Evaluación asimétrica de una red neuronal artificial:
Aplicación al caso de la inflación en Colombia

Maria Clara Aristizábal Restrepo*

Resumen

El objetivo de este trabajo es explorar la relación no lineal entre el dinero y la inflación en Colombia a través de una red neuronal artificial (RNA), utilizando información mensual de la variación del IPC y del agregado monetario M3, desde enero de 1982 hasta febrero de 2005.

La Constitución de 1991 le otorgo al Banco de la República la responsabilidad de velar por la estabilidad de precios. Este hecho, sumado al rezago con el que las políticas monetarias afectan a su variable objetivo, en este caso la inflación, hace indispensable para las autoridades monetarias, contar con los mejores modelos para pronosticarla y guiar sus decisiones de política. Las RNA aparecen como una excelente alternativa para lograr este propósito, dado el comportamiento intrínsecamente no lineal exhibido por la relación entre estas variables.

El presente trabajo incorpora algunas innovaciones en la modelación de dinero e inflación, que permiten generar pronósticos más confiables, debido a que el modelo se aproxima con mayor exactitud a la realidad. Tales innovaciones se refieren a una selección mas sofisticada de los rezagos significativos que deben ser incorporados en el modelo, una construcción de pronósticos que actualiza su base de datos y una función de costos asimétricos para su evaluación.

Palabras Clave:

Red Neuronal Artificial, No linealidad, Unidad Escondida, Función de Activación, *Rolling* de Pronósticos, Función de Perdida Asimétrica.

* Trabajo realizado para optar por el título de Economista en la Universidad Eafit, durante la pasantía en el Banco de la República. Se agradece especialmente a Martha Misas A. su valiosa dirección. Igualmente, se agradecen los comentarios y sugerencias de Enrique López E. y Munir Jalil B. Se agradece a Martha Misas A. por el código en SAS sobre Redes Neuronales y a Munir Jalil B. por el código en SAS sobre evaluación de pronóstico. Los resultados, errores y omisiones son responsabilidad exclusiva de la autora.

1. Introducción

Las redes neuronales artificiales (ANN)[1] son modelos computacionales que tratan de replicar, de manera simplificada, el complejo funcionamiento del cerebro humano. Su capacidad de aprendizaje a través de ensayos repetidos, las ha hecho muy populares en una amplia variedad de aplicaciones en todas las ciencias. Su reciente implementación en economía se debe al hecho de que en las series económicas y financieras es más probable que aparezcan relaciones no lineales que lineales (Granger, 1991)[2] como las exigidas por los modelos econométricos tradicionales. Las ANN han demostrado ser una herramienta muy útil por su sorprendente habilidad para capturar relaciones no lineales entre variables. De hecho, pueden aproximar cualquier función no lineal si son correctamente especificadas (Tkacz y Hu, 1999).

El objetivo de este trabajo es explorar la relación no lineal entre el dinero y la inflación en Colombia a través de una red neuronal artificial, utilizando información mensual de la variación del IPC y del agregado monetario M3, desde enero de 1982 hasta febrero de 2005.

La Constitución de 1991 le otorgo al Banco de la República la responsabilidad de velar por la estabilidad de precios. Este hecho, sumado al rezago con el que las políticas monetarias afectan a su variable objetivo, en este caso la inflación, hace indispensable para las autoridades monetarias, contar con los mejores modelos para pronosticarla y guiar sus decisiones de política. Las ANN aparecen como una excelente alternativa para lograr este propósito, dado el comportamiento intrínsecamente no lineal exhibido por la relación entre estas variables.

Tradicionalmente la meta intermedia de la política económica fueron los agregados monetarios, pero los cambios estructurales experimentados por la economía colombiana a principios de los noventas cuando la Junta Directiva del Banco de la República fue

[1] Artificial Neural Networks.
[2] Citando a Shachmurove, 2000.

instituida como autoridad monetaria, cambiaria y crediticia, dificultaron enormemente el manejo monetario (Hernández y Tolosa, 2001) y condujeron a discusiones acerca de si era preferible continuar con este esquema o guiar la política monetaria a través de la tasa de interés de intervención del Banco. Sin embargo, dada la larga historia entre la inflación y el dinero, esta aproximación a través de redes neuronales, contará con los valores pasados de la inflación misma, como variables explicativas, así como con la historia del agregado monetario M3. Trabajos posteriores podrían además explorar la relación entre inflación y tasa de interés.

De acuerdo con Tkacz (2000) no puede justificarse el supuesto de linealidad, si se cree que los efectos de la política monetaria sobre la inflación son asimétricos. Un enfoque lineal implica que cambios incrementales en la cantidad de dinero tienen el mismo impacto sobre la inflación, independientemente de las cantidades de dinero iniciales. Las asimetrías tienen lugar cuando un estímulo positivo de política tiene un menor impacto sobre la economía, que un estímulo negativo. De esta forma opera precisamente la relación entre dinero e inflación.

Especificaciones no lineales de la inflación en Colombia han sido documentadas en el pasado. Melo y Misas (1998) explicaron el proceso inflacionario como un modelo switching con tres estados, Jalil y Tobón (1999) como un proceso GARCH, Arango y González (1999) y Jalil y Melo (1999) como un proceso Autorregresivo de Transición Suave STAR. Estos últimos incluyen agregados monetarios para explicar el comportamiento de la inflación.

Más recientemente Misas et al (2002) modelaron la relación entre dinero e inflación utilizando un modelo de redes neuronales por su capacidad para capturar las no linealidades entre estas dos variables y por lo tanto generar pronósticos mas precisos de la inflación.

El presente trabajo incorpora algunas innovaciones en la modelación de dinero e inflación, que permiten generar pronósticos más confiables, debido a que el modelo se

aproxima con mayor exactitud a la realidad. Tales innovaciones se refieren a una selección mas sofisticada de los rezagos significativos que deben ser incorporados en el modelo, una construcción de pronósticos que actualiza su base de datos y una función de costos asimétricos para su evaluación. El trabajo de Jalil y Misas (2005) sobre el tipo de cambio es el primero en generar pronósticos mediante un mecanismo de *rolling* y evaluarlos a través de funciones de pérdida asimétrica.

Generalmente las redes neuronales artificiales minimizan una suma de residuales al cuadrado, tanto para su estimación como para la evaluación de sus pronósticos por dentro y fuera de muestra. Sin embargo, Crone (2002) afirma que las aplicaciones han mostrado que los problemas de pronósticos requieren de medidas alternativas del error y por lo tanto para su evaluación se minimizará una función de costos asimétrica que no penalice de igual forma cuando el pronóstico se ubique por encima o por debajo del dato observado, como ocurre en la realidad. Esto sucede porque para la autoridad monetaria resulta mucho más costoso en términos de credibilidad cuando dentro de su esquema de inflación objetivo anuncia una meta inferior a la que posteriormente se registra, que cuando lo contrario ocurre.

Este documento se compone de 5 secciones principales incluyendo esta introducción. En la siguiente sección se justifica el uso de redes neuronales, dado el comportamiento no lineal de la relación entre distintas variables económicas, haciendo particular énfasis en la relación no lineal entre dinero e inflación y en alguna evidencia empírica que sugiera el comportamiento no lineal de la inflación en Colombia. La tercera sección corresponde a una aproximación a las redes neuronales propiamente, a su relación con las redes neuronales biológicas, su arquitectura, estimación y aplicaciones en distintas disciplinas. La cuarta presenta una aplicación de redes neuronales al caso de la inflación en Colombia. La última sección concluye.

2. No linealidad

La creciente popularidad de las ANN en el campo de la economía y las finanzas, se debe a la presencia de comportamientos no lineales en una gran cantidad de relaciones entre variables económicas y financieras, lo que exige tratamientos econométricos distintos a los tradicionales, que sean capaces de capturar adecuadamente las trayectorias no lineales de dichas relaciones.

2.1 Comportamiento no lineal entre variables económicas

Existe una gran variedad de estudios que respaldan la modelación no lineal de series de tiempo económicas. Típicamente se ha considerado que los ciclos económicos exhiben características no lineales, que se evidencian en las diferencias entre una transición desde una expansión hacia una recesión y viceversa. El hecho de que la producción tienda a expandirse lentamente y contraerse rápidamente puede asociarse a dos fuentes. Una de ellas es que la entrada a una industria es más costosa que la salida y la otra es la dificultad que para una firma representa incrementar su producción cuando se encuentra trabajando a capacidad plena, mientras que reducir su producción, cuando las órdenes se reducen, le resulta relativamente fácil.

Trabajos como los de Teräsvirta y Anderson (1992) han documentado este fenómeno. En particular rechazan el supuesto de linealidad para la mayoría de los índices de producción de 13 países y Europa y asumen que si una serie de tiempo no es lineal, entonces puede describirse adecuadamente por un modelo STAR[3], que refleje las respuestas de la producción a choques negativos considerables, de precios del petróleo, por ejemplo. Este tipo de modelos permite que el indicador del ciclo económico alterne suavemente entre dos regímenes distintos que representan dos fases diferentes del ciclo, dando lugar a un continuum de estados entre los dos regímenes extremos.

[3] Por sus siglas en ingles Smooth Transition Autoregressive. Supone una transición gradual entre los distintos regímenes o estados a través de una función de transición continua que cambia suavemente desde 0 hasta 1.

Arango y Melo (2001) probaron la hipótesis de fluctuaciones asimétricas de la actividad económica para algunos países latinoamericanos y encontraron evidencias a favor de un comportamiento asimétrico no lineal tipo STAR para Brasil, Colombia y México. A su vez, las funciones de impulso respuesta FIR[4] mostraron respuestas asimétricas de acuerdo con el signo del choque y el régimen en el que éste ocurrió.

Uno de los primeros trabajos en examinar si la relación entre la tasa de interés e inflación podía mejorarse usando un modelo no lineal fue Tkacz (1999). En esta oportunidad empleó un modelo de cambio discreto entre los regímenes. Posteriormente, Tkacz (2000) extendió este estudio, estimando modelos no paramétricos[5] y de redes neuronales para capturar no linealidades entre los cambios en la inflación y el *spread* (diferencial) de las tasas de interés de largo y corto plazo. Un incremento, por parte de la autoridad monetaria, de la tasa de interés de corto plazo, conduce a un *spread* negativo y viceversa. Las curvas de ambos modelos coinciden en una porción inclinada positivamente para los niveles negativos del *spread* y un tramo plano para los valores positivos, sugiriendo que los *spreads* negativos tienen un impacto marginal mayor sobre la inflación que los *spreads* positivos. La explicación de Friedman (1968) para este fenómeno es que cuando hay un endurecimiento de la política monetaria (la tasa de interés de corto plazo se incrementa) aumenta el costo de endeudarse para financiar proyectos de inversión, por lo que éstos se retrasan o incluso se posponen indefinidamente. Por el contrario, cuando hay un ablandamiento de la política monetaria, no existen incentivos inmediatos para que los individuos decidan incrementar sus niveles de consumo o de inversión. Desde la perspectiva de la autoridad monetaria, esto implica que cuando la política ya es contraccionista, una nueva contracción resultaría en una reducción marginalmente mayor a la inflación. Un endurecimiento similar tendía un menor impacto sobre la inflación, si lo tiene, si se implementa durante un régimen de política expansionista.

Franses y van Dijk (1999) muestran cómo los retornos de los activos financieros también presentan comportamientos erráticos. Ellos observan que las colas de las distribuciones

[4] Esta función describe el efecto en el tiempo de un choque sobre una serie. Se calcula como la diferencia entre el valor esperado condicional de la serie con y sin choque.
[5] No imponen alguna forma funcional sobre los datos.

de las series económicas y financieras son más gruesas que las de una distribución normal, lo que implica que los valores extremos ocurren mucho más a menudo de lo que se esperaría de una distribución normal. Adicionalmente observan que los retornos presentan un sesgo negativo y por lo tanto la cola izquierda de sus distribuciones es más gruesa que la cola derecha. Esto implica que grandes retornos negativos ocurren con mayor frecuencia que grandes retornos positivos. Al igual que Cao y Tsay (1992) también señalan que los valores extremos tienden a ocurrir en grupos, implicando que las series de volatilidad evolucionan en forma no lineal y que los períodos de retornos negativos son seguidos por períodos de alta volatilidad. Arango et al. (2000) presentan evidencia empírica sobre la relación inversa y no lineal entre los precios de las acciones del mercado de valores de Bogotá y la tasa de interés, medida por la tasa interbancaria TIB, que, de alguna forma, se encuentra afectada por la política monetaria.

