Modelos híbridos SARIMA-ANN para pronósticos del contagio por SARS-CoV-2 en el Perú

SARIMA-ANN hybrid models for forecasts of SARS-CoV-2 contagion in Peru

Alipio Ordoñez Mercado[1*]

[1]*Facultad de Ingeniería Económica, Estadística y Ciencias Sociales, Universidad Nacional de Ingeniería, Lima, Perú*

**Email: alorme@uni.edu.pe*

***Recibido (Received): 15/07/2021 Aceptado (Accepted): 01/12/2021 Publicado (Published): 20/12/2021**

RESUMEN

Se construyó modelos híbridos ANN-ARIMA por remodelamiento para realizar los pronósticos de los nuevos casos de contagios por SARS-Co-2 en el Perú. Para ello se extrajo y uso los datos de los casos confirmados de contagiados por SARS-CoV-2 entre el periodo 06/03/20 hasta el 28/ 02/21, desde la plataforma de los datos abiertos del Ministerio de Salud. Los resultados hallados indican que, según los errores de pronósticos medios porcentuales (MAPE), los 02 mejores modelos corresponden al modelo hibrido multiplicativo NNAR (27,1,6)* ARIMA(3,0,2)(1,0,1), y al modelo hibrido aditivo NNAR (27,1,6) + ARIMA(1,0,1), cuyos valores se diferencian en tan solo el 0.575%. Cuando se usa el valor promedio del MAPE para los 03 mejores modelos de cada categoría de modelamiento aplicado, se concluye que los modelos híbridos NNAR-ARIMA tienen los menores errores de pronósticos que los modelos híbridos MLP-ARIMA, los modelos híbridos aditivos NNAR+ARIMA tienen una superioridad del 1.20% sobre los modelos híbridos multiplicativos NNAR*ARIMA; mientras que la superioridad del modelo hibrido aditivo MLP+ARIMA alcanza al 2.31% sobre el modelo hibrido multiplicativo MLP*ARIMA.

Palabras Clave: *modelos ARIMA, redes neuronales autoregresivas, perceptrón multicapas, modelos híbridos NNAR-ARIMA, modelos híbridos MLP-ARIMA*

ABSTRACT

Hybrid ANN-ARIMA models were built by remodeling to forecast new cases of SARS-Co-2 infections in Peru. For this, the data of the confirmed cases of infected by SARS-CoV-2 between the period 03/06/20 until 02/28/21 was extracted and used, from the open data platform of the Ministry of Health. The results found indicate that, according to the mean percentage forecast errors (MAPE), the 02 best models correspond to the multiplicative hybrid model NNAR (27,1,6) * ARIMA (3,0,2) (1,0,1), and the additive hybrid model NNAR (27,1,6) + ARIMA (1,0,1), whose values differ by only 0.575%. When the average MAPE value is used for the 03 best models of each applied modeling category, it is concluded that the NNAR-ARIMA hybrid models have the lowest forecast errors than the MLP-ARIMA hybrid models, the NNAR + ARIMA additive hybrid models they have a superiority of 1.20% over the multiplicative hybrid models NNAR * ARIMA; while the superiority of the MLP + ARIMA additive hybrid model reaches 2.31% over the MLP * ARIMA multiplicative hybrid model.

Keywords*: ARIMA models, autoregressive neural networks, multilayer perceptron, hybrid models NNAR-ARIMA, hybrid models MLP-ARIMA*

1. INTRODUCCIÓN

Las técnicas estadísticas para pronósticos, tienen una gran importancia en el proceso de toma de decisiones, de los diversos procesos, que son necesarios en el desarrollo de las grandes ciudades y los países. Desde los tiempos remotos, aparecen registrados en la literatura estadística estudios que revolucionaron el desarrollo de las áreas correspondientes. Nielsen (2020), presenta la siguiente breve descripción:

Los primeros estudios sobre manchas solares encontradas hace 800 años a.C, en la antigua China. En 1662 se publicó en Londres las primeras tablas de vida por intervalos de tiempo. En 1850 en Inglaterra, se publicó los pronósticos del tiempo en *The Time of London,* un periódico de gran notoriedad. Y En 1877, se registró las impresiones de la trayectoria de un electrocardiograma (pp. 2-10)

Hooker (1905) introdujo el concepto de correlación entre observaciones sucesivas, y años más tarde se aplicó en el estudio de procesos económicos y sociales en el tiempo (Yule, 1909), y en el estudio de la serie de tiempo manchas solares, mediante el modelo auto regresivo (Yule, 1927).

Box & Jenkins (1970) integran el concepto de auto correlación en la teoría de las series de tiempo, y publicaron el texto semillero de la teoría y práctica de los modelos ARIMA. Les siguen muchos otros autores que describieron unos la teoría y otros la aplicación de estos modelos a datos reales, entre ellos se mencionan a: Brokcwell & Davis (1990), Hamilton (1994), Wei (2006), Pankratz (1983), Hydman & Athanasopoluolos (2013), y Uriel (1985).

Al iniciarse en noviembre del 2019 los primeros casos de contagio por la infección del SARS-Cov-2 en China, los modelos estadísticos del tipo ARIMA, y los métodos de suavización de medias móviles y exponenciales fueron aplicados durante el primer semestre del año 2020, determinándose las tendencias del número de infectados por SARS-CoV-2. Estudios bajo este enfoque se realizaron en muchos países (Dehesh, Fard & Desheh, 2020; Gupta & Pal, 2020; Peroné, 2020 a; Tandon, et. Al., 2020; Benvenuto, et.al., 2020; Ceylan, 2020; Ding, Li, Jiao & Shen, 2020; y Ganiny & Nisar, 2021). En la segunda mitad del año 2020 y a inicios del 2021 se aplicaron modelos híbridos del tipo ARIMA-ANN (Peroné, 2020b; y Safi & Sanusi, 2021).

