UNIVERSIDAD CATÓLICA ANDRÉS BELLO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

ANÁLISIS EXPLORATORIO DE LOTERÍAS MEDIANTE MODELOS ESTADÍSTICOS Y REDES NEURALES

TRABAJO DE ASCENSO

Presentado ante la

UNIVERSIDAD CATÓLICA ANDRÉS BELLO

Agregado

Realizado por:
Ing. Evelenir Barreto González

Para optar al escalafón: Agregado

Caracas, Diciembre 2009

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Predecir es "anunciar por revelación, ciencia o conjetura algo que ha de suceder". La predicción ha interesado al hombre desde mucho tiempo atrás, antes que fuese vista con un enfoque científico. Los monarcas de la antigüedad, los magistrados de Esparta y los romanos trataban de predecir los hechos que ocurrirían en base a oráculos o a la posición de las estrellas. Sólo hasta el momento del desarrollo de la estadística se realizaron predicciones con una rigurosidad científica tal, que fuesen totalmente verificables los resultados de las mismas.

Han sido desarrollados métodos como la correlación y el análisis de las series de tiempo, lo cual ha generado resultados muy útiles para el desarrollo científico. Especialmente en las series de tiempo se han elaborado modelos estocásticos y funciones de autocorrelación, como el método ARIMA (Autoregresive Integrated Moving Average) propuesto por Box y Jenkins [Box y Jenkins 1976]. Las series de tiempo son un conjunto de observaciones medidas en momentos específicos con intervalos de separación generalmente iguales. Su análisis ha sido muy útil para la identificación de patrones cíclicos, estaciónales, de tendencias a largo plazo o de movimientos aleatorios. El análisis de las series de tiempo consiste en una descripción matemática de los componentes presentes en la serie. [Spiegel, 1961].

A muchos les gustaría conocer, o tener una idea aproximada, de lo que va a ocurrir en el futuro. Se quisiera anticipar el valor de muchas variables en el campo de las ciencias naturales o de las ciencias económicas. Por ejemplo, anticipar sí las acciones de la bolsa subirán o bajarán, o cuál será el número ganador de la lotería. Por eso siempre habrá un interés de índole práctico pero también de curiosidad científica para predecir lo que esta el futuro. En este sentido, se han desarrollado en el campo teórico modelos de predicción que luego se han aplicado en el pronóstico con resultados variables. Aun existen muchas áreas en las cuales se quiere predecir y quedan muchas preguntas por responder.

En los campos de la medicina, la física, la meteorología y la hidrología entre muchos otros, existen variables cuyos comportamientos no se pueden modelar aun con métodos tradicionales. Por esto, se continúan desarrollan métodos que se acoplen más cercanamente al comportamiento de estas variables. Al conseguir modelos más aproximados a la realidad, se podrá pronosticar el comportamiento de estas variables, lo que será muy provechoso para la mejor toma de decisiones en todos los ámbitos antes mencionados.

Los juegos de azar con autorización gubernamental, en los cuales se apuesta dinero, aparentemente tienen cada día mas adeptos. Se han convertido en una actividad económica significativa por los volúmenes de dinero que manejan. Se estima que los números resultantes en dichos juegos deberían tener un comportamiento totalmente aleatorio, es decir que el resultado tiene la misma probabilidad de ocurrencia para todo el universo de números a seleccionar.

En este sentido se desea conocer sí los resultados de los juegos de azar, en las principales loterías en Venezuela, cumplen con esa aleatoriedad, y en caso contrario, que modelo o combinación de modelos, se ajusta mas a su comportamiento, ya sea por medio de métodos tradicionales estadísticos o métodos modernos, tal como las redes neuronales.

La investigación teórica propuesta puede además tener utilidad práctica como lo demuestra el hecho que ya existen empresas dedicadas al desarrollo y comercialización de software que ejecutan algoritmos de predicción basados en modelos previamente definidos.

A lo largo de toda la investigación se analizará el comportamiento de los resultados obtenidos en los sorteos de loterías venezolanas entre los años 1995 a 2004, mediante un método estadístico y un tipo de red neuronal.

MARCO REFERENCIAL

1. SERIES DE TIEMPO

Una serie de tiempo es "un conjunto de observaciones realizadas secuencialmente en el tiempo" o "un conjunto de datos (observaciones) de una variable medidas en intervalos de tiempo sucesivos e iguales".

Matemáticamente una serie de tiempo representa un conjunto de observaciones y_1, y_2, y_3,..., yn, de una variable **Y** tomadas secuencialmente e igualmente espaciadas en el tiempo $t = 1, 2, 3, \ldots, n$. Por tanto **Y** es una función de t, $\mathbf{Y} = f(t)$.

Existe una gran cantidad de ejemplos de series de tiempo que pueden ser mencionados, entre ellos se tienen:

- La población anual de ganado en un país específico.
- Una secuencia mensual de la cantidad de bienes despachados desde una fábrica.
- Una serie semanal del número de accidentes automovilísticos en una ruta específica.
- Las observaciones horarias hechas sobre el rendimiento de un proceso químico.
- Los precios de las acciones de una compañía específica en la bolsa de valores en un período dado.

Los ejemplos de series de tiempo abundan en áreas tales como la economía, mercadeo, demografía, meteorología, ingeniería, etc. [Collantes 2001]

1.1 Componentes

Las series de tiempo están compuestas por varios componentes principales que son: tendencia, movimientos estacionales, cíclicos y aleatorios. Véase figura 1.

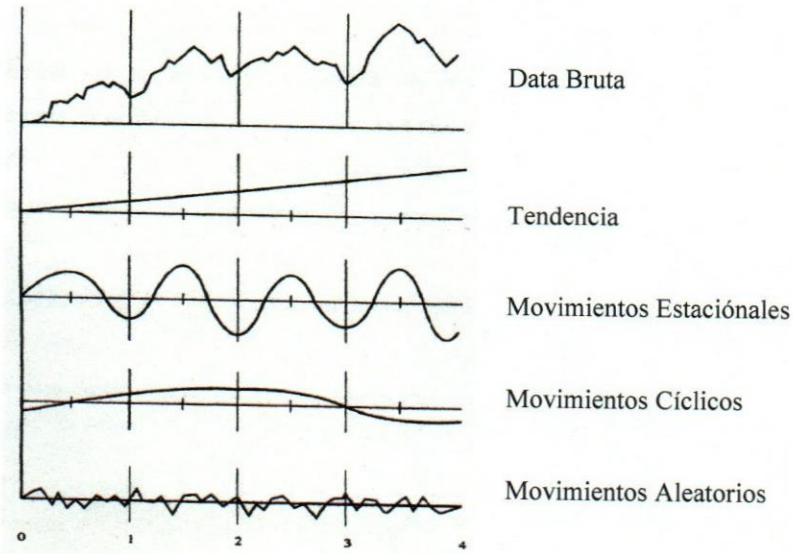

Figura 1. Componentes de una serie de tiempo

La tendencia, es un cambio a largo plazo en la media, se refiere a la dirección general la cual crece o disminuye a lo largo del tiempo. Es útil dibujar una línea de tendencia, que puede ser determinada por el método de los mínimos cuadrados.

Los movimientos o variaciones cíclicas son oscilaciones sobre la línea de tendencia. Estos ciclos pueden periódicos o no.

Los movimientos estaciónales se refieren a patrones idénticos o casi idénticos que se repiten en una serie de tiempo cada determinado período. Aunque los movimientos estacionales, en general, tienen periodicidad anual, también pueden ser diarios, horarios, semanales. [Barreto 2001]

Los movimientos aleatorios o irregulares no responden a ningún comportamiento visible y a menudo no tienen causas específicas.

Los métodos tradicionales de análisis de series de tiempo descomponen las series en las variaciones o movimientos antes mencionados.

1.2 Objetivos del análisis de las series de tiempo

Según Chatfield 1980, el análisis de series de tiempo persigue cuatro grandes objetivos, los cuales son:

- Descripción
- Explicación
- Predicción
- Control

La descripción se refiere a obtener medidas descriptivas de las principales propiedades de la serie, algunas de las cuales se pueden observar con la simple inspección tales como picos, tendencias o variaciones estacionales.

El objetivo de explicar una serie de tiempo se consigue cuando se observa la relación entre dos o más variables de la serie. Los modelos de regresión múltiple son muy útiles en estos casos. También el análisis de sistemas lineales, que convierte una serie de entrada a una serie de salida a través de una función lineal.

La predicción y control se refieren a pronosticar valores futuros de la serie. La predicción esta muy relacionada a los problemas de control en muchas situaciones. Por ejemplo, sí se puede predecir que un proceso manufacturero se saldrá de su curso normal en determinado momento, de acuerdo a datos históricos, entonces, se pueden tomar medidas previsivas antes de que ocurra el problema. En el caso del control estadístico de la calidad, las observaciones de la serie son estudiadas en gráficos de control de tal forma que el controlador tome acciones como resultado del análisis de estos gráficos.

1.3 Clasificación

Las series se pueden clasificar de acuerdo a diferentes criterios, por ejemplo en relación a la certeza de su comportamiento, pudiendo ser determinísticas y estocásticas.

Son series determinísticas aquellas cuyos valores futuros pueden ser calculados con exactitud por una función matemática.

Las series estocásticas son aquellas cuyos valores futuros sólo pueden ser descritos por una distribución de probabilidad. La predicción exacta en las series estocásticas es imposible y por ende debe ser reemplazada por la idea que valores futuros de la serie aparecerán con una cierta probabilidad, lo cual esta condicionado al conocimiento de valores históricos.

1.4 Caracterización de los datos

Un evento aleatorio se describe como aquel que no se puede prever o anticipar. Estadísticamente se han diseñado un conjunto de pruebas o test para verificar la aleatoriedad o independencia que puede existir entre los datos de una serie. Algunas de las pruebas mas frecuentes son enumeradas a continuación:

- **Prueba de la bondad del ajuste:**

Existen varias pruebas de la bondad del ajuste. En esta investigación se trabaja con la prueba Chi-cuadrado. Este test compara las observaciones de una variable con su comportamiento esperado y brinda una respuesta a la hipótesis de que la muestra proviene de la misma población que los datos esperados. La hipótesis se rechaza si el valor del estadístico Chi-cuadrado es mayor al valor crítico para la variable aleatoria Chi-cuadrado con r-1 grados de libertad (siendo r la el número de categorías existentes en la prueba) y probabilidad de 0.95.

- **Prueba de las rachas:**

La prueba de rachas es un procedimiento estadístico que se utiliza para contrastar la hipótesis nula de que la secuencia de apariciones de un valor en el orden observado en la serie es aleatoria, o equivalentemente, que las observaciones son independientes entre sí.

La prueba de rachas se utiliza para contrastar la hipótesis nula de que la secuencia de apariciones de los valores de la muestra en el orden observado es aleatoria. Si el valor asociado al estadístico de contraste es menor que α se rechazara la hipótesis nula al nivel de significación α. [Ferrán 1996]

- **Prueba de intervalos**

La prueba de intervalos persigue observar a los intervalos entre ocurrencias sucesivas del mismo valor. Se cuenta el número de separaciones 1, 2, 3, etc. Luego se calcula el estadístico Chi-cuadrado usando esta cantidad y la cantidad esperada para cada longitud de separación y se compara su valor con el valor crítico de una variable aleatoria Chi-cuadrado para r-1 grados de libertad (siendo r el número de categorías existentes en la prueba) y probabilidad de 0.95.

2 MÉTODO ARIMA

El método ARIMA (de sus siglas en ingles, "Autoregressive Integrated Moving Average") es un método desarrollado por Box y Jenkins (1976) que permite caracterizar y predecir valores futuros de una serie de tiempo, basándose en valores pasados de una sola variable, o de dos variables entre las que exista una relación causal. [Collantes 2001]. En la presente investigación sólo se estudiará una sola variable.

Es uno de los métodos de uso más difundido para el análisis de serie de tiempo, debido a su generalidad y apoyo operacional mediante software bien documentado para su uso. Box y Jenkins fueron quienes lo popularizaron y lo hicieron accesible. [Maddala, 1996, citado por Collantes 2001]

El método de Box y Jenkins deber ser aplicado a procesos estacionarios, en caso contrario se deben realizar procedimientos previos tales como la transformación y diferenciación, las cuales se explicarán mas adelante.

Para utilizar el método, se trabaja con procesos autorregresivos (AR), de promedio móvil (MA), mixtos (autorregresivo y promedio móvil, ARMA) e integrados (autorregresivo integrado de promedio móvil, ARIMA).

El método ARIMA es una herramienta de predicción basada en el análisis de las propiedades probabilísticas de las series de tiempo, bajo la filosofía de "permitir que la información hable por si misma". (Box y Jenkins, 1994) (Gujarati, 1997). Como fue mencionado anteriormente, constituye el método de uso mas frecuente para el análisis de series de tiempo.

Su aplicación se realiza mediante un análisis exploratorio inicial y cuatro etapas principales que son a) identificación de los órdenes del modelo, b) estimación de los parámetros, c) verificación del modelo y d) predicción, tal como se indica en la figura 2.

