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

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

1. Introducción

Los modelos de series de tiempo para predecir brotes de enfermedades infecciosas en América Latina es un área crítica de estudio que integra técnicas de ciencia de datos con prácticas epidemiológicas para mejorar las respuestas de salud pública a las amenazas de enfermedades infecciosas. A medida que la región lidia con numerosos desafíos de salud pública, incluidos brotes de enfermedades como Covid-19, dengue e influenza, la aplicación de modelos de series temporales se ha convertido en una herramienta vital para pronosticar la dinámica de las enfermedades y guiar intervenciones efectivas [1].

Estos modelos utilizan datos históricos para identificar patrones y predecir brotes futuros, lo que ayuda a los funcionarios de salud en sus procesos de toma de decisiones y su asignación de recursos. La importancia de emplear modelos de series de tiempo radica en su capacidad para ofrecer información en tiempo real que sean esenciales para mitigar el impacto de las enfermedades infecciosas [2].

Los métodos tradicionales de vigilancia de salud pública a menudo enfrentan retrasos e inexactitudes, lo que requiere un cambio hacia un análisis más avanzado Técnicas que pueden aprovechar diversas fuentes de datos, incluidas las redes sociales e informes de salud en línea.

En particular, los modelos como el promedio móvil integrado autorregresivo (ARIMA) y sus variantes han ganado prominencia por su robustez en la captura de las intrincadas dependencias temporales características de los datos de enfermedades infecciosas, lo que permite pronósticos a corto y largo plazo. A pesar de los avances en las técnicas de modelado, persisten varios desafíos, incluidas las complejidades de la calidad de los datos, las consideraciones éticas con respecto a la privacidad y la integración de fuentes de datos heterogéneas [3].

Los críticos destacan que la naturaleza no estacionaria de algunos datos de enfermedades infecciosas complica el modelado preciso, mientras que los problemas relacionados con la precisión de los datos y la especificación del modelo continúan planteando barreras para una predicción efectiva [4].

Además, el panorama en evolución de las enfermedades y su dinámica de transmisión subraya la importancia de la colaboración en curso entre los investigadores, los funcionarios de salud pública, y científicos de datos para refinar estos modelos y garantizar su aplicabilidad en escenarios del mundo real [5].

Es por ello, que la aplicación de modelos de series temporales para predecir los brotes de enfermedades infecciosas en América Latina representa una intersección significativa de la ciencia de datos y la salud pública, con el potencial de transformar las estrategias de pronósticos y respuesta epidémicos [6]. A medida que avanza el campo, abordar las limitaciones existentes y mejorar el rigor metodológico de estos modelos será crucial para proteger a las comunidades y mejorar los resultados de salud pública frente a las amenazas emergentes de enfermedades infecciosas.

En los últimos años, el panorama global de la salud pública se ha configurado significativamente por la aparición de enfermedades infecciosas y la amenaza del bioterrorismo. La pandemia Covid-19, junto con brotes de enfermedades como el SAR e influenza, ha destacado la importancia crítica de los robustos sistemas de vigilancia de salud pública para la seguridad nacional y el bienestar de la comunidad [7].

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

Estos sistemas ahora dependen cada vez más de datos en tiempo real de diversas fuentes, incluidos entornos clínicos y centros de telesalud, lo que permite a las agencias de salud pública responder de manera más efectiva a los posibles brotes [8].

Las enfermedades infecciosas surgen cuando los patógenos infectan a las personas, lo que lleva a efectos adversos sobre la salud y la planeación de riesgos para la sociedad. La detección y el monitoreo temprano de estos brotes son esenciales para mitigar las tasas de mortalidad y controlar la propagación de la enfermedad. Con este fin, muchos países han desarrollado mecanismos integrales de vigilancia de enfermedades infecciosas. Estos sistemas implican la colaboración entre los proveedores de atención médica clínica, las agencias de salud locales y estatales, las instituciones federales, los grupos académicos y varias entidades gubernamentales [9].

Además, la tecnología moderna, como las redes sociales y los motores de búsqueda, se está utilizando como herramientas innovadoras para rastrear las tendencias de las enfermedades, complementando las fuentes de datos tradicionales como los registros hospitalarios y el laboratorio. Resultados [9].

