UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

LA SOBREPARAMETRIZACIÓN EN EL ARIMA: UNA APLICACIÓN A DATOS COSTARRICENCES

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar por el grado y título de Maestría Académica en Estadística

CÉSAR ANDRÉS GAMBOA SANABRIA B12672

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

DEDICATORIA

Pendiente

AGRADECIMIENTOS

También pendiente

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Estadística"

Ph.D. Álvaro Morales Ramírez

Decano Sistema de Estudios de Posgrado

MSc. Óscar Centeno Mora

Director de Tesis

Ph.D. Gilbert Brenes Camacho

Lector

Ph.D. ShuWei Chou.

Lector

MSc. Johnny Madrigal Pana

Director Programa de Posgrado en Estadística

César Andrés Gamboa Sanabria

Candidato

Índice

Índice de cuadros

Índice de figuras

RESUMEN

ABSTRACT

1 INTRODUCCIÓN

1.1 Antecedentes

Estimar los valores futuros en un determinado contexto ha producido un aumento en el análisis de los datos referidos en el tiempo, conocido también como series cronológicas. Este tipo de datos se encuentra en diferentes áreas, tanto en investigación académica como en el análisis de datos para la toma de decisiones. En el campo financiero es común hablar de la devaluación del colón con respecto al dólar, cantidad de exportaciones mensuales de un determinado producto o las ventas, entre otros (Hernández, 2011a). Las series cronológicas son particularmente importantes en la investigación de mercados o en las proyecciones demográficas; de manera conjunta apoyan la toma de decisiones para la aprobación presupuestaria en distintas áreas.

En la actualidad, la información temporal es muy relevante: El Banco Mundial[1] cuenta en su sitio web con datos para el análisis de series cronológicas de indicadores de desarrollo, capacidad estadística, indicadores educativos, estadísticas de género, nutrición y población. Kaggle[2], uno de los sitios más populares relacionados con el análisis de información, ofrece una gran cantidad de datos temporales para realizar competencias relacionadas con las series temporales y determinar los modelos ganadores para una determinada temática[3].

Asimismo, los pronósticos (estimación futura de una partícula en una serie temporal) son utilizados por instituciones públicas o del sector privado, centros nacionales o regionales de investigación y organizaciones no gubernamentales dedicadas al desarrollo social. Si las entidades previamente mencionadas cuentan con proyecciones de calidad, la puesta en marcha de sus respectivos planes tendrá un impacto más efectivo.

Los métodos existentes para llevar a cabo un análisis de series cronológicas son diversos, y responden al propio contexto y tipo de datos. Obtener buenos pronósticos o explicar el comportamiento de un fenómeno en el tiempo, siempre será un tema recurrente de investigación. Generar una adecuada estimación es fundamental para obtener un pronóstico de confianza, además resulta importante mencionar que los modelos ARIMA tienen como objetivo explicar las relaciones pasadas de la serie cronológica, para de esta manera conocer el posible comportamiento futuro de la misma (Hyndman & Athanasopoulos, 2018).

Al trabajar con la metodología de Box-Jenkins, uno de las etapas a concretar es identificar los parámetros de estimación que gobiernan la serie temporal. Para indagar los términos en el proceso de investigación se ha utilizado la identificación de parámetros mediante autocorrelogramas par-

[1] https://databank.worldbank.org/home.aspx

[2] Se trata de una subsidiaria de la compañía Google que sirve de centro de reunión para todos aquellos interesados en la ciencia de datos.

[3] Muchas de ellas incluyen recompensas económicas que van desde los $500 hasta los $100,000 para aquellos que logren obtener los mejor pronósticos.

ciales y totales. Sin embargo, los autocorrelogramas formados no analizan de forma exhaustiva y óptima los posibles coeficientes que podrían contemplarse la ecuación de Wold. Según su definición matemática, esta posee infinitos coeficientes, por tanto, se debe buscar una alternativa distinta, que opte por aproximar de una mejor manera la identificación de los parámetros estimados, cubriendo un mayor número de posibilidades. Esto se podría obtener mediante un método analítico de sobreparametrización.

1.2 La problemática

La dificultad visual a la hora de identificar un modelo ARIMA radica en que los autocorrelogramas solo aportan una aproximación al proceso que gobierna la serie. De forma complementaria, es común caer en el problema de la subjetividad, pues a pesar de que alguien proponga un patrón que gobierne la serie, otro analista podría tener una interpretación visual diferente del mismo proceso, proponiendo así distintas identificaciones para un mismo proceso. Además, se posee el inconveniente de que algunos métodos de identificación automática del proceso que gobierna la serie subestiman el número de parámetros que se debería de contemplar.

