Determinación de modelos ARIMA vía sobre parametrización según la temporalidad de la serie cronológica con aplicaciones en datos costarricenses

Universidad de Costa Rica

César Gamboa Sanabria

www.cesargamboasanabria.com

info@cesargamboasanabria.com

03 octubre, 2019

RESUMEN

La metodología de Box-Jenkins busca encontrar el mejor proceso autorregresivo integrado de medias móviles (ARIMA) que explique una serie temporal y_t de T periodos, para pronosticar hasta $T + h$. El paquete `forecast` de R permite hacer uso de la función `auto.arima()` para estimar un modelo ARIMA basado en pruebas de raíz unitaria, minimización del AICc y de la MLE. De esta forma se obtiene un modelo temporal definiendo las diferenciaciones requeridas en la parte estacional d mediante las pruebas KPSS o ADF, y la no estacional D utilizando las pruebas OCSB o la Canova-Hansen, seleccionado el orden óptimo para los términos $ARIMA(p, d, q)(P, D, Q)$ para una serie y_t. Se propone un método de selección fundamentada en las permutaciones de los parámetros de un modelo ARIMA, seleccionando la mejor especificación con base en medidas de rendimiento MAE, RMSE, MAPE y MASE: se comparan todos los posibles términos definiendo una diferenciación adecuada para la serie y permutando hasta un máximo determinado para los términos de especificación de un $ARIMA(p, d, q)(P, D, Q)$. El método propuesto se probó en 6 series cronológicas de distinta temporalidad: mortalidad infantil, mortalidad por causa externa, nacimientos, demanda eléctrica, intereses y comisiones del sector público, e incentivos salariales del sector público.

Palabras clave: ARIMA, R, automatización, selección, estadística

Índice

1 INTRODUCCIÓN

El manejo de información obtenida de manera secuencial a lo largo del tiempo hace referencia al uso de series cronológicas. Este tipo de datos se encuentras en diferentes áreas de investigación. En el campo financiero, por ejemplo, es común hablar de la devaluación del colón con respecto al dólar, cantidad de exportaciones mensuales de un determinado producto o las ventas de este (Hernández 2011a).

En demografía, por ejemplo, el tema de las proyecciones de población tiene un alto impacto a nivel social, pues conocer con anticipación en posible comportamiento de la población en el futuro es clave para una adecuada planificación en diversos proyectos sobre los cuales se debe distribuir un presupuesto que es finito. Durante una emergencia, que difícilmente se sabe cuándo ocurrirá, conocer la posible población que se tiene en una zona es clave para la rápida reacción de las autoridades para el envío de ayuda o para ejecutar planes de evacuación.

El campo actuarial también se ve beneficiado al mejorar sus métodos de pronóstico, pues uno de sus campos de estudio es la mortalidad pues representan 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).

Sin embargo, las series cronológicas por sí solas representan solo un insumo para abordar, como mínimo, tres objetivos básicos: 1) realizar análisis exploratorios de estos datos mediante métodos de visualización y medidas de posición y variabilidad, como ver su crecimiento o decrecimiento a lo largo del tiempo, detectar valores atípicos o cambios drásticos en el nivel o valor medio de la serie, 2) generar modelos estadísticos que sirvan como un simplificación de la realidad, y 3) generar pronósticos para los posibles valores futuros que tomará el problema en cuestión (Hernández 2011b).

Los tres objetivos anteriores se trabajan de manera secuencial, pues es necesario realizar primero el análisis exploratorio de los datos para tener una noción global del panorama y así conocer la serie cronológica con la que se está trabajando. Una vez hecho esto, existen múltiples formas de generar modelos para estos datos, como por ejemplo los métodos de suavizamiento exponencial desarrollados en la década de 1950 (Brown 1956), modelos de regresión para series temporales (Kedem and Fokianos 2005) o los procesos autorregresivos integrados de medias móviles (ARIMA) (Box, Jenkins, and Reinsel 1994). Cuando se ha establecido el modelo, los pronósticos son utilizados en instituciones públicas, gobiernos municipales, instituciones del sector privado, centros académicos, población civil, centros nacionales o regionales de investigación y ONG 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 mayor y más efectivo.

