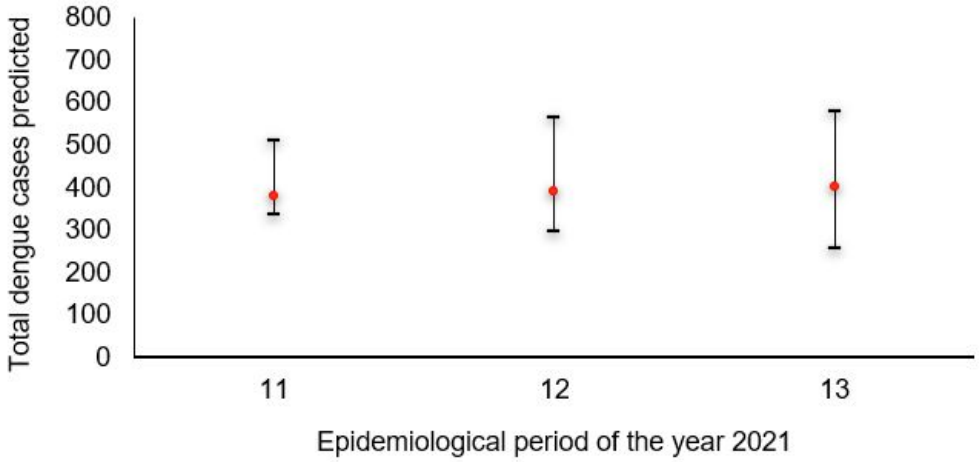


Figure 5. Prediction intervals of total dengue cases for the epidemiological periods $\{11, 12, 13\}$ of the year 2021, using the best model SARIMAX$(4, 1, 3)(1, 0, 1)_{[13]}$ of scenario 2 and the 192-3 configuration. The red dot in each interval denotes the actual number of reported dengue cases.

4. Discussion and conclusions

In this work, we proposed a methodological framework to forecast the incidence rate of dengue in the department of Córdoba, Colombia. The appropriate forecast horizon is determined between $H = \{3, 6, 13\}$ for the incidence rate of dengue by adjusting and evaluating various SARIMAX models for each data configuration determined in the cross-validation stage.

The use of the SARIMAX model to forecast the incidence rate of dengue requires us to not only consider the historical data of this rate, but the values that the exogenous variables will take in the different periods of the forecast. Thus, in this work, we investigated the impact that the values assumed for the exogenous variables have on the performance of a SARIMAX model. It was found that the use of real values of the exogenous variables led to better results in the performance metrics of the adjusted SARIMAX models, regardless of the size of the forecast horizon. However, in practice, usually the future real values of the exogenous variables are unknown before the forecast. Thus, we considered two options to estimate the exogenous values from the available historical information. We found that when the time series models were correctly adjusted for the relative humidity and precipitation

variables, and the predicted values (of the exogenous variables) were used as inputs to the SARIMAX model, then better results were obtained compared to when the averages of the historical values of the exogenous variables were taken as the input variables. The results show the following: *(i)* the values of the performance metrics of the models of configurations 189-6 and 192-3 turned out to be consistent, when evaluating them for both training and test sets; and *(ii)* the residuals of all fitted models satisfied the test of independence and a zero mean. The results of this research are in accordance with what was found in [55]. These results indicate that among the arbovirus prediction models published in the literature, stochastic models, such as SARIMAX, are suitable tools for capturing trends, seasonal changes and random distortions in historical series. Likewise, it is shown that the use of climate variables was essential to obtain a numerical prediction of dengue cases with a greater accuracy.

Regarding the climate variables, it was identified that the relative humidity and precipitation had significant effects on the incidence rate of dengue in the department of Córdoba, Colombia. The results of this paper support the findings presented in [56], which show that there is a high and significant correlation between dengue cases, precipitation, and relative humidity. This can be explained because the Aedes aegypti mosquito lays its eggs in clean water, such as rainwater, and precipitation and the relative humidity are vital for the survival of the mosquito, in the juvenile and adult stages, respectively. Furthermore, the department of Córdoba has favorable environments for the presence of the Aedes aegypti mosquito: most of its territory is in the tropical rainy zone, which has an average annual precipitation of 1262 mm, an average temperature of $27.8°C$, and a relative humidity between 76 and 82%.

