Are Neural Networks Able To Forecast Nonlinear Time Series With Moving Average Components?

M. R. Cogollo and J. D. Velásquez, *Senior Member*, *IEEE*

***Abstract*— In nonlinear time series forecasting, neural networks are interpreted as a nonlinear autoregressive models because they take as inputs the previous values of the time series. However, the use of neural networks to forecast nonlinear time series with moving components is an issue usually omitted in the literature. In this article, we investigate the use of traditional neural networks for forecasting nonlinear time series with moving average components and we demonstrate the necessity of formulating new neural networks to adequately forecast this class of time series. Experimentally we show that traditional neural networks are not able to capture all the behavior of nonlinear time series with moving average components, which leads them to have a low capacity of forecast.**

***Keywords*— Artificial neural networks; prediction; nonlinear time series; forecasting; moving averages.**

I. Introducción

SI bien el pronóstico de series de tiempo generalmente se ha realizado bajo el supuesto de linealidad, lo cual ha promovido el estudio y uso de modelos lineales tales como el autorregresivo (AR), promedios móviles (MA, por su sigla en inglés), autorregresivo de promedios móviles (ARMA, por su sigla en inglés) y autorregresivo integrado de promedios móviles (ARIMA, por su sigla en inglés) [1,2], se ha encontrado que en la realidad los sistemas a menudo presentan una estructura no lineal desconocida [3]. Para abordar este tipo de problema, se han propuesto varios modelos no lineales, como son los modelos bilineales, autorregresivo de heterocedasticidad condicional (ARCH, por su sigla en inglés) y sus extensiones, autorregresivo de transición suave (STAR, por su sigla en inglés), no lineal autorregresivo (NAR, por su sigla en inglés), de redes wavelets y de redes neuronales artificiales (ANN, por su sigla en inglés) [1-7].

Con relación a las ANN, se encuentra que su teoría es muy amplia, y han sido aplicados en el modelado y pronóstico de datos de distintas áreas del conocimiento [1-3,8-14]; sin embargo, en la literatura se encuentra que gran parte de los modelos ANN propuestos están basadas exclusivamente en una estructura no lineal autorregresiva, y sólo unos pocos consideran el hecho de que el proceso generador de la serie no lineal tenga, además de la parte autorregresiva, una componente de promedios móviles. Específicamente, para abordar este caso, algunos autores sugieren usar la red neuronal NARMA y la red neuronal autorregresiva ARNN de alto orden; en [15,16] se presentan casos específicos.

Sin embargo, al revisar la literatura más relevante se encuentra que:

- La teoría del modelo NARMA(p, q) considera que el proceso generador de los datos corresponde a una estructura no lineal con componentes tanto autorregresivas como de promedios móviles; esto hace, que al ignorar la componente autorregresiva (haciendo $p = 0$) se obtenga un modelo no lineal de promedios móviles o NLMA; sin embargo, en la literatura no hay estudios que examinen la capacidad de pronóstico del modelo NARMA(0,q) cuando es aplicado en series de tiempo no lineales que presentan una componente inherente MA.
- No hay evidencias reportadas de que un modelo MA no lineal pueda ser aproximado por un modelo AR no lineal de orden infinito, como si pasa en el caso de los modelos lineales cuando se cumplen ciertas condiciones de invertibilidad.

El objetivo de esta investigación es responder las preguntas de investigación presentadas a continuación con el fin de esclarecer los vacíos anteriores:

1. ¿Un modelo no lineal AR de alto orden, representado por una red ARNN, puede aproximar bien un modelo no lineal MA de orden reducido?
2. ¿Cuándo en una red recurrente NARMA se asume que no hay un proceso autorregresivo, se pueden pronosticar adecuadamente series de tiempo no lineales que contengan componentes inherentes de promedios móviles?

Estas preguntas serán resueltas partiendo del planteamiento de la invertibilidad de los modelos no lineales MA y de simulaciones usando datos experimentales.

La importancia y originalidad de este trabajo se fundamenta en el hecho de que hasta la fecha no hay evidencia en la literatura de estudios que analicen e identifiquen los problemas que surge al modelar y pronosticar series de tiempo con componentes inherentes MA usando redes neuronales. El artículo está organizado como sigue: en las Secciones II y III se presentan el modelo no lineal MA, y las redes neuronales NARMA y NAR, respectivamente. Posteriormente, en la Sección IV, se muestra la metodología empleada y resultados obtenidos para evaluar la capacidad de éstas redes para pronosticar series de tiempo no lineales con componente MA. En la Sección V se presentan los resultados obtenidos, mientras que en la Sección VI se responden las preguntas de

M. R. Cogollo, Universidad EAFIT, Medellín, Colombia, mcogollo@eafit.edu.co
J. D. Velásquez, Universidad Nacional de Colombia, Sede Medellín, Medellín, Colombia, jdvelasq@unal.edu.co

investigación planteadas. Finalmente, se concluye en la Sección VII.

II. EL MODELO NO LINEAL DE PROMEDIOS MÓVILES

En el modelo no lineal de promedios móviles de orden q, denotado como $\mathrm{NLMA}(q)$, el valor actual de la serie de tiempo, y_t, es una función no lineal conocida $h(\cdot)$ de las q innovaciones pasadas $\{\varepsilon_{t-1}, \dots, \varepsilon_{t-q}\}$ y la innovación actual ε_t. Esto es:

$$y_t = \varepsilon_t + h\left(\varepsilon_{t-1}, \dots, \varepsilon_{t-q}; \boldsymbol{\theta}\right); \quad t = 1,2, \dots \tag{1}$$

donde $\boldsymbol{\theta}$ representa el vector de parámetros de la función $h(\cdot)$ y $\{\varepsilon_t\}$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas, centradas en cero y con varianza constante.

Dependiendo de la forma que adopte la función $h(\cdot)$, se han propuesto los siguientes modelos NLMA:

- Polinomial de promedios móviles propuesto por Robinson [17].
- Asimétrico de medias móviles propuesto por Wecker [18].
- No lineal de medias móviles con respuesta de largo alcance propuesto por Robinson y Zaffaroni [19].
- No lineal de medias móviles integrado de Engle y Smith [20].