Históricamente, la detección de brotes de enfermedades se ha visto obstaculizada por desafíos relacionados con la puntualidad y especificidad de las fuentes de datos convencionales, que pueden sufrir retrasos burocráticos y altas demandas de recursos [9].

Como resultado, los funcionarios de salud pública están recurriendo cada vez más a las técnicas de aprendizaje automático y ciencia de datos para mejorar las predicciones en tiempo real de los brotes de enfermedades. Por ejemplo, estudios recientes han demostrado la eficacia de varios modelos de aprendizaje automático, como la máquina de vectores de soporte (SVM) y las redes neuronales profundas (DNN), en la utilización de informes de medios para detectar los primeros signos de brotes de enfermedades infecciosas, logrando niveles de precisión notables [9], [10].

Las complejidades de la epidemiología de enfermedades infecciosas complican aún más los esfuerzos de detección de brotes. Los factores epidemiológicos, incluido el modo de transmisión, los períodos latentes y la presencia de portadores asintomáticos, deben analizarse cuidadosamente para comprender la dinámica de la enfermedad entre poblaciones. Además, la variabilidad en la respuesta de salud pública y la infraestructura de atención médica en diferentes regiones requiere enfoques personalizados para la vigilancia de la enfermedad y las estrategias de intervención [11].

A medida que el campo de la epidemiología continúa evolucionando, existe una necesidad continua de colaboración entre investigadores, funcionarios de salud pública y científicos de datos para mejorar los modelos de predicción de brotes e integrarlos en procesos prácticos de toma de decisiones de salud pública [12].

Este enfoque interdisciplinario tiene como objetivo garantizar que las ideas basadas en datos conduzcan a intervenciones oportunas y efectivas que puedan proteger a las comunidades de los impactos adversos de las enfermedades infecciosas [9].

Modelos de series de tiempo

Los modelos de series de tiempo juegan un papel fundamental en la predicción y el análisis de los brotes de enfermedades infecciosas. Estos modelos aprovechan los datos históricos para pronosticar tendencias futuras, ofreciendo ideas que son esenciales para la planificación de la salud pública y las estrategias de respuesta. Entre los Varias metodologías, el modelo de promedio móvil integrado autorregresivo (ARIMA) es particularmente prominente, ya que captura las dependencias únicas que se encuentran en los datos de la serie temporal, lo que permite pronósticos

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

efectivos de las tendencias epidémicas a corto y largo plazo [13].

Tipos de modelos de series de tiempo

Arima y sus variantes

El modelo ARIMA es una técnica estadística fundamental que combina la autorregresión (AR), la diferencia (I) y los promedios móviles (MA) para modelar datos de series temporales no estacionarias. Sus parámetros se determinan utilizando la metodología Box-Jenkins, que implica identificar estructuras de modelo apropiadas a través de análisis gráficos y funciones de autocorrelación [14]. Variantes como la ARIMA estacional (Sarima) se utilizan cuando los datos exhiben características estacionales, agregando parámetros adicionales para tener en cuenta los efectos estacionales [15].

Técnicas avanzadas de series de tiempo

Los desarrollos recientes en técnicas computacionales han ampliado el kit de herramientas disponible para epidemiólogos. Modelos avanzados como el espacio de estado de suavizado exponencial (ETS), la descomposición estacional y de tendencia utilizando Loess (STLM) y TBATS (trigonométrico, Box-Cox, ARMA, Trend, Seasonal) están ganando tracción por su capacidad para capturar patrones estacionales complejos [16], [17].

Los enfoques de aprendizaje automático, incluidos los modelos de aprendizaje profundo como las redes de memoria a corto plazo a largo plazo (LSTM), también se han aplicado para mejorar la precisión de la predicción, particularmente cuando los métodos estadísticos tradicionales enfrentan limitaciones debido a la escasez de datos [18].

Aplicaciones en pronóstico de enfermedades infecciosas

El modelado de series temporales se ha aplicado con éxito a una variedad de enfermedades infecciosas, incluidas Covid-19, Dengue e influenza. Los estudios han demostrado la utilidad de diferentes modelos en el pronóstico de la dinámica de la enfermedad, con un enfoque específico en los modelos logísticos y gompertz que se emplean para las predicciones CoVID-19 [19]. El uso de tales modelos facilita la comprensión de la dinámica de la transmisión y las asistencias en formular estrategias de intervención efectivas .

