Vol. 9 - pp. 1-20
Junio - Noviembre 2005 - ISSN: 0329-5583
http://vaca.agro.uncor.edu/~revistasae/

MODELADO Y PRONÓSTICO DE UNA SERIE DE TIEMPO CONTAMINADA EMPLEANDO REDES NEURONALES Y PROCEDIMIENTOS ESTADÍSTICOS TRADICIONALES

Silvia Joekes[(1)]; Emanuel P. Barbosa[(2)]; Walter Robledo[(3)]

(1) Fac. de Ciencias Económicas - Univ. Nac. de Córdoba - joekess@eco.unc.edu.ar
(2) IMECC, Unicamp. Brasil - emanuel@ime.unicamp.br
(3) Fac. Ciencias Agropecuarias - Univ. Nac. de Córdoba - wrobledo@agro.uncor.edu.ar

RESUMEN

En este trabajo se presenta la aplicación de una red neuronal feedforward (AHA) a una serie de tiempo previamente estudiada en la literatura por diferentes investigadores (Brubacher, 1974; Martín, 1980; Stahlbut, 1985; Allende, 1989) llamada serie RESEX que presenta estacionalidad y valores extremos (*outliers*). Algunos elementos de la arquitectura de la red tales como la definición de las variables de entrada fueron sugeridos por el análisis previo de la serie y otros como el número de capas ocultas y de neuronas por capa surgieron a través de procedimientos numéricos. Para el entrenamiento de la red se utilizó el algoritmo de retropropagación (*backpropagation*) estándar y también un algoritmo de aprendizaje robusto para tratar adecuadamente con los valores extremos aditivos. Todos los modelos y métodos (tradicionales y de redes neuronales) fueron comparados considerando diferentes medidas de habilidad predictiva. En general, los resultados finales no fueron muy diferentes. El modelo estadístico mostró un menor error residual y menor porcentaje de error absoluto en el ajuste, mientras que la red robustamente entrenada presentó mejores pronósticos.

PALABRAS CLAVE: series de tiempo, redes neuronales feedforward, retropropagación, estimación robusta, predicción.

1. INTRODUCCIÓN

Las redes neuronales artificiales (RNA) fueron desarrolladas a partir de los trabajos de investigación sobre la fisiología de las neuronas biológicas (NB) de Erlonger y Gasser en 1924, y de los estudios sobre la actividad de los neurotransmisores de Hadgkin y Huxley en 1956 (Kovács, 1996).

Una RNA simula una RNB (RN Biológica) en la que cada nodo (unidad de procesamiento) se corresponde con una neurona y tiene como función recibir varios impulsos (entradas), procesarlos y transmitir un resultado a otro u otros nodos (salidas).

La implementación de las RNA como modelo matemático para las RNB fue desarrollado por McCulloch y Pitts (1943). Sin embargo, como entidad matemática, ellas tienen un interés intrínseco y su importancia y aplicación se ha extendido más allá de lo originalmente imaginado (Kovács, 1996).

Las RN han resultado apropiadas para el análisis de datos generados en una amplia variedad de disciplinas. En este sentido, las redes no sólo han sido empleadas de forma innovativa e imaginativa para analizar grandes y complejas bases de datos, sino que también han sido utilizadas para resolver problemas tradicionalmente ligados al análisis estadístico (Blough y Anderson, 1984).

La aplicación de las RN a la predicción con series de tiempo no es nueva. Existen numerosos trabajos al respecto, siendo probablemente los más conocidos los de Werbos (1974, 1988), Lapedes (1987), Weigend *et al.* (1990), entre otros (Weigend y Gershenfeld, 1994).

Desde la estadística, Box y Jenkins (1976) desarrollaron la metodología de los modelos autorregresivos integrados de promedios móviles (ARIMA) para ajustar una clase de modelos lineales para series de tiempo. Posteriormente surgieron versiones robustas de modelos ARIMA y de series de tiempo no lineales (Allende y Moranga, 2000) tendientes a resolver los problemas que introducen la presencia de valores aberrantes o extremos en los datos.

Más recientemente, las RN han sido consideradas como una alternativa para modelar series de tiempo no lineales.

Los modelos de RN se ajustan tradicionalmente por mínimos cuadrados y por lo tanto carecen de robustez en presencia de valores extremos o aberrantes (*outliers*). Como algunos de los procedimientos que tratan con RN surgen como una generalización natural de los modelos estadísticos lineales AR y ARMA al caso no lineal NAR y NARMA, los procedimientos para ajustar RN robustas suelen estar relacionados con los procedimientos empleados para modelar series de tiempo robustas (Connor *et al.*, 1994). El algoritmo se basa el filtrado de los valores extremos de los datos, estimando luego los parámetros del modelo con los datos filtrados.

El objetivo principal de este trabajo consistió en proponer métodos alternativos, basados en RN, para el análisis de series de tiempo con valores extremos. La aplicación de esta metodología a una serie de tiempo con datos reales (conocida como serie RESEX), posibilitó la comparación de la habilidad predictiva de ambos modelos: Estadístico y de RN.

Se pretendió demostrar efectivamente que ambas metodologías no son sustitutivas sino más bien complementarias.

2. MODELOS AUTOREGRESIVOS NO LINEALES Y REDES NEURONALES

Una generalización natural del modelo lineal AR(p) al caso no lineal podría ser el siguiente modelo autoregresivo no linear (NAR):

$$x_t = h(x_{t-1}, x_{t-2}, \ldots, x_{t-p}) + e_t \quad , \qquad 2.1$$

donde *h* es una función de suavizado desconocida.

Se supone que $E(e_t / x_{t-1}, x_{t-2}, \ldots) = 0$ e_t y que σ^2 tiene varianza fini

Bajo estas condiciones el predictor que minimiza el error cuadrático medio de x_t dad $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ es la media condicional

$$\hat{x}_t = E\left[x_t / x_{t-1}, \ldots, x_{t-p}\right] = h\left[x_{t-1}, \ldots, x_{t-p}\right] \quad t \geq p+1 \qquad 2.2$$

Las redes neuronales feedforward (o redes que avanzan hacia delante AHA), fueron propuestas como un modelo NAR para la predicción con series de tiempo por Lapedes y Farber (Connor *et al.*, 1994). Una red AHA se define como la aproximación no lineal de *h* dada por:

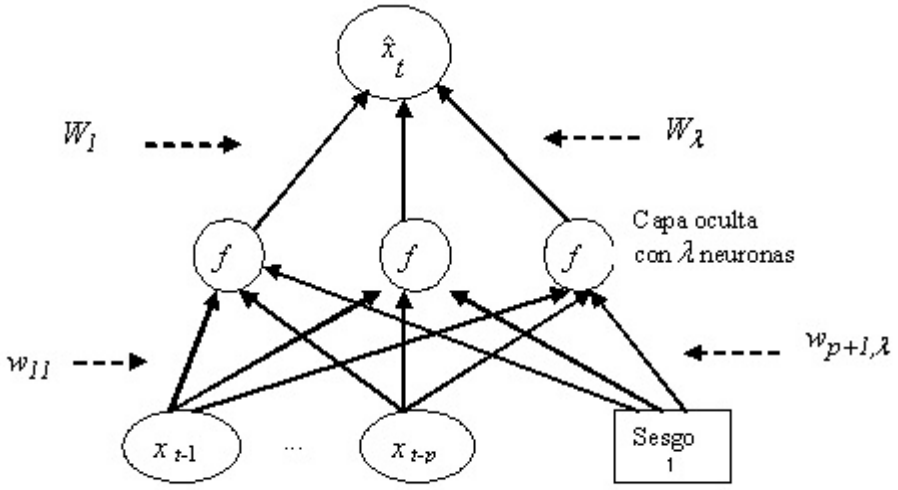
$$\hat{x}_t = h(x_{t-1}, \ldots, x_{t-p}) = \sum_{j=1}^{\lambda} W_j \; f\left(\sum_{i=1}^{p} w_{ij} \, x_{t-i} + w_{p+1,j}\right) , \qquad 2.3$$

donde la función $f(x)$ es una función de suavizado monótona de tipo sigmoide y la función de activación de la salida es supuesta la identidad. w_{ij} representa los parámetros de las conexiones entre las entradas y las neuronas de la capa oculta, W_j representa los parámetros de las conexiones entre las neuronas ocultas y la salida y λ es el número de neuronas en la capa oculta. Los parámetrcW_j w_{ij} y son estimados a partir de una muestra $x_1^0, \ldots, x_N^0$, permitiendo obtener una es$\hat{h}$nación de *h*. Las estimaciones se obtienen minimizando la suma de cuadrados residuales $\sum (x_t - \hat{x}_t)^2$, la que al ser no lineal en sus parámetros se puede minimizar mediante un procedimiento numérico como el del gradiente descendente, conocido también como retro-propagación (*backpropagation*) o por un método de segundo orden (Rumerhalt *et al.*, 1986; Allende, 2000).

Existen redes más complejas. En general se trata de evitar el uso de redes de múltiple capa a causa de que no dan mayores ventajas para problemas de series de tiempo simples. En situaciones más complejas como en problemas de reconocimiento de voz, se utilizan redes con varias capas como las redes de tiempo retrasado (*time-delay*) (TDNN).

Sin embargo, no importa cuan compleja sea la arquitectura de una red AHA, esta es siempre un miembro de una clase de modelos no lineales como el descrito en 2.3 con algún valor finito de *p*.

Figura 1: Red Neuronal AHA para modelos NAR(p) con λ neuronas en la capa oculta.



Para el caso de los modelos lineales ARMA, una generalización natural al caso no lineal está dada por:

$$x_t = h\left(x_{t-1}, x_{t-2}, \ldots, x_{t-p}, e_{t-1}, e_{t-2}, \ldots, e_{t-q}\right) + e_t$$

donde h es una función de suavizado desconocida.

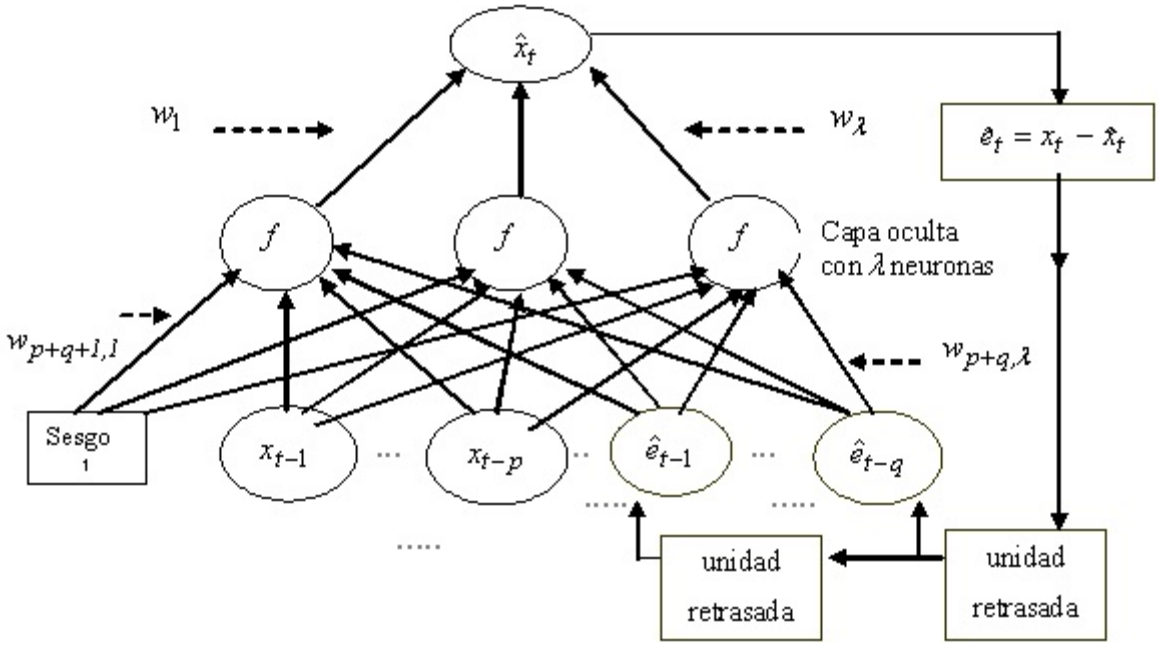
Este modelo puede ser aproximado con el siguiente modelo de RN recurrente NARMA(p, q),

$$\hat{x}_t = \sum_{j=1}^{\lambda} W_j \; f\left(\sum_{i=1}^{p} w_{ij}\, x_{t-i} + \sum_{i=p+1}^{p+q} w_{ij}^{T}\, \hat{e}_{t+p-i} + w_{p+q+1,j}\right), \qquad 2.5$$

donde: $\hat{e}_{t+p-i} = x_{t+p-i} - \hat{x}_{t+p-i}$

Los parámetros W_j w_{ij} w_{ij}^T y son estimados por mínimos cuadrados, al igual que en el caso de redes AHA, es decir, eligiendo aquellos parámetros que minimicen la suma de cuadrados de error $\sum (x_t - \hat{x}_t)^2$.