A diferencia del modelo no lineal autorregresivo (NAR, por su sigla inglés), el modelo NLMA ha sido poco explorado, tanto empírica como teóricamente. Este hecho se debe, en parte, a la dificultad que se presenta para establecer la propiedad de invertibilidad del modelo [21]; dicha propiedad se refiere a la posibilidad de realizar la reconstrucción de las innovaciones ε_t a partir de las observaciones y_t, suponiendo que el verdadero modelo es conocido. Sin embargo, Chan y Tong [22] establecieron que el modelo NLMA puede llegar a ser localmente invertible; es decir, que se pueden establecer condiciones iniciales que permiten reconstruir asintóticamente las innovaciones a partir de las observaciones.

El hecho de que el modelo NLMA no sea globalmente invertible, hace que, al menos teóricamente, no sea equivalente a un modelo NAR de alto orden, como si pasa en el caso lineal. Es importante verificar la invertibilidad del modelo NLMA para garantizar que es apropiado para fines de pronóstico y, además, hacer posible su diagnóstico.

III. MODELOS DE REDES NEURONALES ASOCIADOS CON COMPONENTES DE PROMEDIOS MÓVILES

Matemáticamente, una neurona es una función no lineal, acotada y parametrizada de la forma [23]:

$$o = f\left(x_1, x_2, \dots, x_n;\ \omega_1, \omega_2, \dots, \omega_p\right) = f(\boldsymbol{x}; \boldsymbol{\omega})$$

donde:

- $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ es el vector de variables de entrada a la neurona.
- $\boldsymbol{\omega} = (\omega_2, \dots, \omega_p)$ es el vector de pesos (parámetros) asociados a las conexiones de entrada de la neurona.
- $f(\cdot)$ es una función no lineal de activación.

A su vez, una red neuronal artificial se define como una composición de funciones no lineales de la forma:

$$y = g_1 \circ g_2 \circ \cdots \circ g_N \left(f_1(\boldsymbol{x}; \boldsymbol{\omega}), f_2(\boldsymbol{x}; \boldsymbol{\omega}), \dots, f_p(\boldsymbol{x}; \boldsymbol{\omega})\right)$$

donde:

- y es la variable respuesta o salida de la red neuronal artificial.
- g_i para $i = 1, \dots, N$, son funciones no lineales.
- $f_j(\boldsymbol{x}; \boldsymbol{\omega})$ para $j = 1, \dots, p$, son funciones definidas como en (1).
- N representa el número de capas ocultas en la red.
- p denota el número de neuronas en las capas ocultas.
- El símbolo $\circ$ entre las funciones indica la operación composición.

Las redes neuronales, según su arquitectura e interconexión entre neuronas, se pueden clasificar en dos clases: redes de alimentación hacia adelante (feedforward) y redes retroalimentadas (recurrentes o feedback). La red feedforward, también conocida como estática, constituye una función no lineal de sus entradas, y es representada como un conjunto de neuronas conectadas entre sí, en la cual la información fluye sólo en la dirección hacia adelante, desde las entradas hacia las salidas. Específicamente en [24] se define de la siguiente forma un modelo de red feedforward con una sola neurona de salida y q capas ocultas:

$$o_t = \Phi\left(\beta_0 + \sum_{i=1}^{q} \beta_i \Psi\left(\alpha_i + \sum_{j=1}^{n} \omega_{ij} x_{j,t}\right)\right) =: f(\boldsymbol{x}_t; \boldsymbol{\theta}) \tag{2}$$

donde

- o_t es el estimador de la variable objetivo y_t.
- $\boldsymbol{x}_t = (x_{1,t}, \dots, x_{n,t})$, son n variables de entradas medidas en el tiempo t.
- $\Phi(\cdot)$ y $\Psi(\cdot)$ son las funciones de activación de la red neuronal.
- $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots \beta_q, \alpha_1, \dots, \alpha_q, \omega_{11}, \dots, \omega_{qn})$ representa el vector de parámetros de la red neuronal, el cual es estimado a partir de la minimización de la suma de residuales al cuadrado $\sum_{t=1}^{n} (y_t - \hat{o}_t)^2$.

Es de resaltar que éste es el tipo de redes neuronales más estudiado y aplicado en la literatura, debido principalmente a que son un aproximador universal de funciones [25-27]; y además, porque en la práctica son las redes más sencillas en cuanto a su implementación y simulación. Por su parte, la red feedback, también conocida como dinámica o recurrente, se caracteriza porque su arquitectura presenta ciclos: las salidas de las neuronas de una capa pueden ser entradas a la misma neurona o entradas a neuronas de capas previas. Para mayor información de este tipo de redes se sugiere examinar [23] y [28].

A continuación se describen casos particulares de éstos tipos de redes: la red neuronal autorregresiva ARNN, la cual es de tipo feedforward y la red neuronal recurrente NARMA.

A. Red neuronal autorregresiva (ARNN)

El modelo autorregresivo no lineal de orden p, NAR (p), definido como:

$$y_t = h(y_{t-1}, \dots, y_{t-p}) + \varepsilon_t \tag{3}$$

es una generalización directa del modelo lineal AR, donde $h(\cdot)$ es una función no lineal conocida. Se asume que $\{\varepsilon_t\}$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas con media cero y varianza finita σ^2.

La red neuronal autorregresiva (ARNN), es una red feedforward que constituye una aproximación no lineal para $h(\cdot)$, la cual es definida como:

$$\hat{y}_t = \hat{h}(y_{t-1}, \dots, y_{t-p}) = \beta_0 + \sum_{i=1}^{I} \beta_i f\left(\alpha_i + \sum_{j=1}^{p} \omega_{ij} y_{t-j}\right) \tag{4}$$

donde la función $f(\cdot)$ es la función de activación y $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots \beta_I, \alpha_1, \dots, \alpha_I, \omega_{11}, \dots, \omega_{Ip})$ es el vector de parámetros.

B. Red neuronal recurrente NARMA

Una generalización del modelo lineal ARMA al caso no lineal está dado por

$$y_t = h(y_{t-1}, \dots, y_{t-p}, \varepsilon_{t-1}, \dots, \varepsilon_{t-q}) + \varepsilon_t$$

donde $h(\cdot)$ es una función no lineal conocida y $\{\varepsilon_t\}$ se define como en (3). Este modelo se denomina NARMA (p, q).