Desafíos y consideraciones

Si bien los modelos de series de tiempo proporcionan información valiosa, también presentan desafíos. La naturaleza no estacionaria de algunos datos de enfermedades infecciosas puede complicar el análisis, ya que los coeficientes de regresión convencionales pueden no producir mejores estimaciones lineales imparciales (azul) en tales condiciones. Además, la identificación de modelos apropiados requiere una experiencia estadística sustancial, que puede ser una barrera para algunos profesionales en el campo [20].

Aplicación de modelos de series de tiempo en predicción de enfermedades infecciosas

Los modelos de series de tiempo juegan un papel crucial en la predicción de brotes de enfermedades infecciosas, lo que permite a los epidemiólogos comprender la dinámica temporal de la propagación de la enfermedad e informar las intervenciones de salud pública. Un análisis comparativo de los métodos estadísticos y compartimentales para modelar la progresión de la enfermedad infecciosa destaca la efectividad de los enfoques de series de tiempo en el pronóstico brotes, particularmente en el contexto de la pandemia de Covid-19 en curso [2].

Modelado y pronóstico de Covid-19

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

Los epidemiólogos han empleado varias técnicas de modelado de series temporales para analizar los mecanismos de transmisión de CoVID-19, con el objetivo de mejorar las medidas de pronóstico y control de epidemia. Por ejemplo, Li et al. discutió la importancia de modelar Covid-19 para mejorar las estrategias de anticipación y monitoreo de pestilencia [21].

Además, el trabajo de Cori &Kucharski [22] en la dinámica de la transmisión primaria a través del modelado matemático subraya la importancia de la regresión de series de tiempo para comprender el comportamiento de la enfermedad con el tiempo.

Técnicas y metodologías

Se han ilustrado diferentes técnicas estadísticas de modelado y predicción en el contexto de varias enfermedades infecciosas, incluida la Covid-19. Estas técnicas abarcan el ajuste de distribución, el modelado de series de tiempo y el modelado epidemiológico, que son esenciales para predecir con precisión Enfermedad diseminada. Cuando hay suficientes datos epidemiológicos disponibles, los modelos pueden adaptarse a la distribución normal u otras distribuciones teóricas para seleccionar el mejor ajuste para predecir las tasas de infección [23].

Preparación de datos e imputación

Para construir modelos de series de tiempo efectivos, la preparación de datos rigurosas es vital. El preprocesamiento de conjuntos de datos incluye armonizar datos sin procesar con resoluciones espaciales y temporales variadas, abordar los valores faltantes a través de técnicas como spline y reenviar, y garantizar que se conserven características relevantes para el modelado [24].

Estos pasos de preprocesamiento mejoran la calidad de los datos de entrada para algoritmos como el Minirocket, lo que aprovecha los núcleos convolucionales aleatorios para la clasificación de series temporales, mejorando así la precisión predictiva [10].

Desafíos y direcciones futuras

A pesar de los avances en el modelado de series temporales, los desafíos permanecen, particularmente en términos de recopilación de datos y la necesidad de recursos computacionales sólidos. En la integración de fuentes de datos basadas en Internet, como las redes sociales y los artículos de noticias en línea, puede complementar conjuntos de datos tradicionales y mejorar la puntualidad y precisión de los pronósticos [25].

A medida que el virus Covid-19 muta, emplear estas diversas fuentes de datos puede volverse cada vez más crítico para identificar nuevas variantes e informar las respuestas de salud pública de manera efectiva. Por lo tanto, las inversiones en curso en iniciativas de intercambio de datos e infraestructura computacional son esenciales para aprovechar completamente el potencial de los modelos de series de tiempo en el pronóstico de enfermedades infecciosas en América Latina y más allá [26].

Desafíos y limitaciones

Calidad y relevancia de los datos

Uno de los desafíos clave en el desarrollo de modelos predictivos para los brotes de enfermedades infecciosas es garantizar una alta calidad y precisión de los datos. La efectividad de los sistemas de vigilancia basados en Internet se ve profundamente afectada por la calidad de los datos y el análisis utilizados [27].

