

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

LA SOBREPARAMETRIZACIÓN EN EL ARIMA: UNA APLICACIÓN A
DATOS COSTARRICENSES

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

2021

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

DEDICATORIA	I
AGRADECIMIENTOS	II
RESUMEN	1
ABSTRACT	2
1 INTRODUCCIÓN	3
1.1 Antecedentes	3
1.2 El problema	4
1.3 Objetivos del estudio	4
1.4 Justificación del estudio	5
1.5 Organización del estudio	6
2 MARCO TEÓRICO	7
2.1 Componentes de una serie cronológica	7
2.1.1 La tendencia-ciclo	8
2.1.2 Componentes estacionales	10
2.1.3 Componente irregular	11
2.2 Supuestos en el análisis de series cronológicas	11
2.3 Modelos Autorregresivos Integrados de Medias Móviles	12
2.3.1 Ecuación de Wold	13
2.3.2 Modelos Autorregresivos	14
2.3.3 Modelos de Medias Móviles	14
2.3.4 Metodología Box-Jenkins	15
2.3.5 Modelos ARIMA	15
2.4 Identificación del modelo	16
2.5 Los autocorrelogramas	17
2.6 La sobreparametrización y el análisis combinatorio	18
3 METODOLOGÍA	19
3.1 Medidas de bondad de ajuste y rendimiento	19
3.1.1 AIC	19
3.1.2 AICc	20
3.1.3 BIC	20
3.1.4 MAE	20
3.1.5 MASE	20

3.1.6	RMSE	20
3.2	La sobreparametrización	21
4	RESULTADOS	24
4.1	Introducción	24
4.2	Análisis de simulación en series cronológicas	24
4.3	Estimaciones en series cronológicas costarricenses reales	26
4.3.1	Tasa de mortalidad infantil interanual	26
4.3.2	Mortalidad por causa externa	29
4.3.3	Incentivos salariales del sector público	33
4.3.4	Intereses y comisiones del sector público	36
5	CONCLUSIONES	39
6	ANEXOS	41
6.1	Función de sobreparametrización	41
6.2	Función de simulación de series cronológicas	46
7	REFERENCIAS	48

Índice de cuadros

1	Medidas de rendimiento según método de estimación	26
2	Medidas de rendimiento según método de estimación para la TMII	30
3	Medidas de rendimiento según método de estimación para la Mortalidad por causa externa	32
4	Medidas de rendimiento según método de estimación para los incentivos salariales	36
5	Medidas de rendimiento según método de estimación para los intereses y comisiones del sector público	39

Índice de figuras

1	Número de matrimonios en Costa Rica para el periodo 1978-1983	8
2	Número de turistas en Costa Rica para el periodo 1991-2000	9
3	Tendencia del número de matrimonios en Costa Rica para el periodo 1978-1983	9
4	Índice bursatil NASDAQ-100 para el periodo enero 1990 - junio 2021	10
5	Valores de referencia para la simulación de series cronológicas	25
6	Series cronológicas simuladas	25
7	Tasa de Mortalidad Infantil Interanual 1989 - 2017	28
8	Tasa de Mortalidad Infantil Interanual 1989 - 2017 según periodos	29

9	Descomposición de la TMII en el periodo 2000 - 2017	29
10	Pronósticos de la TMII según método de estimación	30
11	Mortalidad por causa externa 2000 - 2017	31
12	Mortalidad por causa externa 2000 - 2017 según mes	32
13	Descomposición de las defunciones por causa externa en el periodo 2000-2017	33
14	Pronósticos de la TMII según método de estimación	33
15	Incentivos salariales en el sector público 2007 - 2018	34
16	Incentivos salariales en el sector público 2007 - 2018 según mes	35
17	Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018	35
18	Pronósticos de los incentivos salariales según método de estimación	36
19	Intereses y comisiones del sector público en el periodo 2007-2018	37
20	Intereses y comisiones del sector público en el periodo 2007-2018 según mes	38
21	Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018	38
22	Pronósticos de los intereses y comisiones del sector público según método de estimación	39

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 Mundial¹ 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. Kaggle², 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 temática³.

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. Es importante resaltar que las técnicas de proyección ARIMA tienen como objetivo explicar las relaciones pasadas de la serie cronológica, para de esta manera conocer el posible comportamiento futuro de la misma (R. J. Hyndman & Athanasopoulos, 2018a).

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 suele utilizar la identificación de parámetros mediante autocorrelogramas parciales

¹<https://databank.worldbank.org/home.aspx>

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

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

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, y la aproximación mediante los autocorrelogramas no es una forma exacta de aproximar el proceso que gobierna la serie. Por lo 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. Una alternativa al problema de aproximar los parámetros del proceso que gobierna la serie cronológica puede ser la sobreparametrización.

1.2 El problema

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 R⁴ (R. 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), seleccionando 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 de descartar ciertos modelos, y mantener otros con un criterio más científico y una evidencia numérica que respalde 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

⁴Descarga gratuita en <https://cran.r-project.org/>

ecuación del ARIMA.

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 en distintos campos de investigación: el objetivo es poder pronosticar de forma correcta lo que sucederá dentro de los próximos períodos. 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 las series temporales: 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()` y un modelo estándar como el $ARIMA(1, 1, 1)(1, 1, 1)_s$.

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

2 MARCO TEÓRICO

Los modelos de series cronológicas han sido un importante tema de investigación durante décadas ([De Gooijer & Hyndman, 2006](#)). 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 periódica.

Sin embargo, encontrar un modelo que presente un buen comportamiento con respecto a los datos no es sencillo, pues deben considerarse diversos aspectos teóricos, prácticos, y de la temática de estudio para así obtener un modelo adecuado que logre generar pronósticos realistas y pertinentes para la toma de decisiones ([Rezaee, Aliabadi, Dorestani, & Rezaee, 2020](#)).

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.

El presente capítulo consta de seis apartados. El primer apartado abarca los cuatro componentes de una serie cronológica. Posteriormente, la segunda sección repasa los supuestos fundamentales en el análisis de series cronológicas. Con los elementos más básicos introducidos, el tercer apartado cubre el eje central de esta investigación: Los modelos Autorregresivos Integrados de Medias Móviles y sus componentes, los modelos autorregresivos y los modelos de medias móviles, así como la metodología Box-Jenkins y el proceso para la identificación de los modelos. En el cuarto apartado se introducen los métodos para la identificación de los modelos. El quinto apartado abarca los componentes relacionados a los autocorrelogramas, la forma más difundida para la selección de modelos y, finalmente, el sexto apartado introduce el principal aporte de este estudio, la sobreparametrización como método para la selección de modelos.

2.1 Componentes de una serie cronológica

En el análisis de series cronológicas existen dos grandes corrientes de estudio: Los componentes inherentes a la serie cronológica y el estudio de las autocorrelaciones. Según el primer enfoque, de acuerdo con [Hernández \(2011a\)](#), las series cronológicas poseen tres componentes principales: Tendencia-ciclos, Estacionalidad e Irregularidad. Considerando estos tres elementos, las series cronológicas pueden ser *aditivas*, como se muestra en la ecuación 1, en cuyo caso se asume que los tres componentes son independientes entre sí; o *multiplicativa*, donde, por el contrario, los tres

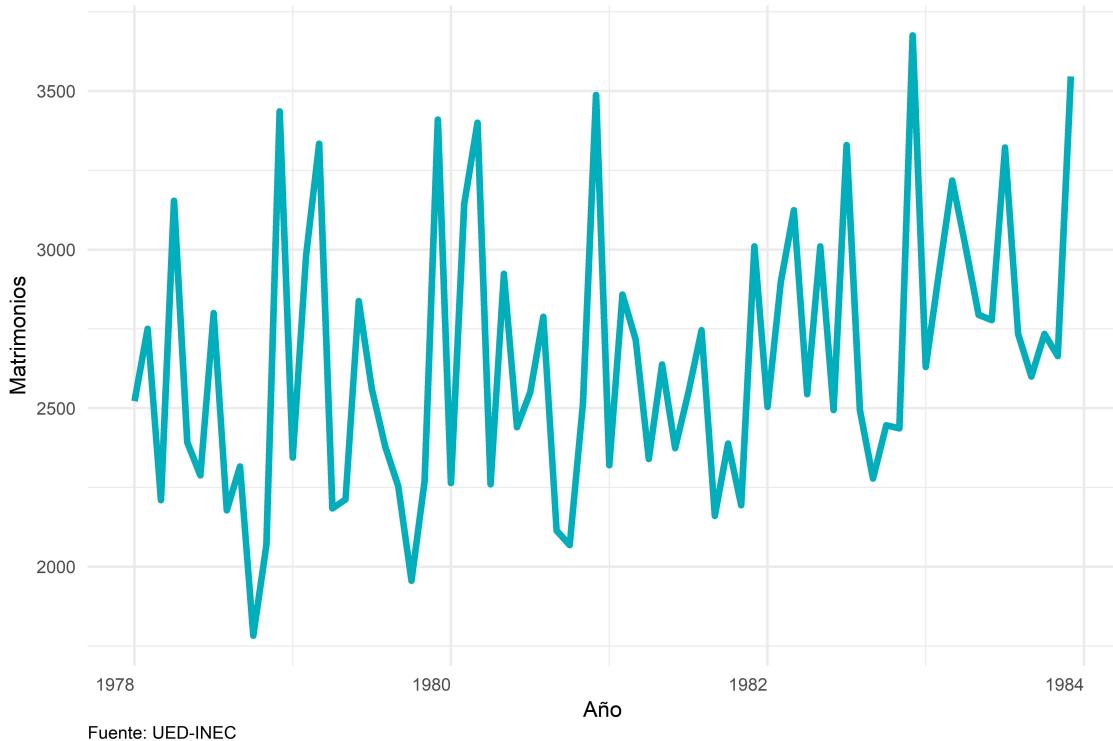
componentes no son independientes, como muestra la ecuación 2.

$$Y(t) = T(t) + S(t) + I(t) \quad (1)$$

$$Y(t) = T(t) \times S(t) \times I(t) \quad (2)$$

Donde Y es la serie cronológica, T es la tendencia-ciclo, S es la parte estacional, I la parte irregular o aleatoria, y t es el momento en el tiempo. Esta perspectiva clásica del análisis de series de tiempo permite realizar un análisis descriptivo del comportamiento de la serie en cuestión; cada una de sus partes se definen en posteriores apartados. De manera visual, una serie cronológica aditiva posee un comportamiento similar al mostrado en la figura 1, mientras que un comportamiento multiplicativo puede apreciarse en la figura 2.

Figura 1: Número de matrimonios en Costa Rica para el periodo 1978-1983



2.1.1 La tendencia-ciclo

A partir del texto de Calderón (2012), la tendencia general de una serie cronológica se refiere al crecimiento, decrecimiento o lateralización de sus movimientos a lo largo del periodo de estudio. Un ejemplo es la serie cronológica del número de matrimonios en Costa Rica para el periodo 1978-1983, que con el tiempo su crecimiento suele comportarse de una forma creciente tal y como muestra la figura 3.

