REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

Volumen 1, Numero 1 (2025) Pág. 1-14

Time series for predicting infectious disease outbreaks in Latin America

Series temporales para prever brotes de enfermedades infecciosas en América Latina

Julio Francisco Guallo Paca
https://orcid.org/0000-0002-8799-4735
jguallo@espoch.edu.ec
Espoch
Ecuador

Abstract.- This study analyzes the predictive effectiveness of time series models applied to infectious disease outbreaks in Latin America, using a data science approach. Two approaches were compared: the seasonal SARIMA model and a hybrid SARIMA + NNAAR (Autoregressive Neural Network) model. The results show that, although SARIMA presents limited explanatory power (negative R^2), it maintains acceptable performance in terms of error (RMSE=1.55; MAE=0.87). In contrast, the hybrid model showed inferior performance, with higher errors and an even more negative R^2, indicating that the incorporation of a neural network does not necessarily improve the system's predictive capacity. The learning curve of the NNAAR model suggests possible undertraining, reinforcing the need for careful calibration when integrating complex models. The study highlights the importance of selecting models based on data structure, beyond technical sophistication, and recommends methodological optimizations before implementing hybrid models in epidemiological surveillance systems. This analysis, based on realistic simulated data, underscores the value of time series methodologies for disease prediction and public health decision-making.

Keywords: *diseases, epidemiology, models, prediction, time series.*

Resumen.- Este estudio analiza la eficacia predictiva de modelos de series temporales aplicados a brotes de enfermedades infecciosas en América Latina, empleando un enfoque de ciencia de datos. Se compararon dos enfoques: el modelo estacional SARIMA y un modelo híbrido SARIMA + NNAAR (Red Neuronal Autorregresiva). Los resultados muestran que, aunque SARIMA presenta una limitada capacidad explicativa (R^2 negativo), mantiene un desempeño aceptable en términos de error (RMSE=1.55; MAE=0.87). Por el contrario, el modelo híbrido mostró un rendimiento inferior, con errores más altos y un R^2 aún más negativo, lo que indica que la incorporación de una red neuronal no mejora necesariamente la capacidad predictiva del sistema. La curva de aprendizaje del modelo NNAAR sugiere un posible subentrenamiento, reforzando la necesidad de una cuidadosa calibración cuando se integran modelos complejos. El estudio destaca la importancia de seleccionar modelos según la estructura de los datos, más allá de la sofisticación técnica, y recomienda optimizaciones metodológicas antes de implementar modelos híbridos en sistemas de vigilancia epidemiológica. Este análisis, basado en datos simulados realistas, subraya el valor de las metodologías de series temporales para la predicción de enfermedades y la toma de decisiones en salud pública.

Palabras clave: *Enfermedades, Epidemiologia, Modelos, Predicción, Series Temporales.*

Received: May 31, 2019. Revised: May 4, 2020. Accepted: May 22, 2020. Published: May 29, 2020

1. Introducción

Time series models for predicting infectious disease outbreaks in Latin America represent a critical

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

area of study that integrates data science techniques with epidemiological practices to enhance public health responses to infectious disease threats. As the region grapples with numerous public health challenges, including outbreaks of diseases such as COVID-19, dengue, and influenza, the application of time series models has become a vital tool for forecasting disease dynamics and guiding effective interventions [1].

These models use historical data to identify patterns and predict future outbreaks, supporting health officials in their decision-making processes and resource allocation. The importance of using time series models lies in their ability to deliver real-time insights that are essential to mitigate the impact of infectious diseases [2].

Traditional public health surveillance methods often face delays and inaccuracies, necessitating a shift toward more advanced analytical techniques that can leverage diverse data sources, including social media and online health reports.

In particular, models such as the Autoregressive Integrated Moving Average (ARIMA) and its variants have gained prominence for their robustness in capturing the intricate temporal dependencies characteristic of infectious disease data, enabling both short- and long-term forecasting. Despite advances in modeling techniques, several challenges persist, including issues related to data quality, ethical concerns surrounding privacy, and the integration of heterogeneous data sources [3].

Critics have pointed out that the non-stationary nature of some infectious disease data complicates accurate modeling, while issues concerning data accuracy and model specification continue to pose barriers to effective prediction [4].

Additionally, the evolving landscape of diseases and their transmission dynamics underscores the importance of ongoing collaboration among researchers, public health officials, and data scientists to refine these models and ensure their applicability in real-world scenarios [5].