Dado que la secuencia $\varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ no es observable directamente, entonces se debe hallar $\hat{y}_t$ empleando un algoritmo recursivo de estimación que considere los siguientes cálculos:

$$\hat{y}_t = h(y_{t-1}, \dots, y_{t-p}, \hat{\varepsilon}_{t-1}, \dots, \hat{\varepsilon}_{t-q}) \tag{5}$$

$$\hat{\varepsilon}_j = y_j - \hat{y}_j, \quad j = t-1, \dots, t-q \tag{6}$$

bajo condiciones iniciales apropiadas [16]. Justamente, el modelo de red neuronal recurrente NARMA (p, q) surge al aproximar (5) y (6) empleando la red recurrente:

$$\hat{y}_t = \alpha_0 + \sum_{j=1}^{h} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-i} + \sum_{i=p+1}^{p+q} \beta_{ij} \hat{\varepsilon}_{t+p-i}\right) \tag{7}$$

donde $\hat{\varepsilon}_{t+p-i} = y_{t+p-i} - \hat{y}_{t+p-i}$.

Al observar la formulación matemática del modelo (7), se podría considerar que una alternativa para modelar una serie de tiempo no lineal con una componente inherente de promedios móviles es emplear un modelo NARMA (0,q). Éste hecho se discutirá en la siguiente sección.

IV. METODOLOGÍA EMPLEADA

La evaluación de la capacidad de pronóstico de los modelos de redes neuronales NARMA (p,q) y ARNN (p) se realizó usando dos conjuntos de datos experimentales provenientes de los modelos descritos en la Tabla I. En el Modelo 1 se define $\{\varepsilon_t\}$ como en (3), y corresponde al modelo NLMA (2) examinado por Zhang et al. [29]. Por otra parte el Modelo 2 fue considerado por Burges y Refenes [15] para ejemplificar el uso de redes neuronales con error feedback bajo una variante del algoritmo de Esperanza- Maximización en el proceso de entrenamiento.

TABLA I. MODELOS GENERADORES DE DATOS.

Modelo	Estructura del modelo
1	$y_t = \varepsilon_t - 0.3\varepsilon_{t-1} + 0.2\ \varepsilon_{t-2} + 0.4\varepsilon_{t-1}\varepsilon_{t-2}$
2	$y_t = \varepsilon_t + 0.5\varepsilon_{t-1} + 06\varepsilon_{t-1}\varepsilon_{t-2}$

Nótese que los dos modelos no contienen términos autorregresivos (no consideran valores pasados de y_t), y además corresponden a distintos niveles de complejidad de la función $h(\cdot)$ en (1).

Se generaron 100 series temporales a partir de cada Modelo. De las cuales, en cada serie generada, las primeras observaciones fueron usadas para la estimación de los parámetros del modelo y las restantes se usaron como conjunto de validación. En la Figura 1 se grafica una de las series del Modelo 1 con $n = 360$ observaciones. En proceso generados de datos, se usaron diferentes inicios aleatorios muestreados de una distribución N(0; 1,5) para el término del error del Modelo 1, y se supuso en el Modelo 2 que $\varepsilon_{-1} = \varepsilon_{-2} = 0$ y $y_0 = \varepsilon_0 = rand(\ )$. Siendo rand( ) un número aleatorio uniforme estándar.

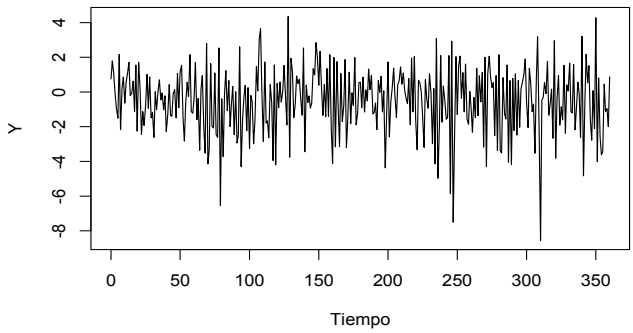

Figura 1. Ejemplo de la serie de tiempo generada por el Modelo 1.

La experimentación se enfocó en dos aspectos: (i) el análisis de la capacidad para capturar todo el proceso no lineal de promedios móviles empleando una red neuronal recurrente NARMA $(0, q)$ o una ARNN (p) con p suficientemente grande, (para lo cual se usó el Modelo 1), y (ii) comparar los resultados obtenidos con alguna de las redes consideradas en este trabajo con los hallados en la literatura para modelar procesos NLMA. En este caso se usó el Modelo 2, y se compararon los

resultados de Burges y Refenes [15] con los obtenidos por una red ARNN (p).

En ese sentido, la metodología usada para cada modelo tiene algunos aspectos diferenciadores:

Modelo 1:

- Se consideraron distintos tamaños muestrales $n = \{100; 200; 360\}$ y porcentajes de datos para entrenamiento de la red (50, 65 y 80), para examinar el efecto que tiene la elección de éstos sobre los valores pronosticados.
- Para el modelo ARNN se examinaron valores de rezagos grandes $p = \{10; 15; 25; 50; 100\}$ con la finalidad de responder la primera pregunta de investigación.
- La estructura de la red a emplear fue considerada con base en los resultados hallados por Zhang et al. [29], quienes vía simulación muestran que la mejor estructura de la red corresponde a una capa oculta con un máximo de dos neuronas. La función objetivo fue minimizar el error cuadrático medio (MSE).
- Para el caso del modelo NARMA, además de la estructura de red anterior, se consideraron los siguientes rezagos para el proceso de promedios móviles $q = \{1; 2; 3; 4; 5; 6; 7; 8; 9; 10\}$.
- Se generó una serie adicional de 150 observaciones que fue empleada como datos de prueba.

Modelo 2: Se consideraron las mismas condiciones experimentales empleadas por Burges y Refenes [15] a fin de poder comparar los resultados:

- El tamaño de las series fue 400 observaciones, de las cuales el 70% inicial se emplea para entrenar la red y el 30% restante para validación.
- La función objetivo fue minimizar el error cuadrático medio normalizado (ECMN).
- Se emplearon en todas las redes una capa oculta con cuatro neuronas.
- Se consideraron los siguientes valores de rezagos grandes $p = \{10; 25; 50\}$.
- Se generaron 100 datos adicionales, que fueron tomados como datos de prueba.