Figura 2: Número de turistas en Costa Rica para el periodo 1991-2000

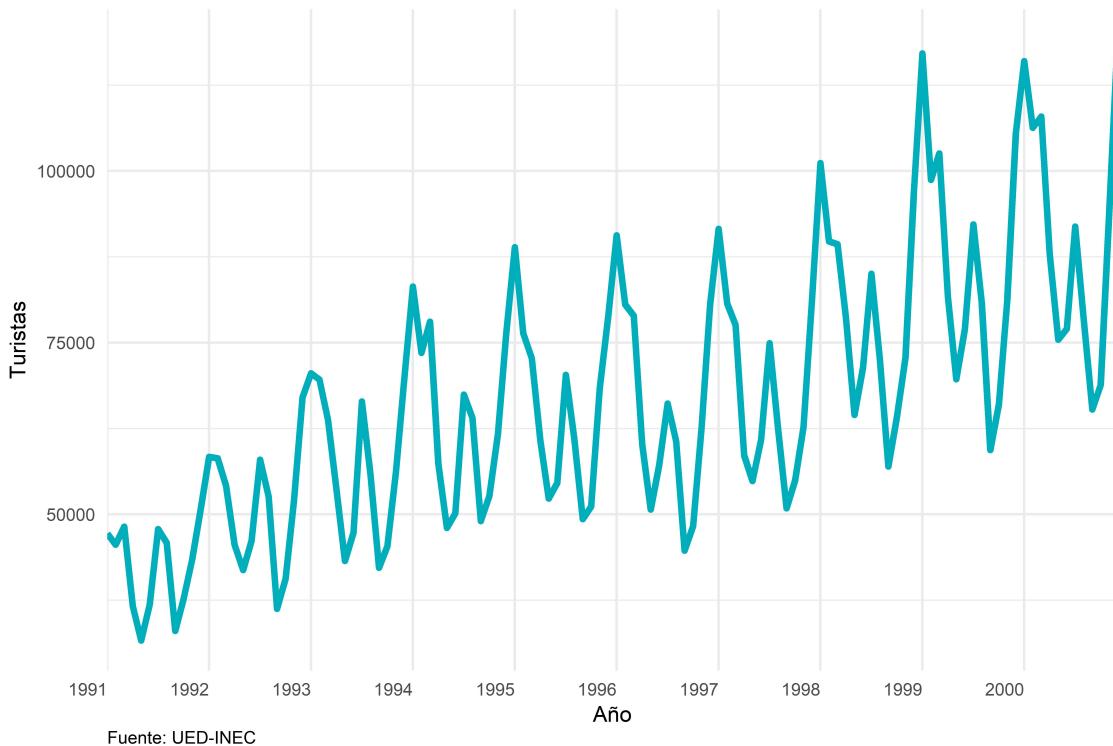
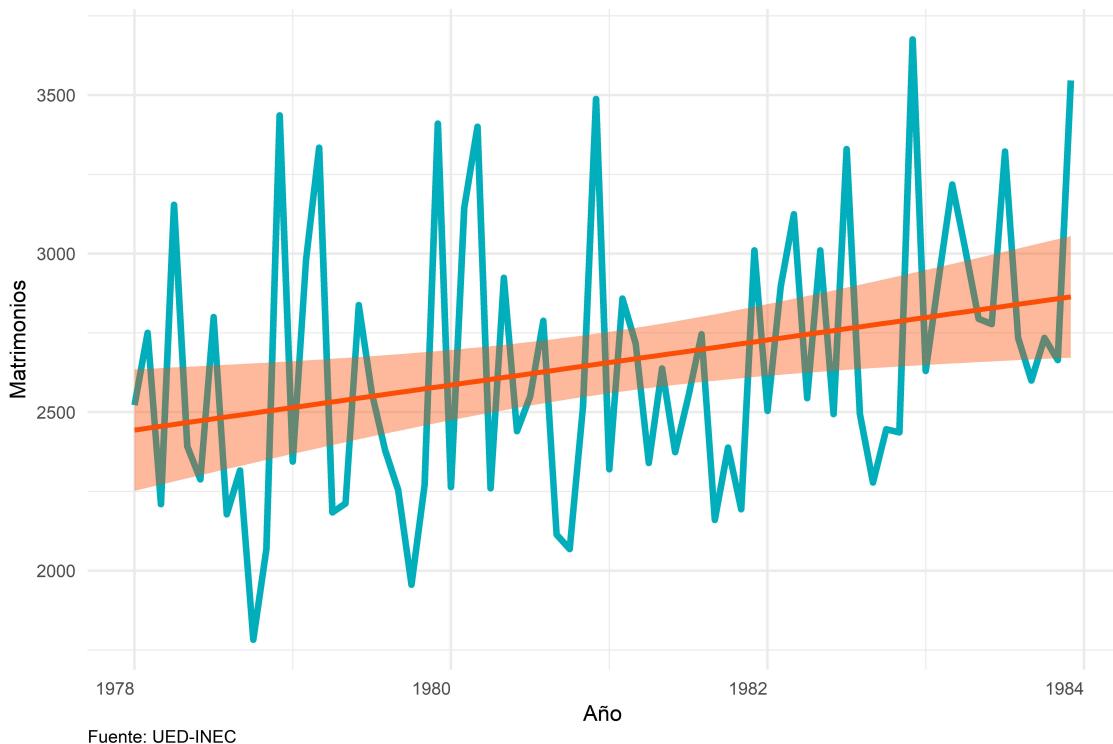


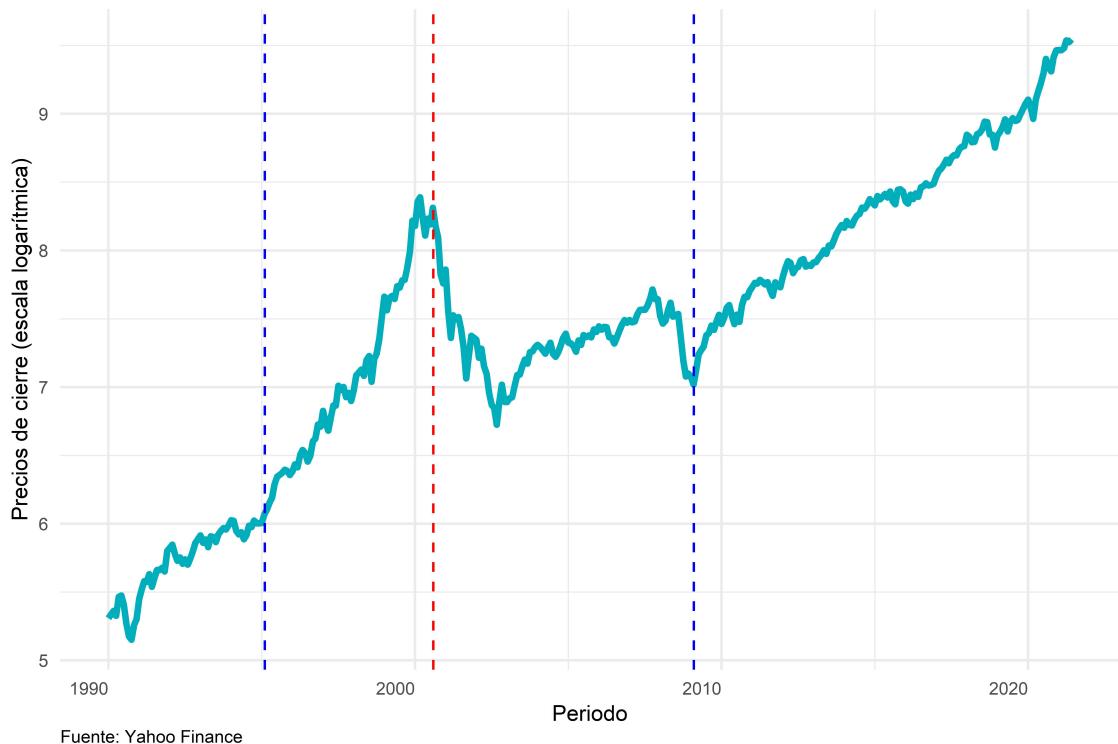
Figura 3: Tendencia del número de matrimonios en Costa Rica para el periodo 1978-1983



Del informe elaborado también por Calderón (2012) se desprende que los períodos cíclicos, por su parte, se refieren a los cambios que se dan en una serie cronológica en el mediano-largo 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 NASDAQ-100. Este indicador resume el estado los 100 valores de las compañías más importantes del sector de la industria de la tecnología, 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. La figura 4 muestra como el índice NASDAQ-100 inicia un auge alrededor de enero de 1995 (primera línea azul punteada), para luego experimentar una fuerte caída a partir de junio del año 2000 (línea roja punteada) y posteriormente iniciar un periodo de recuperación en enero del año 2009 (segunda línea azul punteada).

Figura 4: Índice bursatil NASDAQ-100 para el periodo enero 1990 - junio 2021



La descomposición clásica de la tendencia-ciclo que este componente se mantiene constante de un periodo al siguiente y se obtiene a partir de una media móvil de m periodos (\bar{y}_m). De esta manera la forma matemática de la tendencia-ciclo para una serie cronológica se muestra en la ecuación 3.

$$T(t) = \begin{cases} 2\bar{y}_m, & \text{si } m \text{ es par} \\ \bar{y}_m, & \text{si } m \text{ es impar} \end{cases} \quad (3)$$

2.1.2 Componentes estacionales

Calderón (2012) también se refiere a los cambios estacionales que se presentan en una serie de tiempo, los cuales se relacionan con las fluctuaciones naturales del fenómeno dentro de una temporada de observaciones. Visualmente los efectos estacionales puede apreciarse en la figura 2, en

donde los picos más altos de turistas siempre se ubican entre los meses de diciembre y enero. Matemáticamente, el componente estacional puede definirse como se indica en la ecuación 4:

$$S(t) = \bar{y}_{s_t} - \bar{y}_k; \quad \begin{cases} \bar{y}_{s_t} = \frac{\sum \bar{y}_s}{n} \\ \bar{y}_k = \frac{\sum y_k}{n-m} \\ \bar{y}_s = \sum_{j=1}^s y_{k_j} \\ y_k = y_t - \bar{y}_{mct} \\ \bar{y}_{mct} = \frac{\bar{y}_{m_t} + \bar{y}_{m_{t-1}}}{2} \\ \bar{y}_{m_t} = \frac{\sum_{t=1}^m y_t}{m} \end{cases} \quad (4)$$

donde m representa la cantidad de periodos y s la frecuencia estacional.

