Heliyon 8 (2022) e09897

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Time series forecasting for tuberculosis incidence employing neural network models

Alvaro David Orjuela-Cañón [a,*], Andres Leonardo Jutinico [b], Mario Enrique Duarte González [b], Carlos Enrique Awad García [c], Erika Vergara [b], María Angélica Palencia [c]

[a] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, D.C., Colombia
[b] Mechanical, Electronics and Biomedical Engineering Faculty, Universidad Antonio Nariño, Bogotá, D.C., Colombia
[c] Subred Integrada de Servicios de Salud Centro Oriente, Bogotá, D.C., Colombia

ARTICLE INFO

Keywords:
Tuberculosis
Time series
Forecasting
Neural networks
Machine learning

ABSTRACT

Every effort aimed at stopping the expansion of Tuberculosis is important to national programs' struggle to combat this disease. Different computational tools have been proposed in order to design new strategies that allow managing potential patients and thus providing the correct treatment. In this work, artificial neural networks were used for time series forecasting, which were trained with information on reported cases obtained from the national vigilance institution in Colombia. Three neural models were proposed in order to determine the best one according to their forecasting performance. The first approach employed a nonlinear autoregressive model, the second proposal used a recurrent neural network, and the third proposal was based on radial basis functions. The results are presented in terms of the mean average percentage error, which indicates that the models based on traditional methods show better performance compared to connectionist ones. These models contribute to obtaining dynamic information about incidence, thus providing extra-help for health authorities to propose more strategies to control the disease's spread.

1. Introduction

Tuberculosis (TB) is an infectious disease considered to be the 13th leading cause of death and the second leading infectious mortal illness after COVID-19 around the world by the World Health Organization (WHO). It has been estimated that this disease caused around 1.5 million deaths and 10 million people fell ill from TB in 2019. These illness indicators have decreased slowly in recent years despite the different efforts made against it (Organization & others, 2021). This infection is caused by *Mycobacterium tuberculosis* spreading through the air when infected people cough and expel the bacteria (Lienhardt et al., 2012), affecting the lungs in most cases, which is known as pulmonary TB. Due to its simple propagation mode, the incidence rate of TB is falling at about 2% per year, which is less than expected for health authorities. The international reported numbers are between 5 and more than 500 new cases for every 100000 inhabitants. Because of that, the efforts led by the WHO and initiatives such as End TB have tried to decelerate the increase in TB cases faster and more effectively (Organization & others, 2020).

Geographically, the region of the Americas had 2.9% of the reported cases in 2019, with Brazil being inside the 30 high-TB-burden countries, with an incidence rate of 46 new cases for every 100000 inhabitants (Organization & others, 2020). In Colombia, the reported cases reached an incidence rate of 20 new cases every 100,000 inhabitants and 14684 reported new cases in 2019 (Rincón-Torres et al., 2021; Salud, 2020).

In light of the above, innovative approaches have been studied, proposing more instruments to contribute to the WHO's objectives. In recent decades, time series forecasting (TSF) has been considered for statistical methods based on observed data in order to provide additional knowledge about the behavior of the disease. To this effect, mainly Box-Jenkins or autoregressive integrated moving average (ARIMA) techniques have been used (Helfenstein, 1986, 1996; Nelson, 1998). Recently, the field of artificial intelligence (AI) has contributed with models based on artificial neural networks (ANNs), which are considered to be part of the machine learning theory of supervised learning models. ANNs have different architectures which have been applied to TSF, such as those based on nonlinear autoregressive (NAR) models, on radial basis functions (RBF), or on long short-term memory (LSTM) (Box et al., 2015; Che et al., 2018;

* Corresponding author.
E-mail address: alvaro.orjuela@urosario.edu.co (A.D. Orjuela-Cañón).

https://doi.org/10.1016/j.heliyon.2022.e09897
Received 25 October 2021; Received in revised form 22 February 2022; Accepted 1 July 2022

Greff et al., 2016; Palit and Popovic, 2006; Rivero et al., 2019; Tealab, 2018).

For the specific case of TB, TSF has been the subject of research in different places around the world, mainly in high TB burden countries. For example, studies conducted in China (Wang et al., 2017; Whang et al., 2018) have involved hybrid techniques based on ARIMA and NAR models. Furthermore, morbidity has been forecasted in the same region by employing similar strategies (Zheng et al., 2015). Iraq is another country that addressed the TB forecasting problem, proposing the prediction of the number of smear positive TB cases. To this effect, traditional techniques have been applied, such as Box-Jenkins and nonlinear models, as is the case of China (Moosazadeh et al., 2015; Moosazadeh et al., 2014). Another country with a high TB burden is South Africa, where a model based on seasonality was utilized to determine new TB cases (Azeez et al., 2016). In Latin America, Brazil has been mentioned as the most relevant country in terms of TB. There, studies based on TSF have been reported, in which different univariate models have been employed (Ribeiro et al., 2019). In addition, ARIMA and Holt-Winters models have been used to analyze the incidence reported by the Brazilian Unified Health System (Achcar et al., 2021), as well as the effectiveness of the use of GeneXpert within TSF (Berra et al., 2021). However, the use of AI tools is rather lacking in TSF applications to determine TB incidence in the country, which is the largest South America.