De lo anterior, generar un modelo adecuado es fundamental para obtener un pronóstico de calidad, y es aquí donde resulta importante mencionar una diferencia clave entre los dos modelos clásicos más comúnmente utilizados: los modelos de suavizamiento y los modelos ARIMA. Ambos representan enfoques complementarios a un problema, pues según Hyndman (Hyndman and Athanasopoulos 2018), los modelos de suavizamiento exponencial se fundamentan en un enfoque más descriptivo de los componentes de la serie cronológica en estudio; mientras que los modelos ARIMA tienen como objetivo explicar las relaciones pasadas de ésta. La importancia de la metodología de Box-Jenkins radica en que no supone ningún patrón en particular en la serie histórica que se busca pronosticar, sino que contempla un proceso iterativo para identificar un posible modelo a partir de una clase general de modelos y luego someter dicho modelo a diferentes pruebas y medidas de rendimiento para evaluar su ajuste. Al trabajar la metodología de Box-Jenkins, uno de los pasos es identificar el los parámetros del proceso ARIMA(p,d,q)(P,D,Q) que gobiernan la serie, siendo la manera clásica de trabajar este paso, el análisis visual de las funciones de autocorrelación parcial y total.

El gran obstáculo que presenta esta identificación visual es que en la actualidad contar con una gran cantidad de series cronológicas para analizar es algo muy común. Incluso con cantidades moderadas de series cronológicas a analizar, es difícil contar con personal capacitado para realizar este análisis visual y poder identificar los modelos, por lo que la generación de algoritmos que ayuden a dicha identificación se vuelven cada vez más necesarios (Hyndman and Khandakar 2008).

Han sido varias las aproximaciones a un método que genere de manera automática un modelo ARIMA, como por ejemplo los propuestos por Hannan y Rissanen (Hannan and Rissanen 1982), la extensión de dicha propuesta realizada por Gómez (Gómez 1998) y posteriormente aplicada (Gómez and Maraval 1998) en los software **TRAMO** y **SEATS**; de manera similar se planteó una aplicación en los software **SCA-Expert** (Liu 1989) y **TSE-AX** (Mélard and Pasteels 2000). Otros algoritmos implementados en programas de cómputo de paga son **Forecast Pro** (Goodrich 2000) y **Autobox** (Reilly 2000). Uno de los métodos automatizados de estimación es el que ofrece el paquete `forecast` (Hyndman and Khandakar 2008) del lenguaje de programación R[1] permite hacer uso de la función `auto.arima()` para estimar un modelo ARIMA basado en pruebas de raíz unitaria, minimización del AICc y de la MLE. De esta forma se obtiene un modelo temporal definiendo las diferenciaciones requeridas en la parte estacional d mediante las pruebas KPSS o ADF, y la no estacional D utilizando las pruebas OCSB o la Canova-Hansen, seleccionado el orden óptimo para los términos $ARIMA(p,d,q)(P,D,Q)_s$ para una serie cronológica determinada.

Es a partir de esta necesidad que se propone una metodología para la estimación del mejor modelo ARIMA para una serie cronológica determinada cuya temporalidad sea mensual, bimensual, trimestral o cuatrimestral mediante un proceso de selección fundamentada en las permutaciones de todos los pará-

[1] https://cran.r-project.org/

metros de un modelo ARIMA hasta un cierto límite, considerando además la inclusión semi-automática 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 involucran, entre otras medidas de rendimiento, el MAE, RMSE, MAPE y MASE, las cuales sirven de insumo para utilizar un método de consenso entre ellas para seleccionar el modelo más adecuado: se comparan todos los posibles términos definiendo una diferenciación adecuada para la serie y permutando hasta un máximo definido para los términos de especificación de un $ARIMA(p,d,q)(P,D,Q)_s$ para así seleccionar la especificación que ofrezca mejores resultados al momento de pronosticar valores futuros de la serie cronológica. El método propuesto se probará comparándose con los resultados de 6 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.