De otro lado, ampliamente se acepta que los tipos de cambio son procesos *I(1)* o integrados de orden 1 y que los cambios en dichas tasas no están correlacionados en el tiempo. Por lo tanto estas series no son, generalmente, linealmente predecibles (Kuan y Liu, 1995). Imbs et al (1996) encuentran que el tipo de cambio real también presenta patrones no lineales debido a que los costos de transacción del proceso transitorio hacia el equilibrio de largo plazo hacen que el arbitraje no sea rentable como respuesta a pequeñas desviaciones, mientras que diferenciales considerables de precios sí inducen al arbitraje y por lo tanto a que éstos se ubiquen de nuevo en su valor de equilibrio. Este fenómeno es modelado a través de un modelo TAR[6], el cual asume que puede haber una zona de diferenciales de precios para la cual no existe una tendencia de regresar hacia la media, mientras que por fuera de esta zona los precios relativos son reversibles.

[6] Por sus siglas en inglés Threshold Autoregressive. Es un caso especial del modelo STAR que asume que el proceso sólo puede encontrarse en alguno de los regímenes extremos.

2.2 Relación no lineal entre dinero e inflación.

Un enfoque lineal implica una serie de supuestos a los que no obedece la relación entre dinero e inflación. Por ejemplo, las funciones de impulso respuesta derivadas de este análisis, son simétricas, implicando que un choque monetario positivo y uno negativo de igual magnitud, conducirán a efectos idénticos, pero con signo opuesto. Adicionalmente, son lineales, de tal forma que los efectos serán siempre proporcionales a la magnitud del choque y finalmente, son independientes del momento en el que éste ocurre, es decir que les es indiferente si el choque ocurre en un momento de baja o elevada inflación.

Claramente, estos tres rasgos van en contravía de la forma en la que realmente opera la relación entre dinero e inflación. En primer lugar, los agentes económicos son menos sensibles a estímulos de política positivos que negativos, luego el valor absoluto de los efectos de dichos choques no es de la misma magnitud. Además, las repercusiones de un choque monetario se encuentran estrechamente asociadas a las condiciones y ambiente inflacionario del momento en el que ocurren. Por ejemplo, cuando la inflación es más alta, los choques monetarios afectan el nivel de precios más que proporcionalmente.

Investigaciones como las de Dabús y Tohme (2003) encuentran una relación convexa entre los dos variables, indicando que la magnitud del efecto depende del nivel de inflación al momento del choque. Así, una expansión de dinero conduce a un mayor efecto en períodos de alta inflación, puesto que la alta volatilidad e incertidumbre sobre el futuro desenvolvimiento de ésta y otras variables económicas decisivas, ya han alterado bastante las expectativas de los agentes llevándolos a asumir comportamientos impredecibles e irracionales.

Una clave para explorar la posibilidad de un comportamiento no lineal entre el dinero y la inflación es observar los diferentes regímenes que exhibe un proceso inflacionario. Intuitivamente, la diversidad de efectos de los choques monetarios en distintos ambientes inflacionarios puede explicarse por las expectativas de los agentes, quienes al enfrentarse a altos niveles de inflación, que implican mayor inestabilidad macroeconómica y baja

capacidad de predicción de la economía, suelen responder de manera errática e impredecible para protegerse de ella.

Siguiendo a Franses y van Dijk (1999) un acercamiento natural a la modelación de series de tiempo económicas con modelos no lineales parece definir distintos estados o regímenes y permitir la posibilidad de que el comportamiento dinámico de las variables económicas dependa del régimen que ocurre en un momento determinado. Por comportamiento dinámico dependiente del estado se quiere decir que algunas propiedades de las series de tiempo, como su media, su varianza y autocorrelación son distintas en diferentes regímenes.

Dabús y Tohme (2003) exploraron la hipótesis de existencia de no linealidades en la relación entre dinero e inflación para distintos niveles de ésta. En efecto, en un estudio para Argentina, encontraron evidencia de que la inflación y la oferta de dinero exhiben un comportamiento no lineal, puesto que los choques monetarios afectaron el nivel de precios más que proporcionalmente cuando los niveles de inflación fueron elevados. En particular, ellos dividieron el período muestral en cuatro regímenes: inflación moderada, alta, muy alta e hiperinflación y para la mayor parte de los períodos de inflación moderada y alta, la oferta de dinero varió más que la inflación, mientras que lo contrario se observó para los períodos de muy alta e hiperinflación. Estos resultados indican una relación convexa entre el valor promedio de la serie de inflación y la variación de la oferta de dinero, es decir, que un choque monetario similar puede provocar un mayor efecto cuando los niveles de inflación son altos e inducir a una respuesta inflacionaria más que proporcional, mientras que cuando los niveles son bajos, el efecto es más pequeño. Adicionalmente el coeficiente de correlación positivo y significativo para altos niveles de inflación, es un argumento más, a favor de la existencia de un comportamiento no lineal entre ambas variables.

En Colombia, especificaciones no lineales de la inflación han sido abordadas directamente desde distintas aproximaciones. Melo y Misas (1998) explicaron el proceso inflacionario como un modelo *switching* de Hamilton con tres estados, que parte de un

modelo autorregresivo y de unos valores iniciales para las medias, las varianzas y las probabilidades de transición asociadas a cada uno de los regímenes. La inestabilidad de los parámetros asociados a la muestra completa y la estabilidad de los parámetros de las pruebas asociadas a cada una de las submuestras, fue un claro indicador de la existencia de distintos patrones en lo referente al nivel y volatilidad de la inflación trimestral, lo que justificó la aplicación de técnicas econométricas que consideraran cambios de régimen.

Por su parte, Jalil y Tobón (1999) la modelaron como un proceso GARCH [7]. Su propósito era encontrar una medida para Colombia de la incertidumbre inflacionaria aproximada por la varianza de la inflación. Los autores verificaron empíricamente para el caso colombiano dos hipótesis comunes en la literatura. La primera de ellas es que una inflación alta precede una mayor incertidumbre inflacionaria. La segunda, que la incertidumbre inflacionaria afecta el nivel de inflación.

Como un proceso Autorregresivo de Transición Suave STAR, fue explicando por Arango y González (1999) y Jalil y Melo (1999) con la diferencia de que estos últimos incluyeron agregados monetarios para explicar el comportamiento de la inflación. Este tipo de modelos supone que el proceso generador de la serie oscila de forma suave entre dos regímenes extremos a través de una función de transición, a diferencia del modelo TAR que da lugar a un cambio abrupto desde un régimen a otro.

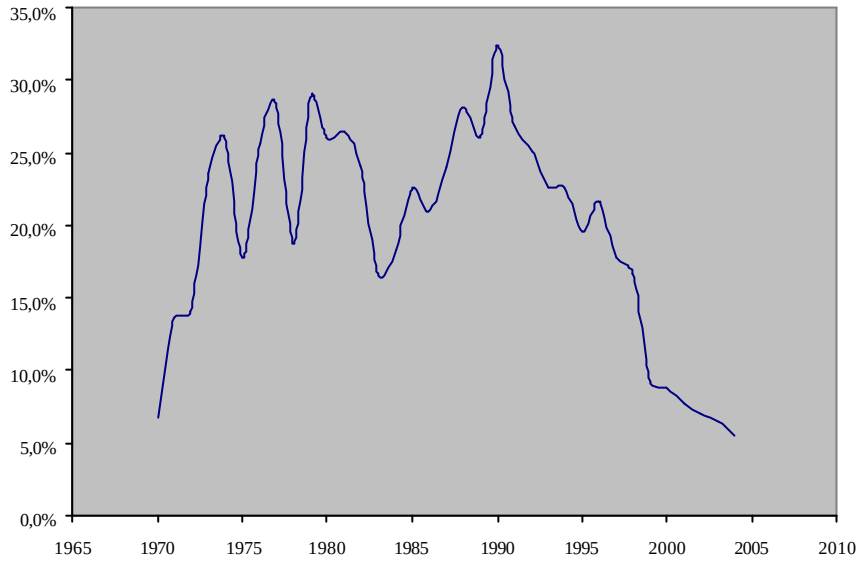
2.3 Alguna evidencia empírica sobre el comportamiento no lineal de la inflación en Colombia

El proceso inflacionario colombiano ha sido un caso muy particular de inflaciones moderadas y persistentes. En los setentas y ochentas, Colombia constituía un ejemplo de estabilidad en el contexto latinoamericano de hiperinflaciones.

[7] Por sus siglas en inglés Generalized Autoregressive Condicional Heteroskedasticy. Permite que la varianza condicional de una serie cambie con el tiempo.

En el gráfico 1 puede observarse cómo la inflación en Colombia durante los años setentas y ochentas se caracterizó por alcanzar rápidamente niveles moderadamente altos, mientras que niveles bajos fueron más lentos y difíciles de obtener. De acuerdo con Arango y González (1999) lo anterior es un claro indicador de las asimetrías intrínsecas al proceso inflacionario colombiano.

Gráfico 1: Crecimiento Anual del IPC Total
1970 - 2005



Sin embargo, la importancia concedida a nivel internacional, durante la década de los noventas, a la estabilidad de precios y la adopción de estrictas medidas monetarias, condujeron a las economías de los países en desarrollo a inflaciones de un solo dígito. En ese nuevo contexto, comparada con economías similares, la inflación en Colombia aparecía relativamente alta.

Los altísimos costos económicos y sociales asociados a la inflación y un consenso a nivel internacional alrededor de privilegiar la estabilidad de precios, entre otras razones, condujeron a que la constitución de 1991 le otorgara al Banco de la República la responsabilidad de velar por la estabilidad de precios.

Tales costos se refieren, entre otros, a que la inflación constituye un impuesto sobre los saldos nominales en poder de individuos y empresas, que afecta en particular a aquellos agentes que no pueden reajustar rápidamente sus contratos nominales. Adicionalmente, las altas tasas de inflación y en particular su variabilidad, reducen la capacidad de predicción de la economía y obligan a los agentes a invertir recursos para protegerse de ella. Tal incertidumbre se traduce, finalmente, en costos en el crecimiento del largo plazo.

La prioridad que para la autoridad monetaria tiene el velar por mantener la inflación en niveles bajos a través de la adopción del programa de "inflación objetivo", sumado a la dificultad de modelar la relación entre dinero e inflación con las técnicas convencionales, dado su carácter no lineal, justifican la exploración de alternativas de modelación exitosas como las son las redes neuronales artificiales. En la cuarta sección se presentan los resultados de un test de portmanteau para no linealidad, basado en redes neuronales artificiales, desarrollado por White y Lee (1989) y White y Granger (1993).

3. Redes Neuronales Artificiales.

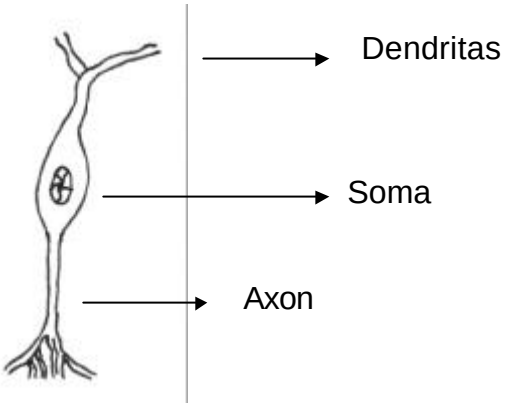
Las redes neuronales artificiales son sistemas de procesamiento de información, desarrolladas por científicos cognitivos con el propósito de entender el sistema nervioso biológico e imitar los métodos computacionales del cerebro (Shachmurove, 2002) y su impresionante habilidad para reconocer patrones (Tkacz, 1999).

