

La Especificación de Modelos de Series de Tiempo Una Propuesta para el Caso no Lineal

Myladis Cogollo Flórez
Departamento de Ciencias Básicas
Universidad EAFIT
Medellín, Colombia
E-mail: mcogollo@eafit.edu.co

Alejandro Palacio Vásquez
Departamento de Ciencias Básicas
Universidad EAFIT
Medellín, Colombia
E-mail: apalac19@eafit.edu.co

Resumen—En este trabajo se propone una metodología para la identificación correcta de modelos de series de tiempo no lineales que captura el comportamiento real de las series temporales y le permita al investigador obtener mayor comodidad en el tratamiento de datos no lineales. Garantizándole la elección del mejor modelo basado en diferentes pruebas de hipótesis de linealidad y criterios de decisión. La metodología es validada por medio de datos simulados y datos reales, los resultados obtenidos son satisfactorios y muestran el buen comportamiento de la misma.

Index Terms—Identificación de modelos no lineales, modelos de series de tiempo no lineales, pruebas de hipótesis de linealidad, criterios de decisión.

I. INTRODUCCIÓN

Es común, en distintas áreas del conocimiento, obtener una muestra de datos de series temporales, producidos bien sea en distintos períodos de tiempo o en más de un proceso, y modelar dichos datos con el fin de identificar un patrón de comportamiento que permita a su vez, la optimización de procesos o el pronóstico del comportamiento futuro (véase por ejemplo [1], [2], [3]). Usualmente el estudio de las series temporales se hace a través de modelos lineales; sin embargo en la literatura se encuentra que este tipo de modelos no son capaces de capturar todo el comportamiento de los datos, lo que implicaría considerar modelos no lineales.

Gooijer y Kumar [4] realizan una revisión de la literatura de modelos de series de tiempo no lineales y pruebas de hipótesis para detectar no linealidad. Ellos estudian diferentes modelos, sus supuestos, su formulación matemática y tanto sus ventajas como desventajas. Los autores resaltan los siguientes modelos: el modelo bilineal (BL, por su sigla en inglés), modelo autorregresivo exponencial (EAR, por su sigla en inglés), modelo autorregresivo con umbral (TAR), y finalmente los modelos autorregresivo con umbral *Self-Exciting* (SETAR) y el modelo autorregresivo de transición suave (STAR), los cuales son variaciones del modelo TAR. Éstos modelos a pesar de su antigüedad siguen siendo hoy en día los modelos comúnmente usados en las diferentes aplicaciones existentes y han perdurado sin modificaciones significativas en sus especificaciones (véase por ejemplo [5], [6], [7] y [8]).

Por otra parte, la estrategia de modelado de series de tiempo se puede dividir en varias etapas principales: identificación, estimación, diagnóstico y finalmente la validación, cada etapa

es igual de complicada y abarca una serie de falencias metodológicas que dificultan el proceso de modelado como tal. Específicamente en el caso no lineal para hacer la identificación del modelo, se han propuesto diferentes estrategias paso a paso. Por ejemplo, en [9] se plantea la siguiente estrategia:

1. Hacer tests o pruebas para verificar la no linealidad de las series.
2. Considerar un pequeño grupo de modelos no lineales.
3. Hacer la estimación con dichos modelos y verificar los resultados por medio de pruebas de identificación.

Sin embargo, los autores recalcan que dicha estrategia no es totalmente exitosa debido a que hay diferentes casos de no linealidad en los datos, y por ende diferentes modelos, métodos de estimación y diferentes pruebas de hipótesis para verificarlo. Lo cual hace que la cantidad posible de combinaciones pruebas-modelos-estimación hagan muy variable la calidad del resultado. Luukkonen [10], resalta que es necesario abordar con mayor profundidad el tema de la identificación de modelos no lineales, puesto que es una temática poco estudiada y deben existir otras pruebas de identificación que funcionen correctamente.

En la literatura, se observa que si bien muchas pruebas de hipótesis para detectar no linealidad han sido propuestas con el propósito de ayudar a solucionar los problemas de la identificación de modelos no lineales (ver [4]), y sus propiedades han sido investigadas, se nota que hace falta una especie de guía o algoritmo, entendido como una secuencia de pasos previamente establecidos, que recopile toda la información existente para poder seleccionar apropiadamente el modelo que mejor se adapte al comportamiento de los datos.

En este trabajo se propone una metodología de especificación de modelos de series de tiempo no lineales, que capture el comportamiento real de las series temporales, y le permita al investigador obtener mayor comodidad en el tratamiento de datos no lineales garantizándole la especificación o elección del mejor modelo basado en diferentes pruebas de hipótesis.

La originalidad e importancia de la metodología propuesta se basa en los siguientes aspectos:

1. En la literatura se encontró que si bien los modelos lineales de series de tiempo han sido ampliamente estudiados (véase por ejemplo [11], [12] y [13]); Para este caso, existen diferentes modelos, enfoques y pruebas de hipótesis que permiten hacer un estudio muy amplio y

detallado. Sin embargo, en el caso de los modelos no lineales, a pesar de que ellos han sido estudiados por varios autores (véase por ejemplo [4], [14] y [9]) no han logrado tener tanto auge como en el caso lineal debido a los problemas que surgen en cuanto a su identificación, estimación y validación.

2. Aunque en la literatura se han formulado distintos modelos no lineales, junto con sus métodos de estimación y validación, así como pruebas de hipótesis para identificar la no linealidad de los datos; el proceso de identificación del modelo se hace por tanteo y error, es decir, el investigador debe jugar con todas las posibles combinaciones de estos modelos y pruebas, y comparar los resultados obtenidos. Esto debido a que no existe una metodología que recopile éstos modelos y pruebas, y constituya una guía paso a paso amigable que permita identificar no sólo si se debe emplear un modelo no lineal sino también seleccionar el mejor modelo para los datos objeto de estudio.

La hoja de ruta está estructurada como sigue: en la sección 2 se presenta la identificación de modelos no lineales, haciendo énfasis en los modelos no lineales más estudiados en la literatura, las pruebas de linealidad existentes y los métodos propuestos de identificación. En la sección 3 se presenta la metodología propuesta. En la sección 4 se presentan los resultados de los diversos experimentos realizados con la metodología propuesta. En la sección 5 la discusión de los resultados y las conclusiones pertinentes, para finalmente presentar la bibliografía consultada.