Alternativas como la función `auto.arima()`, que ofrece el paquete `forecast` del lenguaje de programación R[4] (Hyndman & Khandakar, 2008), permite estimar un modelo ARIMA basado en pruebas de raíz unitaria y minimización del AICc (Burnham & Anderson, 2007). Así se obtiene un modelo temporal definiendo las diferenciaciones requeridas en la parte estacional d mediante las pruebas KPSS (Xiao, 2001) o ADF (Fuller, 1995), y la no estacional D utilizando las pruebas OCSB (Osborn, Chui, Smith, & Birchenhall, 2009) o la Canova-Hansen (Canova & Hansen, 1995), seleccionado el orden óptimo para los términos $ARIMA(p,d,q)(P,D,Q)_s$ para una serie cronológica determinada.

Sin embargo, estas pruebas suelen ignorar diversos términos que bien podrían ofrecer mejores pronósticos; no someten a prueba las posibles especificaciones de un modelo en un rango determinado, sino que realizan aproximaciones analíticas para definir el proceso que gobierna la serie cronológica, dejando así un vacío en el cual se corre el riesgo de no seleccionar un modelo que ofrezca mejores pronósticos. Poner a prueba un mayor número de posibilidades para la especificación de los modelos tiene la ventaja descartar ciertos modelos, y mantener otros con un criterio más científico y una evidencia numérica que despalde esa decisión.

1.3 Objetivos del estudio

El objetivo general de la presente investigación es proponer un algoritmo alternativo más exhaustivo para la selección de modelos ARIMA mediante la sobreparametrización de los términos de la ecuación del ARIMA.

[4]Descarga gratuita en https://cran.r-project.org/

Para lograr esto, se pretende:

1. Generar los escenarios de estimación de los distintos modelos ARIMA mediante permutaciones de los términos (p, d, q) y (P, D, Q) para la estimación de los posibles procesos que gobiernan una determinada serie temporal.

2. Aplicar diversos métodos de validación en la estimación de procesos que gobiernan la serie cronológica.

3. Contrastar la precisión de la estimación así como la generación de pronósticos con otros métodos similares, aplicados en datos costarricenses.

4. Integrar el desarrollo de la metodología de análisis de series temporales en una librería del lenguaje estadístico R.

1.4 Justificación del estudio

El accionar de políticas gubernamentales, así como de otro tipo de sectores, se apoyan cada vez más en un acertado análisis de la información temporal. En demografía, uno de los principales temas de investigación son las proyecciones de población; durante una emergencia, conocer la posible cantidad de población que habita una zona es clave para la rápida reacción de las autoridades en el envío de ayuda o en la ejecución de planes de evacuación. Asimismo, los análisis actuariales se ven beneficiados al mejorar sus métodos de pronóstico. Una de sus principales áreas de estudio es la mortalidad, ya que representa un insumo de vital importancia para la planificación y sostenibilidad de los sistemas de pensiones, servicios de salud tanto pública como privada, seguros de vida y asuntos hipotecarios (Rosero-Bixby, 2018).

La estimación de series de tiempo es una labor común es distintos campos de investigación: el objetivo es poder pronosticar de forma correcta lo que sucederá dentro de los próximos periodos. Métodos actuales como el `auto.arima()` solamente realizan aproximaciones analíticas no óptimas, por lo que suelen omitir procesos que describirían de una mejor manera el comportamiento futuro de una serie cronológica.

Estimar modelos ARIMA considerando diversas permutaciones en sus estimadores, permite mitigar las falencias de otras aproximaciones analíticas que no analizan de forma exhaustiva todos los posibles parámetros a estimar, o escenarios de selección de la mejor serie que gobierne el proceso de interés. El desarrollo y evaluación del método propuesto, la sobreparametrización, mostrará el potencial de esta metodología en la calidad de los pronósticos. El principal aporte de este estudio es brindar evidencia sobre cómo la sobreparametrización puede contribuir a definir la especificación de un modelo ARIMA que genere pronósticos más precisos.

1.5 Organización del estudio

El presente trabajo de investigación consta de cinco capítulos. El primer ofreció una contextualización del uso de las series de tiempo, así como la importancia de poder contar con pronósticos de calidad. Se presentó el objetivo del estudio, así como una breve descripción de la metodología empleada en la aplicación de series temporales, y cómo se planea modificar el método de estimación en los modelos ARIMA. Se concluye esta sección con hechos que justifican la importancia de esta investigación.

El siguiente capítulo consiste en el marco teórico, abarcando aspectos fundamentales de la ecuación de Wold, la metodología Box-Jenkins, la selección de los procesos que gobiernan la serie, la descripción del proceso iterativo, el análisis combinatorio que aborda los escenarios de estimación, entre otros.

El tercer capítulo describe la metodología relacionada al estudio. Se inicia con una descripción global de los conceptos más fundamentales del análisis de series cronológicas, pasando por los componentes fundamentales de las mismas. Se discuten también los supuestos clásicos del análisis de series cronológicas, los distintos tipos de modelos, el análisis de intervención, los métodos de validación y las medidas de rendimiento; aspectos cruciales para obtener un modelo ARIMA vía sobreparametrización. La sección metodológica culmina con la descripción del proceso de simulación que se utilizará, así como la discusión del método propuesto.