En este escenario, se elaboró este proyecto con fines de extender las aplicaciones de los modelos híbridos para el pronóstico del número de confirmados de infección por SARS-CoV-2 en el Perú, mediante el uso de la combinación de la metodología ARIMA y las redes neuronales artificiales- ANN. La hipótesis de investigación es que los modelos híbridos SARIMA-RNA son más efectivos en la obtención de los pronósticos diarios de los nuevos casos de infección por SARS-CoV-2 en el Perú, que el modelo SARIMA y el modelo ANN. La estimación se realizó mediante dos particiones: la clase de modelos con estructura aditiva,

y la de estructura multiplicativa, para luego ser comparados según sus errores absolutos medios porcentuales (MAPE).

2. MATERIALES Y MÉTODOS

Esta investigación es observacional, y descriptiva. Se observa la variable de interés conforme es registrada en la plataforma de los datos abiertos del Ministerio de Salud (MINSA), y se determina los rezagos más importantes que permitan construir la ecuación del modelo estadístico para obtener los pronósticos del número de infectados diarios por SARS-CoV-2 en el Perú. La población bajo estudio, abarca a la totalidad de habitantes al instante de capturar la información, al respecto no hubo una metodología regular sobre la toma de las pruebas de diagnóstico. Al inicio se efectuaron bajo criterios epidemiológicos, y luego según los requerimientos de descarte de los posibles infectados, la distancia y la accesibilidad de los lugares desde donde se originó la solicitud, afecto a los resultados de diagnóstico con respecto al día de ocurrencia. En el Perú, para determinar el número de casos diarios de infección confirmados, se utilizó de aproximadamente 56000 pruebas diarias, distribuidas en las 03 pruebas: serológicas, antígenas y moleculares, cuyo total alcanzaron al 16/06/21 a 13620937 pruebas de diagnóstico. Una limitación a superar sigue siendo el tiempo de diagnóstico, y depende de cuan alejados desde las ciudades importantes se encuentran los pobladores con calificación de pobreza.

2.1 MATERIALES E INSTRUMENTOS

Para resolver las interrogantes formuladas en el problema de investigación se han usado los siguientes materiales:

a. Registró diario de las personas diagnosticadas positivas con la infección por SARS-CoV-2 desde el 06/03/2020 al 28/02/2021, tomados desde la plataforma de datos abiertos del MINSA[1], que al 28/02/21 abarcan un total de 52 semanas para el estudio. Se ha considerado las primeras 49 semanas para el tramo de entrenamiento o estimación de los parámetros de los posibles mejores modelos y las últimas 03 semanas (50,51,52), para efectos de validación y selección de los mejores modelos.
b. Hojas electrónicas en el programa informático Microsoft Excel, para filtrar los casos diarios positivos de infección en el Perú, y cualquier ciudad de interés.
c. Teoría de la metodología ARIMA, y sus respectivos scripts en el lenguaje del programa estadístico R.
d. Teoría sobre el modelamiento de series de tiempo con modelos de redes neuronales artificiales: Perceptrón multicapas (MLP), redes neuronales auto regresivas (NNAR), y sus respectivos programas fuente.

[1] (https://www.datosabiertos.gob.pe/dataset/casos-positivos-por-covid-19-ministerio-de-salud-minsa)

e. Combinación de los modelos SARIMA-MLP y SARIMA-NNAR en las estructuras de combinación aditiva y multiplicativa con sus respectivos programas fuente lenguaje “R”.

2.2 MÉTODOS

Para obtener los modelos híbridos, se combinan según las estructuras aditivas o multiplicativas, usando la arquitectura de una red perceptrón multicapas (MLP) y la red neuronal auto regresiva (NNAR) con la metodología ARIMA, según se describe a continuación.

2.2.1 MODELOS ARIMA

La teoría de los modelos ARIMA fue presentada y publicada en el texto seminal de Box & Jenkins (1970), quienes estipulan la representación multiplicativa del modelo en la siguiente ecuación matemática:

$$\Delta^{d}\Delta_{s}^{D}\phi(B)\Phi_{s}(B^{s})Z_{t}^{(\lambda)} = \theta_{0} + \theta(B)\Theta_{s}(B^{s})\square_{t} \qquad (1)$$

En esta estructura general se considera todos los elementos que definen a los modelos particulares ARMA, ARIMA y SARIMA, permitiendo realizar transformaciones para estabilizar a la varianza de la serie de tiempo (parámetro λ), y el mecanismo de las diferencias regulares y estacionales que hacen posible tratar con una serie de tiempo estacionaria en media, pendiente y estación (“d”, “D”). En la literatura estadística existen estudios que han extendido a otras estructuras alternativas inmediatas del modelo de una serie de tiempo, como la incorporación de variables regresoras para explicar a la variación de la serie de tiempo, dando origen a los modelos ARMAX, ARIMAX y SARIMAX.

2.2.2 MODELOS DE REDES NEURONALES PARA SERIES DE TIEMPO-ANN

De otro lado, los modelos de las redes neuronales artificiales, se usan para el modelamiento de las componentes no lineales. Se propone la expresión matemática no lineal para procesar el valor de la salida de cada neurona mediante la ecuación;

$$Y_{i} = \left(W_{0,j} + \sum_{i=1}^{n} W_{ji}X_{i} \right), \qquad (2)$$

Los $w_{ji,}$ representan a los pesos sinápticos que ponderan las “n” neuronas de entradas “X_i” y al nivel del umbral o sesgo $w_{0,j}$. La función “φ” es la función de activación que transforma la información que envía a la neurona siguiente.

Para su adaptación al caso del modelamiento de las series de tiempo en el instante de tiempo "t", la salida de la red neuronal es calculada, mediante las redes neuronales perceptrones multicapas y las redes neuronales autoregresivas, por la ecuación:

$$Z_t = W_0 + \sum_{j=1}^{q} W_j g\left(W_{0,j} + \sum_{i=1}^{p} W_{ij} Z_{t-j} \right) + \varepsilon_t (3) \qquad (3)$$

Donde:

W_{ij}, $W_{0,j}$, i = 0,1,2, …,p; j = 1,2,…,q; W_j, j = 0,1,2,…, q son los parámetros de la red neuronal y representan las ponderaciones de las conexiones sinápticas, y son determinadas de forma que una función objetivo sea optimizada. Esta función por lo general es el error cuadrático medio:

"p": es el número de neuronas (nodos) de la capa de entrada,
"q": es el número de neuronas (nodos) de la capa oculta, y
"g": es la función de activación, que por lo general pueden ser cualquiera de las siguientes: tangente hiperbólica, sigmoidea, gaussiana, lineal, escalón o la identidad.

