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](#_bookmark37)).  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](#_bookmark51)).  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](#_bookmark38)).  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](#_bookmark32)), modelos de regresión para series temporales (Kedem y Fokianos 2005[)](#_bookmark45) o los procesos autorregresivos integrados de medias móviles (ARIMA) (Box, Jenkins, y Reinsel [1994).](#_bookmark31) 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 debe rescatarse que 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 utilizados: los modelos de suavizamiento exponencial y los modelos ARIMA. Ambos representan enfoques complementarios a un problema, pues según Hyndman (R. J. Hyndman and Athanasopoulos [2018a](#_bookmark41)), 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, para luego evaluar su ajuste sometiéndolo a diferentes pruebas y medidas de rendimiento. Al trabajar con la metodología de Box-Jenkins, uno de los pasos a concretar es identificar 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, actualmente, es común disponer de una gran cantidad de series cronológicas para analizar, lo que dificulta tener suficiente personal capacitado para realizar exitosamente este análisis visual e identificar adecuadamente los modelos. Esto sucede, incluso, con cantidades moderadas de series cronológicas a analizar, por lo que la generación de algoritmos que ayuden a esta identificación se vuelve cada vez más necesaria (Hyndman y Khandakar [2008](#_bookmark43)).  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 y Rissanen [1982),](#_bookmark36) la extensión de dicha propuesta realizada por Gómez (Gómez [1998)](#_bookmark34) y posteriormente aplicada (Gómez and Maraval [1998](#_bookmark35)) en los software **TRAMO** y **SEATS**; de manera similar se planteó una aplicación en los programas **SCA-Expert** (Liu [1989)](#_bookmark46) y **TSE-AX** (Mélard y Pasteels [2000](#_bookmark47)). Otros algoritmos desarrollados en programas de cómputo de paga son **Forecast Pro** (Goodrich [2000)](#_bookmark33) y **Autobox** (Reilly[2000](#_bookmark50)). Uno de los métodos automatizados de estimación es el que ofrece el paquete **forecast** (Hyndman y Khandakar [2008](#_bookmark43)) del lenguaje de programación R y 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 el presente trabajo propone una metodología para la estimación un modelo ARIMA de una serie cronológica determinada abarcando más posibilidades que os enfoques tradicionales, cuya temporalidad sea mensual, bimensual, trimestral o cuatrimestral, mediante un proceso de selección fundamentada en las permutaciones de todos los parámetros de un modelo ARIMA hasta un cierto límite, 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),](#_bookmark42) 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 describa de manera adecuada el comportamiento de una serie cronológica (Hernández [2011c](#_bookmark39)).  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. |
| Fuente(s) de datos (si ya están recolectados) o diseño del estudio (si no están recolectados) | Para iniciar, los datos reales que se propone utilizar pertenecen a la Unidad de Estadísticas demográficas del Instituto Nacional de Estadística y Censos y a la Contraloría General de la República de Costa Rica:   1. Tasa de mortalidad infantil interanual, iniciando desde enero de 1989 hasta diciembre de 2017. 2. Tasa global de fecundidad, iniciando desde enero del 2000 hasta diciembre del 2017. 3. Mortalidad por causa externa, iniciando desde enero del 2000 hasta diciembre del 2017. 4. Incentivos salariales del sector público, iniciando en enero del 2007 hasta junio del 2018. 5. Intereses y comisiones del sector público, iniciando en enero del 2007 hasta junio del 2018. 6. Demanda eléctrica, iniciando en enero del 2011 hasta junio del 2018 |