Figura 2: Red Neuronal recurrente para modelos NARMA (p,q) con $\hat{\lambda}$ neuronas en la capa oculta.



Las redes recurrentes como la dada en (2.5) pueden ser vistas como un caso especial del algoritmo de backpropagation a través del tiempo.

La red recurrente NARMA (2.5) es un caso especial de alguna red recurrente más general. Para esta red:

$$\hat{x}_t = \sum_{j=1}^{\lambda} W_j \, g_j(t) + w_{\lambda+1} \quad , \qquad 2.6$$

donde las λ unidades oculta$g_j(t)$ son computadas recursivamente en el tiempo de la siguiente manera:

$$g_j(t) = f\left(\sum_{i=1}^{\max(p,q)} \tilde{w}_{ij} \, x_{t-i} + \sum_{k=1}^{q} \sum_{l=1}^{\lambda} \tilde{w}_{kjl} \, g_l(t-k) + \tilde{w}_{\max(p,q)+1,j} \right), \qquad 2.7$$

siendo las ponderaciones $\tilde{w}_{kjl}$ son distintas de las ponderacio$\tilde{w}_{ij}$.

3. MODELOS DE PREDICCIÓN ROBUSTA PARA REDES NEURONALES

En esta sección se presenta un procedimiento robusto para el ajuste de redes AHA y recurrentes de tipo NAR y NARMA respectivamente. como una extensión de los procedimientos robustos de modelos ARMA. El método consiste en un procedimiento de filtrado robusto por interpolación del tiempo de ocurrencia de los valores extremos, y está relacionado con un procedimiento de estimación de máxima verosimilitud no Gaussiano.

Supongamos que se parte del siguiente modelo NARMA (*p, q*);

$$x_t = f(x_{t-1},\ldots,x_{t-p},e_{t-1},\ldots,e_{t-q}) + e_t \quad , \quad 3.1$$

el cual puede ser expresado como un vector de dimensión (*p*+*q*). Se define a los vectores columna de dimensión (*p*+*q*), $\mathbf{x}_t$ $\mathbf{f}(\mathbf{x}_{t-1})$ $\mathbf{e}_t$ y del siguiente modo:

$$\mathbf{x}_t = (x_t,..,x_{t-p+1},\ e_t,\ldots,e_{t-q+1})^T \quad , \quad 3.2$$

$$\mathbf{f}(\mathbf{x}_{t-1}) \quad \left[f(x_{t-1}).x_{t-1},..,x_{t-p+1},\ 0,\ e_{t-1},..,e_{t-q+1}\right]^T$$

$$\mathbf{e}_t \quad (e_t,0,\ldots,0,\ e_t,\ 0,\ldots,0)^T \quad . \quad 3.4$$

Considerando ahora a la función f como una función no lineal, se puede expresar al modelo NARMA con valores extremos aditivos del siguiente modo:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}) + \mathbf{e}_t \quad , \quad 3.5$$

$$y_t \quad \mathbf{h}_t^T \mathbf{x}_t \quad + \quad 3.6$$

donde v_t representa valores extremos aditivos $\mathbf{h}$ = (1, 0, 0,...., 0)^T es una vector columna de dimensión (*p*+*q*).

Sea $\hat{\mathbf{x}}_t$ la estimación robusta filtrada de $\mathbf{x}_t$, basado s$(y_1,\ldots,y_t)$ $\hat{\mathbf{x}}_t^{t-1}$ y sea el predictor robusto del paso siguiente de $\hat{\mathbf{x}}_t$ basado sobre las observaciony_1 hasta y_{t-1}.

Similarmente, sean $\hat{x}_t^{t-1}$ $\hat{y}_t^{t-1}$ los predictores del paso siguiex_t e y_t e respectivamente, basados sobre las observaciones y_1 hasy_{t-1}.

La forma del filtro robusto para este modelo consiste en simular el llamado filtro de Kalman "extendido" para modelos de espacio de estado no lineales, mediante el uso de la linealización de $f(\mathbf{x}_{t-1})$ con las estimaciones robustamente filtradas de $\hat{\mathbf{x}}_{t-1}$.

$$\mathbf{f}(\mathbf{x}_{t-1}) \approx \mathbf{f}(\hat{\mathbf{x}}_{t-1}) + \mathbf{Df}(\hat{\mathbf{x}}_{t-1})(\mathbf{x}_{t-1} - \hat{\mathbf{x}}_{t-1}) \quad ,$$

donde $\mathbf{Df}(\mathbf{x}_{t-1})$ es una matriz de dimensión (*p*+*q*) x (*p*+*q*) de las derivadas parciales de $\mathbf{f}$ evaluadas e$\hat{\mathbf{x}}_{t-1}$.

La predicción del paso siguiente del modelo NARMA (*p,q*) está dado por:

$$\hat{\mathbf{x}}_t^{t-1} \quad \mathbf{f}(\hat{\mathbf{x}}_{t-1}) \quad 3.8$$

$$\hat{y}_t^{t-1} \quad \mathbf{h}^T \mathbf{f}(\hat{\mathbf{x}}_{t-1}) \quad f(\hat{\mathbf{x}}_{t-1}) \quad = \quad 3.9$$

donde la estimación en el tiempo $t-1$ está dada por $\hat{\mathbf{x}}_{t-1}$. Nótese que (3.9) es el primer elemento del vector en (3.8).

El filtro robusto recursivo para la estimación de $\hat{\mathbf{x}}_t$ está dado por;

$$\hat{\mathbf{x}}_t = \mathbf{f}(\hat{\mathbf{x}}_{t-1}) \quad \frac{M_t \boldsymbol{h}\, s_t}{s_t^2}\ \Psi\left(\frac{y_t - \hat{y}_t^{t-1}}{s_t}\right) \qquad 3.1$$

donde la función de robustificación Ψ es la función bi-parte redescendente de Hampel. ψ_{HA}, $\psi_{a,m}(r) = r \quad 0 < |r| < a \quad \psi_{a,m}(r) = 0 \quad |r| \geq m$ y

M_t es la covarianza de la predicción, tal qM_{t+1} se presenta en (3.11).