Colombia has conducted studies related to the TSF problem for specific diseases such as Zika and Dengue. In the first case, the number of infections was analyzed based on a generalized Richards model using data obtained from the Antioquia region (Chowell et al., 2016). As for Dengue, some studies have dealt with weekly reported incidence cases within a seven-year period (2008–2015), albeit by applying Bayesian hierarchical dynamic generalized linear models (Martínez-Bello et al., 2017). For the same disease, a forecasting model based on regression models and climate data was used to understand how these variables influence the number of new cases in the Risaralda region (Quintero-Herrera et al., 2015). Unfortunately, despite the cited advances in computer science worldwide, there is not enough information on TSF studies based on public data which are specifically related to TB and the use of artificial intelligence techniques in Colombia.

The objective of this work is to present a comparison between the classical forecasting of reported new cases of TB based on ARIMA models and the use of ANNs-based models for the same task. These strategies make their contribution in the context of a developing country, where precarious health infrastructure is the norm and alternative techniques to identify effective interventions and control disease propagation from the field of computational tools can be useful (Ghassemi et al., 2015; Mai et al., 2015; Tealab, 2018; Thorve et al., 2018). This study was conducted on dataset of publicly available data for the city of Bogotá. This city has the largest population in the country, with almost seven million people, which implies an interesting group of study. Interdisciplinary professionals, including expert physicians in TB diagnosis were involved. This allowed understanding the potential of this type of analysis in the detection of the disease.

2. Materials and methods

2.1. Database

Information extracted from the National Health Institute (*Instituto National de Salud*, INS) was used during this study (de Salud, 2020). The number of new reported cases related to pulmonary TB in Bogotá from 2007 to 2020 was employed in order to establish the time series to be analyzed. According to the Colombian Public Health Monitoring System (*Sistema de Vigilancia en Salud Pública*, SIVIGILA) all confirmed cases must be reported to the INS, and then they are published on a weekly basis in different channels of communication (*e.g.*, the website). Taking advantage of the public open data policies (Ministerio de Tecnologías de la Información y las Comunicaciones, 2016), yearly reports were collected in order to arrange the dataset. Figure 1 shows the time series with data on weeks for the 2007–2020 period in Bogotá, Colombia.

A sequence of 694 values was taken into account during this study. This average length was obtained from 52 weekly reports published per year. A set of twelve complete years was chosen in order to use complete registers and avoid biases due to incomplete information. In addition, only the four first months of 2020 were considered, *i.e.*, the months prior to the national government's declaration of the sanitary emergency caused by COVID-19.

In order to avoid saturation in the weights of the neural models, the time series was normalized into an interval [0, 1] through the following formula:

$$y_{normalized} = \frac{y_i - y_{min}}{y_{max} - y_{min}} \tag{1}$$

where y_i is the original value, and y_{min} and y_{max} are the minimum and maximum values of the time series, respectively. In order to generalize the analysis of the obtained models, it was necessary to implement a holdout validation technique to evaluate the ability to process unseen inputs. This technique was applied by dividing the entire time series into a portion for training and another for testing. The training subset was obtained by using 70% of the series, fixing a 10% to validate the training to avoid overfitting in the models. The testing subset was generated by using the 30% of the sequence, which corresponds to 208 values from the 2016 to 2020.

2.2. Autoregressive integrated moving average

ARIMA or Box-Jenkins models have been traditionally used for TSF in different sectors, such as those related to the economy, the industry, or the environment, among others. However, the medical field can benefit from the application of this kind of models (Box et al., 2015; Helfenstein, 1996). In this way, the ARIMA approach is commonly used to analyze data representations that depend on consecutive measurements. Therefore, the representation of TSF based on ARIMA models was taken as reference in this work.

ARIMA models can be interpreted as autoregressive moving average (ARMA) models with a previous preprocessing based on differentiated values from the original time series. In this way, it is possible to express these models as follows:

$$d_i = \mathrm{c} + \sum_{k=1}^{p} a_i d_{i-k} + \sum_{k=1}^{q} b_i \varepsilon_{i-k} + \varepsilon \tag{2}$$

where d_i represents the differentiated time series $y - y_{i-k}$, a_i are the coefficients associated to the p-order values of the AR model, and b_i is related to the q-order values of the moving average (MA) model, which are based on the computation of values of the white noise ε. i corresponds to the time samples of the time series. Thus, the ARIMA (p,d,q) is a model composed of the AR(p) and MA(q) models applied to the differentiated (d) time series.

Box and Jenkins proposed a technique to determine orders p, q, and d in the models from autocorrelation and partial correlation functions (Helfenstein, 1986, 1996). However, in this case, the available computational resources were exploited in order to find the best model, modifying the orders from one to ten for the p and q values, as well as from zero to five for the d value. Finally, the models were implemented for econometric and statistical modeling via the *statsmodels* tool in Python (Seabold and Perktold, 2010).