2.1.3 Componente irregular

Finalmente, la irregularidad de una serie cronológica, siguiendo a [Calderón \(2012\)](#), 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. Matemáticamente su descomposición se obtiene a partir de los otros componentes así como de la propia serie cronológica $y(t)$, tal y como se muestra en la ecuación 5.

$$I(t) = \begin{cases} y(t) - T(t) - S(t), & \text{si la serie es aditiva} \\ \frac{y(t)}{T(t)S(t)}, & \text{si la serie es multiplicativa} \end{cases} \quad (5)$$

2.2 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 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 estacionariedad. De acuerdo con [Agrawal & Adhikari \(2013\)](#), 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.

2.3 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.

2.3.1 Ecuación de Wold

Según [Sargent \(1979\)](#), cualquier proceso estacionario puede ser representado mediante la ecuación 6:

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t \quad (6)$$

donde $\forall \psi_j \in \mathbb{R}, \psi_0 = 1, \sum_{j=0}^{\infty} \psi_j^2 < \infty$, y ε_t representa un ruido blanco i.i.d., es decir, $\varepsilon_t \sim N(0, \sigma^2)$; además, κ_t es el componente lineal determinístico de forma tal que $cov(\kappa_t, \varepsilon_{t-j}) = 0$, lo cual implica que este componente determinístico es independiente de la suma infinita de los choques pasados.

De lo anterior, si se omite la parte determinística κ_t de 6, el remanente es la suma ponderada infinita, lo cual implica que si se conocen los ponderadores ψ_j , y si además se conoce σ_ε^2 , es posible obtener una representación para cualquier proceso estacionario; este concepto es conocido como *media móvil infinita*.

Sabiendo que $\varepsilon_t \sim N(0, \sigma^2)$, se tiene que ε_t tiene media 0, es decir, está centrado en este valor. De esta manera el ruido blanco es por definición un proceso centrado, lo cual implica que la suma ponderada infinita está centrada en sí misma. De esta manera, la representación de Wold de un proceso x_t supone que se suman los choques pasados más un componente determinístico que no es otro que el valor esperado del proceso: $\kappa_t = m$, donde m es una constante cualquiera. Así, la ecuación 6 puede sustuirse por:

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + m \quad (7)$$

y de 7 puede verificarse que,

$$E(x_t) = E \left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + m \right) = \sum_{j=0}^{\infty} \psi_j E(\varepsilon_{t-j}) + m = m \quad (8)$$

La principal consecuencia del teorema de Wold es que, si se conocen los ponderadores ψ_j , y además σ_ε^2 es ruido blanco es posible conocer el proceso por medio del cual se rige la serie cronológica. Esto permite realizar cualquier previsión, denotada por \hat{X}_{T+h} para el proceso de interés x_T en el momento $T + h$ para una muestra cualquiera de T observaciones de x_t . De acuerdo con [Sargent \(1979\)](#), basado en el teorema de Wold, la mejor previsión posible para un proceso x_t para el momento $T + h$, denotado por \hat{x}_{T+h} , la predicción está dada por:

$$\hat{x}_{T+h} = \sum_{j=1}^{\infty} \psi_j \varepsilon_{T-j+1} \quad (9)$$

De la ecuación 9 se desprende que el error de previsión asociado está dado por:

$$x_{T+h} - \hat{x}_{T+h} = \sum_{j=1}^{\infty} \psi_j \varepsilon_{T-h+1} \quad (10)$$

2.3.2 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 11.

$$y_t = c + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t \quad (11)$$

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 $i = 1, 2, \dots, p$ 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](#)).

2.3.3 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 12.

$$y_t = \mu + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (12)$$

Donde μ representa el valor medio de la serie cronológica y cada valor de $\theta_j (j = 1, 2, \dots, q)$ 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 aproximadamente 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 ([Agrawal & Adhikari, 2013](#)).

2.3.4 Metodología Box-Jenkins

La combinación de un $AR(p)$ y un $MA(q)$, descritos en las ecuaciones 11 y 12 respectivamente, como se mencionó al inicio de esta sección, generan los modelos autorregresivos de medias móviles, $ARMA(p, q)$, representados mediante la ecuación 13.

$$y_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (13)$$

Cochrane (1997) menciona que los modelos $ARMA(p, q)$ suelen manipularse mediante lo que se conoce como operador de rezagos, denotado como $L y_t = y_{t-1}$. 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 cuyo 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, Jenkins, & Reinsel, 1994).

2.3.5 Modelos ARIMA

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 14.

$$\varphi(L)(1 - L)^d y_t = \theta(L)\varepsilon_t \left(1 - \sum_{i=1}^p \varphi_i L^i\right) (1 - L)^d y_t = \left(1 + \sum_{j=1}^q \theta_j L^j\right) \varepsilon_t \quad (14)$$

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, Jenkins, & Reinsel (1994), lo anterior puede generalizarse aún más al considerar los efectos estacionales de la serie cronológica. Si se considera una serie cronológica con observacio-

nes mensuales, una diferenciación de primer orden es igual a la diferencia entre una observación y la observación correspondiente 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, Jenkins, & Reinsel \(1994\)](#) 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 aadecue 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 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 $ARIMA$ ([R. J. Hyndman & Athanasopoulos, 2018b](#)).

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 15.

$$r_{X,Y} = \frac{E(XY)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (15)$$

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 ($Y_{t-2}, (Y_{t-3}), \dots, (Y_{t-n})$). 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 parcial⁵, como menciona Hernández (2011b), busca medir la asociación lineal entre las observaciones Y_t y Y_{t-k} , descartando los efectos de los rezagos $1, 2, \dots, k-1$.

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 go-

⁵ *PACF* por sus siglas en inglés

bierna 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 que a su vez elimine el criterio subjetivo del observador, 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

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 la serie cronológica. 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.

Esta sección aborda la metodología propuesta como método de estimación y pronóstico de series cronológicas. En la búsqueda de un modelo adecuado de entre varios candidatos, se cubren en un primer apartado las medidas de bondad de ajuste y de precisión a utilizar.

El segundo apartado describe en detalle el uso de la sobreparametrización como herramienta para la generación de pronósticos de series cronológicas con temporalidades mensuales, bimensuales, trimestrales, cuatrimestrales o anuales mediante un proceso de selección fundamentada en las permutaciones de todos los parámetros de un modelo ARIMA hasta un rango determinado. Las medidas de precisión y de bondad de ajuste 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 un 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.

3.1 Medidas de bondad de ajuste y rendimiento

El objetivo último al estimar un modelo ARIMA es obtener los pronósticos de dicho modelo Sin embargo, estos pronósticos no son pueden asumirse como correctos, sino que se debe evaluar su calidad con las llamadas medidas de bondad de ajuste y de rendimiento. Existen múltiples medidas, [Adhikari, K, & Agrawal \(2013\)](#) menciona, entre otras, las siguientes:

3.1.1 AIC

Se calcula de la siguiente manera:

$$AIC = -2\log L(\hat{\theta}) + 2k \quad (16)$$

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

3.1.2 AICc

Su forma de cálculo se muestra en la ecuación 17

$$AICc = -2\log L(\hat{\theta}) + 2k + \frac{2k+1}{n-k-1} \quad (17)$$

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

3.1.3 BIC

El último estadístico de bondad de ajuste se calcula como se muestran en la ecuación 18.

$$BIC = -2\log L(\hat{\theta}) + k \cdot \log(n) \quad (18)$$

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

3.1.4 MAE

El error absoluto medio se define mediante la ecuación 19

$$\frac{1}{n} \sum_{t=1}^n |e_t| \quad (19)$$

3.1.5 MASE

Esta medida de rendimiento tiene dos casos, uno para series cronológicas no estacionales y otro para series cronológicas estacionales, como se muestra en las ecuaciones 20 y 21.

$$\frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-1} \sum_{t=2}^T |Y_t - Y_{t-1}|} \quad (20)$$

$$\frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-m} \sum_{t=m+1}^T |Y_t - Y_{t-m}|} \quad (21)$$

Donde m es la temporalidad de la serie.

3.1.6 RMSE

Es la raíz del error cuadrático medio, como se define en la ecuación 22.

$$\sqrt{\frac{1}{n} \sum_{t=1}^n |e_t^2|} \quad (22)$$

3.2 La sobreparametrización

La estimación de los modelos y posterior selección de los mismos vía sobreparametrización es un proceso que requiere de distintas etapas. El procedimiento completo fue programado utilizando el lenguaje R⁶ y su código se muestra en el Código 1, la cuál fue construida haciendo uso de los paquetes de R `tidyR` (Wickham & Henry, 2019), `dplyr` (Wickham, François, Henry, & Müller, 2019) y `parallel` (R Core Team, 2019), los procesos internos de esta función son descritos a continuación.

A partir de una serie cronológica y_t , se realiza una partición de los datos para tener dos conjuntos distintos. Uno de ellos servirá para entrenar y estimar los distintos modelos; mientras que el segundo servirá para validar los pronósticos y posteriormente seleccionar el modelo más adecuado. De manera predeterminada, se utiliza una partición del 80% de los datos para el conjunto de entrenamiento y un 20% para los datos de validación, sin embargo, esto puede cambiar de acuerdo al interés propio del investigador(a).

Una vez que se define la partición que tendrá la serie cronológica, se prosigue con la selección de los escenarios para estimar los modelos de ARIMA. Es en esta instancia en donde se decide en valor máximo de los parámetros p, d, q, P, D, Q del modelo $ARIMA(p, d, q)(P, D, Q)_s$ que serán sujetos al análisis. Si se cuenta con una serie sin patrones estacionales y cuyo modelo con mayor cantidad de parámetros es un $ARIMA(3, 1, 4)$, la matriz de valores paramétricos es la que se muestra en 23:

$$\begin{array}{c} p, d, q \quad P, D, Q \\ \hline \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 3 & 1 & 4 & 0 & 0 & 0 \end{bmatrix} \end{array} \quad (23)$$

De manera análoga, al trabajar con un modelo con algún efecto estacional en una determinada periodicidad, como por ejemplo mensual, la matriz de valores paramétricos al definir el modelo con mayor número de parámetros como un $ARIMA(6, 1, 6)(6, 1, 6)_{12}$ es la mostrada en 24:

⁶<https://cran.r-project.org/>

$$\begin{array}{cc}
 \overbrace{p, d, q} & \overbrace{P, D, Q} \\
 \left[\begin{array}{ccc|ccc}
 0 & 0 & 1 & 0 & 0 & 1 \\
 0 & 0 & 1 & 0 & 0 & 2 \\
 0 & 0 & 1 & 0 & 0 & 3 \\
 0 & 0 & 1 & 0 & 0 & 4 \\
 0 & 1 & 1 & 0 & 0 & 5 \\
 0 & 1 & 1 & 0 & 0 & 6 \\
 0 & 1 & 1 & 0 & 1 & 0 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 6 & 1 & 6 & 6 & 1 & 6
 \end{array} \right] & (24)
 \end{array}$$

Con la matriz de valores paramétricos, como las mostradas en 23 y 24, se estiman los modelos en orden descendente, del modelo con menos parámetros al que tiene más parámetros. Al estimar un nuevo modelo, se evalúa mediante una prueba t ([Stoffer, 2020](#)) para verificar el nuevo término incorporado al modelo es estadísticamente distinto de cero, es decir, el nuevo parámetro está generando un impacto en el modelo. AL tratarse de un proceso iterativo, los cálculos son pueden volverse computacionalmente pesado, es por esta razón que la programación del proceso fue habilitada para realizar procesamiento paralelo y de esta manera reducir el consumo de tiempo en la obtención de resultados.