El capítulo cuatro consiste en la presentación de los resultados, tanto en los datos simulados como en la aplicación a datos costarricenses y se contrastarán contra los obtenidos por otros métodos como el de la función `auto.arima()`, entre otros.

El último capítulo busca discutir los principales resultados, así como señalar las conclusiones más importantes y ofrecer algunas recomendaciones que orienten futuros estudios relacionados.

2 MARCO TEÓRICO

2.1 Introducción

Los modelos de series cronológicas han sido un importante tema de investigación durante décadas. Su objetivo principal consiste en obtener simplificaciones de la realidad mediante el ajuste de diversos modelos, los cuales se ajustan a datos recolectados a lo largo del tiempo de forma regular. Estos modelos son luego utilizados para generar pronósticos sobre el comportamiento futuro del fenómeno de interés.

Sin embargo, encontrar un modelo que presente un buen comportamiento con respecto a los datos no es tarea fácil, pues deben considerarse diversos aspectos teóricos para obtener un modelo adecuado que logre generar pronósticos realistas y pertinentes para la toma de decisiones.

2.1.1 Definición de una serie cronológica

Una serie temporal se define como una secuencia de datos observados, cuyas mediciones ocurren de manera sucesiva durante un periodo de tiempo. Los registros de estos datos pueden referirse a una única variable en cuyo caso de dice que es una serie univariada; o bien, pueden registrarse distintas variables para el mismo periodo de tiempo, conocida como serie temporal multivariada. Según Hipel & McLeod (1994), cada observación puede ser continua o discreta, como la temperatura de una ciudad durante el día o las variaciones diarias del precio de un activo financiero, respectivamente; las observaciones continuas, además, pueden ser convertidas a su vez en observaciones discretas.

2.2 Componentes de una serie cronológica

De acuerdo con Hernández (2011a), las series cronológicas poseen cuatro componentes principales: Tendencia, Ciclos, Estacionalidad e Irregularidad. Considerando estos cuatro elementos, las series cronológicas pueden ser *aditivas*, como se muestra en la ecuación 1, en cuyo caso se asume que los cuatro componentes son independientes entre sí; o *multiplicativa*, donde, por el contrario, los cuatro componentes son independientes, como muestra la ecuación 2.

$$Y(t) = T(t) + S(t) + C(t) + I(t) \tag{1}$$

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t) \tag{2}$$

Donde Y es la serie cronológica, T es la tendencia, S es la parte estacional, C el componente cíclico, I la parte irregular o aleatoria, y t es el momento en el tiempo. Cada una de sus partes se definen

a continuación.

2.2.1 La tendencia

La tendencia general de una serie cronológica se refiere al crecimiento, decrecimiento o lateralización de sus movimientos a lo largo del periodo de estudio. Una tendencia bastante marcada es la del comportamiento poblacional, que con el tiempo su crecimiento suele comportarse de una forma muy similar a una exponencial.

2.2.2 Componentes estacionales

Los cambios estacionales que se presentan en una serie de tiempo se relaciona con las fluctuaciones naturales del fenómeno dentro de una temporada de observaciones. Ejemplos comunes de esto son las condiciones climáticas, consumo de alimentos en fechas festivas, etc.

2.2.3 Componente cíclico

Los periodo cíclicos, por su parte, se refieren a los cambios que se dan en una serie cronológica en el mediano plazo, que son causados por determinados eventos que suelen repetirse. Estos ciclos suelen tener una duración determinada, como es el caso de los índice bursátil S&P 500. Este indicador resume el estado de las 500 empresas más importantes de Estados Unidos, y sus ciclos suelen presentar un auge, seguido por un descenso que, posteriormente, se vuelve una depresión, y que finalmente se convierte en una recuperación a su estado inicial.

2.2.4 Componente irregular

Finalmente, la irregularidad de una serie cronológica se refiere a las fluctuaciones propias de un fenómeno que no pueden ser predichas. Estos cambios no se dan de manera regular, es decir, no siguen un patrón determinado.

2.3 Supuestos en el análisis de series cronológicas

El análisis de series temporales, según Hipel & McLeod (1994), representa un método para comprender la naturaleza de la serie en cuestión y poder utilizarla para generar pronósticos. Es en este sentido que entran en escena las observaciones recolectadas de la serie, pues ellas son analizadas y sujetas a modelados matemáticos que logren capturar el proceso que gobierna a toda la serie cronológica (Zhang, 2003). Los pronósticos se generan a partir de este modelo, es decir, pronosticar el futuro, se utilizan las correlaciones con las observaciones pasadas.

En un proceso determinístico, es posible predecir con certeza lo que ocurrirá en el futuro; las series cronológicas, sin embargo, carecen de esta condición. El análisis de series cronológicas asume que las observaciones pueden ajustarse a un determinado modelo estadístico, esto se conoce como

un proceso estocástico. Es de esta manera que Hipel & McLeod (1994) sugieren que una serie cronológica puede considerarse como una muestra aleatoria de una serie mucho más grande.

