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, 2011). 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 ARIMA, son los de uso más extendido en el análisis de series cronológicas. El nombre ARIMA es la abreviatura inglesa para AutoRegresive Integrated Moving Average, y son aplicados mediante la metodología de Box-Jenkins.

De esta manera, el método de Box-Jenkins inicia con el análisis exploratorio de la serie cronológica de interés, 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 adecúe 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.

El método ARIMA se fundamenta 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 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.

2.2 Investigaciones relacionadas

2.3 Observaciones finales sobre la revisión bibliográfica

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, Jenkins, & Reinsel, 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.

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

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

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

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

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

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

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

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