3.1. Relación entre las redes neuronales biológicas y las redes neuronales artificiales.

El elemento funcional básico del cerebro es la neurona. La neurona, a su vez, está conformada por un cuerpo o soma, unas dendritas y un axón. Cada neurona recibe estímulos eléctricos de otras neuronas a través de las dendritas. En el soma se lleva a cabo la integración de toda la información obtenida en las dendritas. Estos estímulos son amplificados o disminuidos durante la sinapsis y luego sumados. Finalmente, si la suma de todos los estímulos es mayor que el umbral de resistencia máximo de la neurona,

entonces el axón transmite a otras células el mensaje resultante de la integración. Estas conexiones sinápticas, cuya intensidad es variable, se usan para enviar mensajes entre neuronas. Las neuronas coleccionan la información y aprenden patrones al reforzar sus conexiones.

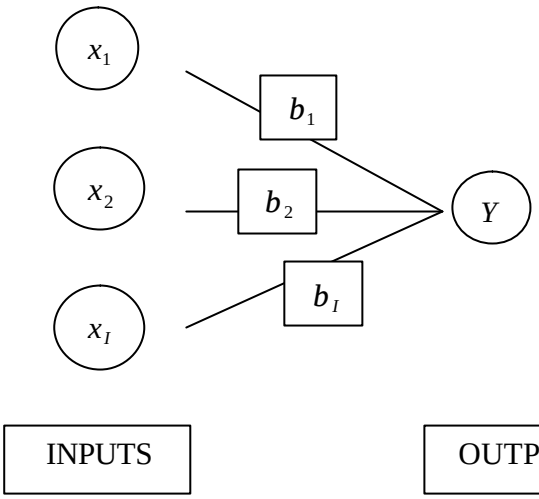
Figura 1: Neurona Biológica



Las redes neuronales artificiales se inspiran en la estructura y funciones de las neuronas biológicas. Una red neuronal artificial es esencialmente una colección de neuronas interconectadas, agrupadas en capas. Haciendo un paralelo con el esquema recién descrito de procesamiento del cerebro, la neurona artificial recibe distintos valores de entrada –*inputs*- que son multiplicados por una ponderación. En el escenario más simple, estos productos son sumados para obtener un valor de salida – *output* - . La forma más básica de red neuronal se encuentra estrechamente vinculada con las técnicas econométricas de regresión estándar. Este tipo de red simplificada posee dos capas, una de *inputs* y otra de *output.* La figura 2 ilustra la representación gráfica estándar de una red neuronal *feedforward* (alimentada hacia delante, es decir que la información fluye desde los *inputs* hacia el *output.*)

Cada neurona está representada por un círculo y las flechas indican conexiones entre ellas. El *output* y_t y los *inputs* $x_{1,}, x_2, ..., x_I$ son vectores de *n x 1* donde *n* es el número de observaciones.

Figura 2: Red Neuronal Artificial Simple



Cada conexión entre un *input* y un *output* está caracterizada por un peso b_i que expresa la importancia relativa de un *input* particular en el cálculo del *output*. Para calcular el valor del *output* en el momento *t*, la neurona *output* colecciona los valores de cada neurona *input* en la observación *t* y multiplica cada uno de ellos por un peso asociado con la conexión relevante. A continuación se suman estos productos y se obtiene

$$y_t = b_1 x_1 + b_2 x_2 + ... + b_I x_I = \sum_{i=1}^{I} b_i x_i \qquad (1)$$

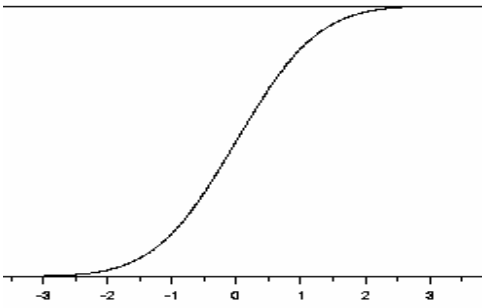
La ecuación 1 indica que *y* es una suma ponderada de x_i, donde cada x_i (las neuronas *input*) se vincula con *y* (la neurona *output*) por los parámetros b_i (las ponderaciones). En este modelo lineal, cuando x_i cambia en una unidad, *y* cambia b_i unidades.

La neurona *output* luego procesa este valor usando una función de activación. En la forma más simple de la red neuronal, la función de activación es la identidad. En este caso, el valor dador en (1) constituiría el *output* final de la red para la observación en *t*. En sus cálculos, la red tratará de reproducir el valor del *output*, dados los valores de los *inputs*.

Ahora, si se cree que existen asimetrías entre los *inputs* o variables de política y el *output*, es decir que la relación entre estas variables depende de la magnitud y la dirección de los *inputs*, entonces (1) debe generalizarse con la introducción de no linealidades en la relación. Esto puede lograrse incorporando una función de umbral, que permita que una suma ponderada de los *inputs* suficientemente grande active un cambio de régimen discreto. Sin embargo, el cambio de régimen no tiene que ser abrupto y para ello se emplean funciones de activación suaves, tales como la función logística

$$G(z) = \frac{1}{1+\exp(-z)} \qquad (2)$$

Figura 3: Función Logística.



La función (2) puede aplicarse al modelo lineal en (1) para permitir una relación no lineal entre los *inputs* y el *output*. Si además se cree que el efecto de los *inputs* sobre el *output* no es directo, como usualmente ocurre en las relaciones económicas, sino que existen variables intermedias que operan entre ellas; entonces el uso de unidades escondidas como etapas intermedias donde los *inputs* x_i y sus pesos son sometidos a una nueva

ponderación antes de afectar al *output* permite que la red capture la relación no lineal entre las variables *input* y el *output*.

Existe una amplia variedad de alternativas para explicar el comportamiento de una variable y_t en función de sus propios valores pasados o de los rezagos de otras variables X_t. De acuerdo con Granger y Teräsvirta (1993) estas alternativas podrían clasificarse de acuerdo con la forma funcional mediante la cual se aproxima esta relación. Si asumen una forma funcional específica en la que usualmente deben estimarse parámetros, se trataría, naturalmente, de modelos paramétricos. Si, por el contrario, esta forma funcional no se encuentra restringida a pertenecer a ninguna clase particular de función, el modelo sería no paramétrico. En el medio se encontrarían especificaciones semiparamétricas dado que algunas variables entran de forma paramétrica y otras lo hacen no parametricamente. A su vez, los modelos paramétricos pueden clasificarse de acuerdo con el carácter de la relación entre las variables objeto de estudio.

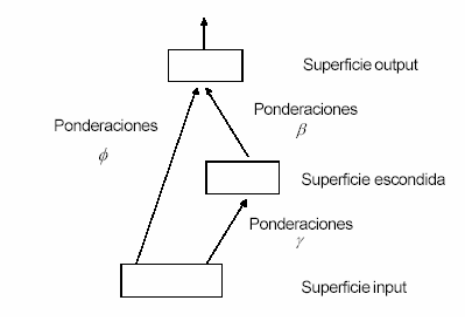
Un modelo lineal $y_t = b' X_t + e_t$ es apto para modelar relaciones en las que choques positivos y negativos de igual magnitud producen efectos idénticos, pero en direcciones opuestas. Así mismo, es adecuado para describir relaciones en las que los efectos son siempre proporcionales a la magnitud del choque y además independientes del momento en el que éste ocurre. Si, efectivamente, la relación entre las variables objeto de estudio es no lineal, el investigador se enfrenta a una amplia gama de posibilidades entre las que se encuentran las redes neuronales artificiales.

3.2. Arquitectura

Por arquitectura de una red neuronal artificial se entiende el conjunto de *inputs* incluidos en la parte no lineal, *p*, que puede estar completamente contenido en el conjunto de *k* variables asociadas a la parte lineal; así como el número de unidades escondidas, *q*, y el número de superficies escondidas, necesarias para la determinación del componente no lineal (Misas et al., 2003)

El tipo de red neuronal que se aborda en este trabajo es multiplica (posee tres superficies), se alimenta hacia delante (la información fluye desde la superficie *input* hacia la superficie *output*) y con una única superficie escondida o *"single hidden layer feedforward network"*. En la figura 4 puede observarse una representación gráfica del tipo de red neuronal descrito.

Figura 4: Representación de una red neuronal alimentada hacia adelante con una única superficie escondida o *"single layer feedforward network"*.



En la base de esta red se encuentra una superficie *input* conformada por el conjunto de variables explicativas X_t. Estas *k* variables explicativas pueden ser rezagos de la misma variable dependiente, *l*, así como de las variables exógenas, *m*, $X_t = \{y_{t-1},..., y_{t-l}, w_{1t},..., w_{mt}\}$ y relacionarse tanto lineal como no linealmente con la variable explicada. Estas últimas conforman el conjunto Z_t, donde $Z_t \subseteq X_t$.

Cada uno de estos *inputs* es multiplicado por una ponderación. Como es de esperarse, la estimación de las ponderaciones iniciales correspondientes a la parte lineal f_i, para $i = 1,...,l+m$ se lleva a cabo a través de mínimos cuadrados ordinarios y la sumatoria de los productos de estas variables por sus respectivos pesos se va directamente al *output* como lo ilustra la figura 4.

$$\hat{y}_{t(lineal)} = f_0 + f_1 y_{t-1} + ... + f_l y_{t-l} + f_{l+1} w_{1t} + ... + f_{l+m} w_{mt} \qquad (3)$$

Por su parte, los pesos asociados a la parte no lineal $g_{i,j}$ para $i=0,\ldots,p$ y $j=1,\ldots,q$ son obtenidos aleatoriamente a partir de una distribución uniforme en el intervalo $[-a,a]$. Antes de entrar a la superficie oculta, estos pesos se encargan de amplificar o disminuir el efecto de las señales originales. En la superficie oculta existen unidades escondidas que pueden estar asociadas a una variedad de funciones que permiten la transición suave o discreta desde un régimen a otro. En este trabajo se emplearon funciones logísticas como la presentada en la ecuación (2). Allí, las funciones de activación transforman las combinaciones $Z_t'g_{i,j}$ en los valores comprendidos entre cero y uno. Finalmente, estos valores son multiplicados por b_j para $j=1,\ldots,q$. Los valores iniciales de b_j se hallan, una vez más, por mínimos cuadrados ordinarios.

$$\hat{y}_{t(no\,lineal)} = \sum_{j=1}^{q} b_j G\left(g_{0,j} + g_{1,j}\, y_{t-1} + \cdots + g_{l,j}\, y_{t-l} + g_{l+1,j}\, w_{1t} + \cdots + g_{l+m,j}\, w_{mt}\right) \quad (4)$$

La suma de la parte lineal y la parte no lineal produce el *output* $y_t = \hat{y}_t + e_t$, donde

$$\hat{y}_t = f_0 + X_t' f_i + \sum_{j=1}^{q} b_j G(Z_t' g_{i,j}) \quad (5)$$

Una especificación con un número adecuado de unidades escondidas puede aproximar cualquier función no lineal con un grado arbitrario de precisión (Tkacz, 1999). Esto se conoce como la propiedad universal de aproximación de las redes neuronales y tal aproximación no tendría lugar en ausencia de la capa de unidades escondidas (White, 1992).

Siguiendo a Swason y White (1995), en la terminología de redes, los parámetros asociados con la parte no lineal, g_{ij} y b_j se conocen como ponderaciones *input to hidden layer* y *hidden layer to output*, respectivamente, mientras que los parámetros correspondientes a la parte lineal de la red, f_i, se conocen como ponderaciones *input to output.* La red adquiere conocimiento a través del conjunto de parámetros

$\Theta = \{\gamma_{ij}, \beta_j, \phi_i\}$. El vector Θ tiene *(p+1)*q + q + (k+1)* parámetros. Los primeros *(p+1)*q* elementos se relacionan con los parámetros asociados el intercepto más las *p* variables incluidas en la parte no lineal, que a su vez se hallan vinculados con las distintas unidades escondidas *q*. Los siguientes *q* elementos se refieren a las ponderaciones que van desde las *q* unidades escondidas hasta el *output.* Los restantes *(k+1)* elementos corresponden a los parámetros asociados al intercepto más las *k* variables de la parte lineal. El aprendizaje de la red consiste en el ajuste repetido de estos parámetros hasta alcanzar un nivel de convergencia deseado. Este proceso será explicado en detalle en la siguiente sección.