2.3. Nonlinear autoregressive model (NAR)

The applications of ANNs vary according to their ability to learn and detect patterns to perform approximations of functions, especially

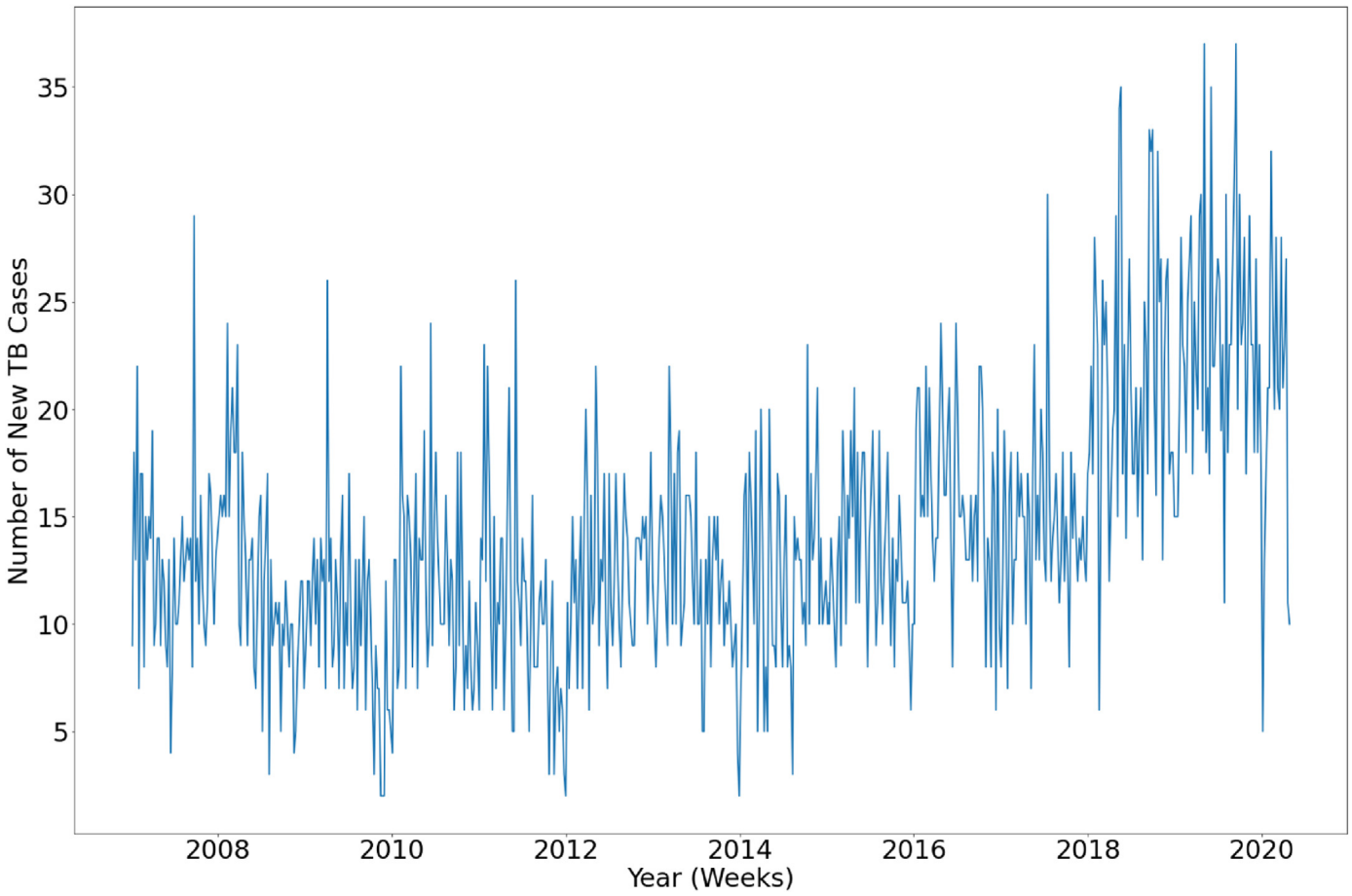

Figure 1. Number of reported TB new cases in Bogotá for the 2007–2020 period. Information is collected weekly by the National Health Institute in Colombia.

nonlinear ones (Haykin, 2009; Palit and Popovic, 2006). In the field of time series, forecasting is one of such applications, where NAR models are employed to obtain the series behavior from data.

As in autoregressive (AR) models, the past values of the time sequence y_i are utilized to adjust the a_i coefficients:

$$y_i = tanh\left(\sum_{k=1}^{p} a_i y_{i-k} + b\right) \tag{3}$$

where y_i is the current time output estimation, and a_i are the coefficients of the model, which are known as synaptic weights (a_{ij}) in the proposed ANN model. Parameter b is employed in the model as a bias and helps improve the estimation. The main difference to the AR model is that the NAR model is nonlinear, which holds this nonlinearity in the hyperbolic tangent (tanh) in Eq. (3). It is also known as a transfer function of the units or neurons in the neural model. The inputs used to train the NAR model are delayed samples of time series $y_{i\text{-}k}$, the hidden layer sets the nonlinear computation, and the output is computed just by one time unit, which results in the future value (forecasting) of sequence y_i.

In order to obtain the number of inputs and units in the hidden layer, a heuristic process was followed. For a comparison with the ARIMA model, delays caused by lags and hidden unit values from one to ten were tested. Then, the model with the lowest error was preferred. For training, the development subset and the use of a resilient backpropagation algorithm were selected due to their performance when compared to other training algorithms in terms of speed (Günther and Fritsch, 2010). The maximum number of epochs was adjusted to 50 according to experimental findings in the error curve during training, and, in this way, the model was compared to other ANN proposals with the same training parameters.

