Evaluación de pronósticos del tipo de cambio utilizando redes neuronales y funciones de pérdida asimétricas

Munir A. Jalil B. y Martha Misas A.♦

Febrero de 2006

Resumen

El presente trabajo compara especificaciones lineales y no lineales (expresadas en redes neuronales artificiales) ajustadas a la variación porcentual diaria del tipo de cambio utilizando para ello funciones de costo tradicionales (simétricas) a la vez que se introduce el análisis asimétrico. Los resultados muestran que las redes neuronales permiten obtener mejores pronósticos con ambos tipos de funciones de costos. Sin embargo, es de anotar que cuando se evalúan los pronósticos con funciones asimétricas, el modelo no lineal supera ampliamente a su contraparte lineal.

Clasificación JEL: C22, C53, F31.

Palabras Clave: Tipo de cambio, redes neuronales artificiales, evaluación de pronóstico.

♦ Investigador y Econometrista Principal del Banco de la República, respectivamente. Los resultados y opiniones son responsabilidad exclusiva de los autores y su contenido no compromete al Banco de la República ni a su junta directiva. Los autores agradecen los comentarios y sugerencias de Andrés González.

1. Introducción

Tanto para el sector privado como para la autoridad monetaria es útil tener información sobre la evolución futura del tipo de cambio ya que con ésta es posible establecer la respuesta óptima ante el comportamiento predicho. Por lo anterior, en la literatura internacional se cuenta con un gran número de trabajos en los que se intenta generar *buenos* pronósticos de dicha serie (e.g. Diebold y Nasson 1990; Meese y Rogoff 1983 y Meese y Rose 1991). Estos trabajos se pueden clasificar en dos grupos: los que usan especificaciones lineales y los que utilizan especificaciones no lineales. Una pregunta que surge entonces es ¿por qué se dan distintas especificaciones para la misma serie?. A la luz de la Econometría, toda serie tiene asociado un *proceso generador de datos* y lo que se observa son las realizaciones de éste. Sin embargo, dicho proceso es desconocido y lo que los investigadores deben hacer es aproximarlo, de tal manera que, la especificación obtenida se ajuste a las realizaciones observadas y pueda ser utilizado para pronosticar de manera adecuada. Históricamente, el dominio lo han tenido las especificaciones lineales, las cuales han presentado un mayor desarrollado teórico y una mayor difusión. Además, la dificultad computacional de los procesos de estimación no lineal ha hecho que solamente en tiempos recientes se consideren tales especificaciones que, en principio, pueden representar mejor ciertas características observadas de las series[1].

Otro hecho interesante tiene que ver con la capacidad predictiva de los dos tipos de especificación. Con trabajos como los de Stock (2001), se ha generado la idea que las especificaciones no lineales no son mucho mejores que las lineales para pronosticar y por ello la relación costo-beneficio jugaría en contra de los modelos no lineales. En este sentido, la evaluación tradicional del pronóstico se ha concentrado en medidas simétricas de pérdida, en las que las magnitudes idénticas de errores positivos y negativos tienen el mismo costo asociado. Está por confirmarse, sin embargo, cómo se diferencian estas dos especificaciones a la luz de funciones de pérdida asimétricas en las que los errores tienen valoraciones distintas dependiendo de si son positivos o negativos.

Con respecto a la determinación del tipo de cambio, antes de los setentas, el modelo dominante para su determinación fue el modelo del flujo de bienes. De acuerdo con este modelo, la demanda por moneda extranjera viene principalmente de compras y ventas de bienes. Por ejemplo, un incremento en las exportaciones, incrementa la demanda externa por moneda doméstica con el fin de pagar por los bienes exportados. La implicación de lo anterior es muy intuitiva: países con superávit comercial tendrán apreciaciones (la cual viene de la demanda de doméstica creada por el superávit.) Con todo y lo intuitivo del esquema, al ser esta falla confrontada con los datos: los balances comerciales tienen una correlación muy baja con

[1] Es de anotar que, tal como lo menciona Watson (2005), en determinadas ocasiones las especificaciones no lineales pueden servir para replicar de manera adecuada procesos lineales cuya especificación es desconocida para el investigador.

movimientos del tipo de cambio en los mercados de divisas más importantes. Este resultado negativo no es tan sorpresivo cuando se tiene en cuenta que el mercado de bienes y servicios representa una fracción muy pequeña del total de transacciones de moneda.

En los setentas, para resolver el anterior inconveniente, surgió el modelo asociado al mercado de activos. Este se construyó utilizando la intuición anterior, reconociendo que la demanda de moneda extranjera proviene no solamente de compras y ventas de bienes sino también de compras y ventas de activos. Por ejemplo, con el fin de comprar TES, un inversionista extranjero localizado en los Estados unidos debe comprar los pesos. Adicionalmente, el retorno en dólares del inversionista dependerá de los movimientos del peso, así su demanda por los bonos depende en parte de su deseo de especular en esos movimientos. Este cambio de perspectiva trajo un cambio en la estrategia de modelación. Los modelos comenzaron a incluir nociones tales como "eficiencia" especulativa: los tipos de cambio comenzaron a ser modelados como eficientes, en el sentido de que ellos incorporaban toda la información públicamente disponible, haciendo la información pública inútil para producir retornos extra. Esta es una característica que el modelo de mercado de bienes no presentaba.

El trabajo empírico no ha confirmado la idea de los mercados de activos. Las variables macroeconómicas que están detrás del mismo no mueven el tipo de cambio de la manera predicha. La referencia clásica en este sentido es Meese y Rogoff (1983), quienes muestran que el modelo de mercado de activos no puede explicar los tipos de cambio más importantes mejor que un modelo tan simple como el de "no cambio" o caminata aleatoria. Peor aún, los modelos de mercados de activos no permiten obtener consistentemente la dirección correcta. Lo anterior es recopilado por Meese en su revisión de literatura de 1990, en la que escribe que "la proporción de cambios en el tipo de cambio que pueden ser explicadas por los modelos actuales es esencialmente cero". (La literatura que documenta el comportamiento "tan pobre" de este modelo es vasta. Para revisiones verse Frankel y Rose 1995; Isard 1995 y Taylor 1995).

Estas observaciones negativas no implican que el modelo de mercado de activos esté completamente equivocado. Por el contrario, en la academia existe un consenso que señala, en términos generales, que el modelo es apropiado. Aparentemente hay algo que hace falta para determinar la forma cómo el tipo de cambio es determinado, hecho que hace parte de agendas muy importantes de investigación en muchas partes del mundo.

Por lo anterior, dado que las explicaciones económicas tradicionales necesitan completarse, el trabajo de pronóstico con modelos econométricos de series de tiempo autorregresivos, aparece como una primera alternativa para pronosticar el comportamiento del tipo de cambio.

De acuerdo con Kuan y Liu (1995), es ampliamente aceptado que la tasa de cambio es un proceso integrado de orden uno, $I(1)$, y por ende, su cambio es no correlacionado en el tiempo.

Así, los cambios de la tasa de cambio podrían no ser linealmente predecibles[2]. Surge la inquietud de si el problema de la no predictibilidad está asociado a limitaciones en los modelos lineales. Es decir, si la existencia de no linealidades en el comportamiento del crecimiento de la tasa de cambio conlleva la dificultad que existe para alcanzar pronósticos adecuados al no reconocer tal comportamiento.

Adicionalmente, como lo señala Tenti (1996), la existencia de evidencia que apoya la hipótesis de caminata aleatoria implica que los cambios en la tasa de cambio son independientes e idénticamente distribuidos. Así, la única información relevante de su historia, para la predicción de movimientos futuros, es aquella más reciente.

Por otro lado, tal como lo muestran West et al. (1993), es posible que de maximizar el rendimiento de un portafolio específico que contiene el tipo de cambio, se obtenga que las medidas óptimas de tal procedimiento consistan en analizar los pronósticos de volatilidad a través de una medida asimétrica.

El presente trabajo compara especificaciones lineales y no lineales (expresadas en redes neuronales artificiales) ajustadas a la variación porcentual diaria del tipo de cambio utilizando para ello funciones de costo tradicionales (simétricas) a la vez que se introduce el análisis asimétrico. Los resultados muestran que las redes neuronales permiten obtener mejores pronósticos con ambos tipos de funciones de costos. Sin embargo, es de anotar que cuando se evalúan los pronósticos con funciones asimétricas, el modelo no lineal supera ampliamente a su contraparte lineal.

La manera como se procede consiste en describir brevemente el método de identificación y estimación de una red neuronal artificial a la vez que se describe la metodología "*rolling*" que se siguió para llevar a cabo la evaluación de pronósticos. Luego se hace una descripción de la teoría general de pronósticos, con el fin de (i) señalar el conjunto de supuestos que tradicionalmente se hacen cuando de evaluación de pronósticos se trata y (ii) mostrar las ventajas que se pueden generar al llevar a cabo cambios en dichos supuestos. Lo anterior desde la perspectiva de la teoría de la decisión. Posteriormente, se lleva a cabo la comparación de los pronósticos obtenidos a través de una red neuronal con un modelo ARIMA y una caminata aleatoria utilizando para ello funciones de pérdida simétricas y asimétricas.