3.3 Estimación

De la misma manera que en el caso lineal, la estimación de parámetros en modelos intrínsecamente no lineales, se basa en la minimización o maximización de una función objetivo como la suma de errores al cuadrado o la función de verosimilitud. Las ponderaciones $\Theta = \{\gamma_{ij}, \beta_j, \phi_i\}$ de la red presentada en este trabajo, se obtienen minimizando la suma de las desviaciones al cuadrado entre el output y el pronóstico de dicha red, es decir la suma de residuales al cuadrado.

$$S(\Theta) = \sum_{t=1}^{n} [y_t - f(X_t, \Theta)]^2 \qquad (6)$$

donde $f(X_t, \Theta) = f_0 + X_t' f_i + \sum_{j=1}^{q} \beta_j G(Z_t' \gamma_{i,j}) + e_t \qquad (7)$

Precisamente, el aprendizaje de la red, se encuentra en el proceso de entrenamiento durante el cual se estiman y ajustan sucesivamente los parámetros $\Theta = \{\gamma_{ij}, \beta_j, \phi_i\}$ con el fin de minimizar el error y obtener el modelo de red neuronal que mejor capture el comportamiento de la serie bajo estudio. Para empezar, se requiere una conjetura sobre los valores iniciales de este vector. A cada iteración, este vector actualiza el conocimiento adquirido en el momento r, de acuerdo con una regla de aprendizaje adecuada $\Delta^{(r)}$.

$$\hat{\Theta}^{(r+1)} = \hat{\Theta}^{(r)} + \Delta^{(r)} \qquad (8)$$

La definición de esta regla de aprendizaje conduce a una diversidad de formas para aproximarse al problema de optimización declarado en la ecuación 6. Este problema de optimización puede resolverse mediante la aplicación de métodos de direccionamiento, cuyo objetivo es reducir un problema multidimensional a una serie de problemas unidimensionales, mediante la determinación de un conjunto de direcciones hacia las cuales moverse para llevar a cabo búsquedas lineales en cada una de ellas, con la esperanza de encontrar un punto en el cual el gradiente desaparezca $\nabla S\left(\hat{\Theta}^{(r)}\right) = 0$, es decir un punto óptimo.

Entre estos métodos se encuentran el direccionamiento genérico, direccionamiento por coordenadas o los métodos de descenso. Estos últimos generan una dirección de búsqueda $d^{(r)}$ tal que un ligero movimiento en esa dirección, desde un punto inicial $\hat{\Theta}^{(r)}$, haga decrecer el valor de la función objetivo $\nabla S\left(\hat{\Theta}^{(r)}\right)$. Del cálculo se sabe que la dirección más rápida de descenso es la negativa del gradiente.

$$d^{(r)} = -\nabla S\left(\hat{\Theta}^{(r)}\right) \qquad (9)$$

Si $d^{(r)} \neq 0$ ésta es una dirección de descenso y todavía es posible, mediante iteraciones, mejorar el valor de la función objetivo. Este tipo de métodos se conoce como *steepest descen*t (descenso más pronunciado) y entre ellos se encuentra el procedimiento más ampliamente usado para el entrenamiento supervisado[8] de redes neuronales multicapa alimentadas hacia adelante.

[8] Entrenamiento supervisado implica conocimiento sobre los valores efectivamente observados, contra los cuales es posible calcular una medida de error.

Una vez se ha obtenido la dirección del descenso, debe encontrarse el *step length* l (longitud de paso) que responda al problema de cuánto debe ser el desplazamiento en esa dirección. Esto se logra a través de una búsqueda lineal en la que se minimiza

$$f(l) = S\left(\Theta^{(r)} + l d^{(r)}\right) \qquad \text{sujeto a} \qquad l \geq 0 \tag{10}$$

Si $f(l)$ es una función convexa, la condición suficiente de optimalidad es $f'(l) = 0$, y se resuelve para l. Ahora con este valor se actualiza

$$\hat{\Theta}^{(r+1)} = \hat{\Theta}^{(r)} + l d^{(r)} \tag{11}$$

Obsérvese cómo, para este caso, la regla de aprendizaje corresponde a un movimiento de longitud l en la dirección opuesta del gradiente

$$\Delta^{(r)} = l d^{(r)} \qquad \text{donde} \qquad d^{(r)} = -\nabla S\left(\hat{\Theta}^{(r)}\right) \tag{12}$$

Recuérdese que el gradiente de una función evaluada en un punto, se define como el vector de derivadas de la función con respecto a cada una de las variables. Teniendo en cuanta que estas derivadas parciales miden el efecto que un cambio marginal de estas variables tiene sobre la función, en este caso, la suma de residuales al cuadrado $S(\Theta)$, este tipo de algoritmos, que usa información sobre el gradiente, lo que hace es mirar cómo cambia el error cuando cambia el valor de las variables. Dado que no es posible cambiar el valor de las variables, se cambia el valor de sus pesos. Cuando se aplica una regla de aprendizaje como la presentada en la ecuación 12, lo que realmente se está haciendo es tomar la información o conocimiento disponible en r, y ajustarla en la dirección contraria del impacto que sobre el error tiene dicha ponderación, con el fin de reducirlo a cada iteración.

El valor de la función objetivo evaluada en este nuevo punto $\hat{\Theta}^{(r+1)}$, será inferior que aquél evaluado en la iteración anterior $\hat{\Theta}^{(r)}$

$$S\left(\hat{\Theta}^{(r+1)}\right) < S\left(\hat{\Theta}^{(r)}\right) \qquad (13)$$

Este proceso iterativo debe continuar hasta que las condiciones de convergencia deseadas sean alcanzadas. Idealmente, el gradiente debería desaparecer por completo, pero dada la complejidad de ciertos problemas de optimización, es difícil que esta condición pueda satisfacerse y por lo tanto, en ocasiones, sólo se exige que se aproxime a cero.

El método recién descrito corresponde al algoritmo de *backpropagation* (propagación hacia atrás). Su nombre se deriva del hecho de que las señales de error son propagadas hacia atrás, a través de la red, capa por capa.

Este algoritmo fue popularizado en 1986 por Rumelhart, Hinton y Williams, aunque se trata realmente de una sofisticada aplicación de la regla de la cadena del cálculo elemental de Werbos (1974).

Utilizando una analogía biológica, las ponderaciones representan un estado de memoria, el "mejor cálculo" de cómo hacer predicciones a partir de los resultados de los nodos. Una vez que la entrada de *inputs* se procesa a través del sistema, puede compararse con el valor del resultado efectivo (aprendizaje supervisado). Los valores resultantes y efectivos se comparan. Si existe alguna diferencia entre los dos valores (parecida a un valor residual) entonces sería deseable ajustar el modelo con la esperanza de mejorarlo. Una vez calculado el error en el valor del *output*, éste es distribuido hacia atrás en el sistema. Como funciona por su vía a través de nodos, las ponderaciones cambian proporcionalmente, aumentando o disminuyendo dependiendo de la dirección del error.

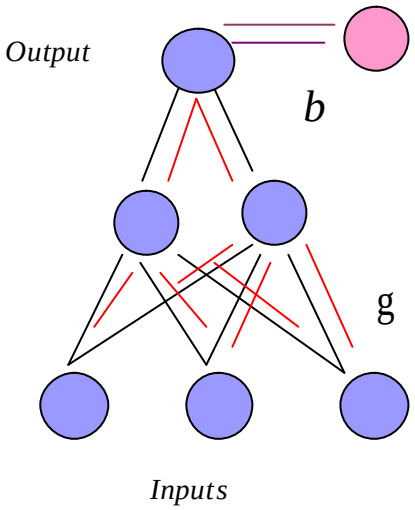
En la fase de preparación, el objetivo es procesar un gran número de casos a través de la red neuronal, de tal forma que ésta pueda hacer las mejores predicciones para todas las pautas de entrada de datos.

La figura 5 ilustra los dos pasos de cómputo del entrenamiento de *backpropagation.*

- El *forward pass* (pase hacia adelante): En éste, la red se alimenta de los *inputs* y produce un *output*. Durante esta etapa los pesos sinápticos son fijos.

- El *backward pass* (pase hacia atrás): Los pesos sinápticos son todos ajustados de acuerdo con la señal de error, la cual es propagada hacia atrás por toda la red en dirección opuesta a las conexiones sinápticas.

Figura 5: Representación del Algoritmo de *Backpropagation*



En general, en los métodos de descenso, la búsqueda de la dirección viene dada por

$$d^{(r)} = -A\left(\hat{\Theta}^{(r)}\right)^{-1} \nabla S\left(\hat{\Theta}^{(r)}\right) \qquad (14)$$

donde $A\left(\hat{\Theta}^{(r)}\right)$ es una matriz de dirección.

Obsérvese que si $A\left(\hat{\Theta}^{(r)}\right)$ es igual a la matriz idéntica, la ecuación (14) converge a (9) y se tiene el método de descenso más pronunciado. El algoritmo de *backpropagation* hace parte de este tipo de algoritmos. Como ya se anotó, este método es tan sólo un caso especial de los métodos de descenso, que a su vez son un tipo de métodos de direccionamiento.

La mayoría de estos algoritmos tiene la forma

$$\hat{\Theta}^{(r+1)} = \hat{\Theta}^{(r)} - lA\left(\hat{\Theta}^{(r)}\right)^{-1} \cdot \nabla S\left(\hat{\Theta}^{(r)}\right) \qquad (15)$$

El rasgo que diferencia los algoritmos alternativos es la definición de $A\left(\hat{\Theta}^{(r)}\right)$. Los métodos de Newton hacen $A\left(\hat{\Theta}^{(r)}\right)$ igual a la matriz Hessiana, es decir que computan derivadas de segundo orden, y luego proceden en dirección descendiente para localizar un mínimo después de un número de iteraciones.

$$A\left(\hat{\Theta}^{(r)}\right) = \left[\nabla^2 S\left(\hat{\Theta}^{(r)}\right)\right] = H\left(\hat{\Theta}^{(r)}\right) \qquad (16)$$

Dado que el cálculo numérico de la matriz Hessiana es computacionalmente muy costoso, incluso en problemas de tamaño moderado, y que, adicionalmente, la dirección de la búsqueda requiere que esta matriz sea invertible, los métodos Quasi-Newton tienen como punto inicial a una matriz simétrica definida positiva, por ejemplo la matriz identidad y a partir de la información sobre la función $S\left(\hat{\Theta}^{(r)}\right)$ y el gradiente $\nabla S\left(\hat{\Theta}^{(r)}\right)$ construyen, a cada iteración, información sobre la curvatura $\nabla^2 S\left(\hat{\Theta}^{(r)}\right)$ y hacen una aproximación de la inversa de la matriz Hessiana usando una técnica de actualización apropiada.

Uno de los primeros esquemas para construir la inversa de la matriz Hessiana fue el DFP, propuesto originalmente por Davidon (1959) y posteriormente desarrollado por Fletcher y Powell (1963). Experimentos numéricos han mostrado que el desempeño de la fórmula de Broyden, Fletcher, Goldfarb y Shanno (BFGS) es superior, lo que la ha hecho particularmente popular en el trabajo de redes neuronales. En este trabajo se empleó la subrutina NLPQN[9] implementada en el paquete estadístico SAS. Esta subrutina permite la especificación de distintas fórmulas de actualización, entre ellas BFGS, presentada en la siguiente ecuación.

$$H_{k+1} = H_k + \frac{q_k q_k^T}{q_k^T s_k} - \frac{H_k^T s_k^T s_k H_k}{s_k^T H_k s_k} \tag{17}$$

donde

$$s_k = \Theta^{(r+1)} - \Theta^{(r)} \tag{18}$$

y

$$q_k = \nabla S\left(\Theta^{(r+1)}\right) - \nabla S\left(\Theta^{(r)}\right) \tag{19}$$

Dado que la suma de residuales al cuadrado $S(\Theta)$ puede poseer numerosos mínimos locales, Franses y van Dijk (1999) sugieren el uso de distintos valores iniciales para el vector de parámetros Θ y elegir aquellos que conducen al menor valor de $S(\Theta)$ con el fin de mejorar las posibilidades de encontrar un mínimo global. Adicionalmente proponen otros métodos para mejorar las propiedades numéricas de os estimadores, entre los que se encuentran reescalar las variables y_t y X_t de tal forma que tengan media cero y desviación estándar igual a uno, así como evitar que los estimadores asuman valores demasiado grandes. Esto podría lograrse aumentando la función objetivo en (6) con un término penalizador conocido como weight decay. La función objetivo a minimizar sería entonces

[9] Nonlinear Optimization by Quasi-Newton Method.

$$S(\Theta)=\sum_{t=1}^{n}\left[y_t - f\left(X_t,\Theta\right)\right]^2 + r_f\sum_{i=0}^{k} f_i^2 + r_b\sum_{j=0}^{q} b_j^2 + r_g\sum_{j=0}^{q}\sum_{i=1}^{p} g_{ij}^2 \qquad (20)$$

donde r_f, r_b y r_k deben ser especificados[10]

3.4 Aplicaciones

Las redes neuronales artificiales han sido ampliamente usadas en una variedad de disciplinas. Sus aplicaciones van desde convertir texto escrito a voz (Sejnowsky y Rosenberg, 1986), reconocer caracteres escritos a mano (LeCun et al., 1990), jugar Backgammon (Tesauro, 1989), tocar música (Brecht y Aiken, 1995)[11] hasta pronosticar el tiempo de supervivencia de pacientes enfermos.

