TOWARDS TUBERCULOSIS INCIDENCE TIME SERIES FORECASTING IN COLOMBIAN REGIONS: THE ANTIOQUIA CASE

Ivan Camilo Sanchez-Vega[1], Alvaro David Orjuela-Cañon[2], Carlos Enrique Awad Garcia[3] and Erika Vergara[4],
Maria Angelica Palencia[3] and Andres L. Jutinico[1]

[1]Mechanical, Electronics and Biomedical Engineering Faculty,
Universidad Antonio Nariño, Bogota D.C., Colombia
[2]School of Medicine and Health Sciences,
Universidad del Rosario, Bogota D.C., Colombia
[3]Subred Integrada de Servicios de Salud Centro Oriente,
Bogota D.C., Colombia
[4]Universidad Antonio Nariño, Bogota D.C., Colombia

ABSTRACT

Antioquia is a Colombian department where 6.7 million people live. Currently, it is the region of the country with the newest cases of tuberculosis reported in 2021, about 18.8%. In addition, the incidence rate of tuberculosis was 36.8 per 100,000 inhabitants. Public government health policy regarding tuberculosis should aim to prevent the uninfected community, in addition to detecting and treating people with tuberculosis. In this sense, the study of algorithms to predict the epidemic trend should be promoted. This work addresses the prediction of tuberculosis cases in Antioquia, considering data from the health surveillance system between 2007 and 2021. For the prediction, the Kalman filter and the autoregressive model are considered. The results show a better performance using the Kalman filter for the prediction of tuberculosis cases at six weeks.

KEYWORDS

Tuberculosis, Forecasting, Autoregressive Model, Kalman Filter, Performance.

1. INTRODUCTION

Tuberculosis (TB) is an infectious disease causing 1.3 million deaths worldwide in 2020 [1]. Moreover, in 2020 was the second leading cause of death from a single infectious agent after COVID-19. According to the national Public Health Surveillance System (SIVIGILA), in Colombia, 14.060 TB cases were reported in 2021, being 91.98% new cases. Besides, in this country, the incidence rate of tuberculosis was 25.8 per 100,000 inhabitants [2] [3]. The above indicates a growth in the contagion concerning 2020 [4]. This disease is contracted by airways when people infected with TB cough expel the bacteria Mycobacterium tuberculosis.

The more known form of TB affects the lungs (pulmonary TB). The contagion and transmission of the disease could be controlled if public health entities had the necessary tools. There are several traditional methods for control, infection minimization, and treatment. However, in Colombia, none are oriented to predict how many TB cases may occur during each week of the

David C. Wyld et al. (Eds): MLDS, IBCOM, CSIA, NLAI, SPPR, ICCSEA - 2022
pp. 117-126, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.122110

year. This information would help make administrative decisions to minimize infection rates through prevention. In state-of-the-art, several forecasting methods exist for infectious diseases, for instance [5]. However, few studies deal with TB prediction in Colombia [6], [7]. Also, TB forecasting is still a challenge given the abrupt changes in trend and dispersion of TB incidence relative to each region or country. This epidemic depends on several factors related to the contagious disease, some of these random and variables in time, such as the increase of immigrant populations, representing a significant portion of TB epidemics [8]. This study addressed TB case prediction in the Antioquia department in Colombia using the Kalman Filter (KF) and the Auto-Regressive (AR) models.

Authors of [9] conducted a study in Brazil with a database of confirmed cases of Pulmonary TB in the period from 2011 to 2018. The authors considered three statistical models for prediction: the Simple Exponential Smoothing Model, Autoregressive Integrated Moving Average Model, and Holt-Winters Exponential Smoothing Model.