2.4. Long short-term memory model (LSTM)

LSTM is based on recurrent neural networks, where feedback connections within the architecture evidence its main differences with NAR models with feedforward weights. In addition, this model can be considered to be a deep learning one, which depends on the number of layers. However, only one hidden layer was employed in this case.

This architecture is composed of a cell with input, output, and forget gates. The latter has a function for determining what information must be remembered by the network. Three gates are used to control the flow of information that enters and exits the cell (Greff et al., 2016). The expressions of the LSTM model are formulated as follows:

$$i(t) = \theta_i(W_{xi}x(t) + W_{hi}h(t-1) + W_{ci}c(t-1)) \tag{4}$$

$$f(t) = \theta_f\left(W_{xf}x(t) + W_{hf}h(t-1) + W_{cf}c(t-1)\right) \tag{5}$$

$$c(t) = f(t)c(t-1) + i(t)\theta_c(W_{xc}x(t) + W_{hc}h(t-1)) \tag{6}$$

$$q(t) = \theta_q\left(W_{xq}x(t) + W_{hq}h(t-1) + W_{cq}c(t)\right) \tag{7}$$

$$h(t) = q(t)\theta_h(c(t)) \tag{8}$$

where $i(t)$, $f(t)$, $q(t)$, and $c(t)$ in Eqs. (4), (5), (6), and (7) correspond to the input, forget, and output gates, as well as the cell activation vectors of the same size for the hidden vector $h(t)$ (8). The θ symbol labels the activation functions.

Traditional recurrent models suffer from the gradient vanish problem when the backpropagation algorithm is used for training, which is

due to its deep connections over long periods in time. LSTM shows an improved behavior with regard to this problem, employing a cell structure based on control gates that do not affect the training (Greff et al., 2016; Sutskever et al., 2014). For the sake of comparison with the NAR approach, the number of training epochs was the same in this present case (50).

As in NAR model, the delays caused by lags and the number of hidden cells for the network were varied from one to ten in order to compare the models using the same values for the parameters to be adjusted.

2.5. Radial basis functions (RBF)

The RBF architecture is based on three layers, similar to the NAR model, but it can be regarded as linear combinations from nonlinear functions with a radial basis, which is represented in the hidden layer (Haykin, 2009). To obtain these combinations, n functions are proposed as follows:

$$y_i = \sum_{k=1}^{n} a_k \rho(\|\mathbf{x} - c_k\|) \tag{9}$$

where ρ is the radial basis function –commonly a Gaussian function– and a_k are the weights associated to each unit or neuron k. As the functions have a radial basis, c_k represents the center of the n functions. The nonlinear function is applied to the distance of each input x in relation to the center c. In this case, as a comparative method, this number of functions was modified from one to ten, finding the best parameters in the training process.

As with the previous models, the parameters to find the best performance were experimentally found. The number of lags or inputs and the number of radial basis functions or hidden layer units were modified within the same interval: from one to ten. To maintain the training parameters controlled for comparison, 50 epochs were employed to adjust the weights.

All neural networks models were implemented by using classes of the Keras API in the Google Colab environment (Chollet & others, 2015; Manaswi, 2018).

2.6. Evaluation methods

The prediction accuracy was evaluated by determining the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the mean absolute error (MAE), which are employed to assess forecasting models (Shcherbakov et al., 2013). These values can be computed according with expressions (10), (11), and (12).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{10}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} 100\% \tag{11}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |e_i| \tag{12}$$

where N is the number of the values or points in the sequence, $\hat{y}_i$ is the forecasted series, y_i is the original series, and e_i is the error between samples of the series.

For the three techniques described, and according to the parameter exploration, the best model was chosen based on three error metrics in the training set. Then, the data from the test subset was applied to assess model generalization. This allowed determining what model provides the best performance when unseen data are used.

3. Results and discussion

Table 1 shows the results for the best ARIMA models according to the RMSE, MAPE, and MAE values. The parameters of the best model are presented in terms of the p, d, and q orders employed in the adjusting process. Tables 2, 3, and 4 show the ANN proposals, visualizing the results for the NAR, LSTM, and RBF models. Here, the information is similar to Table 1, albeit with units and cells in the hidden layer and the number of input lags employed in the training.

The ARIMA model exhibited the use of seven lags or inputs and differentiated time series of one and two orders. The best RMSE result was achieved without any differences computed by the time series and an order of nine for the MA model associated with white noise. In the case of the NAR model, the best performance, taking the RMSE as reference, was reached by a neural network with nine inputs and seven units in the hidden layer. For the LSTM model, a recurrent neural network with eight inputs and eight cells in the hidden layer obtained the best result in terms of the RMSE. Finally, the RBF model employed an architecture based on eight inputs and nine units (functions) in the hidden layer.

The difference between the ANN models was reflected on the inputs, with a minimum eight inputs for LSTM and RBF models, compared to nine lags for the NAR model. In addition, the number of units in the hidden layer showed a minimum of seven with the NAR model, as well as a maximum of ten cells for the LSTM. The RBF model required nine functions for forecasting.

The ARIMA models obtained higher error values for the RMSE and MAE metrics, but a slightly lower complexity for the models, employing just seven lags from the time series. Despite this, the performance of the three models was computed for the test subset, choosing the best model based on the three metrics. A comparison of the errors in the test subset is shown in Table 5 for the selection based on RMSE, where the lowest values were reached by the ARIMA (7,0,9) model. However, the error difference was not more than one unit for the RMSE and MAE metrics, obtaining more than six units with regard to the MAPE metric. Tables 6 and 7 show the comparison of the three metrics for model selection based on MAE and MAPE values, taking the model column in Tables 1, 2, 3, and 4 as reference. There, it can be observed that, for the NAR and RBF models, the result was the same when the RMSE and MAE metrics were considered.

