UNIVERSIDAD DE COSTA RICA

Solicitud de Aprobación de Tema y Comité Asesor para Tesis

Maestría Académica en Estadística

PPEs Programa de Posgrado en Estadística

17 de diciembre del 2019

Miembros de la Comisión
Programa de Posgrado en Estadística

Estimados señores:

Yo, **César Andrés Gamboa Sanabria**, carné **B12672**, estudiante de la Maestría Académica en Estadística, solicito la aprobación de mi Tema y Comité Asesor para trabajar la Tesis.

El tema propuesto es ***"La sobreparametrización en el ARIMA: una aplicación a datos costarricenses"***, cuya explicación se adjunta en la Propuesta de Tema de Tesis. Además, la propuesta del Comité Asesor es la siguiente: Director de Tesis **MSc. Oscar Centeno Mora** y como asesores **Ph.D. Gilbert Brenes Camacho** y **Ph.D. ShuWei Chou.**

Atentamente,

Firma
César Andrés Gamboa Sanabria
B12672

Firma
MSc. Oscar Centeno Mora
V°B° Profesor tutor

Firma
Ph.D. Gilbert Brenes Camacho
V°B° Lector

Firma
Ph.D. ShuWei Chou.
V°B° Lector

Importante:

- Si uno de los miembros del Comité Asesor no pertenece al **Posgrado en Estadística** debe adjuntar el Curriculum Vitae que incluya copia de títulos universitarios.
- Debe adjuntar a esta solicitud el formulario de Propuesta de Tema de Tesis.

Estimado(a) estudiante:

- Se solicita completar la siguiente información de la manera más concreta posible en al menos tres páginas y hasta en un máximo de cinco.
- Su solicitud será revisada en la siguiente reunión de la Comisión del Posgrado, siempre y cuando la documentación sea recibida en la Administración una semana antes de esa reunión.
- Se recomienda revisar el Reglamento del Posgrado en Estadística, en todo lo referente a la realización de las tesis, para evitar inconvenientes. Abajo se especifica parte del artículo 15, para que con su firma Usted haga constar que se ajustará a los plazos establecidos.

Nombre del estudiante		César Andrés Gamboa Sanabria
Título de la tesis		La sobreparametrización en el ARIMA: una aplicación a datos costarricenses
Introducción	Justificación/importancia del tema	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 encuentra 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, el tema de las proyecciones de población tiene un alto impacto social y económico, pues conocer con anticipación el 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, generalmente, es finito. Durante una emergencia, que difícilmente se sabe cuándo ocurrirá, 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. El campo actuarial también se ve beneficiado al mejorar sus métodos de pronóstico, pues uno de sus campos 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). Sin embargo, las series cronológicas por sí solas representan solo un insumo para abordar, como mínimo, tres objetivos básicos: 1) realizar análisis exploratorios usando 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 una 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 y Fokianos 2005) o los procesos autorregresivos integrados de medias móviles (ARIMA) (Box, Jenkins, y 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

		considerando además 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 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 y seleccionar el modelo más adecuado mediante la sobreparametrización: 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 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.
	Contribución de la tesis a la Estadística como disciplina	El principal aporte de este estudio es, por medio de un proceso de simulación, aportar evidencia sobre cómo la sobreparametrización puede contribuir a definir la especificación de un modelo ARIMA que genere pronósticos adecuados, contrastando la calidad de estos con respecto a otros métodos similares, como lo son las funciones auto.arima() o seas().
Objetivos	Objetivo general	Evaluar la calidad de los pronósticos realizados con modelos ARIMA especificados vía sobreparametrización para proponer un modelo adecuado en una serie cronológica.
	Objetivos específicos	1. Diseñar un algoritmo para la selección del mejor modelo ARIMA según la temporalidad de la serie. 2. Aplicar validación cruzada en distintos horizontes de pronóstico para identificar la mejor especificación de un modelo ARIMA. 3. Comparar la precisión de los pronósticos con métodos similares, como el propuesto por Rob Hyndman, de la Oficina de Censos de los Estados Unidos, entre otros. 4. Integrar el desarrollo de la metodología de análisis de series temporales en una librería del lenguaje estadístico R.
Metodología	Referentes o elementos teóricos que va a utilizar	**Modelos Arima** Los modelos ARIMA, junto con los de suavizamiento exponencial, 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. Como menciona Rob. Hyndman (R. J. Hyndman y Athanasopoulos 2018b), la metodología de Box-Jenkins difiere a los demás métodos porque no supone un determinado patrón en la serie cronológica, sino que parte de un proceso iterativo para identificar el modelo de un gran grupo de estos para luego ponerlo a prueba según varias medidas de rendimiento. Un proceso ARIMA es caracterizado por dos funciones: la autocorrelación y la autocorrelación parcial; el enfoque Box-Jenkins compara estas funciones con el objetivo de identificar el proceso que