Las aplicaciones más comunes en medicina clínica corresponden al diagnóstico de enfermedades. El trabajo de Ravdin et al. fue uno de los primeros estudios en usar las redes neuronales para el análisis de supervivencia y producir estimadores precisos para pacientes con cáncer. Ohno-Machado et al. estimaron el tiempo de supervivencia de pacientes infectados con sida. La información con la que entrenaron la red corresponde al seguimiento de individuos por tantos intervalos como categorías de *output* tiene el modelo. En su modelo los nodos *output* corresponden a la probabilidad de que un individuo muera durante el primer intervalo, el segundo y así sucesivamente, siendo estos intervalos mutuamente excluyentes. Algunas variaciones de este modelo se refieren a la predicción de la supervivencia acumulada, es decir, si un paciente determinado estará muerto después de un intervalo de tiempo dado y la supervivencia condicional, es decir, la probabilidad de que cierto individuo que ha sobrevivido hasta cierta fecha seguirá vivo en el siguiente intervalo.

[10] En este trabajo se siguen las recomendaciones de Franses y van Dijk (1999) y se establece $r_f = 0{,}01$, $r_b = r_g = 0{,}0001$.

[11] Citando a González (2000)

Otra aplicación reciente de las redes neuronales, se refiere a un problema que desde la antigüedad ha atraído el interés de los científicos: el pronóstico del flujo de los ríos. Este problema ha sido atacado con técnicas lineales como modelos AR, ARMAX y Filtros de Kalman y hasta hace muy poco se han venido explotando el potencial de las redes neuronales en este campo. Particularmente Atiya et al. (1999) pronosticaron el flujo del río Nilo. Egipto depende casi exclusivamente del este río para la irrigación agrícola. Su flujo está lejos de ser estable y exhibe un comportamiento estacional, bajo durante los meses de invierno y alto en Agosto y Septiembre. La represa de Aswan retiene el agua que llega y la libera de una forma más uniforme para cubrir de manera óptima las necesidades agrícolas y de generación de electricidad, luego el pronóstico del flujo del río ha ayudado a determinar la cantidad óptima de agua a liberar y por lo tanto a manejarla de manera más eficiente.

Las anteriores hacen parte de un sinnúmero de aplicaciones en las que las redes neuronales han demostrado su increíble capacidad para capturar comportamientos atípicos durante su fase de entrenamiento y generar pronósticos acertados al generalizar el conocimiento adquirido por fuera de dicho conjunto de información.

4. Aplicación de redes neuronales al caso de la inflación en Colombia[12]

La información utilizada en este trabajo corresponde a datos mensuales desde enero de 1982 hasta febrero de 2005 de la primera diferencia del logaritmo del Índice de Precios al consumidor (DLIPC). Como variables explicativas se consideraron 18 rezagos de la variable endógena, así como igual número de rezagos de la variable exógena M3, definida también como la primera diferencia del logaritmo de esta serie (DLM3). Ambas series fueron normalizadas (NDLIPC y NDLM3), para que tuvieran media cero y desviación estándar igual a uno, tal como lo sugieren Franses y van Dijk (1999), con el propósito de mejorar las propiedades numéricas de los estimadores.

[12] Para esta aplicación se emplearon 5 códigos programados en SAS (ver anexo 6 para una explicación detallada de cada uno de ellos). Los programas de *stepwise*, simulación, evaluación por dentro de muestra y *rolling* de pronósticos fueron desarrollados por Martha Misas A. La evaluación por fuera de muestra fue programada por Munir Jalil B. Todos los programas cuentan con pruebas de escritorio y documentación desarrolladas por la autora.

Dado que se calcularon 18 rezagos de ambas series, se contó con información completa a partir de agosto de 1983, punto en el cual empezaría la base de datos de entrenamiento de la red neuronal. Adicionalmente, los últimos 18 datos fueron removidos de esta muestra, para un total de 241 observaciones. Lo anterior se hizo con el fin de tener información suficiente para evaluar el desempeño de cada una de las redes estimadas por fuera de muestra.

Naturalmente, el primer paso en la selección de la mejor red neuronal artificial, consiste en la elección del conjunto de variables explicativas. Éstas pueden ser rezagos de las variables endógenas como variables exógenas y/o sus rezagos. Lo usual es emplear la estrategia *stepwise* cuyos procedimientos básicos incluyen la identificación de un modelo inicial, la iteración de pasos, esto es, la alteración repetida del modelo en el paso anterior, adicionando o removiendo una variable explicativa de acuerdo con un criterio de selección[13] y la terminación de la búsqueda cuando no sea posible dar más pasos, dado el criterio o cuando el número máximo de pasos especificado haya sido alcanzado.

Particularmente, la elección del conjunto de *inputs* para esta red, es el resultado de la intersección de una serie de búsquedas que trae programadas SAS[14], como *stepwise selection*, *forward entry* o *backward removal*, que hacen uso de distintos criterios tales como el R^2 más alto, el R^2 ajustado o el estadístico de Mallow *C(p)*[15].

El procedimiento *forward* (hacia adelante) adiciona una variable explicativa en cada paso, la cual sólo es incluida en el conjunto de *inputs* si su entrada mejora los criterios de selección del modelo anterior. Si ninguna variable tiene un valor que exceda el valor crítico especificado para entrar en el modelo, entonces el proceso concluye, de lo contrario la variable con el valor más alto en la estadística de entrada, entra en el modelo.

[13] En la adopción de esta estrategia, se consideraron significancias al 5% como es tradicional.
[14] Statistical Analysis System.
[15] El criterio de información de Mallow.

El esquema *backward* (hacia atrás), en cambio, parte de un modelo que contiene todas las variables explicativas posibles y a cada paso remueve las variables que menor aporte le hacen al modelo anterior. Si ninguna variable tiene un valor que sea menor que el valor crítico para ser removida del modelo, entonces el proceso concluye, de lo contrario, la variable con el menor valor es removida del modelo.

Después de la adopción de las anteriores estrategias, a partir de la base de datos de entrenamiento, se encontró que el mejor modelo en la parte lineal debía incluir como variables explicativas a los rezagos uno, cuatro, once y doce de NDLIPC y a los rezagos dos, cuatro y trece de NDLM3, para un total de siete variables en X_t. Los resultados se presentan en el anexo 1.

Una vez se han elegido la k variables que conformaran el conjunto X_t que se relaciona de forma lineal con el *output* y_t, debe decidirse el número p óptimo de estas variables que entrarán a conformar el conjunto Z_t que se relaciona de manera no lineal con y_t. Lo anterior se logra paso a paso, incluyendo en primera instancia tan sólo a la primera variable del conjunto de *inputs* y adicionando cada vez, una variable más, hasta incluir a la totalidad de variables en X_t, es decir hasta que Z_t sea igual a X_t.

Así mismo, debe decidirse sobre el número de unidades escondidas q, que responda adecuadamente a la disyuntiva entre capturar el comportamiento no lineal entre las variables mediante un número elevado de unidades escondidas, sin que esto conduzca a un sobre ajuste del modelo que le impida hacer pronósticos acertados. Para evitar que esto ocurra se deben probar simultáneamente todas las combinaciones posibles de p y q, es decir, de variables en la componente no lineal y de unidades escondidas.

Para este trabajo se estimaron 28 arquitecturas distintas, una para cada combinación posible entre número de variables en la parte no lineal, p =1,…,7; y número de unidades escondidas, q =1,…,4; de tal forma que la primera red incluiría tan sólo a la primera variable del conjunto X_t y una unidad escondida, hasta completar 7 variables en la parte

no lineal y 4 unidades escondidas. Recuérdese que en la parte lineal siempre se considerará la totalidad de elementos del conjunto X_t.

Dado que la función que se desea minimizar puede presentar diversos mínimos locales, el hecho de que el algoritmo numérico empleado converja, no significa que se haya encontrado un mínimo global. Por lo tanto, se siguió la recomendación de Franses y van Dijk (1999) en cuanto a estimar las redes partiendo de múltiples valores iniciales del vector de parámetros Θ. En particular, cada una de las 28 redes fue estimada a partir de 30 valores iniciales[16] distintos del vector de parámetros, obtenidos de forma aleatoria a partir de una distribución uniforme en el intervalo [-2,2]. Una vez estimados los 30 vectores de parámetros, resultantes del proceso de optimización, se verificó que cada uno de ellos cumpliera con la condición de primer orden[17], para garantizar que ese vector de parámetros estimados, efectivamente condujera a un punto crítico. Si un vector de parámetros no satisfacía esta condición, entonces era rechazado. Los vectores de parámetros restantes se ordenaron de forma ascendente de acuerdo con el valor de la función objetivo evaluada en cada uno de ellos. Así se completaron los 5 mejores vectores de parámetros para cada una de las 28 arquitecturas, para un total de 140 vectores de parámetros. A continuación, cada uno de estos vectores de parámetros fue sometido a evaluaciones por dentro y por fuera de muestra. Usualmente, los trabajos de redes neuronales han partido de un solo conjunto de parámetros por cada arquitectura: aquél que al ser evaluado en la función objetivo conduce a su menor valor. La elección de los 5 mejores vectores para cada arquitectura, constituye una innovación que pretende responder al problema de modelos cuyo ajuste por dentro de muestra es el mejor, pero que presentan pobres desempeños por fuera de ella.

[16] La decisión de generar 30 vectores de parámetros iniciales se debió a que con este número de replicaciones, generalmente se logran obtener al menos 5 vectores que satisfacen la condición de primer orden.

[17] En este trabajo se exigió que los elementos de los vectores gradientes evaluados en cada uno de los vectores de parámetros resultantes del proceso de optimización, fueran menores o iguales a 0,04.

4.1 Evaluación por dentro de muestra

La evaluación de las distintas arquitecturas posibles que conduce a la elección de la mejor red neuronal, debe llevarse a cabo tanto dentro de muestra como por fuera de ella, dado que una red cuyo desempeño sea excelente dentro de muestra puede presentar problemas de pronóstico por fuera de muestra. Esto podría deberse a una sobre especificación atribuida a un número elevado de unidades escondidas que le daría una gran flexibilidad a la red, permitiéndole capturar y memorizar perfectamente el comportamiento no lineal de la serie bajo estudio, pero le impediría predecir su comportamiento futuro.

La evaluación del desempeño dentro de muestra para cada una de las arquitecturas estimadas se lleva a cabo sobre la variación anual del IPC, calculada como la primera diferencia del logaritmo del IPC en un mes dado y el logaritmo del IPC del mes correspondiente del año anterior. Recuérdese que en la estimación esta variable fue reescalada para que tuviera media cero y desviación estándar uno. En la evaluación por dentro de muestra esta variable deberá ser desnormalizada es decir que NDLIPC deberá multiplicarse por su desviación estándar y sumársele su media. El anexo 2 presenta las medidas calculadas en ésta etapa para cada uno de los vectores de parámetros Θ asociados a cada una de las arquitecturas. En el anexo 3 se presenta un cuadro con las medidas de las tres mejores arquitecturas de acuerdo con cada uno de los criterios.

A la luz de criterios como *AIC, RMSE,* y *MAPE* la arquitectura 6_4 supera a las demás y es segunda en *BIC* y *RMSPE*. Ésta tiene seis variables en la parte no lineal *(p=6)*, cuatro unidades escondidas en la capa oculta *(q=4)* y por supuesto siete variables en la parte lineal *(k=7)*. Corresponde además a la primera replicación del vector de parámetros *(w=1)*. La arquitectura 7_4, la más compleja de todas, supera a las demás cuando se observan criterios como *RMSPE*, es segunda en *RMSE* y tercera en *AIC* y *MAE*.