Cuando se han realizado las pruebas de significancia estadística a los modelos, son calculadas las medidas de bondad de ajuste y de rendimiento mencionadas con anterioridad. Tras esto, se aplica un método de concenso para seleccionar el modelo más adecuado. Este criterio consiste en darle mayor o menor ponderación a los resultados obtenidos con el conjunto de datos de entrenamiento y el de validación; de forma predeterminada se le da una ponderación de 0.8 a los resultados de validación y un 0.2 a los de entrenamiento, esto porque en la práctica, los datos de validación son considerados como datos más recientes y que, mientras más cercanos sean los pronósticos a estos datos, mejores resultados ofrece el modelo. El método de concenso es utilizado para obtener un puntaje de cada modelo ARIMA, su cálculo se obtiene de la ecuación 25:

$$\min \left(\sum_i m_i \cdot w_j \right) \quad (25)$$

Donde m_i representa cada una de las medidas de rendimiento y w_j es el valor de ponderación de los conjunto de entrenamiento y validación mencionados anteriormente. El valor más bajo de todos los modelos es el que se define como el modelo más adecuado.

Como parte de esta investigación, es necesario validar la estimación de modelos ARIMA mediante sobreparametrización no solo con datos reales, sino también mediante datos simulados. Para ello es necesario obtener series cronológicas que son gobernadas por un proceso determinado.

Con este fin, se programó un función haciendo uso del lenguaje R que toma una serie de valores, los

cuales pueden ser reales o simulados. Además, se especifican los valores p, d, q, P, D, Q del modelo ARIMA a partir del cual se desea obtener los datos, así como los valores de los coeficientes presentes en el modelo. Es partiendo de este modelo que se simulan los datos de las series cronológicas que serán insumo para la prueba la selección mediante sobreparametrización, este procedimiento se encuentra en el [Código 2](#).

4 RESULTADOS

4.1 Introducción

La metodología propuesta será puesta a prueba en una primera etapa con series cronológicas simuladas a partir de distintos modelos. Los resultados obtenidos al utilizar la sobreparametrización serán contrastados con otros dos métodos: La función `auto.arima()` de R y un modelo ARIMA estándar, que se trata de un $ARIMA(1, 1, 1)$ en el caso de series no estacionales, y un $ARIMA(1, 1, 1)(1, 1, 1)_{12}$ sobre los datos simulados de series mensuales.

De forma similar, en la segunda parte de este capítulo se comparan los resultados obtenidos mediante la sobreparametrización con los obtenidos utilizando la función `auto.arima()` de R, aplicado a series cronológicas costarricenses.

Los paquetes utilizados para la obtención de estos resultados, aparte de los ya mencionados, son `knitr` (Xie, 2014), `kableExtra` (Zhu, 2021), `readxl` (Wickham & Bryan, 2019), `gridExtra` (Auguie, 2017), `ggpubr` (Kassambara, 2020), `ggplot2` (Wickham, 2016), `lubridate` (Grolemund & Wickham, 2011), `ggseas` (Ellis, 2018), `ggpmisc` (Aphalo, 2021) y `forecast` (Rob J. Hyndman & Khandakar, 2008).

4.2 Análisis de simulación en series cronológicas

Como se mencionó en el capítulo anterior, se utiliza la función Código 2 para generar las distintas series cronológicas que evaluarán el método. El primer insumo de esta función consiste en una muestra de valores aleatorios; en este caso se seleccionaron 100 valores de una distribución $N(10, 1)$ para generar series de 200 observaciones. Los valores paramétricos de los modelos ARIMA fueron seleccionados de manera aleatoria. Los valores iniciales para obtener las series cronológicas simuladas se resumen en la figura 5; donde las regiones azules oscuro representan la densidad de datos entre los percentiles 25 y 75, las líneas punteadas de color naranja marcan la cantidad de desviaciones estándar que los datos se alejan del promedio, y las líneas punteadas de color azul marcan los puntos de corte mínimo, percentiles 25, 50 y 75, y el máximo.

Por su lado, se simularon un total de ocho series cronológicas, cuyo comportamiento y proceso gobernante se muestran en la figura 6.

El cuadro 1 muestra los valores del MAE, MASE y RMSE para cada proceso estimado y, además, muestra el proceso de origen de las series cronológicas simuladas. Al aplicar la función `auto.arima()` y la sobreparametrización sobre una serie cronológica generada a partir de un $ARIMA(1, 0, 0)$, se obtienen los mismos resultados, y esto son ligeramente inferiores a los obtenidos mediante un $ARIMA(1, 1, 1)$. Al generar datos a partir de un $ARIMA(1, 0, 1)$ los mejores

Figura 5: Valores de referencia para la simulación de series cronológicas

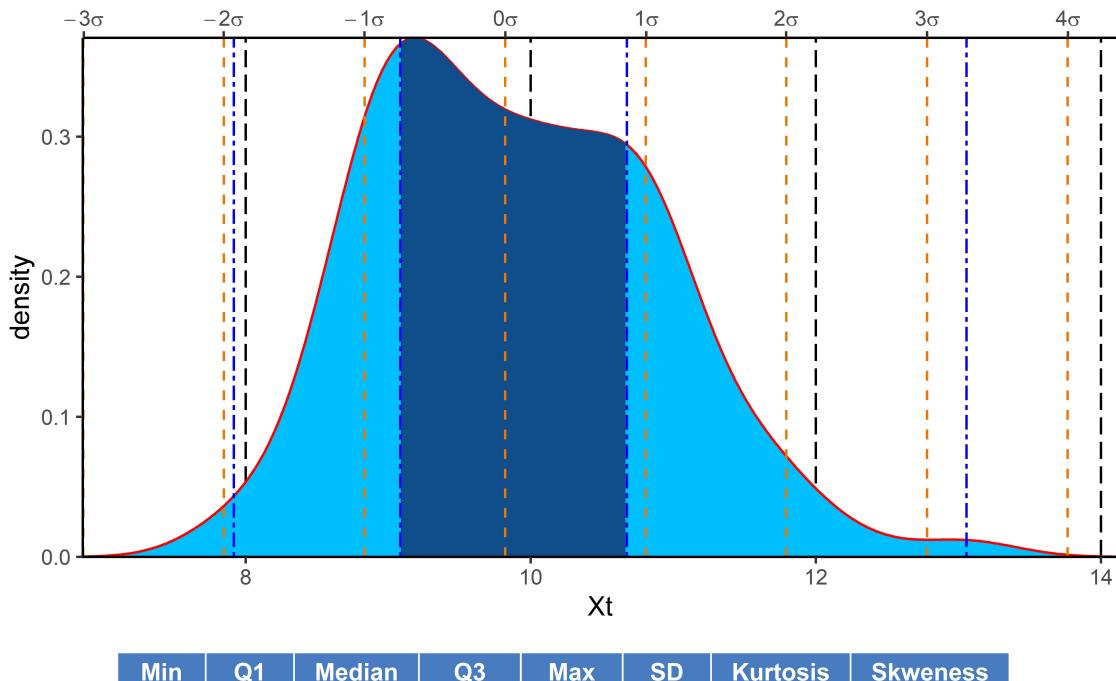
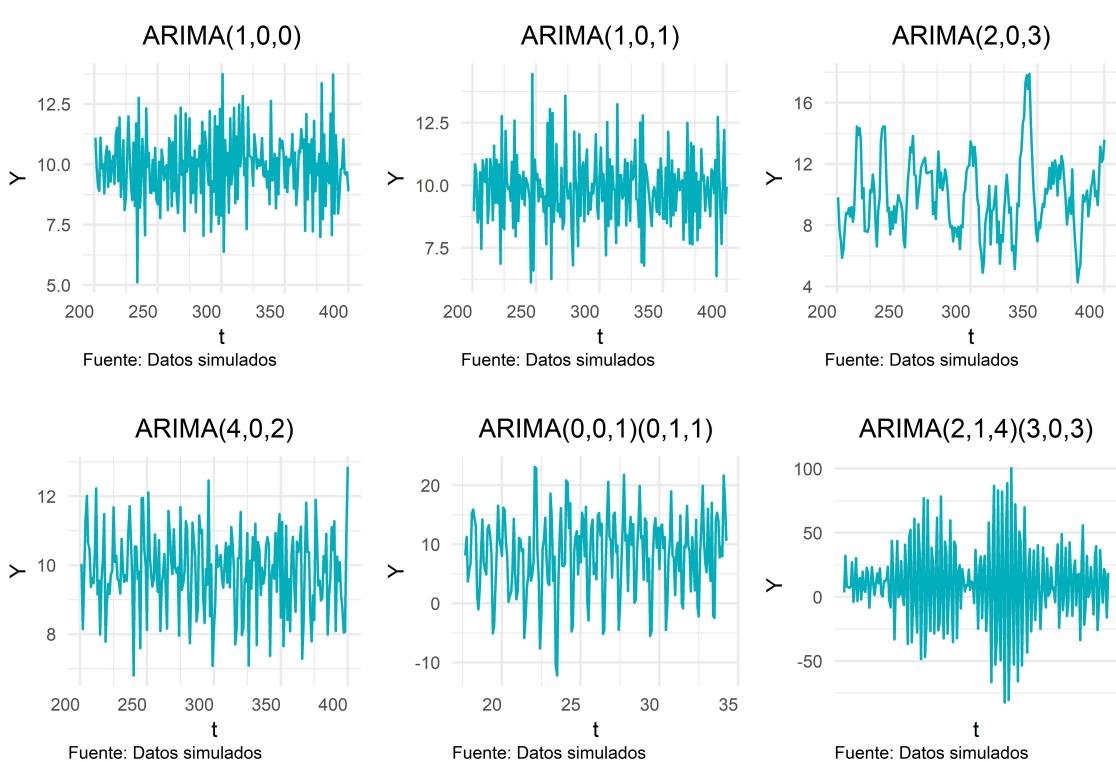


Figura 6: Series cronológicas simuladas



pronósticos se obtienen tanto con el `auto.arima()` como con la sobreparametrización, pues ambos son superiores a los obtenidos mediante un $ARIMA(1,1,1)$. Al ir incorporando términos en las series no estacionales, como es el caso de los datos simulados a partir de un $ARIMA(2,0,3)$, los

mejores pronósticos se obtienen mediante el uso de la sobreparametrización, pues la magnitud de sus errores son siempre menores.