II. IDENTIFICACIÓN DE MODELOS NO LINEALES

En esta sección se presenta un breve resumen de las herramientas principales utilizadas en la modelación de series temporales no lineales. Se hace un breve estudio sobre los modelos no lineales más relevantes que han sido propuestos en la literatura, su formulación, parámetros y funcionamiento. Se identifican las pruebas de hipótesis existentes para determinar linealidad y también aquellas pruebas que son específicas entre modelos, además, se identifican cuáles son los criterios de selección más utilizados e importantes en el momento de decidir el modelo ideal. Finalmente se presentan diferentes metodologías halladas en la literatura de cómo hacer el proceso de identificación y especificación de series no lineales.

En general en cuanto al modelado no lineal de las series temporales, se destacan varios problemas (ver [9]). El primero, es que no existe en la práctica algún tipo de teoría o estudio que permita decidir a partir de la observación de la serie si el fenómeno subyacente es no lineal. Segundo, la selección de un modelo no lineal apropiado, que sea capaz de capturar totalmente el comportamiento de los datos y, tener algún atractivo intuitivo. Una vez que un modelo es seleccionado, el siguiente problema es encontrar suficiente evidencia en los datos para sustentar la decisión de abandonar el enfoque lineal y dar paso al no lineal.

II-A. Modelos No Lineales

Sea $\{Y_t\}$ una serie de tiempo observada en el tiempo en los puntos $t = 1, 2, 3, \dots$. Una relación entre $\{Y_t\}$ y los valores

finitos de su historia $\{Y_{t-1}\}$ y $\{\epsilon_t\}$ se define de la forma:

$$Y_t = h(Y_{t-1}, \dots, Y_{t-p}, \epsilon_{t-1}, \dots, \epsilon_{t-q}) + \epsilon_t \quad (1)$$

donde ϵ_t es un ruido blanco, es decir, una secuencia de variables aleatorias independientes, idénticamente distribuidas, con media 0 y varianza σ^2 , y $h(\cdot)$ es una función no lineal definida en los reales.

Basados en la ecuación (1) y en transformaciones de la función $h(\cdot)$, se han propuesto diferentes alternativas de modelos no lineales, los cuales varían tanto en su especificación como estimación (véase por ejemplo [4] y [15]). Entre éstos modelos se destacan los expuestos en la tabla I. Dicha tabla presenta el nombre del modelo no lineal junto con el año en que se propuso y su autor. Además en la última columna se muestra la formulación matemática de cada modelo.

Estos modelos han logrado persistir a lo largo del tiempo como los más empleados en el estudio de series temporales no lineales. Diversas aplicaciones con datos reales y simulados se han realizado y además existen librerías e implementaciones en softwares especializados (tales como R, Matlab, EViews, Stata, entre otros) de estos modelos. Con respecto al proceso de estimación de los parámetros de los modelos presentados en la tabla I, en [9] se presenta el estado del arte de las diferentes técnicas o métodos de estimación para los diferentes modelos no lineales presentados anteriormente. En general se observa para los modelos no lineales paramétricos, que el método de estimación más común son los mínimos cuadrados no lineales condicionales, el cual es equivalente al método de máxima verosimilitud condicional. Mientras que para los modelos no paramétricos, el método más común, es aplicar el método de Kernel.

II-B. Pruebas de Hipótesis de No Linealidad

Las pruebas de hipótesis de no linealidad consisten en un proceso mediante el cual se desea determinar, bajo un nivel de significancia dado, determinar la no linealidad de una serie temporal en particular. En general, para una serie de tiempo $\{X_t\}$ se desea probar:

$H_0 : X_t$ es un proceso lineal.

$H_1 : X_t$ no es un proceso lineal.

Hasta la fecha, varios investigadores han estudiado y propuesto el conjunto de las pruebas de hipótesis para detectar no linealidad en los datos (véase por ejemplo [4] y [15]). En la literatura, se destacan los siguientes tres grandes grupos de pruebas de hipótesis para obtener el estadístico:

1. Multiplicadores de Lagrange (*LM*)
2. Pruebas de Razón de Verosimilitud (*LR*)
3. Pruebas de Wald (*W*)

En la prueba de multiplicadores de Lagrange (*LM*) [16], requiere que la hipótesis nula sea expresada como una o más restricciones en los valores de los parámetros, además, estimar los parámetros de las restricciones. El estadístico de prueba está dado por:

Modelo	Autor (Año)	Formulación Matemática
Modelo Estado Dependiente (SDM)	Maurice Priestley (1980)	$Y_t = \mu(z_{t-1}) + \sum \phi_i(z_{t-1})Y_{t-1} + \epsilon_t + \sum \theta_j(z_{t-1})\epsilon_{t-1}$
Modelo Bilineal (BL)	Granger y Andersen (1978)	$Y_t = \sum_u \sum_v c_{uv} Y_{t-v} \epsilon_{i-u}$
Modelo Autoregresivo Exponencial (EAR)	Jones, Ozaki y Oda (1978)	$Y_t = \mu + \sum_i \{\phi_i + \pi_i \exp(-\gamma Y_{t-1}^2)\} Y_{t-1} + \epsilon_t$
Modelo Autoregresivo con Umbral (TAR)	Howell Tong (1977)	$Y_t = \phi_0^{(j)} + \sum_t \phi_t^{(j)} Y_{t-1} + \epsilon_t^{(j)}$
Modelo SETAR	Howell Tong (1980)	$Y_t = \phi_0^{(j)} + \sum_t \phi_t^{(j)} Y_{t-1} + \epsilon_t^{(j)}$
Modelo STAR	Howell Tong (1986)	$Y_t = \mu + \sum_i \phi_i Y_{t-i} + \{\pi_0 + \sum_i \pi_i Y_{t-i}\} F(z_t) + \epsilon$

Tabla I: Modelos No Lineales más Relevantes

$$LM = \tilde{q}' I^{-1} \tilde{q} = \tilde{\lambda}' \tilde{H}' I^{-1} \tilde{H} \tilde{\lambda}$$

donde

- $\tilde{q} = q \begin{pmatrix} \tilde{\theta} \end{pmatrix}$
- $\tilde{I} = I \begin{pmatrix} \tilde{\theta} \end{pmatrix}$
- $\tilde{H} = H \begin{pmatrix} \tilde{\theta} \end{pmatrix}$.

Cabe resaltar que:

- $L(\theta)$, es una función de verosimilitud logarítmica de un vector de parámetros θ .
- $q(\theta) = \frac{\partial L(\theta)}{\partial \theta}$, es la función de puntuación.
- $I(\theta) = -E \left[\frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'} \right]$, es la la matriz de información.
- $H(\theta) = \frac{\partial k(\theta)'}{\partial \theta}$ y $k(\theta)$ es el vector de restricciones.