De los anteriores resultados es evidente que para que la red logre el mejor ajuste dentro de muestra, requiere la mayor complejidad posible, esto es, un número alto de variables

explicativas en la parte no lineal, así como el mayor número de unidades escondidas, que le permitan capturar con exactitud, el complejo comportamiento de la serie bajo estudio.

Sin embargo, como se verá en la siguiente sección, arquitecturas muy complejas le restan flexibilidad a la red neuronal para hacer pronósticos por fuera de la muestra de entrenamiento.

4.2 Evaluación por fuera de muestra

La evaluación del desempeño por fuera de muestra representa un verdadero avance en el estudio que en Colombia se ha hecho sobre redes neuronales artificiales. El trabajo de Jalil y Misas (2005) para el tipo de cambio, es el primero en generar pronósticos mediante un mecanismo de *rolling* y evaluarlos a través de funciones de pérdida asimétrica. Tradicionalmente, esta evaluación se ha llevado a cabo comparando los valores observados (que se dejaron por fuera de la muestra de entrenamiento con este propósito) frente a los valores estimados de la variable endógena a partir de los distintos conjuntos de parámetros estimados.

Sin embargo, esta metodología desconoce que la entrada sucesiva de observaciones adicionales modifica el conjunto de información sobre el cual se estimó el vector de parámetros. Por lo tanto, lo ideal en estos casos consiste en reestimar este vector de parámetros cada vez que un nuevo dato es incorporado en la base de datos inicial.

Trabajos anteriores habrían empleado el vector de parámetros $\Theta_{\_0}$, calculado a partir de la base de datos de entrenamiento, es decir con las 241 observaciones que se dejaron por dentro de muestra, para hacer los pronósticos que parten de cada una de las 18 observaciones que se dejaron por fuera de muestra, a un horizonte *h* determinado[18].

El esquema de *rolling* adoptado en este trabajo se ilustra a continuación. Nótese cómo el vector de parámetros $\Theta_{\_0}$ sólo es empleado para hacer pronósticos *h* períodos hacia

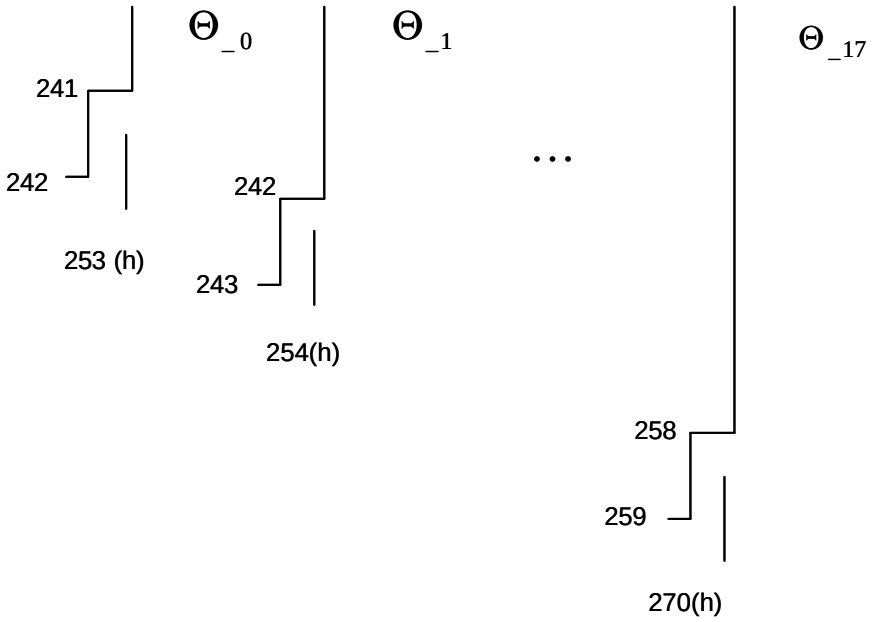
[18] El horizonte de pronóstico utilizado en este trabajo fue de 12 meses.

adelante desde la última observación que se dejó por dentro de muestra, de tal forma que los pronósticos realizados con dicho vector se extienden desde la observación 242 hasta la observación 253, como puede observarse en la figura 6.

Una vez se adiciona el dato 242 en la muestra de entrenamiento, se reestima el vector de parámetros y se encuentra a $\Theta_{\_1}$. A su vez, este nuevo vector es empleado para realizar los pronósticos desde la observación 243, *h* períodos hacia delante, hasta la observación 254.

El proceso continúa, reestimando vectores de parámetros cada vez que un nuevo dato es incorporado y proyectando *h* períodos hacia adelante a partir de cada uno de ellos. Finalmente se adiciona la penúltima observación que se dejó por fuera de muestra, se encuentra a $\Theta_{\_17}$ y con él se proyecta información desde el dato en 259 hasta el dato en 270. No tendría sentido incorporar el último dato para reestimar parámetros y con ellos hacer mas pronósticos, puesto que éstos no tendrían valores observados contra los cuales pudieran ser evaluados.

Figura 6: Esquema de *rolling* de pronósticos



Al concluir este proceso, para cada arquitectura considerada, se obtiene una matriz de 18 x 12, compuesta por 18 vectores, uno por cada conjunto de parámetros que se reestimó al incluir una observación adicional de aquellas que se dejaron por fuera de muestra; del tamaño del horizonte de pronóstico, en este caso 12.

Es decir que la primera columna contiene los pronósticos, 12 períodos hacia adelante, realizados con los parámetros calculados a partir de la muestra de entrenamiento de la red. La última columna contiene los pronósticos hechos a partir del vector de parámetros que se calculó al incorporar los datos de las observaciones que se dejaron por fuera de muestra sin incluir el último.

Una vez se han obtenido los pronósticos de cada una de las arquitecturas simuladas, debe construirse una medida del error de pronóstico de la red neuronal, para luego hallar cuál de ellas está arrojando los valores pronosticados más cercanos a los observados.

Existen diversas medidas del error. El error de pronóstico básico e_{t+h} donde h denota el horizonte de pronóstico, se calcula como la diferencia entre el dato pronosticado $\hat{y}_{t+h}$ y dato observado y_{t+h}

$$e_{t+h} = \hat{y}_{t+h} - y_{t+h} \qquad (21)$$

De este error básico se derivan mediadas alternativas del error, tales como el error absoluto

$$AE_{t+h} = \left|\hat{y}_{t+h} - y_{t+h}\right| \qquad (22)$$

o el error cuadrático

$$SE_{t+h} = \left(\hat{y}_{t+h} - y_{t+h}\right)^2 \qquad (23)$$

Aunque estas medidas estadísticas del error producen un valor de cero para un pronóstico óptimo y son simétricas alrededor de este punto, cada una de ellas implica una ponderación distinta para las desviaciones del valor del pronóstico con respecto al valor observado. Las medidas de error cuadráticas o cúbicas, así como otras de potencias mayores tienen la ventaja de que penalizan más a las desviaciones extremas que a las pequeñas, mientras que medidas de error absolutas le dan pesos idénticos a todos los errores sin importar su tamaño.

Sin embargo, todas estas medidas, desconocen que los costos de que los pronósticos se ubiquen por debajo o por encima del dato observado, son frecuentemente no simétricos y típicamente no cuadráticos. Por ejemplo, en el manejo de inventarios médicos, los costos de subestimar o sobreestimar la cantidad necesaria de sangre de determinado grupo sanguíneo puede resultar en costos altamente asimétricos. La sobreestimación puede causar costos de almacenamiento de inventarios, mientras que la subestimación puede ser fatal.

La calidad de un pronóstico debe evaluarse considerando su habilidad para mejorar la calidad de las decisiones que soportan y por lo tanto la evaluación del desempeño de un método particular debe medirse por los costos en los que se incurra por la toma de decisiones basada en pronósticos incorrectos.

En países como el nuestro, que todavía deben enfrentar procesos desinflacionarios, a la autoridad monetaria le resulta mucho más costoso, en términos de credibilidad, cuando anuncia una meta de inflación inferior a la que posteriormente tiene lugar, que cuando lo contrario ocurre. En otras palabras, su evaluación de pronósticos por fuera de muestra, debe penalizar más duramente cuando el dato efectivo supere al dato estimado.

Una vez se ha reconocido que los costos en los que la autoridad monetaria debe incurrir como resultado de una subestimación de la inflación, no guardan simetría con los costos derivados de una sobreestimación de esta variable, debe seleccionarse una medida de error compatible con estas características.

En un plano cuya abcisa mida la distancia entre el dato pronosticado y el dato efectivamente observado, es decir, la magnitud del error del modelo y cuya ordenada mida el grado de penalización de tales errores, este fenómeno podría capturarse a través de una función LINLIN cuya pendiente del tramo a la derecha de cero fuera inferior a la pendiente del tramo a su izquierda.

En otras palabras, bajo un esquema de inflación objetivo, donde el Banco Central anuncia una meta o incluso un rango en el que deberá situarse la inflación en el siguiente período, resulta mucho más costoso que el modelo que ayuda a soportar dicho anuncio, arroje un valor inferior al que ocurre posteriormente.

Esto sucede porque los agentes económicos que inicialmente creyeron en el anuncio de inflación y con base en él negociaron sus contratos laborales para el siguiente período, pero posteriormente debieron soportar un incremento inesperado en el nivel de precios, que redujo sus salarios reales y su poder adquisitivo, perderán su confianza en la autoridad monetaria y ésta a su vez perderá uno de sus activos más valiosos: credibilidad. En el siguiente período, el público hará caso omiso de la meta o rango anunciado y su expectativa de inflación superará a aquella anunciada, luego pactará salarios nominales compatibles con su propia expectativa. Estos se traducirán en mayores costos laborales que finalmente se verán reflejados en los precios, confirmándose así sus expectativas.

Este juego repetido en el que el Banco Central le falló en una oportunidad al público, le significó el fracaso de su programa desinflacionario en los períodos posteriores. Por lo tanto, la red aquí propuesta minimizará una función de costos asimétrica. Muy poco se ha investigado sobre funciones de error no cuadráticas para el entrenamiento de las redes o sobre funciones de costos asimétricas para la evaluación por fuera de muestra.

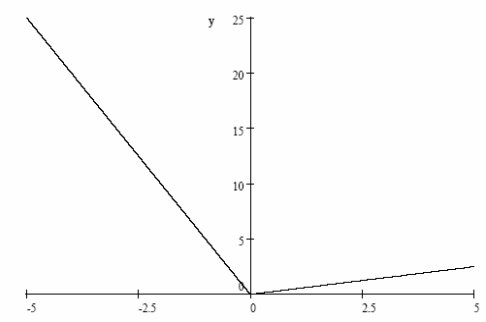
Granger desarrolló una función lineal asimétrica de costos para pronósticos de manejos de inventarios. La función LINLIN de costos es lineal a la izquierda y a la derecha de cero. Los parámetros *a* y *b* son las pendientes para cada tramo de la función. El parámetro

a corresponde los costos asociados a la pérdida de ingresos por ventas no realizadas como resultado de una subestimación, mientras que *b* se relaciona con los costos de almacenamiento de inventario resultantes de una sobreestimación. Para $a \neq b$ estas funciones de costos son asimétricas alrededor de cero y su grado de asimetría está dado por el ratio *a/b*.

$$LLC(\hat{y}_{t+h}, y_{t+h}) = \begin{cases} a|\hat{y}_{t+h} - y_{t+h}| & para \quad \hat{y}_{t+h}, < y_{t+h} \\ 0 & para \quad \hat{y}_{t+h} = y_{t+h} \\ b|\hat{y}_{t+h} - y_{t+h}| & para \quad \hat{y}_{t+h} > y_{t+h} \end{cases} \qquad (24)$$

Para ilustrar este fenómeno, considere el popular ejemplo de una aerolínea que debe destinar aviones de distintos tamaños, de tal forma que logre satisfacer la demanda de vuelos. Se asume que viajar con una silla vacía por sobreestimar la demanda de tiquetes y emplear un avión más grande de lo necesario, es menos costoso que, por una subestimación, emplear un avión pequeño y dejar de vender un tiquete por falta de cupo. Esto equivale a tener un beneficio marginal mayor que un costo marginal. Por tanto, los costos *a* de subestimar la demanda y dejar de percibir ingresos por ventas, son mayores que aquellos derivados de una sobreestimación, es decir *a>b*.