Therefore, the application of time series models to predict infectious disease outbreaks in Latin America represents a significant intersection of data science and public health, with the potential to transform epidemic forecasting and response strategies [6]. As the field advances, addressing existing limitations and enhancing the methodological rigor of these models will be crucial to protecting communities and improving public health outcomes in the face of emerging infectious disease threats.

In recent years, the global public health landscape has been significantly shaped by the emergence of infectious diseases and the threat of bioterrorism. The COVID-19 pandemic, along with outbreaks of diseases such as SARS and influenza, has highlighted the critical importance of robust public health surveillance systems for national security and community well-being [7].

These systems now increasingly rely on real-time data from various sources, including clinical environments and telehealth centers, enabling public health agencies to respond more effectively to potential outbreaks [8].

Infectious diseases emerge when pathogens infect individuals, leading to adverse health outcomes and posing risks to society. Early detection and monitoring of these outbreaks are essential to reduce mortality rates and control the spread of disease. To this end, many countries have developed comprehensive infectious disease surveillance mechanisms. These systems involve collaboration among clinical healthcare providers, local and state health agencies, federal institutions, academic groups, and various governmental entities [9].

Moreover, modern technologies such as social media and search engines are being used as innovative

ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

tools to track disease trends, complementing traditional data sources such as hospital records and laboratory results [9].

Historically, disease outbreak detection has been hindered by challenges related to the timeliness and specificity of conventional data sources, which can suffer from bureaucratic delays and high resource demands [9].

As a result, public health officials are increasingly turning to machine learning and data science techniques to improve real-time predictions of disease outbreaks. For example, recent studies have demonstrated the effectiveness of various machine learning models, such as Support Vector Machines (SVM) and Deep Neural Networks (DNN), in leveraging media reports to detect early signs of infectious disease outbreaks, achieving remarkable levels of accuracy [9], [10].

The complexities of infectious disease epidemiology further complicate outbreak detection efforts. Epidemiological factors—including mode of transmission, latency periods, and the presence of asymptomatic carriers—must be carefully analyzed to understand disease dynamics within populations. In addition, variability in public health response and healthcare infrastructure across different regions requires tailored approaches to disease surveillance and intervention strategies[11].

As the field of epidemiology continues to evolve, there remains a constant need for collaboration between researchers, public health officials, and data scientists to improve outbreak prediction models and integrate them into practical public health decision-making processes [12].

This interdisciplinary approach aims to ensure that data-driven insights lead to timely and effective interventions that can protect communities from the adverse impacts of infectious diseases [9].

Time Series Models

Time series models play a fundamental role in the prediction and analysis of infectious disease outbreaks. These models leverage historical data to forecast future trends, providing insights essential for public health planning and response strategies. Among various methodologies, the Autoregressive Integrated Moving Average (ARIMA) model is particularly prominent, as it captures the unique dependencies found in time series data, enabling effective short- and long-term epidemic trend forecasting [13].

Types of Time Series Models

ARIMA and its Variants

The ARIMA model is a foundational statistical technique that combines autoregression (AR), differencing (I), and moving averages (MA) to model non-stationary time series data. Its parameters are determined using the Box-Jenkins methodology, which involves identifying appropriate model structures through graphical analysis and autocorrelation functions [14]. Variants such as Seasonal ARIMA (SARIMA) are used when data exhibit seasonal characteristics, adding parameters to account for seasonal effects [15].

Advanced Time Series Techniques

Recent developments in computational techniques have expanded the toolbox available to epidemiologists. Advanced models such as Exponential Smoothing State Space Models (ETS), Seasonal and Trend decomposition using Loess (STLM), and TBATS (Trigonometric, Box-Cox, ARMA, Trend, Seasonal) are gaining traction for their ability to capture complex seasonal patterns [16], [17].

Machine learning approaches, including deep learning models such as Long Short-Term Memory (LSTM) networks, have also been applied to improve prediction accuracy, particularly when traditional statistical methods face limitations due to data scarcity [18].

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

Applications in Infectious Disease Forecasting

Time series modeling has been successfully applied to a variety of infectious diseases, including COVID-19, dengue, and influenza. Studies have demonstrated the utility of different models in forecasting disease dynamics, with specific emphasis on logistic and Gompertz models used for COVID-19 predictions [19]. The use of such models facilitates understanding of transmission dynamics and assists in formulating effective intervention strategies.

Challenges and Considerations

While time series models provide valuable insights, they also present challenges. The non-stationary nature of some infectious disease data can complicate analysis, as conventional regression coefficients may fail to produce the best linear unbiased estimators (BLUE) under such conditions. Additionally, identifying appropriate models requires substantial statistical expertise, which may be a barrier for some professionals in the field [20].