2. Clasificación de los modelos econométricos

Como lo presentan Misas et al. (2002), la construcción de un modelo que relacione a una variable y_t con su propia historia y/o con la historia de otras variables, X_t, puede llevarse a cabo a través de una variedad de alternativas, Granger y Teräsvirta (1993). Estas dependen de la forma

[2] Véanse, Baillie y McMahon (1989), citados por Kuan y Liu (1995).

funcional mediante la cual se aproxima la relación, como también, de la relación existente entre dichas variables, es decir, de si ésta es de carácter lineal o no lineal. Las diferentes alternativas pueden ser clasificadas de la siguiente forma:

- No paramétrico: $y_t = f(X_t) + e_t$ donde f no está restringida a pertenecer a una clase específica de funciones.
- Paramétricos: supone una forma funcional específica para $f(\ )$ usualmente con parámetros que deben ser estimados. Por ejemplo:
    - Lineales: $y_t = \beta' X_t + e_t$
    - No lineales:
        - Transición suave: $y_t = \beta_1' X_t + F(X_t)\beta_2' X_t + e_t$ donde la función $F(\ )$ captura la transición del modelo
        - Redes Neuronales: $y_t = \Phi_0 + X_t'\Phi + \sum_{j=1}^{Q} \beta_j G(Z_t'\gamma_j) + e_t$

Semiparamétricos: $y_t = \beta' \overline{X}_t + f(Z_t) + e_t$ donde las variables entran en el modelo de forma paramétrica y no paramétrica.

En nuestro caso, se consideran los pronósticos generados a través de modelos paramétricos lineales como son ARIMA y caminata aleatoria. Dichos pronósticos se contrastan con aquellos obtenidos mediante un modelo paramétrico no-lineal de redes neuronales.

2.1. Modelo paramétrico lineal

Dentro del grupo de modelos paramétricos lineales se consideran: (i) el modelo ARIMA, donde el comportamiento de una serie de tiempo, y_t, se explica a través de sus valores pasados y de una suma ponderada de errores, ε_t, pasados y presentes: $\Phi(L)(1-L)^d y_t = \delta + \Theta(L)\varepsilon_t$; con $\{\varepsilon_t\}$ serie de perturbaciones ruido blanco y d número de diferenciaciones requeridas para que $\{y_t\}$ alcance un comportamiento estacionario y (ii) la caminata aleatoria donde se tiene que $y_t = y_{t-1} + v_t$, con $\{v_t\}$ serie de perturbaciones ruido blanco.

2.2. Modelo paramétrico no lineal

Siguiendo la literatura internacional (Dijk, Terasvirta y Franses, 2001), en los últimos años el uso de modelos no lineales de series de tiempo se ha incrementando de manera considerable y

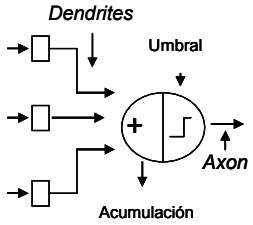
dentro de ellos, los de redes neuronales artificiales (ANN[3]). En el contexto de análisis de series de tiempo, las ANN se clasifican como modelos entrenados para (i) realizar conexiones entre los valores pasados y presentes de una serie de tiempo, aprendiendo de su error de pronóstico y (ii) extraer estructuras y relaciones escondidas que gobiernan el sistema de información (Azoff,1996). Su utilización está primordialmente motivada por la capacidad de aproximarse a cualquier función medible de Borel con un muy buen grado de exactitud, como lo señala, entre otros, Rech (2002)[4].

3. Redes neuronales artificiales

Como lo señalan Swanson y White (1995, 1997a, 1997b), Plasmans et al. (1998) entre otros, los modelos de redes neuronales artificiales se definen como una clase de modelos no lineales flexibles desarrollados por científicos cognitivos. Tales modelos están inspirados en ciertas características asociadas al procesamiento de información en el cerebro humano. El elemento central de este tipo de modelo es la estructura novedosa del sistema de procesamiento de la información, la cual está compuesta por un gran número de elementos interconectados de procesamiento que operan al mismo tiempo para resolver un problema específico. Dichos modelos son capaces de aprender mediante la interacción con su ambiente, tal aprendizaje puede ser entendido como un procedimiento estadístico de estimación recursiva. En particular, una red neuronal artificial se configura para una aplicación específica, de tal forma que, el reconocimiento de patrones y la clasificación de información se alcanza a través de un proceso de aprendizaje. Es de señalar que, el aprendizaje tanto en sistemas biológicos como en las redes neuronales artificiales conlleva ajustes en las conexiones sinápticas entre neuronas[5].

[3] Del inglés, Artificial Neural Network
[4] Citando a Hornik et al. (1989).
[5] Es importante recordar como funciona el cerebro humano en el proceso del aprendizaje. En el cerebro humano, una neurona típica recibe señales de otras a través de receptores de estructura fina, conocidos como *dendrites*. La neurona envía picos de actividad eléctrica a través de un canal llamado *axon*, el cual se divide en cientos de ramas. Al final de cada rama, una estructura llamada *synapse* convierte la actividad del *axon* en un efecto eléctrico que inhibe o estimula la actividad en las neuronas conectadas. Cuando una neurona recibe un input que la estimula, es decir, suficientemente fuerte comparado con aquellos inhibitorios, ésta envía un pico de actividad eléctrica a través de su *axon,* continuando de esta forma se cubren millones de neuronas. El aprendizaje ocurre cambiando la efectividad de la *synapses* , de tal forma que, la influencia de una neurona sobre otra sufra cambios. Una explicación sencilla se da a través de la siguiente figura:

Esquema neuronal

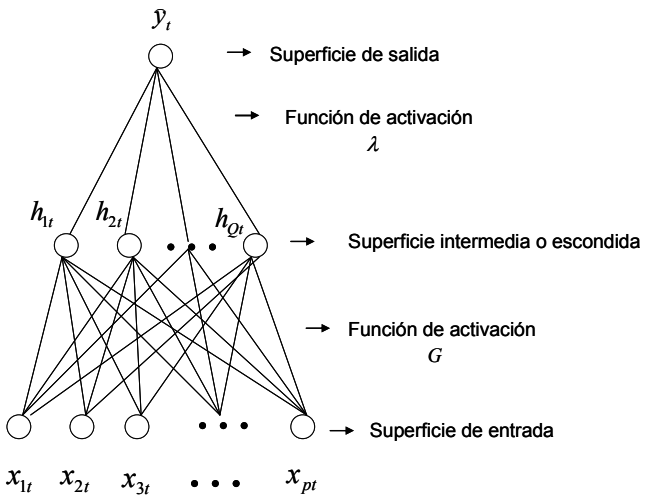


Las redes neuronales artificiales han mostrado, en diferentes campos del conocimiento, una gran capacidad predictiva. Hecho que hace que, en la actualidad, se les considere una herramienta importante en la elaboración de pronósticos de variables macro económicas y financieras. Una posible explicación de tal éxito, es su gran habilidad para aproximar cualquier función dado un número amplio de términos no lineales y una adecuada selección de parámetros.

3.1. Representación

De acuerdo con señalan Kuan y Liu (1995), una red neuronal artificial es un tipo de modelo entrada – salida (*input-output*), que puede ser entendido como una función de regresión no lineal que caracteriza la relación entre una variable dependiente o *output* y_t y un vector de variables explicativas o *inputs* $X_t = \left(x_{1t}, \cdots, x_{pt}\right)$. De tal forma que, sin considerar una función específica no lineal, el modelo se construye combinando muchas funciones no lineales a través de una estructura multicapa (*multilayer structure*).

Gráfico 1



Una clase de ANN, ampliamente estudiada e implementada en este trabajo, es la conocida como de alimentación hacia delante con una única capa o superficie escondida (*single hidden layer feedforward network*), Gráfico 1.

En este tipo de red, las variables explicativas o *inputs* $\left\{x_{1t}, x_{2t}, \cdots, x_{pt}\right\}$ activan de manera simultánea a las Q unidades escondidas en la superficie intermedia, a través de una función G, dando como resultado Q unidades escondidas de activación h_{it}, $i = 1, \cdots, Q$, de tal forma que,

posteriormente estas unidades se activan a través de una función λ para producir el output $\hat{y}_t$. Simbólicamente, lo anterior se describe a través de las siguientes ecuaciones:

$$h_{it} = G\left(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_{jt}\right) \quad \forall i = 1, \cdots, Q \qquad (1)$$

$$\hat{y}_t = \lambda\left(h_{1t}, \cdots, h_{Qt}\right) = \lambda\left(\beta_0 + \sum_{j=1}^{Q} \beta_J h_{jt}\right) \qquad (2)$$

Las funciones de activación son funciones no lineales que pueden ser seleccionadas de manera arbitraria, con una restricción de acotamiento sobre G. Por ejemplo, es usual considerar a G como la función logística: $G(w) = \dfrac{1}{1+\exp(-w)}$ y a λ como la función idéntica, es decir: $\lambda\left(\beta_0 + \sum_{j=1}^{Q} \beta_J h_{jt}\right) = \beta_0 + \sum_{j=1}^{Q} \beta_J h_{jt}$ .Como lo señalan Plasmans et al. (1998) y Franses y Dijk (2000), es conveniente incluir una conexión directa entre la superficie *input* y la *output* para incorporar de manera explicita el modelo lineal básico. Así,

$$\hat{y}_t = \beta_o + \tilde{X}_t' \Phi + \sum_{j=1}^{Q} \beta_J G\left(X_t' \gamma_j\right) + \varepsilon_t \qquad (3)$$