Qiao Liu in Jiangsu, China, conducted a study to predict the seasonality and trend of pulmonary TB. In this paper are shown two methods: Back Propagation Neural Network (BPNN) and autoregressive integrated moving average (ARIMA) [10]. The ARIMA and the nonlinear autoregressive neural network (NAR) were used to predict the incidence of TB cases in Jiangsu, China, in 2017. This province of China has one of the highest pulmonary TB infection rates in the world [11]. In South Africa, performed research using Pulmonary TB data collected from 2010 to 2015 and implementing the seasonal ARIMA model and NAR for TB prediction in the Eastern Cape region [12]. In 2020 a study conducted in Delhi, India, states that patients with latent or confirmed TB are at increased risk of contracting SARS-CoV-2 infection and developing severe pneumonia due to COVID-19. The study shows that with the support of public entities, at the peak of the COVID-19 pandemic, a 25.35% reduction in the TB & SARS-CoV-2 co-infection rate was achieved. The authors confirm that hospitals and medical centers should be prepared for early forecasting and diagnosis of COVID-19 in TB patients [13]. In [5] performed the time series analysis under the hybrid method Box-Jenkins and Elman neural network to predict TB incidence in Kashgar (China), where there is a high rate of TB occurrence. They conclude that this AR hybrid method is effective and can predict TB incidence in this place.

The rest of this paper is organized as follows: Section 2 introduce the Auto-Regressive model; Section 3 introduce the Kalman filter; in Section 4 are explained the results and in Section 5 the conclusions.

2. AUTO-REGRESSIVE MODEL

Antioquia was the Colombia region with more new cases of TB in 2021. In this research, we use the number of reported cases of pulmonary TB in this department from 2007 to 2021, as shown in Fig. 1. This data set is formed by confirmed cases week by week. The time series has been extracted from the Health National Institute (INS in Spanish), reported by the SIVIGILA. The time series is split into two data groups to identify and validate a state-space model. From the identification process,

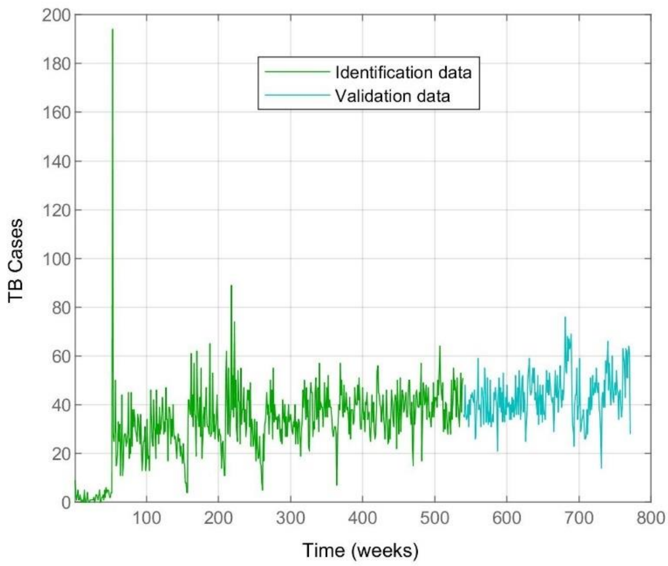

Figure 1. Reported TB cases in Antioquia for 2007-2021 period.

we obtain models based on the Numerical algorithms for Subspace State Space System Identification (N4SID) [14]. These AR models are defined in terms of their order n, and in a canonical way, which allows forecasting states, as is described in (1)-(2). In this model $xk \in Rn$ is the state vector to be estimated with its corresponding state variables (χ), $yk \in Rp$ is the output variable, $\omega k \in R^{m2}$ is the disturbance given by the estimation error and $vk \in Rt$ is the output noise. The nominal matrices F, G, and C have appropriate dimensions.

$$\begin{bmatrix} x_{k+1} \\ x_{k+2} \\ \vdots \\ x_{k+n-1} \\ x_{k+n} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ f_0 & f_1 & f_2 & \cdots & f_{n-1} \end{bmatrix}}_{F} \begin{bmatrix} x_k \\ x_{k+1} \\ x_{k+2} \\ \vdots \\ x_{k+n-1} \end{bmatrix} + \underbrace{\begin{bmatrix} g_{n-1} \\ g_{n-2} \\ g_{n-3} \\ \vdots \\ g_0 \end{bmatrix}}_{G} w_k \text{ , (1)}$$

$$y_k = \underbrace{[1 \quad 0 \quad 0 \quad \cdots \quad 0]}_{C} \underbrace{\begin{bmatrix} x_k \\ x_{k+1} \\ x_{k+2} \\ \vdots \\ x_{k+n-1} \end{bmatrix}}_{x_k} + v_k \text{ , (2)}$$