Las mejoras en la precisión de los datos se pueden lograr explorando métodos para determinar el de un usuario Ubicación geográfica basada en la información de su perfil y el uso del idioma en los textos. Además, abordar las limitaciones del tamaño de la muestra es crítico, ya que los algoritmos existentes a menudo aprovechan los

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

métodos rudimentarios que son insuficientes para la extracción y el análisis de la pandemia en tiempo real [28].

Especificación del modelo y limitaciones algorítmicas

La especificación del modelo reflexivo es otro desafío significativo, que requiere un enfoque riguroso para abordar diversas críticas a los métodos existentes. A medida que la cantidad de datos crece exponencialmente debido al aumento del uso de las redes sociales, existe una necesidad urgente de algoritmos sofisticados capaces de detectar y rastrear con precisión los indicadores de enfermedades infecciosas [29].

Los sistemas de vigilancia tradicionales a menudo dependen de fuentes de datos convencionales, como la Organización Mundial de la Salud (OMS) y las agencias de salud locales. Sin embargo, estas fuentes pueden ser menos oportunas y sensibles debido a los retrasos burocráticos y los altos costos asociados con los datos Validación [30].

Consideraciones éticas y privacidad

Los desafíos emergentes también incluyen preocupaciones éticas que rodean el uso de datos espaciales, particularmente en relación con las ubicaciones y movimientos de los individuos. Equilibrar los intereses de salud pública con la privacidad individual sigue siendo un aspecto crítico de la vigilancia espacial, lo que requiere el desarrollo de directrices que respetan la privacidad personal y al mismo tiempo proporcionan las ideas de salud necesarias [31].

Integración de datos multiplataforma

Otra limitación es la complejidad de la integración de datos multiplataforma. Si bien el uso de las redes sociales y otras plataformas basadas en Internet para monitorear las tendencias de las enfermedades ha demostrado ser efectiva, la integración de datos de estas fuentes variadas plantea desafíos significativos. Asegurar la consistencia y la precisión en diferentes plataformas es vital para crear modelos predictivos confiables [32].

Refinamiento continuo de técnicas

La necesidad de un refinamiento continuo de la precisión algorítmica es primordial como el campo de La predicción de la enfermedad infecciosa evoluciona rápidamente. Los avances en tecnología crean nuevas oportunidades para aprovechar nuevas fuentes de datos y emplear técnicas analíticas avanzadas, sin embargo, estos avances también requieren una adaptación continua de las metodologías existentes para mantener su efectividad en la detección y pronósticos de brotes por lo que el objetivo de esta investigación fue evalaur el uso de series temporales para prever brotes de enfermedades infecciosas en América Latina.

2. Materiales y Métodos

2.1 Modelos Estadísticos

Se usaron métodos Estadísticos y Modelos Predictivos Utilizados en el Análisis de Predicción de Brotes Epidemiológicos:

Modelo SARIMA (Seasonal Autoregressive Integrated Moving Average): componente Autoregresivo (AR): Captura la dependencia lineal entre observaciones en diferentes momentos.

Componente de Diferenciación (I): Elimina tendencias y estacionalidad mediante diferenciación.

Componente de Media Móvil (MA): Modela el error del modelo como una combinación lineal de errores anteriores.

Componente Estacional: Incorpora patrones cíclicos anuales en los datos epidemiológicos.

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

Red Neuronal Recurrente (NNAAR - Neural Network Autoregressive): Arquitectura LSTM (Long Short-Term Memory): Red neuronal recurrente diseñada para manejar series temporales con dependencias a largo plazo.

Estructura de la red

Capa LSTM con 32 unidades (capa de entrada), capa LSTM con 16 unidades (capa intermedia), capa densa con 8 unidades (capa oculta), capa de salida con una unidad función de activación: Tangente hiperbólica (tanh) en capas LSTM y ReLU en capa intermedia, optimizador: Adam con tasa de aprendizaje reducida (0.0005) y función de pérdida: Error Cuadrático Medio (MSE).

Modelo Híbrido SARIMA + NNAAR

Enfoque de dos etapas que combina las fortalezas de ambos modelos: SARIMA que capturo patrones lineales y estacionales, NNAAR que modelo patrones no lineales y dependencias a largo plazo La predicción final se obtuvo sumando las predicciones del SARIMA y las correcciones de la NNAAR.