| Definición de variable(s) de estudio | 1. **Mortalidad infantil interanual**: El cálculo de este indicador consiste en tomar siempre un periodo de 12 meses; es decir, cuando se obtiene la información de los nacimientos y las defunciones infantiles de un determinado mes, se toman los datos de 11 meses hacia atrás para completar un año. Se calcula como el total de defunciones de menores de un año en un mes dado y los 11 precedentes, entre el total de nacimientos en el mismo periodo de tiempo. 2. **Tasa global de fecundidad**: Número de hijos que en promedio tendrían las mujeres al final de su vida reproductiva si durante la misma estuvieran expuestas a las tasas de fecundidad por edad del período de estudio. Se calcula como el número total de nacimientos entre la población femenina de 15 a 49 años. 3. **Mortalidad por causa externa**: Defunciones debidas a suicidios, homicidios y accidentes en general. Muertes que a nivel nacional e internacional corresponden a causas prevenibles y evitables. Se calcula como el total de defunciones por las causas previamente citadas entre la población total. 4. **Incentivos salariales del sector público**: Retribuciones que de conformidad con la legislación vigente se asignan al servidor por sus 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](#_bookmark44)). Por su parte, la Contraloría General de la República La CGR dispone de información completa, pública, proveniente de [fuentes primarias](https://es.wikipedia.org/wiki/Fuente_primaria), actualizada en tiempo, de libre acceso, procesable tanto por humanos como por máquinas ([human-machine readable](https://es.wikipedia.org/wiki/Datos_legibles_por_m%C3%A1quina)), en [formatos abiertos](https://es.wikipedia.org/wiki/Formato_abierto), con licencia libre de uso, permanente y sin costos de utilización.  La validación de los datos se realiza en dos vías: un protocolo de detección semiautomático de inconsistencias el cual consiste en la tabulación de cuadros que sirvan para visualizar los cruces más importantes entre variables y que dejen en evidencia casos extraños, como que ocurran nacimientos en un asilo de ancianos, o verificar la edad de la madre con la cantidad de hijos tenidos. También se realiza una segunda verificación manual para revisar casos que por la forma en que fueron digitados luzcan sospechosos, como un peso al nacer inferior a 500 gramos o números de cédula repetidos en nacimientos que no son múltiples. En casos muy puntuales en los que se disponga de un medio de comunicación con el informante, este es contactado para corroborar datos que luzcan sospechosos o bien para completar información faltante. |
| Modelo(s) estadísticos o técnica(s) estadística(s) de análisis a emplear | **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),](#_bookmark42) 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 describa de manera adecuada el comportamiento de una serie cronológica (Hernández [2011c](#_bookmark39)).  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.  Los modelos ARIMA hacen uso de las variaciones y correlaciones en el tiempo de una serie cronológica para encontrar patrones que sirvan para fundamentar predicciones para valores futuros. Estos modelos necesitan identificar los coeficientes y la cantidad de puntos pasados que se van a utilizar para ajustar el modelo que mejor se adapte a los datos. La forma más común de representar los modelos ARIMA es mediante la expresión ARIMA(p,d,q)(P,D,Q), donde los componentes p,d,q son indican el orden en la parte no estacional, mientras que los P,D,Q se relacionan con la parte estacional.  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 puede expresarse de la siguiente manera:  Donde el término representa ruido blanco. El modelo anterior es muy similar a una regresión lineal múltiple, donde cada coeficiente *φ* va acompañado por su correspondiente rezago . 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:  Donde el término representa nuevamente el ruido blanco. La ecuación anterior representa un modelo de medias móviles de orden *q*, en la cual cada término 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](#_bookmark28)). 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](#_bookmark30)), el cual se define para dos variables *X* e *Y* como sigue:  Este mismo concepto puede aplicarse a las series cronológicas para comparar el valor de esta en el tiempo, con su valor en el tiempo , es decir, se comparan las observaciones consecutivas  *con* . Esto también es aplicable no solo a una observación rezagada (), sino también con múltiples rezagos . Para esto se hace uso del coeficiente de autocorrelación.  