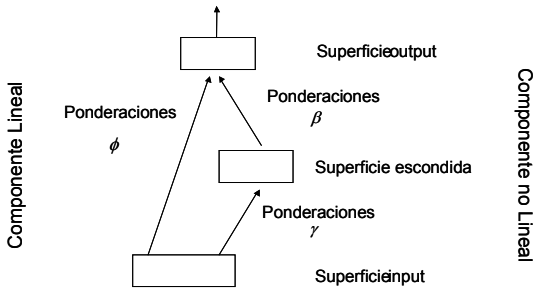
donde $\tilde{X}_t' = \left(x_{1t}, \cdots, x_{pt}\right)$ y $X_t' = \left(1, x_{1t}, \cdots, x_{pt}\right)$. En general, la ecuación (3) se re escribe de la forma presentada en (4) y corresponde al Gráfico 2.

$$\hat{y}_t = F\left(X_t; \Theta\right) = \phi_o + \sum_{r=1}^{P} \phi_r x_{rt} + \sum_{l=1}^{Q} \beta_l G\left(Z_t' \gamma_l\right) + \varepsilon_t \quad ; Z_t \subseteq X_t \qquad (4)$$

siendo $\Theta = \left(\phi_r, r = 0, \cdots, p; \beta_l, \gamma_l; l = 1, \cdots, Q\right)$ donde $p \in \{0,1, \cdots, P\}$, de tal forma que cuando $p = P \Rightarrow Z_t = X_t$, de lo contrario Z_t es un subconjunto propio de X_t. Así, cada combinación p y Q determinan una arquitectura particular[6].

[6] Es de señalar que el número total de arquitecturas puede determinarse a través del contador J el cual enumera las diferentes arquitecturas. Así, $J = 1$ se refiere a la arquitectura correspondiente a

Gráfico 2



3.2. Aprendizaje

El proceso de aprendizaje de las ANN es de carácter secuencial, Kuan y White (1994). Así, el aprendizaje es un proceso donde la red adquiere conocimiento momento a momento, siendo éste definido como la acumulación de experiencias ocurridas. El conocimiento se adquiere a través de los conectores o parámetros de la red. Así, el conocimiento en el momento $(t+1)$, $\hat{\Theta}_{t+1}$ depende del conocimiento en el momento $(t), \hat{\Theta}_t$ .Es decir, $\hat{\Theta}_{t+1} = \hat{\Theta}_t + \Delta_t$,donde el término Δ_t está asociado a un incremento en el conocimiento o aprendizaje, de tal forma que, éste depende del conocimiento previo obtenido de experiencias ocurridas $\{(X_1, y_1), \cdots, (X_t, y_t)\}$ y de los nuevos valores observados (X_{t+1}, y_{t+1}) a través de una función apropiada: $\Delta_t = \Psi((X_{t+1}, y_{t+1}), \Theta_t)$.

El aprendizaje, en los modelos de redes neuronales, se centra en encontrar aquellos valores del conjunto de parámetros que hace mínima la siguiente diferencia:

$$S(\Theta) = \sum_{i=1}^{T} (y_t - F(X_t; \Theta))^2 \qquad (5)$$

al considerar T observaciones de la forma $\{(X_t, y_t)\}_{t=1}^{T}$, donde y_t es la variable *output* o variable objetivo que la red neuronal debe generar cuando el $t-ésimo$ vector *input* X_t aparece. Es decir, el aprendizaje puede ser visto como un problema general de minimización. Por consiguiente, este puede ser abordado a través de diferentes métodos de optimización no restringida, los cuales se llevan a cabo a través de algoritmos recursivos. Así, dado $\hat{\Theta}^{(r)}$ vector de parámetros estimados en la r-ésima iteración, $S(\hat{\Theta}^r)$ suma de residuales al cuadrado y $\nabla S(\hat{\Theta}^r)$ gradiente asociado, la estimación r+1-ésima se obtiene a través de la siguiente formulación:

$(p=1, Q=1)$ en tanto que $J = M$ es : $(P =$ máximo número de variables en la componente no lineal, $Q = 4)$

$\hat{\Theta}^{r+1} = \hat{\Theta}^{r} - \lambda A\left(\hat{\Theta}^{r}\right)^{-1} \nabla S\left(\hat{\Theta}^{r}\right)$, de esta forma, el error cometido en (r) es de vital importancia para la estimación en (r+1), hecho que coincide con el proceso de aprendizaje del cerebro humano en lo referente a la consideración de los errores pasados. Es importante señalar que la diferencia entre algoritmos radica en la determinación de la matriz $A\left(\hat{\Theta}^{r}\right)$, Franses y van Dijk (2000) sugieren el uso particular del método BFGS de los algoritmos quasi newton[7]. Es de señalar que, las propiedades numéricas de dicho método pueden mejorar si las variables son re-escaladas de tal forma que tengan media cero y desviación estándar unitaria.

3.3. Generalización

De acuerdo con Plasmans et al. (1998), una red neuronal tiene una buena capacidad de generalización si ella alcanza un buen desempeño fuera de la muestra en la cual recibió su entrenamiento[8]. Así, existen diferentes enfoques que tratan de incrementar este aspecto. Un enfoque ampliamente conocido es el *weight decay,* en el cual se adicionan términos de penalización a la componente original del error:

$$S(\Theta) = \sum_{t=1}^{n} \left[y_t - F(X_t;\Theta)\right]^2 + r_\phi \sum_{i=0}^{k'} \phi_i^2 + r_\beta \sum_{j=1}^{Q} \beta_j^2 + r_\gamma \sum_{j=1}^{Q} \sum_{i=0}^{k} \gamma_{ij}^2 \qquad (6)$$

de tal manera que (6) se convierte en la función objetivo del proceso de optimización. Dicha penalización garantiza que cada conexión tiende a cero a menos de que ésta sea realmente importante. Es de señalar que, en general, los algoritmos de optimización sobre redes neuronales son sensibles a la selección de los valores iniciales, Rech (2002). Por consiguiente, en la fase de estimación de cada arquitectura se consideran diferentes vectores de valores iniciales[9], Θ^{O}, seleccionando aquel que converja al mínimo valor de la función objetivo Ec (6).

De lo anterior puede concluirse que en la implementación de una red neuronal artificial se requiere de la selección de cuatro elementos básicos: (i) la función de activación G, (ii) el número de unidades ocultas (*hidden units*), (Q), (iii) el número de variables input, (p) y (iv) la determinación de dos conjuntos de información: el primero definido como información de entrenamiento y el segundo como información de evaluación.

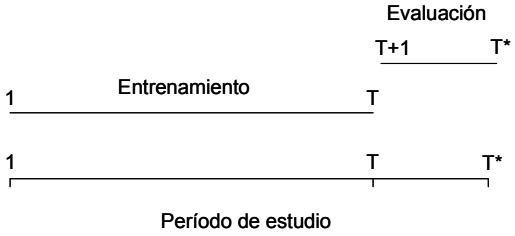
[7] Véase, para una explicación detallada M.C. Aristizábal (2005).
[8] Fuera del período en el cual se llevó a cabo la estimación de sus parámetros. Es decir, aprendizaje adaptativo o capacidad para aprender como se realiza una tarea basado en la información del conjunto de entrenamiento.
[9] Muestreados de manera aleatoria de una distribución uniforme con límites [-2,2]

Existen diferentes estrategias que pueden ser utilizadas para determinar el conjunto de variables *input*, una de ellas es la considerada por Swanson y White (1995, 1997a). Estos autores proponen, como esquema para definir el conjunto de variables *input*, la adopción de un estrategia *step-wise* en la componente lineal. Así, a partir de un sistema de información amplio, X_t^*, se pueda, a través de criterios de selección[10], determinar las variables relevantes, y sus rezagos, en la explicación de la variable objetivo y_t. Dichas variables conformarán el conjunto de información $\tilde{X}_t$. Una vez definido el conjunto de variables *input* de la componente lineal, $\tilde{X}_t$, ecuación (3), se conforman subconjuntos de éste. En primera instancia, el subconjunto incluye tan sólo la primera variable del conjunto *input*, y se define $p=1$, luego se adiciona a dicho conjunto la segunda variable *input* de tal forma que, continuando de esta manera, en el último paso se tiene a $\tilde{X}_t$. La estimación de la red se lleva a cabo considerando cada uno de estos subconjuntos y diferentes números de unidades ocultas, (Q) $(Q=1,\cdots,Q^*)$[11]. De acuerdo con Gradojevic y Yang (2000), un número alto de unidades ocultas Q conducen a un sobre entrenamiento o sobre ajuste que evita que la red alcance una generalización óptima; muy pocas unidades ocultas por otro lado, inhiben el aprendizaje del patrón entre el *input* y el *output*. La selección del mejor modelo se lleva a cabo a través de un esquema *rolling* de evaluación de pronóstico fuera de muestra[12].

3.4. Evaluación fuera de muestra

Como es ampliamente conocido, en el contexto de las ANN es habitual subdividir el período de estudio en dos partes, de tal forma que, en la primera se lleva a cabo el proceso de entrenamiento o aprendizaje y en la segunda el de evaluación , Gráfico 3.