La estructura matemática de la ecuación 03, se puede modificar para permitir otras capas ocultas adicionales que permita un aprendizaje profundo de la red neuronal. No obstante, una capa oculta permite obtener una buena aproximación de cualquier función no lineal (Cybenko, 1989; Hornik, Stinchcombe, & White, 1989; 1990).

2.2.3 MODELOS HÍBRIDOS PARA SERIES DE TIEMPO

Usando una estructura aditiva, Zhang (1998, 2003), y Zhang & Qi (2005), combinaron las ecuaciones (1) y (3) para obtener un modelo aditivo hibrido al que denominaron modelo hibrido aditivo ARIMA+ANN, y cuyos pronósticos resultaron mejor que los modelos combinados por separado. Luego, Wang et.al (2013), emplearon la estructura multiplicativa para combinar la metodología ARIMA con los modelos neuronales artificiales (ANN), al cual denominaron modelo hibrido multiplicativo ARIMA*ANN. De esta forma para realizar un estudio sobre los modelos híbridos ARIMA-ANN o viceversa, se usa las ecuaciones siguientes:

Donde:

L_t: Componente de variación lineal y es estimada por el modelo ARIMA.

N_t: Componente de variación no lineal y es estimada por el modelo de ANN.

En esta investigación, la ecuación (4) se ha usado mediante el proceso de remodelamiento para construir los modelos de series de tiempo, que permitan obtener los mejores pronósticos para el número diario de infectados por SARS-CoV-2 en el Perú, dicha actividad se viene incrementando en muchos países con el fin de mejorar los pronósticos para enfrentar el impacto del SARS-CoV-2.

La construcción de los mejores modelos híbridos se determinó según el error absoluto medio porcentual (MAPE) mínimo en el tramo de validación o testeo, es decir, se siguió los siguientes pasos:

a. En la etapa del pre-procesamiento se realizó a estimación de las unidades perdidas, mediante un promedio móvil de la unidad referida a un día de la semana, para las últimas 03 semanas
b. Determinar según los valores del MAPE, los 03 mejores modelos para pronósticos, en cada categoría de modelamiento: metodología SARIMA, redes neuronales artificial perceptrón multicapas (MLP), y redes neuronales autoregresiva (NNAR)
c. Realizar la combinación aditiva para cada modelo SARIMA seleccionado (L_t), con el modelo extraído del residual mediante el modelo MLP(N_t), y en viceversa
d. Realizar la combinación multiplicativa para cada modelo SARIMA seleccionado (L_t), con el modelo extraído del residual mediante el modelo MLP(N_t), y en viceversa
e. Realizar la combinación aditiva para cada modelo SARIMA seleccionado (L_t), con el modelo extraído del residual mediante el modelo NNAR(N_t), y en viceversa
f. Realizar la combinación multiplicativa para cada modelo SARIMA seleccionado (L_t), con el modelo extraído del residual mediante el modelo NNAR(N_t), y en viceversa
g. Realizar un resumen con los valores del MAPE a fin de verificar las hipótesis de investigación y el mejor modelo para pronósticos del número diario de infectados por el SARS-CoV-2 en el Perú.

3. RESULTADOS

El modelamiento y re-modelamiento de los diversos modelos para pronósticos por separado, y los híbridos se ejecutó en conformidad con la teoría general, y la herramienta fundamental del correlograma, establecidas y publicadas en Box & Jenkins (1970).

3.1. MODELAMIENTO ARIMA Y ANN

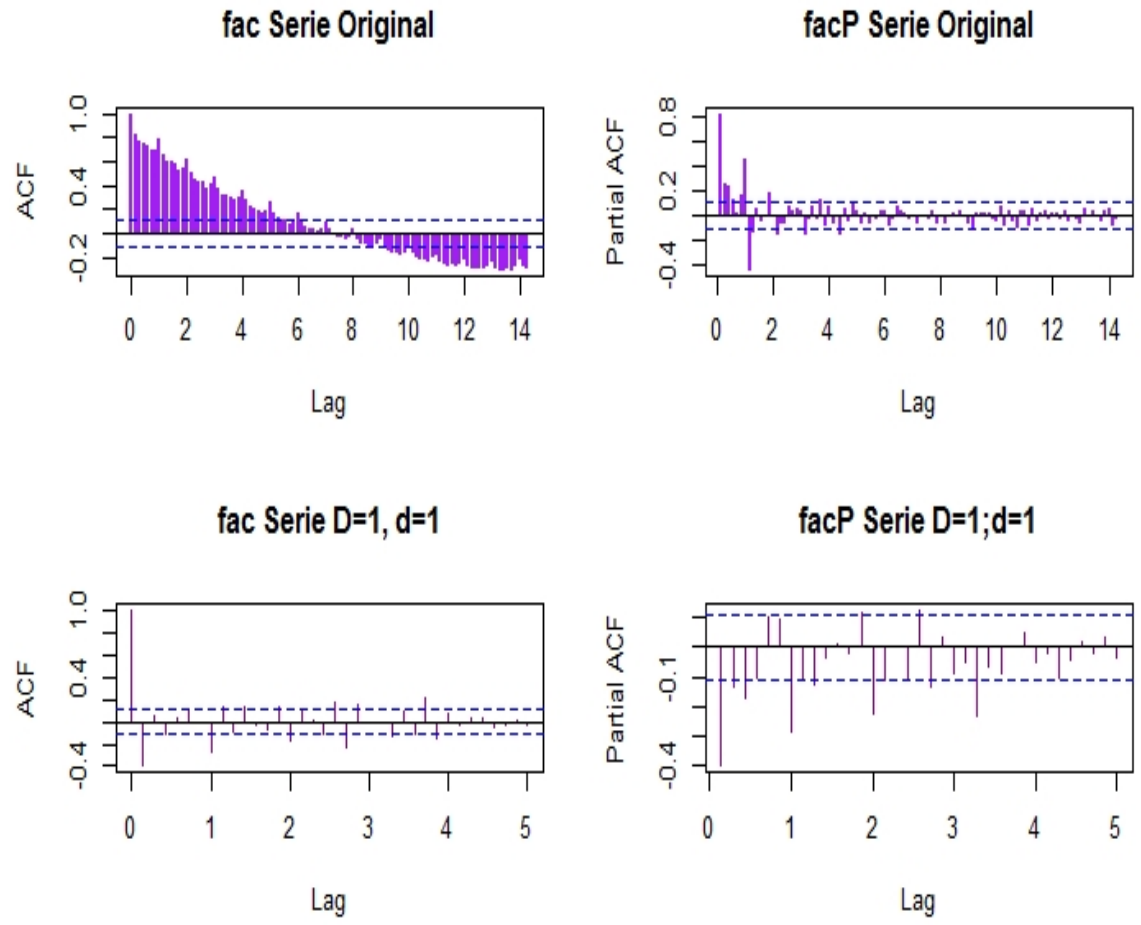
En esta sección, según los valores del MAPE, se determinó a los 03 mejores modelos para pronósticos según la metodología SARIMA, MLP, y NNAR.