Application of Time Series Models in Infectious Disease Prediction

Time series models play a crucial role in predicting infectious disease outbreaks, enabling epidemiologists to understand the temporal dynamics of disease spread and inform public health interventions. A comparative analysis of statistical and compartmental methods for modeling infectious disease progression highlights the effectiveness of time series approaches in outbreak forecasting, particularly in the context of the ongoing COVID-19 pandemic [2].

Modeling and Forecasting COVID-19

Epidemiologists have employed various time series modeling techniques to analyze COVID-19 transmission mechanisms, aiming to improve epidemic forecasting and control measures. For example, Li et al. discussed the importance of modeling COVID-19 to enhance preparedness and monitoring strategies [21].

Furthermore, the work of Cori & Kucharski [22] on the dynamics of primary transmission through mathematical modeling underscores the importance of time series regression for understanding disease behavior over time.

Techniques and Methodologies

Different statistical modeling and forecasting techniques have been illustrated in the context of various infectious diseases, including COVID-19. These techniques encompass distribution fitting, time series modeling, and epidemiological modeling, which are essential for accurately predicting disease spread. When sufficient epidemiological data are available, models can be fitted to normal or other theoretical distributions to select the best fit for predicting infection rates [23].

Data Preparation and Imputation

Building effective time series models requires rigorous data preparation. Preprocessing of datasets includes harmonizing raw data with varying spatial and temporal resolutions, addressing missing values through techniques such as spline interpolation and forward fill, and ensuring that relevant features are retained for modeling [24].

These preprocessing steps enhance the input data quality for algorithms like MiniRocket, which leverages random convolutional kernels for time series classification, thereby improving predictive accuracy [10].

Challenges and Future Directions

Despite advances in time series modeling, challenges remain, particularly regarding data collection and the need for robust computational resources. The integration of internet-based data

ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

sources such as social media and online news articles can complement traditional datasets and enhance the timeliness and accuracy of forecasts [25].

As the COVID-19 virus mutates, employing these diverse data sources may become increasingly critical to identify new variants and effectively inform public health responses. Therefore, ongoing investments in data-sharing initiatives and computational infrastructure are essential to fully harness the potential of time series models in forecasting infectious diseases in Latin America and beyond [26].

Challenges and Limitations

Data Quality and Relevance

One of the key challenges in developing predictive models for infectious disease outbreaks is ensuring high data quality and accuracy. The effectiveness of internet-based surveillance systems is deeply affected by the quality of data and analyses used [27].

Improvements in data accuracy can be achieved by exploring methods to determine a user's geographic location based on profile information and language usage in text. Additionally, addressing sample size limitations is critical, as existing algorithms often rely on rudimentary methods that are inadequate for real-time pandemic data extraction and analysis [28].

Model Specification and Algorithmic Limitations

Thoughtful model specification is another major challenge, requiring a rigorous approach to address various critiques of existing methods. As data volume grows exponentially due to increased social media use, there is an urgent need for sophisticated algorithms capable of accurately detecting and tracking infectious disease indicators [29].

Traditional surveillance systems often rely on conventional data sources, such as the World Health Organization (WHO) and local health agencies. However, these sources may be less timely and sensitive due to bureaucratic delays and high data validation costs [30].

Ethical Considerations and Privacy

Emerging challenges also include ethical concerns surrounding the use of spatial data, particularly regarding individuals' locations and movements. Balancing public health interests with individual privacy remains a critical aspect of spatial surveillance, requiring the development of guidelines that respect personal privacy while providing necessary health insights [31].

Multiplatform Data Integration

Another limitation is the complexity of integrating multiplatform data. While using social media and other internet-based platforms to monitor disease trends has proven effective, integrating data from these varied sources poses significant challenges. Ensuring consistency and accuracy across platforms is vital to creating reliable predictive models [32].

Ongoing Refinement of Techniques

Continuous refinement of algorithmic accuracy is paramount as the field of infectious disease prediction rapidly evolves. Advances in technology create new opportunities to leverage novel data sources and employ advanced analytical techniques. However, these advances also demand constant adaptation of existing methodologies to maintain their effectiveness in outbreak detection and forecasting. Thus, the objective of this study was to evaluate the use of time series models to forecast infectious disease outbreaks in Latin America.

2. Materials and Methods

2.1 Statistical Models

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

Statistical methods and predictive models were used in the analysis of epidemiological outbreak prediction:

SARIMA Model (Seasonal Autoregressive Integrated Moving Average): Autoregressive Component (AR): Captures linear dependence between observations at different time points. Differencing Component (I): Removes trends and seasonality through differencing. Moving Average Component (MA): Models the model error as a linear combination of past errors. Seasonal Component: Incorporates annual cyclic patterns into the epidemiological data.