The states of the AR model (1) are estimated considering, $\omega k = yk - C\,\hat{x}_{k|k-1}$, as follows,

$$\hat{x}_{k+1|k} = F\,\hat{x}_{k|k-1} + Gk\omega k.$$

3. The Kalman Filter

The Kalman filter aims to find the optimal states estimates $\hat{x}_k{}^*$, $\hat{x}_{k+1}{}^*$ of model (1)-(2)
by minimize the following quadratic criterion,

$$\min_{\widehat{w}_k, \hat{v}_k, \hat{x}_k, \hat{x}_{k+1}} = \left\{ \left\| \hat{x}_k - \hat{x}_{k|k-1} \right\|_{P_{k|k-1}{}^{-1}}{}^{2} + \begin{bmatrix} \widehat{w}_k \\ \hat{v}_k \end{bmatrix}^T \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{bmatrix} \widehat{w}_k \\ \hat{v}_k \end{bmatrix} \right\}, (4)$$

The update of the matrix $P_{k|k}$ is given by the Riccati equation (5)-(6),

$$P_{k|k} = P_{k|k-1} - P_{k|k-1}C^T(R + CP_{k|k-1}C^T)^{-1}CP_{k|k-1}, (5)$$

$$P_{k+1|k} = Q + F\,P_{k|k}F^T, (6)$$

where matrices $P_{k|k} \succ 0$, $Q \succ 0$ and $R \succ 0$ are the posteriori estimate covariance, the covariance of the process noise and the covariance of the observation noise respectively. The Kalman gain L_k is given by:

$$L_k = P_{k|k-1}C^T(R + CP_{k|k-1}C^T)^{-1}, (7)$$

Thus, the current states of the model (1)-(2) are estimated by the Kalman filter as follows,

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k(y_k - C\hat{x}_{k|k-1}), (8)$$

More details about the Kalman filter can be found in [15] and [16].

4. RESULTS

In order to evaluate the performance of forecasting models, we use the Mean Absolute Error (MAE) measure, the Tracking Signal (TRS), and Root Mean Square Error (RMSE). These rates are commonly employed to assess forecasting models [17], [18] and computed according to:

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|e_k| , (9)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N} e_k{}^2} , (10)$$

$$TRS = \frac{\sum_{k=1}^{N} e_k}{MAE}, (11)$$

where N is the maximum step number or points in the sequence and $e_k = y_k - \hat{y}_k$ is the errorbetween the actual and the forecasted series.

The model for predicting TB cases in the Antioquia region is based on the N4SID algorithm. By this algorithm, we get a state-space model by training with the 70% of time series, as shown in Figure 1 in green color. However, several options to define the model mainly depend on the number of steps predicted and the order of the system. Therefore, we consider AR models with different order n to forecast the TB incidence between one to ten weeks. Additionally, a Kalman filter with covariance Q = 400 and R = 0.1 was employed for data estimation. The above, given that we assume unknown disturbances concerning imperfections from the identification process. Hence, it is considered a standard deviation of 20 TB cases. In this sense, we perform three heat maps considering several system sizes and weeks for each performance indices (9)-(11).

Figure 2 shows the heat map for the RMS values, the map relates step prediction forward m, and the α variable, which is proportional with system order, where $n = m + \alpha - 1$. Notice that the results with the lowest values in the graph have dark blue colors representing predictions with less error. Figure 3 and Figure 4 show the heat map for the MAE and TRS values, respectively. From the observation of the heat

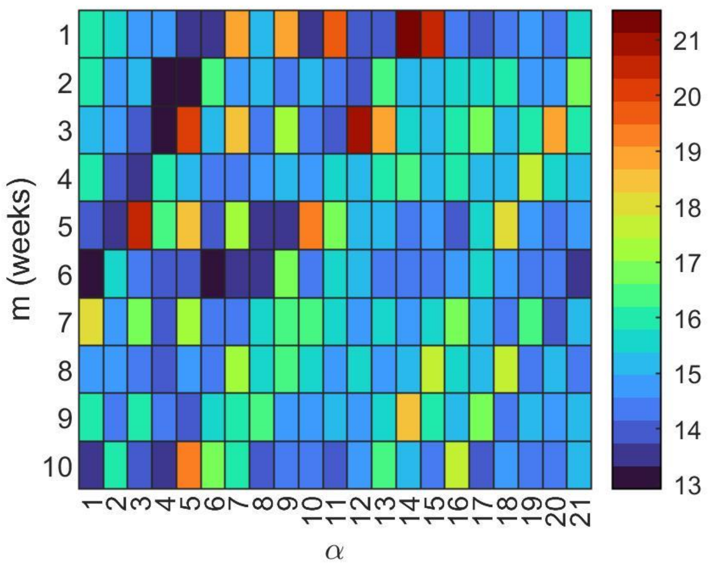

Figure 2. Heat map for the RMS values. Notice that n is the system order, m is the step prediction forward, and α is a variable related, where $n = m + \alpha - 1$.