The dynamics of the disease from 2007 to 2021 showed that despite the actions carried out by the Departmental Health Secretary of Córdoba, there was a significant increase in dengue cases in the department in these periods. This means that surveillance and prevention actions should be improved. To do this, departmental health institutions could rely on research carried out by research institutions such as universities. Specifically, with the results of this research, the epidemiological behavior of the disease can be known and the total number of dengue cases can be predicted with a high precision in at least 3 epidemiological periods, that is, in 12 calendar weeks. For example, if it is assumed that we only have access to information from the period 2007-01 to 2021-10, and we are interested in knowing the total number of dengue cases for the remainder of the year 2021, then the best adjusted model for Scenario 2 could be used with the 192 training data to predict the incidence rates of dengue in the epidemiological periods 11,12,13 of the year 2021 and the total number of dengue cases for those periods. Figure 5 shows the prediction intervals obtained for the periods 11,12 and 13 of the year 2021. It is observed that all the intervals contained the real value, which shows that the model provides timely and reliable information for decision making, and could help establish the necessary policies to prepare for future dengue cases in advance.

It is important to remark that for this model to remain updated, it is necessary to have timely access to information on dengue cases, population size, and exogenous variables, which could be achieved with the support of government entities. Furthermore, the following is of interest: (i) to examine the effect that other climate variables, such as temperature, might have, of which it is known that the range $18°C$ to $31°C$ is optimal for mosquito incubation and survival, while the average temperature has a minor effect on dengue transmission in tropical countries because they have relatively constant temperatures [56]; (ii) propose new studies that model dengue cases at the level of municipalities in the department of Córdoba, thereby considering variables such as the number of prevention campaigns,

socioeconomic variables, and population size, among others; and (iii) consider hybrid or artificial neural network models, which have proven to have good performances [26]. The execution of these new investigations is subject to the availability of the information. Similar to other works, the methodology presented in this work has limitations. Some of these limitations are highly related to the availability of data. However, there are contributions that help to get a better insight into the topic. Finally, as future work of this research, we propose a methodology to identify and estimate the model parameters through the use of metaheuristics, such that these processes can be simplified so that any non-statistician can make use of the model.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. *World Health Organization*, Dengue guias para el diagnóstico, tratamiento, prevención y control: nueva edición, 2009. https://apps.who.int/iris/handle/10665/44504
2. *Pan American Health Organization*, Dengue, 2022. https://www.paho.org/en/topics/dengue
3. K. Fatima, N. I. Syed, Dengvaxia controversy: impact on vaccine hesitancy, *J. Glob. Health*, **8** (2018). https://doi.org/10.7189/jogh.08-020312
4. A. C. Jaramillo, Infecciones por arbovirus, *Rev. MVZ Córdoba*, **5** (2000), 51–56.
5. M. F. Suárez, M. J. Nelson, Registro de altitud del aedes aegypti en Colombia, *Biomédica*, **1** (1981), 1–225.
6. J. C. Castrillón, J. C. Castaño, S. Urcuqui, Dengue en Colombia: diez años de evolución, *Rev. chilena de infectología*, **32** (2015), 142–149.
7. B. Castillo, J. Castillo, M. Salas, El dengue y su incidencia en la salud de niños y adolescentes frente a la economía colombiana, *Rev. Med. - Facultad de Ciencias de la Salud - Programa de Medicina*, **14** (2015), 38–44.
8. A. M. Sánchez, Análisis de la respuesta del estado colombiano frente al fenómeno de la niña 2010–2011 : el caso de Santa Lucía, Technical report, 2014.
9. V. N. Valencia, Y. Díaz, J. M. Pascale, M. F. Boni, J. E. Sanchez-Galan, Assessing the effect of climate variables on the incidence of dengue cases in the metropolitan region of panama city, *Int. J. Environ. Res. Public Health*, **18** (2021), 12108. https://doi.org/10.3390/ijerph182212108
10. G. Box, G. Jenkins, G, Reinsel, G. Ljung, *Time Series Analysis Forecasting and Control*, John Wiley & Sons, Inc, 5th edition, 2016.
11. Z. Li, H. Gurgel, L. Xu, L. Yang, J. Dong, Improving dengue forecasts by using geospatial big data analysis in google earth engine and the historical dengue information-aided long short term memory modeling, *Biology*, **11** (2022), 169. https://doi.org/10.3390/biology11020169