Como una serie de tiempo puede considerarse como un proceso estocástico, éstas se encuentran sujetas a múltiples supuestos. El más fundamental de ellos es que todas las observaciones son independientes e idénticamente distribuidas (i.i.d.) siguiendo una distribución aproximadamente Normal, con una media y variancia dadas. Lo anterior, sin embargo, es contrario al uso de las observaciones pasadas para pronosticar el futuro, por lo que este supuesto, según Cochrane (1997), no es exacto pues una una serie de tiempo no es exactamente, i.i.d., sino que siguen un patrón medianamente regular en el largo plazo.

Otro concepto de interés en las series cronológicas es el de estacionaridad. Una serie se considera estacionaria cuando su nivel medio y su variancia son aproximadamente las mismas durante todo el periodo, es decir, el tiempo no afecta a estos estadísticos de variabilidad. Este supuesto busca simplificar la identificación del proceso estocástico con el objetivo de obtener un modelo adecuado para generar los pronósticos. Sin embargo, y de una manera similar al supuesto de i.i.d., si una serie cronológica posee tendencias o patrones estacionales hace que esta sea no estacionaria. En la práctica, una serie puede volverse estacionaria al aplicarle transformaciones o diferenciaciones de distinto orden.

El último supuesto, y quizá el que más debate genera, es el criterio de parsimonia. Como mencionan Zhang (2003) y Hipel & McLeod (1994), este principio sugiere que se prioricen modelos sencillos, con pocos parámetros, para representar una serie de datos. Mientras más grande y complicado sea el modelo, mayor será el riesgo de sobre ajuste, lo que implica que el ajuste sea muy bueno en el conjunto de datos con que se generó el modelo, pero que los pronósticos generados sean pobres ante nuevos conjuntos de datos. Este problema, sin embargo, se presenta al considerar un único modelo con muchos parámetros; pero si se consideran varios modelos y estos son sometidos a distintos criterios, puede obtenerse un modelo sobreparametrizado que ofrezca buenos pronósticos.

2.4 Modelos Autorregresivos Integrados de Medias Móviles

Hay dos grandes grupos de modelos lineales de series cronológicas: Los modelos Autorregresivos (AR) (Lee, s. f.) y los modelos de Medias Móviles (MA) (Box, Jenkins, & Reinsel, 1994). La combinación de estos dos grandes grupos forman los Modelos Autorregresivos de Medias Móviles (ARMA) (Hipel & McLeod, 1994) y los modelos Autorregresivos Integrados de Medias Móviles (ARIMA), siendo este último de particular interés en esta investigación.

Los modelos ARIMA son los de uso más extendido en el análisis de series cronológicas. Se fundamentan en las autocorrelaciones pasadas, y contempla un proceso iterativo para identificar un posible proceso óptimo a partir de una clase general de modelos. El teorema de Wold (Surhone,

Timpledon, & Marseken, 2010) sugiere que todo proceso estacionario puede ser determinado de una forma específica y cuya ecuación posee, en realidad, infinitos coeficientes, pero que debe ser reducido a una cantidad finita para luego evaluar su ajuste sometiéndolo a diferentes pruebas y medidas de rendimiento.

2.4.1 Modelos Autorregresivos

Un modelo autorregresivo de orden p, denotado como $AR(p)$, considera los valores futuros de una serie cronológica como una combinación lineal las p observaciones predecesoras, un componente aleatorio y un término constante. Hipel & McLeod (1994) y Lee (s. f.) emplean la notación de la ecuación 3.

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t \tag{3}$$

Donde y_t y ε_t corresponden al valor de la serie y al componente aleatorio en el momento actual t, mientras que φ_i, con $i = 1, 2, \cdots, p$ son los parámetros del modelo, y c es su término constante, que en ciertas ocasiones se suele omitir para simplificar la notación. Los parámetros de esta clase de modelos suelen estimarse mediante la ecuación de Yule-Walker (Brockwell & Davis, 2009).

2.4.2 Modelos de Medias Móviles

De manera similar a como un $AR(p)$ utiliza los valores pasados para pronosticar los futuros, los modelos de medias móviles de orden q, denotados como $MA(q)$, utilizan los errores pasados de las variables independientes. Estos modelos se describen mediante la ecuación 4.

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{4}$$

Donde μ representa el valor medio de la serie cronológica y cada valor de $\theta_j (j = 1, 2, \cdots, q)$ son los parámetros del modelo. Como los $MA(q)$ utilizan los errores pasados de la serie cronológica, se asume que estos son i.i.d. centrados en cero y con una variancia constante, siguiendo una distribución aproximada mente Normal, con lo cual este tipo de modelos pueden considerarse como una regresión lineal entre una observación determinada y los términos de error que le preceden.