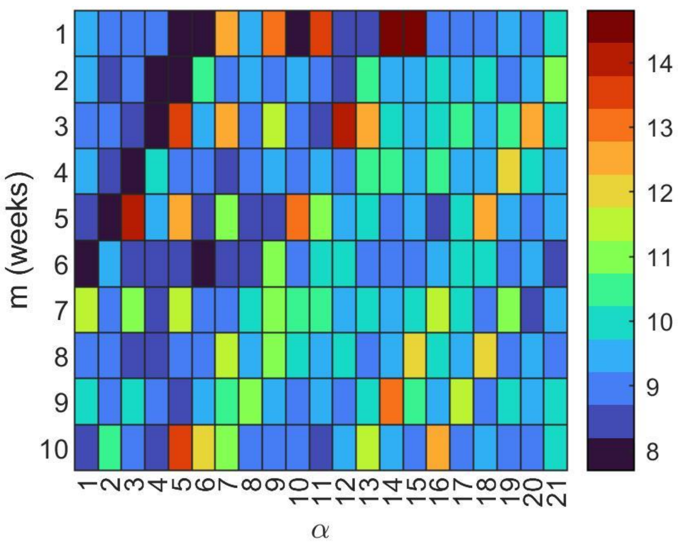

Figure 3. Heat map for the MAE values. Notice that n is the system order, m is the step prediction forward, and α is a variable related, where $n = m + \alpha - 1$.

maps, we select the system order $n = 6$ and the step prediction forward $m = 6$ with $\alpha = 1$. Since this is the case with better behavior, as shown in Figure 5.

Table 1 shows training results for six weeks TB cases prediction in Antioquia Department. The identification data serie has the number of confirmed case of pulmonary TB during 539 weeks. Notice that the KF model achieved similar índices RMSE and MAE in cooperation with the AR model. However, the TRS index is more reliable for the KF approach, lower by 30%. Fig. 6 shows the measured signal in green color and the forecasting TB cases given by the AR and KF in black and blue colors, respectively.

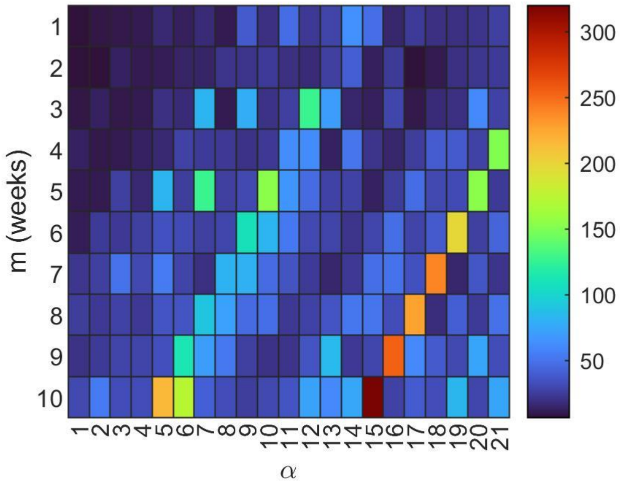

Figure 4. Heat map for the TRS values. Notice that n is the system order, m is the step prediction forward, and α is a variable related, where $n = m + \alpha - 1$.