Por otra parte, en cuanto a las pruebas de razón de verosimilitud (LR) se han utilizado para comparar dos modelos anidados. Para mayor información véase [15]. El estadístico de prueba está dado por:

$$LR = -2 \log_e \left(\frac{L_s(\hat{\theta})}{L_g(\hat{\theta})} \right)$$

donde

- $L_s(\hat{\theta})$ es la verosimilitud del modelo nulo.
- $L_g(\hat{\theta})$ es la verosimilitud del modelo alternativo.

La prueba de Wald [17] es una prueba estadística paramétrica que se utiliza para poner a prueba el verdadero valor del parámetro basado en la estimación de la muestra. En el caso univariado, el estadístico de Wald está dado por:

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var(\hat{\theta})}$$

Donde, la estimación por máxima verosimilitud $\hat{\theta}$ del parámetro de interés θ se compara con el valor propuesto bajo la hipótesis nula H_0 ; asumiendo que la diferencia entre ambos se distribuye aproximadamente normal. Lo que conlleva a que el cuadrado de la diferencia se compara con una distribución chi-cuadrado. En general, en la detección de series de tiempo no lineales, se observa que las pruebas de Wald son las menos estudiadas. De manera que las Pruebas *LR* requieren la estimación de los parámetros del modelo bajo las hipótesis nulas y alternas presentadas, mientras que las pruebas *LM* requieren la estimación sólo de la hipótesis nula. Por ende, en el caso de una prueba de hipótesis alterna no lineal muy

compleja, la prueba *LM* sería la más usual. En relación a los modelos de series de tiempo no lineales presentados en el cuadro I, es común utilizar la prueba *LM* cuando se está trabajando con los modelos: *BL*, *EAR*, *ARCH* y *STAR*. Mientras que la prueba *LR* es comúnmente utilizada al trabajar con el modelo *SETAR*.

Es prudente resaltar la tercera metodología, *multiplicadores de Lagrange*, puesto que en esta la hipótesis alterna se modifica, de tal modo que se chequea es linealidad Vs los modelos no lineales puntuales. Para identificar la elección del mejor modelo se necesitan utilizar los criterios de selección. En la siguiente subsección se presentarán dichos criterios junto con sus propiedades y variaciones.

II-C. Criterios de Selección de Series de Tiempo

Los criterios de selección son una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos (véase por ejemplo [4] y [15]). Entre éstos criterios de selección se destaca el “Criterio de Información de Akaike”, AIC, el cual se define matemáticamente como:

$$AIC = 2k - 2 \ln(L)$$

donde, k es el número de parámetros en el modelo estadístico y L es el valor máximo de la función de verosimilitud. El resto de criterios de selección son básicamente modificaciones del criterio *AIC*, algunos de estos son:

- *AIC_C*: es una corrección para el AIC en el caso de tener muestras finitas y se define como: $AIC_C = AIC + \frac{2k(k+1)}{n-k-1}$ donde n es el tamaño de la muestra.
- *BIC*: el criterio de información bayesiano se define de la forma: $-2 \ln \hat{L} + k \ln(n)$, donde, k es el número de parámetros libres a estimar, \hat{L} es el valor máximo de la función de verosimilitud.
- *DIC*: el criterio de información de desviación es una generalización jerárquica de los criterios *AIC* y *BIC*. Es útil en problemas selección de modelos Bayesianos donde la distribución de los modelos ha sido obtenida por medio de cadenas de Markov o el método de Monte Carlo y se define como $DIC = p_D + \bar{D}$ donde p_D es el número de parámetros efectivos y $p_D = \bar{D} - D(\hat{\theta})$, donde θ son los parámetros desconocidos.
- *FIC*: el criterio de información enfocada es un método para seleccionar el modelo más adecuado entre un conjunto de competidores para un determinado conjunto de

datos. A diferencia de la mayoría de las otras estrategias de selección de modelos, como el *AIC*, *BIC* o *DIC*, el *FIC* no intenta evaluar el ajuste global de los modelos candidatos, sino que centra la atención directamente en el parámetro de interés primario por medio de un análisis estadístico.

- *HQC* : el criterio de información de Hannan-Quinn es una alternativa a los criterios presentados anteriormente.

Estos criterios hacen parte de los métodos de identificación de modelos lineales y no lineales, y por ende, son un punto de alta importancia en dicho proceso y deben ser utilizados cuidadosamente.

II-D. Métodos de Identificación

En la literatura se destacan dos aproximaciones o metodologías, propuestas por Gooijer-Kumar (1992) y por Terasvirta-Tjstheim-Granger (1992), para hacer el modelado de series de tiempo no lineales (ver [4] y [9]). Ambas metodologías son esencialmente análogas y en general se plantea lo siguiente:

1. Identificar y estimar el modelo autorregresivo de medias móviles (ARMA) lineal. Aunque se recomienda emplear sólo modelos autorregresivos (AR) debido a que son más fáciles de interpretar y permiten una comparación directa con la mayoría de modelos AR no lineales.
2. Realizar el diagnóstico del modelo, es decir, examinar la gráfica de la serie, los correlogramas, residuales y hacer pruebas de diagnóstico como la de Ljung-Box.
3. Seleccionar alguna clase de modelo AR no lineal. Esta selección se hace de manera empírica, a no ser que existan pruebas teóricas suficientes para justificar la elección de un modelo conciso. Posteriormente realizar pruebas de linealidad tipo LM para determinar el orden del mejor modelo a utilizar como referencia.
4. Seleccionar un modelo AR no lineal diferente al previamente seleccionado que sea similar al mejor modelo obtenido en la etapa anterior. Usando nuevamente pruebas de linealidad como la LM por ejemplo. Se mide el desempeño de este nuevo modelo a comparación del mejor resultado obtenido previamente.
5. Si el nuevo presenta mejores resultados que el anterior se debe utilizar los criterios de selección para escoger el mejor modelo, y posteriormente hacer la estimación de los parámetros.
6. Finalmente se deben hacer pruebas de diagnóstico en el modelo no lineal elegido en la etapa anterior. En caso de no pasar dichas pruebas se debe repetir todo el proceso.

Cabe destacar que la metodología propuesta en éste trabajo se centra en la parte de especificación del modelo, es decir, todo el proceso anterior a la parte de la estimación. De tal modo que en el momento de hacer la estimación del modelo se esté totalmente seguro del rendimiento óptimo que éste tendrá.

En la siguiente sección se presenta la metodología propuesta, su justificación y su procesamiento paso a paso.