Gráfico 3



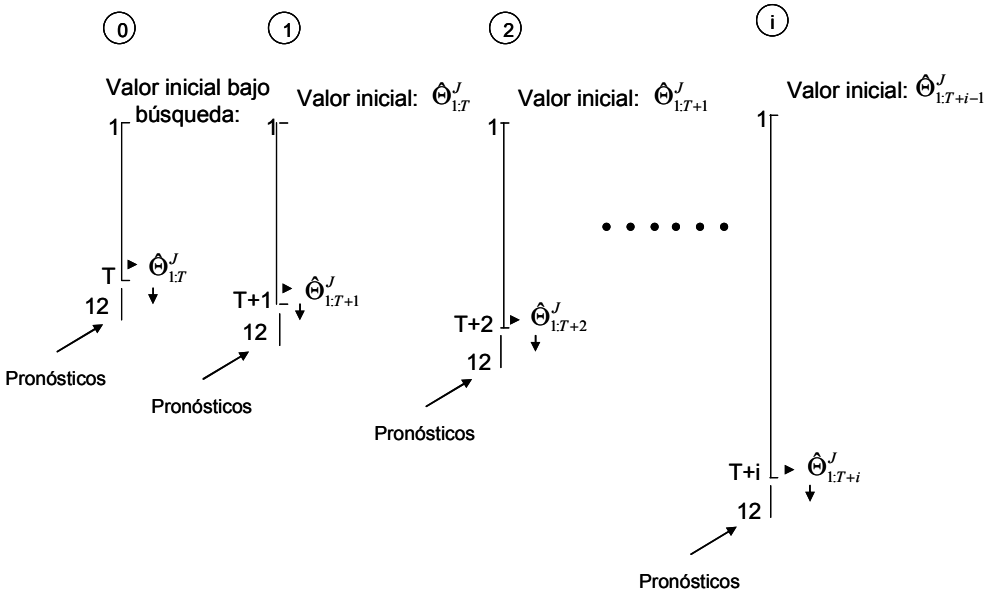
[10] Criterios tales como AIC y BIC.

[11] Donde $Q*$ es el máximo número de unidades ocultas.

[12] Swanson y White (1995, Op.Cit.) señalan que al existir modelos que resultan mejores de acuerdo a alguna medida de evaluación pero no de acuerdo a otras, se deben mirar todas las medidas y se podría incluso especificar una función de pérdida que asigne ponderaciones a las diferentes medidas.

Una vez se lleva a cabo la estimación de las diferentes arquitecturas dentro de la muestra de entrenamiento se pasa a la evaluación por fuera de ésta bajo un esquema de *rolling.* En dicho esquema, se parte del conjunto de parámetros estimados para cada arquitectura en el período de entrenamiento, es decir, $\hat{\Theta}^{J}_{1:T}\ J=1,.....,M$, y se generan pronóstico de horizonte 12 por arquitectura (J), $\hat{y}^{J,1}_{1:T},\cdots,\hat{y}^{J,12}_{1:T}$. Posteriormente, se re-estiman los parámetros de cada arquitectura considerando (i) como conjunto de información aquella que reúne la inicial o muestra de entrenamiento $\{Obs\,1:Obs\,T\}$ con la observación inmediatamente siguiente $\{Obs\,T+1\}$, es decir, $\{Obs\,1:Obs\,T\}\cup\{Obs\,T+1\}=\{Obs\,1:Obs\,T+1\}$, y (ii) como valores iniciales de los parámetros a aquellos obtenidos en el período de entrenamiento $\hat{\Theta}^{J}_{1:T}\ J=1,.....,M$. De esta forma se produce un nuevo conjunto de parámetros: $\hat{\Theta}^{J}_{1:T+1}\ J=1,.....,M$ con los cuales se lleva a cabo un pronóstico de horizonte 12 por arquitectura (J), $\hat{y}^{J,1}_{1:T+1},\cdots,\hat{y}^{J,12}_{1:T+1}$. Así, en el $i-ésimo$ paso se considera como conjunto de Información a $\{Obs\,1:Obs\,T\}\cup\{Obs\,T+1\}\cup\cdots\cup\{Obs\,T+i\}=\{Obs\,1:Obs\,T+i\}$ para la re-estimación y como valores iniciales de los parámetros por arquitectura a aquellos estimados en el paso anterior, es decir, $\hat{\Theta}^{J}_{1:T+i-1}\ J=1,.....,M$. Con tales parámetros se generan, por arquitectura (J), pronósticos de horizonte 12, $\hat{y}^{J,1}_{1:T+i},\cdots,\hat{y}^{J,12}_{1:T+i}$. Este procedimiento se lleva a cabo hasta el momento $(T*-1)$ para tener información observada de la variable y_t y construir las medidas de evaluación simétricas y asimétricas por horizonte $h=1,2,\cdots,12$ y por arquitectura $(J=1,2,\cdots,M)$. El Gráfico 4 presenta la metodología anteriormente explicada.

Gráfico 4



Una vez se tiene el conjunto de pronósticos para cada arquitectura (J) cubriendo el período de evaluación se calculan las medidas de comparación tanto simétricas[13] como asimétricas para cada uno de los horizontes, $(h = 1,2,\cdots,12)$. Así, las medidas de evaluación permiten seleccionar la mejor arquitectura por horizonte de pronóstico.

4. Teoría sobre decisiones y pronóstico

Cuando se pronostica, existe una cantidad desconocida de interés, Y_{t+h}, donde h indica el horizonte de pronóstico. Debido a que esta variable no ha sucedido y estará influenciada por muchos eventos no esperados, se puede considerar como un vector aleatorio indexado por el tiempo. Los usuarios de los pronósticos necesitan tomar decisiones basados en el conjunto posible de realizaciones de esta variable aleatoria. Para ello una caracterización completa de la variable aleatoria es su función de distribución. Entonces es útil caracterizar Y_{t+h} por su distribución condicional, $F_{Y|\Omega_t}(y) = \Pr(Y_{t+h} \leq y|\Omega_t),$ el condicionamiento hecho con respecto a la información conocida en el tiempo t , denotada por Ω_t.

La regla de decisión en este ambiente es el predictor (o predicción cuando se utilizan observaciones). La predicción se denotará como $f_{t+h,t},$ un vector $p \times 1.$ Los errores de pronóstico están definidos como $e_{t+h,t} = f_{t+h,t} - y_{t+h}.$ En el período t tenemos información $x_t, z_t.$ Las variables $\mathbf{x}$ pueden incluir realizaciones presentes y pasadas de las variables de interés u otras variables (tales como variables exógenas), las variables $\mathbf{z}$ son variables de estado. La función de pérdida para pronóstico se puede definir como:

DEFINICIÓN: *Una función de costos es una* ***función valorada en los reales*** $L : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^+$ *denotada por* $L(y_{t+h}, f_{t+h,t}, z_t)$ *donde* $y_{t+h}, f_{t+h,h}$ *son* $p \times 1$ *y* z_t *es* $k \times 1$ *tal que: (i)* $L(x_1, x_2, x_3) = 0, x_1 = x_2$ *y (ii)* $L(x_1, x_2, x_3) > 0, x_1 \neq x_2.$

[13] Véase, Anexo1.

4.1. Una teoría general de pronóstico en economía

4.1.1. Caracterización del pronóstico óptimo

Para encontrar el pronóstico óptimo utilizando teoría de la decisión, la idea es minimizar el riesgo (o maximizar la utilidad esperada)

DEFINICIÓN: *Un **pronóstico óptimo** $f^*_{t+h,t}$ es el pronóstico que minimiza la pérdida esperada (riesgo) i.e.*

$$f^*_{t+h,t} = \arg \min_{f_{t+h,t}} E[L(y_{t+h}, f_{t+h,t}, z_t)|\Omega_t]$$

donde

$$E[L(y_{t+h}, f_{t+h,t}, z_t)|\Omega_t] = \int_y L(y_{t+h}, f_{t+h,t}, z_t) dF_{Y|\Omega_T}(y).$$

Esta definición aparece en varios artículos, entre ellos Granger (1999, 2001). Nótese que este método tal como está planteado no puede ser implementado. Para su desarrollo, es necesario hacer algunos supuestos extra sobre las formas funcionales de la función de pérdida, la regla de decisión y tal vez alguna parametrización de la parte probabilística. Antes de analizar el rol de cada una de las anteriores partes en la definición, es importante mencionar una definición alternativa que hace énfasis sobre el rol de la forma funcional del pronóstico:

DEFINICIÓN: *Un **pronóstico óptimo** $f^*_{t+h,t}$ es el pronóstico que minimiza la pérdida esperada (riesgo) i.e.*

$$f^*_{t+h,t} = \arg \min_{f_{t+h,t}} E[L(y_{t+h}, f_{t+h,t}, z_t)|\Omega_t]$$

$$s.a. \quad f_{t+h,t} = g(x_t, \beta),$$

donde

$$E[L(y_{t+h}, f_{t+h,t}, z_t)|\Omega_t] = \int_y L(y_{t+h}, f_{t+h,t}, z_t) dF_{Y|\Omega_T}(y),$$

$x_t \in \Omega_t$, β es un parámetro.

Alternativamente, es posible tener un problema más familiar

$$\beta^* = \arg \min_{\beta} E[L(y_{t+h}, g(x_t, \beta), z_t)|\Omega_t]$$

y por consiguiente

$$f^*_{t+h,t} = g(x_t, \beta^*).$$

La definición anterior supone un modelo paramétrico para el pronóstico. Es posible utilizar una forma no paramétrica (y la primera definición así lo permite) pero por el resto de esta sección, la especificación paramétrica será utilizada. Es importante recalcar que la forma del modelo puede ser lineal o no lineal.