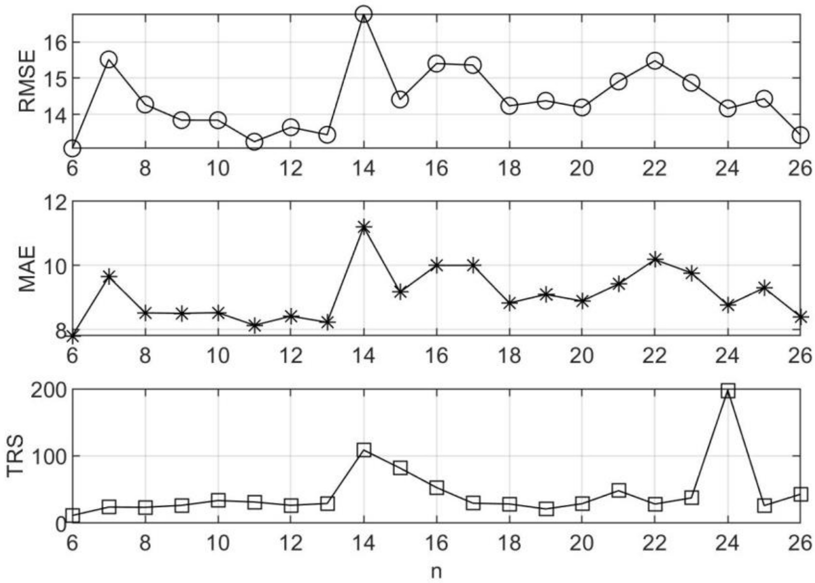

Figure 5. Selection of the order of model using the Kalman filter and N4SID algorithm. The selected order of model is $n = 6$

The validation data have the information of 232 weeks. Fig. 7 shows the measured signal in green color and the forecasting TB cases given by the AR and KF in black and blue colors, respectively. Notice that the KF and AR allowed forecasting the details from the original series related to the trend. The above is according to RMSE and MAE indices shown in Table 2, which are similar. However, looking at Fig. 7 is evident that prediction does not achieve a hit when the signal has abrupt changes and signal dispersion is higher. Notice that the TRS index is more suitable to rate this phenomenon. The above, since TRS measures how well the forecast is predicting the actual value. In this sense, the KF approach has a lower TRS than the AR, as shown in Table 2.

Table 1. Results of the training for six-week TB cases prediction.

	RMSE	*MAE*	*TRS*
KF	13.05	7.81	13.56
AR	12.51	7.31	43.57

Table 2. Results of the test for six-week TB cases prediction.

	RMSE	*MAE*	*TRS*
KF	9.22	7.41	3.34
AR	9.21	7.24	12.27

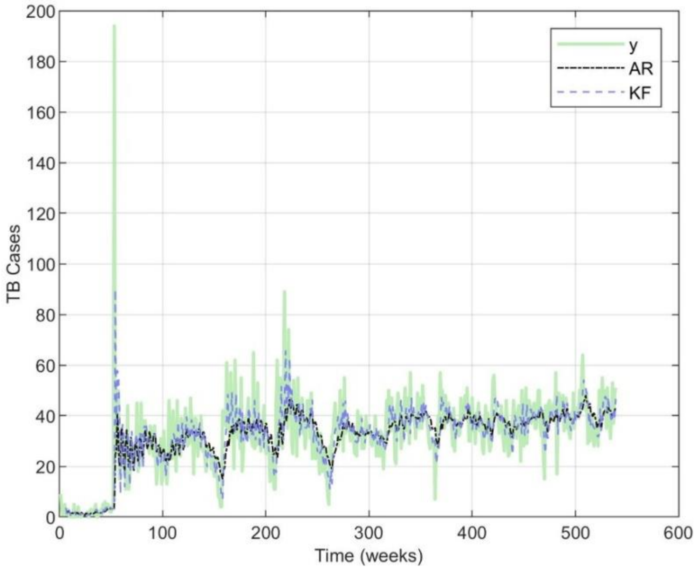

Figure 6. Results of the training for six-week TB cases prediction.