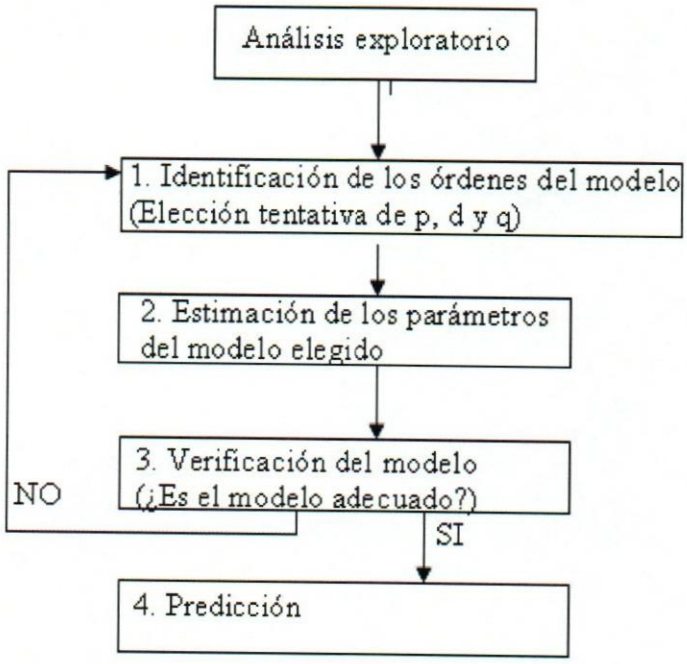

Figura 2. Etapas en el método iterativo para construir el modelo

2.1 Análisis Exploratorio

La etapa preparatoria consiste en realizar un análisis exploratorio del comportamiento de la serie a través de gráficos, histogramas y estadísticos descriptivos, que permitan formular conjeturas iniciales acerca del proceso estocástico. Además, en la etapa preparatoria se separan los datos en dos grupos: el primero, para ajustar el modelo y el segundo para validar el modelo (Figura 3). Se recomienda utilizar 80% de los datos para el "ajuste" y 20% para la "validación".

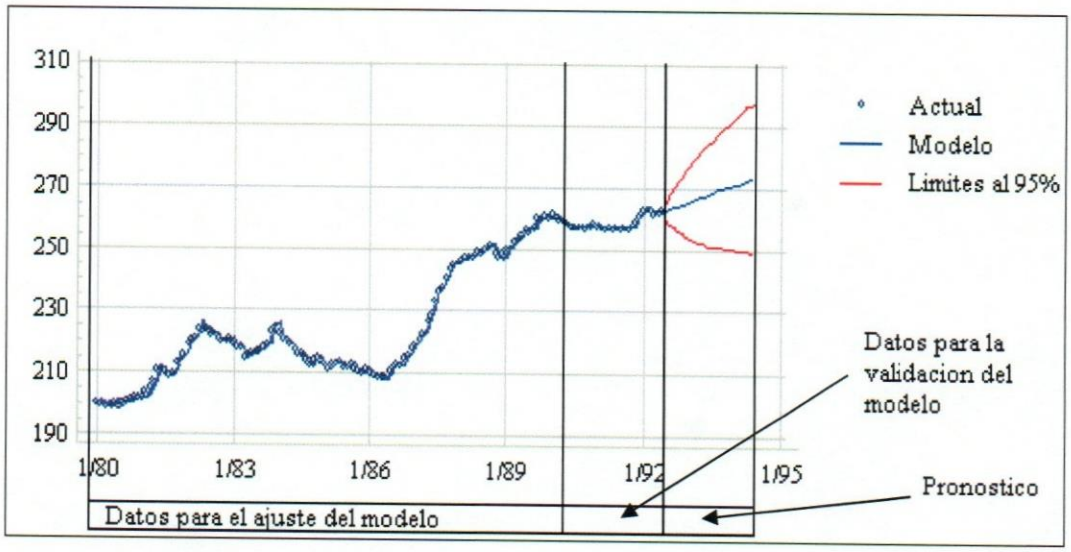

Figura 3. Serie de tiempo observada y pronosticada con límites al 95% de probabilidad

2.2 Identificación de los órdenes del modelo

Esta etapa persigue el propósito de identificar las siguientes variables:

s = Periodicidad estacional de la serie

d = Ordenes de diferenciación regular

D = Ordenes de diferenciación estacional

p = Cantidad de términos autorregresivos regulares

P = Cantidad de términos autorregresivos estaciónales

q = Cantidad de términos de promedio móvil regular

Q = Cantidad de términos de promedio móvil estacional

Para determinar los órdenes del modelo se requiere realizar previamente "las pruebas de estacionariedad", es decir verificar el cumplimiento de las siguientes condiciones:

- La media de Y_t sea constante: $E(Y_t) = \mu_t = \mu \qquad \forall t$
- La varianza de Y_t sea constante: $V(Y_t) = \sigma^2_t = \sigma^2 \qquad \forall t$

- La correlación entre Y_t y Y_{t+k} depende únicamente del numero de retardos que las separa: $\rho(t,k) = \rho_k \qquad \forall k$

Para probar la estacionariedad, se aplican procedimientos estadísticos los cuales se interrelacionan como se indica en la figura 4:

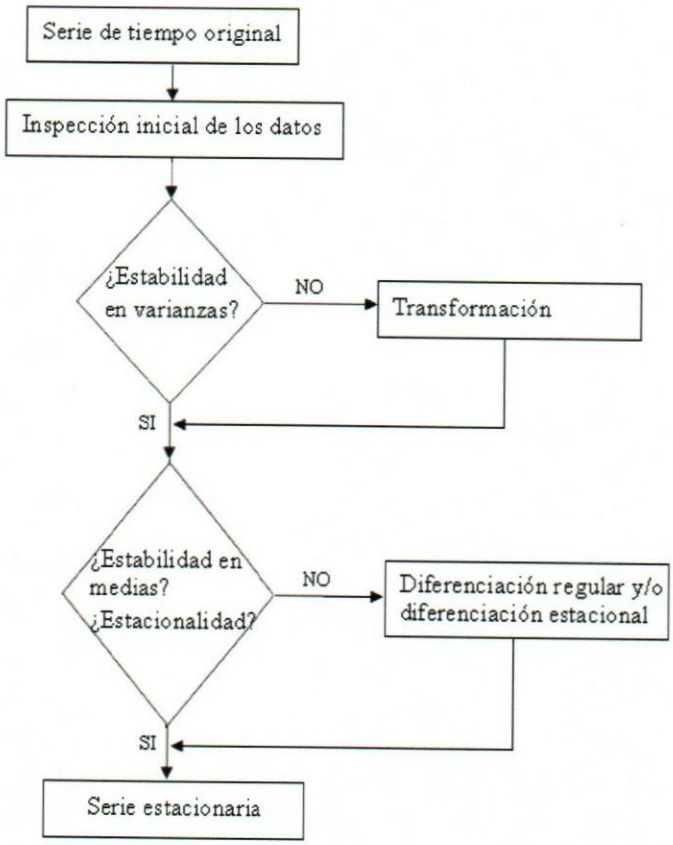

Figura 4. Interrelaciones para probar la estacionariedad

- **Inspección inicial de los datos**

Es una inspección visual del gráfico de la serie, la cual proporciona una clave inicial respecto a la posible naturaleza de la serie de tiempo. Este análisis intuitivo sirve como comienzo para las pruebas más formales de estacionariedad, que se describen a continuación.

- **Análisis de la estabilidad en varianzas (Prueba de homocedasticidad)**

La homocedasticidad es la condición que tienen las series de poseer varianza (σ^2) constante en el tiempo. Pero las series pueden presentar una varianza cambiante. La estabilidad en varianzas se puede detectar agrupando las observaciones en K períodos de tiempo con el mismo número de observaciones y realizando la prueba de Levene que permite comprobar la hipótesis de que los K grupos proceden de poblaciones con varianza común, o lo que es lo mismo, presentan homocedasticidad.

Si la hipótesis de homogeneidad de varianzas fuera rechazada, a la serie original se le aplican transformaciones logarítmicas que estabilizan la varianza. Una de las transformaciones más comunes es la de Box-Cox. Si λ es el estadístico de Box-Cox, entonces la serie Y_t se transforma de la manera siguiente:

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{\left(Y_t^{(\lambda)} - 1\right)}{\lambda} & \text{Para } \lambda \neq 0 \\ \ln Y_t & \text{Para } \lambda = 0 \end{cases}$$

Obsérvese que para $\lambda = 1$, la transformación apenas tiene influencia sobre los valores originales. El valor que mas se suele utilizar es el de $\lambda = 0$ porque simplemente consiste en calcular logaritmos neperianos de los valores originales y obteniéndose con frecuencia una transformación exitosa. Posteriormente se debe realizar nuevamente la prueba de Levene para verificar la estabilidad en varianzas. [Collantes 2001]

En caso de que la hipótesis de homogeneidad de varianzas se acepte, entonces se prosigue con el análisis de estabilidad en medias.

- **Análisis de la estabilidad en medias y de la estacionalidad**

La condición de estacionariedad con respecto a la media establece que la media debe ser estable o constante durante el todo el tiempo *t*.

Si al graficar la serie se nota por simple inspección una tendencia creciente o decreciente, se puede descartar la condición de estacionariedad, pero sí por el contrario, la gráfica no muestra señales aparentes de alguna tendencia, no seria suficiente para emitir juicio alguno.

La estabilidad en medias se evalúa a partir de la función de autocorrelación simple. Si el correlograma presenta autocorrelaciones positivas en los primeros retardos y una disminución leve hacia cero a medida que aumentan los retardos, como se observa en la Figura 5, la serie no es estable en medias y debe ser diferenciada regularmente. De lo contrario se asume que la media es constante en el tiempo.

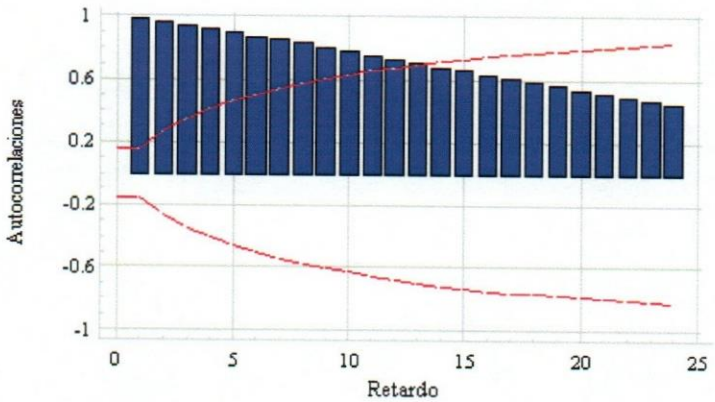

Figura 5. Función de autocorrelación simple (ACF) de una serie no estacionaria

Si la "función de autocorrelación simple"(ACF) muestra un comportamiento estacional, entonces se debe diferenciar estacionalmente.

El análisis se continua con el estudio conjunto de la "función de autocorrelación simple" y la "función de autocorrelación parcial"(PACF) conjuntamente para establecer los órdenes del modelo.

Un proceso estacionario y autorregresivo AR(p) se caracteriza porque los primeros coeficientes de la función de autocorrelación parcial son no nulos y el resto es cero, y la función de autocorrelación simple, presenta muchos coeficientes no nulos que decrecen con el retardo, con mezcla de exponenciales y sinusoidales.

Todo proceso de promedio móvil es estacionario. Un proceso de promedio móvil MA(q) se caracteriza porque los q primeros coeficientes de la función de autocorrelación simple son no nulos y el resto cero, y la función de autocorrelación parcial, en general, presenta muchos coeficientes no nulos que decrecen con el retardo con mezcla de exponenciales y sinusoidales.

Dado que los procesos ARMA son una mezcla de procesos AR y procesos MA, la función de autocorrelación simple y la función de autocorrelación parcial de un proceso estacionario ARMA(p, q) será una superposición de la función de autocorrelación simple y función de autocorrelación parcial de los procesos AR(p) y MA(q) correspondientes.

Lo descrito anteriormente se resume en la Tabla 1 que se presenta a continuación:

Tabla 1. Relaciones entre los correlogramas simple y parcial

Modelo	**Función de autocorrelación simple (ACF)**	**Función de autocorrelación parcial (PACF)**
Ruido blanco a_t	Nulo en todos los retardos	Nulo en todos los retardos
AR(p)	Decae lentamente	Se corta después del retardo p
MA(q)	Se corta después del retardo q	Decae lentamente
ARMA(p,q)	Decae lentamente	Decae lentamente

Según el comportamiento de los coeficientes de autocorrelación simple y parcial, se pueden determinar los ordenes p, d y q del modelo. Si la serie presenta estacionalidad, y ha sido diferenciada estacionalmente, se realiza el estudio de la función de autocorrelación simple y la función de autocorrelación parcial sólo para los retardos estacionales s, 2s, 3s, etc. Se realiza el mismo análisis para obtener los órdenes p y q, del modelo.

En resumen, si la varianza de la serie no es constante se requiere realizar una "transformación" de la serie para estabilizar la varianza. En caso de conseguir la estabilidad, el análisis continuará con la serie transformada. Si la media no es constante, se "diferenciará regularmente" la serie hasta estabilizar la media. Si la serie fuera estacional, se tomarían "diferencias estacionales" hasta eliminar la estacionalidad. Si la serie no es estacionaria en medias y es estacional, en ocasiones, al diferenciar estacionalmente, además de eliminar la estacionalidad, estabiliza la media. [Ferran, 1996]

En base a esta determinación se define el tipo de modelo a utilizar, ya sea de caminata aleatoria ARIMA(0,1,0), autorregresivo integrado de primer orden ARIMA(1,1,0), o de suavizamiento exponencial simple ARIMA(0,1,1) u otro.