4.2. El rol de la función de costos (L), la densidad condicional (F), el conjunto de información (Ω) y la forma funcional (g)

4.2.1. La función de costos

La función de costos relaciona los resultados y los pronósticos y puede derivarse de la teoría económica. Sin embargo esta no es la manera en que la literatura ha procedido. Los pronosticadores han utilizado funciones de costos que son matemáticamente convenientes o que tienen otras características llamativas pero que rara vez están relacionadas a los costos económicos o a funciones de utilidad. (Pesaran y Skouras, 2002).

Como se definió hasta ahora, el costo depende de la variable aleatoria de interés (y es así una variable aleatoria por sí misma), el predictor y otras variables. Si el problema particular no necesita otras variables (por ejemplo, no necesita distintas valoraciones para estados diferentes) entonces estas variables pueden salir de la integral.

Las restricciones en la definición son importantes, dado que no cualquier función puede ser una función de costos. En particular una función de costos debería existir si existen errores de pronósticos como tal. La primera restricción $(L(x_1, x_2, x_3) = 0)$ es sólo una normalización. Nótese que los costos están representados como cantidades positivas, así los costos son positivos desde el comienzo o ellos deberían ser multiplicados por -1.

Un subconjunto de funciones de costos es $L(y_{t+h}, f_{t+h,t}, z_t) = L(f_{t+h,t} - y_{t+h}) = L(e_{t+h,t})$. Este costo es llamado *función de pérdida de error de predicción* por Christoffersen y Diebold (1997). Granger y Newbold (1986) sugieren que mirar los errores de pronóstico es sensible porque tanto los pronósticos como las variables de interés tienen propiedades estadísticas diferentes y compararlas puede prestarse para confusiones.

Granger (1999) discute un conjunto más fuerte de condiciones para las funciones de costos. Las restricciones pueden o no pueden incluirse como parte del ejercicio de pronóstico:

- $L(0) = 0$
- $L(e) > 0$ para $e \neq 0$ y
- $L(e)$ monotónicamente no decreciente en $|e|$.

Como la función de costos es sólo una función con algunas restricciones, todas las características usuales asociadas a funciones aparecen pero con la simplificación que el rango es la línea real de los no negativos. Las propiedades adicionales de una función de costos pueden ser (véase Granger, 1999)

- Simetría: $L(-e) = L(e)$ para $p = 1$.
- Homogeneidad: $L(\alpha e) = h(a)L(e)$.
- Convexidad
- Continuidad y
- Diferenciabilidad.

A continuación se presentan un conjunto de funciones de costos muy usado. Sólo se presentan funciones convexas y con excepción del último ejemplo, el vector aleatorio es tan sólo una variable aleatoria $(p = 1)$.

4.2.1.1. Error Cuadrático Medio (ECM)

$$L(e_{t+h,t} : \alpha) = \alpha e_{t+h,t}^2 \quad \alpha > 0$$

α. es una constante libre que no es de importancia. Típicamente, $\alpha = \frac{1}{2}$ con el fin de ayudar con la derivadas. Este costo es el más popular en la literatura debido a su tractabilidad matemática. Es monotónicamente creciente, simétrico, homogéneo de grado 2 y diferenciable en todo su rango.

4.2.1.2. Error Absoluto Medio (EAM)

$$L(e_{t+h,t} : \alpha) = \alpha|e_{t+h,t}| \quad \alpha > 0.$$

Esta función de costos es monotónicamente creciente, simétrica, homogénea y diferenciable en todo su rango con la excepción de $e_{t+h,t} = 0$.

4.2.1.3. Función Linex (Linex)

Introducida por Varian (1974) y estudiada en detalle por Zellner (1986)

$$L(e_{t+h,t} : \alpha_1, \alpha_2) = \alpha_1[\exp(\alpha_2 e_{t+h,t}) - \alpha_2 e_{t+h,t} - 1] \quad \alpha_1 \geq 0, \alpha_2 \neq 0.$$

La función está normalizada de tal manera que $L(0) = 0.$ Esta función de costos es asimétrica. Si $\alpha_2 > 0$ es casi lineal a la izquierda del eje y y casi exponencial a la derecha. Esta función se voltea si $\alpha_2 < 0.$ La función es diferenciable en todo su rango. Nótese que si $\alpha_1 = \frac{1}{\alpha_2^2}$ entonces $\lim_{\alpha_2 \to 0} L(e) = \frac{e^2}{2}$ así para α_2 pequeño, el costo cuadrático está aproximadamente anidado dentro del costo linex.

4.2.1.4. Función doble exponencial

$$L(e_{t+h,t} : \alpha_1, \alpha_2, \beta_1, \beta_2) = L(e_{t+h,t} : \alpha_1, \alpha_2) - L(e_{t+h,t} : \beta_{1,} - \beta_2) \quad \alpha_1, \beta_1 \geq 0, \quad \alpha_2, \beta_2 \neq 0.$$

Esta función de costos no es simétrica si $\alpha_2 > 0$ y $\beta_2 > 0, \alpha_2 \neq \beta_2.$ La misma se convierte en simétrica si $\alpha_2 = \beta_2.$ La función es casi exponencial para errores tanto positivos como negativos.

4.2.1.5. Familia de funciones de pérdida particionadas

$$L(e_{t+h,t} : a, b, \rho) = \begin{cases} aL_1(e_{t+h,t} : \rho) & e_{t+h,t} > 0 \\ bL_2(e_{t+h,t} : \rho) & e_{t+h,t} < 0 \end{cases} \quad a, b, \rho > 0.$$

Típicamente se elige

$$L_1(e_{t+h,t} : \rho) = L_2(e_{t+h,t} : \rho) = |e_{t+h,t}|^\rho \quad \rho \in \mathbb{N}.$$

Casos especiales son:

- $\rho = 1$: caso lin-lin
- $\rho = 2$: caso Quad-Quad.

Las dos son continuas pero no diferenciables en cero. Para $a \neq b$ son asimétricas. Excepto por la no diferenciabilidad en cero, la Lin-Lin es diferenciable una vez y la Quad-Quad es doblemente diferenciable.

Trabajar con esta familia se hace más fácil llevando a cabo una transformación. Para ello hacemos $\alpha = \frac{a}{a-b}, \alpha \in (0,1)$ de tal manera que:

$$L(e_{t+h,t} : \alpha, \rho) = (a+b)[\alpha + (1-2\alpha)1_{(e<0)}]|e_{t+h,t}|^{\rho}$$

donde $1_{(e<0)}$ es una función indicadora que toma el valor de 1 si $e < 0.$ Dado que la función es homogénea es posible obviar el término $(a+b)$ y trabajar con:

$$L(e_{t+h,t} : \alpha, \rho) = [\alpha + (1-2\alpha)1_{(e<0)}]|e_{t+h,t}|^{\rho}$$

bajo esta especificación, otros casos especiales son:

- $\rho = 1, \alpha = \frac{1}{2}$: EAM.
- $\rho = 2, \alpha = \frac{1}{2}$: ECM.

4.2.2. La distribución condicional de predicción

La distribución condicional caracteriza completamente la variable aleatoria de interés. Si el interés es en un vector aleatorio (porque existen varias variables de interés o porque la idea es pronosticar varios períodos de tiempo) una distribución conjunta es apropiada. Desde el punto de vista de la teoría de la decisión, la distribución condicional describe la incertidumbre asociada al problema. Sin embargo, la mayoría de la literatura de pronóstico no considera la distribución dado que únicamente están interesados en pronósticos puntuales, aunque en ocasiones, con el desconocimiento de los autores, existe un supuesto sobre la misma implícito. Desarrollos más recientes, estiman el total de la distribución condicional. Esto puede ser hecho, por ejemplo, utilizando regresión por percentiles o análisis no paramétrico. Para una discusión ver Diebold et al (1998) y Elliott y Timmermann (2002).

Cuando la distribución de predicción es estimada, los tomadores de decisiones con diferentes funciones de costos pueden tomar decisiones óptimas. Por esta razón, Granger y Pesaran (2000) argumentan que todo lo que se necesita para pronosticar es un pronóstico de la distribución. Con la disminución en el costo y tiempo de cómputo, esta alternativa está ganando cada día más adeptos. Entre otros, el Banco de la República publica pronósticos de densidad trimestrales de la inflación y el producto.

4.2.2.1. Interacción entre la función de costos y la distribución condicional.

La importancia de la interacción entre el costo y la distribución condicional es bien sabida al menos desde el artículo de Granger (1969). Por ejemplo, asimetrías en ambos interactúan para definir el predictor óptimo. Asimetrías en la función de costos indican si existen costos diferentes asignados a sobre o sub-predicciones, mientras la asimetría en la distribución indica si la realización de la variable de interés tiende a estar por encima o por debajo de la media.

Si sobre-predicciones (errores de pronóstico positivos) son igual de costosos que sub-predicciones (errores de pronóstico negativos), y la distribución condicional es además simétrica, entonces el pronóstico óptimo debería ser la media. (Gráfico 5).

Si las sub-predicciones son más costosas que las sobre-predicciones (con una función lin-lin por ejemplo) y la distribución condicional es simétrica, entonces el pronóstico óptimo debería estar por encima de la mediana con el fin de hacer las sub-predicciones menos recurrentes. Qué tan lejos de la mediana, depende de los parámetros de la función de costos. El Gráfico 6 muestra un ejemplo donde el pronóstico óptimo está ubicado en el tercer cuartil.