Figure 2 shows the segment of series used to test the models. The black series corresponds to the original data, the red time series is associated with the results of the ARIMA model. This segment is compared to the best forecasts of the NAR (blue), LSTM (green) and RBF models (purple). In this Figure, it is difficult to observe differences in the performance of the three models. The ANN proposals managed to forecast the details from the original series regarding trend and speed changes. These models did not reach the high values of the original series in the test set, thus failing to obtain peak values of incidence higher than 25 TB cases. This can be explained by the division of the sequence (for training and testing), an aspect visible in the training series, with 486 week values with no points higher than 25 TB cases, as seen in Figure 1. The TB incidence changed abruptly in the test set before week 100 (Figure 2), reaching peaks with more than 35 TB cases in the last samples. This means that the models do not reach the highest values in the training due to the division made, even after the normalization process. Nevertheless, the ARIMA model could obtain values higher than 25 new cases,

Table 1. Results for the ARIMA model.

Type of Error	Training Set	
	Value	Model
RMSE	4.010	ARIMA (7,0,9)
MAE	3.155	ARIMA (7,1,7)
MAPE	34.138	ARIMA (7,2,2)

Table 2. Results for the NAR model.

Type of Error	Training Set	
	Value	Model
RMSE	3.9001	Lags: 9 Units: 7
MAE	3.0491	Lags: 9 Units: 7
MAPE	37.4935	Lags: 1 Units: 9

Table 3. Results for the LSTM model.

Type of Error	Training Set	
	Value	Model
RMSE	3.9058	Lags: 8 Cells: 8
MAE	3.0599	Lags: 8 Cells: 10
MAPE	38.7197	Lags: 3 Cells: 1

Table 4. Results for the RBF model.

Type of Error	Training Set	
	Value	Model
RMSE	4.0154	Lags: 8 Units: 9
MAE	3.1404	Lags: 8 Units: 9
MAPE	37.0049	Lags: 1 Units: 7

Table 5. Comparison of results for all models in the test subset when the RMSE was considered.

Type of Error	Models			
	ARIMA	NAR	LSTM	RBF
RMSE	5.8916	6.4775	6.5928	7.2267
MAE	4.5547	4.9573	5.0602	5.5030
MAPE	26.9366	36.1461	33.6701	31.2997

Table 6. Comparison of results for all models in the test subset when the MAE was considered.

Type of Error	Models			
	ARIMA	NAR	LSTM	RBF
RMSE	5.7444	6.4775	7.2643	7.2267
MAE	4.4707	4.9573	5.5583	5.5030
MAPE	26.9926	36.1461	32.7088	31.2997

Table 7. Comparison of results for all models in the test set when the MAPE was considered.

Type of Error	Models			
	ARIMA	NAR	LSTM	RBF
RMSE	5.8509	7.5124	9.6832	7.8548
MAE	4.5748	9.3237	7.7452	7.4045
MAPE	28.1752	34.1471	36.5180	36.1315

showing a lower difference after the week 100, with a performance higher than that of the ANNs. This allows observing that the ARIMA model had a better generalization of the time series phenomena than the ANN models.

Despite this last disadvantage, the obtained models could be useful in applications where the behavior of the phenomenon is more important, especially in critical situations requiring the healthcare system to work better, anticipating actions in the form of policies or educational stages. In other words, the identification of possible changes in the tendency and obtaining information about important slopes regarding the number of TB cases is preferable for institutions and administrative entities involved in healthcare. Moreover, it can be seen that the models learned the seasonal behavior of the series, providing the necessary incidence data to establish timely control strategies. According to this, it was observed that the models with the best results required between seven and nine lags in the input for TSF, which implies that information for at least two months (eight weeks) is required before the possible intervention to modify the new TB cases time series. This is important to determine the effective design of strategies to control TB incidence in terms of possible new cases.

The differences in the results for both employed ANN models can be explained by the characteristics of each architecture. The explorations of the training parameters were similar for both models, and the number of epochs and stopping criteria were the same. The lags in the input and units in the hidden layer were modified using the same intervals. The LSTM model had a larger architecture, with ten units in the hidden layer and comparable results, which indicates that the forecasting task was more difficult. Previous works have shown how the NAR model is adequate for obtaining information on TB (Wang et al., 2017; Whang et al., 2018). The LSTM did not have a smaller architecture, as could be expected due to its recurrent behavior. Despite this, the errors in Tables 2 and 3 show a minimal difference between both analyzed ANN proposals.

In terms of the RMSE and MAE values, these errors were close to four to five units regarding new TB cases in the training set, and they reached about six to eight units in the test subset –almost nine for the LSTM. This means that the differences between the forecasted series can be inaccurate with respect to the original series around this number, but it is also important to understand the performance of the series in order to identify periods or moments for interventions related to control or health policies –which, in this case, requires an interval of eight weeks.