		características laborales que complementan las remuneraciones básicas. Los incentivos se reconocen tanto a profesionales como a no profesionales, facultados por disposiciones jurídicas que así lo autorizan. Algunos de estos incentivos son: anualidades, dedicación exclusiva, salario escolar, carrera profesional, carrera técnica, zonaje, desarraigo, regionalización, riesgo policial, riesgo penitenciario, riesgo de seguridad y vigilancia, peligrosidad, incentivo didáctico. Se calcula como la suma total de los rubros anteriores y se da en millones de colones. 5. **Intereses y comisiones del sector público**: Comprende el pago de los intereses de la deuda del gobierno, esto es, las erogaciones de intereses y comisiones destinadas por las instituciones públicas para cubrir el pago a favor de terceras personas, físicas o jurídicas, del sector privado o del sector público, residentes en el territorio nacional o en el exterior, por la utilización en un determinado plazo de recursos financieros provenientes de los conceptos de emisión y colocación de títulos valores, contratación de préstamos directos, créditos de proveedores, depósitos a plazo y a la vista, intereses por deudas de avales asumidos, entre otros pasivos de la entidad transados en el país o en el exterior. Incluye, el pago por concepto de otras obligaciones contraídas entre las partes, que no provienen de las actividades normales de financiamiento. Además, los intereses y comisiones por las operaciones normales de los bancos comerciales del sector público, así como las diferencias por tipo de cambio por operaciones financieras; y también el pago de intereses moratorios correspondientes a la deuda pública. Se calcula como la suma total de los rubros anteriores y se da en millones de colones. 6. **Demanda eléctrica**: Datos obtenidos de los informes anuales que reporta el Instituto Costarricense de Electricidad, de una serie mensual de la demanda eléctrica nacional en Megavatios hora (MWh). Se calcula como la suma total del consumo, en este caso a nivel nacional.
	Evidencias de calidad de la medición para la(s) variable(s) del estudio	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 o áreas 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, esto porque 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). Por su parte, la Contraloría General de

El componente **AR** de los modelos ARIMA hace referencia al uso de modelos autorregresivos, en los cuales los pronósticos para la variable de interés utilizan una combinación lineal de las observaciones previas, llamándose así *autorregresivos* porque se aplica una regresión de dicha variable de interés con respecto a sí misma; caso contrario a la regresión múltiple, en donde los pronósticos se realizan con respecto a una combinación lineal de distintos predictores. Un modelo autorregresivo de orden *p* para una serie cronológica y_t puede expresarse de la siguiente manera:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

Donde el término ϵ_t representa ruido blanco. El modelo anterior es muy similar a una regresión lineal múltiple, donde cada coeficiente φ va acompañado por su correspondiente rezago y_{t-p}. De manera muy similar, el término **MA** en los modelos ARIMA se refiere a los modelos de medias móviles, los cuales hacen uso de los errores para pronosticar; el modelo de medias móviles puede representarse de la siguiente manera:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

Donde el término ϵ_t representa nuevamente el ruido blanco. La ecuación anterior representa un modelo de medias móviles de orden *q*, en la cual cada término ϵ_{t-q} se entiende como una media móvil de los *t* previos errores de predicción.

El componente **I** de los modelos ARIMA se refiere a "Itegrated", es decir, a la estacionariedad de la serie cronológica. Tradicionalmente, la metodología de Box-Jenkins consiste en visualizar la serie cronológica con el objetivo de, en caso de ser necesario, transformar los datos para estabilizar la variancia y generar así un proceso estacionario. Se dice que una serie posee un comportamiento estacionario si el comportamiento de esta no depende del tiempo, por lo que en principio no presentaría ningún patrón particular con respecto al tiempo; en otras palabras, la serie posee un movimiento bastante horizontal.

Cuando la serie cronológica muestre indicios de tendencia o patrones estacionales que resulten en un conjunto de datos que no es estacionario por naturaleza, es necesario realizar transformaciones sobre los datos para hacer que la serie se vuelva estacionaria (Adhikari, K, and Agrawal 2013a). Estas transformaciones hacen referencia al uso de logaritmos o alguna potencia que logre estabilizar la variabilidad de la serie. Los métodos más clásicos para identificar la no estacionariedad en una serie cronológica son las previamente mencionadas 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 previamente mencionados **AR** y **MA**.