Gráfico 5

Distribución condicional simétrica y pérdida simétrica

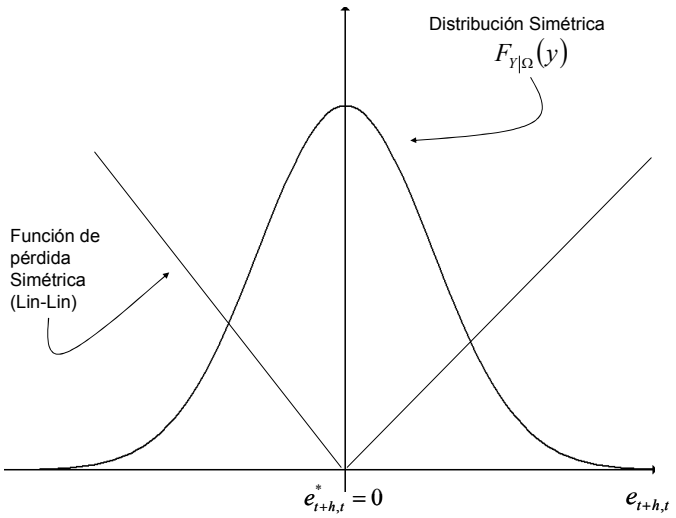
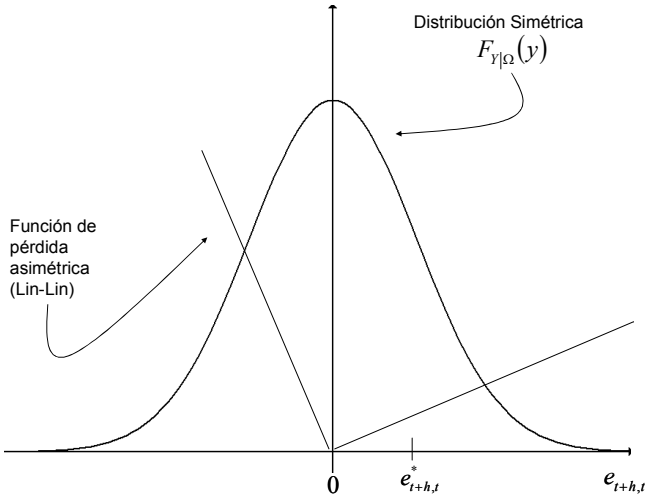


Gráfico 6

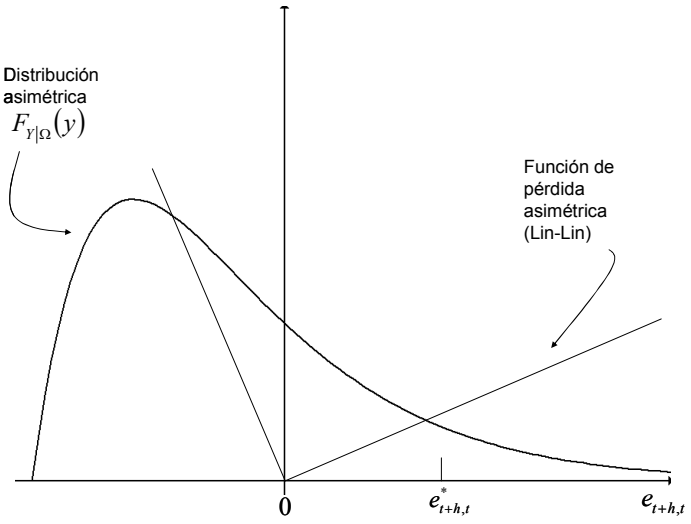
Distribución condicional simétrica y pérdida asimétrica



Como un ejemplo final, nótese que con la misma función de costos asimétrica del Gráfico 5, pero con una distribución condicional asimétrica, el pronóstico óptimo estará aún en el tercer cuartil pero, dado que la distribución es diferente, la realización del pronóstico como tal debe cambiar. El Gráfico 7 muestra una distribución con la media por encima de la mediana (con asimetría derecha).

Gráfico 7

Distribución condicional asimétrica y pérdida asimétrica



Nótese que un pronóstico óptimo puede ser diferente de la media condicionada ya sea por asimetrías en la función de costos o debido a asimetrías en la distribución condicional.

Elliott y Timmermann (2002) muestran que otra manera de ver la interacción entre la función de costos y la distribución se logra a través de realizar una expansión de Taylor alrededor de la media condicional del error de pronóstico: $\mu_e = E[f_{t+h,t} - y_{t+h}|\Omega_t]$.

$$L(e_{t+h,t}) = L(\mu_e) + L'(\mu_e)(e_{t+h,t} - \mu_e) + \frac{1}{2}L''(\mu_e)(e_{t+h,t} - \mu_e)^2 + \sum_{k=3}^{\infty} \frac{1}{k!}L^k(\mu_e)(e_{t+h,t} - \mu_e)^k$$

y tomando expectativas

$$E[L(e_{t+h,t})|\Omega_t] = L(\mu_e) + \frac{1}{2}L''(\mu_e)E\left[(e_{t+h,t} - \mu_e)^2 \,\middle|\, \Omega_t\right] + \sum_{k=3}^{\infty} \frac{1}{k!}L^k(\mu_e)E\left[(e_{t+h,t} - \mu_e)^k \,\middle|\, \Omega_t\right]$$

La interacción entre la forma de la función de costos (las derivadas) y los momentos (centrales) de la distribución se observa claramente en la expresión. Combinaciones entre los valores de las derivadas y los momentos de la distribución, determinarán qué tanto de la una o la otra se necesita para aproximar la función de costos.

4.2.3. El conjunto de información

Los contenidos del conjunto de información son una de las elecciones más importantes del pronosticador dado que cada pronóstico está condicionado al mismo. Conjuntos distintos de información generarán pronósticos óptimos distintos, aún cuando esto es usualmente olvidado cuando éstos se interpretan o se evalúan. La elección del conjunto de información es una de las partes del proceso de pronóstico donde la teoría económica puede ser muy útil, indicando las variables relevantes a incluir.

Los contenidos del conjunto de información Ω_t son usualmente denotados por $\Omega_t : Y_{t-j}, X_{i,t-j}, i = 1, \ldots, l, j \geq 0,$ véase Granger (1999). Si el conjunto de información contiene el pasado y el presente de las series de interés se le llama un subconjunto propio del total de datos. La idea es que, como mínimo, el conjunto de información sea un subconjunto propio con el fin que este contenga los errores de pronóstico pasados. Esto es fundamental cuando se consideran las propiedades de pronósticos óptimos y además para la evaluación de éstas

propiedades. Cuando hablamos de series univariadas $(l = 0)$, el subconjunto propio incluirá el presente y el pasado de la serie de interés. El trabajo clásico de pronóstico utilizando el anterior conjunto de información (hecho en un ambiente cuadrático-lineal) es el hecho por Box y Jenkins (1994).

Cuando los pronósticos son multivariados $(l > 0)$, la teoría económica puede ayudar a decidir qué variables incluir y cuáles de ellas pueden ser consideradas como exógenas o endógenas. Una referencia sobre la elección de variables y exogeneidad es Clements y Hendry (1998).

4.2.4. La forma funcional del modelo de pronóstico

Existen algunos cuestionamientos sobre la forma funcional del modelo de pronóstico y en general éstas interactúan con la función de pérdida, la distribución condicional y la elección de la información utilizada.

Tal vez las dos más importantes son las decisiones sobre el uso de un modelo paramétrico de la forma $g(x_t, \beta)$ o el uso de un modelo no paramétrico (o semi paramétrico) y la elección de la forma funcional. La forma funcional puede ser lineal o no lineal. En la literatura de pronóstico, el modelo lineal paramétrico es el más usado, pero en tiempos recientes alternativas no lineales paramétricas han ganado popularidad. Alternativas recientes incluyen además la búsqueda por una forma funcional dadas las ventajas asociadas a los bajos costos de computación disponible (las redes neuronales artificiales son un ejemplo), aunque a este nivel, las ventajas de tales procedimientos son objeto de estudio.

La elección de la forma funcional es donde la relación con la literatura econométrica es más clara. En particular, bajo algunas condiciones el pronóstico óptimo es la expectativa condicionada y entonces $g(x_t, \beta)$ está directamente relacionado al campo de regresión (Como se encuentra explicado por Elliott y Timmermann (2002) y Granger (1969)). Otros vínculos se pueden establecer, por ejemplo, cuando se usan las funciones de costos lin-lin y se prueba que la regresión por percentiles es muy útil (véase Elliott y Timmermann (2002)).}

4.3. La elección del pronóstico óptimo

La elección del pronóstico óptimo involucra entonces un conjunto de interacciones que tradicionalmente no son tenidas en cuenta. La mayoría de la literatura sobre pronóstico supone una distribución simétrica para la generación de los mismos y evalúan éstos con una función simétrica

de costos. Lo que hemos visto en los apartados inmediatamente anteriores es que, aparte de este esquema convencional, se pueden presentar situaciones en las que la utilización tanto de funciones condicionales como de funciones de costos asimétricas generan pronósticos óptimos distintos a los obtenidos bajo esquemas simétricos. Es por esto que en la sección siguiente, aparte del análisis tradicional de pronóstico, se añadirá el análisis con funciones de pérdida asimétricas con el fin de ilustrar este punto.