En los dos modelos la función de activación empleada fue la logística, para cada entrenamiento, los pesos y sesgos iniciales de la red fueron generados de una distribución uniforme continua en el rango de $(-5; 5)$, y además la elección del mejor modelo se realizó considerando las 100 series y distintas configuraciones de la red, bajo el procedimiento de validación cruzada sugerido por Zemouri et al. [30], a saber:

1. Realizar desde i=1 hasta M=1000 veces, desde distintos puntos iniciales:
   - Entrenar la red usando los datos de entrenamiento.
   - Validar la red entrenada usando los $n.val$ datos de validación. Calcular el error medio de pronóstico $E(i)$ y la desviación estándar $std(i)$ sobre el conjunto de validación:

$$E(i) = \frac{1}{n.val}\sum_{j=1}^{n.val}(y_j - \hat{y}_j) \tag{9}$$

$$std(i) = \sqrt{\frac{1}{n.val}\sum_{j=1}^{n.val}(y_j - \hat{y}_j)^2} \tag{10}$$

2. Calcular las siguientes medidas para evaluar el desempeño de los pronósticos de la red:
   - $M1 = \bar{E} = \frac{1}{M}\sum_{i=1}^{M} E(i)$. Corresponde a una estimación de la media global de los errores medios de pronóstico, y evalúa la cercanía entre los valores pronosticados y los reales. Si $M1 = 0$, entonces la probabilidad de que el pronóstico esté centrado alrededor de los datos reales es muy alta.
   - $M2 = \overline{std} = \frac{1}{M}\sum_{i=1}^{M} std(i)$. Mide la precisión (en términos de variabilidad) de los pronósticos. El valor ideal es $M2 = 0$, debido a que éste indica que se tiene una probabilidad significativa de que los valores pronosticados no están dispersos (es decir, tienen baja variabilidad).
   - $M3 = \frac{\sqrt{\frac{1}{M}\sum_{i=1}^{M}[E(i)-\bar{E}]^2} + \sqrt{\frac{1}{M}\sum_{i=1}^{M}[std(i)-\overline{std}]^2}}{2}$. Sirve para indagar si el proceso de entrenamiento de la red es repetible (en cuyo caso $M3 = 0$), de modo que se obtenga siempre la misma estructura de la red neuronal en cada corrida del proceso de entrenamiento, independientemente de los valores iniciales.
   - $M4 = \frac{1}{M1+M2+M3}$. Examina la exactitud del pronóstico. Si las salidas de la red son muy cercanas a los valores reales, entonces las medidas $M1$, $M2$ y $M3$ son cercanas a cero, y en ese caso M4 tomará valores muy grandes, de modo que $M4 >> 0$ es el valor ideal para tener confianza en los pronósticos.
3. Elegir como mejor candidata la red que tenga el mayor valor $M4$ y menores valores $M1, M2, M3$, sobre el conjunto de validación. De esta forma se evitan problemas de sobreajuste y subajuste. Finalmente, de las M corridas realizadas para dicha red, se selecciona el modelo con menor E(i).
4. Realizar la verificación en los datos de prueba: calcular E(i) y std(i) para cada configuración seleccionada (una por cada serie considerada). Elegir como modelo final aquel que proporcione menor E(i).

Las medidas anteriores, se utilizaron para validar la precisión de los resultados obtenidos con las redes objeto de estudio.

V. RESULTADOS

Los resultados hallados se presentan a continuación para cada modelo considerado.

A. Modelo 1

Las Figuras 2, 3 y 4, muestran los valores obtenidos para las medidas $M1 - M4$ sobre el conjunto de validación de la red ARNN para cada tamaño muestral, bajo los distintos números de rezagos y porcentajes de entrenamiento considerados. A su vez la Figura 5 contiene los valores de las medidas de rendimiento $E(i)$ y $std(i)$, obtenidos en el conjunto de validación. La Tabla II muestra los resultados hallados, para la red ARNN, en los datos de prueba bajo los nueve escenarios considerados y los valores de rezagos grandes $p = \{10; 15; 25; 50; 100\}$. La primera columna contiene el tamaño de muestra, la segunda el número de rezagos p, y las últimas tres columnas muestran los valores hallados de las medidas $E(i)$ y $std(i)$ sobre el conjunto de prueba para cada porcentaje de entrenamiento. En esta tabla el símbolo * indica que el valor del rezago p es superior al tamaño de la muestra del conjunto de validación, por lo cual no se puede examinar la capacidad de pronóstico en ese conjunto de datos.

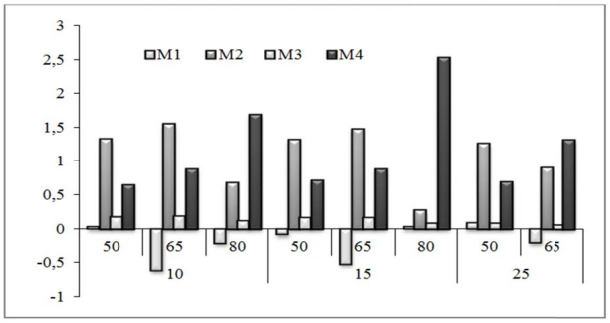

Figura 2. Medidas de rendimiento para el modelo ARNN con $n = 100$, $p = \{10,15,25\}$ y %entrenamiento (50, 65,80).

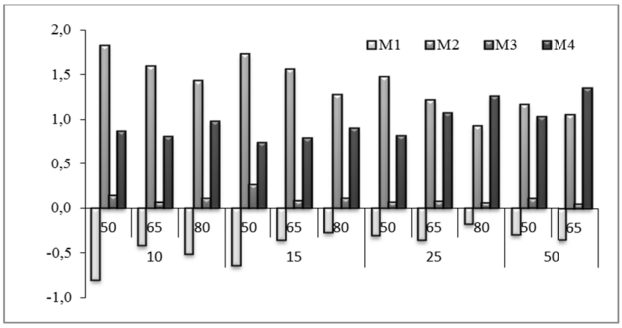

Figura 3. Medidas de rendimiento para el modelo ARNN con $n = 200$, $p = \{10,15,25,50\}$ y %entrenamiento (50, 65,80).