2.4.3 Metodología Box-Jenkins

La combinación de un $AR(p)$ y un $MA(q)$, descritos en las ecuaciones 3 y 4 respectivamente, como se mencionó al inicio de esta sección, generan los modelos autorregresivos de medias móviles, $ARMA(p, q)$, representados mediante la ecuación 5.

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{5}$$

Cochrane (1997) menciona que los modelos $ARMA(p,q)$ suelen manipularse mediante lo que se conoce como operador de rezagos, denotado como $Ly_t = y_{t-1}$. Esto significa que en un $AR(p)$ se tiene que $\varepsilon_t = \varphi(L)y_t$, mientras que en $MA(q)$ se tiene que $y_t = \theta(L)\varepsilon_t$, y por consiguiente en un $ARMA(p,q)$ se tiene $\varphi(L)y_t = \theta(L)\varepsilon_t$. Por lo tanto, de lo anterior se desprende que $\varphi(L) = 1 - \sum_{i=1}^{p} \varphi_i L^i$, y que $\theta(L) = 1 + \sum_{j=1}^{q} \theta_j L^j$.

Los modelos $ARMA$, sin embargo, solamente pueden ser utilizados en series cronológicas suyo proceso es estacionario. Esto, en la práctica, es poco común, pues una serie de tiempo a menudo posee tendencias y ciertos patrones estacionales y, además, como menciona Hamzaçebi (2008), presentan procesos no estacionarios por naturaleza. Esta condición hace necesaria la introducción de una generalización de los modelos $ARMA$, la cual se conoce como los modelos $ARIMA$ (Box et al., 1994).

Partiendo de una serie con un proceso no estacionario, es posible aplicar transformaciones o diferenciaciones (d)a los datos con el objetivo de convertirlos en un proceso estacionario. Utilizar la notación de rezagos descrita anteriormente, según Flaherty & Lombardo (2000), permite plantear un modelo $ARIMA(p,d,q)$ como se describe en la ecuación 6.

$$\varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \left(1 - \sum_{i=1}^{p} \varphi_i L^i\right)(1-L)^d y_t = \left(1 + \sum_{j=1}^{q} \theta_j L^j\right)\varepsilon_t \tag{6}$$

Donde los términos p, d y q son positivos y mayores a cero y corresponden al modelo autorregresivo, a la diferenciación y al modelo de medias móviles, respectivamente. El componente d es el número de diferenciaciones, si $d = 0$ se tiene un modelo ARMA, y $d \geq 1$ representa el número de diferenciaciones; en la mayoría de casos $d = 1$ suele ser suficiente. Así, un $ARIMA(p,0,0) = AR(p)$, $ARIMA(0,0,q) = MA(q)$, y un $ARIMA(0,1,0) = y_t = y_{t-1} + \varepsilon_t$, es decir, un modelo de caminata aleatoria.

Como sugieren Box et al. (1994), lo anterior puede generalizarse aún más al considerar los efectos estacionales de la serie cronológica. Si se considera una serie cronológica con observaciones mensuales, una diferenciación de primer orden es igual a la diferencia entre una observación y la observación correpondiente al mismo mes pero del año anterior; es decir, si el periodo estacional es de $s = 12$ meses, entonces esta diferencia estacional aplicada a un $ARIMA(p,d,q)(P,D,Q)_S$ es calculada mediante $z_t = y_t - y_{t-s}$.

De esta manera, el método de Box et al. (1994) inicia con el análisis exploratorio de la serie cronológica, teniendo un interés particular en identificar si hay presencia de factores no estacionarios

en la misma. Si en efecto se cuenta con una serie no estacionaria, ésta debe volverse estacionaria mediante algún tipo de transformación, típicamente el logaritmo natural. Con la serie ya transformada, se busca identificar el proceso que gobierna la serie, la forma clásica de hacer esto es mediante los gráficos de autocorrelación y autocorrelación parcial. Cuando se logra identificar un proceso que se adecue más a la serie cronológica, se deben realizar los diagnósticos para evaluar la calidad del ajuste del modelo, así como las medidas de rendimiento referentes a los pronósticos que genera el modelo estimado hasta un horizonte determinado.

2.4.4 Identificación del modelo

Los métodos más clásicos para la identificación del proceso que gobierna a una serie cronológica son las funciones de autocorrelación y autocorrelación parcial, las cuales sirven de indicador acerca de qué tan relacionadas están las observaciones unas de otras. Estas funciones ofrecen indicios sobre el orden de los términos para los modelos $AR(p)$, $MA(q)$ y para la diferenciación y, por ende, para la identificación de un modelo $ARIMA$.

Para medir la relación lineal entre dos variables cuantitativas es común utilizar el coeficiente de correlación r de Pearson (Benesty & Chen, 2009), el cual se define para dos variables X e Y como se muestra en la ecuación 7.

$$r_{X,Y} = \frac{E(XY)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}} \tag{7}$$

Este mismo concepto puede aplicarse a las series cronológicas para comparar el valor de la misma en el tiempo t, con su valor en el tiempo $t-1$, es decir, se comparan las observaciones consecutivas Y_t con Y_{t-1}. Esto también es aplicable a no solo una observación rezagada (Y_{t-1}), sino también con múltiples rezagos $(Y_{t-2}), (Y_{t-3}), \cdots, (Y_{t-n})$. Para esto se hace uso del coeficiente de autocorrelación.