5. Resultados

De las secciones anteriores se observa que un análisis de pronóstico con un modelo no lineal que tenga en cuenta las posibles asimetrías en la función de costos podría brindar información útil para la toma de decisiones por parte de los encargados de las mismas. Es por ello que se decidió, a manera de ilustración, tomar el cambio de la tasa de cambio nominal y realizar un ejercicio en el que se le ajusta tanto un modelo lineal como uno no lineal evaluándose ambos con medidas simétricas y asimétricas.

La evaluación de pronóstico del cambio de la tasa de cambio nominal se lleva a cabo sobre los pronósticos de un modelo lineal ARIMA y de uno no lineal que considera una red neuronal artificial autorregresiva. Este trabajo se lleva a cabo con información diaria correspondiente a la primera diferencia del logaritmo de la tasa de cambio nominal. El estudio abarca el período comprendido entre el 8 de febrero de 2000 y el primero de marzo de 2005. Intervalo de tiempo en el cual, tanto para el modelo ARIMA como para ANN, las últimas 60 observaciones son utilizadas para la evaluación *rolling* fuera de muestra.

5.1. Modelo no lineal

El período de entrenamiento corresponde a la muestra comprendida entre el 8 de febrero de 2000 y el 2 de diciembre de 2004, con un total de 1176 observaciones. La evaluación rolling fuera de muestra considera el período entre el 3 de diciembre y el primero de marzo de 2005, es decir 60 observaciones. Con el propósito de mejorar las propiedades de estimación, como se mencionó anteriormente, la variable crecimiento de la tasa de cambio, $\Delta LTCN_t$, es re-escalada en el intervalo (0,1).

La determinación de las variables *inputs* de la componente lineal o determinación del conjunto X_t, en cada red, se lleva a cabo mediante la estrategia ***stepwise***[14], propuesta por

[14] A pesar de ser la estrategia *Stepwise* de carácter lineal, es frecuentemente utilizada como mecanismo de selección en el contexto de redes neuronales. Como lo expresa Franses frente a una consulta de Arango et.

Swanson y White (1995,1997a). Así, en dicha estrategia se parte de una regresión lineal cuya variable dependiente es $\Delta LTCN_t$ y cuyas posibles variables explicativas son seleccionadas dentro de sus primeros 24 rezagos.

Una vez definido el conjunto de variables *input* de la componente lineal, X_t, se realiza el proceso de estimación de la red neuronal mediante el proceso de optimización numérica Quasi-Newton de Broyden, Fletcher, Goldfarb y Shano[15], ecuación (6), para las diferentes configuraciones del conjunto de información Z_t, $Z_t \subseteq X_t$, de la componente no lineal[16] y para un número de unidades ocultas Q que varían desde uno hasta cuatro[17].

Franses y van Dijk (2000) señalan cómo la convergencia en el proceso de estimación no garantiza la obtención del mínimo global. Por consiguiente, se llevan a cabo 30 estimaciones de cada una de las diferentes arquitecturas utilizando distintos valores iniciales del vector de parámetros γ. Tales valores iniciales son obtenidos aleatoriamente a partir de una distribución uniforme entre [-2,2]. Los parámetros del término de *weight decay* en la función objetivo $S(\Theta)$ son $r_\phi = 0.01$, $r_\beta = r_\gamma = 0.0001$.

La selección de las cinco estimaciones óptimas por arquitectura se realiza considerando dos criterios: (i) menor valor de la función objetivo y (ii) vector de gradientes, asociado a los parámetros de la estimación, sin elementos superiores a 1×10^{-3}. Una vez llevada a cabo dicha selección se lleva a cabo el procedimiento de pronósticos fuera de muestra bajo el esquema *rolling.* Para finalmente, proceder a calcular las medidas de evaluación simétricas y asimétricas.

El cuadro 1 presenta, por horizonte, las medidas de evaluación RMSPE y MAPE, de carácter simétrico, de los modelos ARIMA, ANN y caminata aleatoria. Es de señalar que en el caso

---

Al. (2004): *"As nonlinear functions can appear in dozens of formats, it is difficult to make a selection first. Hence, one usually starts with the first order linear approximation".*

[15] Como lo sugieren Franses and Dick V. Dijk (2000) y Rech (2002), este es uno de los algoritmos más utilizados en el contexto de redes neuronales para solucionar el problema de minimización, planteado en la ecuación (2).

[16] La especificación del conjunto Z_t se lleva a cabo de la siguiente manera: en un primer paso, o $p=1$, el conjunto Z_t incluye la primera variable del conjunto X_t, luego, en un segundo paso, o $p=2$, se adiciona al conjunto Z_t la segunda variable de X_t de tal forma que en el último paso, $p=P$, se tiene la igualdad de los conjuntos, $Z_t = X_t$. Es de resaltar que, el conjunto X_t que conforma la componente lineal permanece invariante a través de las diferentes arquitecturas

[17] La selección de Q desde uno hasta cuatro es una regularidad empírica observada en trabajos similares.

de las ANN se reporta aquella red o arquitectura que es la mejor en el sentido de mínima medida de evaluación. La existencia de evidencia a favor de la caminata aleatoria como proceso generador del nivel de la tasa de cambio lleva a que el mejor pronóstico de $\Delta LTCN$ es cero, valor frente al cual se lleva a cabo la evaluación. Como puede observarse, en el caso del RMSPE a horizontes menores a cinco días el mejor pronóstico se obtendría a través del supuesto de caminata aleatoria. Para los restantes horizontes, claramente se obtienen reducciones del error de pronóstico al considerar las diferentes arquitecturas de la red neuronal.

CUADRO 1

Medidas simétricas de Evaluación de pronósticos bajo *rolling* $\Delta LTCN$

			RMSPE		
HORIZONTE	**P**	**Q**	**ANN**	**ARIMA**	**RW**
1	4.00	3.00	119.80	125.72	100.00
2	2.00	3.00	108.73	118.98	100.00
3	2.00	3.00	102.27	125.05	100.00
4	2.00	3.00	103.40	125.11	100.00
5	4.00	2.00	102.92	124.84	100.00
6	3.00	4.00	96.65	125.11	100.00
7	3.00	4.00	97.82	126.55	100.00
8	5.00	4.00	95.40	130.15	100.00
9	5.00	4.00	92.46	131.22	100.00
10	5.00	4.00	107.90	145.55	100.00
11	9.00	2.00	99.70	102.58	100.00
12	3.00	3.00	100.71	102.66	100.00

			MAPE		
HORIZONTE	**P**	**Q**	**ANN**	**ARIMA**	**RW**
1	5.00	4.00	95.99	103.06	100.00
2	2.00	4.00	93.69	103.38	100.00
3	2.00	3.00	90.19	109.18	100.00
4	2.00	2.00	92.92	110.59	100.00
5	4.00	2.00	91.35	110.26	100.00
6	3.00	4.00	89.83	110.43	100.00
7	5.00	4.00	90.60	111.15	100.00
8	4.00	2.00	86.80	112.38	100.00
9	5.00	4.00	84.80	112.84	100.00
10	5.00	4.00	97.97	116.31	100.00
11	1.00	3.00	89.25	100.97	100.00
12	3.00	3.00	87.39	101.05	100.00

En lo referente a la medida MAPE, claramente se observa, para todo horizonte, la ventaja de trabajar con redes neuronales artificiales. Las dos medidas señalan la poca conveniencia de trabajar con modelos lineales como lo es el modelo ARIMA.

El Cuadro 2 presenta las medidas simétricas de evaluación RMSE y MAE. Se observa, para todo horizonte de pronóstico, un mejor comportamiento de los pronósticos obtenidos a través de las redes neuronales.

CUADRO 2

Medidas simétricas de Evaluación de pronósticos bajo *rolling* $\Delta LTCN$

			RMSE		
HORIZONTE	**P**	**Q**	**ANN**	**ARIMA**	**RW**
1	2.00	2.00	0.0072	0.0076	0.0081
2	3.00	2.00	0.0075	0.0078	0.0081
3	1.00	4.00	0.0079	0.0081	0.0082
4	3.00	4.00	0.0079	0.0081	0.0082
5	7.00	3.00	0.0072	0.0081	0.0083
6	9.00	4.00	0.0078	0.0081	0.0083
7	9.00	4.00	0.0076	0.0080	0.0082
8	9.00	4.00	0.0075	0.0079	0.0081
9	2.00	4.00	0.0077	0.0080	0.0082
10	2.00	4.00	0.0079	0.0082	0.0082
11	7.00	4.00	0.0081	0.0083	0.0083
12	7.00	3.00	0.0080	0.0082	0.0082

			MAE		
HORIZONTE	**P**	**Q**	**ANN**	**ARIMA**	**RW**
1	2.00	2.00	0.0047	0.0052	0.0057
2	3.00	4.00	0.0050	0.0055	0.0058
3	5.00	4.00	0.0054	0.0057	0.0058
4	4.00	2.00	0.0055	0.0057	0.0058
5	7.00	3.00	0.0053	0.0058	0.0059
6	7.00	4.00	0.0053	0.0057	0.0058
7	9.00	4.00	0.0053	0.0056	0.0057
8	6.00	2.00	0.0051	0.0055	0.0056
9	2.00	4.00	0.0052	0.0055	0.0056
10	6.00	3.00	0.0054	0.0057	0.0056
11	4.00	4.00	0.0054	0.0057	0.0057
12	9.00	2.00	0.0053	0.0055	0.0056

En el Cuadro 3 se consignan los resultados de la estrategia asimétrica de evaluación de pronósticos del modelo lineal ARIMA, el modelo no lineal de redes neuronales artificiales, y los resultados de evaluación de los pronósticos de una camita aleatoria. Como medida asimétrica se eligió una función Lin-Lin en la que se le asigna un costo más alto a sub-predicciones que a sobre predicciones[18]. Del cuadro se deduce que si existe un interés en un modelo que brinde pronósticos que no arroje demasiadas sub-predicciones (dado que éstas son costosas) una ANN es mucho más eficiente para este propósito que un modelo ARIMA o una caminata aleatoria a todo horizonte.