2.3 Estimación de parámetros del modelo elegido

El modelo ARIMA, tiene la expresión matemática que se enuncia a continuación:

$$\Phi_P(B^S)\phi_p(B)(1 - B)^d (1 - B^S)^D Y_t = \delta + \theta_q(B)\Theta_Q(B^S)a_t$$

donde:

$\Phi_P(B^S)$ $= 1 - \Phi_1 B^S - \ldots - \phi_P B^{PS}$

$\phi_p(B)$ $= 1 - \phi_1 B - \ldots - \phi_p B^p$

$\theta_q(B)$ $= 1 - \theta_1 B - \ldots - \theta_q B^q$

$\Theta_Q(B^S)$ $= 1 - \Theta_1 B^S - \ldots - \Theta_Q B^{QS}$

B = Operador de retardo ➔ $BY_t = Y_{t-1}$

Y_t = Valores medidos de la serie.

a_t = Perturbación aleatoria o ruido blanco.

δ = Constante.

El procedimiento busca la determinación de los parámetros Φ, ϕ, Θ y θ, que se obtienen por la minimización de la suma de cuadrado de los errores a_t, a través del método de los mínimos cuadrados condicional o incondicional. Generalmente la determinación de los valores de los parámetros se realiza con la ayuda de herramientas computacionales, tales como el *Statistical Package for Social Science (SPSS)*, *Statistical Analysis SAS*, etc.

2.4 Verificación del modelo

La tercera etapa del método ARIMA tiene dos fases. La primera trata de comprobar si los residuos del modelo seleccionado se aproximan a un comportamiento de ruido blanco, si los criterios de comparación tienen valores aceptables y si los parámetros del modelo son significativamente distintos de cero. La segunda fase se refiere al "sobreajuste".

2.5 Predicción

En esta etapa final se predicen los valores de la serie, utilizando el modelo seleccionado. En primer lugar se comparan los resultados obtenidos por el modelo con las observaciones separadas en el "análisis exploratorio" para verificar la validez del modelo obtenido. Luego se calculan los errores de predicción, tales como el error medio absoluto, la raíz del error cuadrado medio porcentual y el error medio, entre otros, para poder así juzgar la bondad del pronóstico.

En segundo lugar y como fase final del método ARIMA, sí el modelo se considera ya definitivo y adecuado para predecir, se pronostican los valores futuros para un periodo de tiempo previamente seleccionado.

2.6 Materiales utilizados en las series de tiempo

Los resultados de las loterías venezolanas fueron descargados desde la página http://www.loterias.com.ve .

El software utilizado para los cálculos y generación de modelos fue el *SPSS* versión 10.0 para Windows.

3 REDES NEURALES ARTIFICIALES (RNA)

Las redes de neuronas humanas sirvieron de base para la creación de un modelo matemático para transmisión de información, que intenta emular el comportamiento de estas, en principio como una neurona artificial y luego expandiéndose a modelos más complejos conocidos como redes neuronales artificiales (RNA).

Las redes neuronales están compuestas por neuronas, el modelo de una neurona artificial tiene los siguientes elementos:

- **Entradas**: escalares que se le proporcionan a la red, de acuerdo al problema en estudio. Biológicamente representan las señales que provienen de otras neuronas y son capturadas por las dendritas.

- **Salidas**: son los valores que arroja la red como resultado del aprendizaje.

- **Pesos sinápticos**: son valores numéricos que corresponden a la fuerza de la sinapsis. Expresan la importancia de la entrada correspondiente.

- **Punto de suma**: realiza la combinación lineal o suma de todas las entradas multiplicadas por sus correspondientes pesos.
- **Función de activación**: es una función que limita el rango de salida de la neurona.

- **Sesgo**: valor formado por una entrada fija e igual a uno.

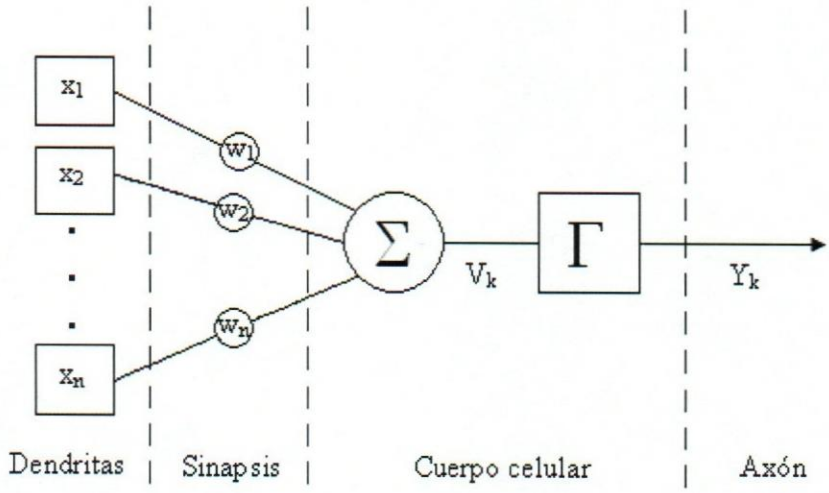

Figura 6. Relación entre la neurona biológica y la artificial

En la figura 6, se muestra como el modelo de la neurona artificial tiene sus equivalentes biológicos. La notación es la siguiente:

x_i = entradas a la neurona k; i = 1, 2,…, n

w_{ik} = pesos de la neurona k. i = 1, 2,…, n

$V_k = \sum_{i=0}^{n} x_i w_{ik}$ = suma de las entradas multiplicadas por los pesos correspondientes

Γ_k = función de activación

$Y_k = \Gamma(V_k)$ = Salida de la neurona k.

La figura muestra un conjunto de entradas (x_i) que provienen de otras neuronas o de algún estimulo externo, un conjunto de pesos (w_{ik}) que indican las fuerzas de las conexiones sinápticas. El cuerpo celular de la neurona, donde se acumulan las señales ponderadas que pasan luego por una función de activación (Γ); y por último el axón, encargado de transmitir las salidas de la neurona activada por otras neuronas.(Aguilar et al 2001)

3.1 Propiedades de las redes neuronales artificiales

- **Capacidad de adaptación:** en general, los sistemas neuronales son capaces de aprender. Algunas redes tienen la capacidad de auto-organizarse, asegurando su estabilidad como sistemas dinámicos. Una red auto-organizada puede reconocer un cambio en el problema que esta resolviendo o inclusive, puede conseguir una nueva manera de resolverlo.

- **Memoria distribuida:** el término memoria corresponde a un mapa de activación de las neuronas. La memoria es entonces distribuida sobre muchas unidades creando así resistencia al ruido. En las memorias distribuidas, es posible empezar con datos con ruido para luego conseguir los datos correctos.

- **Paralelismo:** las redes neuronales son esencialmente paralelas y con tendencia a expresarse en una "notación paralela" e implantación en hardware paralelo.

- **Tolerancia a las fallas:** la memoria distribuida también es responsable por la tolerancia a las fallas. En la mayoría de las redes neuronales, si algún nodo se destruye, o sus conexiones son alteradas levemente, entonces el comportamiento de la red como un todo se degrada muy poco. Esta característica, hace que los sistemas de computo neuronal sean extremadamente útiles para aplicaciones donde la falla en el control de los equipos sea un punto critico.

- **Capacidad de generalizar:** los diseñadores de sistemas expertos tienen dificultades en la formulación de reglas que encapsulen el conocimiento de los expertos en relación a algún problema. Una red neuronal puede aprender de reglas solo a partir de un grupo de ejemplos. La capacidad de generalización de una red neuronal es la capacidad de dar una respuesta satisfactoria a una entrada la cual no es parte del conjunto de ejemplos con la cual se entreno.

- **Facilidad de construcción:** las simulaciones computacionales de pequeñas aplicaciones pueden ser implementadas relativamente rápido.

3.2 Limitaciones en el uso de redes neurales artificiales

- Los sistemas neurales son meramente paralelos pero normalmente son simulados en máquinas secuenciales.
- El desempeño de la red es sensible a la calidad de los datos de entrada y al tipo de preprocesamiento que se le haya hecho a los mismos.
- Las redes neurales no pueden explicar el resultado que ellas obtienen; sus reglas de operación son completamente desconocidas.
- El desempeño es medido por métodos estadísticos, que genera desconfianza en algunos potenciales usuarios.
- Muchas de las decisiones requeridas en el desarrollo de una aplicación no son claras.

3.3 Arquitectura de las redes neurales artificiales

La arquitectura de la red se refiere a la forma o estructura de una red neuronal artificial. Dentro de una red neuronal, los elementos de procesamiento se encuentran agrupados por capas, una capa es una colección de neuronas, y de acuerdo a su ubicación, cada capa tiene un nombre: capa de entrada, capa oculta y capa de salida. La interconexión entre capas y la dirección en que viaja la información también son consideradas como parte de la arquitectura de la red.

3.4 Análisis de series de tiempo mediante redes neurales artificiales

Las series de tiempo han sido analizadas mediante redes neurales artificiales (RNA), las cuales son modelos computacionales basados en una aproximación a la estructura y

funcionamiento de las neuronas del sistema nervioso humano. El modelo generado, se ajusta al proceso estocástico observado y permite pronosticar el comportamiento futuro de la serie.

La RNA en su forma más simple, tiene una serie de componentes que definen su arquitectura (véase Figura 7). Los componentes son:

- Entradas
- Salidas
- Pesos
- Nodo sumatorio (punto de suma)
- Función de activación
- Sesgo

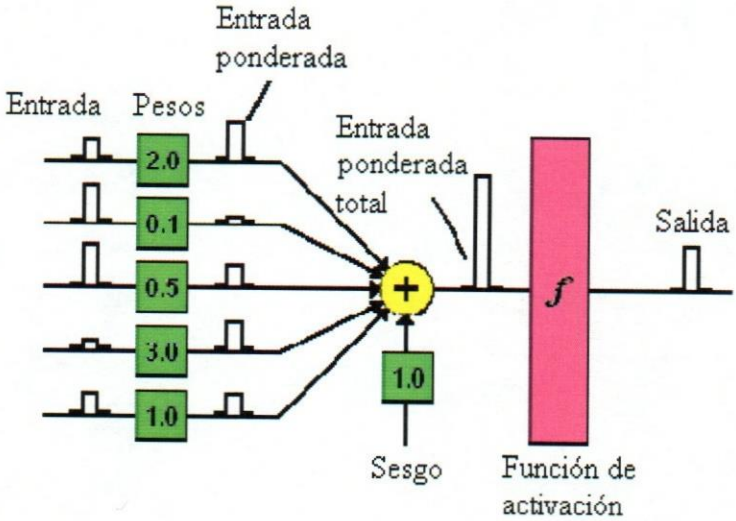

Figura 7. Arquitectura y proceso de una RNA simple

El proceso de funcionamiento de la red neuronal se describe en forma simplificada de la siguiente manera: las variables de entrada son afectadas por un valor denominado "peso", que puede ser positivo (excitatorio) o negativo (inhibitorio). Las entradas ponderadas fluyen a un nodo sumatorio donde se acumulan todas las señales de entrada multiplicadas por los pesos o ponderaciones, además éste nodo recibe la influencia del

valor sesgo (b), también afectado por un peso. La entrada ponderada total (V_k), fluye a través de una "función de activación", que transforma dicho valor en el valor de salida (Y_k).

Un modelo más complejo de red neural artificial es el que se muestra en la Figura 8. Allí se aprecia que existen R entradas, R*S pesos o ponderaciones, tres capas con S puntos de suma de las diferentes entradas y sesgos y S funciones de activación. Las salidas de las funciones de activación de la primera capa, son la entrada a los nodos de acumulación de la segunda capa y así recorre la señal a las otras capas de la red.

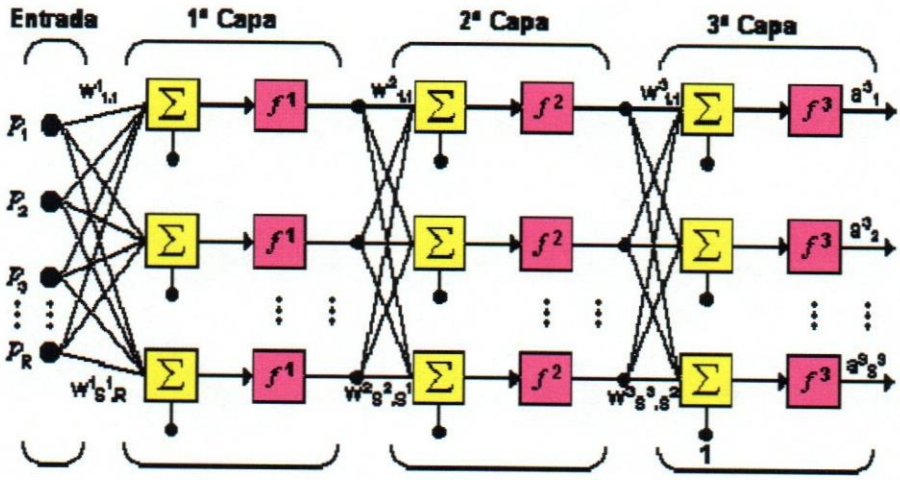

Figura 8. Red Neural de tres capas

El procedimiento para determinar la RNA mas adecuada, es un proceso mediante el cual se modifican los pesos en respuesta a una información de entrada. Los pesos pueden tener un valor de cero cuando la conexión entre las capas se elimina o un valor distinto de cero cuando se modifican o se crean conexiones entre las capas. El proceso de entrenamiento o aprendizaje es una secuencia de pasos para modificar los pesos en respuesta a una información de entrada, se realiza mediante la aplicación de un algoritmo de aprendizaje. Se trata de entrenar la red para que los pesos de interconexión minimicen el error.