III. METODOLOGÍA PROPUESTA

Con el fin de brindar una alternativa de solución a algunas falencias detectadas en la identificación de series de tiempo,

se propone una metodología de o una guía paso a paso automatizada y amigable que permita al investigador hacer la identificación del modelo tipo *TAR* exitosamente. Es de resaltar que el proceso propuesto en este trabajo, se puede extender fácilmente a la identificación de otras familias de modelos no lineales.

La metodología propuesta fue desarrollada, implementada y examinada en el software estadístico *R*, el cual es un entorno de software libre para computación y gráficos estadísticos entre muchas otras herramientas que posee. Todo el proceso se limita a una función llamada *ESP_ST* la cual recibe como argumentos una base de datos objeto de estudio y de análisis, además de la frecuencia de los mismos. La metodología propuesta se presenta en la figura 1 y consta básicamente de los siguientes pasos.

- *Paso 1*: Seleccionar el mejor modelo autorregresivo de medias móviles integrado lineal (*ARIMA*) para los datos trabajados. Esto se hace considerando n modelos *ARIMA* distintos entre si y seleccionando el que mejor se adapte a los datos por medio del criterio *AIC*.
- *Paso 2*: Realizar el diagnóstico del mejor modelo *ARIMA* seleccionado anteriormente. Esto se hace esencialmente revisando los residuales del mismo y por medio de una prueba de *Box*. Se debe establecer la funcionalidad del mejor modelo lineal seleccionado.
- *Paso 3*: En caso de que no haya problemas de ningún tipo con el diagnóstico del modelo, se asume que la serie es lineal y restaría realizar la etapa del pronóstico del modelo y de control de estabilidad, dando por terminada la tarea de identificación.
- *Paso 4*: En caso de que haya problemas con el diagnóstico del modelo, se procede a realizar las pruebas de linealidad (generales y específicas).
 - Prueba de la red neuronal de Terasvirta.
 - Prueba de la red neuronal de White.
 - Prueba de no linealidad de un grado de Keenan.
 - Prueba de McLeod-Li.
 - Prueba de no linealidad de Tsay.
 - Prueba de máxima verosimilitud de no linealidad con umbrales.
- *Paso 5*: La metodología trabaja conjuntamente con las pruebas de hipótesis descritas anteriormente, y en caso de que no rechace ninguna de ellas se debe considerar transformaciones de la serie para pasar de nuevo al *paso 1*.
- *Paso 6*: En caso de que rechace alguna de las pruebas de hipótesis descritas anteriormente y como se consideró la prueba específica de modelos *AR* contra modelos *TAR*, se pasa a evaluar seis modelos: cinco no lineales y el mejor modelo *ARIMA* encontrado en el *paso 1*. Los modelos no lineales considerados en este paso son:
 - Modelo Autorregresivo no-lineal con Red Neuronal (nnetTs).
 - Modelo SETAR.
 - Modelo LSTAR (Logistic Smooth Transition Autoregressive Model).
 - Modelo STAR.

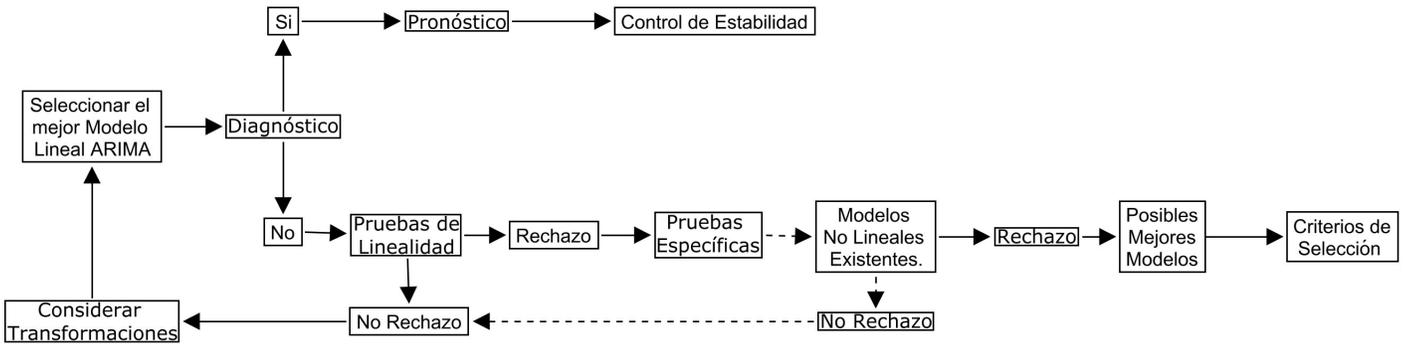


Figura 1: Metodología Propuesta

- Modelo AAR (Additive Nonlinear Autoregressive Model).
- *Paso 7*: Criterios de selección y cálculos de error. Después de tener los modelos identificados y aplicados a los datos, se procede a calcular tanto los criterios de selección AIC y BIC como las medidas de error MSE (error cuadrático medio) y $MAPE$ (error medio absoluto porcentual) de los modelos ajustados.
- *Paso 8*: La metodología arroja una lista con la siguiente información:
 - El mejor modelo en cada uno de los cuatro ítems descritos en el *paso 8*: AIC , BIC , MSE , $MAPE$.
 - El tiempo de ejecución de la metodología.
 - La gráfica de la serie de tiempo.

Cabe resaltar que lo que realmente interesa a esta metodología es la identificación del modelo más no la especificación del mismo. Dicha tarea se podría proponer como un trabajo futuro.

IV. EXPERIMENTACIÓN Y RESULTADOS

En esta sección se presenta la parte de aplicación y experimentación de la metodología propuesta con los resultados obtenidos. La experimentación se realiza bajo dos enfoques: datos simulados y datos reales. Así mismo, se examina la capacidad de identificación de la metodología al considerar una serie no lineal generada por un modelo que no esté incluido dentro de familia de modelos de la metodología.

IV-A. Datos Simulados

En un primer experimento se generaron cinco series de tiempo, en el software estadístico R , por medio del modelo no lineal $SETAR$ definido en la ecuación (2).

$$Y_t = \gamma_0^{(j)} + \sum_t \gamma_t^{(j)} Y_{t-1} + \epsilon_t^{(j)} \quad (2)$$

Los parámetros fueron tomados aleatoriamente de una distribución uniforme en intervalos dados. Cada serie cuenta con 200 observaciones, dos rezagos en cada régimen y un umbral de 2. Los valores iniciales fueron los mismos en las cinco simulaciones. En la tabla II se presentan los rangos en que dichos parámetros fueron generados y algunas de sus respectivas gráficas se pueden observar en las figuras 2 y 3.