12. C. M. Benedum, K. M. Shea, H. E. Jenkins, L. Y. Kim, N. Markuzon, Weekly dengue forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore, *PLOS Neglect. Trop. D.*, **14** (2020), e0008710. https://doi.org/10.1371/journal.pntd.0008710

13. R. Bomfim, S. Pei, J. Shaman, T. Yamana, H. A. Makse, J. S. Andrade, et al., Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas, *J.Roy. Soc. I.*, **17** (2020), 20200691. https://doi.org/10.1371/journal.pntd.0008710

14. S. Polwiang, The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017), *BMC Infect. Dis.*, **20** (2020), 208. https://doi.org/10.1186/s12879-020-4902-6

15. L. Thiruchelvam, S. C. Dass, V. S. Asirvadam, H. Daud, B. S. Gill, Determine neighboring region spatial effect on dengue cases using ensemble Arima models, *Sci. Rep.*, **11** (2021), 5873. https://doi.org/10.1038/s41598-021-84176-y

16. F. Cortes, C. M. T. Martelli, R. X. Arraes de Alencar, U. Ramos, J. Bosco, O. Gongalves, et al., Time series analysis of dengue surveillance data in two brazilian cities, *Acta Trop.*, **182** (2018), 190–197. https://doi.org/10.1016/j.actatropica.2018.03.006

17. T. Chakraborty, S. Chattopadhyay, I. Ghosh, Forecasting dengue epidemics using a hybrid methodology, *Phys. A*, **527** (2019), 121266. https://doi.org/10.1016/j.physa.2019.121266

18. N. Zhao, K. Charland, M. Carabali, E. O. Nsoesie, M. Maheu-Giroux, E. Rees, et al., Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia, *PLOS Neglect. Trop. D.*, **14** (2020), e0008056. https://doi.org/10.1371/journal.pntd.0008056

19. L. López, A. Pulecio, G. Marcillo, Dengue cases in Colombia: Mathematical forecasts for 2018–2022, *MEDICC Rev.*, **21** (2019), 38–45.

20. A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, E. Guven, Ensemble method for dengue prediction, *PLOS ONE*, **13** (2018), e0189988. https://doi.org/10.1371/journal.pone.0189988

21. A. F. B. Gabriel, A. P. Alencar, S. G. E. K. Miraglia, Dengue outbreaks: unpredictable incidence time series, *Epidemiol. Infect.*, **147** (2019), e116. https://doi.org/10.1017/S0950268819000311

22. V. J. Jayaraj, R. Avoi, N. Gopalakrishnan, D. B. Raja, Y. Umasa, Developing a dengue prediction model based on climate in Tawau, Malaysia, *Acta Trop.*, **197** (2019), 105055. https://doi.org/10.1016/j.actatropica.2019.105055

23. W. Liang, A. Hu, P. Hu, J. Zhu, Y. Wang, Estimating the tuberculosis incidence using a SARIMAX-NNARX hybrid model by integrating meteorological factors in Qinghai Province, China, *Int. J. Biometeorol.*, **67** (2023), 55–65. https://doi.org/10.1007/s00484-022-02385-0

24. P. Manigandan, M. D. S. Alam, M. Alharthi, U. Khan, K. Alagirisamy, D. Pachiyappan, et al., Forecasting natural gas production and consumption in United States-Evidence from SARIMA and SARIMAX models, *Energies*, **14** (2021), 6021. https://doi.org/10.3390/en14196021

25. M. J. Llop, A. Gómez, P. Llop, M. S. López, G. V. Müller, Prediction of leptospirosis outbreaks by hydroclimatic covariates: a comparative study of statistical models, *Int. J. Biometeorol.*, **66** (2022), 2529–2540. https://doi.org/10.1007/s00484-022-02378-z