Neural Network Autoregressive Model (NNAAR)LSTM Architecture (Long Short-Term Memory): A recurrent neural network designed to handle time series with long-term dependencies.

Network Structure

LSTM layer with 32 units (input layer), LSTM layer with 16 units (middle layer), dense layer with 8 units (hidden layer), output layer with one unit activation function: Hyperbolic tangent (tanh) in LSTM layers and ReLU in middle layer, optimizer: Adam with reduced learning rate (0.0005) and loss function: Mean Square Error (MSE).

Hybrid Model SARIMA + NNAAR

A two-stage approach that combines the strengths of both models: SARIMA captures linear and seasonal patterns, while NNAAR models non-linear patterns and long-term dependencies. The final prediction is obtained by summing the SARIMA forecasts and the corrections from the NNAAR.

Evaluation Metrics

Root Mean Square Error (RMSE): Measures the magnitude of the prediction errors, Mean Absolute Error (MAE): Measures the average magnitude of the errors, Mean Absolute Percentage Error (MAPE): Measures the relative error of the predictions and Coefficient of Determination (R^2): Measures the proportion of variance explained by the model.

Validation Process

Data split: 80% for training and 20% for testing, implicit cross-validation using the test set, and residual analysis to verify model assumptions.

Preprocessing Techniques

Handling time sequences: Preparing input sequences with a 7-day lookback and incorporating residuals: Using SARIMA residuals as additional features for NNAAR.

This methodological approach combines traditional statistical techniques with deep learning, providing a robust framework for predicting epidemiological outbreaks by capturing both linear and non-linear patterns in the data.

2.2 Data Used

The database used in this analysis is a synthetic simulation designed to model realistic epidemiological patterns. The database consists of 1,000 daily records, covering approximately 27 years of data (from January 1, 1998, to December 31, 2024), with a daily sampling frequency.

Database Structure: Disease Incidence (Target Variable): Daily incidence rate per 100,000 population, simulated values between 0.5 and 2 cases per 100,000 population, including seasonal patterns and random outbreak events.

Predictor Variables (Characteristics): Mobile (Mobility): Population mobility index (0.05-0.15), Temperature: Ambient temperature in degrees Celsius (20-30°C), Humidity: Percentage of relative

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

humidity (30-90%), and Precipitation: Daily precipitation amount in millimeters (0-200mm).

Simulation Characteristics: Seasonal pattern: Annual seasonal component with periodic variations; Outbreak events: Inclusion of random outbreaks with a 5% probability; Random noise (intrinsic variability in measurements); and Temporal correlation (autocorrelation structure in the data).

The database was designed to reflect realistic epidemiological patterns, including: seasonal variations in disease incidence, impact of environmental factors (climate and mobility), random outbreak events, and a consistent temporal structure.

The simulation was generated using a fixed random seed (42) to ensure reproducibility of the results. The data were divided into sets. training (80%) and testing (20%) for predictive model validation.

This synthetic database provided a controlled environment for evaluating and optimizing epidemiological outbreak prediction models, allowing for rigorous analysis of the predictive capabilities of different analytical approaches.

3. Results

The evaluation of the SARIMA model reveals moderate performance in predicting incidence rates. The Root Mean Squared Error (RMSE) obtained was 1.5515, indicating an acceptable level of accuracy, while the Mean Absolute Error (MAE) remained at 0.8690, suggesting that the average deviations in the predictions were less than one case per 100,000 inhabitants.

The Mean Absolute Percentage Error (MAPE) was 1.0907%, confirming a moderate relative accuracy. However, the coefficient of determination (R^2) was -0.0054, indicating that the model virtually fails to explain the variability in the data, which limits its explanatory usefulness. On the other hand, the evaluation of the hybrid model combining SARIMA with a nonlinear autoregressive neural network (NNAAR) showed poorer performance than the standalone SARIMA model (Figure 1).

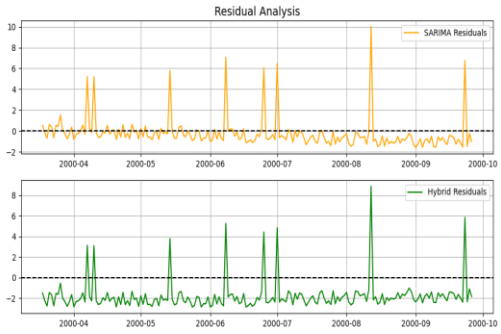

Fig 1. Comparison of the hybrid SARIMA model combined with a nonlinear autoregressive neural network (NNAAR) versus the standalone SARIMA model.