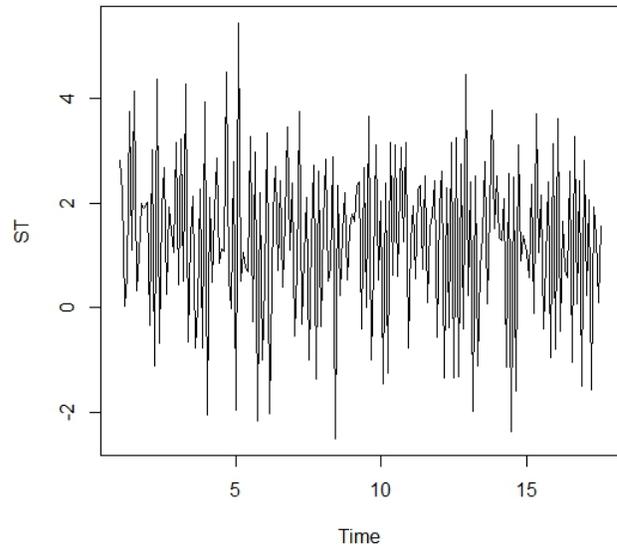


Figura 2: Serie Simulada #2

Parámetro	Rango
γ_0	[2, 3]
γ_1	[-1, 0]
γ_2	[-1, 0]
γ_3	[-2, -1]
γ_4	[0, 1]
γ_5	[0, 1]

Tabla II: Rango de los parámetros de las series $SETAR$

En la tabla III se presentan los resultados arrojados por la metodología propuesta para cada una de las series simuladas tanto para los criterios AIC y BIC , como para los medidores de error MSE y $MAPE$, recordar que la metodología se encarga de seleccionar, después de una serie de pasos y pruebas, el mejor modelo no lineal para trabajar los datos entregados y muestra los resultados obtenidos en esos cuatro criterios. Al final, consideramos que el resultado válido es el entregado por el AIC aunque los arrojados por los demás criterios también se deben considerar.

En la última columna de dicha tabla se muestra el tiempo en segundos que tardó la metodología en encontrar el resultado.

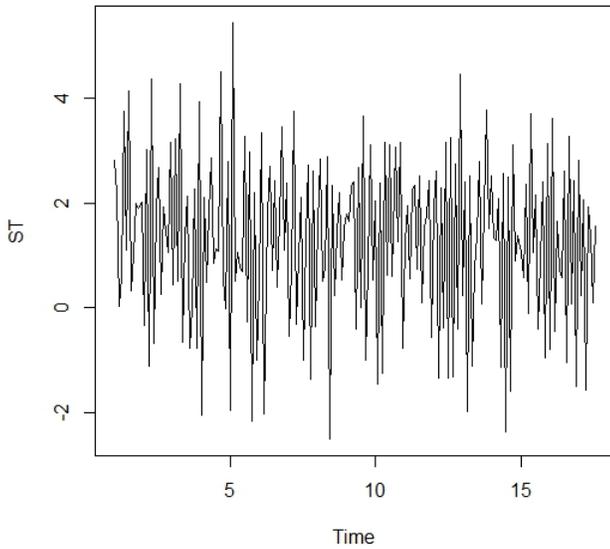


Figura 3: Serie Simulada #3

# de Serie	<i>AIC</i>	<i>BIC</i>	<i>MSE</i>	<i>MAPE</i>	<i>t</i> (s)
1	<i>SETAR</i>	<i>SETAR</i>	<i>LSTAR</i>	<i>SETAR</i>	4.53
2	<i>SETAR</i>	<i>SETAR</i>	<i>NNET</i>	<i>AAR</i>	4.48
3	<i>SETAR</i>	<i>SETAR</i>	<i>AAR</i>	<i>AAR</i>	4.67
4	<i>SETAR</i>	<i>SETAR</i>	<i>SETAR</i>	<i>NNET</i>	4.48
5	<i>SETAR</i>	<i>SETAR</i>	<i>NNET</i>	<i>AAR</i>	2.53

Tabla III: Resultados obtenidos con la metodología propuesta

Los resultados obtenidos por medio de la metodología en este primer experimento son bastante satisfactorios. Se puede observar que la metodología en casi todos los casos seleccionó como mejor modelo para trabajar las series el modelo no lineal *SETAR* (por medio de los criterios *AIC* y *BIC*), a pesar de que en algunos casos fueron otros de los modelos trabajados quienes dieron un menor error la conclusión siempre fue que el modelo *SETAR* era el encargado de trabajar las series propuestas lo cual era lo esperado que pasara puesto que ya sabíamos que dichas series fueron generadas bajo este modelo. En segundo lugar se puede notar la eficiencia de la metodología puesto que tardó en dar una solución para cada serie un tiempo promedio de 4.138 segundos.

Se comprobó entonces el buen funcionamiento de la metodología propuesta con datos simulados. Por otra parte, se consideró un segundo experimento en el cual se trabajan datos generados por medio de un modelo que no pertenezca a la familia de modelos propuestos en la metodología. De este modo se puede analizar el comportamiento de la metodología al analizar los resultados obtenidos.

El modelo trabajado en este caso es conocido como “*VAR* multivariado con Umbral” y se le denomina *TVAR*. El *VAR*, por sus siglas en inglés “*Value at Risk*”, es usualmente usado como el medidor de riesgo por excelencia, pero en este caso, se propone usarlo como una serie no lineal multivariada definida con uno o varios umbrales. Para el experimento se generaron