Modelamiento SARIMA

Las herramientas principales para iniciar el modelamiento SARIMA se presenta en la Figura 1. Se visualizan las funciones de autocorrelación de la serie original, las cuales tienen una intensa variación de estación y tendencia. Por ello es necesario aplicar una diferencia estacional, y luego una diferencia regular a fin de arribar a una estacionariedad tanto en estación y en media.

A partir del correlograma de la serie de tiempo estacionarizada se identifican al menos 05 modelos siguientes: SARIMA (3;1;0) $(2;1;0)_7$, SARIMA (0;1;1) $(0;1;2)_7$, SARIMA (3;1;0) $(0;1;2)_7$, SARIMA (0;1;1) $(2;1;0)_7$, SARIMA (3;1;1) $(2;1;2)_7$, y SARIMA (1,0,1) $(0,1,2)_7$, cuya evaluación en el tramo de testeo o validación según el valor del MAPE se presentan en la TABLA 1. En la cual se identifican a los siguientes 03 mejores modelos según el MAPE: SARIMA (1,0,1) $(0,1,2)_7$, SARIMA (3,1,0) $(0,1,2)_7$ y el SARIMA (3,1,0) $(2,1,0)_7$, en orden de importancia. Estos modelos fueron usados para realizar las combinaciones con los mejores modelos de redes neuronales respectivas.

Figura 1. Correlograma de la serie original y diferenciada regular y estacional



Modelamiento Perceptrón Multicapa (MLP)

La red neuronal MLP es representada por el acrónimo, MLP (k, h, o), y para el caso de aplicar a una serie de tiempo, se considera el modelo constituido por 03 capas, una capa de entrada con "k" neuronas, una capa oculta con "h" neuronas y una capa de salida con "o=1" neurona. Para determinar en número de neuronas en la capa de entrada, se usó la teoría general de las series de tiempo, que propone a los primeros pocos rezagos, por ejemplo k<=15 o 20. El número de neuronas de la capa oculta, fue determinado de forma a minimizar el error de pronóstico, MAPE entre las propuestas de Shibata & Ikeda (2009), $n_h = (ne\ n_s)^{1/2}$ y de Trenn (2008), $n_h = (n_e + n_s - 1)/2$, para este estudio se tomó k=15, así n_h fluctúa entre 04 a 10 neuronas. De esta manera, usando el lenguaje estadístico R y su función *Accuracy*, se determinó los 03 mejores modelos con valores del MAPE menores en el tramo de validación, y que corresponden a modelos: MLP(9,9,1), MLP(9,6,1), y el MLP(9,4,1), cuyos valores del MAPE son: 11.5271, 11. 9712, y 12.3282 respectivamente.

TABLA 1. Tres mejores modelos por categoría de modelamiento.

MODELOS, l = 1	RMSE	MAE	MAPE	MAPE MEDIO
SARIMA(1,0,1)(0,1,2)$_7$	1072.896	838.6980	19.77682	
SARIMA(3,1,0)(0,1,2)$_7$	2104.103	1745.6888	38.30800	
SARIMA(3,1,0)(2,1,0)$_7$	2311.3799	2034.1059	40.50732	32.86405
MLP(9,9,1)$_7$	825.3103	663.5425	11.52713	
MLP(9,6,1)$_7$	747.4627	606.1292	11.97118	
MLP(9,4,1)$_7$	758.9584	578.9416	12.32816	11.94216
NNAR(27,1,6)$_7$	772.0379	625.7362	10.18269	
NNAR(27,1,8)$_7$	876.3024	679.25235	10.56613	
NNAR(27,2,5)$_7$	762.2316	633.2234	10.87282	10.54055

Modelamiento según las Redes neuronales Autoregresivas-NNAR

Las redes neuronales autoregresivas fueron creadas para que se adapten al impacto de los pocos primeros rezagos sobre el valor de la serie temporal en el tiempo "t", permitiendo además del impacto de los rezagos estacionales, su representación usa el acrónimo NNAR(p,P,S)$_{[m]}$ donde "p" representa a las entradas no estacionales y sirven para capturar las variaciones de la tendencia mediante los rezagos (Z_{t-1}, Z_{t-2}, …., Z_{t-p}). "P" es el orden estacional y captura las variaciones estacionales mediante los rezagos (Z_{t-P}, Z_{t-2P}, Z_{t-mP}), en donde h representa al número de neuronas en la capa oculta, y m a la longitud del periodo estacional. Adicionalmente es posible realizar una transformación "λ" dentro de la función NNETAR del programa estadístico "R". De esta forma, esta red neuronal tiene como mínimo

03 capas, una de entrada con "p+P" neuronas de entrada, una capa oculta con (p*P+1)/2 neuronas, y una capa de salida con una neurona. Cabe la posibilidad de tener varias otras capas ocultas, y que, en la capa de entrada, se acepten variables regresoras que ayuden a explicar al valor de la serie en el tiempo "t". Para nuestro caso se han usado p=27, P=1, y el número de neuronas en la capa oculta optimizada por la propia red neuronal. Según el criterio del MAPE los 03 mejores modelos de redes neuronales auto regresivas son: NNAR(27,1,6)$_7$ con un MAPE=10.18269, NNAR(27,1,8)$_7$ con un MAPE=10.56613 y NNAR(27,2,5)$_7$ con un MAPE=10.87282, los mismos que se describen en la TABLA 1.