Figura 7: Función LINLIN[19]



[19] Análisis de Pronóstico con Funciones de Pérdida Asimétrica. Jalil (2005).

Teniendo en cuenta las anteriores consideraciones, los pronósticos por fuera de muestra, de las distintas redes neuronales estimadas en este trabajo, se evaluarán a través de una función de costos asimétrica como la LINLIN que describe adecuadamente la naturaleza del fenómeno bajo estudio. Dicha evaluación se hará por horizonte de pronóstico y con parámetros que varían con la lejanía en el tiempo de dicho pronóstico. Es decir que la red que mejor pronostica, un período hacia adelante, es aquella que minimiza la función de costos elegida bajo unos parámetros *a* y *b* determinados. De igual forma, se hallarán las redes con mejor desempeño de pronóstico en cada uno de los horizontes siguientes. Sin embargo, los parámetros de esta función se suavizarán, reflejando el hecho de que la información que la red tiene disponible para pronosticar a horizontes más elevados, es cada vez más difusa.

La primera tabla del anexo 5, muestra la arquitectura que minimiza el error asimétrico, en cada uno de los horizontes. Puede observarse, cómo, según este criterio, la arquitectura 3_2 tiene el mejor desempeño pronosticando un período hacia adelante, la 6_2 y la 2_2 son las mejores pronosticando 2 y 3 períodos en el futuro respectivamente y del horizonte 4 en adelante, la mejor red es la 2_3. Esta es definitivamente una arquitectura sencilla, con tan sólo 2 variables en la parte no lineal, a diferencia de la complejidad de las mejores redes por dentro de muestra.

El anexo 4 incluye las mejores redes para cada uno de los horizontes, de acuerdo con medidas simétricas tradicionales del error como lo son el RMSE, el MAE, el RMSPE y el MAPE. Puede observarse cómo la elección de la mejor red, en los primeros 3 horizontes, difiere de la decisión adoptada a través de la minimización de una función de costos asimétrica.

Hasta ahora tan sólo se ha elegido, dentro de las distintas redes neuronales aquí estimadas, la arquitectura con el menor error de pronóstico por horizonte; sin embargo no se ha contrastado aún su desempeño frente a otros modelos.

Se ha estimado un modelo ARIMA[20] cuyos pronósticos se han hecho también siguiendo un esquema de *rolling* como el adoptado para la red neuronal. La última columna de la segunda tabla del anexo 5 contiene las medidas del error asimétrico calculadas para este modelo. La comparación de ambas tablas comprueba la superioridad de los pronósticos hechos por las redes neuronales en cada uno de los horizontes. En el primer horizonte, por ejemplo, la medida de la mejor red, de acuerdo con una medida de error asimétrica, es 0.172 contra 0.44 arrojado por el ARIMA. Pronosticando 6 períodos hacia adelante, la medida de la mejor red es 0.197 frente a 1.089 del ARIMA para una diferencia de 0.892. Similarmente sucede con el resto de medidas por horizonte. Tan sólo en el último, el error del ARIMA se encuentra 0.015 por debajo de la red.

5. Conclusiones

Las redes neuronales artificiales son modelos computacionales que tratan de replicar, de manera simplificada, el complejo funcionamiento del cerebro humano. De acuerdo con Tkacz y Hu (1999) pueden aproximar cualquier función no lineal si son correctamente especificadas. Dado que en las series económicas, es más probable que aparezcan relaciones no lineales que lineales (Granger, 1991), como las exigidas por los modelos econométricos tradicionales, las ANN han ganado una inmensa popularidad en este campo de estudio.

En términos generales una red neuronal se compone de nodos, que actúan como *inputs*, *outputs* o procesadores intermedios. En la base de este modelo se encuentra la superficie de *inputs* que contiene a las variables explicativas en x_t. Ésta a su vez se conecta con el siguiente conjunto mediante una serie de trayectorias ponderadas o fuerzas conectoras $g_{i,j}$ (parecidas a las ponderaciones en un modelo de regresión). En la superficie oculta, se forman las combinaciones lineales de $x_t' g_{i,j}$ y se transforman en un valor entre 0 y 1 por las funciones de activación $G(.)$. Finalmente, éstas son multiplicadas por pesos b_j para producir el output y_t.

[20] Por sus siglas en inglés, Autoregressive Integrated Moving Average.

Las ponderaciones $\Theta = \{\gamma_{ij}, \beta_j, \phi_i\}$ de la red presentada en este trabajo, se obtuvieronminimizando la suma de las desviaciones al cuadrado entre el output y el pronóstico de dicha red, es decir la suma de residuales al cuadrado. Precisamente, el aprendizaje de la red, se encuentra en el proceso de entrenamiento durante el cual se estiman y ajustan sucesivamente estos parámetros con el fin de minimizar el error y obtener el modelo de red neuronal que mejor capture el comportamiento de la serie bajo estudio.

Particularmente, este trabajo exploró la relación entre el dinero y la inflación a través de una red neuronal artificial. Intuitivamente, dicha relación parece presentar comportamientos no lineales, que motivaron este ejercicio.
La evaluación de las distintas arquitecturas posibles que condujo a la elección de las mejores redes neuronales, fue llevada a cabo tanto dentro de muestra como por fuera de ella. Pudo confirmarse que aquellas redes cuyo desempeño era el mejor dentro de muestra presentaban un número elevado de unidades escondidas y en consecuencia una gran flexibilidad que les permitía capturar y memorizar perfectamente el comportamiento no lineal de la serie bajo estudio, pero les impedía predecir su comportamiento futuro.

La evaluación por fuera de muestra incorporó una serie de innovaciones en el estudio que en Colombia se ha hecho sobre redes neuronales artificiales. Primero, se adoptó un esquema de *rolling* de pronósticos que actualiza la estimación de parámetros cada vez que un nuevo dato es incorporado en la base de datos. Segundo, además de las tradicionales medidas simétricas para evaluar el desempeño de pronóstico de un modelo, se minimizó también una función de costos asimétrica, puesto que para la autoridad monetaria resulta mucho más costoso en términos de credibilidad cuando dentro de su esquema de inflación objetivo anuncia una meta inferior a la que posteriormente se registra, que cuando lo contrario ocurre.

El desempeño de las mejores redes neuronales, de acuerdo con criterios tanto simétricos como asimétricos, fue comparado contra el de un modelo ARIMA, mostrando resultados claramente superiores para el caso de las redes seleccionadas.

Bibliografía

Abrahart R.J. y See L. (1998). Neural Networks vs. ARMA Modelling: constructing benchmark case studies of river flow prediction.

Arango L.E. y A. González. (1999). Some Evidence of Smooth Transition Nonlinearity in Colombian Inflation. Borradores de Economía. 105. Banco de la República.

Arango L.E., A. González y C.E. Posada. (2000). Returns and Interest Rate: A Nonlinear Relationship in the Bogotá Stock Market. Borradores de Economía. 169. Banco de la República.

Arango L.E. y L.F. Melo. (2001). Expansions and Contractions in Brazil, Colombia and Mexico: A view through non linear models. Borradores de Economía. 186. Banco de la República.

Atiya A.F., S.M. El-Shoura, S.I. Shaheen y M.S. El-Sherif. (1999). A Comparison between Neural Networks Forecasting Techniques. Case Study: River Flow Forecasting. IEEE. Vol. 10, No. 2.

Cao C.Q. y R.S. Tsay. (1992). Non linear Time Series Análisis of Stock Volatility. Journal of Applied Econometrics. Vol. 7.

Clements M.P., P.H. Franses y N. R. Swanson. (2004). Forecasting Economic and Financial Time Series with non linear Models. International Journal of Forecasting. 20. 169 - 183.

Christoffersen P.F. y F.X. Diebold. (1994). Optimal Prediction under Asymmetric Loss. National Bureau of Economic Research. Technical Working Paper No. 167.

Christoffersen P.F. y F.X. Diebold. (1996). Further Results on Forecasting and Model Selection under Asymmetric Loss. Journal Applied of Econometrics. Vol. 11. 561 - 572.

Crone S.F. (2002). Trainning Artificial Neural Networks for Time Series Prediction using Asymmetric Cost Functions.

Crone S.F. (2002). Prediction of White Noise Time Series using Artificial Neural Networks and Asymmetric Cost Functions.

Dabús C. y F. Tohme. (2003). Non-Linearities in the Relation between Inflation and Money Suply in Argentina: A SOC Approach.

Franses P.H. y D. van Dijk. (2000). Non-linear Time Series Models in Empirical Finance. Cambridge University Press.

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

González S. (2000). Neural Networks for Macroeconomic Forecasting: A complementary Approach to Linear Regression Models. Working Papers. 2000-07. Department of Finance Canada.

Imbs J., H. Mumtaz, M.O. Ravn y H. Rey. (1996). Non linearities and Real Exchange Rate Dynamics.

Kuan C.M. y T. Liu. (1995). Forecasting Exchange Rates using Feedforward and Recurrent Neural Networks. Journal of Applied Econometrics. Vol. 10. No. 4.

Jalil M.A. y L.F. Melo. (1999). Una relación no lineal entre inflación y los medios de pago. Borradores de Economía. 145. Banco de la República.

Jalil M.A. y C. Tobón. (1999). Incertidumbre Inflacionaria en Colombia: Una Aproximación a través de Modelos GARCH.

Jalil M.A. y M. Misas. (2006). Evaluación de pronósticos del tipo de cambio utilizando redes neuronales y funciones de pérdida asimétrica. Borradores de Economía. 376. Banco de la República.

McMillan D.G. (2003). Non-Linear Predictability of U.K Stock Market Return. Oxford Bulletin of Economics and Statistics. Vol. 65. No. 5. 557 - 573.

Melo L.F. y M.A. Misas. (1997). Análisis del comportamiento de la inflación trimestral en Colombia bajo cambios de régimen: Una evidencia a través del modelo switching de Hamilton. Borradores de Economía. 86. Banco de la República.

Misas M.A., E. López y P. Querubín. (2002). La Inflación en Colombia: Una aproximación desde las Redes Neuronales. Borradores de Economía. 199. Banco de la República.

Misas M.A., E. López, C.A. Arango y N. Hernández. (2003). La Demanda de Efectivo en Colombia: Una Caja Negra a la luz de las Redes Neuronales. Borradores de Economía. 268. Banco de la República.

Moshiri S. y N. Cameron. (1998) Neural Networks versus Econometric Models in Forecasting Inflation. University of Manitoba.

Nakamura E. (2004). Inflation Forecasting using a Neural Network. Havard University.

Ohno-Machado L., M.G. Walker y M.A. Musen. (1994). Hierarchical Neural Networks for Survival Analysis. Section of Medical Informatics, Stanford University School of Medicine.

Shachmurove Y. (2000). Utilizing Artificial Neural Network Model to Predict Stock Markets. CARESS Working Paper 00-11.

Shachmurove Y. (2002). Applying Artificial Neural Networks to Business, Economics and Finance. University of Pennsylvania.

Swanson N.R. y H. White. (1995). A Model Selection Approach to Real Time Macroeconomic Forecasting using Linear Models and Artificial Neural Networks. The Review of Economics and Statistics. No. 79.

Teräsvirta T. y H.M. Anderson. (1992). Characterizing Nonlinearities in Business Cycles using Smooth Transition Autoregressive Models. Journal of Applied Econometrics. Vol. 7. 119 - 136.

Tkacz G. y S. Hu. (1999) Forecasting GDP growth using Artificial Neural Networks. Working Paper 99-3. Bank of Canada.

Tkacz G. (2000). Non-Parametric and Neural Network Models of Inflation Changes. Working Paper 00-7. Bank of Canada.