1.1 Contribución de la tesis a la Estadística como disciplina

El principal aporte de este estudio es, por medio de un estudio de simulación, aportar evidencia sobre cómo la sobreparametrización puede representar una herramienta para definir la especificación de un modelo ARIMA que genere pronósticos adecuados, contrastando la calidad de estos con respecto a otros métodos similares, como lo es la función auto.arima().

1.2 Objetivos

1.2.1 Objetivos general

- Evaluar la calidad de los pronósticos realizados con modelos ARIMA especificados vía sobre parametrización.

1.2.2 Objetivos específicos

- Diseñar un algoritmo para la selección del mejor modelo ARIMA según la temporalidad de la serie.
- Aplicar validación cruzada en distintos horizontes de pronóstico para identificar la mejor especificación de un modelo ARIMA.
- Comparar la precisión de los pronósticos con el método propuesto por Rob Hyndman.
- Integrar la metodología de análisis de series temporales en una librería del lenguaje estadístico R.

2 REFERENCIAS

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

Brown, Robert G. 1956. *Exponential Smoothing for Predicting Demand*. A.D.Little. https://www.industrydocuments.ucsf.edu/docs/jzlc0130.

Goodrich, RL. 2000. "The Forecast Pro Methodology." *International Journal of Forecasting* 16 (4): 533–35. http://www.forecasting-competition.com/downloads/NN3/methods/Goodrich%20(2000)%20The%20Forecast%20Pro%20methodology%20science.pdf.

Gómez, V. 1998. "Automatic Model Identification in the Presence of Missing Observations and Outliers." Edited by Dirección General de Análisis y Programación Presupuestaria Ministerio de Economía y Hacienda. Working paper D-98009.

Gómez, V., and A. Maraval. 1998. "Programs Tramo and Seats, Instructions for the Users." Edited by Dirección General de Análisis y Programación Presupuestaria Ministerio de Economía y Hacienda. Working paper 97001.

Hannan, E. J., and J. Rissanen. 1982. "Recursive Estimation of Mixed Autoregressive-Moving Average Order." *Biometrika* 69 (1): 81–94. http://www.jstor.org/stable/2335856.

Hernández, O. 2011a. "Introducción a Las Series Cronológicas." In, 1st ed., 1. Editorial Universidad de Costa Rica. http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html.

———. 2011b. "Introducción a Las Series Cronológicas." In, 1st ed., 2. Editorial Universidad de Costa Rica. http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html.

Hyndman, R.J., and G. Athanasopoulos. 2018. *Forecasting: Principles and Practice*. OTexts. https://books.google.co.cr/books?id=/_bBhDwAAQBAJ.

Hyndman, Rob, and Yeasmin Khandakar. 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.

Kedem, B., and K. Fokianos. 2005. *Regression Models for Time Series Analysis*. Wiley Series in Probability and Statistics. Wiley. https://books.google.co.cr/books?id=8r0qE35wt44C.

Liu, Lon-Mu. 1989. "Identification of Seasonal Arima Models Using a Filtering Method." *Communications in Statistics - Theory and Methods* 18 (6): 2279–88. https://doi.org/10.1080/03610928908830035.

Mélard, G., and J.-M. Pasteels. 2000. "Automatic Arima Modeling Including Interventions, Using Time Series Expert Software." *International Journal of Forecasting* 16 (4): 497–508. https://doi.org/https://doi.org/10.1016/S0169-2070(00)00067-4.

Reilly, D. 2000. "The Autobox System." *International Journal of Forecasting* 16 (4): 531–33. https://

ideas.repec.org/a/eee/intfor/v16y2000i4p531-533.html.

Rosero-Bixby, L. 2018. “Producto c Para Supen. Proyección de La Mortalidad de Costa Rica 2015-2150.” CCP-UCR. http://srv-website.cloudapp.net/documents/10179/999061/Nota+t%C3%A9cnica+tablas+de+vida+segunda+parte.