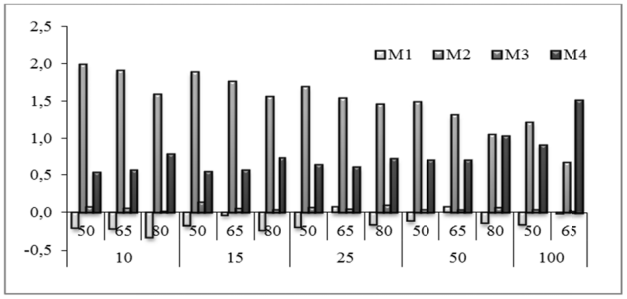

Figura 4. Medidas de rendimiento para el modelo ARNN con $n = 360$, $p = \{10,15,25,50,100\}$ y %entrenamiento (50, 65,80).

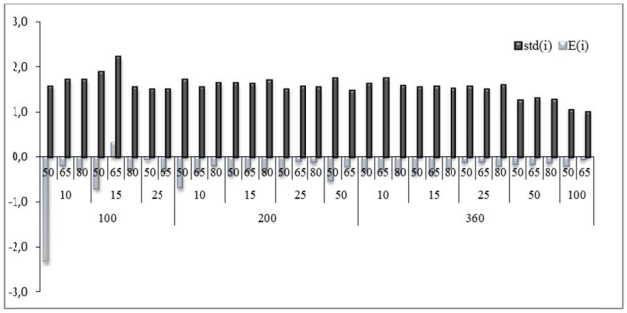

Figura 5. Medidas E(i) y std(i) en el conjunto de validación para el modelo ARNN, según el tamaño muestral, rezagos y porcentaje de entrenamiento.

TABLA II. MEDIDAS DE RENDIMIENTO PARA EL MODELO ARNN EN LOS DATOS DE PRUEBA.

n	p	Medida	Porcentaje de entrenamiento 50	65	80
100	10	$E(i)$	-2.334	-0.1958	-0.2934
		$Std(i)$	1.5836	1.7273	1.7324
	15	$E(i)$	-0.7113	0.3311	-0.3016
		$Std(i)$	1.8923	2.2368	1.5623
	25	$E(i)$	-0.3702	-0.3160	*
		$Std(i)$	1.5233	1.5139	*
200	10	$E(i)$	-0.6903	-0.339	-0.2041
		$Std(i)$	1.7291	1.5689	1.6601
	15	$E(i)$	-0.3939	-0.3065	-0.3379
		$Std(i)$	1.6468	1.6451	1.7154
	25	$E(i)$	-0.3945	-0.1167	-0.1290
		$Std(i)$	1.5177	1.5716	1.5575
	50	$E(i)$	-0.5299	-0.2362	*
		$Std(i)$	1.756	1.4851	*
360	10	$E(i)$	-0.3284	-0.299	-0.346
		$Std(i)$	1.6458	1.7647	1.5909
	15	$E(i)$	-0.3678	-0.371	-0.2601
		$Std(i)$	1.559	1.5777	1.5391
	25	$E(i)$	-0.1201	-0.123	-0.2232
		$Std(i)$	1.5785	1.5136	1.6092
	50	$E(i)$	-0.1744	-0.1713	-0.1388
		$Std(i)$	1.2746	1.3208	1.2823
	100	$E(i)$	-0.2222	**-0.05965**	*
		$Std(i)$	1.06824	**1.01172**	*

A partir de las Figuras 2 a 5, se observa que independientemente del valor del rezago, existe una relación directa entre el porcentaje de entrenamiento y la precisión del pronóstico. Con respecto a la reproducibilidad del modelo, se observa que en general las redes ajustadas siempre satisfacen esta condición. Finalmente, la mayor exactitud del pronóstico se obtiene al combinar el máximo rezago permitido con el máximo porcentaje de entrenamiento y tamaño de muestra. Nótese además, que la calidad del pronóstico, en términos de disminución de los valores $E(i)$ y $std(i)$, es mejor a medida

que $p \to \infty$. Lo cual hace a su vez que la media global y la precisión de los pronósticos converjan a sus valores ideales.

Adicionalmente, de la Tabla II y las Figuras 2, 3 y 4 se deduce que el número de rezagos seleccionados en el modelo ARNN final depende del tamaño de la serie y el porcentaje de datos usados para entrenar la red: para que la red sea capaz de pronosticar adecuadamente, es necesario elegir el máximo número de rezagos permitido y el mayor conjunto de entrenamiento; lo cual conlleva a sospechar que el uso de redes ARNN para pronosticar series con componente inherente MA, tiende a sufrir de problemas de sobre parametrización. Este hecho se corroboró al examinar el comportamiento del MSE según el número de rezagos y capas dela red. Se obtuvo que a medida que aumentaba el orden del modelo AR no lineal, el MSE tiende a disminuir independientemente de los nodos considerados; sin embargo, los menores MSE se obtienen al considerar la red con dos nodos en la capa oculta (véase la Figura 6).

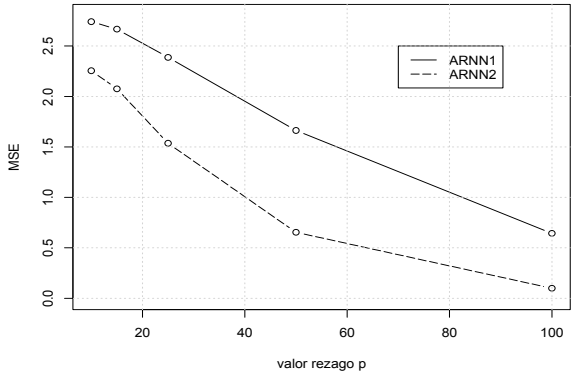

Figura 6. Número de regazos del modelo no lineal versus el MSE de la red ARNN con uno (ARNN1) y dos (ARNN2) nodos en la capa oculta.

El mejor resultado hallado para la red ARNN (en cuanto a mejores resultados en las medidas sobre los datos de prueba) se obtuvo al considerar 360 observaciones, de las cuales el 65% se usaron para entrenar la red con el máximo número de rezagos (100) y 2 nodos en la capa oculta. Sin embargo, ésta no es capaz de capturar todo el proceso no lineal de promedios móviles (véase el gráfico (a) de la Figura 7).