Cuando se consideran series cronológicas estacionales, se presenta un comportamiento similar a lo previamente descrito. Al tener una baja cantidad de parámetros en los datos simulados de un proceso estacional $ARIMA(0,0,1)(0,1,1)_{12}$, la sobreparametrización iguala los resultados obtenidos mediante la función `auto.arima()`; y además, cuando se incorporan más parámetros al modelo generador de los dato como en el caso del $ARIMA(2,1,4)(3,0,3)_{12}$, la sobreparametrización logra captar de mejor manera el comportamiento de los datos, obteniendo así menores mediciones del error, es decir, los pronósticos se acercan más a los datos reales.

Cuadro 1: Medidas de rendimiento según método de estimación

Proceso original	Método	Proceso	MAE	MASE	RMSE
ARIMA(1,0,0)	auto.arima	ARIMA(1,0,0)	0.4942305	0.5528034	0.6402943
	Sobreparametrización	ARIMA(1,0,0)	0.4942305	0.5528034	0.6402943
	ARIMA estándar	ARIMA(1,1,1)	0.4765113	0.5329843	0.6543140
ARIMA(1,0,1)	auto.arima	ARIMA(0,0,1)	0.8070104	0.5896820	1.1022758
	Sobreparametrización	ARIMA(0,0,1)	0.8070104	0.5896820	1.1022758
	ARIMA estándar	ARIMA(1,1,1)	1.0726758	0.7838035	1.1821687
ARIMA(2,0,3)	auto.arima	ARIMA(2,0,2)	0.5128772	1.1656878	0.6699120
	Sobreparametrización	ARIMA(5,1,1)	0.4933922	1.1214016	0.6197431
	ARIMA estándar	ARIMA(1,1,1)	0.5390294	1.2251276	0.6233500
ARIMA(4,0,2)	auto.arima	ARIMA(2,0,3)	0.5454557	1.1463361	0.7156326
	Sobreparametrización	ARIMA(4,0,2)	0.4570216	0.9604820	0.6649190
	ARIMA estándar	ARIMA(1,1,1)	0.6131072	1.2885133	0.6860296
ARIMA(0,0,1)(0,1,1)[12]	auto.arima	ARIMA(0,0,1)(0,1,1)[12]	0.3877273	0.0547085	0.8137802
	Sobreparametrización	ARIMA(0,0,1)(0,1,1)[12]	0.3877273	0.0547085	0.8137802
	ARIMA estándar	ARIMA(1,1,1)(1,1,1)[12]	0.3941180	0.0556102	0.9030985
ARIMA(2,1,4)(3,0,3)[12]	auto.arima	ARIMA(2,0,4)(0,0,2)[12]	3.6020919	0.1811366	4.8015578
	Sobreparametrización	ARIMA(2,1,3)(0,1,3)[12]	3.5539765	0.1787170	4.7836465
	ARIMA estándar	ARIMA(1,1,1)(1,1,1)[12]	9.6771221	0.4866286	14.3620737

4.3 Estimaciones en series cronológicas costarricenses reales

4.3.1 Tasa de mortalidad infantil interanual

La Tasa de Mortalidad Infantil (TMI) es uno de los indicadores demográficos más importantes, pues es utilizado como un parámetro de referencia sobre la calidad del sistema de salud, tanto a nivel nacional como regional. Si bien este indicador se construye relacionando las defunciones de menores de un año con el total de nacimientos, también involucra de manera implícita otras condiciones tales como las económicas, sociales y culturales, así como la efectividad en los métodos preventivos y curativos de esta categoría poblacional ([León, 1998](#)). Debido a esto, el fallecimiento de un niño menor de un año se traduce en una falla del sistema de salud, por lo que estos casos son sujetos de estudio con el fin de conocer las causas que desencadenaron el evento.

En algunos países en vías de desarrollo de Asia, África y América Latina, la mortalidad infantil alcanza valores elevados pues la desnutrición, ausencia de asistencia médica y mala calidad de las

condiciones sanitarias son, a diferencia de los países más desarrollados, algo muy común ([Donoso, 2004](#)). En el caso de Costa Rica, la unidad de estadísticas demográficas del Instituto Nacional de Estadística y Censos⁷ (INEC) es el ente encargado de reportar este indicador con el fin de dar seguimiento y control al comportamiento del mismo a lo largo del tiempo con el objetivo de llegar a los niveles más bajos posibles.

En el INEC, cada mes se publica el boletín de la TMII interanual (TMII), que analiza la TMII de un mes y los 11 meses previos para comparar los períodos correspondientes ([INEC, 2004](#)). Este apartado busca hacer un análisis de la TMII para los 12 períodos desde el año 1989 y hasta 2017, y no de manera mensual simple, pues dada la volatilidad del fenómeno de estudio, hacer un estudio interanual permite analizar de una mejor manera los cambios entre períodos. Es decir, se analizará la TMII desde el período Febrero 1989 – Enero 1990 hasta el período Enero 2017 – Diciembre 2017.

La importancia de este proceso, aparte de servir de parámetro para evaluar el sistema de salud, está en su estrecha relación con las proyecciones de población, pues como se mencionó previamente, la TMII analiza la mortalidad en el grupo de edad de menores de un año, que es el primer grupo al generar tablas de mortalidad, ya sea de la forma clásica o mediante la mortalidad óptima ([Villalón, 2006](#)). Uno de los métodos más conocidos para realizar estas estimaciones es el método de los componentes de cambio demográfico, que son la fecundidad, la mortalidad y la migración. En el caso de la mortalidad, uno de los puntos de partida es la estimación de las tasas de mortalidad por grupos de edad, siendo de particular interés la de menores de cinco años, pues esta a su vez se subdivide en los grupos de menores de un año y el de uno a cuatro años. Conocer el comportamiento de la mortalidad infantil es importante porque es en este grupo de edad en el que pueden existir cambios muy bruscos en la mortalidad y la fecundidad ([Rincon, 2000](#)).

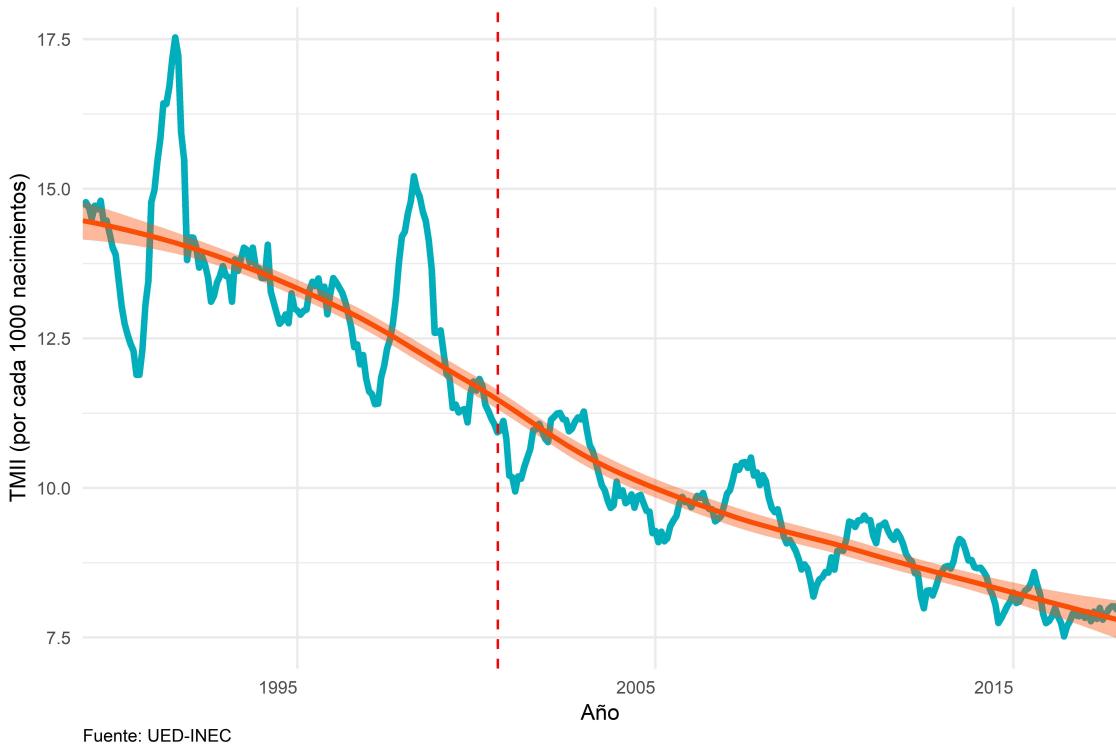
Dado que la medición de la TMII se hace partiendo de un determinado mes y a partir de éste se consideran los 11 meses anteriores, el primer valor de la base de datos fue medido a partir de Enero de 2000, que corresponde al período interanual Febrero 1999 – Enero 2000. Todos los períodos siguientes se muestran en la figura 7.

La serie muestra picos y valles pronunciados a lo largo de todo el período. A modo de visualización, se ajustó un suavizamiento de Loess para buscar señales de tendencia y concavidad en los datos temporales. La línea roja punteada se ubica aproximadamente en el mes de Julio del año 2000, pues a partir de ese punto el suavizamiento de Loess muestra un ligero cambio en la concavidad, lo cual sugiere que a partir ese punto será más difícil que la TMII vuelva a alcanzar valores similares a los mostrados al inicio de la serie. Además, al presentarse dos caídas y subidas abruptas en la TMII, esta tiende a estabilizarse.

Mediante un análisis visual, la figura 8 parece respaldar el supuesto de que la mortalidad no posee

⁷<http://www.inec.go.cr/>

Figura 7: Tasa de Mortalidad Infantil Interanual 1989 - 2017

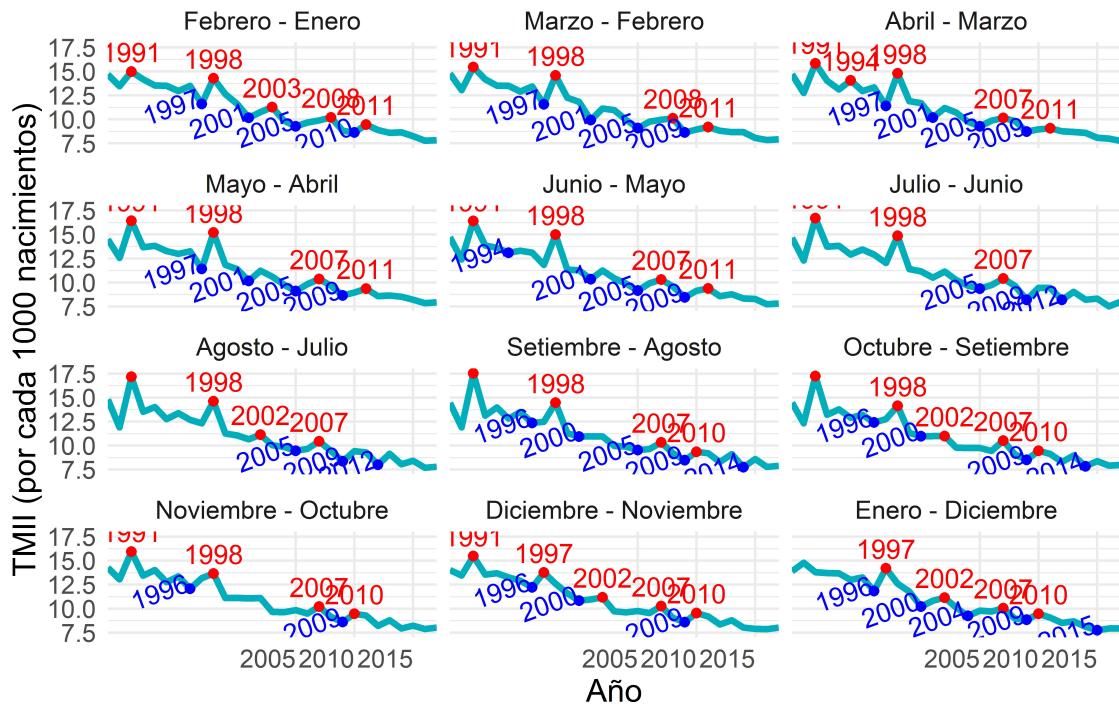


efectos estacionales determinantes, pues para cada uno de los 12 períodos, en ninguno parecen existir mayores diferencias. El efecto que se mantiene en cada uno de los períodos es el de la tendencia, pues en cada uno ésta sigue descendiendo con el pasar de los años. Este hecho coincide con lo observado en la figura 7.