Función de autocorrelación

Para medir la relación lineal entre dos variables cuantitativas es común utilizar el coeficiente de correlación *r* de Pearson (Benesty y Chen 2009), el cual se define para dos variables *X* e *Y* como sigue:

MAE

El error absoluto medio se define como $\frac{1}{n}\sum_{t=1}^{n}|e_t|$.

MAPE

El porcentaje promedio de error absoluto se define como $\frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|$.

RMSE

Es la raíz del error cuadrático medio $\sqrt{\frac{1}{n}\sum_{t=1}^{n}|e_t^2|}$.

MASE

Para series no estacionales $\frac{\frac{1}{J}\sum_{J}|e_j|}{\frac{1}{T-1}\sum_{t=2}^{T}|Y_t-Y_{t-1}|}$.

Para series estacionales $\frac{\frac{1}{J}\sum_{J}|e_j|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|Y_t-Y_{t-m}|}$

Donde m es la temporalidad de la serie.

AIC

Se calcula de la siguiente manera: $AIC = -2logL(\hat{\theta}) + 2k$

Donde k es el número de parámetros y n el número de datos.

AICc

Se calcula de la siguiente manera: $AIC = -2logL(\hat{\theta}) + 2k + \frac{2k+1}{n-k-1}$

Donde k es el número de parámetros y n el número de datos.

BIC

Se calcula de la siguiente manera: $AIC = -2logL(\hat{\theta}) + 2k \cdot \log(n)$

Donde k es el número de parámetros y n el número de datos.

Estudio de simulación

Inicialmente se simulan series cronológicas partiendo de valores aleatorios de alguna distribución de probabilidad, o bien, de datos reales de alguna serie. Con estos valores iniciales se generarán n valores aleatorios que sigan un proceso específico $ARIMA(p,d,q)(P,D,Q)_s$.

Para generar el proceso en cuestión, se deben fijar los valores de p,d,q en la parte no estacional y P,D,Q en la parte estacional de un modelo ARIMA, así como la temporalidad que se desea para la misma. Además, se ofrece la posibilidad de definir el valor de los coeficientes del modelo para cada orden del proceso; por ejemplo, si se desea generar una serie cuyo proceso es un $ARIMA(2,1,1)(1,1,3)_s$, el 2 indica que se pueden fijar los valores de los

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

———. 2011c. "Introducción a Las Series Cronológicas." In, 1st ed., 77. Editorial Universidad de Costa Rica.http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html.

———. 2011d. "Introducción a Las Series Cronológicas." In, 1st ed., 69. 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. 2018a. *Forecasting: Principles and Practice*. OTexts.https:

//books.google.co.cr/books?id=/_bBhDwAAQBAJ.

———. 2018b. *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.

INEC. 2017. "Población, Nacimientos, Defunciones Y Matrimonios."http://inec.cr/sites/default/files/documetos-biblioteca-virtual/repoblacev2017_0.pdf.

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:

	Resultados en casos reales	1	Nov-20	Nov-20	Aplicación del método en series reales.
	Redacción de conclusiones	1	Dic-20	Dic-20	Conclusiones y limitaciones finales de la investigación.

Artículo 15. Para el Programa de Maestría Académica el periodo máximo entre el ingreso del estudiante a la segunda etapa (fecha del primer curso matriculado) y la presentación del examen de candidatura es de cuatro años. Si no lo aprobara en ese periodo quedará automáticamente fuera del Programa, pudiendo solicitar a la Comisión traslado al Programa de la Maestría Profesional en Estadística. El estudiante tendrá un plazo de tres ciclos lectivos para completar la tercera etapa, a partir de la fecha de aprobación del examen de candidatura. Si al cabo de este periodo el estudiante no ha presentado la tesis, la Comisión podrá conceder una única prórroga de un ciclo lectivo, al cabo del cual, el estudiante que no ha defendido su tesis será separado del programa.

FIRMA DEL ESTUDIANTE

Asunto: sobre el examen de candidatura de César
De: SHUWEI CHOU <SHUWEI.CHOU@ucr.ac.cr>
Fecha: 9/12/2019 05:55
Para: POSGRADO EN ESTADISTICA - SEP <ESTADISTICA.SEP@ucr.ac.cr>
CC: César Gamboa <info@cesargamboasanabria.com>

Estimada Cindy,

Envío este correo para confirmar que estoy de acuerdo con la propuesta del tema "La sobreparametrización en el ARIMA: una aplicación a datos costarricenses".

Saludos

Shu Wei