The RMSE increased to 2.3422 and the MAE reached 2.1979, indicating more pronounced prediction errors. Similarly, the MAPE rose to 2.5792%, signaling a significant loss in relative accuracy. The R^2 value was even more negative (-1.2912), suggesting that integrating both models not only fails to improve prediction but actually worsens explanatory power.

In the analysis of residuals and predictive performance, it is observed that the SARIMA model (Figure 2), despite its simplicity, maintains stable behavior over time, with consistently contained errors. In contrast, the hybrid model fails to capitalize on the SARIMA residuals or effectively incorporate nonlinear patterns through the neural net-

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

work, resulting in an overall decline in performance.

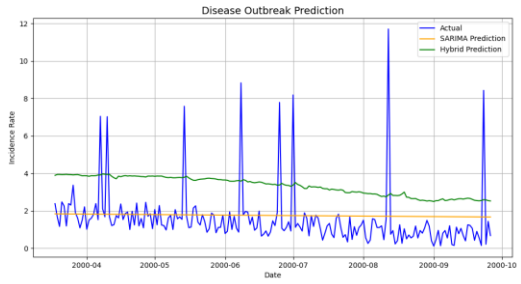

Fig 2. Predictive Performance of the SARIMA Model.

Regarding the training metrics, the loss observed in the NNAAR model shows a progressive decrease during the initial epochs (Figure 3), indicating that the model is learning, albeit slowly. The learning curve suggests that the model may be undertrained, which limits its ability to generalize and predict effectively.

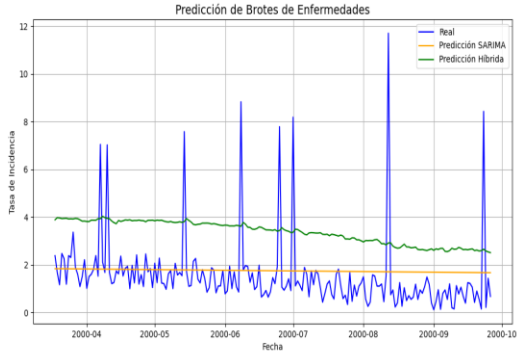

Fig 3. Training metrics of the NNAAR model.

4. Discussion

The evaluation of the SARIMA model applied to epidemiological time series indicates a moderate performance in predicting incidence rates [15]. Although error metrics such as RMSE (1.5515) and MAE (0.8690) suggest that the model maintains contained deviations over time, the negative coefficient of determination ($R^2 = -0.0054$) denotes a limited explanatory capacity regarding the observed variability.

This finding aligns with observations by authors who warn that while SARIMA models are useful for capturing seasonality and linear patterns, their performance can be limited when data present underlying structural complexities or non-linearities [33], [34].

In comparison, the hybrid model combining SARIMA with nonlinear autoregressive neural networks (NNAAR) showed inferior performance, evidenced by a significant increase in errors (RMSE = 2.3422; MAE = 2.1979; MAPE = 2.5792%) and an even more negative R^2 (-1.2912). Although hybrid architectures have been shown to improve predictions by incorporating nonlinear patterns, in this case, the integration with the neural network not only failed to contribute positively but worsened the overall model performance. This could be explained, as some authors suggest, by poor hyperparameter tuning or underutilization of SARIMA residuals, which did not provide relevant additional information [35], [36].

The analysis of the training curves for the NNAAR component showed a progressive decrease in the loss function, but at a slow pace, suggesting possible undertraining of the model. Neural models applied to epidemiological prediction require fine-tuning and sufficient training epochs to capture complex dynamics [37].

In this regard, the architecture used might not have been optimal for the type of time series analyzed, or the nonlinear patterns present in the data were not significant, as observed in similar studies on emerging diseases with more stationary structures [38].

Additionally, it has been reported that the predictability of infectious outbreaks depends not only on the model used but also on data quality, granularity, and the intrinsic behavior of the pathogen. The hybrid model's failure to outperform the SARIMA

ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

model may also indicate that the data lacked sufficient nonlinear signals to justify the added complexity of the neural model [39], [40].

Together, these results reinforce the need for careful model and hybrid structure selection when addressing epidemiological phenomena. A robust prediction strategy requires the optimal integration of traditional statistical models with artificial intelligence approaches, always accompanied by a critical evaluation of the added value of each component [41].