El coeficiente de autocorrelación recibe su nombre debido a que se utiliza el coeficiente de correlación para pares de observaciones de la serie cronológica. Al conjunto de todas las autocorrelaciones se le llama función de autocorrelación.  **Función de autocorrelación parcial**  La función de autocorrelación parcial busca medir la asociación lineal entre las observaciones y , descartando los efectos de los rezagos (Hernández [2011d)](#_bookmark40).  **Modelos ARIMA no estacionales**  Como se mencionó anteriormente, los modelos ARIMA aplicados a una serie cronológica *Yt* son una combinación de un modelo autorregresivo, uno de medias móviles y alguna clase de diferenciación (logarítmica, exponenciación) para así obtener una serie diferenciada . Si se juntan ambas se obtiene un modelo ARIMA(p,d,q) que no cubre los efectos estacionales, donde *p* es el orden del modelo autorregresivo, *d* el grado de la diferenciación y *q* es el orden del modelo de medias móviles, cuya estructura se muestra en la ecuación:  **Modelos ARIMA estacionales**  Los modelos ARIMA son también capaces de cubrir los efectos estacionales, es decir, particularidades de la serie cronológica que se repiten periódicamente con una cierta temporalidad (mensual, bimensual, etc.). Para ello se incorporan términos adicionales al modelo relacionados con la parte estacional de una manera similar a como se incorporan en el modelo ARIMA no estacional, pero ahora considerando retrocesos según sea la temporalidad estacional, pasando así de un a un , donde se refieren a la parte estacional y *S* a la temporalidad presente en la serie.  **Medidas de rendimiento**  Cuando se tiene el modelo ARIMA estimado, 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 y Adhikari (Adhikari, K, y Agrawal [2013b)](#_bookmark29) menciona las siguientes:  **MAE**  El error absoluto medio se define como .  **MAPE**  El porcentaje promedio de error absoluto se define como .  **RMSE**  Es la raíz del error cuadrático medio .  **MASE**  Para series no estacionales .  Para series estacionales  Donde m es la temporalidad de la serie.  **AIC**  Se calcula de la siguiente manera:  Donde k es el número de parámetros y n el número de datos.  **AICc**  Se calcula de la siguiente manera:  Donde k es el número de parámetros y n el número de datos.  **BIC**  Se calcula de la siguiente manera:  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 .  Para generar el proceso en cuestión, se deben fijar los valores de en la parte no estacional y 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 , el 2 indica que se pueden fijar los valores de los coeficientes y en, digamos 2 y 46 respectivamente; de forma análoga, pueden definirse los coeficientes en*,* 4*,*1 y 3 respectivamente.  Para esta simulación, se trabajarán dos tipos de series escenarios: estacionales y no estacionales. El valor máximo del proceso ARIMA que se generará es un .  **Selección de ARIMA vía sobreparametrización**  A partir de las series cronológicas simuladas en la sección anterior, se prueba el método propuesto, el cual consiste en una 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 .  Tras ejecutar el método con los escenarios de simulación mencionados, se validará el mismo con las seis series cronológicas reales previamente mencionadas. |
| Software estadístico a utilizar | Se utilizará el lenguaje de programación R (R Core Team [2019a)](#_bookmark48) mediante su interfaz RStudio para todos los procesos relacionados con la estimación y la simulación. Se hará uso de funciones contenidas en los paquetes tidyr (Wickham and Henry [2019](#_bookmark53)), dplyr (Wickham et al. [2019)](#_bookmark52) y parallel (R Core Team [2019b](#_bookmark49)). |
| Referencias bibliográficas | | Adhikari, R., A. R. K, and R. K. Agrawal. 2013a. *An Introductory Study on Time Series Modeling and Forecasting*. Lap Lambert Academic Publishing GmbH KG.h[ttps://arxiv.org/ftp/arxiv/papers/1302/](https://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf) [1302.6613.pdf](https://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf).  ———. 2013b. *An Introductory Study on Time Series Modeling and Forecasting*. Lap Lambert Academic Publishing GmbH KG.h[ttps://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf.](https://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf)  Benesty, J., and Y.and Cohen Chen J.and Huang. 