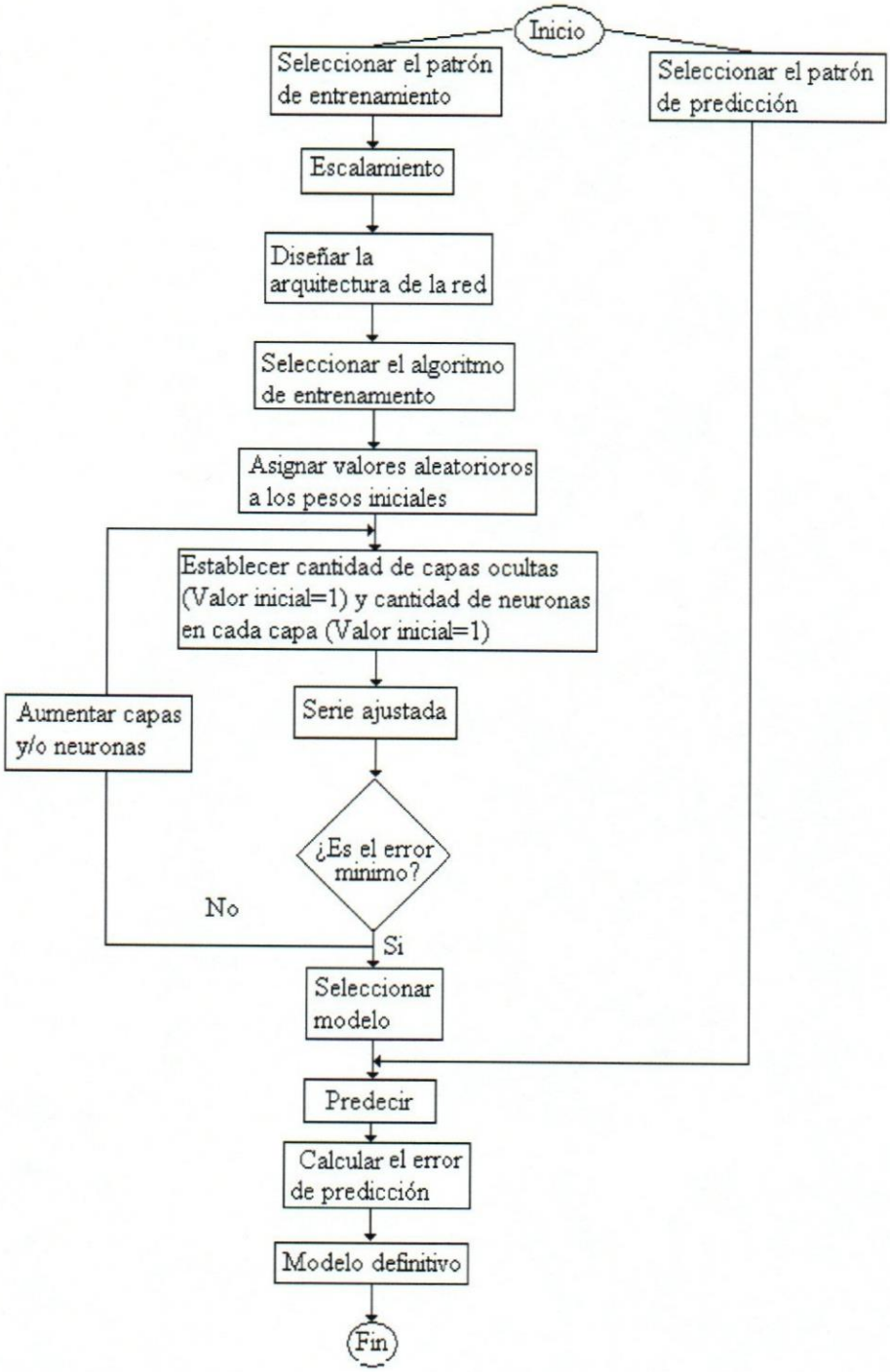

Figura 9. Diagrama de la metodología para análisis de series de tiempo mediante RNA

3.6 Materiales utilizados en redes neurales

Para la ejecución de los pasos metodológicos descritos y que se representan en el diagrama de flujo de la Figura 9, existen en la actualidad varios programas tales como *Matlab, C* y *Statistica Neural Networks*, entre otros. Este último requiere información previa sobre el patrón de entrenamiento, el patrón de predicción, el escalamiento, el número de entradas y el número de salidas. Luego el programa inicia los ciclos de entrenamiento de la red, iterando en base a diferentes funciones de activación, número de neuronas y capas y diferentes algoritmos de entrenamiento.

El programa puede ejecutar varios tipos de funciones de activación y varios tipos de entrenamientos, para escoger la red óptima. También se puede trabajar con una red personalizada por el diseñador para tener un entrenamiento específico. En todo caso es deseable que el diseñador escoja el algoritmo de entrenamiento y la función de activación de acuerdo a la función principal de la red y a la experiencia que el diseñador tenga en el área.

Evaluando los errores de entrenamiento y predicción, el programa determina la ecuación del modelo y realiza la predicción sólo en caso que éstos errores sean mínimos.

EXPLORACIÓN DE RESULTADOS

En esta sección se presentan los resultados de los experimentos aplicados a las series de números ganadores de dos loterías (Véase Tabla 2). Los experimentos se realizaron utilizando dos métodos:

- El método estadístico ARIMA (Autorregresive integrated moving average).
- El método de Redes Neuronales Artificiales (RNA).

Al final se realiza un análisis comparativo entre los resultados para detectar cual método se ajusta mejor a las series observadas.

Tabla 2. Casos de estudio

Caso	Ítem	Período
1	Valores ganadores de la Lotería del Zulia en el triple A de las 12pm	11/02/2003 - 01/09/2004
2	Valores ganadores del triple de la Lotería de Caracas en su sorteo diario	03/06/1995 – 07/04/1998

4. Caso 1. Valores Ganadores De La Lotería Del Zulia

4.1 Caracterización de los datos

Es conveniente, antes de aplicar los métodos que se mencionaron anteriormente, realizar un análisis exploratorio inicial de la serie con el fin de caracterizarlas en sus parámetros estadísticos más significativos y probar la aleatoriedad de los mismos.

El tamaño de la muestra fue de 465 observaciones. El valor del estadístico moda de la serie es 147, es decir, el número que más se repitió durante el período de evaluación. El valor mínimo fue 001 y el máximo 998. La media 481, con una desviación estándar de 288.04. El análisis exploratorio de la serie de los resultados del triple "A" de la Lotería del Zulia en su sorteo de las 12:00 m. no mostró ningún patrón significativo. Se observó una aparente aleatoriedad, tal como se muestra en el gráfico de secuencia de la figura 10.

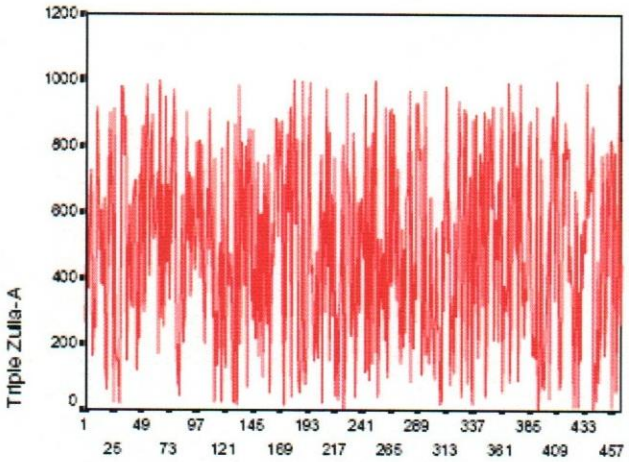

Figura 10. Serie de los resultados del triple "A" de la lotería del Zulia

La serie de los datos originales se transformó en una serie estandariza en valores entre cero y uno, que facilitará análisis posteriores (Véase figura 11).

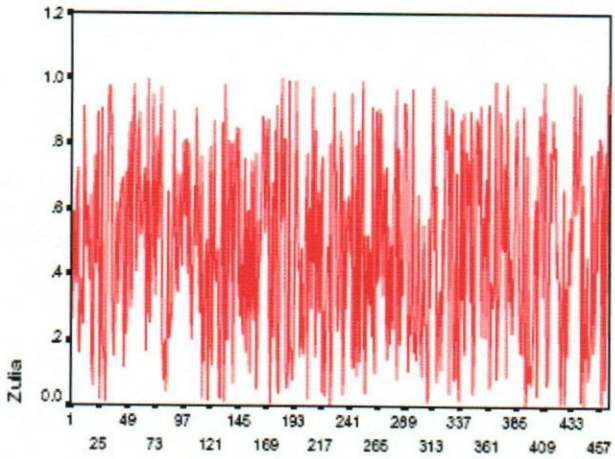

Figura 11. Serie estandarizada

4.2 Pruebas de aleatoriedad

A continuación se investiga si los valores de la serie en estudio, en el orden en que ellos han aparecido, muestran algún tipo de comportamiento significativo. Generalmente se realizan tres tipos de pruebas: la prueba Chi-cuadrado, la prueba de las rachas y la prueba de intervalos. Los resultados de dichas pruebas se muestran a continuación:

- **Prueba Chi-cuadrado**

La muestra se dividió en 25 grupos de 40 números cada uno, por lo que se tienen 24 grados de libertad. Se prueba si las frecuencias observadas y las esperadas obedecen a la misma distribución.

Tabla 3. Prueba de la Chi-cuadrado para los datos de la Lotería de Zulia

Chi-cuadrado	Grados de libertad	P
31.93548387	24	0.12860064

Como el valor P no es significativo al 5%, indica que no existen razones para sospechar que la muestra de los valores de la Lotería del Zulia en el período evaluado no sea aleatoria. Se realizó la prueba a cada dígito por separado, siendo el dígito 1 el mas significativo o las centenas, dígito 2, las decenas y el dígito 3 las unidades.

Tabla 4. Prueba Chi-cuadrado dígito por dígito

	Chi-cuadrado	Grados de Libertad	P
Digito 1	14.73626374	9	0.098439798
Digito 2	23.06451613	9	0.006053038
Digito 3	6.591397849	9	0.679576022

La hipótesis de aleatoriedad se rechazó para el dígito 2, pero no se rechaza para los otros dígitos. Para el dígito 2 se pudo observar que los números 7 y 9 aparecieron en 13.97% de las observaciones cada uno, donde el número de apariciones teórico es de 10%. El número 5 sólo apareció en el 7% de los casos y esas son las razones por la cual ésta hipótesis se rechazo.

- **Prueba de rachas**

Tomando como valor de prueba la mediana (al hacer esto se asume que la mitad de los valores debe tener valores menores que ella y la otra mitad valores superiores), se realizó la prueba de rachas, dando como resultados la tabla siguiente:

Tabla 5. Prueba de rachas para los resultados de la lotería del Zulia

	Zulia
Valor de prueba	476
Casos < Valor de prueba	232
Casos ≥ Valor de prueba	233
Casos en total	465
Número de rachas	228
Z	-0.511
Sig. asintót. (bilateral)	0.610

Se acepta la hipótesis de aleatoriedad en la secuencia de apariciones de los valores de la muestra. Igualmente se realizó la prueba de rachas dígito por dígito.

Tabla 6. Prueba de rachas digito por digito

	Dígito 1	Dígito 2	Dígito 3
Valor de prueba	4	5	5
Casos < Valor de prueba	203	207	225
Casos ≥ Valor de prueba	262	258	240
Casos en total	465	465	465
Número de rachas	218	227	238
Z	-1.110	-0.348	0.441
Sig. asintót. (bilateral)	0.267	0.728	0.659

Los tres dígitos muestran aleatoriedad en su orden de aparición.

- **Prueba de intervalos**

Este análisis fue hecho para cada posible valor del último dígito del triple. Los resultados de esta prueba se muestran en la tabla 7. No se rechazó la hipótesis de que los datos observados siguen un comportamiento similar al teórico. En conclusión, para esta prueba se asumió aleatoriedad de los datos.

Tabla 7. Resultados de la prueba de intervalos

Dígito 3	Chi-cuadrado	Grados de libertad	p
0	0.32366156	3	0.95551865
1	0.19573896	1	0.65818244
2	2.90224123	3	0.40694455
3	0.06326531	1	0.8014075
4	4.02828997	5	0.54535013
5	1.38295584	1	0.23959832
6	3.45940919	3	0.32606491
7	11.5103803	6	0.07382649
8	5.75910176	3	0.1239371
9	0.02658146	2	0.9867972

4.3 Método ARIMA

Para el ajuste del modelo ARIMA se usaron 372 observaciones, de un total de 465, que comprenden el período 11/02/2003 al 13/05/2004. El grupo de datos para la validación abarcó los valores entre 14/05/2004 y el 01/09/2004, es decir un total de 93 observaciones. A continuación se presentan los resultados obtenidos en la aplicación de la secuencia de pasos metodológicos descritos en la sección del método de ARIMA:

4.4 Identificación de los órdenes del modelo

Previo a la identificación de los órdenes del modelo, se verificó la estacionariedad de la serie, que es una condición necesaria para la aplicación del método ARIMA, mediante las siguientes pruebas:

- **Prueba de la homocedasticidad**

La inspección del "diagrama de cajas" (Véase figura 12) no es suficiente para probar que la varianza es invariante en el tiempo, por lo que se realizó la prueba de Levene (Tabla 8), con una significancia del 5%; no se rechazó la hipótesis de homocedasticidad.