Notemos que ahora, el residual robustamente estimado está dado por:

$$s_t\ \Psi\left(\frac{y_t - \hat{y}_t^{t-1}}{s_t}\right),$$

La matriz de covarianza de la predicción del paso siguiente está dada por:

$$M_{t+1} = Df(\hat{\mathbf{x}}_t)\, P_t\, Df^T(\hat{\mathbf{x}}_t) + Q \qquad , \qquad 3.11$$

siendo Q la matriz de covarianza de $\boldsymbol{e}_t$ P_t la matriz de covarianza condicional de ($\hat{\mathbf{x}}_t \,|r_t\ \mathbf{y_{t-1}}$,) dada por:

$$P_t = M_t - M_t \boldsymbol{h}\ (s_t^2)^{-1} w\left(\frac{y_t - \hat{y}_t^{t-1}}{s_t}\right)\ \boldsymbol{h}^T M_t$$

donde al igual que antes; $w(r) = \Psi(r)/r \quad s_t^2 = (M_t)_{11}$ y

El uso de $\Psi = \Psi_{HA}$ determina el siguiente comportamiento para el filtro robusto:

a) Cuando $|y_t|$ es suficientemente grande debido a la presencia de un valor extremo aditivo, es decir, cuando $\left|y_t - \hat{y}_t^{t-1}\right| > ms_t$ $\hat{x}_t$, entonces, es la p ción de x_t dada por:

$$\hat{x}_t = f(\hat{\mathbf{x}}_{t-1}).$$

b) Cuando la magnitud de $\left|y_t - \hat{y}_t^{t-1}\right|$ es pequeña, es decir, cuando $\left|y_t - \hat{y}_t^{t-1}\right| < as_t$ $\hat{x}_t$, entonces, es l$x_t$predicción de dada por:

$$\hat{x}_t = (\hat{\mathbf{x}}_t)_1 = y_t.$$

Los resultados de Martin (1980) y Masreliez (1975) sugieren que si Ψ es la función score para la densidad de predicción de las observaciones $P(y_t \mid \mathbf{y_{t-1}})$ entonces $\hat{\mathbf{x}}_t$ $\hat{\mathbf{x}}_t^{t-1}$ son estimaciones aproximadas de medias condicionales. Esto determina una buena racionalización para la recursión del filtro robusto. (Ver Anexo III, Joekes S. 2002)

Desde otro punto de vista, la recursión anterior representa una versión más robusta del filtro de Kalman extendido para modelos no lineales por analogía con la versión robusta del filtro de Kalman para modelo lineales.

4. RED NEURONAL ROBUSTA

En esta sección se traslada el concepto de filtro recursivo robusto como ha sido definido en la sección anterior, para ser aplicado al caso de RN.

Supongamos ahora que f se aproxima mediante una red neuronal recurrente $g(\hat{\mathbf{x}}_{t-1})$ y consideremos el conjunto de parámetros de la $\boldsymbol{\delta}$ d por ;

$$\boldsymbol{\delta} = \left(W^T,\ w_1^T, \ldots, w_\lambda^T,\ w_1'^T, \ldots, w_\lambda'^T,\ w'_{p+q+1} \right)$$

Dada la dependencia de $\boldsymbol{g}$ sobre $\boldsymbol{\delta}$ se puede escribi $g(\hat{\mathbf{x}}_{t-1}) = g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})$

Un estimador de máxima verosimilitud aproximado (M-estimador) de $\hat{\delta}$ como fue desarrollado por Martín (1980) y Martín y Yohai, (1985) para el caso de modelos ARMA, está dado por:

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{n} \rho \left(\frac{y_t - \hat{y}_t^{t-1}(\boldsymbol{\delta})}{s_t} \right) \right\} \quad . \quad 4.2$$

Donde la función de robustificación que se usa es = $_{\mathrm{H}}$ (Hampel)

El estimador s_t es un estimador de escala altamente robusto (MADM) para la predicción de los residuales:

$$r_t(\boldsymbol{\delta}) = y_t - \hat{y}_t^{t-1}(\boldsymbol{\delta}) \qquad 4.3$$

Aunque la estimación dada en (4.2) fue originalmente desarrollada para la estimación robusta de modelos lineales AR y ARMA con valores extremos aditivos, su aplicación al ajuste robusto de redes AHA y recurrentes mediante modelos NAR y NARMA es algo natural.

Por lo tanto, el modelo predictor recurrente toma ahora la forma;

$$\hat{y}_t^{t-1}(\boldsymbol{\delta}) = g(\hat{\mathbf{x}}_{t-1},\ \boldsymbol{\delta}) = \sum_{j=1}^{\lambda} W_j\, f \left(\sum_{i=1}^{p} w_{ij}\ \hat{x}_{t-i} + \sum_{i=1}^{q} w_{ij}^T\ \hat{e}_{t-i} + \ w_{p+q+1,j} \right)$$

Para resolver el problema de optimización (4.2) es necesario recurrir a un procedimiento iterativo. Supongamos por el momento que x_t s_t no dependen de $\boldsymbol{\delta}$.

Diferenciando el lado derecho de (4.2) con respecto a $\boldsymbol{\delta}$ da la ecuación de estimación:

$$\sum_{t=1}^{n} \mathrm{D}\mathbf{g}(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta}) \; \Psi\left(\frac{y_t - g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})}{s_t}\right) = 0$$

Consideremos ahora el primer elemento del vector de recursión (3.10) con $y_t^{t-1} = f(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})$ re$g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})$ o por , es decir:

$$\hat{x}_t = g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta}) + s_t \; \Psi\left(\frac{y_t - g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})}{s_t}\right)$$

Multiplicando ambos miembros por $\mathrm{D}\mathbf{g}(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})$ y sumando, se obtiene:

$$\sum_{t=1}^{n} \mathrm{D}\mathbf{g}(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta}) \; s_t \; \Psi\left(\frac{y_t - g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})}{s_t}\right) =$$

$$\sum_{t=1}^{n} \mathrm{D}\mathbf{g}(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta}) \; \left[\hat{x}_t - g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})\right] .$$

El lado izquierdo de la ecuación, debería ser equivalente a (4.5) si s_t tomara un valor constante, es decir, $s_t = \hat{s}$. En ef$s_t$to, ser$\hat{s}$igual la estimación robusta de σ_e para la mayoría de las observaciones cuany_t esté libre de valores extremos. Entonces, es razonable suponer que el lado izquierdo de (4.7) es una buena aproximación de (4.5) por un multiplicador constante.

El lado derecho de (4.7) es la ecuación de estimación para la red ajustada por mínimos cuadrados basada sobre datos robustamente filtrados. Esto es, el estimador $\hat{\boldsymbol{\delta}}$ obtenido resolviendo el lado derecho de la ecuación (4.7) está dado por:

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} = \sum_{t=1}^{n} \left[\hat{x}_t - g(\hat{\mathbf{x}}_{t-1}, \boldsymbol{\delta})\right]^2 \qquad 4.8$$

En el argumento que conduce a (4.8) se ha tratado a $\hat{x}_t$ y s_t como fijos e independientes de $\boldsymbol{\delta}$. Esto no es siempre cierto, debido a que el filtro robusto depende de $\boldsymbol{\delta}$. Sin embargo, muchos valores d$\hat{x}_t$ son igualy_ta y tales valores son esencialmente independientes de $\boldsymbol{\delta}$, como lo son los correspondientes valores de s_t .