El coeficiente de autocorrelación (ACF por sus siglas en inglés) recibe su nombre debido a que se utiliza el coeficiente de correlación para pares de observaciones $r_{Y_t, Y_{t-1}}$ de la serie cronológica. Al conjunto de todas las autocorrelaciones se le llama función de autocorrelación.

La función de autocorrelación parcial ($PACF$ por sus siglas en inglés), como menciona Hernández (2011b), busca medir la asociación lineal entre las observaciones Y_t y Y_{t-k}, descartando los efectos de los rezagos $1, 2, \cdots, k-1$.

Cuando se tiene el modelo ARIMA debidamente identificado, es importante realizar los pronósticos. Sin embargo, estos pronósticos no son imperativos, sino que se debe evaluar su calidad con las llamadas medidas de rendimiento. Estas mediciones son hechas comparando el pronóstico y su diferencia con el valor real. Existen múltiples medidas de rendimiento, Adhikari, K, & Agrawal

(2013) menciona entre ellas el *MAE*, *MAPE*, *RMSE*, *MASE*, *AIC*, *AICc* y el *BIC*.

2.5 Los autocorrelogramas

El uso del *ACF* y el *PACF* se suele aplicar de manera visual. Sin embargo, hacer usos de estos elementos implica considerar múltiples condiciones. En el caso de la identificación del orden de la diferenciación:

- Si la serie posee autocorrelaciones positivas en un amplio número de rezagos, entonces es posible que se requiera un orden más alto en el valor de d.
- Si la autocorrelación en $t-1$ es menor o igual a cero, o si las autocorrelaciones resultan ser muy bajas y sin seguir algún patrón en particular, entonces no se requiere un alto orden para la diferenciación.
- Una desviación estándar baja suele ser indicador de un orden adecuado de integración.
- Si no se utiliza ninguna diferenciación, se asume que la serie cronológica es estacionaria. Aplicar una diferenciación asume que la serie cronológica posee una media constante, mientras que dos diferenciaciones sugiere que la tendencia varía en el tiempo.

Para la identificación de los términos p y q:

- Si la *PACF* de la serie cronológica diferenciada muestra una diferencia marcada y si, además, la autocorrelación en $t-1$ es positiva, entonces debe considerarse aumentar el valor de p.
- Si la *PACF* de la serie cronológica diferenciada muestra una diferencia marcada y si, además, y la autocorrelación en $t-1$ es negativa, entonces debe considerarse aumentar el valor de q.
- Los términos p y q pueden cancelar sus efectos entre sí, por lo que si se cuenta con un modelo $ARMA$ más mixto que parece adaptarse bien a los datos, puede deberse también a que p o q deben ser menores.
- Si la suma de los coeficientes del modelo AR es muy cercana a la unidad, es necesario reducir la cantidad de términos en uno y aumentar el orden de la diferenciación en uno.
- Si la suma de los coeficientes del modelo MA es muy cercana a la unidad, es necesario reducir la cantidad de términos en uno y disminuir el orden de la diferenciación en uno.

Tener en consideración estos y otros posibles criterios para la identificación del proceso que gobierna la serie cronológica puede fácilmente volverse algo subjetivo, pues dos personas diferentes pueden llegar a dar distintas interpretaciones a las visualizaciones de los autocorrelogramas. Estas interpretaciones pueden sesgar la identificación de los modelos y, además, no considerar otros escenarios para los términos de un modelo $ARIMA$; para solventar esto es necesario considerar un abanico más amplio de opciones, lo cual se puede lograr al considerar múltiples permutaciones de términos, es decir, empleando la sobreparametrización.

2.6 La sobreparametrización y el análisis combinatorio

La identificación visual mediante los autocorrelogramas puede llevar a decisiones erradas acerca del proceso que gobierna la serie cronológica. Una alternativa es considerar estimaciones procesos de ordenes bajos, como un $ARMA(1,1)$ y poco a poco ir incorporando términos, este proceso de revisión permite encontrar los puntos en que agregar un coeficiente más al modelo no aporta ninguna mejora en los resultados del pronóstico, y así considerar únicamente aquellos modelos que tengan coeficientes con un aporte estadísticamente significativo. Este procedimiento es conocido como sobreparametrización. Dependiendo de la cantidad de observaciones y del rango con que se trabajen los coeficientes, la comparación de los modelos puede volverse muy extensa y complicada, razón por la cual resulta imperativo generar un procedimiento sistemático que logre seleccionar el mejor modelo con base en sus medidas de ajuste y rendimiento del modelo.