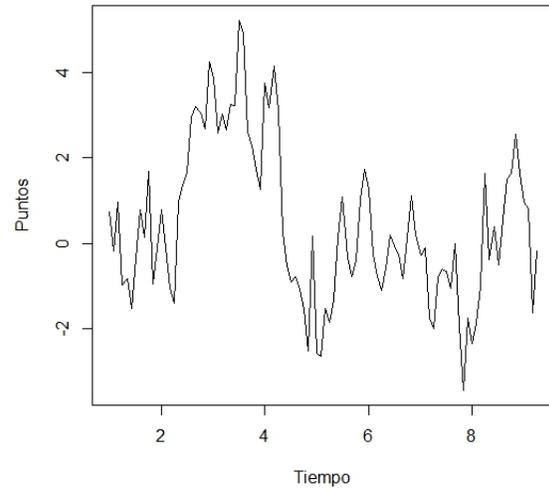


Figura 4: TVAR #1

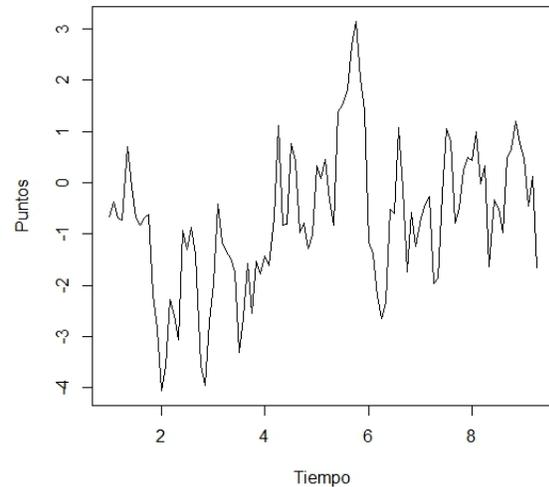


Figura 5: TVAR #2

dos series por medio del modelo *TVAR* en el software *R* y se quiere observar y analizar el desempeño de la metodología propuesta. Dichas series cuentan con 100 observaciones, parámetros aleatorios, ningún umbral y la misma matriz de coeficientes a ser simulados. Las gráficas de las series se pueden ver en las figuras 4 y 5, además, los resultados obtenidos con la metodología propuesta se muestran en la tabla IV.

Serie	<i>AIC</i>	<i>BIC</i>	<i>MSE</i>	<i>MAPE</i>	<i>t</i> (seg)
TVAR #1	<i>SETAR</i>	<i>STAR</i>	<i>NNET</i>	<i>NNET</i>	1.44
TVAR #2	<i>STAR</i>	<i>STAR</i>	<i>NNET</i>	<i>NNET</i>	1.4

Tabla IV: Resultados TVAR obtenidos con la metodología

Analizando los resultados, se puede observar que al no

considerar los modelos *TVAR* en la metodología, el resultado para ambas series arroja (en un tiempo de ejecución demasiado corto) que el modelo autorregresivo no lineal con red neuronal (*NNET*) es el modelo que menor error tiene en ambas medidas, sin embargo, los criterios de selección determinan que el modelo *SETAR* en la serie 1 y el modelo *STAR* en la serie 2, son los modelos pertinentes para hacer la identificación. En el caso de la segunda serie, ambos criterios de selección dan por resultado el modelo *STAR* lo cual hace pensar que es posible la existencia de una relación entre el modelo *TVAR* utilizado y definido en dicho caso y el modelo *STAR*.

IV-B. Datos Reales

Se aplicó la metodología propuesta a un conjunto de series de tiempo de la vida real que representan series financieras, económicas y actuariales. Las series aquí trabajadas han sido catalogadas por varios autores como no lineales debido a la alta volatilidad que presentan y a su comportamiento atípico y difícil de modelar o predecir. Sin embargo, fue considerada aparte una serie lineal para comprobar el buen funcionamiento de la metodología.

A continuación se presentan las cinco series que fueron tomadas para trabajar como ejemplo y una breve descripción de ellas:

1. Índice COLCAP: Es el índice bursátil más representativo del mercado de renta variable Colombiano. Dicho índice está compuesto por las 20 acciones más importantes del país y se encarga de medir las captaciones diarias del mercado. Tomado entonces como un índice referente que muestra el comportamiento del mercado diariamente.
2. Índice COLTES: Es el índice bursátil más representativo del mercado de renta fija Colombiano. Dicho índice está compuesto por los bonos gubernamentales (mejor conocidos como tesoros o TES) más importantes del país y por importancia se entiende de mayor volumen, liquidez y rentabilidad. Se encarga de medir las captaciones diarias del mercado y es tomado entonces como un índice referente que muestra el comportamiento del mercado diariamente.
3. Esperanza de vida en Colombia: La esperanza de vida es básicamente la media de la cantidad de años que vive una determinada población absoluta o total en un cierto periodo, en este caso, la población colombiana. De ante mano se conoce que esta serie es lineal.
4. Índice de precios del consumidor (IPC): es un índice en el que se valoran los precios de un conjunto de productos que una cantidad de consumidores adquiere de manera regular. El índice mide la variación del precio con respecto al de una muestra anterior.
5. Índice Standard & Poor's 500: es uno de los índices bursátiles más importantes de Estados Unidos. Se le considera el índice más representativo de la situación real del mercado.

En la tabla V se presenta el número de observaciones que se tiene para cada serie y algunas gráficas se pueden ver en las figuras 6, 7, 8, 9 y 10.

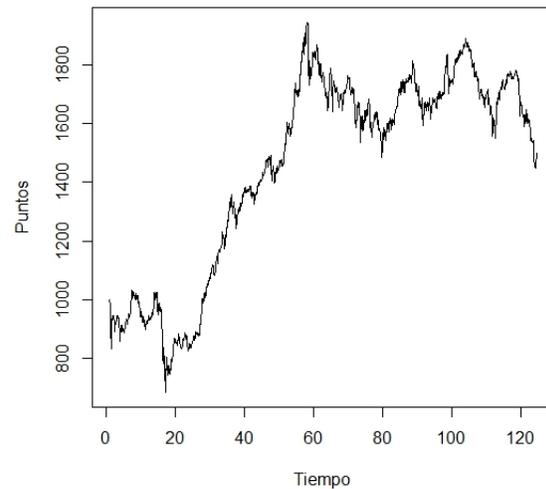


Figura 6: COLCAP

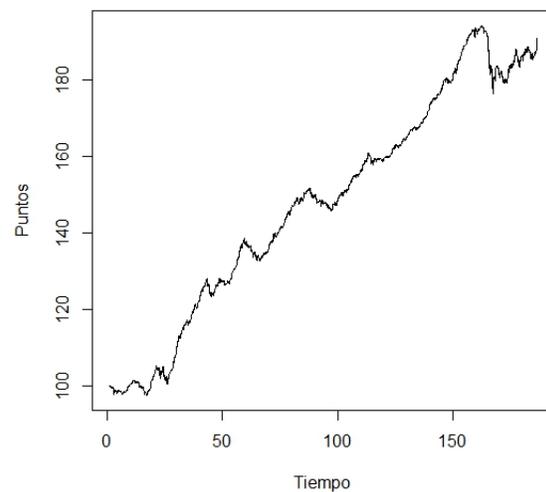


Figura 7: COLTES

Nombre de la Series	# de Observaciones
COLCAP	1,485
COLTES	2,228
Esperanza_COL	51
IPC	715
S&P 500	16,156

Tabla V: Series de Tiempo reales trabajadas

Con respecto a los resultados obtenidos, en la tabla VI se presentan los resultados arrojados por la metodología en cada una de las series simuladas para los criterios *AIC*, *BIC* y los medidores de error *MSE* y *MAPE*, recordar que la metodología se encarga de seleccionar, después de una serie de pasos y pruebas, el mejor modelo no lineal para trabajar los datos entregados y muestra los resultados obtenidos en esos cuatro criterios.