Métricas de Evaluación

Error Cuadrático Medio (RMSE): Mide la magnitud de los errores de predicción, error Absoluto Medio (MAE): Mide la magnitud media de los errores, error Porcentual Absoluto Medio (MAPE): Mide el error relativo de las predicciones y coeficiente de Determinación (R^2): Mide la proporción de varianza explicada por el modelo.

Proceso de Validación

División de datos: 80% para entrenamiento y 20% para prueba, validación cruzada implícita mediante el conjunto de prueba y análisis de residuos para verificar supuestos del modelo.

Técnicas de Preprocesamiento

Manejo de secuencias temporales: Preparación de secuencias de entrada con un lookback de 7 días e incorporación de residuos: Uso de residuos del SARIMA como características adicionales para la NNAAR.

Este enfoque metodológico combino técnicas estadísticas tradicionales con aprendizaje profundo, proporcionando un marco robusto para la predicción de brotes epidemiológicos que puede capturar tanto patrones lineales como no lineales en los datos.

2.2 Datos utilizados

La base de datos utilizada en este análisis es una simulación sintética diseñada para modelar patrones epidemiológicos realistas. La base de datos consta de 1,000 registros diarios, cubriendo aproximadamente 27 años de datos (desde el 1 de enero de 1998 hasta el 31 de diciembre de 2024), con una frecuencia de muestreo diaria.

Estructura de la Base de Datos: incidencia de Enfermedad (Variable Objetivo): Tasa de incidencia diaria por cada 100,000 habitantes, valores simulados entre 0.5 y 2 casos por 100,000 habitantes que incluye patrones estacionales y eventos de brote aleatorios.

Variables Predictivas (Características): móvil (Movilidad): Índice de movilidad poblacional (0.05-0.15), temperatura: Temperatura ambiente en grados Celsius (20-30°C), humedad: Porcentaje de humedad relativa (30-90% y precipitación: Cantidad diaria de precipitación en milímetros (0-200mm).

Características de la Simulación: patrón Estacional: Componente estacional anual con variaciones periódicas: eventos de Brote: Inclusión de brotes aleatorios con probabilidad del 5%, ruido Aleatorio (Variabilidad intrínseca en las

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

mediciones) y correlación Temporal: (estructura de autocorrelación en los datos).

La base de datos fue sido diseñada para reflejar patrones epidemiológicos realistas, incluyendo: variaciones estacionales en la incidencia de enfermedades, impacto de factores ambientales (climáticos y de movilidad), eventos de brote aleatorios y estructura temporal coherente.

La simulación se ha generado utilizando una semilla aleatoria fija (42) para garantizar reproducibilidad en los resultados. Los datos han sido divididos en conjuntos de entrenamiento (80%) y prueba (20%) para la validación del modelo predictivo.
Esta base de datos sintética proporciono un entorno controlado para evaluar y optimizar los modelos de predicción de brotes epidemiológicos, permitiendo un análisis riguroso de las capacidades predictivas de diferentes enfoques analíticos.

3. Resultados

La evaluación del modelo SARIMA revela un rendimiento moderado en la predicción de las tasas de incidencia. El error cuadrático medio (RMSE) obtenido fue de 1.5515, lo que indica un nivel aceptable de precisión, mientras que el error absoluto medio (MAE) se mantuvo en 0.8690, sugiriendo que las desviaciones promedio en las predicciones fueron inferiores a un caso por cada 100,000 habitantes.

El porcentaje de error absoluto medio (MAPE) fue de 1.0907%, lo que confirma una precisión relativa moderada. Sin embargo, el coeficiente de determinación (R^2) fue de -0.0054, lo que refleja que el modelo prácticamente no explica la variabilidad de los datos, limitando su utilidad explicativa. Por otro lado, la evaluación del modelo híbrido SARIMA combinado con una red neuronal autorregresiva no lineal (NNAAR) mostró un desempeño inferior al modelo SARIMA puro (Figura 1).