3 METODOLOGÍA

3.1 Introducción

La aplicación de las series cronológicas tiene tres objetivos: 1) el análisis exploratorio de la serie en cuestión, 2) estimar modelos de proyección, y 3) generar pronósticos para los posibles valores futuros que tomará el problema en cuestión. Asimismo, existen múltiples formas de proceder mediante la etapa de estimación, como lo son los métodos de suavizamiento exponencial (Brown, 1956), modelos de regresión para series temporales (Kedem & Fokianos, 2005), redes neuronales secuenciales aplicadas a datos longitudinales (Tadayon & Iwashita, 2020), estimaciones bayesianas (Jammalamadaka, Qiu, & Ning, 2018), y finalmente, los procesos autorregresivos integrados de medias móviles o ARIMA por sus siglas en inglés (Box et al., 1994), siendo estos últimos el foco de interés en este estudio.

Un proceso ARIMA es caracterizado por dos funciones: la autocorrelación y la autocorrelación parcial; mediante la comparación de dichas funciones se busca la identificación del proceso que describa de manera adecuada el comportamiento de una serie cronológica.

En la búsqueda de un modelo adecuado entre varios candidatos, se llevan a cabo comparaciones de medidas de bondad de ajuste y de precisión. Se consideran temporalidades mensuales, bimensuales, trimestrales o cuatrimestrales, mediante un proceso de selección fundamentada en las permutaciones de todos los parámetros de un modelo ARIMA hasta un rango determinado, considerando la inclusión semiautomática de intervenciones en periodos específicos y la validación cruzada para evaluar la calidad de las particiones de la base de datos en conjuntos para entrenar y probar el rendimiento del modelo. Dichas pruebas sirven de insumo para utilizar un método de consenso entre ellas y seleccionar el modelo más adecuado mediante la sobreparametrización: se comparan todos los posibles en in intervalo específico de términos definiendo una diferenciación adecuada para la serie y permutando hasta un máximo definido para los términos autorregresivos y de medias móviles especificados para así seleccionar la especificación que ofrezca mejores resultados al momento de pronosticar valores futuros de la serie cronológica.

3.2 Conceptos y definiciones en el análisis de series cronológicas

3.2.1 Definición de una serie cronológica

3.2.2 Procedimiento al analizar series cronológicas

3.2.3 Estacionaridad

3.2.4 La parsimonia

3.3 Componentes de una serie cronológica

3.3.1 La tendencia

3.3.2 Componentes estacionales

3.3.3 Componente cíclico

3.3.4 Componente irregular

3.4 Modelos de series cronológicas

3.5 Modelos Autorregresivos Integrados de Medias Móviles

3.5.1 Modelos Autorregresivos

3.5.2 Modelos de Medias Móviles

3.5.3 Metodología Box-Jenkins

3.5.4 Etapa 1 - Identificación

3.5.5 Etapa 2 - Estimación y diagnóstico

3.5.6 Etapa 3 - Pronóstico

3.5.7 Notación de los modelos ARIMA

3.5.8 Diferenciación

3.6 Análisis de intervención

3.7 Validación cruzada

3.8 Medidas de rendimiento

3.8.1 MFE

3.8.2 MAE

3.8.3 MAPE

3.8.4 MPE

3.8.5 MSE

4 RESULTADOS

4.1 Introducción

El método propuesto se probará comparándose con los resultados de seis series con distintas temporalidades: mortalidad infantil, mortalidad por causa externa, nacimientos, demanda eléctrica, intereses y comisiones del sector público e incentivos salariales del sector público.

4.2 Datos simulados

4.2.1 Comparación en datos simulados - Sobreparametrización vs auto.arima

4.3 Estimaciones en datos costarricenses

En el campo demográfico, por ejemplo, las estadísticas vitales son sistematizadas y divulgadas año tras año, por tanto, revelan los cambios acontecidos durante este periodo. Esta información junto con la proveniente de los censos de población constituye la base para construir los diferentes índices, tasas y otros indicadores que revelan la situación demográfica del país, información de gran relevancia para la planificación nacional, regional y local en diversos campos. Uno de estos principales campos de acción es la salud pública, para la cual la tasa de mortalidad infantil se considera uno de los indicadores prioritarios dado que refleja no solo las condiciones de salud de la población infante, sino también los niveles de desarrollo del país, pues depende de la calidad de la atención de la salud, principalmente de la prenatal y perinatal, así como de las condiciones de saneamiento. Por tanto, su continuo monitoreo es fundamental para diseñar, implementar y evaluar políticas de salud pública orientadas a disminuir y erradicar aquellas que son prevenibles (INEC, 2017).