Por otra parte, los resultados hallados sobre la capacidad predictiva de la red neuronal recurrente NARMA ante la presencia de promedios móviles se muestran en la Tabla III y la gráfica (b) de la Figura 7. En la tabla III, la primera columna muestra el tamaño de muestra, y las últimas tres columnas muestran para cada porcentaje de entrenamiento los siguientes resultados: configuración seleccionada (número de rezagos p y número de nodos en la capa oculta k), valores obtenidos para las medidas $M1 - M4$ sobre el conjunto de validación, y los valores $E(i)$ y $std(i)$ para el conjunto de prueba y las últimas tres columnas muestran los valores hallados de éstas medidas para cada porcentaje de entrenamiento.

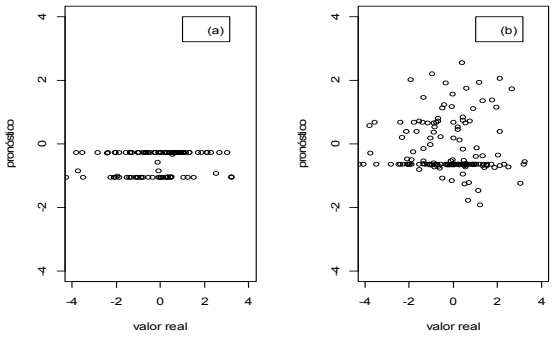

Figura 7. Comparación entre los datos de prueba y sus pronósticos hallados con la mejor red (a) ARNN (100) y (b) NARMA ($q = 2, k = 2$).

De ésta tabla se concluye que la red NARMA requiere considerar tamaños muestrales grandes para ajustar modelos capaces de disminuir la heterogeneidad en los pronósticos hallados para el conjunto de prueba. Así mismo, al igual que en las redes ARNN, el porcentaje de datos empleados para entrenar la red tiene una relación directa con la exactitud del pronóstico, para cualquier tamaño muestral.

Se encontró que el mejor resultado para la red NARMA (en cuanto a las medidas sobre los datos de prueba) fue proporcionado al considerar dos nodos en la capa oculta, $q = 2$ rezagos y 360 observaciones, de las cuales el 80% fue empleado para entrenar la red.

Es de resaltar que si bien la red NARMA tampoco es capaz de capturar todo el comportamiento de los datos con comportamiento de promedios móviles (véase la gráfica (b) de la Figura 7), se encontró que ella (empleando un menor número de parámetros a estimar tiene un mejor desempeño que la red ARNN.

TABLA III. MEDIDAS DE RENDIMIENTO PARA EL MODELO NARMA.

			Porcentaje de entrenamiento		
n	p	Medida	50	65	80
100	10	$E(i)$	-2.334	-0.1958	-0.2934
		$Std(i)$	1.5836	1.7273	1.7324
	15	$E(i)$	-0.7113	0.3311	-0.3016
		$Std(i)$	1.8923	2.2368	1.5623
	25	$E(i)$	-0.3702	-0.3160	*
		$Std(i)$	1.5233	1.5139	*
200	10	$E(i)$	-0.6903	-0.339	-0.2041
		$Std(i)$	1.7291	1.5689	1.6601
	15	$E(i)$	-0.3939	-0.3065	-0.3379
		$Std(i)$	1.6468	1.6451	1.7154
	25	$E(i)$	-0.3945	-0.1167	-0.1290
		$Std(i)$	1.5177	1.5716	1.5575
	50	$E(i)$	-0.5299	-0.2362	*
		$Std(i)$	1.756	1.4851	*

360	10	$E(i)$	-0.3284	-0.299	-0.346
		$Std(i)$	1.6458	1.7647	1.5909
	15	$E(i)$	-0.3678	-0.371	-0.2601
		$Std(i)$	1.559	1.5777	1.5391
	25	$E(i)$	-0.1201	-0.123	-0.2232
		$Std(i)$	1.5785	1.5136	1.6092
	50	$E(i)$	-0.1744	-0.1713	-0.1388
		$Std(i)$	1.2746	1.3208	1.2823
	100	$E(i)$	-0.2222	**-0.05965**	*
		$Std(i)$	1.06824	**1.01172**	*

B. Modelo 2

La tabla IV contiene los valores del error cuadrático medio normalizado (ECMN) hallados por Burges y Refenes [15] para los modelos NARMA con 1, 2 y 3 rezagos (primeras tres filas), y los obtenidos en este trabajo al usar la red ARNN de alto orden y la red NARMA con 1, 2 y 3 rezagos. La información sobre los modelos ARNN y NARMA considerados en esta Tabla es complementada con la Tabla V, la cual contiene para cada modelo la información de las medidas de rendimiento sugeridas por Zemouri et al. [30]. Los valores reales de prueba versus los pronósticos de las mejores redes ARNN y NARMA se muestran en la Figura 8.

TABLA IV. COMPARACIÒN DE RESULTADOS PARA DATOS SIMULADOS DEL MODELO (11).

Modelo	Datos de entrenamiento	Datos de validación	Datos de prueba
NARMA(1) [15]	0.813	0.846	NA
NARMA(2) [15]	0.692	0.755	NA
NARMA(3) [15]	0.689	0.789	NA
ARNN(10)	0.714	0.858	0.0858
ARNN(25)	0.636	0.864	**0.0198**
ARNN(50)	0.623	0.767	0.1390
NARMA(1)	0.743	0.783	0.909
NARMA(2)	0.773	0.714	0.876
NARMA(3)	0.757	0.787	**0.855**

En la Tabla IV se observa que las redes NARMA ajustadas en este trabajo, para cada rezago, tienen ECMN más bajos en el conjunto de validación que sus correspondientes hallados por Burges y Refenes [15]; Para el caso de las redes ARNN, se tiene que ninguna de ellas produce (bajo el conjunto de validación) un ECMN inferior al mejor valor hallado por los autores.

En este segundo experimento, se evidencia nuevamente el problema de sobreparametrización que sufren las redes ARNN, lo cual conlleva a la inconsistencia observada entre los valores del ECMN hallados para los tres conjuntos de datos (véase la Tabla IV).