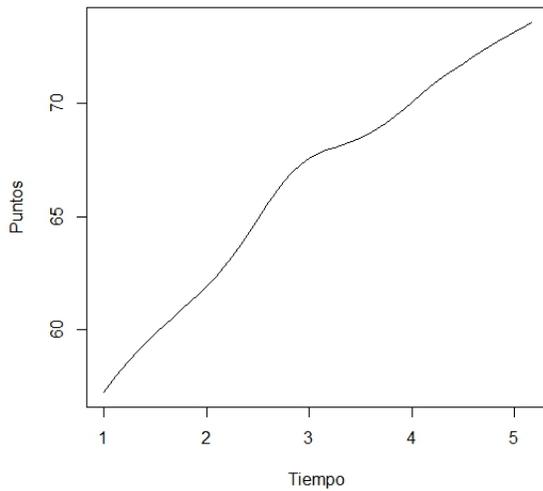


Figura 8: Esperanza de Vida

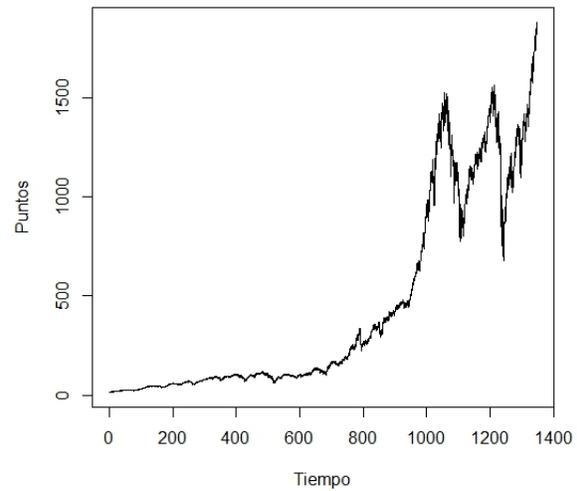


Figura 10: S&P 500

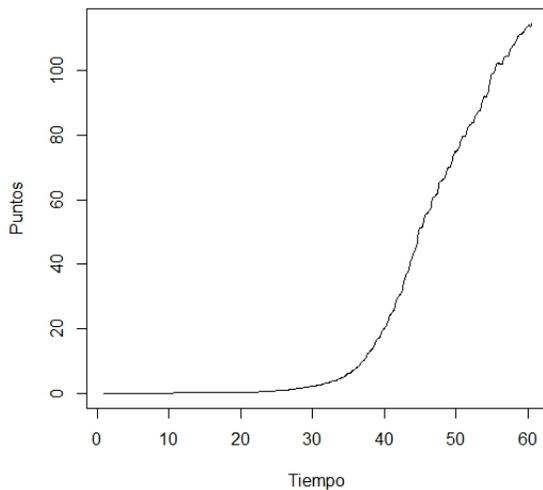


Figura 9: IPC

Las primeras columnas muestran para cada una de las 5 series trabajadas cual fue el modelo no lineal seleccionado en cada uno de los 4 criterios, mientras que en la última columna de dicha tabla se muestra el tiempo que tardó la metodología en encontrar el resultado.

Serie	<i>AIC</i>	<i>BIC</i>	<i>MSE</i>	<i>MAPE</i>	<i>t</i> (min)
COLCAP	<i>SETAR</i>	<i>SETAR</i>	<i>SETAR</i>	<i>SETAR</i>	0.44
COLTES	<i>SETAR</i>	<i>SETAR</i>	<i>SETAR</i>	<i>SETAR</i>	0.64
Esp_COL	"Realice el Pronóstico"				0.06
IPC	<i>STAR</i>	<i>STAR</i>	<i>STAR</i>	<i>AAR</i>	0.94
S&P 500	<i>SETAR</i>	<i>SETAR</i>	<i>SETAR</i>	<i>AAR</i>	5.82

Tabla VI: Resultados obtenidos con la metodología propuesta

Los resultados obtenidos por medio de la metodología en este caso son bastante satisfactorios puesto que en cada

una de las cinco series la metodología logró encontrar un resultado que fue congruente con tres de los cuatros ítems de la respuesta. Para el índice COLCAP los dos criterios de selección (*AIC* y *BIC*) junto con los dos medidores de error (*MSE* y *MAPE*) dieron como resultado el modelo *SETAR* como mejor modelo para llevar a cabo la identificación del índice. Cabe resaltar que este es el resultado ideal puesto que es el escenario más confiable y es aquel en que los cuatro criterios están de acuerdo con el mismo modelo. Esto ocurre similarmente para el índice COLTES, se puede observar que el modelo *SETAR* es seleccionado en los cuatro criterios. Esto no ocurre para el IPC pues como se puede observar en la tabla el *MAPE* arroja un resultado diferente al de los demás. Los criterios de selección y la primera medida de error establecen el modelo *STAR* como el modelo óptimo para llevar a cabo la identificación, esto ocurre igualmente con el S&P 500 puesto que los tres primeros criterios definen el modelo *SETAR* mientras que el *MAPE* nuevamente resalta un modelo diferente. En estos casos es confiable trabajar con el modelo arrojado por los tres primeros criterios puesto que su procedimiento es confiable y al estar relacionados entre sí, el resultado es consistente.

Por otro lado, se puede observar que el tiempo de ejecución es proporcional al número de observaciones puesto que a mayor número de observaciones mayor será el tiempo que tarda en ejecutarse la metodología. Por ejemplo, en el índice COLCAP que se cuenta con 1,485 datos el tiempo de ejecución fue de 0.44 minutos mientras que en el S&P 500 que cuenta con 16,156 observaciones el tiempo de ejecución fue de 5,82 minutos, unas 13 veces mayor que en el primer caso. Se puede ver que en la serie de Esperanza de Vida en Colombia, la cual sabemos es una serie lineal, obtuvo un resultado totalmente acertado pues fue el de pasar directamente al pronóstico, es decir, la serie resultó ser lineal y el mejor modelo *ARIMA* estimado fue excelente para sacar dicha conclusión (ver figuras 1 y 8).

En general la metodología fue congruente con sus resultados y logro establecer después de un proceso analítico con varias pruebas y escenarios, de una forma rápida y amigable, cual es el mejor modelo no lineal para tratar dichas series reales logrando de este modo nuestro objetivo propuesto

V. CONCLUSIONES

Se propuso una metodología de identificación de modelos de series de tiempo no lineales pertenecientes a la familia *TAR*, que captura el comportamiento real de las series temporales y permite seleccionar el modelo más apropiado según el caso. Dicha metodología es una guía paso a paso amigable que permite identificar no sólo si se debe emplear un modelo no lineal sino también seleccionar el mejor modelo de la familia *TAR* para los datos objeto de estudio.