The results obtained from the neural model proposals compared to traditional strategies were unsatisfying, as the ARIMA model showed a better performance. However, the simplicity of the models can explain the results, which were employed to assess the possibility of using AI techniques without any complications. The results from traditional and AI techniques were contrasted, concluding that the statistical methods were better (Makridakis et al., 2018). Moreover, the results depend on the application in terms of data quality, quantity, length, and availability, but, at the same time, on the interactions of the specific field (Ray et al., 2021), namely the healthcare area.

The novelty of this work is the use of ANN techniques such as recurrent neural networks and nonlinear autoregressive models to analyze the TB incidence time series. There are similar studies in the region, where context similarities can be applied; for example, models based on statistical models such as ARMA, ARIMA, simple exponential smoothing, Holt-Winters, and its modified exponential smoothing were studied in Brazil, reaching the best results with the implementation of the ARIMA (4,1,5) model (Ribeiro et al., 2019). In addition, a complementary study compared the effectiveness of the GeneXpert technology with an ARIMA (5,0,0) model (Berra et al., 2021). Nevertheless, AI strategies to propose or obtain this kind of model have not been reported in the same region. This study presented the comparison of two ANN models to carry out this task, as well as to determine the advantages and disadvantages AI techniques in this context. Furthermore, based on this proposal, the researchers believe that interdisciplinary work in the field can be effective, as it involves synergy between the establishment of the models and their application with direct users, with the purpose of improving models, methods, and performances. Physicians believe that the results could be beneficial for TB diagnosis in locations with lacking resources, where the advantages of employing neural networks models have been evidenced (Orjuela-Cañón et al., 2018).

The limitations of this study are associated with the results obtained from the ARIMA model in comparison with the ANN ones. Traditional strategies to fit the ARIMA models are based on analyzing total and partial correlation, in addition to complementary techniques for

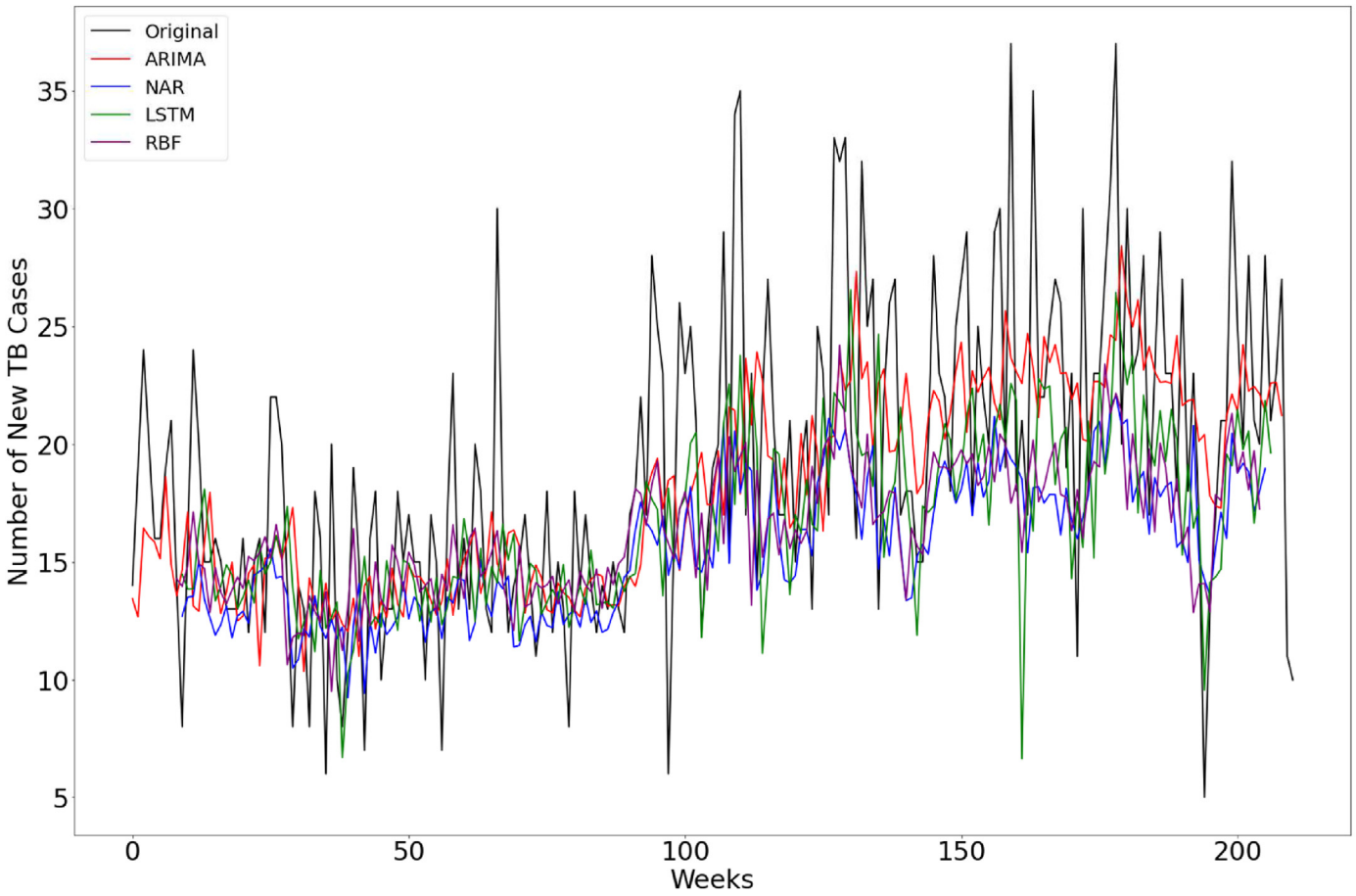

Figure 2. Results of the time series forecasting comparison for the three employed models. The ARIMA model showed the best performance, as represented by the red line.