26. M. R. Cogollo, G. González-Parra, A. J. Arenas, Modeling and forecasting cases of RSV using artificial neural networks, *Mathematics*, **9** (2021), 2958. https://doi.org/10.3390/math9222958

27. G. González-Parra, J. F. Querales, D. Aranda, Predicción de la epidemia del virus sincitial respiratorio en Bogotá, DC, utilizando variables climatológicas, *Biomédica*, **36** (2016), 378–389. https://doi.org/10.7705/biomedica.v36i3.2763

28. N. B. D. Campos, M. H. F. Morais, A. P. R. Ceolin, M. C. M. Cunha, R. R. Nicolino, O. L.Schultes, et al., Twenty-two years of dengue fever (1996-2017): an epidemiological study in a Brazilian city, *Int. J. Env. Health Res.*, **31** (2021), 315–324. https://doi.org/10.1080/09603123.2019.1656801

29. R. Chumpu, N. Khamsemanan, C. Nattee, The association between dengue incidences and provincial-level weather variables in Thailand from 2001 to 2014, *Plos One*, **14** (2019), e0226945. https://doi.org/10.1371/journal.pone.0226945

30. F. J. Colón-González, C. Fezzi, I. R. Lake, P. R. Hunter, The effects of weather and climate change on dengue, *PLoS Neglect. Trop. Dis.*, **7** (2013), e2503. https://doi.org/10.1371/journal.pntd.0002503

31. Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, J. Rocklöv, Forecast of dengue incidence using temperature and rainfall, *PLoS Neglect. Trop. Dis.*, **6**(2012), e1908. https://doi.org/10.1371/journal.pntd.0001908

32. W. A. Abualamah, N. A. Akbar, H. S. Banni, M. A. Bafail, Forecasting the morbidity and mortality of dengue fever in KSA: A time series analysis (2006–2016), *J. Taibah Univ. Med. Sci.*, **16** (2021), 448–455. https://doi.org/10.1016/j.jtumed.2021.02.007

33. N. Mohammed, M. Z. Rahman, Forecasting dengue incidence in Bangladesh using seasonal ARIMA model, a time series analysis, in *International Conference on Machine Intelligence and Emerging Technologies*, (2022), 589–598. https://doi.org/10.1007/978-3-031-34622-4-47

34. M. Panja, T. Chakraborty, S. S. Nadim, I. Ghosh, U. Kumar, N. Liu, An ensemble neural network approach to forecast Dengue outbreak based on climatic condition, *Chaos, Solitons & Fractals*, **167** (2023), 113124. https://doi.org/10.1016/j.chaos.2023.113124

35. M. Tejaswi, V. Supritha, S. Thangam, Early prediction of dengue cases using time series model, in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, (2023), 1–6. https://doi.org/10.1109/I2CT57861.2023.10126405

36. L. A. Barboza, S. W. Chou-Chen, P. Vásquez, Y. E. García, J. G. Calvo, H. G. Hidalgo, et al., Assessing dengue fever risk in Costa Rica by using climate variables and machine learning techniques, *PLOS Neglect. Trop. Dis.*, **17** (2023), e0011047. https://doi.org/10.1371/journal.pntd.0011047

37. E. Dantas, M. Tosin, A. Cunha Jr, Calibration of a SEIR–SEI epidemic model to describe the Zika virus outbreak in Brazil, *Appl. Math. Comput.*, **338** (2018), 249–259. https://doi.org/10.1016/j.amc.2018.06.024

38. J. A. Gutierrez, K. Laneri, J. P. Aparicio, G.J . Sibona, Meteorological indicators of dengue epidemics in non-endemic Northwest Argentina, *Infect. Dis. Mod.*, **7** (2022), 823–834. https://doi.org/10.1016/j.idm.2022.10.004

39. S. S. Md-Sani, J. Md-Noor, W. H. Han, S. P. Gan, N. S. Rani, H. L. Tan, et al., Prediction of mortality in severe dengue cases, *BMC Infect. Dis.*, **18** (2018), 1–9. https://doi.org/10.1186/s12879-018-3141-6