Anexo 1

Resultados de la Estrategia Stepwise

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.011	0.038	0.028	0.090	0.771
Y1	0.501	0.049	34.668	105.990	<.0001
Y4	-0.150	0.041	4.402	13.460	0.000
Y11	0.151	0.058	2.238	6.840	0.010
Y12	0.155	0.061	2.148	6.570	0.011
X2	0.192	0.043	6.482	19.820	<.0001
X13	0.123	0.045	2.495	7.630	0.006

Resultados de la Estrategia Forward

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.011	0.038	0.028	0.090	0.771
Y1	0.501	0.049	34.668	105.990	<.0001
Y4	-0.150	0.041	4.402	13.460	0.000
Y11	0.151	0.058	2.238	6.840	0.010
Y12	0.155	0.061	2.148	6.570	0.011
X2	0.192	0.043	6.482	19.820	<.0001
X13	0.123	0.045	2.495	7.630	0.006

Resultados de la Estrategia Backward

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.006	0.038	0.009	0.030	0.867
Y1	0.556	0.044	52.823	157.670	<.0001
Y4	-0.153	0.041	4.610	13.760	0.000
Y11	0.228	0.050	6.900	20.590	<.0001
X2	0.214	0.043	8.401	25.080	<.0001
X13	0.121	0.045	2.395	7.150	0.008

Anexo 2

Medidas de Evaluación dentro de Muestra

$$AIC\,[21](k) = n\,\ln\left(\hat{s}^{\,2}\right) + 2k$$

$$BIC\,[22](k) = n\,\ln\left(\hat{s}^{\,2}\right) + k\ln(n)$$

$$RMSE\,[23] = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2}$$

$$RMSPE\,[24] = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2}$$

$$MAE\,[25] = \frac{1}{n}\sum_{t=1}^{n}\left|\hat{y}_t - y_t\right|$$

$$MAPE\,[26] = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right|$$

$$SRP = \frac{1}{n}\sum_{t=1}^{n} I_t\left[\left\{\left(y_t - y_{t-1}\right)\cdot\left(\hat{y}_t - \hat{y}_{t-1}\right)\right\} > 0\right]$$

$$SRN = \frac{1}{n}\sum_{t=1}^{n} I_t\left[\left\{\left(y_t - y_{t-1}\right)\cdot\left(\hat{y}_t - \hat{y}_{t-1}\right)\right\} < 0\right]$$

$$SR = SRP + SRN$$

[21] Akaike Information Criterion.
[22] Bayesian Information Criterion.
[23] Root Mean Squared Error.
[24] Root Mean Squared Prediction Error.
[25] Mean Absolute Error.
[26] Mean Absolute Prediction Error.

donde

n es el número de observaciones.

$\hat{e}_t$ son los errores estimados, entendidos como la diferencia entre los datos observados y los valores estimados por la red neuronal.

$\hat{s}^2 = \frac{\sum_{t=1}^{n} \hat{e}_t^2}{n}$ es la varianza estimada.

k es el número de parámetros de la red neuronal.

$\hat{y}_t$ es el valor estimado por la red neuronal.

Anexo 3

Resultados de la Evaluación dentro de Muestra

CRITERIO	P	Q	W	AIC	BIC	R2	RMSE	RMSPE	MAE	MAPE	SR
AIC	**6**	**4**	**1**	**-1.311**	-0.848	0.996	0.455	0.026	0.346	0.020	75.417
	6	4	2	-1.214	-0.751	0.996	0.477	0.027	0.356	0.020	73.750
	7	**4**	**1**	**-1.199**	-0.678	0.996	0.473	0.025	0.352	0.019	73.333
BIC	4	3	1	-1.137	-0.877	0.995	0.526	0.029	0.410	0.023	70.000
	6	**4**	**1**	-1.311	**-0.848**	0.996	0.455	0.026	0.346	0.020	75.417
	1	3	1	-0.961	-0.831	0.993	0.596	0.031	0.453	0.025	71.250
R2	1	1	4	-0.837	-0.794	0.992	0.650	0.034	0.460	0.025	71.667
	1	1	5	-0.837	-0.794	0.992	0.650	0.034	0.460	0.025	71.667
	1	1	1	-0.842	-0.799	0.992	0.648	0.035	0.464	0.026	72.500
RMSE	**6**	**4**	**1**	-1.311	-0.848	0.996	**0.455**	0.026	0.346	0.020	75.417
	7	**4**	**1**	-1.199	-0.678	0.996	**0.473**	0.025	0.352	0.019	73.333
	6	4	2	-1.214	-0.751	0.996	0.477	0.027	0.356	0.020	73.750
RMSPE	**7**	**4**	**1**	-1.199	-0.678	0.996	0.473	**0.025**	0.352	0.019	73.333
	6	**4**	**1**	-1.311	-0.848	0.996	0.455	**0.026**	0.346	0.020	75.417
	7	4	2	-1.140	-0.620	0.996	0.487	0.027	0.347	0.019	73.750
MAE	**6**	**4**	**1**	-1.311	-0.848	0.996	0.455	0.026	**0.346**	0.020	75.417
	7	4	2	-1.140	-0.620	0.996	0.487	0.027	0.347	0.019	73.750
	7	**4**	**1**	-1.199	-0.678	0.996	0.473	0.025	**0.352**	0.019	73.333
SR	3	3	5	-0.948	-0.732	0.994	0.585	0.033	0.435	0.024	67.083
	7	2	2	-0.955	-0.695	0.994	0.576	0.032	0.421	0.024	67.083
	6	2	3	-0.934	-0.703	0.994	0.587	0.033	0.434	0.025	67.500

Anexo 4

Resultados de la Evaluación Simétrica por fuera de Muestra

Redes Neuronales Artificiales

HORIZONTE	P	Q	W	RMSE	MAE	RMSPE	MAPE
1	4	1	1	0.2490	0.2021	4.0762	3.3349
2	2	4	2	0.3681	0.3185	6.3011	5.4419
3	2	3	1	0.3720	0.3165	6.4628	5.4287
4	2	3	1	0.4117	0.3687	7.1503	6.3414
5	2	3	1	0.4883	0.4374	8.6535	7.6232
6	2	3	1	0.4443	0.3666	7.9307	6.4480
7	2	3	1	0.5453	0.4476	9.8373	7.9431
8	2	3	1	0.6378	0.5082	11.6177	9.1173
9	2	3	1	0.5159	0.4419	9.2044	7.8375
10	2	3	1	0.6036	0.4987	10.6953	8.7682
11	2	3	1	0.7644	0.6638	13.7793	11.8034
12	2	3	1	0.8459	0.7494	15.4157	13.4371

Modelo ARIMA

HORIZONTE	RMSE	MAE	RMSPE	MAPE
1	0.4510	0.3828	7.4341	6.3498
2	0.8308	0.7063	13.8123	11.8229
3	1.1353	0.9978	19.1540	16.9003
4	1.3902	1.2441	23.8510	21.2853
5	1.5752	1.4253	27.4459	24.7200
6	1.6915	1.5493	29.6783	27.0396
7	1.7197	1.5807	30.2533	27.7165
8	1.6319	1.4500	28.8165	25.5411
9	1.4857	1.2840	26.6110	22.7299
10	1.2728	1.0073	23.2610	18.0817
11	0.9857	0.7305	18.2815	13.3316
12	0.7474	0.7189	13.4908	12.8177

Anexo 5

Resultados de la Evaluación Asimétrica por fuera de Muestra

Redes Neuronales Artificiales

HORIZONTE	P	Q	W	RMSE	MAE	RMSPE	MAPE	LINLIN
1	3	2	2	0.280	0.222	4.671	3.702	0.172
2	6	2	2	0.424	0.325	7.149	5.470	0.192
3	2	2	1	0.443	0.355	7.664	6.101	0.190
4	2	3	1	0.412	0.369	7.150	6.341	0.228
5	2	3	1	0.488	0.437	8.654	7.623	0.261
6	2	3	1	0.444	0.367	7.931	6.448	0.197
7	2	3	1	0.545	0.448	9.837	7.943	0.239
8	2	3	1	0.638	0.508	11.618	9.117	0.269
9	2	3	1	0.516	0.442	9.204	7.837	0.222
10	2	3	1	0.604	0.499	10.695	8.768	0.249
11	2	3	1	0.764	0.664	13.779	11.803	0.332
12	2	3	1	0.846	0.749	15.416	13.437	0.375

Modelo ARIMA

HORIZONTE	RMSE	MAE	RMSPE	MAPE	LINLIN
1	0.4510	0.3828	7.4341	6.3498	0.4437
2	0.8308	0.7063	13.8123	11.8229	0.5010
3	1.1353	0.9978	19.1540	16.9003	0.7026
4	1.3902	1.2441	23.8510	21.2853	0.8749
5	1.5752	1.4253	27.4459	24.7200	1.0019
6	1.6915	1.5493	29.6783	27.0396	1.0889
7	1.7197	1.5807	30.2533	27.7165	1.0938
8	1.6319	1.4500	28.8165	25.5411	0.9813
9	1.4857	1.2840	26.6110	22.7299	0.8339
10	1.2728	1.0073	23.2610	18.0817	0.6024
11	0.9857	0.7305	18.2815	13.3316	0.3715
12	0.7474	0.7189	13.4908	12.8177	0.3595

Anexo 6
Códigos en SAS

A continuación se explica brevemente lo que hace cada uno de los 5 programas empleados para la aplicación de redes neuronales artificiales al caso de la inflación en Colombia.

El primero es un programa desarrollado por Martha Misas A. para la elección del mejor conjunto de variables explicativas. Este programa hace uso de las observaciones del período de entrenamiento de la variable bajo estudio y de aquellas variables que se cree, pueden explicar su comportamiento, en este caso la información mensual rezagada un año y medio de la normalización de la diferencia de logaritmos del IPC y de M3. Particularmente, la elección del conjunto de *inputs* para esta red, es el resultado de la intersección de una serie de búsquedas que trae programadas SAS, como *stepwise selection*, *forward entry* o *backward removal*, explicados en la cuarta sección.

El programa de simulación es uno de los más extensos y computacionalmente costosos. El *input* de este programa lo constituye la información por dentro de muestra del conjunto de variables seleccionado en la etapa anterior. Como su nombre lo indica, este programa simula cada una de las posibles combinaciones de variables en la parte no lineal y de unidades escondidas. El número de variables en la parte lineal es siempre fijo y corresponde al número de variables explicativas encontradas en el anterior programa. Para la parte no lineal se probó desde una variable hasta el total de variables explicativas, siete en este caso, a la vez que se probaba con una unidad escondida hasta cuatro, para un total de 28 arquitecturas distintas. El resultado es un vector de parámetros iniciales por cada una de las arquitecturas simuladas. En este caso se simularon 30 vectores por cada arquitectura con el fin de seleccionar aquellos 5 que al ser evaluados en la función objetivo arrojaran los menores valores.

La evaluación por dentro de muestra calcula para cada uno de estos 140 vectores de parámetros iniciales (5 replicaciones por cada una de las 28 arquitecturas) una variedad

de medidas, a partir de las cuales es posible seleccionar aquellos vectores cuyo ajuste es el mejor por dentro de muestra. Nótese que las observaciones son mensuales, mientras que para esta aplicación interesan los pronósticos anuales de la inflación, luego las evaluaciones por dentro y fuera de muestra requieren una transformación de los datos que es llevada a cabo al interior de los mismos programas.

El programa de *rolling* de pronósticos es tal vez el más complejo y novedoso en la aplicación de redes neuronales. Este requiere no sólo las observaciones con las que la red fue entrenada, sino aquellas contra las que se pretende evaluar los pronósticos por fuera de muestra y naturalmente el conjunto de parámetros simulados para cada arquitectura. Lo interesante de este programa es que aborda el tema de pronósticos de una manera más dinámica

Al concluir este proceso, para cada arquitectura considerada, se obtiene una matriz de 18 x 12, compuesta por 18 vectores, uno por cada conjunto de parámetros que se reestimó al incluir una observación adicional de aquellas que se dejaron por fuera de muestra; del tamaño del horizonte de pronóstico, en este caso 12. Es decir que la primera columna contiene los pronósticos, 12 períodos hacia adelante, realizados con los parámetros calculados a partir de la muestra de entrenamiento de la red. La última columna contiene los pronósticos hechos a partir del vector de parámetros que se calculó al incorporar los datos de las observaciones que se dejaron por fuera de muestra sin incluir el último.

Una vez se han obtenido los pronósticos de cada una de las arquitecturas simuladas, debe construirse una medida del error de pronóstico de la red neuronal, para luego hallar cuál de ellas está arrojando los valores pronosticados más cercanos a los observados.