3.2 MODELAMIENTO HÍBRIDO SARIMA-ANN

Los modelos híbridos con menores valores del MAPE, se determinan combinando los 03 modelos con menor MAPE, según sus categorías de modelamiento descritos en la tabla 1, de acuerdo a las estructuras de combinación aditiva y multiplicativa, y según la componente a extraer, lineal (L_t) con el modelo ARIMA, y no lineal (N_t) con el modelo de redes neuronales ANN. Este proceso se describe en las próximas secciones.

Modelamiento hibrido SARIMA-ANN

Bajo esta categoría de modelamiento se combinan los 03 mejores modelos SARIMA, primero con la red neuronal MLP (k, h, o), y luego con la red neuronal auto regresiva NNAR (p, P, h), con estructuras aditivas (+) y multiplicativas (*). De esta forma, se extraen en primer lugar la componente lineal mediante la aplicación de las herramientas del modelamiento SARIMA (p, d, q) (P, D, Q), y luego en la serie de residuales se Aplican las redes neuronales para extraer a las componentes no lineales que aún permanecen en la serie residual. Los resultados de realizar esta combinación SARIMA-ANN, se presentan en la TABLA 2.

TABLA 2. Mejores modelos híbridos SARIMA-ANN.

Modelo Híbrido **ARIMA - ANN**	RMSE	MAE	MAPE	MAPE MEDIO
SARIMA(3,1,0)(2,1,0)7 + MLP(4,5,1)	2467.732	2146.518	43.46059	35.13386
SARIMA(3,1,0)(0,1,2)7 + MLP(7,5,1)	2245.47	1874.799	40.53245	
SARIMA(1,0,1)(0,1,2)7 + MLP(4, 5, 1)	1168.942	929.3524	21.40854	
SARIMA(1,0,1)(0,1,2)7 * MLP(3,5,1)	3036.427	2755.838	49.40641	49.196315
SARIMA(3,1,0)(2,1,0)7 * MLP(2,5,1)	2830.986	2572.897	48.98622	
SARIMA(3,1,0)(0,1,2)7 *MLP(1,5,1)	*	*	*	
SARIMA(3,1,0)(2,1,0)7 + NNAR(20,1,5)	2402.994	2098.388	41.1505	33.2722967
SARIMA(3,1,0)(0,1,2)7 + NNAR(20,1,5)	2131.052	1767.427	38.25838	
SARIMA(1,0,1)(0,1,2)7 + NNAR(20,1,6)	1103.697	876.8934	20.40801	
SARIMA(3,1,0)(2,1,0)7 * NNAR(20,1, 7)	2240.05	1784.666	35.83145	30.02048
SARIMA(3,1,0)(0,1,2)7 * NNAR(20,1, 7)	2079.168	1652.571	31.66039	
SARIMA(1,0,1)(0,1,2)7 * NNAR(20,1, 8)	1664.294	1230.983	22.5696	

Usando el MAPE como criterio de selección en el tramo de validación, se observa que el mejor modelo alcanza un valor de 20.40801 y corresponde a la combinación aditiva SARIMA (1, 0, 1) (0, 1, 2)+NNAR (20, 1, 6). El segundo mejor modelo alcanza un valor de 21.40854 y corresponde a la combinación híbrida aditiva SARIMA (1, 0, 1) (0, 1, 2) + MLP (4, 5, 1), y finalmente el tercer mejor modelo alcanza un valor de 22.95696 y corresponde a la combinación híbrida multiplicativa SARIMA (1, 0, 1) (0, 1, 2)*NAAR (20, 1, 8). De otro lado, usando los valores promedios de los MAPEs, se observa que existe una superioridad del 5.11 %, de los modelos híbridos aditivos SARIMA+ NNAR sobre la combinación de los modelos híbridos aditivos SARIMA+MLP, y de los modelos híbridos aditivos SARIMA+ANN sobre los modelos híbridos multiplicativos SARIMA*ANN en el 5.41%.

No obstante, los modelos híbridos multiplicativos SARIMA*NNAR, son superiores a las demás estructuras de los modelos SARIMA-ANN en más del 3.25%.

Modelamiento híbrido ANN - SARIMA

En esta categoría de modelamiento se inicia extrayendo la componente no lineal (N_t), mediante los 03 mejores modelos de redes neuronales MLP, y NNAR, para luego combinarlos según una estructura aditiva o multiplicativa, con la extracción de la componente lineal (L_t) de la serie residual mediante el modelo SARIMA. Los resultandos en términos de los valores de los errores de pronósticos usados con mayor frecuencia, se presentan en la Tabla 3.

TABLA 3. Mejores modelos híbridos ANN-SARIMA.