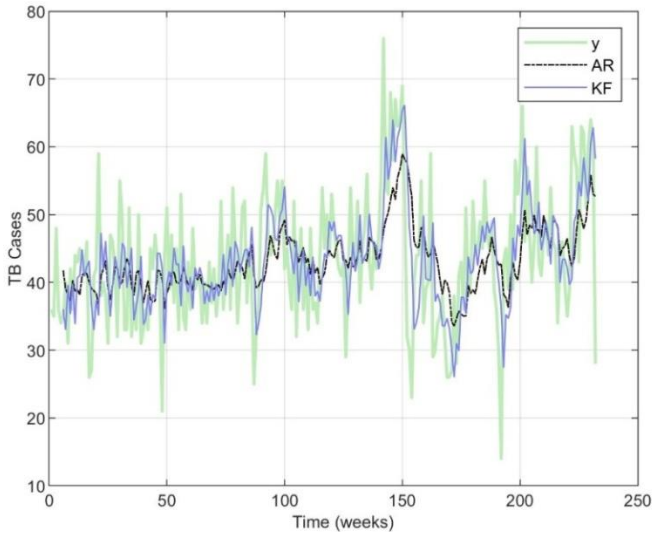

Figure 7. Results of the test for six-week TB cases prediction.

5. CONCLUSIONS

In this article, we made weekly forecasting of TB cases in Antioquia, Colombia. In addition, a comparative study between an Autoregressive Model and the Kalman filter was carried on. The KF outperformed the AR for the six-week TB case prediction accuracy. We select the model order and the step prediction via the RMSE, the MAE, and the TRS. We highlight that the six-week step forward could allow taking decisions for the health system for prevention. The prediction of pulmonary TB worldwide is very diverse and may vary due to economic, social, environmental, and legal factors.

With the unified effort of academic institutions and public health entities, we hope this study can contribute to accurately predicting confirmed cases of pulmonary TB in Colombia regions. Furthermore, we hope to study other methods to fit the incidence of TB, considering more extended periods forward. For future works, we intend to extend this study using deep learning networks, the Robust Kalman Filter for discrete-time Markovian jump linear systems [19], also the Unscented Kalman Filter, the Extended Kalman filter, and Particle filtering.

ACKNOWLEDGEMENTS

Authors acknowledge the support of the Ministerio de Ciencia y Tecnología – Minciencias from Colombia, through founded project 123380762899. In addition, institutions as Universidad Antonio Nariño, the Subred Integrada de Servicios de Salud Centro Oriente and Universidad del Rosario were relevant for the development of this work.

REFERENCES

[1] World Health Organization, *Global tuberculosis report 2021*. Geneva: World Health Organization, 2021.

[2] S. E. Fadul Perez, "Protocolo de vigilancia de tuberculosis," *Inst. Nac. salud*, pp. 1–29, 2022, [Online]. Available: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/PROTOCOLOS/Protocolo de Vigilancia de Tuberculosis.pdf.

[3] S. Fadul Perez, "Protocolo de vigilancia en salud pública Tuberculosis farmacorresistente," vol. 3, pp. 1–31, 2020.

[4] Instituto Nacional de Salud and Ministerio de Salud y Protección Social, "Comportamiento de la Vigilancia de Tuberculosis, Colombia,2020," *Bol. epidemiológico Sem.*, vol. 11, pp. 1–31, 2022.

[5] Y. Zheng, X. Zhang, X. Wang, K. Wang, and Y. Cui, "Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China," *BMJ Open*, vol. 11, no. 1, pp. 1–8, 2021, doi: 10.1136/bmjopen-2020-041040.

[6] A. L. Jutinico, E. Vergara, C. E. A. Garciá, M. A. Palencia, and A. D. Orjuela-Cañon, "Robust kalman filter for tuberculosis incidence time series forecasting," *IFAC-PapersOnLine*, vol. 54, no. 15, pp. 424–429, 2021, doi: 10.1016/j.ifacol.2021.10.293.

[7] A. D. Orjuela-Cañón, A. L. Jutinico, M. E. Duarte González, C. E. Awad García, E. Vergara, and M. A. Palencia, "Time series forecasting for tuberculosis incidence employing neural network models," *Heliyon*, vol. 8, no. 7, 2022, doi: 10.1016/j.heliyon.2022.e09897.

[8] M. M. Kavanagh, L. O. Gostin, and J. Stephens, "Tuberculosis, human rights, and law reform: Addressing the lack of progress in the global tuberculosis response," *PLoS Med.*, vol. 17, no. 10, pp. 1–9, 2020, doi: 10.1371/journal.pmed.1003324.

[9] R. C. M. Ribeiro *et al.*, "Forecasting incidence of tuberculosis cases in Brazil based on various univariate time-series models," *Int. J. Innov. Educ. Res.*, vol. 7, no. 10, pp. 894–909, 2019, doi: 10.31686/ijier.vol7.iss10.1841.