Se comprobó la viabilidad y la eficiencia de la metodología propuesta empleando datos simulados y datos reales. Los resultados encontrados satisfacen las expectativas en ambos casos experimentales, comprobando entonces el buen funcionamiento de la metodología en el caso que se conocía el resultado (datos simulados) y en el caso de datos experimentales al ser capaz la metodología de tratar diversas series financieras o económicas reales.

Como trabajo futuro se plantea la posibilidad de que la metodología considere otras familias de modelos no lineales para que sea más completa y abarque mayores posibilidades en sus resultados, además, se plantea la posibilidad de no solo realizar la identificación del mejor modelo sino también la estimación del mismo.

REFERENCIAS

- [1] E. Parzen, "Some recent advances in time series modeling," *Automatic Control, IEEE Transactions on*, vol. 6, pp. 723–730, 1974.
- [2] L. C. Alwana and H. V. Roberts, "Time-series modeling for statistical process control," *Journal of Business & Economic Statistics*, vol. 3, pp. 1988.
- [3] G. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.
- [4] J. G. D. Gooijer and K. Kumar, "Some recent developments in non-linear time series model, testing and forecasting," *International Journal of Forecasting*, vol. 8, pp. 135–156, 1992.
- [5] J. G. D. Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, pp. 443–473, 2006.
- [6] N. Teruia and H. K. van Dijk, "Combined forecasts from linear and nonlinear time series models," *International Journal of Forecasting*, vol. 18, pp. 421–438, 2002.
- [7] H. Feng and J. Liu, "A setar model for canadian gdp: Non-linearities and forecast comparisons," *Applied Economics*, vol. 18, pp. 1957–1964, 2003.
- [8] H. Tong, "Nonlinear dynamical approach and self-exciting threshold model in forecasting daily stream-flow," *Fresenius Environmental Bulletin*, vol. 10, pp. 2836–2847, 2013.
- [9] T. Terasvirta, D. Tjstheim, and C. W. J. Granger, "Aspects of modelling nonlinear time series," *University Of Goteborg Department Of Statistics*, 1992.
- [10] R. Luukkonen and T. Terasvirta, "Testing linearity of economic time series against cyclical asymmetry," *Annales D'Economie Et De Statistique*, 1991.
- [11] J. D. Hamilton and B. E. Hansen, *Time Series Analysis*. Princeton University Press, 1994.
- [12] D. Poskitt and A. Tremayne, "The selection and use of linear and bilinear time series models," *International Journal of Forecasting*, vol. 2, pp. 101–114, 1986.
- [13] G. M. Kuersteiner, "Kernel-weighted gmm estimators for linear time series models," *Journal of Econometrics*, vol. 170, pp. 399–421, 2012.
- [14] W. A. Brock and P. J. F. de Lima, "Nonlinear time series, complexity theory, and finance," *Santa Fe Institute*, 1995.
- [15] M. P. Clements, P. H. Franses, and N. R. Swanson, "Forecasting economic and financial time-series with non-linear models," *International Journal of Forecasting*, vol. 20, pp. 169–183, 2004.
- [16] M. Arellano, "Lagrange multiplier test," 2002.
- [17] R. Engle, "Chapter 13 wald, likelihood ratio, and lagrange multiplier tests in econometrics," *Handbook of Econometrics*, vol. 2, 1984.
- [18] B. Abraham and J. Ledolter, "Statistical methods for forecasting," *John Wiley & Sons*, pp. 3–4, 2005.
- [19] H.-Z. An, L.-X. Zhu, and R.-Z. Li, "A mixed-type test for linearity in time series," *Journal of Statistical Planning and Inference*, vol. 88, pp. 339–353, 2000.
- [20] A. Berg, E. Paparoditis, and D. N. Politis, "A bootstrap test for time series linearity," *Journal of Statistical Planning and Inference*, vol. 140, pp. 3841–3857, 2010.
- [21] G. Box and G. M. Jenkins, "Time-series analysis: forecasting and control," *San Francisco, CA: Holden Day*, 1970.
- [22] J. L. Castle and D. F. Hendry, "A low-dimension portmanteau test for non-linearity," *Journal of Econometrics*, vol. 158, pp. 231–245, 2010.
- [23] J. Deeks and D. Altman, "Statistics notes - diagnostic tests 4: Likelihood ratios," *British Medical Journal*, vol. 329, 2004.
- [24] S. Giannerini, "The quest for nonlinearity in time series," *Time Series Analysis: Methods and Applications*, vol. 30, pp. 43–63, 2012.
- [25] H. Kaufmann, R. Kruse, and P. Sibbertsen, "On tests for linearity against star models with deterministic trends," *Economics Letters*, vol. 117, pp. 268–271, 2012.
- [26] E. Marrocu, "An investigation of the effects of data transformation on nonlinearity," *Empirical Economics*, vol. 31, pp. 801–820, 2006.
- [27] D. Massacci, "A simple test for linearity against exponential smooth transition models with endogenous variables," *Economics Letters*, vol. 117, pp. 851–856, 2012.
- [28] S. Resnick and E. V. D. Berg, "A test for nonlinearity of time series with infinite variance," *Kluwer Academic Publishers*, pp. 145–172, 2000.
- [29] E. Slutsky, "The summation of random causes as the source of cyclic processes," *Econometrica*, vol. 5, pp. 105–146, 1937.
- [30] T. Suzuki, T. Ikeguchi, and M. Suzuki, "Evaluation of nonlinearity and validity of nonlinear modeling for complex time series," *Physical Review*, vol. 76, 2007.
- [31] P. Teles and W. W. Wei, "The effects of temporal aggregation on tests of linearity of a time series," *Computational Statistics & Data Analysis*, vol. 34, pp. 91–103, 2000.
- [32] A. M. Walker, "On the periodicity in series of related terms," *Proceedings of the Royal Society of London A*, vol. 131, pp. 518–532, 1931.
- [33] Y. Yuan and Z. Q. Huang, "V -test for determining the degree of nonlinearity of time series," *College of Engineering Shantou University*, pp. 546–553, 2012.
- [34] G. U. Yule, "Why do we sometimes get nonsense-correlations between time series? a study in sampling and the nature of time series," *Journal of the Royal Statistical Society*, vol. 89, pp. 1–64, 1926.