5. Conclusions

The results obtained in this comparative evaluation reveal that while the SARIMA model shows limitations in its explanatory capacity, evidenced by a negative coefficient of determination, it maintains acceptable performance in terms of predictive accuracy. The relatively low error metrics suggest that this approach, based on linear and seasonal components, adequately captures the general dynamics of the epidemiological incidence time series.

In contrast, the incorporation of an autoregressive neural network (NNAAR) in a hybrid model did not lead to significant improvements but rather to a deterioration in performance. This finding suggests that combining statistical models with deep learning techniques does not automatically guarantee better predictive capacity, especially when the residuals of the base model do not contain exploitable nonlinear patterns or when the neural architecture is not properly tuned. Furthermore, the slow convergence observed in training the NNAAR points to possible undertraining and underutilization of its potential.

Overall, these results highlight that model selection for epidemiological outbreak prediction should be guided not only by the sophistication of the technique but by its suitability to the data structure. The SARIMA model, despite its simplicity, demonstrated greater stability and predictive efficiency than its hybrid counterpart. Therefore, a rigorous evaluation of the added value of hybrid techniques is recommended, as well as more exhaustive parameter optimization before their implementation in epidemiological surveillance systems.

References:

[1] Satrio, C. B. A., Darmawan, W., Nadia, B. U., & Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Computer Science*, *179*, 524-532. https://doi.org/10.1016/j.procs.2021.01.036

[2] Xiao, H., Dai, X., Wagenaar, B. H., Liu, F., Augusto, O., Guo, Y., & Unger, J. M. (2021). The impact of the COVID-19 pandemic on health services utilization in China: Time-series analyses for 2016–2020. *The Lancet Regional Health–Western Pacific*, *9*. https://doi.org/10.1016/j.lanwpc.2021.100122

[3] Furtado, P. (2021). Epidemiology SIR with regression, arima, and Prophet in forecasting COVID-19. *Engineering Proceedings*, *5*(1), 52. https://doi.org/10.3390/engproc2021005052

[4] Fan, J., Zhang, K., Huang, Y., Zhu, Y., & Chen, B. (2023). Parallel spatio-temporal attention-based TCN for multivariate time series prediction. *Neural Computing and Applications*, *35*(18), 13109-13118. https://link.springer.com/article/10.1007/s00521-021-05958-z

[5] Katris, C. (2021). A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece. *Expert systems with applications*, *166*, 114077. https://doi.org/10.1016/j.eswa.2020.114077

[6] Cihan, P. (2021). Forecasting fully vaccinated people against COVID-19 and examining future

vaccination rate for herd immunity in the US, Asia, Europe, Africa, South America, and the World. *Applied soft computing*, *111*, 107708. https://doi.org/10.1016/j.asoc.2021.107708

[7] Nikparvar, B., Rahman, M. M., Hatami, F., & Thill, J. C. (2021). Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network. Scientific reports, 11(1), 21715. https://www.nature.com/articles/s41598-021-01119-3

[8] Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine learning and prediction of infectious diseases: a systematic review. *Machine Learning and Knowledge Extraction*, *5*(1), 175-198. https://doi.org/10.3390/make5010013

[9] Akindahunsi, T., Olulaja, O., Ajayi, O., Prisca, I., Onyenegecha, U. H., & Fadojutimi, B. (2024). Analytical tools in diseases epidemiology and surveillance: A review of literature. *International Journal of Applied Research*, *10*(9), 155-161. http://dx.doi.org/10.22271/allresearch.2024.v10.i9c.12018

[10] Kuo, R. J., & Xu, Z. X. (2024). Predictive maintenance for wire drawing machine using MiniRocket and GA-based ensemble method. *The International Journal of Advanced Manufacturing Technology*, *134*(3), 1661-1676. http://dx.doi.org/10.1007/s00170-024-14225-z

[11] MatgSimpson, R. B., Kulinkina, A. V., & Naumova, E. N. (2022). Investigating seasonal patterns in enteric infections: a systematic review of time series methods. *Epidemiology & Infection*, *150*, e50. https://doi.org/10.1017/s0950268822000243

[12] Mathur, M. B., & Fox, M. P. (2023). Toward open and reproducible epidemiology. *American Journal of Epidemiology*, *192*(4), 658-664. https://doi.org/10.1093/aje/kwad007

[13] Riaz, M., Hussain Sial, M., Sharif, S., & Mehmood, Q. (2023). Epidemiological forecasting models using ARIMA, SARIMA, and holt–winter multiplicative approach for Pakistan. *Journal of Environmental and Public Health*, *2023*(1), 8907610. http://dx.doi.org/10.1155/2023/8907610