Modelo Híbrido **SARIMA - ANN**	RMSE	MAE	MAPE	MAPE MEDIO
MLP(9,9,1) + ARIMA(32,0,0)	815.3445	655.047	16.7214	13.60655
MLP(9,6,1) + ARIMA(62,0,0)	722.695	586.0924	11.394	
MLP(9,4,1) + RIMA(0,0,19)	783.4243	604.9407	12.70426	
MLP(9,9,1) * SARIMA(5,0,0)(2,0,0)	1249.699	923.082	21.91613	15.92384
MLP(9,6,1) * SARIMA(5,0,0)(2,0,0)	875. 95	670.8607	14.91325	
MLP(9,4,1) * SARIMA(5,0,0)(2,0,0)	694.15	509.5406	10.94214	
NNAR(27,1,6) + ARIMA(1,0,1)	786.8008	597.1452	9.236697	10.50343
NNAR(27,1,8) + ARIMA(25,0,0)	876.7596	678.745	10.55018	
NNAR(27,2,5) + ARIMA(55,0,0)	790.2368	671.1552	11.72343	
NNAR(27,1,6) * SARIMA(3,0,2)(1,0,1)	786.9704	592.7973	9.183901	11.70405
NNAR(27,1,8) * SARIMA(1,0,2)(1,0,1)	1032.528	743.6361	11.59612	
NNAR(27,2,5) * SARIMA(2,0,4)(1,0,1)	955.8425	795.7162	14.33215	

Se observan en la Tabla 3, que la combinación híbrida en estructura multiplicativa NNAR(27,1,6)*SARIMA(3,0,2)(1,0,1) alcanza el menor valor del MAPE= 9.18390, el segundo mejor modelo hibrido aditivo NNAR(27,1,6)+ARIMA(1,0,1) con un valor del MAPE=9.236697, y el tercer mejor modelo hibrido aditivo NNAR(27,1,8)+ARIMA(25,0,0)

cuyo valor del MAPE iguala a 10.55018, siendo los coeficientes f_1,f_3,f_{14} y f_{25} diferentes de "0", y los demás rezagos iguales a "0". La diferencia entre los dos mejores modelos híbridos es de apenas el 0.57%, indicando que producen prácticamente los mismos pronósticos.

Usando los valores de los MAPEs medios, se concluye que los modelos híbridos NNAR (p, P, h)-SARIMA (p, d, q) (P, D, Q) producen mejores pronósticos que los MLP (k, h, o)-SARIMA (p, d, q) (P, D, Q), porque reducen en promedio el valor del MAPE en más del 3.66%. Dentro de cada categoría de combinación, los modelos híbridos aditivos NNAR(p, P, h) + SARIMA (p, d, q)(P,D,Q) son superiores, en sus pronósticos, que los modelos híbridos multiplicativos NNAR (p, P, h)*SARIMA (p, d, q) (P,D,Q) porque reducen en promedio el valor del MAPE en más del 1.20%; mientras que para el caso de los modelos híbridos aditivos MLP(k,h,o)+SARIMA(p, d, q)(P,D,Q) son superiores en pronósticos que los modelos híbridos aditivos MLP(k,h,o) * SARIMA(p, d, q)(P,D,Q) porque reducen en promedio el valor del MAPE en más del 2.31%. De esta forma, los modelos híbridos aditivos, en promedio producen mejores pronósticos que los modelos híbridos multiplicativos del número de infectados del SARS-CoV-2 en el Perú.

3.3 ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Se diseñó la capacidad predictiva de los modelos SARIMA, modelos neuronales MLP (k, h, o), NNAR (p, P, h), los modelos híbridos aditivos y multiplicativos SARIMA-ANN, y ANN-SARIMA para los datos de los nuevos casos diarios confirmados de infección del SARS-CoV-2 en el Perú. Se concluyó que los modelos híbridos ANN-SARIMA tienen una superioridad predictiva cuando se usa el error absoluto medio porcentual (MAPE), sobre los modelos neuronales NNAR(p, P, h), y MLP(k,h,o) respetivamente. Esto se interpreta que cuando la serie de tiempo tiene patrones de no linealidad fuertes, entonces resulta más conveniente, iniciar la extracción de la componente N_t con los modelos de redes neuronales NNAR y MLP, y proseguir con la extracción de la componente lineal L_t vía la metodología SARIMA.

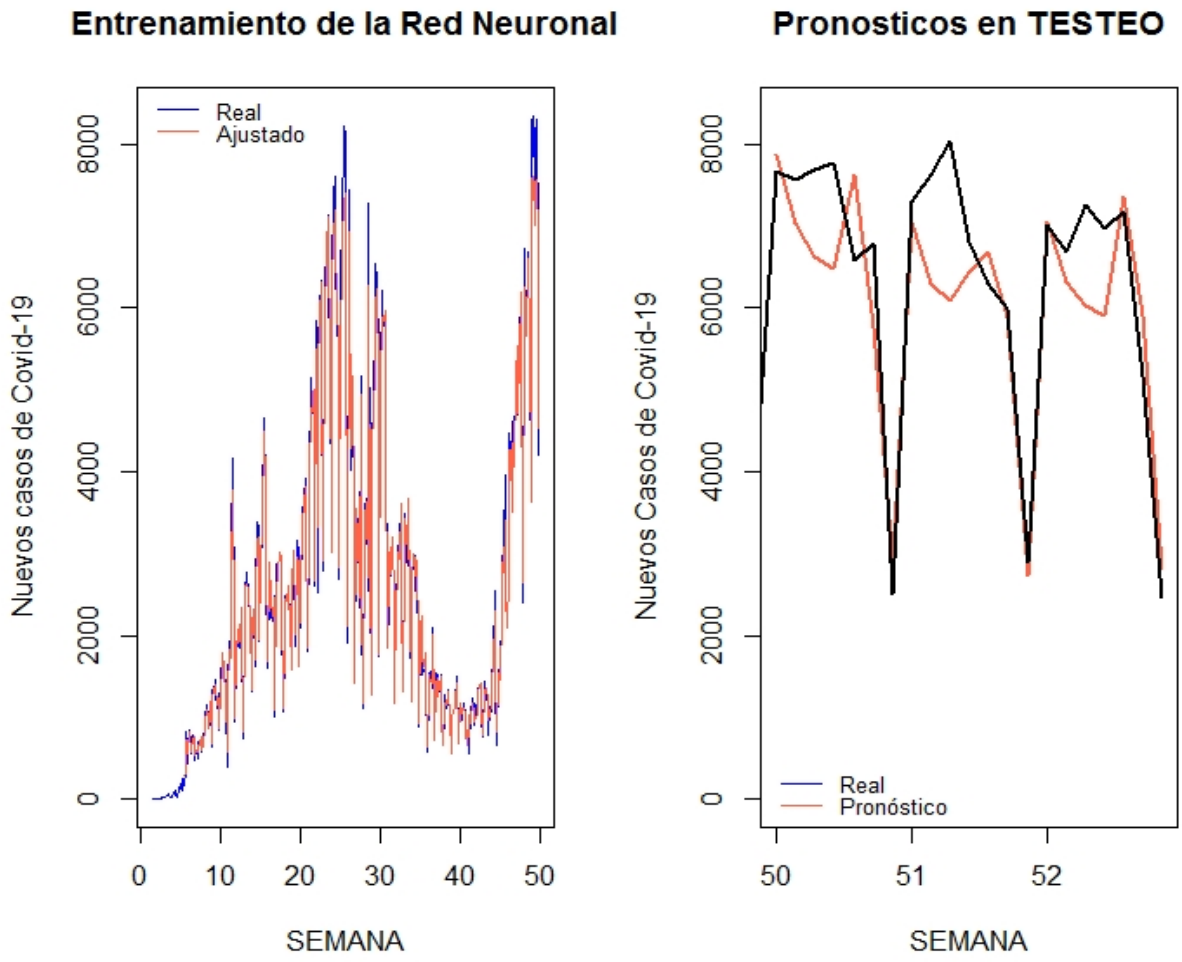
La combinación híbrida aditiva ANN+SARIMA produce los mejores resultados, con respecto al valor de los errores medios porcentuales de la combinación entre las categorías de modelamiento realizado. Esta conclusión es consistente con la fundamentación teórica de que los modelos de redes neuronales son buenos aproximadores de cualquier función no lineal, con los resultados hallados en el estudio sobre el pronóstico del número de hospitalizaciones durante la segunda ola en Italia (Peroné, 2020b), y con los resultados obtenidos bajo las estructuras aditiva y multiplicativa respectivamente (Zhang, 2003; y Wang et.al., 2013).