Tabla 8. Prueba de la homogeneidad de varianzas

	Estadístico de Levene	Significación
Basándose en la media	0.897	0.690

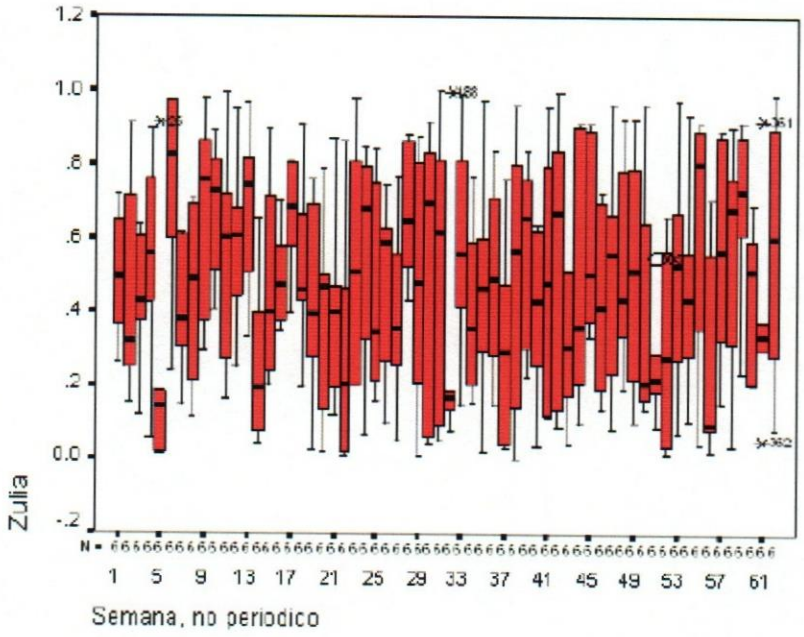

Figura 12. Diagrama de cajas

- **Estabilidad en medias y estacionalidad**

Al igual que la varianza, la media debe ser estable. Para verificar la estabilidad en medias, se utilizan los correlogramas simple y parcial, en los cuales no se observó un comportamiento que indique cambios importantes en la media, ni tampoco la existencia de estacionalidad.

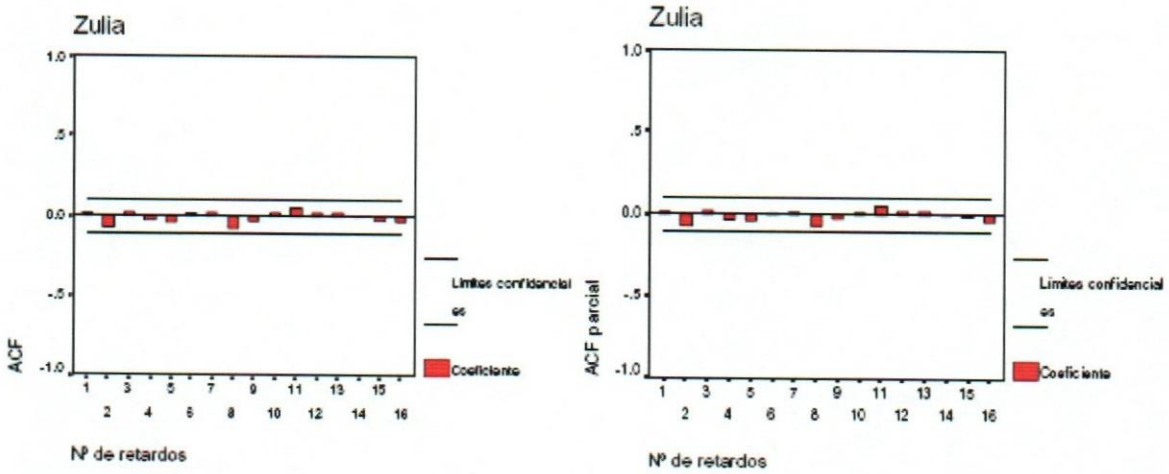

Figura 13. Correlograma simple y parcial de la serie de la Lotería del Zulia

Aplicando el método explicado en el capítulo de Metodología, se estima la cantidad de términos AR y MA necesarios para modelar la serie a partir de los correlogramas respectivos. En este caso no existe ningún coeficiente de autocorrelación simple ni de autocorrelación parcial significativo que dé sospechas de un proceso autorregresivo o de medias móviles. Por esto, se asume un modelo preliminar **ARIMA(1,0,1) sin constante.**

4.5 Estimación de los parámetros

Los parámetros asociados a los términos AR y MA son φ y θ respectivamente. Se calcularon utilizando el programa *SPSS*, a través de los métodos de mínimos cuadrados.

Tabla 9. Valores para los parámetros del modelo

	B	SEB	Proporcion-t
AR1 (φ)	0.99987182	0.00039650	2521.7282
MA1 (θ)	0.97153982	0.01987810	48.8749

4.6 Verificación del modelo

La verificación del modelo implica comprobar si los residuos tienen una naturaleza de ruido blanco y están incorrelacionados. Se verifica también si los parámetros son significativamente diferentes de cero.

En este caso, los residuos tienen una media de 0.0082, aproximadamente igual a cero.

Tabla 10. Significación de los parámetros

Prob. Aprox.	
AR1 (φ)	0,0000000
MA1 (θ)	0,0000000

La incorrelación de los residuos se muestra en la figura 14.

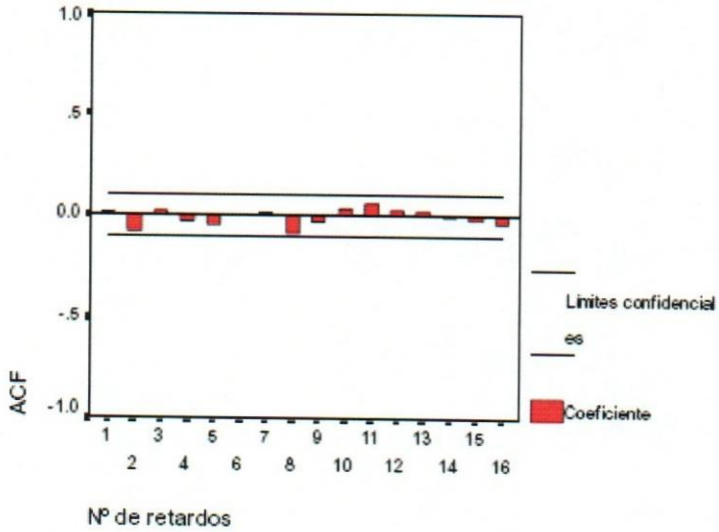

Figura 14. Correlograma simple de los residuos

El error estándar, el criterio de información de Akaike (AIC) y el criterio bayesiano de Schwarz (SBC) sirven como criterios de verificación del ajuste. A continuación se muestran los resultados de la aplicación de estos criterios para el modelo ARIMA(1,0,1) sin constante.

Tabla 11. Criterios de verificación del ajuste

Error estándar	0.29177579
Criterio de información de Akaike (AIC)	145.29331
Criterio bayesiano de Schwarz (SBC)	153.1311

- **Sobreajuste**

Un paso importante cada vez que se aplica la metodología ARIMA es el uso de la técnica del sobreajuste. En ciertos casos, el modelo que mejor se adecua a la serie no es el planteado en un principio y no precisamente el más simple.

Para el sobreajuste se evaluaron los siguientes modelos:

- ARIMA(2,0,1) con constante
- ARIMA(2,0,1) sin constante
- ARIMA(1,0,2) con constante
- ARIMA(1,0,2) sin constante

Realizando las respectivas evaluaciones en cada modelo, ninguno de los ellos mejoró el ajuste de manera significativa, por lo que se seleccionó el modelo **ARIMA(1,0,1) sin constante**, como definitivo.

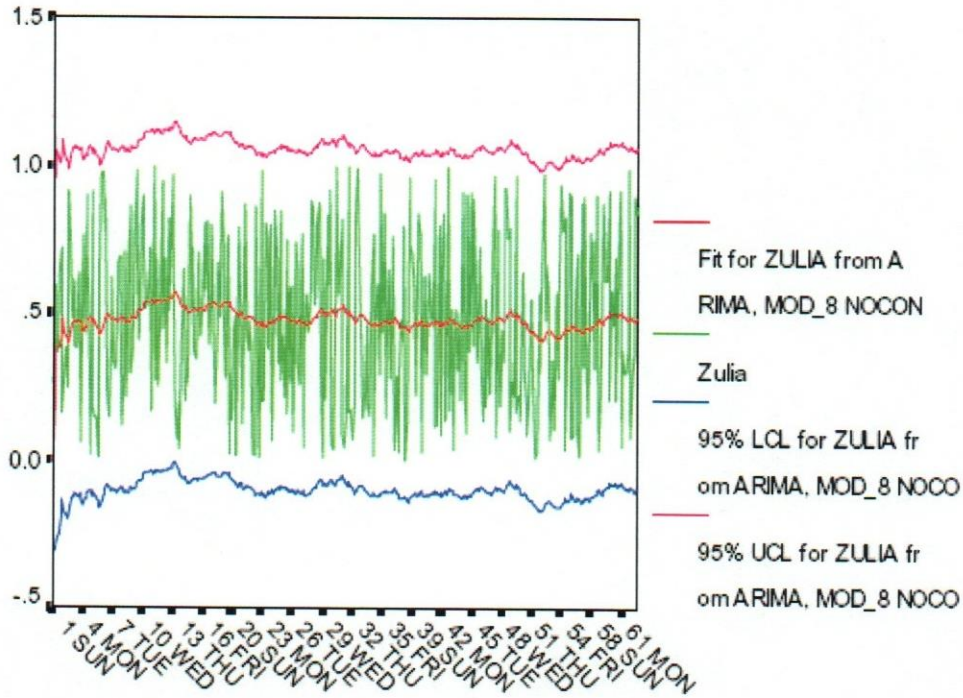

Figura 15. Serie original y serie ajustada Arima (1,0,1) sin constante

En la figura mostrada anteriormente, se representa la serie original de los resultados del triple "A" de la Lotería del Zulia y los valores del modelo ajustado Arima (1,0,1) sin constante.

- **Errores del ajuste:**

La siguiente tabla muestra diferentes tipos de errores de ajuste, que permitirán comparar los resultados del método ARIMA con los resultados del las redes neuronales artificiales.

Tabla 12. Errores del ajuste con ARIMA

Error medio	0.0082
Error medio absoluto	0.2548
Error cuadrático medio	0.0862
Raíz del error cuadrático medio	0.2936

4.7 Predicción

No se realiza la etapa de predicción en vista de que el modelo ajustado no representa adecuadamente la estructura de la serie observada.

4.8 Redes Neuronales Artificiales

La metodología de RNA descrita anteriormente sirvió de guía para el ajuste de la serie de tiempo. Los resultados obtenidos en los pasos aplicados son los siguientes:

Escalamiento y selección del patrón de datos

1. Escalamiento: los datos a los cuales se les aplicaron un escalamiento entre 0 y 1 o estandarización fueron los mismos que se utilizaron como datos de entrenamiento y verificación.

2. Patrones de entrenamiento y prueba: los patrones de entrenamientos constan de 372 observaciones que corresponde al 80% del número total de datos. Los patrones de prueba o verificación son 93 observaciones que equivalen al 20%. De esta forma se mantienen los mismos grupos utilizados en el método ARIMA, lo que hace que la comparación sea consistente.

Diseño de la arquitectura de la red

1. Interconexión: todas las redes neuronales probadas fueron de alimentación adelantada, totalmente conectadas y con dos capas ocultas.

2. Determinación del número de entradas y salidas: el número de entradas para la serie de datos de la Lotería del Zulia fue uno, así como también el número de salidas.

3. Número de capas ocultas y de nodos en cada capa: la red óptima generada por el *Intelligent Problem Solver* del programa *Statistica Neural Networks*, fue un perceptrón multicapas con dos capas ocultas y 20 neuronas en cada capa. La Figura 16 muestra un esquema de esta red.

4. Determinación de la función de activación: se utilizó la curva logística como función de activación en las neuronas de las capas ocultas y de salida, función que fue seleccionada por el *Intelligent Problem Solver* del programa S*tatistica Neural Networks*.

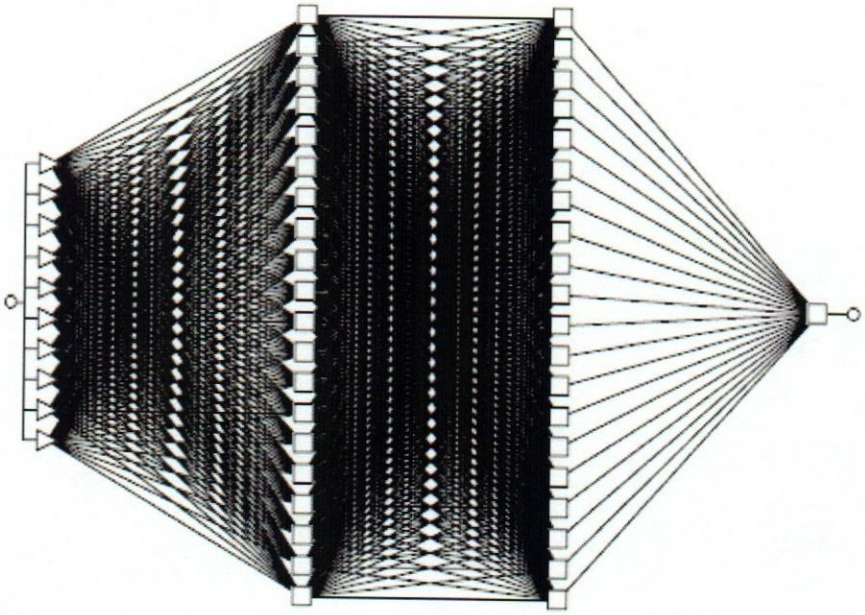

Figura 16. Esquema de la red neural aplicada para el caso de Lotería del Zulia

Entrenamiento

1. Algoritmo de entrenamiento: el algoritmo utilizado para entrenar las redes fue el de Retropropagación, selección que se realizó en base a la bibliografía estudiada.