CUADRO 3

Medida asimétrica de Evaluación de pronósticos bajo *rolling* $\Delta LTCN$

			LINLINP		
HORIZONTE	**P**	**Q**	**ANN**	**ARIMA**	**RW**
1	8	1	0.17	0.25	0.27
2	8	1	0.17	0.26	0.27
3	8	1	0.17	0.28	0.28
4	9	1	0.18	0.28	0.28
5	8	1	0.18	0.28	0.28
6	8	1	0.18	0.28	0.28
7	9	1	0.18	0.29	0.28
8	9	1	0.17	0.29	0.29
9	7	1	0.17	0.30	0.29
10	7	1	0.17	0.30	0.30
11	7	1	0.18	0.30	0.30
12	7	1	0.17	0.31	0.31

6. Conclusiones

Este trabajo compara pronósticos provenientes de un modelo no lineal (Red Neuronal), con los de un modelo lineal tradicional (ARIMA). Los pronósticos son obtenidos a través de una metodología de "rolling" y su evaluación se lleva a cabo con respecto a medidas tanto simétricas (las cuales asignan la misma valoración a errores de la misma magnitud sin importar su signo) como asimétricas (las cuales permiten diferenciar los errores dependiendo no solamente de su magnitud sino de su signo). La literatura siempre ha tenido la visión que los pronósticos no lineales, si son mejores para pronosticar, no lo son de una manera abrumadora con respecto a sus contrapartes lineales. Este argumento ha sido siempre esbozado utilizando para ello funciones de pérdida simétricas. Por lo anterior este trabajo presenta un esquema de la teoría de decisión y pronóstico en economía, con el fin de ilustrar las distintas posibilidades que existen para evaluar

[18] La idea detrás de esta valoración proviene del hecho de que sub-predicciones (pronosticar por debajo del valor efectivamente observado) significan mucho dinero perdido para el sistema financiero.

predicciones. De allí se concluye que las funciones basadas en minimización de un error cuadrático medio son tan sólo una de las muchas posibilidades existentes para evaluar la bondad de un pronóstico. Por lo anterior, en el presente documento se utilizaron, además de las medidas tradicionales, funciones de pérdida asimétricas con el fin de comparar, bajo este esquema, los pronósticos.

Los resultados obtenidos permiten concluir que a la luz de las funciones de pérdida asimétricas, los modelos no lineales tienen una mejora considerable en capacidad de pronóstico, con respecto a los modelos lineales. Este resultado es robusto al horizonte de pronóstico, justificando así el uso de técnicas de estimación más complejas si lo que se necesita es solucionar un problema en el que el pronóstico óptimo deba ser evaluado con funciones de pérdida que no son simétricas.

Bibliografía

Arango, C., M. Misas, E. López y J. N. Hernández (2004), “No-linealidades en la demanda de efectivo en Colombia: las redes neuronales como herramienta de pronóstico”, Ensayos sobre política económica, No. 45, 11- 57.

Azoff, E. M. (1996), *Neural Network. Time Series Forecasting of Financial Markets*, Wiley, A Wiley Finance Edition.

Box, G.E.P., G.M. Jenkins, y G.C. Reinsel (1994), *Time Series Analysis. Forecasting and Control*, 3ra edición. New Jersey: Prentice-Hall.

Christoffersen, P. y F.X. Diebold, (1997), "Optimal Prediction under Asymmetric Loss," Econometric Theory, 13, 806-817.

Clements, M.P. y D.F. Hendry, (1998), *Forecasting Economic Time Series*. Cambridge Cambridge University Press.

Diebold, F.X., T.A. Gunther y A.S. Tay, (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management," International Economic Review, 39, 863-883.

Diebold, F.X. and Nasson, J.M. (1990) “Nonparametric Exchange Rate Prediction” *Journal of International Economics* (28) pp. 315 - 32.

Elliott, G. y A. Timmermann, (2002), "Optimal Forecast Combinations under General Loss Functions and Forecast Error Distributions," UCSD Working Paper.

Frankel, J y A. Rose (1995), “Empirical research on nominal exchange rates”. En *Handbook of International Economics,* editado ´por G. Grossman y K. Rogoff. Amsterdam: Elsevier Science.

Franses P.H. and D. van Dijk (2000), *Non-linear time series models in empirical finance*, Cambridge University Press.

Gradojevic, N. and J. Yang (2000), “The Application of Artificial Neural Networks to Exchange Rate Forecasting: The Role of Market Microstructure Variables”, Working paper 2000-23, Bank of Canada.

Granger, C.W.J., (1969), "Prediction with a Generalized Cost Function," Operational Research 20, 199-207, reimpreso en E. Ghysels, N.R. Swanson y M.W. Watson (eds.), *Essays in Econometrics: Collected Papers of Clive W.J. Granger*, Volume I, 2001. Cambridge: Cambridge University Press.

Granger, C.W.J. y P. Newbold, (1986), *Forecasting Economic Time Series*, 2nd edition. Orlando: Academic Press.

Granger, C.W.J. y T. Teräsvirta (1993), *Modelling Nonlinear Economic Relationships*, Advanced Texts in Econometrics, Oxford University Press.

Granger, C.W.J., (1999), "Outline of Forecast Theory using Generalized Cost Functions," Spanish Economic Review, 1, 161-173.

Granger, C.W.J. y M.H. Pesaran, (2000), "A Decision Theoretic Approach to Forecast Evaluation," en W.S. Chon, W.K. Li, y H. Tong (eds.) *Statistics and Finance: An Interface*, pp. 261-278. London: Imperial College Press.

Granger, C.W.J., (2001), "Evaluation of Forecasts," en: Hendry, D.F. y N.R. Ericsson, *Understanding Economic Forecasts*. Cambridge: The MIT Press.

Isard, P. (1995), *Exchange rate economics.* Cambridge: Cambridge University Press.

Kuan C. M. y H. White (1994), "Artificial Neural Networks: An Econometric Perspective", Econometric Reviews 13.

Kuan C. M. y T. Liu (1995), "Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks", Journal of Applied Econometrics, Vol. 10, 347-364.

Meese, R. y Roose, A. (1991). "An empirical assessment of non-linearities in models of exchange rate determination". en: Review of Econometric Studies, No 58.

Meese, R. y K. Rogoff (1983), "The out-of-sample failure of empirical exchange rate models. En *Exchange Rate and International Marcoeconomics*, editado por J. Frenkel. Chicago: University of Chicago Press.

Plasmans J., W. Verkooijen y H. Daniels (1998), Estimating Structural Exchange Rate Models by Artificial Neural Networks", Applied Financial Economics, 8, 541-551.

Pesaran, M.H. and S. Skouras, (2002), "Decision-Based Methods for Forecast Evaluation," in M.P. Clements and D.F. Hendry (eds.), A Companion to Economic Forecasting, Oxford: Blackwell Publishers.

Swanson, N. R. y H. White (1995), "A Model-Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks", Journal of Business & Economic Statistics, Vol. 13, No.3.

---------- (1997a), "A model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks", The Review of Economics and Statistics, No. 79.

---------- (1997b), "Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models", International Journal of Forecasting, No. 13.

Rech G. (2002), "Forecasting with artificial neural network models", SSE/EFI Working paper Series in Economics and Finance, No. 491.

Stock J. H. (2001), "Forecasting Economic Time Series", en *A Companion to Theoretical Econometrics*, Blackwell Publishers.

Taylor, M. (1995). "The Economics of Exchange Rates". *Journal of Economic Literature* 83: 13-47.

Tenti, P. (1996), "Forecasting Foreign Exchange Rates Using Recurrent Neural Networks", Applied Artificial Intelligence, 10: 567-581.

Van Dijk, D., T Teräsvirta y P. H. Franses (2001), "Smooth Transition Autoregressive Models – A Survey of Recent Developments", Working Paper Series in Economics and Finance, Stockholm School of Economics.

Varian, H., (1974), "A Bayesian Approach to Real Estate Assessment," in S.E. Fienberg and A. Zellner (eds.), Studies in Bayesian Econometrics and Statistics in Honor of L.F. Savage, 195-208. Amsterdan: Noth-Holland.

Watson M. (2005), Comentario sobre “What’s Real about the Business Cycle”, Federal Reserve Bank of St Louis Review, July/August 2005.

West, K., H.J. Edison, and D. Cho, (1993), “A Utility Based Evaluation of Some Models of Exchange Rate Variability," *Journal of International Economics*, 35, 23-46.

Zellner, A., (1986), "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," Journal of Forecasting, 8, 446-451.

Anexo 1

Medidas de evaluación de pronóstico

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}$$

$$RMSPE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|(\hat{y}_t - y_t)\right|$$

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right|$$

donde:

T : es el número de observaciones consideradas
$\hat{y}_t$: valor estimado por el modelo