Por lo tanto (4.8) debería conducir a una buena aproximación de (4.5) al menos cuando la fracción de valores extremos no es muy grande.

5. MODELADO Y PRONÓSTICO DE SERIES DE TIEMPO CON REDES NEURONALES: UN ESTUDIO COMPARATIVO EMPLEANDO LA SERIE RESEX.

5.1. Introducción

En este apartado, se considera el modelado y pronóstico mediante una RN AHA (Feedforward) de una serie de tiempo previamente estudiada en la literatura, llamada serie RESEX, que presenta estacionalidad y valores extremos (outliers). La serie está referida al número de conexiones mensuales de extensiones telefónicas residenciales en cierta área de Canadá. Además, ha sido considerada en la literatura, como una referencia para evaluar procedimientos de estimación robusta aplicados a series de tiempo a causa de sus prominentes valores extremos aditivos.

Algunos autores han analizado estos datos considerando diferentes procedimientos para tratar con los valores extremos, como por ejemplo pre-filtrando los datos extremos antes de emplear otros procedimientos estándar (Brubacher, 1974; Martin, 1980; Stahlbut, 1985; Allende, 1989). Los análisis previos de los datos de la serie RESEX incluyen la identificación de un modelo estacional auto-regresivo de segundo orden AR(2) luego de su diferenciación estacional (Brubacher, 1974). Los resultados del análisis de la serie RESEX usando un modelo estacional AR(2) y diferentes procedimientos de corrección de valores extremos o métodos robustos, fueron presentados por Allende (1989).

En este trabajo se propone un procedimiento alternativo para el modelado y pronóstico de los datos de la serie RESEX. Este procedimiento trata con los valores extremos de una manera diferente mediante la aplicación de una AHA que presenta algunas características ventajosas e interesantes.

En primer lugar, se considera la arquitectura de la red (con 3 variables retrasadas como entrada y una capa oculta intermedia), la cual puede ser interpretada como una extensión no-lineal del modelo estacional AR(2) previo. Esto es, la AHA considerada puede ser vista como una forma de representar e implementar un modelo estacional tipo NAR (*Non-linear Auto-Regresive*) para la serie, generalizando los modelos previamente propuestos. Por otro lado, como una RN puede ser vista como una "función de aproximación universal", de acuerdo al teorema de Kolmogorov-Nielsen (Bishop, 1995), permite aproximar la función no-lineal desconocida asociada al modelo NAR.

En segundo lugar, las RN y particularmente las consideradas en este trabajo, son más resistentes a la presencia de valores extremos en los datos que los métodos tradicionales empleados en el análisis de series de tiempo, aún cuando la red sea estimada o entrenada mediante procedimientos usuales, tales como el algoritmo de retropropagación (*backpropagation*) basado en métodos de gradiente.

Esto se muestra en este trabajo y está relacionado con el hecho de que las RN constituyen un modelo en dos estados, con muchos parámetros y una gran flexibili-

dad, restringiendo la influencia de los valores extremos a un efecto local. En efecto, los parámetros o ponderaciones de la red son estimados aquí considerando dos algoritmos diferentes, el algoritmo estándar de retro-propagación (Haykin, 1999) y un algoritmo de aprendizaje alternativo, aún más robusto, basado sobre un procedimiento iterativo de filtrado robusto (Connor, 1993).

5.2. La serie RESEX y sus análisis preliminares

5.2.1. Descripción de los datos y preparación inicial

La serie RESEX está referida al número de conexiones mensuales de extensiones telefónicas residenciales en una cierta área de Canadá, entre enero de 1966 y mayo de 1973, con un total de 89 observaciones. Las características más destacadas de esta serie son, su estacionalidad anual y la presencia de un par de valores extremos (extremadamente grandes) cercanos al final de la serie. Como se puede ver en el gráfico de la serie de tiempo presentado en la Figura 3, los valores extremos corresponden a los meses de noviembre y diciembre de 1972. La razón de estos valores fue una promoción especial (sin costo) ofrecida por la compañía telefónica en esos meses.

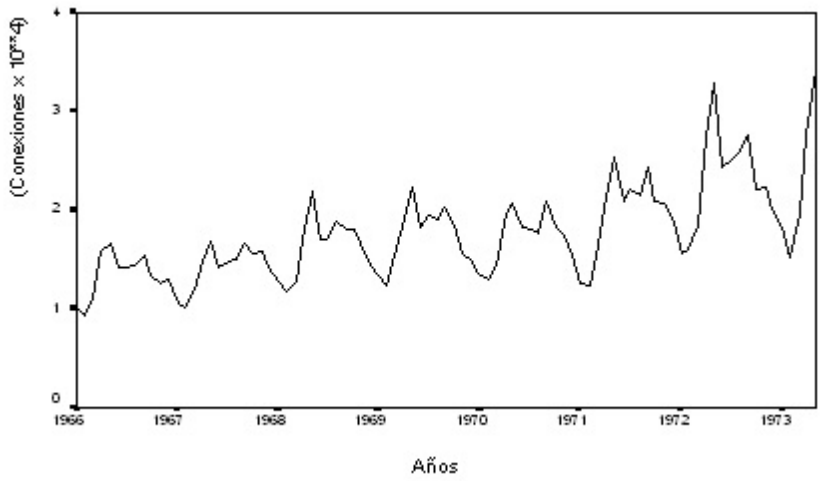
Figura 3: Serie RESEX

En la figura anterior se observa claramente la estacionalidad con períodos de un año o 12 meses. Otra característica sugerida por el gráfico podría ser una eventual y leve componente de tendencia. No obstante, al comprobar la hipótesis de estacionariedad versus la presencia de raíz unitaria o tendencia, empleando el test de Dickey & Fuller (1979, 1981) luego de la diferenciación estacional, no se detectó evidencia (a un nivel del 5% de significación) de la presencia de una componente de tendencia. Por lo tanto, luego de la diferenciación estacional, la serie puede ser

considerada como estacionaria, (para detalles, ver Joekes, S. 2002) resultado que acuerda con los trabajos previos, como por ejemplo, Brubacher (1974) y otros.

En orden a identificar un modelo estacional para los datos, se recomienda un filtrado preliminar e interpolación de los valores extremos de manera de evitar una incorrecta especificación del modelo. En este sentido, los valores extremos de la serie fueron corregidos empleando el filtro de Kalman, logrando de esta manera una versión libre de valores extremos para la serie RESEX, como se muestra en la figura siguiente.