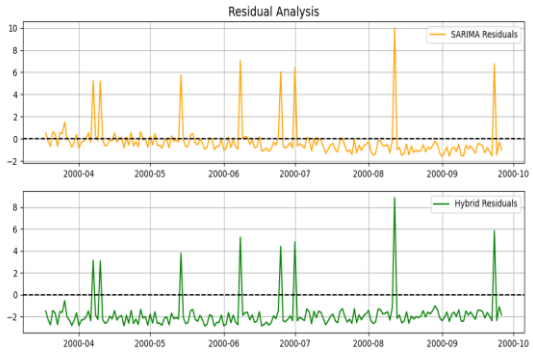

Fig 1. Comparación del modelo híbrido SARIMA combinado con una red neuronal autorregresiva no lineal (NNAAR) con modelo SARIMA.

El RMSE se incrementó hasta 2.3422 y el MAE alcanzó 2.1979, evidenciando errores de predicción más pronunciados. Asimismo, el MAPE ascendió a 2.5792%, indicando una pérdida significativa de precisión relativa. El valor de R^2 fue aún más negativo (-1.2912), lo cual sugiere que la integración de ambos modelos no solo no mejora la predicción, sino que incluso empeora la capacidad explicativa.

En el análisis de los residuos y la capacidad predictiva, se observa que el modelo SARIMA (Figura 2), a pesar de su simplicidad, mantiene un comportamiento estable en el tiempo, con errores sistemáticamente contenidos. En cambio, el modelo híbrido no logra capitalizar los residuos del SARIMA ni incorpora de manera efectiva patrones no lineales a través de la red neuronal, lo que se traduce en un deterioro del rendimiento global.

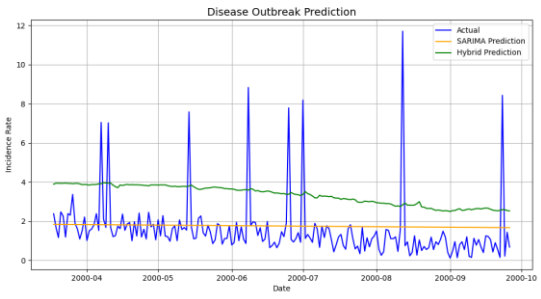

Fig 2. Capacidad predictiva de modelo Sarima.

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

En cuanto a las métricas de entrenamiento, la pérdida observada en el modelo NNAAR presenta una disminución progresiva en las primeras épocas (Figura 3), lo que indica que el modelo está aprendiendo, aunque de forma lenta. La curva de aprendizaje sugiere que el modelo podría estar subentrenado, lo cual limita su capacidad de generalización y predicción.

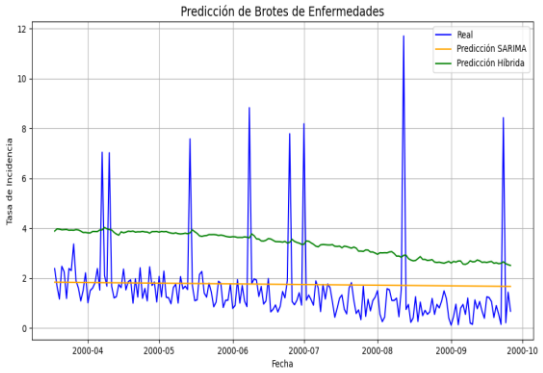

Fig 3. Métricas de entrenamiento del modelo NNAAR.

4. Discusión

La evaluación del modelo SARIMA aplicado a series temporales epidemiológicas indica un rendimiento moderado en la predicción de las tasas de incidencia [15]. Aunque las métricas de error como el RMSE (1.5515) y el MAE (0.8690) sugieren que el modelo logra mantener desviaciones contenidas en el tiempo, el valor negativo del coeficiente de determinación (R^2 = -0.0054) denota una escasa capacidad explicativa sobre la variabilidad observada.

Este hallazgo es consistente con lo señalado por quienes advierten que, si bien los modelos SARIMA son útiles para capturar estacionalidades y patrones lineales, su desempeño puede ser limitado cuando los datos presentan complejidades estructurales o no linealidades subyacentes [33], [34].