2. Pesos iniciales: los pesos se inicializaron con valores aleatorios uniformemente distribuidos. Su rango varió entre 1 y -1.

3. Los parámetros de entrenamiento: fueron establecidos por defecto por el programa, resultando los valores que se enumeran a continuación:

    a. Numero máximo de épocas:100
    b. Tasa de aprendizaje: 0.1
    c. Incremento en la tasa de aprendizaje. 0.1
    d. Momento: 0.3

4. Errores de entrenamiento: a la red neuronal antes descrita se le calcularon los errores típicos (Tabla 13), que permitieron realizar la comparación con el modelo obtenido por el método ARIMA.

Tabla 13. Errores de entrenamiento con RNA

	RNA
Error medio	-0.018
Error medio absoluto	0.2543
Error cuadrático medio	0.0878
Raíz del error cuadrático medio	0.2911

Serie ajustada

La serie ajustada con la red neuronal antes descrita se observa en la siguiente figura:

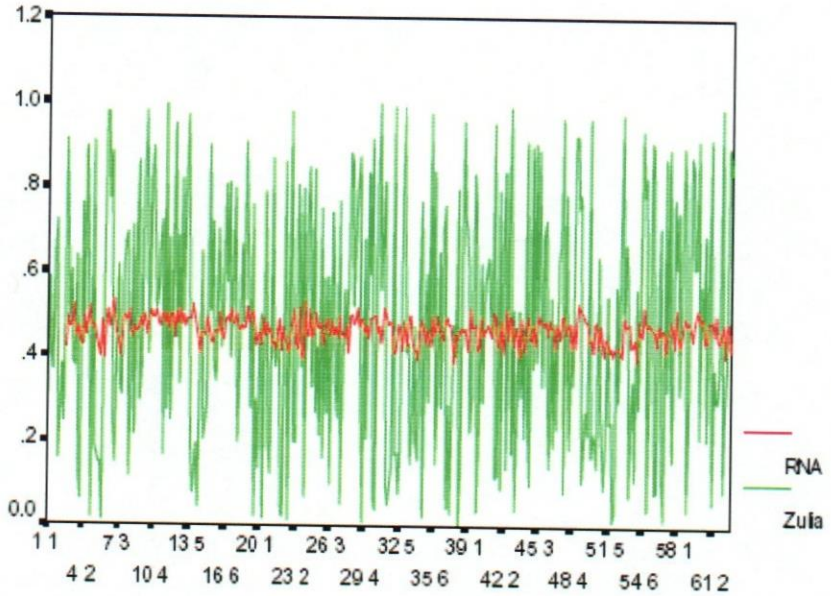

Figura 17. Serie original y serie ajustada con la RNA óptima

Predicción

El ajuste no es suficientemente representativo de la serie observada. Los errores mostrados en la siguiente tabla son bastante altos, por lo que no se realiza ninguna predicción.

Tabla 14. Errores de predicción con RNA

	RNA
Error medio	-0.00352
Error medio absoluto	0.24679
Error cuadrático medio	0.08543
Raíz del error cuadrático medio	0.2892

4.9 Comparación de los resultados entre el método ARIMA y RNA

Observando la tabla 15 se concluye que los errores entre el método ARIMA y RNA son muy grandes y similares. Sin embargo, el error medio absoluto, la raíz del error cuadrático medio son levemente menores en las RNA que los del método ARIMA.

Tabla 15. Comparación entre los errores del método ARIMA y la RNA

	ARIMA (ajuste)	ARIMA (predicción)	RNA (ajuste)	RNA (predicción)
Error medio	0.0082	0.0070	-0.018	-0.0035
Error medio absoluto	0.2548	0.2617	0.2543	0.2467
Error cuadrático medio	0.0862	0.0895	0.0878	0.0854
Raíz del error cuadrático medio	0.2936	0.2945	0.2911	0.2892

En la siguiente figura, se muestran los comportamientos de la serie observada y los ajustes con redes neuronales y con el método ARIMA.

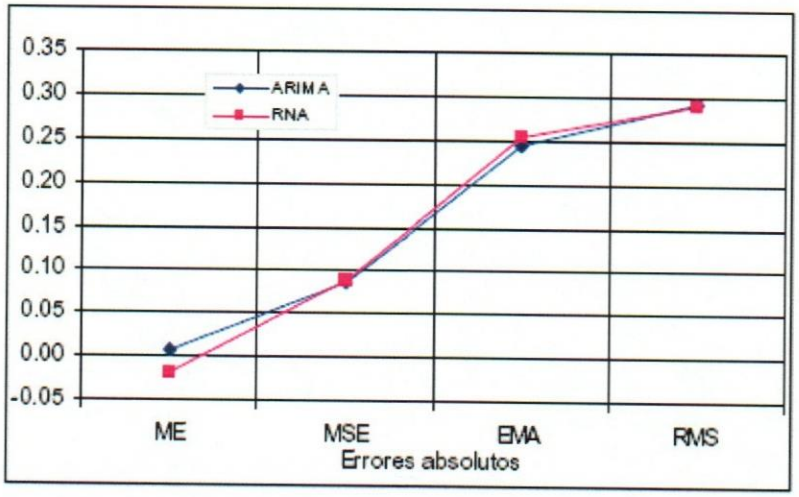

Figura 18. Errores absolutos de ajuste entre ARIMA y RNA

Los resultados de los métodos ARIMA y RNA aplicados no produjeron un ajuste satisfactorio que siguiera el comportamiento de la serie. Las pruebas de aleatoriedad confirmaron que la serie bajo estudio es aleatoria lo que hace que el ajuste y predicción de la misma tenga un error no satisfactorio.

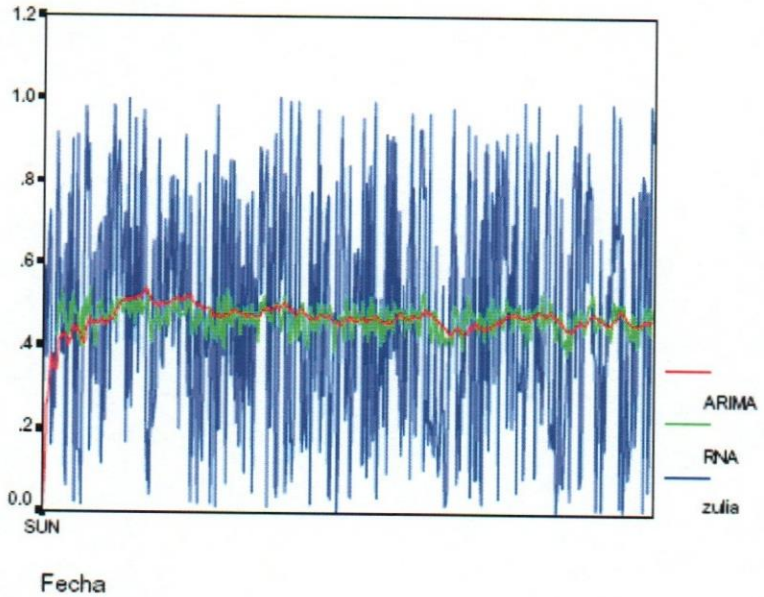

Figura 19. Serie estandarizada y series ajustadas con RNA y el método ARIMA

5. Caso 2. Valores Ganadores Del Triple De La Lotería De Caracas

5.1 Caracterización de los datos

En la siguiente figura se muestra la secuencia de la serie de los valores ganadores del triple de la Lotería de Caracas para el período 03/06/1995 al 07/04/1998, período que contiene la mayor cantidad de datos continuos sin interrupciones de esta lotería. No se observó ningún patrón significativo, por el contrario se detecta una aparente aleatoriedad.

El número más repetido fue el 342, el mínimo 002 y el máximo 998. La media fue 507 con una desviación estándar de 284.31.

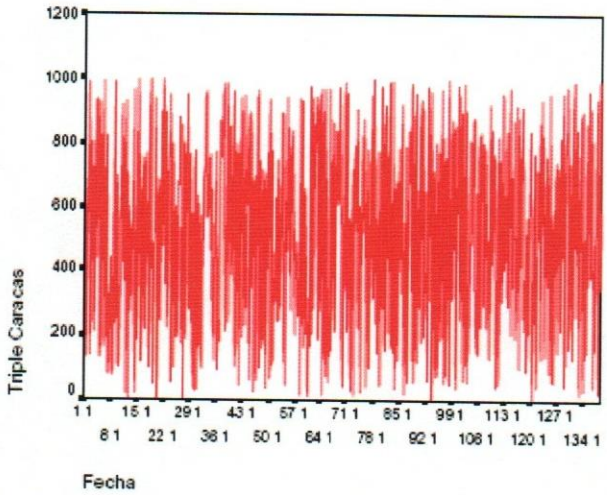

Figura 20. Serie de los resultados del triple de la lotería de Caracas

La serie de los datos originales se transformó en una serie estandariza en valores entre cero y uno, que facilitará análisis posteriores (Véase figura 21).

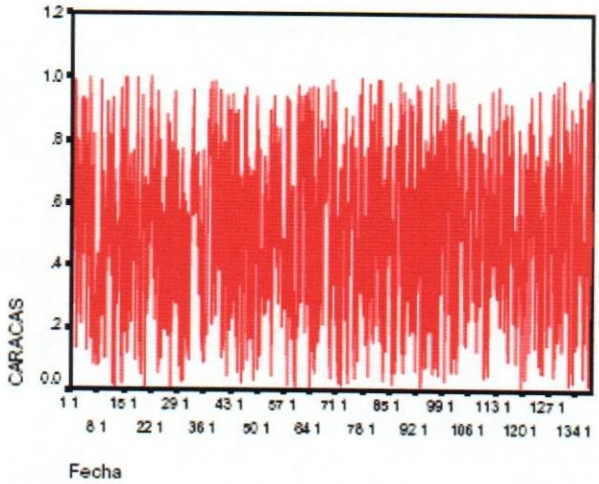

Figura 21. Serie estandarizada

5.2 Pruebas de aleatoriedad

- **Prueba Chi-cuadrado**

La muestra se dividió en 25 grupos de 40 números cada uno, por lo que se tiene una prueba Chi-cuadrado de 24 grados de libertad.

Tabla 16. Resultado de la prueba Chi-cuadrado para los datos de la Lotería de Caracas

Chi-cuadrado	Grados de libertad	P
26	24	0.353164933

Como el valor P no es significativo al 5%, indica que no existen razones para sospechar que la muestra de los valores de la Lotería de Caracas en el período evaluado no sea aleatoria.

También se realizó la prueba para cada dígito por separado, arrojando los siguientes resultados:

Tabla 17. Prueba Chi-cuadrado digito por digito

	Chi-cuadrado	Grados de Libertad	P
Digito 1	4.932038835	9	0.840191892
Digito 2	10.27184466	9	0.328929584
Digito 3	5.830097087	9	0.756795146

Se aprueba la hipótesis para todos los dígitos de la lotería de Caracas.

- **Prueba de las rachas**

Tomando la mediana como valor de prueba se tiene la siguiente tabla de resultados para la prueba de rachas:

Tabla 18. Prueba de rachas para los resultados de la lotería de Caracas

	Caracas
Valor de prueba	514
Casos < Valor de prueba	411
Casos ≥ Valor de prueba	413
Casos en total	824
Número de rachas	405
Z	-0.558
Sig. asintót. (bilateral)	0.577

No se rechaza la hipótesis nula de la aleatoriedad en el orden de aparición de los datos.

Tabla 19. Prueba de rachas digito por digito

	Dígito 1	Díigito 2	Dígito 3
Valor de prueba	5	4	5
Casos < Valor de prueba	394	326	407
Casos ≥ Valor de prueba	430	498	417
Casos en total	824	824	824
Número de rachas	405	375	396
Z	-0.504	-1.461	-1.181
Sig. asintót. (bilateral)	0.614	0.144	0.238

La prueba de rachas para cada uno de los dígitos también muestra que se acepta la hipótesis nula y se asume aleatoriedad en el orden de aparición de los datos.

- **Prueba de intervalos**

Al igual que con los resultados de la Lotería del Zulia, se tomó el último dígito del triple de la Lotería de Caracas para realizar esta prueba. En general todos los resultados indicaron aleatoriedad entre cada aparición. (Véase Tabla 20)

Sólo el resultado de la prueba para el número 7 en el dígito 3 se rechazó debido a que los intervalos de separación entre cada aparición observada fueron en ocasiones muy diferentes a los esperados.