Para hacer la descomposición de la serie se hizo una transformación de Box-Cox con $\lambda = 1$ para aplicarla de manera multiplicativa. Esto se debe a que en la figura 7 pueden observarse cambios considerables en la variabilidad de la serie a lo largo del tiempo. La figura 9 muestra, como se mencionó previamente, una tendencia decreciente y una estacionalidad que no es reiterada a lo largo del tiempo. Además, el componente aleatorio muestra como los errores no son constantes durante todo el período.

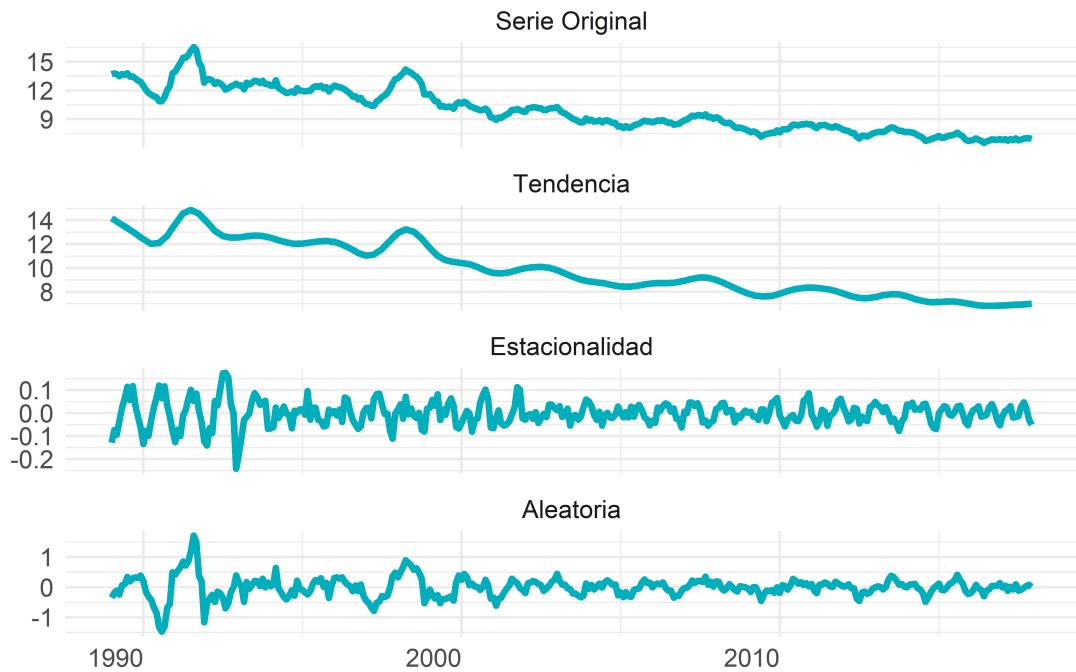
Tras tener un panorama más claro del comportamiento de la serie mediante el análisis descriptivo anterior, se modela la serie mediante la función `auto.arima()`, teniendo como mejor modelo un *ARIMA*(2, 1, 0)(0, 0, 1); y utilizando la sobreparametrización se tiene como mejor modelo un *ARIMA*(4, 1, 0)(4, 1, 0). El cuadro 2 muestra como el uso de la sobreparametrización ofrece pronósticos más cercanos al valor real, y aunque los pronósticos están lejos de ser perfectos, el método propuesto logra reproducir de mejor manera el comportamiento de los datos, tal y como se muestra en la figura 10

Figura 8: Tasa de Mortalidad Infantil Interanual 1989 - 2017 según períodos



Fuente: UED-INEC

Figura 9: Descomposición de la TMII en el periodo 2000 - 2017



Fuente: UED-INEC

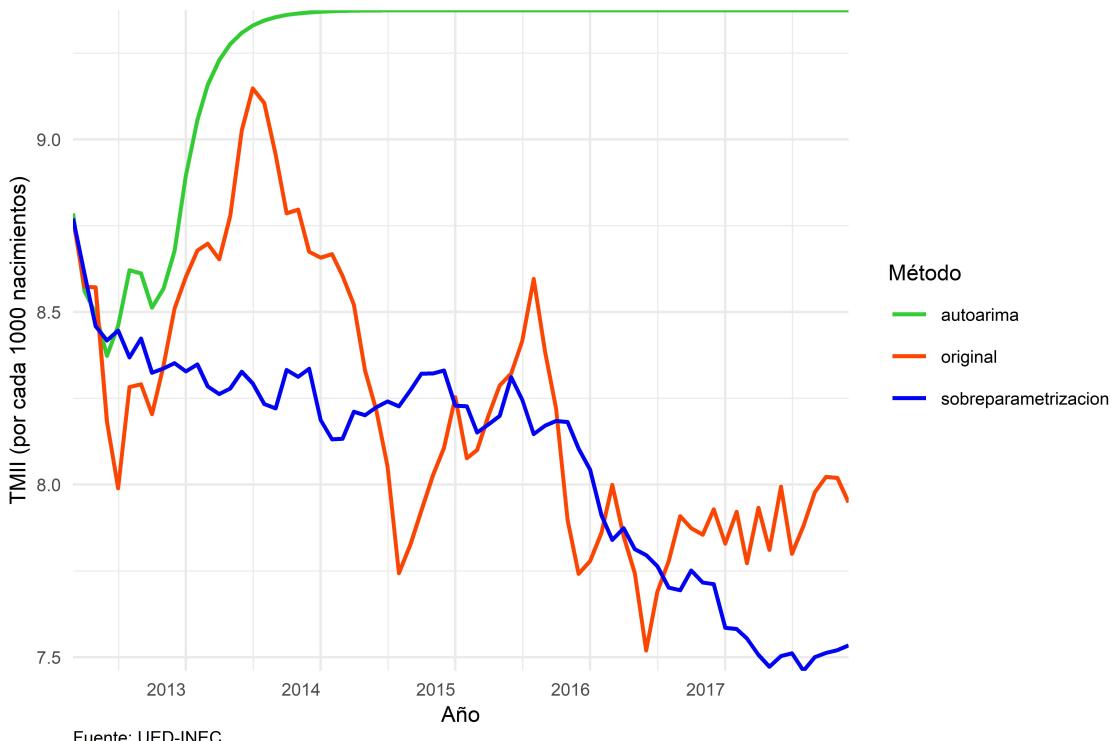
4.3.2 Mortalidad por causa externa

La violencia es un acto tan antiguo como el mundo, sin embargo, la evolución de esta en conjunto con el crecimiento de su relación con las defunciones registradas en una población la vuelven

Cuadro 2: Medidas de rendimiento según método de estimación para la TMII

Método	RMSE	MAE	MAPE
autoarima	1.1481936	1.0192497	12.683182
sobreparametrización	0.3495541	0.2839742	3.418392

Figura 10: Pronósticos de la TMII según método de estimación



un problema de salud pública. En base a la clasificación Internacional de Enfermedades ([OPS, 2016](#)) de la de la Organización Mundial de la Salud⁸, las defunciones pueden clasificarse en cuatro grandes grupos, siendo el más importante el de las causas naturales, el cual incluye enfermedades congénitas, cardiopatías u otras relacionadas con la vejez. En menor cuantía se encuentran las causas de muerte ignoradas, las cuales se dan cuando la causa de muerte es desconocida y de intención indeterminada; y de forma similar se encuentran las causas de muerte que se mantienen en estudio, bien sea por parte de la morgue o de algún otro organismo, esta última tiene pocos registros conforme más se retrocede en el tiempo.

El otro gran grupo, aunque considerablemente menor que las causas naturales, son las causas externas, las cuales son objeto de análisis en este apartado. Este grupo puede a su vez ser clasificado en homicidios, suicidios y las muertes accidentales, esta última comprende los accidentes de tránsito, las muertes por caídas, personas ahogadas, víctimas de incendios, terraplenes u otros similares. Aunado a estas categorías se encuentran también las causas indeterminadas, las cuales se diferencian a las ignoradas en que se sabe que se debe a una causa externa pero no se conoce con certeza a cuál

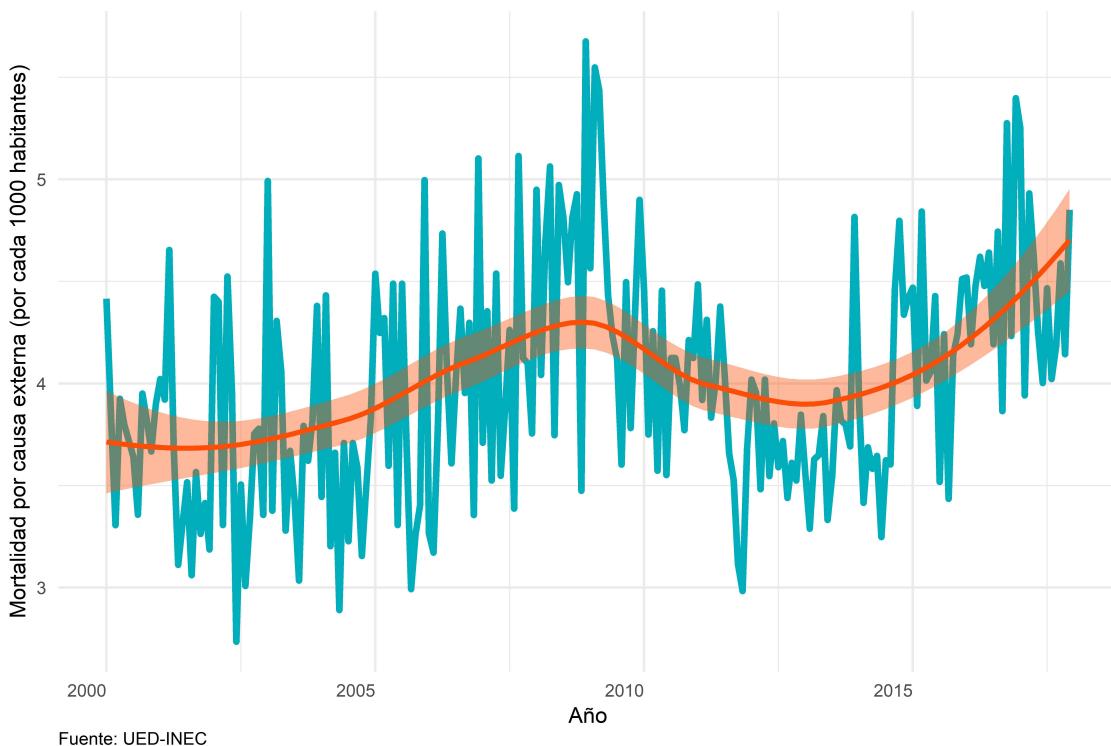
⁸<https://www.paho.org/salud-en-las-americas-2017/?lang=es>

categoría pertenece o aún está en investigación, tal es el caso de una persona que fallece debido a una alta ingesta de drogas o estupefacientes; bien pudo haber consumido intencionalmente hasta morir, lo cual sería un suicidio, o bien el consumo excesivo se debió a un accidente.