40. K. C. Poh, L. F. Chaves, M. Reyna-Nava, C. M. Roberts, C. Fredregill, R. Bueno Jr, et al., The influence of weather and weather variability on mosquito abundance and infection with West Nile virus in Harris County, Texas, USA, *Sci. Total Environ.*, **675** (2019), 260–272. https://doi.org/10.1016/j.scitotenv.2019.04.109

41. C. C. D. Silva, C. L. D. Lima, A. C. G. D. Silva, G. Machado Magalhães Moreno, A. Musah, A. Aldosery, et al., Forecasting Dengue, Chikungunya and Zika cases in Recife, Brazil: a spatio-temporal approach based on climate conditions, health notifications and machine learning, *Res., Soc. Dev.*, **10** (2021), e452101220804. https://doi.org/10.33448/rsd-v10i12.20804

42. B. V. Vishwas A. Patel, *Hands-on Time Series Analysis with Python*, Apress, 2020. https://doi.org/10.1007/978-1-4842-5992-4

43. K. Taegon, J. Minseok, C. J. Hyun, J. Sung-Kwan, Short-term residential load forecasting using 2-step sarimax, *J. Elec. Eng. Tech.*, **17** (2022), 751–758. https://doi.org/10.1007/s42835-021-00917-z

44. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023.

45. J. Breitung P. H. Franses, On Phillips–Perron-type tests for seasonal unit roots, *Economet. Theor.*, **14** (1998), 200–221. https://doi.org/10.1017/S0266466698142032

46. A. Kagalwala, kpsstest: A command that implements the Kwiatkowski, Phillips, Schmidt, and Shin test with sample-specific critical values and reports p-values, *Stata J.*, **22** (2022), 269–292. https://doi.org/10.1177/1536867X221106371

47. E. Kosicka, E. Kozłowski, D. Mazurkiewicz, The use of stationary tests for analysis of monitored residual processes, *Eksploatacja i Niezawodność*, **17** (2015), 604–609. http://dx.doi.org/10.17531/ein.2015.4.17

48. S. Panigrahi, R. M. Pattanayak, P. K. Sethy, S. K. Behera, Forecasting of sunspot time series using a hybridization of arima, ets and svm methods, *Sol. Phys.*, **296** (2021), 1–19. https://doi.org/10.1007/s11207-020-01757-2

49. H. Hewamalage, K. Ackermann, C. Bergmeir, Forecast evaluation for data scientists: common pitfalls and best practices, *Data Min. Knowl. Discovery*, **37** (2023), 788–832. https://doi.org/10.1007/s10618-022-00894-5

50. R. Hyndman, G, Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2 edition, 2018.

51. K. M. Lem, The STL-ARIMA approach for seasonal time series forecast: a preliminary study, in *the 19th IMT-GT International Conference on Mathematics, Statistics and Their Applications (ICMSA 2024)*, **67** (2024), 01008. https://doi.org/10.1051/itmconf/20246701008

52. J. M. González, V. Pakrashi, B. Ghosh, An overview of performance evaluation metrics for short-term statistical wind power forecasting, *Renew. Sust. Energ. Rev.*, **138** (2021), 110515. https://doi.org/10.1016/j.rser.2020.110515

53. S. Engmann, D. Cousineau, Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test, *J, Appl. Quant. Methods*, **6** (2011), 1–18.

54. U. Knief, W. Forstmeier, Violating the normality assumption may be the lesser of two evils, *Beh. Res. Meth.*, **53** (2021), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

55. C. L. Lima, A. C. da Silva, G. M. Moreno, C. Cordeiro da Silva, A. Musah, A. Aldosery, et al., Temporal and spatiotemporal arboviruses forecasting by machine learning: A systematic review, *Front. Pub. Health*, **10** (2022), 900077. https://doi.org/10.3389/fpubh.2022.900077

56. N. A. M. H. Abdullah, N. C. Dom, S. A. Salleh, H. Salim, N. Precha, The association between dengue case and climate: A systematic review and meta-analysis, *One Health*, **15** (2022), 100452. https://doi.org/10.1016/j.onehlt.2022.100452

AIMS Press