La combinación híbrida de los modelos de redes neuronales auto regresivas tiene en promedio una superioridad en capacidad predictiva sobre los modelos de redes neuronales del tipo perceptrón multicapas. Esta conclusión puede atribuirse al hecho de que la red

neuronal perceptrón multicapas no tiene la opción directa para capturar las variaciones estacionales, aunque ellas pueden ser implementadas a través de las variables artificiales de estación, opción que no fue usada en este estudio.

Por otro lado, la comparación de modelos individuales identifica al modelo híbrido multiplicativo, NNAR(27,1,6)*SARIMA(3,0,2)(1,0,1), como el mejor modelo para realizar pronósticos, seguido del modelo hibrido aditivo NNAR(27,1,6) + ARIMA(1,0,1), cuya diferencia del valor del MAPE es pequeña. Indicando que los modelos híbridos NNAR-SARIMA con estructura aditiva o multiplicativa tienen la misma capacidad predictiva.

Figura 2. Pronósticos con el mejor modelo híbrido NNAR (27, 1, 6) * ARIMA (3, 0,2) (1, 0, 0)



En la Figura 2, se observa que los pronósticos con el mejor modelo hibrido multiplicativo NNAR (27, 1, 6) * ARIMA (3, 0, 2) (1, 0, 0), en la etapa de entrenamiento emula casi perfectamente a las realizaciones de la serie de tiempo en estudio, exceptuando puntos atípicos, que necesitan ser tratadas previamente. En la etapa de validación, este modelo no reproduce de forma satisfactoria la variación del cuarto día de la semana, requiriendo algún tipo de variación a ser identificado en más detalle, por lo que es posible mejorar el

modelamiento en investigaciones futuras. Entre estos otros tipos de variaciones se incluyen los factores medio ambientales, las decisiones de distanciamiento, aislamiento, cuarentenas e inamovilidad en los fines de semana implementadas por el Gobierno para contrarrestar el impacto de la acción viral. Estos tipos de variaciones no se usaron en el presente estudio.

CONCLUSIONES

El objetivo de este estudio fue construir un modelo hibrido que combine los modelos SARIMA y los modelos de redes neuronales MLP (k, h, o) y NNAR (p, P, h) con el fin de mejorar las predicciones diarias de los nuevos casos confirmados de infección por SARS-CoV-2 en el Perú. Para ello se ajustó 41 modelos desde los cuales se han evaluado los valores promedios de los MAPEs por categoría de modelamiento a fin de verificar las hipótesis formuladas. Los resultados del procesamiento de la información obtenida desde el 06/03/2020 hasta el 28/02/2021, de 360 observaciones diarias ha permitido concluir que:

1. Según los valores promedios del MAPE de los 03 mejores modelos por categoría de modelamiento, los pronósticos realizados por modelos híbridos aditivos NNAR-SARIMA son mejores predictores que los modelos NNAR, SARIMA, por separados, debido a su reducción en los valores del MAPE del 0.0371%, para el NNAR, y del 22.36% respecto del SARIMA. Esta conclusión no se cumple con los modelos para las otras combinaciones estudiadas.
2. Para el valor promedio del MAPE de los 03 mejores modelos por categoría de modelamiento no se verifica que los modelos híbridos multiplicativos produzcan mejores pronósticos que los modelos NNAR y MLP por separados. No obstante a nivel del modelo individual NNAR27,1,6)*SARIMA(3,0,2) (1,0,1), y MLP(9,4,1)*SARIMA (5,0,0) (2,0,0), sí se verifican que este modelo híbrido multiplicativo, son superiores a los modelos separados que se combinan, debido a sus reducciones en el valor del MAPE.
3. Según el valor promedio del MAPE, para los 03 mejores modelos por categoría de modelamiento se verifica, que los modelos híbridos aditivos NNAR+ARIMA tienen una capacidad predictiva mayor que los modelos híbridos multiplicativos NNAR*SARIMA, cuya reducción alcanza al 1.20%; mientras que los modelos aditivos MLP+SARIMA tienen una capacidad predictiva mayor que los modelos híbridos multiplicativos MLP*ARIMA, debido a que las reducciones en el valor del MAPE son, en promedio del orden del 2.317%. No obstante, en la evaluación para los mejores modelos en cada categoría de modelamiento se verifica que los modelos híbridos multiplicativos NNAR*SARIMA y MLP*ARIMA, tienen una ligera superioridad sobre los modelos híbridos aditivos NNAR+ARIMA y MLP+ARIMA, cuyas reducciones son del orden del 0.052% y 0.451% respectivamente.