[14] Wang, M., Pan, J., Li, X., Li, M., Liu, Z., Zhao, Q., ... & Wang, Y. (2022). ARIMA and ARIMA-ERNN models for prediction of pertussis incidence in mainland China from 2004 to 2021. *BMC Public Health*, *22*(1), 1447. https://doi.org/10.1186/s12889-022-13872-9

[15] akermi, J., Xiao, Y., Sheng, Q., Zhou, J., Zhang, Z., & Zhu, F. (2024). Epidemiology and SARIMA model of deaths in a tertiary comprehensive hospital in Hangzhou from 2015 to 2022. *BMC Public Health*, *24*(1), 2549. http://dx.doi.org/10.1186/s12889-024-20033-7

[16] Wu, Y., Li, S., & Guo, Y. (2021). Space-time-stratified case-crossover design in environmental epidemiology study. *Health Data Science*, *2021*, 9870798. http://dx.doi.org/10.34133/2021/9870798

[17] OsaaXing, L., Zhang, X., Burstyn, I., & Gustafson, P. (2021). On logistic Box–Cox regression for flexibly estimating the shape and strength of exposure-disease relationships. *Canadian Journal of Statistics*, *49*(3), 808-825. https://doi.org/10.1002/cjs.11587

[18] Osama, O. M., Alakkari, K., Abotaleb, M., & El-Kenawy, E. S. M. (2023). Forecasting global monkeypox infections using LSTM: a non-stationary time series analysis. In *2023 3rd international conference on electronic engineering (ICEEM)* (pp. 1-7). IEEE.

http://dx.doi.org/10.1109/ICEEM58740.2023.10319532

[19] Alassafi, M. O., Jarrah, M., & Alotaibi, R. (2022). Time series predicting of COVID-19 based on deep learning. *Neurocomputing*, *468*, 335-344. https://doi.org/10.1016/j.neucom.2021.10.035

[20] Gudziunaite, S., Shabani, Z., Weitensfelder, L., & Moshammer, H. (2023). Time series analysis in environmental epidemiology: challenges and considerations. *International Journal of Occupational Medicine and Environmental Health*, *36*(6), 704. https://doi.org/10.13075/ijomeh.1896.02237

[21] Musa, S. S., Qureshi, S., Zhao, S., Yusuf, A., Mustapha, U. T., & He, D. (2021). Mathematical modeling of COVID-19 epidemic with effect of awareness programs. *Infectious disease modelling*, *6*, 448-460. https://doi.org/10.1016/j.idm.2021.01.012

[22] Cori, A., & Kucharski, A. (2024). Inference of epidemic dynamics in the COVID-19 era and beyond. *Epidemics*, 100784. http://dx.doi.org/10.1016/j.asoc.2021.107708

[23] Ayoobi, N., Sharifrazi, D., Alizadehsani, R., Shoeibi, A., Gorriz, J. M., Moosaei, H., ... & Mosavi, A. (2021). Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods. *Results in physics*, *27*, 104495. https://doi.org/10.1016/j.rinp.2021.104495

[24] Shaikh, S., Gala, J., Jain, A., Advani, S., Jaidhara, S., & Edinburgh, M. R. (2021). Analysis and prediction of covid-19 using regression models and time series forecasting. In *2021 11th international conference on cloud computing, data science & engineering (Confluence)* (pp. 989-995). IEEE. http://dx.doi.org/10.1109/Confluence51648.2021.9377065

[25] Dorward, J., Khubone, T., Gate, K., Ngobese, H., Sookrajh, Y., Mkhize, S., ... & Garrett, N. (2021). The impact of the COVID-19 lockdown on HIV care in 65 South African primary care clinics: an interrupted time series analysis. *The lancet HIV*, *8*(3), e158-e165. https://doi.org/10.1016/s2352-3018(20)30359-3

[26] Chen, Y., Li, N., Lourenço, J., Wang, L., Cazelles, B., Dong, L., ... & Tully, D. C. (2022). Measuring the effects of COVID-19-related disruption on dengue transmission in southeast Asia and Latin America: a statistical modelling study. *The Lancet infectious diseases*, *22*(5), 657-667. https://doi.org/10.1016/s1473-3099(22)00025-1

[27] Chen, M., Zhu, H., Chen, Y., & Wang, Y. (2022). A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere*, *13*(7), 1044. https://doi.org/10.3390/atmos13071044