Tabla 20. Resultados de la prueba de intervalos

Dígito 3	Chi-cuadrado	Grados de libertad	p
0	7.46555332	7	0.38206552
1	1.99049827	4	0.73750662
2	2.68239393	6	0.84752244
3	5.4846072	6	0.48331909
4	1.21979229	6	0.97589583
5	3.27355967	7	0.85859558
6	2.04757415	5	0.84252387
7	29.1438489	8	0.00029918
8	3.84891919	5	0.57136663
9	10.0842444	6	0.12114885

5.3 Método ARIMA

Para el ajuste se utilizaron 660 datos que se refieren al período del 3 de Junio de 1995 al 8 de Septiembre de 1997. Para validar la predicción se utilizaron los datos desde el 9 de Septiembre de 1997 hasta el 7 de Abril de 1998.

5.4 Identificación de los órdenes del modelo

- **Prueba de la homocedasticidad**

El estadístico de Levene para los datos de la Lotería de Caracas fue 1.126, por lo que se acepta la hipótesis de la homogeneidad de varianzas y se continúa con el estudio de la media y la comprobación de la existencia de estacionalidad.

- **Estabilidad en medias y estacionalidad**

Se observa en las siguientes figuras que no existe ningún coeficiente de la función de autocorrelación simple ni de la parcial, por lo que se selecciona el modelo inicial: **ARIMA(1,0,1) con constante.**

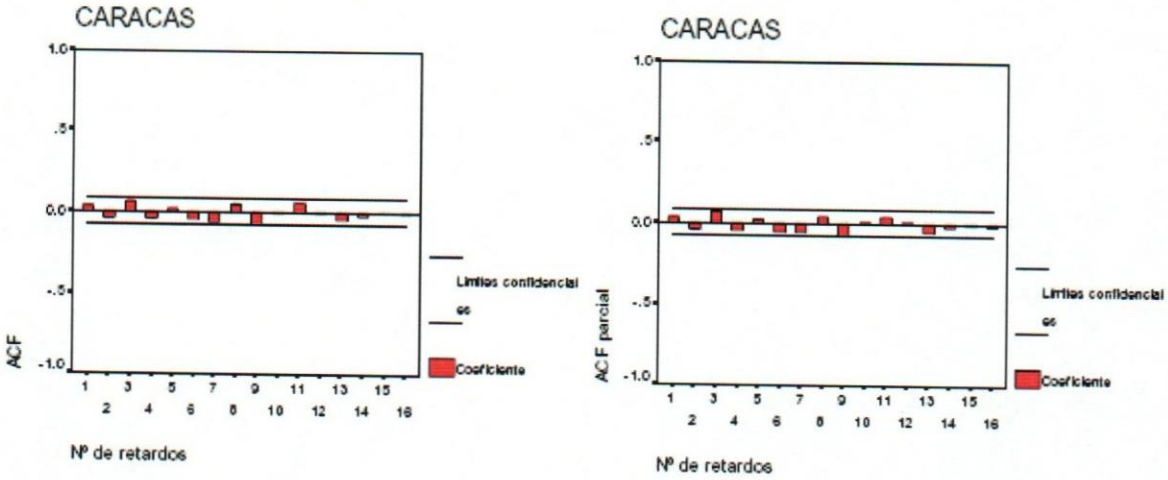

Figura 22. Correlograma simple y parcial de la serie Lotería del Caracas

5.5 Estimación de los parámetros

Los valores de los parámetros para el modelo ARIMA (1,0,1) con constante son los mostrados en la siguiente tabla:

Tabla 21. Valores para los parámetros del modelo

	B	SEB	Proporcion-t
AR1 (φ)	0.25189848	0.25189848	-2.590712
MA1 (θ)	-0.71531362	0.23230074	-3.079257
Constante (δ)	0.50653280	0.01163029	43.552903

5.6 Verificación del modelo

Para verificar el modelo se calcula la media de los residuos que en este caso fue de $3.6x10^5$, lo que es aproximadamente igual a cero. La incorrelación de los residuos se muestra en el siguiente gráfico.

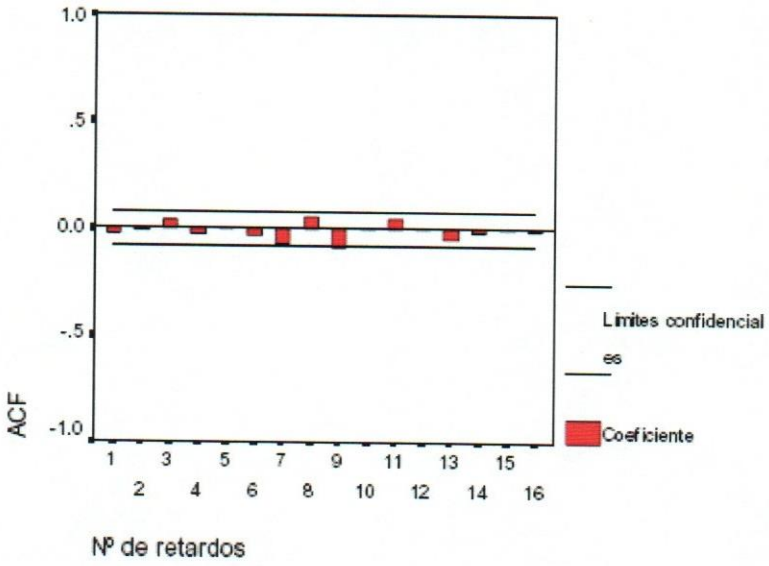

Figura 23. Correlograma simple de los residuos

Las probabilidades aproximadas para estudiar la hipótesis que los parámetros son iguales a cero se muestran en la siguiente tabla. Como son menores que el nivel de significación α, establecido en 0.05, se rechaza la hipótesis.

Tabla 22. Significación de los parámetros

PROB. APROX.	
AR1 (φ)	0.00979030
MA1 (θ)	0.00216170
Constante (δ)	0,00000000

Los resultados de los criterios de verificación del ajuste se muestran en la siguiente tabla:

Tabla 23. Criterios para la verificación del ajuste

Error estándar	0.28787255
Criterio de información de Akaike (AIC)	232.29925
Criterio bayesiano de Schwarz (SBC)	245.77597

Al realizar el sobreajuste no se mejoró la adecuación de la serie por lo que se mantuvo el modelo **ARIMA(1,0,1) con constante.**

La siguiente figura, muestra la serie original, la serie ajustada y los intervalos de confianza al 95%. Se repite la conclusión del caso anterior (Lotería del Zulia), el ajuste no representa adecuadamente a la serie.

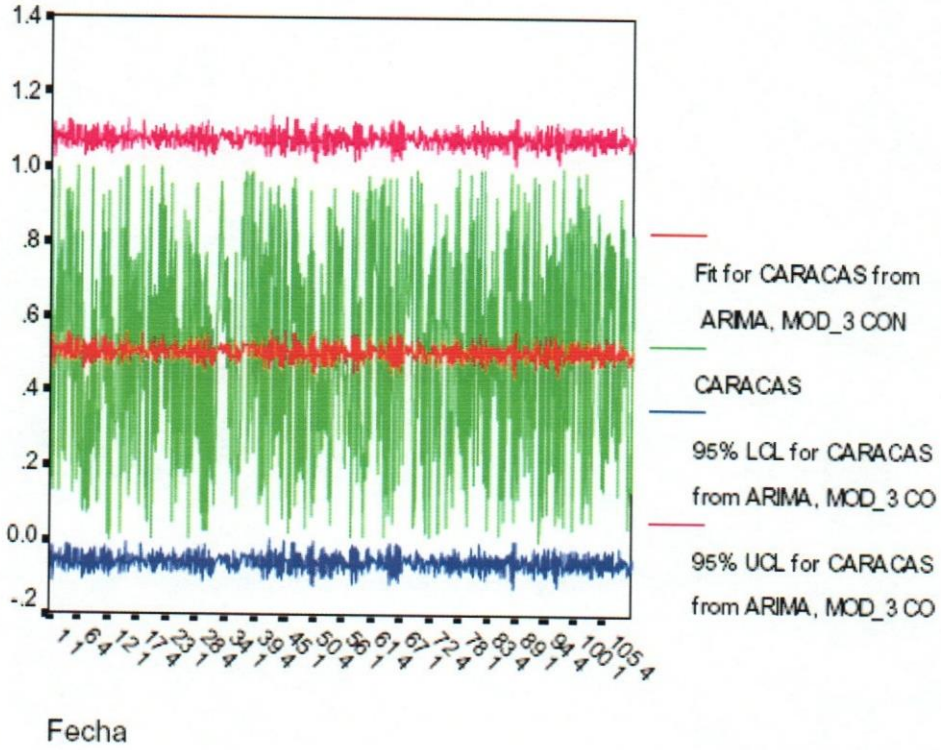

Figura 24. Serie original y serie ajustada Arima (1,0,1) sin constante Lotería de Caracas

- **Errores del ajuste**

A continuación se indican los resultados de los diferentes errores calculados para el modelo seleccionado, los cuales permitirán compararlo con el resultado de las redes neuronales artificiales.

Tabla 24. Errores del ajuste con ARIMA

Error medio	0.0000
Error medio absoluto	0.2494
Error cuadrático medio	0.0829
Raíz del error cuadrático medio	0.2879

Los errores porcentuales del error medio (MPE) y del error en la media absoluta (MAPE) resultaron indeterminados y por lo tanto no se aplican.

5.7 Predicción

No se realiza la etapa de predicción en vista de que el modelo ajustado no representa adecuadamente la estructura de la serie observada.

5.8 Redes Neuronales Artificiales

Los resultados de las RNA aplicados a los valores de la Lotería de Caracas fueron obtenidos siguiendo la misma metodología que se aplicó en el caso de los valores de la Lotería del Zulia. A continuación se detalla:

Escalamiento y selección del patrón de datos

Los datos se sometieron a un proceso de escalamiento entre cero y uno. Los patrones de entrenamientos fueron 660 observaciones que son el 80% del total. Los patrones de prueba fueron 164 observaciones equivalentes al 20% restante.

Diseño de la arquitectura de la red

Todas las redes neuronales probadas fueron de alimentación adelantada, totalmente conectadas y dos capas ocultas. El número de entradas para la serie de datos de la Lotería de Caracas fue uno, así como también el número de salidas. La red óptima generada por el *Intelligent Problem Solver* del programa *Statistica Neural Networks*, fue un perceptrón multicapas con dos capas ocultas y 13 neuronas en cada capa. La función de activación fue la logística para todas neuronas de las capas ocultas y de la capa de salida (Véase Figura 25).

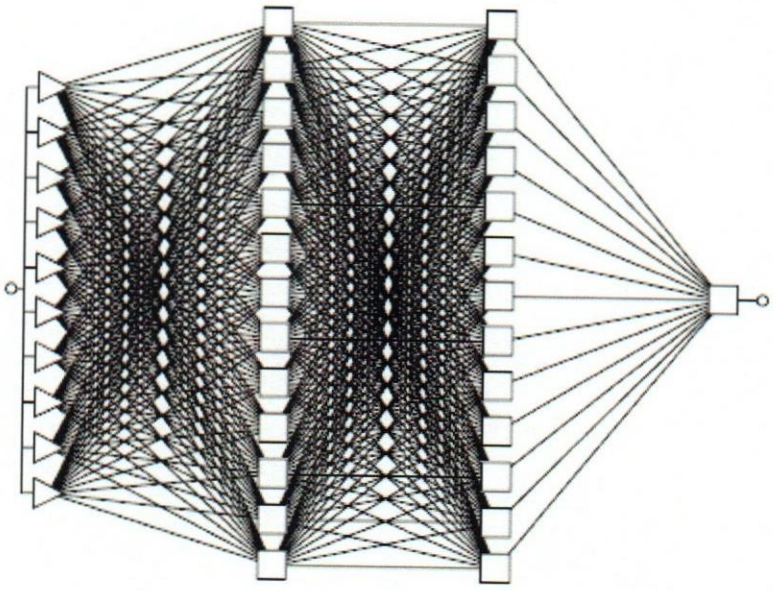

Figura 25. Esquema de la Red Neural aplicada para el caso de la Lotería de Caracas

Entrenamiento

El algoritmo utilizado fue el de Retropropagación, con pesos iniciales aleatorios uniformemente distribuidos entre -1 y 1.

Los parámetros de entrenamiento fueron:

a. Numero máximo de épocas:100

b. Tasa de aprendizaje: 0.1

c. Incremento en la tasa de aprendizaje. 0.1

d. Momento: 0.3

Los errores de entrenamiento aparecen indicados en la siguiente tabla, los cuales permitirán realizar una comparación con los errores obtenidos por el método **ARIMA**.

Tabla 25. Errores de entrenamiento con RNA

	RNA
Error medio	-0.02705
Error medio absoluto	0.2525824
Error cuadrático medio	0.08922
Raíz del error cuadrático medio	0.2916

Serie ajustada

La figura 26 muestra la serie ajustada con la red neuronal antes descrita y los valores de la serie original. Se observa que el ajuste no es suficientemente representativo.