4.3.1 Tasa de mortalidad infantil interanual

4.3.2 Tasa global de fecundidad

4.3.3 Mortalidad por causa externa

4.3.4 Incentivos salariales del sector público

4.3.5 Intereses y comisiones del sector público

4.3.6 Demanda eléctrica

4.3.7 Comparación en datos reales - Sobreparametrización vs auto.arima

4.4 Discusión de los resultados

5 CONCLUSIONES Y RECOMENDACIONES

5.1 Introducción

5.2 Conclusiones

5.3 Recomendaciones

6 ANEXOS

6.1 La función funcion_1

Código 1: Una función

```
funcion_1 <- function(x,y){
    x+y
}
```

7 REFERENCIAS

Adhikari, R., K, A. R., & Agrawal, R. K. (2013). *An Introductory Study on Time Series Modeling and Forecasting* (pp. 42-45). Recuperado de https://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf

Benesty, J., & Chen, Y. C., J.and Huang. (2009). Pearson Correlation Coefficient. En *Noise Reduction in Speech Processing* (pp. 37-38). https://doi.org/10.1007/978-3-642-00296-0_5

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Recuperado de https://books.google.co.cr/books?id=sRzvAAAAMAAJ

Brockwell, P. J., & Davis, R. A. (2009). *Time Series: Theory and Methods*. En *Springer Series in Statistics* (p. 239). Recuperado de https://books.google.co.cr/books?id=/_DcYu/_EhVzUC

Brown, R. (1956). *Exponential Smoothing for Predicting Demand*. Recuperado de https://www.industrydocuments.ucsf.edu/docs/jzlc0130

Burnham, K. P., & Anderson, D. R. (2007). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Recuperado de https://books.google.co.cr/books?id=IWUKBwAAQBAJ

Canova, F., & Hansen, B. E. (1995). Are Seasonal Patterns Constant over Time? A Test for Seasonal Stability. *Journal of Business & Economic Statistics*, *13*(3), 237-252. Recuperado de http://www.jstor.org/stable/1392184

Cochrane, J. H. (1997). *Time Series for Macroeconomics and Finance*. Recuperado de http://econ.lse.ac.uk/staff/wdenhaan/teach/cochrane.pdf

Flaherty, J., & Lombardo, R. (2000, enero). *Modelling Private New Housing Starts in Australia*. Recuperado de http://www.prres.net/papers/Flaherty_Modelling_Private_New_Housing_Starts_In_Australia.pdf

Fuller, W. A. (1995). *Introduction to Statistical Time Series*. Recuperado de https://books.google.co.cr/books?id=wyRhjmAPQIYC

Hamzaçebi, C. (2008). Improving Artificial Neural Networks' Performance in Seasonal Time Series Forecasting. *Inf. Sci.*, *178*(23), 4550-4559. https://doi.org/10.1016/j.ins.2008.07.024

Hernández, O. (2011a). *Introducción a las Series Cronológicas* (1.ª ed.). Recuperado de http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html

Hernández, O. (2011b). *Introducción a las Series Cronológicas*. Recuperado de http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html

Hipel, K. W., & McLeod, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems*. Recuperado de https://books.google.co.cr/books?id=t1zG8OUbgdgC

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. Recuperado de https://books.google.co.cr/books?id=/_bBhDwAAQBAJ

Hyndman, R., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software, Articles*, *27*(3), 1-22. https://doi.org/10.18637/jss.v027.i03

INEC. (2017). *Población, nacimientos, defunciones y matrimonios*. Recuperado de http://inec.cr/sites/default/files/documetos-biblioteca-virtual/repoblacev2017_0.pdf

Jammalamadaka, S. R., Qiu, J., & Ning, N. (2018). *Multivariate Bayesian Structural Time Series Model*. Recuperado de https://arxiv.org/pdf/1801.03222.pdf

Kedem, B., & Fokianos, K. (2005). *Regression Models for Time Series Analysis*. Recuperado de https://books.google.co.cr/books?id=8r0qE35wt44C

Lee, J. (s. f.). Univariate time series modeling and forecasting (Box-Jenkins Method). *Econ 413, lecture 4*.

Osborn, D. R., Chui, A. P. L., Smith, J., & Birchenhall, C. (2009). *Seasonality and the order of integration for consumption*. Recuperado de http://www.est.uc3m.es/esp/nueva_docencia/comp_col_get/lade/tecnicas_prediccion/OCSB_OxBull1988.pdf

Rosero-Bixby, L. (2018). *Producto C para SUPEN. Proyección de la mortalidad de Costa Rica 2015-2150*. Recuperado de CCP-UCR website: http://srv-website.cloudapp.net/documents/10179/999061/Nota+t%C3%A9cnica+tablas+de+vida+segunda+parte

Surhone, L. M., Timpledon, M. T., & Marseken, S. F. (2010). *Wold Decomposition*. Recuperado de https://books.google.co.cr/books?id=7cSqcQAACAAJ

Tadayon, M., & Iwashita, Y. (2020). *Comprehensive Analysis of Time Series Forecasting Using Neural Networks*. Recuperado de https://arxiv.org/pdf/2001.09547.pdf

Xiao, Z. (2001). Testing the Null Hypothesis of Stationarity Against an Autoregressive Unit Root Alternative. *Journal of Time Series Analysis*, *22*(1), 87-105. https://doi.org/10.1111/1467-9892.00213

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.