En Costa Rica para el año 2011, las muertes por causas externas ocuparon el tercer lugar, siendo solo superadas por las enfermedades del sistema circulatorio, en particular las enfermedades cardiovasculares, y los tumores, ambos casos mostraron una tendencia ascendente ([Nación, 2013](#)). Es debido a los elevados costos económicos y sociales ([Cardona, 2013](#)) que se aborda la imperiosa necesidad comprender el comportamiento de las defunciones debido a las causas externas con el fin de contar con un punto de partida para la elaboración de políticas públicas que busquen reducir al mínimo este tipo de eventos.

Dado que los registros de defunciones por causa externa se realizan diariamente, conviene analizar su comportamiento de manera mensual desde inicios del milenio de una manera más general, dicho comportamiento puede observarse en la figura 11.

Figura 11: Mortalidad por causa externa 2000 - 2017

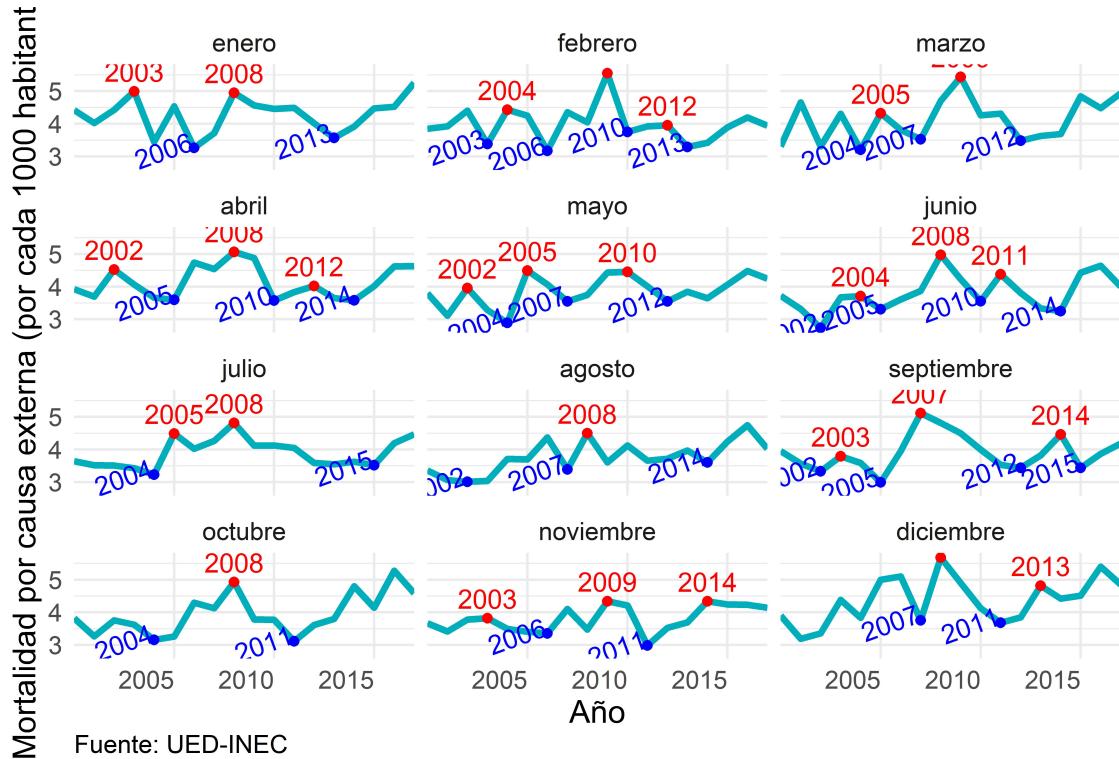


Es importante recalcar que, entre Junio del año 2012 y Diciembre del año 2017, el aumento en la tasa de cambio de la cantidad de defunciones debido a causas externas coincide con el aumento de la flotilla de motocicletas, pues en un período de cinco años esta cifra creció en un 189 % ([Vázquez, 2017](#)). Conviene entonces verificar el comportamiento a lo interno de la serie en referencias a las categorías de las causas externas.

De la figura 12 puede notarse que cada mes tiene sus picos y valles durante cada mes a lo largo

del periodo, siendo los meses de Enero, Abril y Diciembre los que presentaron valores ligeramente más altos entre los años 2000 y 2017.

Figura 12: Mortalidad por causa externa 2000 - 2017 según mes



Fuente: UED-INEC

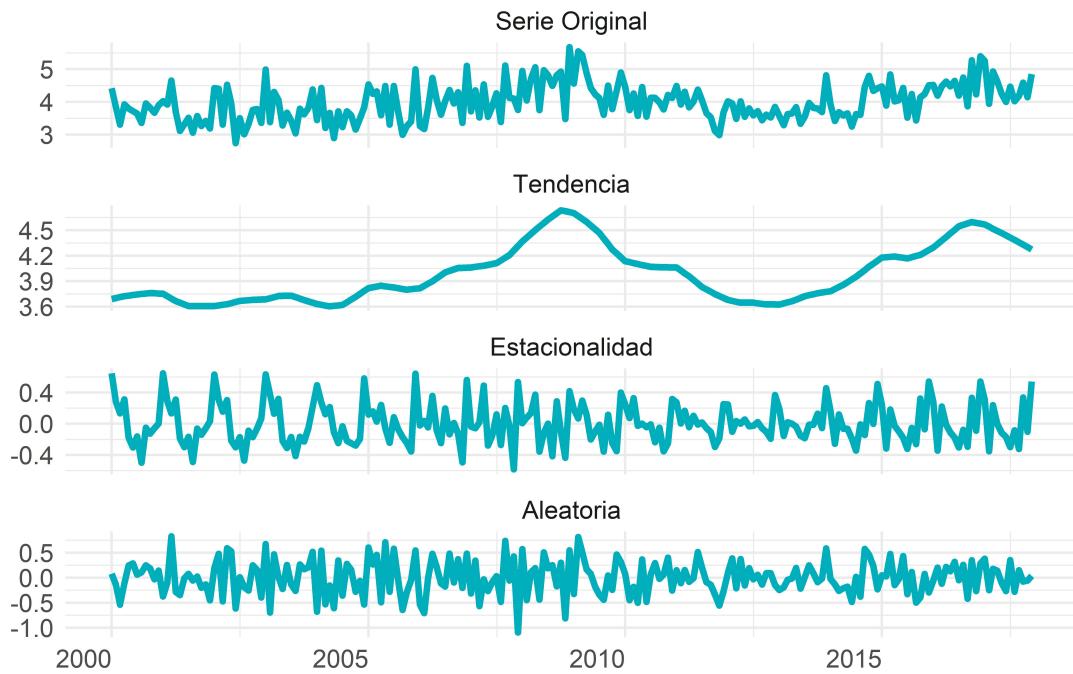
La descomposición de la serie se hará de forma aditiva debido a que en el gráfico 1 no se observan grandes cambios en la variabilidad a lo largo del tiempo. La figura 13 muestra que la tendencia se mantiene casi constante a lo largo del tiempo, mientras que parece haber estacionalidad en ciertos lapsos de la segunda mitad del año. Además, el componente aleatorio muestra como los errores no son constantes a lo largo de todo el período.

De manera similar a lo hecho para pronosticar la TMII, se ajusta un modelo utilizando la función `auto.arima()`, siendo el modelo sugerido un *ARIMA*(1, 1, 1); mientras que la sobreparametrización propone como mejor modelo un *ARIMA*(2, 0, 1)(0, 1, 1)₁₂. El cuadro 3 muestra las medidas de rendimiento obtenidas de los pronósticos realizados con cada método, los cuales son mejores al utilizar la sobreparametrización; este hecho también se observa gráficamente mediante la figura 14

Cuadro 3: Medidas de rendimiento según método de estimación para la Mortalidad por causa externa

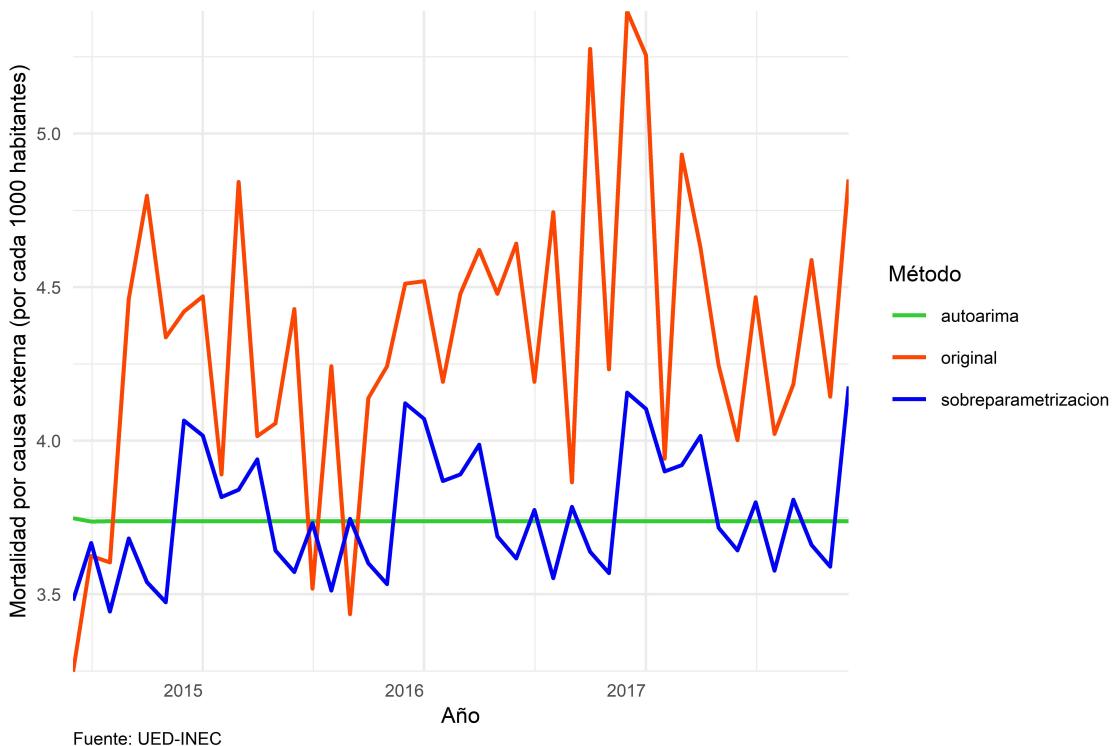
Método	RMSE	MAE	MAPE
autoarima	0.7555003	0.6508858	14.35132
sobreparametrizacion	0.7058923	0.6011625	13.27879