Figura 4: Serie RESEX usando filtro de Kalman



Como puede observarse en la figura anterior, el patrón estacional no es estable o estacionario, sugiriendo la aplicación de una diferenciación estacional.

Luego de esta preparación inicial de los datos (corrección de valores extremos y diferenciación estacional) se logra un modelo de serie de tiempo estacionario para la serie RESEX corregida.

5.2.2. Identificación del modelo

Un análisis de autocorrelación estándar mediante correlograma simple y parcial, efectuado a la serie RESEX (libre de valores extremos) diferenciada estacionalmente, sugieren un modelo AR(2) (auto-regresivo de orden 2) para los datos. Empleando notación de Box-Jenkins para modelos ARIMA estacionales, se identifica un modelo ARIMA(2,0,0)(0,1,0) para la serie, resultado que acuerda con los trabajos previos (Brubacher, 1974 y otros).

Los parámetros estimados (ML) y sus correspondientes errores estándar, acuerdan con los resultados previos de la mencionada literatura. Los residuales del modelo pueden ser considerados satisfactorios (auto-correlación residual no significativa) confirmando la identificación del modelo.

El modelo AR(2) estacional estándar es considerado aquí como una referencia para establecer la performance predictiva de los procedimientos de RN propuestos. En este sentido, el modelo de referencia es ajustado a los datos de los primeros 7 años (84 observaciones), dejando los últimos 5 meses para propósitos pronósticos. Los resultados pronósticos (un paso adelante y multi paso) han sido evaluados mediante el RSE (Residual Estándar Error) y el MAPE (*Mean Absolute Percentage Error*) y se presentan en la sección 5.5 junto con los resultados de la RN.

5.3. Redes Neuronales Feedforward para Series de Tiempo

5.3.1. Elementos básicos

En su forma más simple, una AHA relaciona una respuesta o variable de salida (output) con n predictores o variables de entrada (input) mediante una relación no-lineal representada como una composición a dos niveles de relaciones lineales generalizadas GLR's (Una GLR es una transformación no-lineal aplicada a una combinación lineal de sus entradas o inputs). Esto es, la salida es un GLR de variables intermedias, donde cada una, a su vez, es un GLR de las variables de entrada (Haykin, 1999).

En el contexto de las series de tiempo, es común relacionar una serie dada con su pasado (*lags*) mediante un proceso auto-regresivo lineal, que puede ser extendido a una forma no-lineal (modelo NAR) mediante una implementación vía AHA como se presentó en la Figura 1 de la sección 2.

La especificación de la arquitectura de la red implica, la definición de las variables de entrada (número p de retrasos u orden del modelo NAR), el número de GLR's o nodos en la capa intermedia y la transformación no-lineal aplicada en la capa intermedia (llamada función link o función de activación, generalmente una función sigmoidea). La activación de la salida, usualmente empleada en series de tiempo, es la función identidad. Una vez que la arquitectura de la red está completamente definida, el objetivo consiste en estimar las ponderaciones a partir de los datos.

5.3.2. Algoritmos de entrenamiento

Los procedimientos de aprendizaje iterativo más comúnmente empleados para estimar los parámetros o ponderaciones de una RN, son conocidos como algoritmos de retro-propagación (Bishop, 1995; Haykin, 1999) y están basados en procedimientos de gradiente, principalmente el gradiente descendente (*steepest descent*) o el gradiente conjugado (*conjugate gradient*).

Entre estos procedimientos, se ha adoptado para la implementación de la RN el algoritmo de retro-propagación basado en el gradiente conjugado. Esto se debe a que este método tiene ventajas conocidas en relación con su versión más simple (Haykin, 1999) pero tiene un costo computacional adicional al calcular las matrices Hessianas.

Un aspecto práctico importante de la implementación de estos algoritmos es el criterio de parada adecuado para evitar el sobre-entrenamiento. En este trabajo se considera el llamado criterio de "parada temprana" (*early stopping*) que consiste en detener el entrenamiento cuando el error de validación, monitoreado *on-line*, logra un mínimo.

5.3.3. Implementación y arquitectura de la red

Uno de los primeros elementos a especificar en la red son las variables de entrada. Mediante la información previa de que el modelo que ajusta a los datos de la serie RESEX es un modelo AR(2), estas variables fueron definidas como versiones retrasadas (retrasos 1, 2 y 12) de la serie original.

Un segundo elemento es la especificación del número de capas intermedias (una o dos) y sus correspondientes nodos (generalmente entre uno y cuatro). Se consideraron un total de 9 casos posibles, o combinaciones de estos valores, de manera de elegir el que determine la mejor RN para ajuste y pronóstico. Todos los modelos fueron implementados empleando el software SPSS (*Statistical Package for the Social Science*) y su módulo para RN.

El total de datos de la serie fue dividido en tres partes: un conjunto de entrenamiento (72 observaciones), un conjunto de validación (12 observaciones) y un conjunto de test o pronóstico (últimas 5 observaciones o meses).

Los resultados del ajuste y pronóstico de cada posible arquitectura son presentados junto con ciertas medidas de error, tales como el RSE y el MAPE, como se observa en la Tabla 1, a continuación.

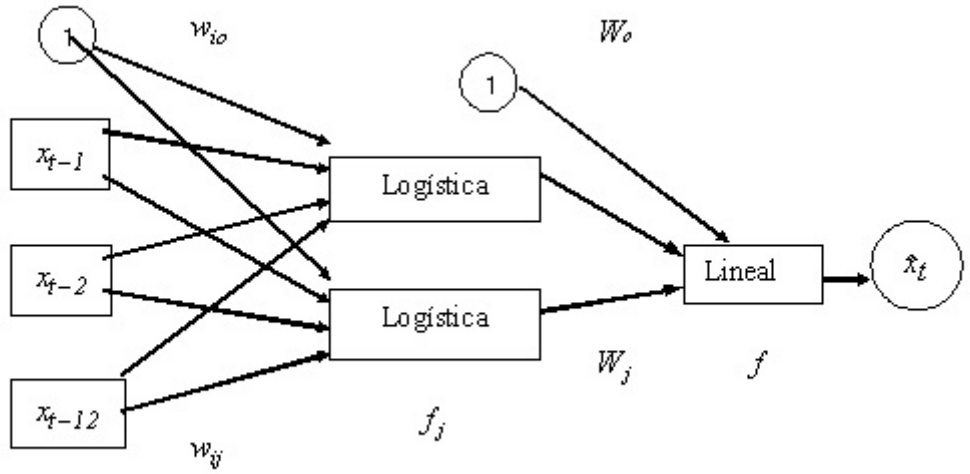
Tabla Nº 1: Comparación de varios modelos AHA (los valores de RSE están expresados en unidades de 10^3 y el MAPE en porcentajes)