Finalmente se recomienda que, en vista que se tienen varios puntos atípicos, se realice un pre procesamiento más exhaustivo sobre el tratamiento de estas observaciones, considerar los diversos factores implementados por el Gobierno para hacer frente la lucha contra el impacto en la salud del SARS-CoV-2, para ello usarlos como variables auxiliares que permitan explicar a las variaciones de la serie de tiempo y posibiliten la combinación de los modelos de redes neuronales con las metodologías ARIMAX y SARIMAX. Por el lado de los modelos de redes neuronales, se deben aplicar otros tipos de arquitecturas como las redes neuronales recurrentes que puedan mejorar el entrenamiento de los modelos neuronales.

REFERENCIAS

Benvenuto D.; Giovanetti M.; Vasallo L.; Angeletti S.; & Ciccozzi M.2020) Application of the ARIMA model on the Covid 19 epidemic dataset. Data Brief. 2020; 29: 105340. Published 2020 Feb 26. Doi: 10.1016 / j .dib.2020.105340.

Brockwell, P. & DAVIS R. (1991) Time Series: Theory and Methods. Colorado State University, second edition, Springer-Verlag, New York Inc.

Box G.E.P & Jenkins G.M (1970) Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

Ceylan Z. (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of the Total Environment 729 (2020) 138817.

Cybenko G. (1989) Approximation by superpositions of a sigmoid function. Mathematics of Control Signals and Systems 2, 303–314.

Dehesh T; Fard H.A. M.; & Dehesh P. (2020) Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. MedRxiv preprint,

DOI: https://doi.org/10.1101/2020.03.13.200 35345.

Ding G. LI X.; Jiao F.; & Shen Y. (2020) Brief Analysis of the ARIMA model on the Covid 19 in Italy. MedRxiv preprint https://doi.org/10.1101/2020.04.08.20058636

Ganiny S. & Nisar O. (2021) Mathematical modeling and a month ahead forecast of the coronavirus disease 2019 (COVID19) pandemic: an Indian

Scenario. Modeling Earth Systems and Environment. Modeling Earth Systems and Environment. 2021 Jan: 1-12. DOI: 10.1007/s40808-020-01080-6.

Gobierno del Perú (2021) Datos abiertos Covid-19 Plataforma Nacional Ministerio de Salud-MINSA. https:// www.datosabiertos.gob.pe/dataset/casos-positivos-por-covid-19-ministerio-de-salud-minsa

Hamilton, J. D. (1994) Time Series Analysis. Princeton University Press, Princeton NJ.

Hydman R.J. and Athanasopoulos G. (2013) Forecasting: Principles and Practice. Monash University, Australia.

Horni K.; Stinchcombe, M.; & White H. (1989) Multilayer feed forward networks are universal approximators. Neural Networks 2, 359–366.

Hooker R. H. (1901) “On the correlation of the marriage-rate with trade." Journal Roy. Stat. Soc., London, vol. 64, p. 485, 1901.

Hornik K.; stinchcombe M.; & White H. (1990) Universal approximation of an unknown mapping and its derivatives using multilayer feed forward networks. Neural Networks 3, 551–560.

Nielsen A. (2020) Practical Time Series Analysis. Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, and Sebastopol. USA

Pankratz A. (1983) Forecasting with Univariate Box-Jenkins Models: Concepts and Cases First edition. John Wiley & Sons, Inc.

Perone G. (2020a) An ARIMA model to forecast the spread and the final size of Covid 19 epidemic in Italy. No. 20/07. HEDG, c/o Department of Economics, University of York,

Perone G. (2020b) Comparison of ARIMA, ETS, NNAR and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. October 2020 https://arxiv.org/ftp/arxiv/papers/2010/2010.11617.pdf

Safi S.K. & Sanusi I.S. (2021) A hybrid of artificial neural network, exponential smoothing, and ARIMA models for COVID-19 time series forecasting. Model Assisted. Statistics and Applications 16 (2021) 25–35 25 DOI 10.3233/MAS-210512 IOS Press.

Shibata K. & Ikeda Y.(2009) “Effect of number of hidden neurons on learning in large-scale layered neural networks,” in Proceedings of the ICROS-SICE International Joint Conference 2009 (ICCASSICE ’09), pp. 5008–5013, August 2009.

Trenn S. (2008) “Multilayer perceptrons: approximation order and necessary number of hidden units,” IEEE Transactions on Neural Networks, vol. 19, no. 5, pp. 836–844.

Uriel J.E.(1985) Análisis de series temporales: Modelos ARIMA. Paraninfo Madrid España

Wang L.; ZOU H.; SU J.; LI L.; & CHAUDHRY S. (2013) An ARIMA-ANN Hybrid Model for Time Series Forecasting. Systems Research and Behavioral Science Syst. Res. 30, 244–259 (2013). Research paper.

Wei W.W.S (1990) Time series analysis univariate and multivariate methods. Temple University. First edition Addison-Wesley, Reading, MA.

Yule G.U. (1909) The Applications of the Method of Correlation to Social and Economic Statistics. Journal of the Royal Statistical Society, Vol. 72, No. 4 (Dec., 1909), pp. 721-730

Yule G.U (1927) On a method of investigations periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philos. Trans. Roy. Soc. London Ser. A 226 267-298.

Zhang G.P. (1998) Linear and nonlinear time series forecasting with artificial neural networks. Ph.D. Dissertation, Kent State University, Kent, OH.

Zhang G. P.(2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50: 159-75.v

Zhang G.P. & QI M. (2005) Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, 160, pp. 501-514.