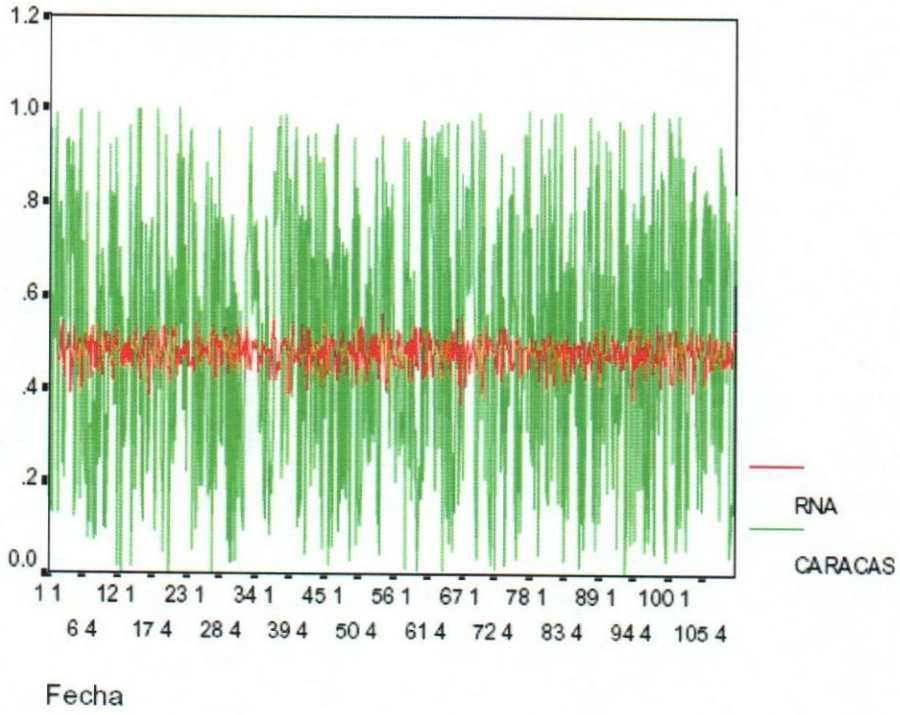

Figura 262. Serie original y serie ajustada con la RNA óptima

Predicción

Tomando la red óptima se generaron los valores de predicción para la Lotería de Caracas. Estos resultados se pueden observar en la figura 27.

Los errores observados luego de la predicción fueron los siguientes:

Tabla 26. Errores de predicción con RNA

	RNA
Error medio	-0.03751
Error medio absoluto	0.2536
Error cuadrático medio	0.0868
Raíz del error cuadrático medio	0.2947

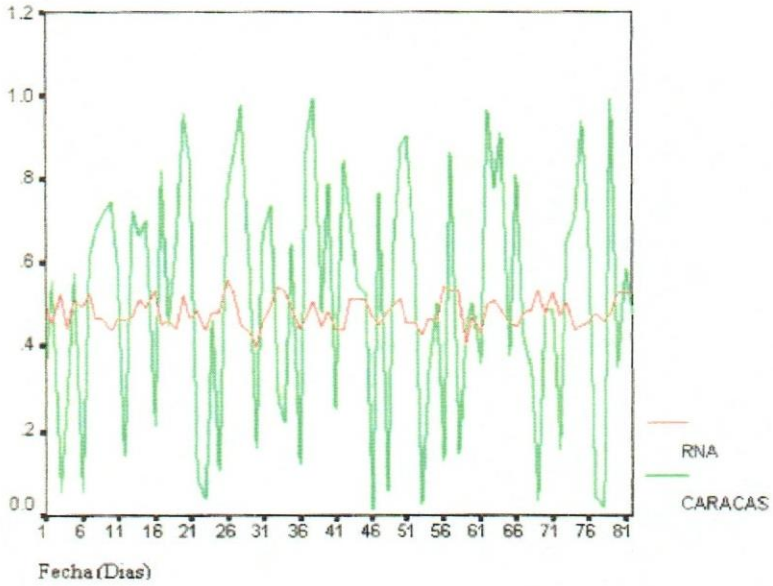

Figura 27. Predicción con RNA

5.9 Comparación de los resultados entre el método ARIMA y RNA

Al aplicar los métodos ARIMA y RNA a los datos de la Lotería de Caracas resultan errores de magnitudes muy grandes y similares. Sin embargo éstos últimos, son levemente inferiores a los del método ARIMA. (Véase Tabla 27).

Tabla 27. Comparación entre los errores del método ARIMA y la RNA

	ARIMA (ajuste)	ARIMA (predicción)	RNA (ajuste)	RNA (predicción)
Error medio	0.0000	0.0024	-0.02705	-0.03751
Error medio absoluto	0.2494	0.2290	0.2525	0.2536
Error cuadrático medio	0.0829	0.0746	0.0892	0.0868
Raíz del error cuadrático medio	0.2879	0.2732	0.2916	0.2947

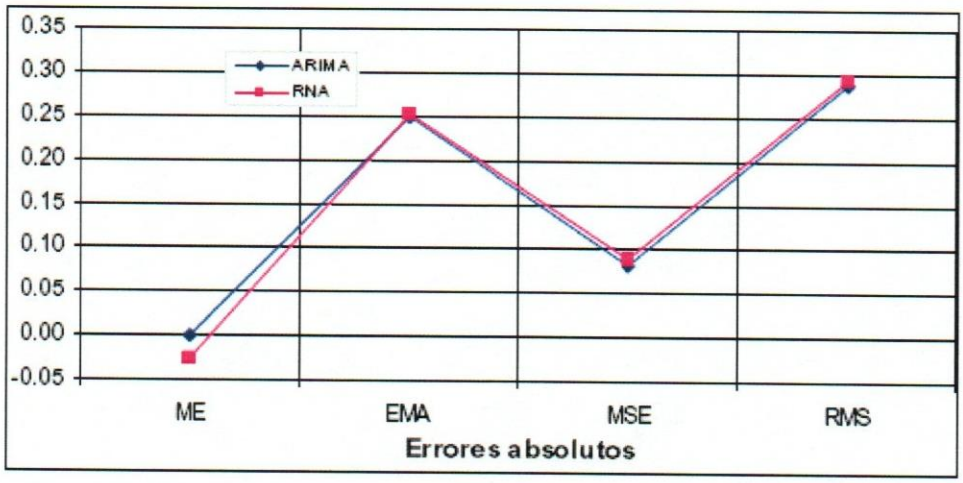

Figura 28. Comparación de los errores absolutos entre el método ARIMA y la RNA

Para finalizar la comparación entre los métodos se presenta en la siguiente figura la serie original y las ajustadas por el método ARIMA y RNA

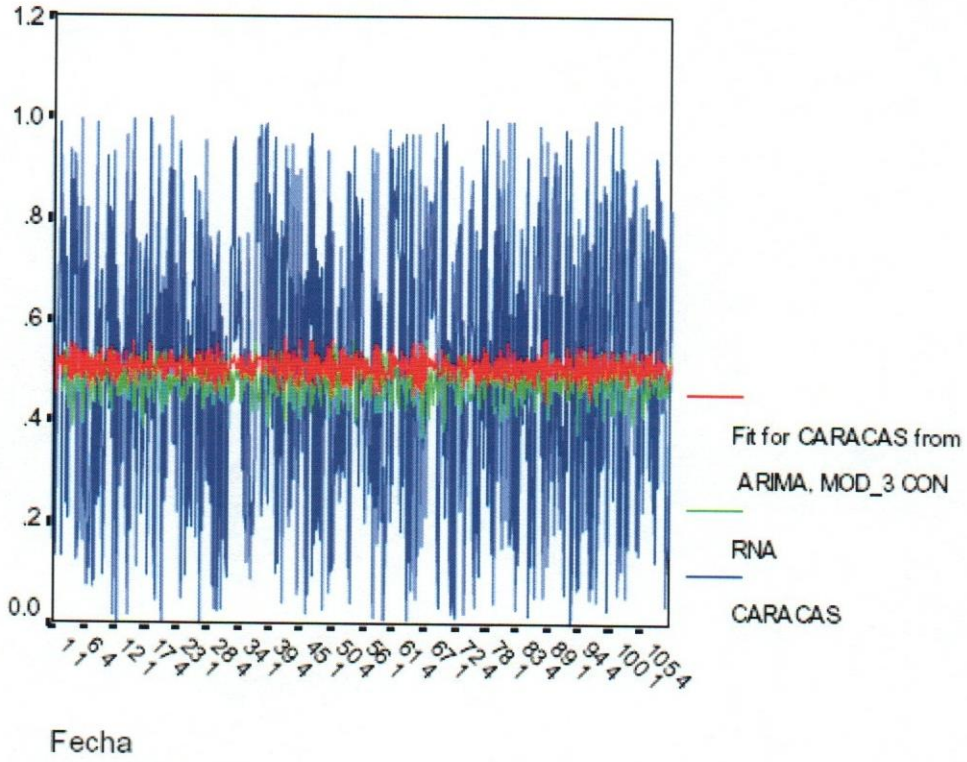

Figura 29. Serie estandarizada y series ajustadas con RNA y el método ARIMA

La condición de aleatoriedad se demostró con las pruebas realizadas; así mismo los modelos de RNA y ARIMA no consiguieron ningún patrón por lo que no hubo un ajuste satisfactorio para la serie de la Lotería de Caracas.

CONCLUSIONES

El objetivo principal de esta investigación fue realizar un análisis exploratorio sobre los datos manejados en las principales loterías de Venezuela, realizando una comparación entre modelos estadísticos y redes neurales artificiales.

Se realizó la caracterización de la serie de datos de dos loterías venezolanas (Lotería del Zulia y Lotería del Caracas). La aplicación de las pruebas de aleatoriedad: la prueba Chi-cuadrado, la prueba de las rachas y la prueba de intervalos; indicaron que los valores no son suficientemente significativos para rechazar la hipótesis de aleatoriedad de los datos.

La comparación de los histogramas de frecuencia entre las dos loterías, mostró que la Lotería de Caracas presenta más uniformidad entre sus valores, sin embargo ambas loterías están cerca de la zona de rechazo para la hipótesis de aleatoriedad.

La red óptima generada para la serie de la Lotería de Caracas, fue un perceptrón multicapas con dos capas ocultas y 13 neuronas en cada capa. Para la Lotería del Zulia, también fue un perceptrón multicapas con dos capas ocultas y 20 neuronas. Ambas redes fueron entrenadas con el algoritmo de retropropagación.

A lo largo del desarrollo de la investigación y con los resultados obtenidos en las diferentes pruebas realizadas, se pudo corroborar el potencial que poseen las técnicas de las redes neurales artificiales. Estas redes mostraron mejor habilidad para representar modelos de predicción debido a la capacidad que poseen para encontrar relaciones inesperadas en la muestra y ajustarlo a un patrón; sin embargo los errores de ajuste y entrenamiento generados por las dos técnicas (ARIMA y RNA) fueron muy similares y de magnitudes muy alta. De igual forma la predicción generó errores altos de tal forma que tendría poca utilidad predecir resultados futuros en estos tipos de datos.

BIBLIOGRAFÍA

Acosta, M. Zuluaga, C. (2000). Redes Neuronales. Disponible en: http://ohm.utp.edu.co/neuronales/main.htm

Aguilar, J. y F. Rivas, (2001). "*Introducción a las técnicas de computación inteligente*". Universidad de Los Andes. Mérida, Venezuela.

Aranaz, M. (1996). "*SPSS para Windows: Programación y análisis estadístico*". Madrid, España. McGraw-Hill/Interamericana de España.

Bao, H. (s.f.). Neural Network Models. Disponible en: http://www.netnam.vn/unescocourse/knowlegde/63.htm.

Barreto, E. y F. Ávila. (2001) "*Reconstrucción de data caótica mediante métodos basados en redes neuronales artificiales y dinámicas no lineales*". Tesis de grado (Ingeniero de Sistemas). Universidad Metropolitana. Caracas, Venezuela.

Bowerman, B. y R. O'Connel, R. (1993). "*Forecasting and time series: an applied approach*". California, USA. Duxbury Press.

Box, G., G. Jenkins y G. Reinsel, (1994). "*Time series analysis, forecasting and control*". New Jersey, USA. Prentince Hall.

Chatfield, C. (1980). "*The analysis of time series: an introduction*". Londres, Inglaterra: Chapman and Hall.

Collantes, J. (2001). "*Predicción con redes neuronales: comparación con las metodologías de Box y Jenkins*". Trabajo de grado (Magíster Scientiae en Estadística Aplicada) Universidad de Los Andes. Mérida – Venezuela.

Ferrán, M. (1996). "SPSS para Windows: Programación y análisis estadístico". Madrid, España. McGraw-Hill/Interamericana de España.

Gomez, M (2004). "Redes de retropropagación (back-prop)". Disponible en: http://www.iiia.csic.es/~mario/rna/tutorial/RNA_backprop.html

Gujarati, D. (1997). "*Econometría*". Santa Fe de Bogota, Colombia. McGraw-Hill

Hagan, M., H. Demuth y Beale, M. (1996). "*Neural networks design*". Massachussets, USA. PWS Publishing Company.

Molinero, L. (2004). Análisis de series temporales. Disponible en: http://www.seh-lelha.org/tseries.htm.

Mora C., (1996) "*Modelos Arima: Poblaciones de pequeños mamíferos en la selva nublada de Mérida*". Trabajo de grado. (Magíster Scientiae en Estadística Aplicada) Universidad de los Andes. Mérida, Venezuela.

Nau, R. (2004). Introduction to ARIMA. Disponible en: http://www.duke.edu/~rnau/411arim.htm

Plummer, E. (2000). "*Time Series Forecasting With Feed-Forward Neural Networks*". Disponible en: http://www.karlbranting.net/papers/plummer/Paper_7_12_00.htm

Spiegel, M. (1961). "*Theory and problems of Statistics*". New York, USA: Schaum Publishing Company.