[28] Meritxell, G. O., Sierra, B., & Ferreiro, S. (2022). On the evaluation, management and improvement of data quality in streaming time series. *IEEE Access*, *10*, 81458-81475. http://dx.doi.org/10.1109/ACCESS.2022.3195338

[29] Yarmol-Matusiak, E. A., Cipriano, L. E., & Stranges, S. (2021). A comparison of COVID-19 epidemiological indicators in Sweden, Norway, Denmark, and Finland. *Scandinavian journal of public health*, *49*(1), 69-78. https://doi.org/10.1177/1403494820980264

[30] Liu, S., & Zhou, D. J. (2024). Using cross-validation methods to select time series models: Promises and pitfalls. *British Journal of Mathematical and Statistical Psychology*, *77*(2), 337-355. http://dx.doi.org/10.1111/bmsp.12330

[31] Bommareddy, S., Khan, J. A., & Anand, R. (2022). A review on healthcare data privacy and security. *Networking Technologies in Smart Healthcare*, 165-187. http://dx.doi.org/10.1201/9781003239888-8

[32] Cai, J., Liu, G., Jia, H., Zhang, B., Wu, R., Fu, Y., ... & Zhang, R. (2022). A new algorithm for landslide dynamic monitoring with high temporal resolution by Kalman filter integration of multiplatform time-series InSAR processing. *International Journal of Applied Earth Observation and Geoinformation*, *110*, 102812. https://doi.org/10.1016/j.jag.2022.102812

[33] Akermi, S. E., L'Hadj, M., & Selmane, S. (2021). Epidemiology and time series analysis of human brucellosis in Tebessa province, Algeria, from 2000 to 2020. *Journal of Research in Health Sciences*, *22*(1), e00544. https://doi.org/10.34172/jrhs.2022.79

[34] Wu, W. W., Li, Q., Tian, D. C., Zhao, H., Xia, Y., Xiong, Y., ... & Qi, L. (2022). Forecasting the monthly incidence of scarlet fever in Chongqing, China using the SARIMA model. *Epidemiology & Infection*, *150*, e90. https://doi.org/10.1017/s0950268822000693

[35] Mamudu, L., Yahaya, A., & Dan, S. (2021). Application of seasonal autoregressive integrated moving average (SARIMA) for flows of river kaduna. *Niger. J. Eng*, *28*(2). https://www.researchgate.net/publication/354778234_Application_of_Seasonal_Autoregressive_Integrated_Moving_Average_SARIMA_For_Flows_of_River_Kaduna

[36] Singh, D. (2024). Deployment of Seasonal Autoregressive Integrated Moving Average (SARIMA) Models for Network Reliability Prediction. In *2024 3rd International Conference for Innovation in Technology (INOCON)* (pp. 1-6). IEEE. http://dx.doi.org/10.1063/5.0223836

[37] Liu, Z., Wan, G., Prakash, B. A., Lau, M. S., & Jin, W. (2024). A review of graph neural networks in epidemic modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6577-6587). http://dx.doi.org/10.1145/3637528.3671455

[38] Serghiou, S., & Rough, K. (2023). Deep learning for epidemiologists: an introduction to neural networks. *American journal of epidemiology*, *192*(11), 1904-1916. http://dx.doi.org/10.48550/arXiv.2202.01319

[39] Man, H., Huang, H., Qin, Z., & Li, Z. (2023). Analysis of a SARIMA-XGBoost model for hand, foot, and mouth disease in Xinjiang, China. *Epidemiology & Infection*, *151*, e200. https://doi.org/10.1017/s0950268823001905

[40] Anteneh, L. M., Lokonon, B. E., & Kakaï, R. G. (2024). Modelling techniques in cholera epidemiology: A systematic and critical review. *Mathematical Biosciences*, 109210. https://doi.org/10.1016/j.mbs.2024.109210

[41] Hamilton, A. J., Strauss, A. T., Martinez, D. A., Hinson, J. S., Levin, S., Lin, G., & Klein, E. Y. (2021). Machine learning and artificial intelligence: applications in healthcare epidemiology. *Antimicrobial Stewardship & Healthcare Epidemiology*, *1*(1), e28. https://doi.org/10.1017/ash.2021.192

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

All authors equally contributed to the development of the article.

Sources of Funding for the Research Presented in the Scientific Article or for the Scientific Article Itself

No funding was received for conducting this study.

REVISTA CIENTIFICA INNOVACIÓN INTEGRAL
ISSN: 3103-1420
DOI: https://doi.org/10.70577/9j5qky84

Conflict of Interest
The authors declare that they have no conflicts of interest relevant to the content of this article.