Nº de capas ocultas	Nº de neuronas por capa	Nº de parámetros	Ajuste				Pronóstico un paso adelante	
			entrenamiento		validación			
			RSE	MAPE	RSE	MAPE	RSE	MAPE
1	1	6	1.66	7.28	1.55	7.58	1.44	4.75
1	2	11	1.60	7.16	1.56	7.56	1.20	4.14
1	3	16	1.67	7.31	1.54	7.49	1.73	5.96
1	4	21	1.49	6.46	1.53	6.57	9.94	36.67
2	1-1	8	1.82	8.22	1.59	8.17	3.12	10.70
2	2-2	17	1.86	8.14	1.31	6.00	4.30	15.13
2	2-3	21	1.71	7.82	1.57	7.92	2.22	8.76
2	3-2	23	1.49	6.65	1.58	7.23	7.34	20.86
2	3-3	28	1.69	7.51	1.55	7.58	2.56	7.82

De la tabla anterior se puede determinar que el modelo de red con una capa intermedia (capa oculta) y con dos nodos o neuronas en esta capa, es el que determina las mejores medidas predictivas y el mejor ajuste, por lo que esta arquitectura

es la que se adopta para los datos de la serie RESEX. Esta arquitectura, con 3 nodos de entrada, 2 nodos en la capa intermedia y 1 nodo de salida se muestra en la figura 5, a continuación.

Figura Nº 5: Arquitectura de la red propuesta para la serie RESEX.



La ecuación de predicción para calcular una estimación con salida unidimensional, sobre la base de un conjunto de observaciones pasadas convenientemente seleccionadas como entradas, se puede expresar como:

$$\hat{x}_t = f\left\{W_o + \sum_j W_j f_j\left(w_{io} + \sum_i w_{ij}\ x_{t-i}\right)\right\}$$

donde w_{io} denota los parámetros para la conexión entre la entrada constante (sesgo) y las neuronas de la capa oculta y W_o denota la ponderación de la conexión externa (sesgo) entre una entrada constante y la salida o pronóstico. Las ponderaciones w_{ij} y W_j son las ponderaciones para las otras conexiones, entre las entradas y las neuronas ocultas y entre las neuronas ocultas y la salida, respectivamente. Las funciones f_j f son las funciones de activación utilizadas en la capa oculta y en la salida, respectivamente. Se debe notar que los niveles de las neuronas ocultas pueden ser intercambiados sin modificar el modelo.

Si se incorporan los sesgos en las sumas, se obtiene la ecuación (2.3) para una red AHA.

5.4. Red neuronal robusta de entrenamiento

En esta Sección, se traslada el concepto de filtro recursivo robusto desarrollado en la Sección 4 para ser aplicado a la red AHA ajustada a la serie RESEX.

El predictor robusto del paso siguiente de y_t usando valores robustamente filtrados está dado por:

$$\hat{y}_t^{t-1}(\boldsymbol{\delta}) = g(\hat{\boldsymbol{x}}_{t-1},\ \boldsymbol{\delta}) \qquad , \qquad 5.2$$

el modelo predictor toma ahora la forma;

$$g(\hat{\boldsymbol{x}}_{t-1},\ \boldsymbol{\delta}) = f\left\{ W_0 + \sum_{j=1}^{\lambda} W_j f_j \left(\sum_{i=1}^{p} w_{ij}\ \hat{x}_{t-i} + \ w_{p+1,j} \right) \right\}$$

Con el propósito de obtener estimaciones "limpias" que estén libres de valores extremos se utiliza el valor robustamente filtrado de $\hat{x}_{t-i}$ en lugar de las observaciones y_{t-i} .

El filtro robusto recursivo para la estimación de $\hat{x}_t$ está dado por:

$$\hat{x}_t = g(\hat{\boldsymbol{x}}_{t-1}, \boldsymbol{\delta}) \ + \ s_t\ \Psi\left(\frac{y_t - g(\hat{\boldsymbol{x}}_{t-1}, \boldsymbol{\delta})}{s_t} \right)$$

donde el uso de $\Psi = \Psi_{\mathrm{HA}}$ determina el comportamiento para el filtro robusto.

En esta aplicación se ha observado que con $N = 3$ iteraciones fue suficiente para estabilizar los parámetros de la red. (para detalles, ver Joekes, S. 2002)

5.5. Resultados comparativos

Se presenta aquí una tabla comparativa con los resultados predictivos obtenidos por aplicación de los tres métodos considerados para el modelado de los datos de la serie RESEX: el modelo lineal estacional AR(2) con corrección para valores extremos, la AHA con el procedimiento de aprendizaje estándar (sin corrección para valores extremos) y la AHA con el procedimiento de aprendizaje robusto. El primer método no es considerado sin corrección para valores extremos debido a que su ajuste y habilidad predictiva son incorrectos a efectos comparativos.

Todos los modelos fueron ajustados considerando los primeros 7 años u 84 observaciones y utilizando los 5 meses finales a efectos pronósticos (predicción multi-paso). Para la predicción un paso adelante, después de la primera predicción, un nuevo dato se adiciona a la muestra de aprendizaje. Las medidas de ajuste consideradas fueron el RSE y el MAPE como se muestra en la Tabla 2, abajo.

Tabla Nº 2: Resultados comparativos (los valores de RSE están expresados en unidades de 10^3 y el MAPE en porcentajes)

Modelos	Medidas de ajuste		Pronóstico Un paso adelante		Pronóstico Multi-paso	
	RSE	MAPE	RSE	MAPE	RSE	MAPE
a) AR(2) - con corrección para valores extremos	1.34	5.78	1.48	5.40	1.12	4.20
b) AHA - sin corrección para valores extremos	1.60	7.16	1.20	4.14	1.17	4.00
c) AHA - robustamente entrenada	1.37	6.48	1.20	4.33	1.06	3.33

5.6. Conclusión y líneas futuras de investigación

Al comparar los resultados de la red robustamente entrenada con los obtenidos ajustando el modelo AR(2) se pudieron observar muy pocas diferencias en el ajuste y pronóstico de ambos procedimientos. El modelo estadístico mostró un menor error residual y menor porcentaje de error absoluto en el ajuste, pero la red robustamente entrenada superó al modelo estadístico en los pronósticos. Previo a efectuar la corrección para los valores extremos, el modelo de red logró un buen ajuste de los datos de la serie y un mejor pronóstico que el modelo estadístico corregido. Esto se debe a que el efecto de los valores extremos en las RN es sólo local en contraposición con lo que sucede con los modelos estadísticos. En general, cuando la serie presenta uno o dos valores extremos, la RN los modela casi completamente determinando que el modelo de RN es un buen predictor para series de tiempo contaminadas, siempre que no se efectúen pronósticos próximos a los valores extremos. El entrenamiento robusto permitió mostrar la importancia que esta técnica tiene sobre el modelado y habilidad predictiva de las RN.