obtaining the final model to represent the data. However, the main objective of this work was to compare ANN models to the common techniques used for the TSF problem. To this effect, the methodology sought equality in the adjusting/training process for all considered models. Additionally, some problems that were not examined are related to the review of data acquisition issues. The TSF problem is affected by spurious events that modify the time series without a natural behavior. In other words, the COVID-19 pandemic could have affected TB services; the reported new cases may show drops in the notifications collected by national healthcare institutions (Organization & others, 2021). Due to this, this study took values from before the pandemic.

4. Conclusions

In order to obtain a model for time series forecasting of the reported new cases of TB for the city of Bogotá, Colombia, models based on artificial neural networks were employed. A comparison between NAR, RBF and LSTM models was performed in order to determine which had the best performance, taking the traditional ARIMA model as reference.

According to the results obtained, the LSTM models exhibited an architecture smaller than that of NAR with similar results; RMSE values of 6.59 and 6.39 were obtained, respectively. The ARIMA model obtained an RMSE of 5.89, the best performance among the analyzed models. This means that traditional models are better than AI-based strategies in this context. Regardless, the information gathered from the implementation of different strategies for TSF is useful in decision-making processes, and it allows planning interventions to mitigate the propagation of TB, especially when national vigilance institutions declare emergencies. In this case, the results allowed establishing an interval of seven to eight weeks for TSF, due to the fact that this value yields the best results for the studied models.

Finally, these preliminary results allowed determining the usefulness of the aforementioned neural models in the study of TB incidence as observed by a time series based on the weekly number of new TB cases. As it was seen, ANN models can be a starting point to continue with proposals for forecasting new reported TB cases. As future work, deep learning models could be employed which are based on LSTM networks, as they control backpropagation vanishing. In addition, hybrid proposals that include traditional and AI-based models could be analyzed.

Declarations

Author contribution statement

Alvaro David Orjuela-Cañón: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Andres Leonardo Jutinico Alarcón and Mario Enrique Duarte González: Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Carlos Enrique Awad García: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Erika Vergara: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Maria Angélica Palencia: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Ministerio de Ciencia, Tecnología e Innovación - Miniciencias, in Colombia (123380762899).

Data availability statement

Data will be made available on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors would like to acknowledge Universidad del Rosario, Universidad Antonio Nariño, and *Subred Intregrada de Servicios de Salud Centro-Oriente* and its *Unidad de Servicios de Salud Santa Clara* for their support, as well as Minciencias.

References

Achcar, J.A., Oliveira, R.P., Barili, E., 2021. The incidence of tuberculosis in Brazil from 2001 to 2018: use of polynomial regression combined with a stochastic volatility model. Int. J. Clin. Biostat. Biom. 7, 35.

Azeez, A., Obaromi, D., Odeyemi, A., Ndege, J., Muntabayi, R., 2016. Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model. Int. J. Environ. Res. Publ. Health 13 (8), 757.

Berra, T.Z., Gomes, D., Ramos, A.C.V., Alves, Y.M., Bruce, A.T.I., Arroyo, L.H., Arcêncio, R.A., 2021. Effectiveness and trend forecasting of tuberculosis diagnosis after the introduction of GeneXpert in a city in south-eastern Brazil. PLoS One 16 (5), e0252375.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y., 2018. Recurrent neural networks for multivariate time series with missing values. Sci. Rep. 8 (1), 6085.

Chollet, F., others, 2015. Keras. GitHub. Retrieved from. https://github.com/fchollet/keras.

Chowell, G., Hincapie-Palacio, D., Ospina, J., Pell, B., Tariq, A., Dahal, S., Viboud, C., 2016. Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics. PLoS Curr. 8.

de Salud, I.N., 2020. Tuberculosis: Protocolo de Vigilancia en Salud Pública. Instituto Nacional de Salud, Colombia.

Ghassemi, M., Pimentel, M.A.F., Naumann, T., Brennan, T., Clifton, D.A., Szolovits, P., Feng, M., 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In: Twenty-Ninth AAAI Conference on Artificial Intelligence.

Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J., 2016. LSTM: a search space odyssey. IEEE Transact. Neural Networks Learn. Syst. 28 (10), 2222–2232.

Günther, F., Fritsch, S., 2010. neuralnet: training of neural networks. Rice J. 2 (1), 30–38.

Haykin, S., 2009. Neural Networks and Learning Machines. Prentice-Hall. Retrieved from. https://books.google.com.co/books?id=K7P36lKzI_QC.

Helfenstein, U., 1986. Box-Jenkins modelling of some viral infectious diseases. Stat. Med. 5 (1), 37–47.

Helfenstein, U., 1996. Box-Jenkins modelling in medical research. Stat. Methods Med. Res. 5 (1), 3–22.

Lienhardt, C., Glaziou, P., Uplekar, M., Lönnroth, K., Getahun, H., Raviglione, M., 2012. Global tuberculosis control: lessons learnt and future prospects. Nat. Rev. Microbiol. 10 (6), 407.