Siguiendo el criterio propuesto por Zemouri et al. [30], los mejores modelos son: ARNN (25) y NARMA (3). Nótese que existe una coherencia al seleccionar el mejor modelo usando la medida ECMN o la E(i) (obtenidas para los datos de prueba).

TABLA V. MEDIDAS DE RENDIMIENTO DE LOS MODELOS NARMAY ARNN.

Modelo	M1	M2	M3	M4	E(i)	std(i)
ARNN(10)	0.115	1.999	0.134	0.445	-0.0394	1.0708
ARNN(25)	0.0904	1.852	0.150	0.478	**0.00544**	**1.101**
ARNN(50)	0.129	1.607	0.0565	0.558	-0.0417	0.153
NARMA(1)	-0.170	2.004	0.0276	0.537	-0.0841	1.890
NARMA(2)	-0.218	1.912	0.202	0.527	-0.0672	1.865
NARMA(3)	0.254	1.211	0.248	0.584	**-0.0249**	**1.852**

No obstante, se evidencia que éstos modelos no tiene una buena capacidad predictiva, dado que en la Figura 8 las nubes de puntos distan mucho de la recta de 45º.

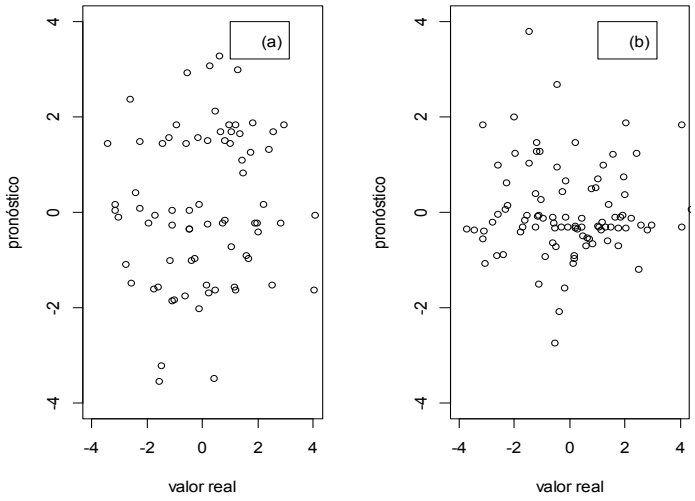

Figura 8. Comparación entre los datos de prueba y sus pronósticos hallados con la red (a) ARNN (25) y (b) NARMA (3).

VI. DISCUSIÓN

En esta sección se responden las preguntas de investigación planteadas.

A. ¿Un modelo no lineal AR de alto orden, representado por una red ARNN, puede aproximar bien un modelo no lineal MA de orden reducido?

Al examinar si la red ARNN con un orden alto para el rezago p, es capaz de aproximar correctamente un NLMA se encontró que si bien a medida que aumenta el número de rezagos p, el MSE de entrenamiento tiende a disminuir (como se mostró en la Figura 6) y las medidas de *E(i)* y *std(i)* presentan mejores resultados, este hecho no se ve reflejado en la capacidad de pronóstico del modelo (véase la gráfica (a) de la Figura 7).

Es de destacar que la capacidad del pronóstico no depende sólo del valor del rezago asumido, sino también del tamaño muestral y el porcentaje de datos usados para entrenar la red. Los mejores resultados de las redes ARNN se obtienen para los valores más grandes de rezagos acompañados de tamaños

muestrales grandes, de los cuales se use un gran porcentaje para entrenamiento. Sin embargo, hay que tener presente que este hecho conlleva a ajustar modelos no parsimoniosos y con problemas de sobreparametrización

Si adicionalmente a esto, se considera el hecho de que modelo NLMA no es globalmente invertible, entonces la respuesta a la pregunta es que un modelo no lineal autorregresivo (en este caso, aproximado por una red ARNN) de un alto orden no es capaz de representar un modelo no lineal de promedios móviles (NLMA) de bajo orden.

B. ¿Cuándo en una red recurrente NARMA se asume que no hay un proceso autorregresivo, se pueden pronosticar adecuadamente series de tiempo no lineales que contengan componentes inherentes de promedios móviles?

En las Figuras 7 y 8, así como en las Tabla II y V, se observa que si bien el modelo *NARMA* seleccionado, tiene un mejor desempeño (en cuanto a las medidas de rendimiento propuestas por Zemouri et al. [30] y acercamiento a la recta de 45°) que las otras redes evaluadas, los valores pronosticados por éste modelo distan mucho de los valores reales de la serie temporal no lineal con componente inherente MA. (véase las gráficas (b) de las Figuras 7 y 8).

Teniendo en cuenta este hecho, la respuesta es que una red recurrente NARMA $(0, q)$ no puede pronosticar adecuadamente series de tiempo no lineales que contengan componentes inherentes de promedios móviles.

Sin embargo, se destaca que en la experimentación se observó que al igual que ocurre con las expresiones matemáticas, en la práctica la red NARMA presenta un mejor acercamiento al modelo NLMA (desde el punto de vista de mejores medidas de capacidad de pronóstico) que la red ARNN. Lo cual nos indica que esta red puede ser una buena candidata para modelar datos no lineales que contengan componentes inherentes de promedios móviles, pero requiere sea estudiada detalladamente, y así surge una nueva pregunta de investigación: ¿Qué consideraciones, desde el punto de vista del planteamiento teórico, debe tener una red recurrente *NARMA (0, q)* para que pueda pronosticar adecuadamente series de tiempo no lineales que contengan componentes inherentes de promedios móviles?

VII. Conclusiones

Se muestra que tanto el modelo de red neuronal recurrente NARMA y el modelo de red neuronal autorregresivo ARNN, no son capaces de capturar en su totalidad el comportamiento de una serie de tiempo no lineal que contenga una componente inherente de promedios móviles (MA). Esto plantea la necesidad de formular un modelo de redes neuronales artificiales que permita pronosticar adecuadamente series de tiempo no lineales con componente inherente MA, el cual puede tener como punto de partida la red NARMA.

Referencias

[1] J. G. De Gooijer and R.J. Hyndman, "25 years of time series forecasting", *International Journal of Forecasting*, vol. 22, no.3, pp. 443-473, 2006.