Finalmente es posible concluir que si bien las RN constituyen una metodología para el análisis de series de tiempo que en muchas situaciones de investigación logran excelentes resultados, también existen modelos estadísticos que logran muy buenos resultados. La razón es simple, no existe un modelo que garantice ser el "mejor" modelo para cualquier serie de tiempo. De todas maneras, considerando las características más destacadas de ambas metodologías en términos de su información interpretativa y explicativa, es posible establecer que requerimientos tales como estimación estadísticamente óptima de parámetros o propiedades de optimalidad de varios algoritmos de entrenamiento, no son considerados usualmente con las RN. y esto constituye la principal ventaja de los modelos estadísticos sobre los modelos de RN.

Algunas de las futuras líneas de trabajo en esta área podrían ser:

1. Estudiar la habilidad predictiva de redes neuronales más complejas como por ejemplo redes recurrentes para modelos ARMA y aplicarlo a series de tiempo no lineales con datos reales.

2. Investigar los algoritmos de aprendizaje de las RN. Estudiar las limitaciones de diferentes software y sugerir la incorporación del cálculo de medidas de variabilidad e intervalos de predicción en la rutina de computación para mejorar la habilidad de generalización de las RN.

3. Futuros trabajos en técnicas de robustez y RN podrían estar centrados en lograr RN robustas a cambios en la varianza del error. Los modelos no lineales, tales como las RN, son particularmente sensitivos a los cambios en la varianza por lo que se podría esperar mejoramientos sustanciales en la predicción de series de tiempo con RN si se pudiera lograr robustificar la varianza.

4. La metodología de modelado de las RN adoptada en este trabajo se sustenta fundamentalmente en el supuesto de estacionariedad de la serie. No se conoce cual es el impacto sobre la metodología adoptada en el caso en que la serie sea no estacionaria. En este sentido sería importante investigar los efectos de posibles transformaciones de los datos para lograr estacionariedad o en su defecto, como modificar la metodología de ajuste de las RN para modelar series no estacionarias.

REFERENCIAS

Allende, H. (1989) "Robust Recursive Estimation of Autoregressive Models" *Rev. Soc. Chilena de Estadística* –6 (1-2): 3-19. Santiago de Chile.

Allende, H. y Moranga, C. (2000): "Time Series Forecasting with Neural Networks". Forschungsbericht 727/2000 – Fachberich Informatik Universitat Dortmund, Germany, March 2000.

Bishop, C. (1995) "Neural Networks for Pattern recognition" Ed. Clarendon Press.- Oxford-

Blough, D. K. y Anderson, K. (1984) "A Comparison of Artificial Neural Networks and Statistical Analysis". U. S. Departament of Energy. Pacific Northwest Laboratory. Richland, Washington.

Blough, D. K. y Anderson. K. (1984) "A Comparison of Artificial Neural Networks and Statistical Analysis". U. S. Departament of Energy. Pacific Northwest Laboratory. Richland, Washington.

Box, G. E. y Jenkins, G. M. (1976) "Time Series Analysis: forecasting and control". Holden-Day.

Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C. (1994) "Time Series: an applied approach". 3ª Ed. Englewood Cliffs. Prentice Hall.

Brubacher, S. R. (1974) "Time series outlier detection and modeling with interpolation". Bell-Laboratories-Technical Memo, USA.

Connor, J. (1993) "Time Series and Neural Network Modeling". Ph.D. Thesis. Univ. of Washington, USA.

Connor, J. T.; Martin, R. D. y Atlas, L. E. (1994) "Recurrent Neural Networks and Robust Time Series Prediction". IEEE Transactions of Neural Networks. Vol. 5 Nº 2- pp. 240-253.

Dickey, D. W. y Fuller, W. (1979) "Distribution of the estimates of AR time series with unit root" *JASA*, vol. 74, 427-431.

Dickey, D. W. y Fuller, W. (1981) "Likelihood Ratio Statistics for AR Time Series with a Unit Root" *Econometrica*, vol. 49, pp 1057-1072.

Haykin, S. (1999) *Neural Networks: a comprehensive foundation*. Macmillan College Publ. Comp., NY, USA.

Joekes, S. (2002) "Aplicación de Redes Neuronales Robustas en Series de Tiempo: un estudio comparativo". Tesis de Maestría en Estadística Aplicada. Universidad Nacional de Córdoba.

Kovács, Z. (1996): "Redes Neurais Artificiais. Fundamentos e Aplicaçones". 2º Edición Collegium Cognitium. SP- Brasil.

Lapedes, A. y Farber, R. (1987) "Nonlinear signal Processing Using Neural Networks. Prediction and modeling" Technical Report Nº LA-UR-87-2662. Los Alamos National Laboratory. Los Alamos, NM.

Martin, R. D. (1980) "*Robust Estimation of Autoregressive Models*" (with discussion) In Directions in Time Series. Dr. Brilliger and G.C. Tiao eds. Inst. of Math. Statist., Hayward, California, 228-262.

Martin, R. D. y Yohai, V. J. (1985) "Highly robust estimation of autoregresion integrated time series models" Tech. Report.

Masreliez, C. J. (1975) "Aproximate Non-Gaussian Filtering with linear State and observation relations". IEEE Trans. Autom. Control – vol. 20 – Nº 1- pp. 107-110.

McCulloch, W. S. y Pitts, W. (1943) "A logical calculus of the ideas immanent in nervous activity" Bulletin of Mathematical Biuphysics. 4, pp.115-133.

Rumelhart, D. E.; Hinton, G. E. y Willams, R. J. (1986.b) "Learning Representations by Back-Propagating Errors". Nature 323. pp. 533-536.

Stahlhut, R. (1985) "Robust Schaetzunger in Zeitreihernmodellen" Diplomarbeit. Abteinlung Statistik Univ. Dortmund.

Weigend, A. S. y Gershenfeld, N. A. (1994) "Time Series Prediction" Reading, MA. Adison-Wesley.

Weigend, A. S.; Huberman, B. A. y Rumelhart, D. E. (1990) "Predicting the Future: A connectionist Approach" Intl. J. Neur. Sys. 1: 193-209.

Werbos, P. J. (1988) "Generalization of Backpropagation with applications to a recurrent Gas-Market model". *Neur. Net.* Vol. 1, pp. 339-356.

Werbos, P. J. (1974). “Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences”. Ph.D. Tesis. Harvard University. Cambridge, MA.