Mai, Q., Aboagye-Sarfo, P., Sanfilippo, F.M., Preen, D.B., Fatovich, D.M., 2015. Predicting the number of emergency department presentations in W estern A ustralia: a population-based time series analysis. Emerg. Med. Australasia (EMA) 27 (1), 16–21.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and Machine Learning forecasting methods: concerns and ways forward. PLoS One 13 (3), e0194889.

Manaswi, N.K., 2018. Understanding and working with Keras. In: Deep Learning with Applications Using Python. Springer, pp. 31–43.

Martínez-Bello, D.A., López-Quílez, A., Torres-Prieto, A., 2017. Bayesian dynamic modeling of time series of dengue disease case counts. PLoS Neglected Trop. Dis. 11 (7), e0005696.

Ministerio de Tecnologías de la Información y las Comunicaciones, 2016. Guía para el uso y aprovechamiento de Datos Abiertos en Colombia.

Moosazadeh, M., Khanjani, N., Nasehi, M., Bahrampour, A., 2015. Predicting the incidence of smear positive tuberculosis cases in Iran using time series analysis. Iran. J. Public Health 44 (11), 1526.

Moosazadeh, M., Nasehi, M., Bahrampour, A., Khanjani, N., Sharafi, S., Ahmadi, S., 2014. Forecasting tuberculosis incidence in Iran using box-jenkins models. Iran. Red Crescent Med. J. 16 (5).

Nelson, B.K., 1998. Time series analysis using autoregressive integrated moving average (ARIMA) models. Acad. Emerg. Med. 5 (7), 739–744.

Organization, W.H., others, 2020. Global Tuberculosis Report 2020: Executive Summary.

Organization, W.H., others, 2021. Global Tuberculosis Report 2021.

Orjuela-Cañón, A.D., Mendoza, J.E.C., García, C.E.A., Vela, E.P.V., 2018. Tuberculosis diagnosis support analysis for precarious health information systems. Comput. Methods Progr. Biomed.

Palit, A.K., Popovic, D., 2006. Computational Intelligence in Time Series Forecasting.

Quintero-Herrera, L.L., Ramírez-Jaramillo, V., Bernal-Gutiérrez, S., Cárdenas-Giraldo, E.V., Guerrero-Matituy, E.A., Molina-Delgado, A.H., 2015. Potential impact of climatic variability on the epidemiology of dengue in Risaralda, Colombia, 2010–2011. J. Infect. Public Health 8 (3), 291–297.

Ray, A., Chakraborty, T., Ghosh, D., 2021. Optimized ensemble deep learning framework for scalable forecasting of dynamics containing extreme events. Chaos: Interdisc. J. Nonlinear Sci. 31 (11), 111105.

Ribeiro, R.C.M., Quadros, T.A., Saldarriaga, J.J., Júnior, S., de Almeida, J.F., Marques, G.T., 2019. Forecasting incidence of tuberculosis cases in Brazil based on various univariate time-series models. Int. J. Innov. Educ. Res. 7 (10), 894–909.

Rincón-Torres, C.E., Rubio, V., Castro, C., Garc\'\ia, I., Cruz, O.A., Trujillo-Trujillo, J., Puerto, G.M., 2021. Red Nacional de Gestión de Conocimiento, Investigación e Innovación en Tuberculosis en Colombia. Rev. Panam. Salud Públic 45, e23.

Rivero, C.R., Pucheta, J., Otaño, P., Orjuela-Cañon, A.D., Patiño, D., Franco, L., Juarez, G., 2019. Time series forecasting using recurrent neural networks modified by Bayesian inference in the learning process. In: 2019 IEEE Colombian Conference on Applications in Computational Intelligence, ColCACI 2019 - Proceedings.

Salud, I.N. de., 2020. Weekly Epidemiological Report - (Boletín Epidemiológico Semanal) - [In Spanish]. Instituto Nacional de Salud.

Seabold, S., Perktold, J., 2010. statsmodels: econometric and statistical modeling with python. In: 9th Python in Science Conference.

Shcherbakov, M.V., Brebels, A., Shcherbakova, N.L., Tyukov, A.P., Janovsky, T.A., Kamaev, V.A., 2013. A survey of forecast error measures. World Appl. Sci. J. 24 (24), 171–176.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112.

Tealab, A., 2018. Time series forecasting using artificial neural networks methodologies: a systematic review. Fut. Comput. Info. J. 3 (2), 334–340.

Thorve, S., Wilson, M.L., Lewis, B.L., Swarup, S., Vullikanti, A.K.S., Marathe, M.V., 2018. EpiViewer: an epidemiological application for exploring time series data. BMC Bioinf. 19 (1), 1–10.

Wang, K.W., Deng, C., Li, J.P., Zhang, Y.Y., Li, X.Y., Wu, M.C., 2017. Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network. Epidemiol. Infect. 145 (6), 1118–1129.

Whang, J., Wang, C., Wenyu, Z., 2018. Data analysis and forecasting of tuberculosis prevalence rates for smart healthcare based on a novel combination model. Appl. Sci. 8 (9), 1–24.

Zheng, Y.-L., Zhang, L.-P., Zhang, X.-L., Wang, K., Zheng, Y.-J., 2015. Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. PLoS One 10 (3), e0116832.