[2] R. S.Tsay, *Analysis of Financial Time Series*. 3rd ed., Jhon Wiley & Sons, 2010.

[3] T. Teräsvirta, "Forecasting economic variables with nonlinear models", SSE/EFI Working Paper in Economics and Finance, Department of Economic Statistics, Stockholm School of Economics, 2005, pp. 598.

[4] R. F. Engle, "Autogressive conditional heteroskedasticity with estimates of the variance of uk inflation," *Econometrica*, vol. 50, pp. 987-1008, 1982.

[5] S. Lundbergh and T. Teräsvirta, *Forecasting with smooth autoregressive models*, in Clements, M., Hendry, D. (Eds.), A Companion to Economic Forecasting. Blackwell. Chap. 21, 2002.

[6] T. Teräsvirta, "Specification, estimation and evaluation of smooth transition autoregressive models," *Journal of the American Statistical Association*, vol. 89, no. 425, pp.208-218, 1994.

[7] M. Clements, P. H. Frances and N. R. Swanson, "Forecasting economic and financial time-series with non-linear models," *International Journal of Forecasting*, vol. 20, no. 2, pp.169-183, 2004.

[8] U. Anders and O. Korn, "Model selection in neural networks," *Neural Networks*, vol. 12, no. 2, pp.309–323, 1999.

[9] M. Paliwal and U.A. Kumar, "Neural networks and statistical techniques: A review of applications," *Expert Systems with Applications,* vol. 36, no. 1, pp. 2-17, 2009.

[10] M. Qi and G.P. Zhang, "An investigation of model selection criteria for neural network time series forecasting," *European Journal of Operational Reserach*, vol. 132, no. 1, pp. 666-680, 2001.

[11] T. Teräsvirta, C.F. Lin, and C.W.J. Granger, "Power of the neural network linearity test," *Journal Time Series Analysis*, vol. 14, no. 2, pp. 209-223, 1993.

[12] F. M. Tseng, H. C. Yu, and G.H. Tzeng, "Combining neural network model with seasonal time series arima model," *Technological Forecasting & Social Change*, vol. 69, no. 1, pp. 71-87, 2002.

[13] G.P. Zhang, B. Patuwo, and M. Hu, "Forecasting with artificial networks: The state of art.;" *International Journal of Forecasting*, vol. 14, no. 1, pp. 35-62, 1998.

[14] J. D. Velásquez and C. J. Franco, "Pronóstico de series de tiempo con tendencia y ciclo estacional usando el modelo airline y redes neuronales artificiales," *Revista Ingeniería y Ciencia*, vol. 8, no. 15, pp. 171-189, 2012.

[15] A. N. Burges and A.-P. N. Refenes, "Modelling non-linear moving average processes using neural networks with error feedback: An application to implied volatility forecasting," *Signal Processing*, vol. 74, no. 1, pp.89-99, 1999.

[16] J. T. Connor and R.D. Martin, "Recurrent neural networks and robust time series prediction," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 240-253, 1994.

[17] P. M. Robinson, "The estimation of a nonlinear moving average model," *Stochastic Processes and their Applications*, vol. 5, no. 1, pp. 81-90, 1977.

[18] W .E. Wecker, "Asymmetric time series," *Journal of the American Statistical Association*, vol. 76, no. 373, pp.16–21, 1981.

[19] P.M. Robinson and P. Zaffaroni, "Modelling nonlinearity and long memory in time series," *Fields Institute Communications*, vol. 11, pp. 161-170, 1997.

[20] R.F. Engle and A. Smith, "Stochastic permanent breaks," *The Review of Economics and Statistics*, vol. 81, no. 4, pp. 553-574, 1999.

[21] J. De Gooijer and K. Brannas, "Invertibility of non-linear time series models," *Communications in Statistics - Theory and Methods*, vol. 24, no. 11, pp. 2701-2714, 1995.

[22] K. Chan and H. Tong, "A note on the invertibility of nonlinear arma models," *Journal of Statistical Planning and Inference*, vol. 140, no. 12, pp. 3709-3714, 2010.

[23] S. Haykin. *Neural Networks a comprehensive foundation*. 2nd Ed. Prentice Hall International, 1999.

[24] R. Gencay and T. Liu, "Nonlinear modelling and prediction with feedforward and recurrent Networks," *Physica D*, vol. 108, no.1-2, pp. 119-134, 1997.

[25] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no.5, pp. 359-366, 1989.

[26] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, no. 5, pp. 551-560, 1990.

[27] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251-257, 1991.

[28] G. Dreyfus. *Neural Networks Methodology and Applications*. 2nd Ed. Springer- Verlag, 2005.

[29] G. P. Zhang, B. E. Patuwo, and M. Y. Hu, "A simulation study of artificial neural networks for nonlinear time series forecasting,", *Computers & Operations Research*, vol. 28, no. 4, pp. 381-396, 2001.

[30] R. Zemouri, R. Gouriveau, and N. Zerhouni, "Defining and applying prediction performance metrics on a recurrente NARX time series model," *Neurocomputing*, vol. 73,no.13-15, pp. 2506-2521, 2010.

Myladis Cogollo is an Assistant Professor at the School of Sciences, Universidad EAFIT, Medellín, Colombia. She received a M.Sc. in Science-Statistics in 2008 from the Universidad Nacional de Colombia. Currently, she is a PhD student in Systems Engineering at Universidad Nacional de Colombia; her current research interests are biostatistics, forecasting, artificial neural networks and nonlinear time series modeling.

Juan D. Velásquez rreceived the Bs. Eng in Civil Engineering in 1994, the MS degree in Systems Engineering in 1997, and the PhD degree in Energy Systems in 2009, all of them from the Universidad Nacional de Colombia. Medellin, Colombia. From 1994 to 1999, he worked for electricity utilities and consulting companies within the power sector and since 2000 for the Universidad Nacional de Colombia. Currently, he is a Professor in the Computing and Decision Sciences Department, Facultad de Minas, Universidad Nacional de Colombia. His research interests include: simulation, modeling and forecasting in energy markets; nonlinear time-series analysis and forecasting using statistical and computational intelligence techniques; and optimization using metaheuristics.