2009. “Pearson Correlation Coefficient.” In *Noise Reduction in Speech Processing*, 37–38. Berlin, Heidelberg: Springer Berlin Heidelberg.h[ttps://doi.org/](https://doi.org/10.1007/978-3-642-00296-0_5) [10.1007/978-3-642-00296-0\_5](https://doi.org/10.1007/978-3-642-00296-0_5).  Box, G. E.P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*. Forecasting and Control Series. Prentice Hall.h[ttps://books.google.co.cr/books?id=sRzvAAAAMAAJ.](https://books.google.co.cr/books?id=sRzvAAAAMAAJ)  Brown, Robert G. 1956. *Exponential Smoothing for Predicting Demand*. A.D.Little.h[ttps://www.](https://www.industrydocuments.ucsf.edu/docs/jzlc0130) [industrydocuments.ucsf.edu/docs/jzlc0130](https://www.industrydocuments.ucsf.edu/docs/jzlc0130).  Goodrich, RL. 2000. “The Forecast Pro Methodology.” *International Journal of Forecasting* 16 (4): 533– 35.h[ttp://www.forecasting-competition.com/downloads/NN3/methods/Goodrich%20(2000)%20The%](http://www.forecasting-competition.com/downloads/NN3/methods/Goodrich%20(2000)%20The%20Forecast%20Pro%20methodology%20science.pdf) [20Forecast%20Pro%20methodology%20science.pdf](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.h[ttp://www.jstor.org/stable/2335856.](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-](http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html)series-cronologicas.[html](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-](http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html)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-](http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html)cronologicas.[html](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-](http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/1985-introduccion-a-las-series-cronologicas.html)cronologicas.[html](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.h[ttps://doi.org/10.18637/jss.v027.i03.](https://doi.org/10.18637/jss.v027.i03)  INEC. 2017. “Población, Nacimientos, Defunciones Y Matrimonios.”h[ttp://inec.cr/sites/default/files/documetos-biblioteca-virtual/repoblacev2017\_0.pdf](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.h[ttps://books.google.co.cr/books?id=8r0qE35wt44C.](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.h[ttps://doi.org/10.1080/03610928908830035.](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.  R Core Team. 2019a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.h[ttps://www.R-project.org/.](https://www.R-project.org/)  ———. 2019b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.h[ttps://www.R-project.org/.](https://www.R-project.org/)  Reilly, D. 2000. “The Autobox System.” *International Journal of Forecasting* 16 (4): 531–33.h[ttps://](https://ideas.repec.org/a/eee/intfor/v16y2000i4p531-533.html) [ideas.repec.org/a/eee/intfor/v16y2000i4p531-533.html](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.h[ttp://srv-website.cloudapp.net/documents/10179/999061/Nota+t%C3%A9cnica+tablas+](http://srv-website.cloudapp.net/documents/10179/999061/Nota%2Bt%C3%A9cnica%2Btablas%2Bde%2Bvida%2Bsegunda%2Bparte) [de+vida+segunda+parte](http://srv-website.cloudapp.net/documents/10179/999061/Nota%2Bt%C3%A9cnica%2Btablas%2Bde%2Bvida%2Bsegunda%2Bparte).  Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2019. *Dplyr: A Grammar of Data Manipulation*.h[ttps://CRAN.R-project.org/package=dplyr.](https://CRAN.R-project.org/package%3Ddplyr)  Wickham, Hadley, and Lionel Henry. 2019. *Tidyr: Tidy Messy Data*.h[ttps://CRAN.R-project.org/](https://CRAN.R-project.org/package%3Dtidyr) [package=tidyr](https://CRAN.R-project.org/package%3Dtidyr). |
| Cronograma | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Actividad** | **Duración** | **Periodo** | | **Descripción** | |  |  | **Inicio** | **Fin** |  | | Revisión bibliográfica | 2 | Ene-20 | Feb-20 | Revisión de fuentes bibliográficas para elaborar el marco teórico | | Elaboración de la introducción y la justificación | 3 | Mar-20 | May-20 | Redacción de la introducción y la justificación del tema de investigación | | Elaboración del marco teórico | 3 | Jun-20 | Ago-20 | Desarrollo de la metodología empleada | | Estudio de simulación | 2 | Set-20 | Oct-20 | Aplicación del método en escenarios de simulación | | 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**