[10] Q. Liu *et al.*, "Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses," *Infect. Drug Resist.*, vol. 12, pp. 2311–2322, 2019, doi: 10.2147/IDR.S207809.

[11] K. W. Wang, C. Deng, J. P. Li, Y. Y. Zhang, X. Y. Li, and M. C. Wu, "Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network," *Epidemiol. Infect.*, vol. 145, no. 6, pp. 1118–1129, 2017, doi: 10.1017/S0950268816003216.

[12] A. Azeez, D. Obaromi, A. Odeyemi, J. Ndege, and R. Muntabayi, "Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model," *Int. J. Environ. Res. Public Health*, vol. 13, no. 8, 2016, doi: 10.3390/ijerph13080757.

[13] Y. Marimuthu, B. Nagappa, N. Sharma, S. Basu, and K. K. Chopra, "COVID-19 and tuberculosis: A mathematical model based forecasting in Delhi, India," *Indian J. Tuberc.*, vol. 67, no. 2, pp. 177–181, 2020, doi: 10.1016/j.ijtb.2020.05.006.

[14] P. Van Overschee and B. De Moor, "N4SID: Subspace Algorithms for the Stochastic Systemst," vol. 30, no. I, pp. 75–93, 1994.

[15] M. S. Grewal and A. P. Andrews, *Kalman Filtering : Theory and Practice Using MATLAB California State University at Fullerton*, vol. 5, no. 5. John Wiley & Sons, Inc, 2001.

[16] A. H. Sayed, *Adaptive Filters*. Hoboken: John Wiley & Sons, Inc, 2008.

[17] F. D. Hanggara, "Forecasting Car Demand in Indonesia with Moving Average Method.pdf," *J. Eng. Sci. Technol. Manag.*, vol. 1, 2021.

[18] M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, and V. A. Kamaev, "A Survey of Forecast Error Measures," *World Appl. Sci. J. 24*, pp. 171–176, 2013, doi: 10.5829/idosi.wasj.2013.24.itmies.80032.

[19] F. M. Escalante, A. L. Jutinico, J. C. Jaimes, M. H. Terra, and A. A. G. Siqueira, "Markovian Robust Filtering and Control Applied to Rehabilitation Robotics," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 1, pp. 491–502, 2021, doi: 10.1109/TMECH.2020.3034245.

AUTHORS

Ivan Camilo Sanchez-Vega received the bachelor's degree in Electronics Engineering from Universidad Popular del Cesar, Colombia, in 2017. He is currently working towards his M.Sc. degree in in instrumentation and automation at Universidad Antonio Nariño, Bogotá, Colombia. His current research interests include Kalman Filtering, robust control and filtering, and disease forecasting with autoregressive algorithms.

Andres L. Jutinico received the Bachelor in Electronic Engineering from The Universidad Distrital Francisco José de Caldas, Colombia, in 2005, the M.Sc. degree in Industrial Automation from the National University of Colombia, in 2012, and the Ph.D. degree in Mechanical Engineering from the University of Sao Paulo, Brazil, in 2019. His current research interests include control systems, computational intelligence and rehabilitation robotics.

Alvaro David Orjuela-Cañon received the B.Sc degree in Electronic Engineering in Bogotá, Colombia in 2006 from Universidad Distrital Francisco José de Caldas. M.Sc degree in Electrical Engineering from Universidade Federal de Rio de Janeiro (COPPE/UFRJ) in Brazil in 2009. In 2015 earned his Ph. D degree in COPPE/UFRJ with study subject related to support the diagnosis of pleural and meningeal Tuberculosis

Carlos Enrique Awad Garcia is a Colombian pulmonologist and university professor with more than three decades of service and several publications to his credit. He graduated as a Physician from Universidad Nacional de Colombia in 1983 and as a Pulmonologist from Universidad del Bosque in 1987.

Erika Vergara in 2005 received her Medical degree from Colegio Mayor Nuestra Señora del Rosario. In 2010 he received his Specialization in Internal Medicine. In 2013 she received his degree of Specialization in Clinical Epidemiology.

Maria Angelica Palencia received his Medical degree from Universidad de la Sabana in 2012. In 2016 he received his Specialization in Internal Medicine. In 2018 she received her master's degree in Epidemiology.