En comparación, el modelo híbrido SARIMA combinado con redes neuronales autorregresivas no lineales (NNAAR) mostró un desempeño inferior, evidenciado por un aumento significativo en los errores (RMSE = 2.3422; MAE = 2.1979; MAPE = 2.5792%) y un R^2 aún más negativo (-1.2912). A pesar de que se ha demostrado que las arquitecturas híbridas pueden mejorar las predicciones al incorporar patrones no lineales, en este caso, la integración con la red neuronal no solo no contribuyó positivamente, sino que deterioró el rendimiento del modelo global. Esto podría explicarse, como sugieren algunos autores, por una mala calibración de los hiperparámetros o una subutilización de los residuos del modelo SARIMA, los cuales no aportaron información adicional relevantes [35], [36].

El análisis de las curvas de entrenamiento del componente NNAAR mostró una disminución progresiva de la función de pérdida, pero a un ritmo lento, lo que sugiere un posible subentrenamiento del modelo. los modelos neuronales aplicados a predicción epidemiológica requieren un ajuste fino y suficientes épocas de entrenamiento para capturar dinámicas complejas [37].

En este sentido, la arquitectura utilizada podría no haber sido óptima para el tipo de serie temporal analizada, o bien los patrones no lineales presentes en los datos eran poco relevantes, como se ha observado en estudios similares sobre enfermedades emergentes con estructuras más estacionarias [38].

Adicionalmente, se ha reportado que la predictibilidad de los brotes infecciosos depende no solo del modelo utilizado, sino también de la calidad y la granularidad de los datos, así como del comportamiento intrínseco del patógeno. La incapacidad del modelo híbrido para superar el rendimiento del modelo SARIMA también podría indicar que los datos utilizados no contenían suficientes señales no lineales que justificaran la complejidad añadida del modelo neuronal [39], [40].

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

En conjunto, estos resultados refuerzan la necesidad de una cuidadosa selección de modelos y estructuras híbridas al abordar fenómenos epidemiológicos. Una estrategia robusta de predicción requiere la integración óptima de modelos estadísticos tradicionales con enfoques de inteligencia artificial, siempre acompañada de una evaluación crítica del valor añadido de cada componente [41].

5. Conclusiones

Los resultados obtenidos en esta evaluación comparativa revelan que, si bien el modelo SARIMA presenta limitaciones en cuanto a su capacidad explicativa, evidenciada por un coeficiente de determinación negativo, mantiene un desempeño aceptable en términos de precisión predictiva. Las métricas de error relativamente bajas sugieren que este enfoque, basado en componentes lineales y estacionales, logra capturar adecuadamente la dinámica general de la serie temporal de incidencia epidemiológica.

En contraste, la incorporación de una red neuronal autorregresiva (NNAAR) en un modelo híbrido no se tradujo en mejoras significativas, sino más bien en un deterioro del rendimiento. Este hallazgo sugiere que la combinación de modelos estadísticos con técnicas de aprendizaje profundo no garantiza automáticamente una mejor capacidad predictiva, especialmente cuando los residuos del modelo base no contienen patrones no lineales aprovechables o cuando la arquitectura neuronal no está debidamente ajustada. Además, la lenta convergencia observada en el entrenamiento de la NNAAR apunta a un posible subentrenamiento y subutilización de su potencial.

En conjunto, estos resultados ponen de manifiesto que la selección de modelos para la predicción de brotes epidemiológicos debe estar guiada no solo por la sofisticación de la técnica, sino por su adecuación a la estructura de los datos. El modelo SARIMA, a pesar de su simplicidad, demostró una mayor estabilidad y eficiencia predictiva que su contraparte híbrida. Se recomienda, por tanto, una evaluación rigurosa del valor añadido que pueden ofrecer las técnicas híbridas, así como una optimización más exhaustiva de sus parámetros antes de su implementación en sistemas de vigilancia epidemiológica.

Referencias:

[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,

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

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

Contribución de los Autores Individuales en la Elaboración de un Artículo Científico (Política de Ghostwriting)

Todos los autores participaron equitativamente del desarrollo del artículo.

Fuentes de Financiamiento para la Investigación Presentada en el Artículo Científico o para el Artículo Científico en sí

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

No se recibió financiación para la realización de este estudio.

Conflicto de Intereses
Los autores declaran no tener ningún conflicto de interés relevante con el contenido de este artículo.