Figura 13: Descomposición de las defunciones por causa externa en el periodo 2000-2017



Fuente: UED-INEC

Figura 14: Pronósticos de la TMII según método de estimación



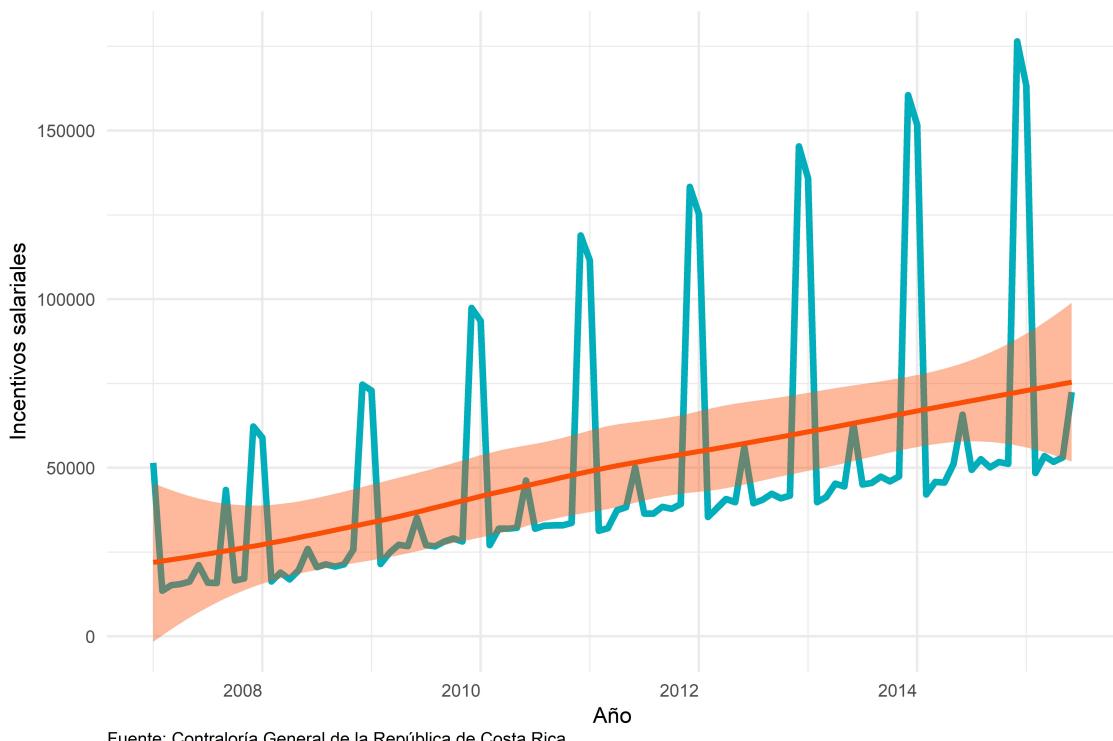
4.3.3 Incentivos salariales del sector público

Los incentivos salariales son 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, desarraigado, regionalización, riesgo policial, riesgo penitenciario, riesgo de seguridad y vigilancia, peligrosidad, incentivo didáctico, entre otros. Esta serie cronológica representa los incentivos salariales en millones de colones del sector público de Costa Rica de enero 2007 a junio 2015.

De manera análoga a las secciones anteriores, la figura 15 muestra el comportamiento general de la serie cronológica. al hacer un suavizamiento Loess hay un ligero cambio de concavidad a partir de Julio 2008, lo cual sugiere que a partir de este momento los incentivos salariales vuelvan a alcanzar valores similares a los mostrados al inicio de la serie. La figura 16 muestra cómo hay un crecimiento sostenido de los incentivos en cada mes a lo largo de todo el periodo. Sin embargo, este crecimiento se da a una tasa mucho mayor en la época de fin y principio de año.

Figura 15: Incentivos salariales en el sector público 2007 - 2018



En la figura 17 se muestra la descomposición de la serie en sus distintos componentes. Pueden observarse, además de un crecimiento a lo largo del tiempo, los picos y las caídas en la parte estacional, esto hace referencia a los meses de Diciembre y Enero; cuando no se está en este periodo los incentivos poseen un comportamiento más estable. El componente aleatorio muestra indicios de que la variabilidad de la serie no es homogénea, sino que cambia conforme pasa el tiempo.

Nuevamente, se ajustaron dos modelos ARIMA, el primero de ellos mediante la función `auto.arima()`, siendo el modelo sugerido un $ARIMA(0,0,1)(1,1,0)_{12}$; y el segundo modelo

Figura 16: Incentivos salariales en el sector público 2007 - 2018 según mes

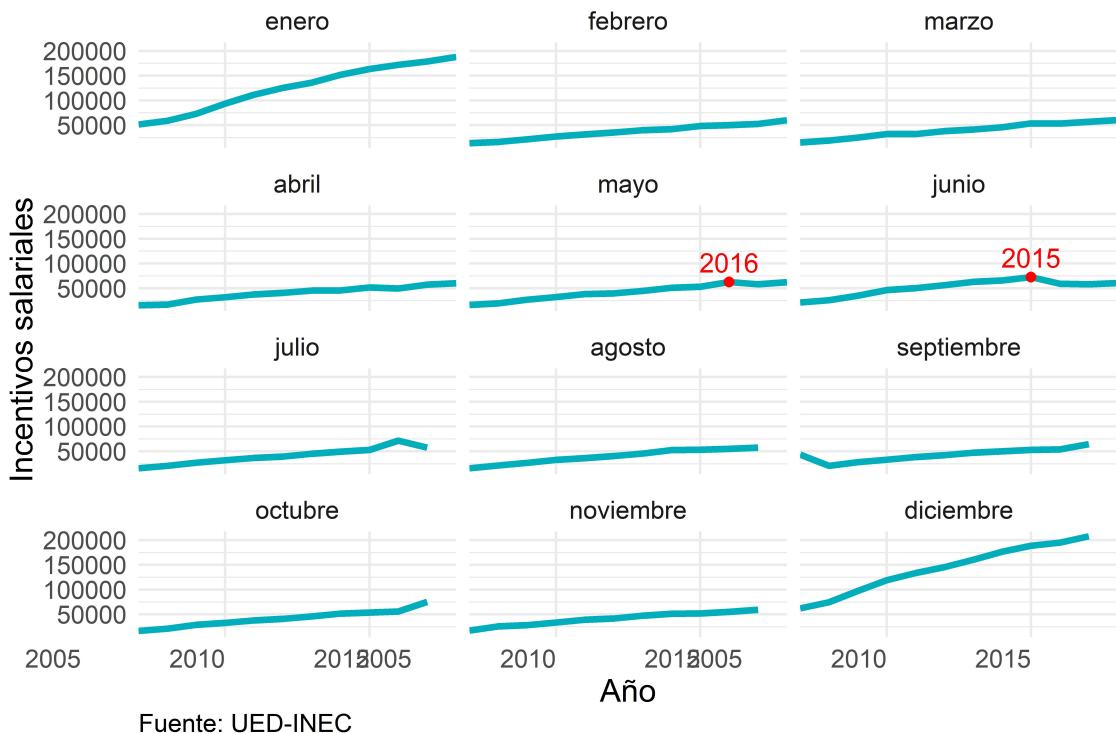
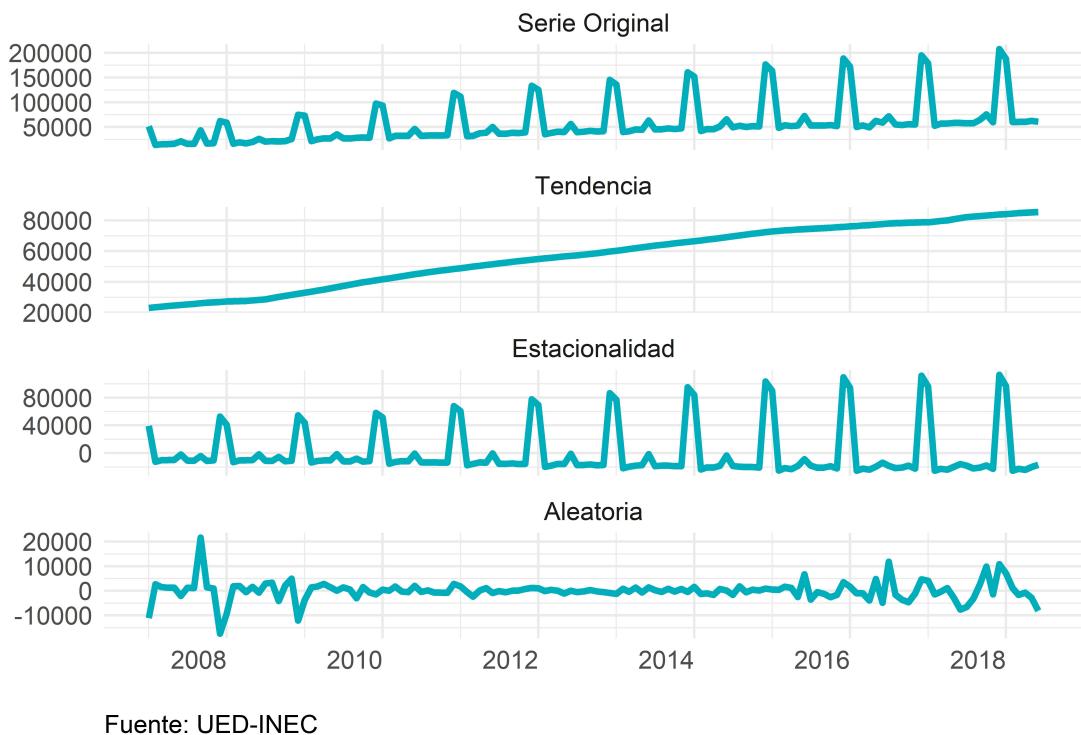


Figura 17: Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018

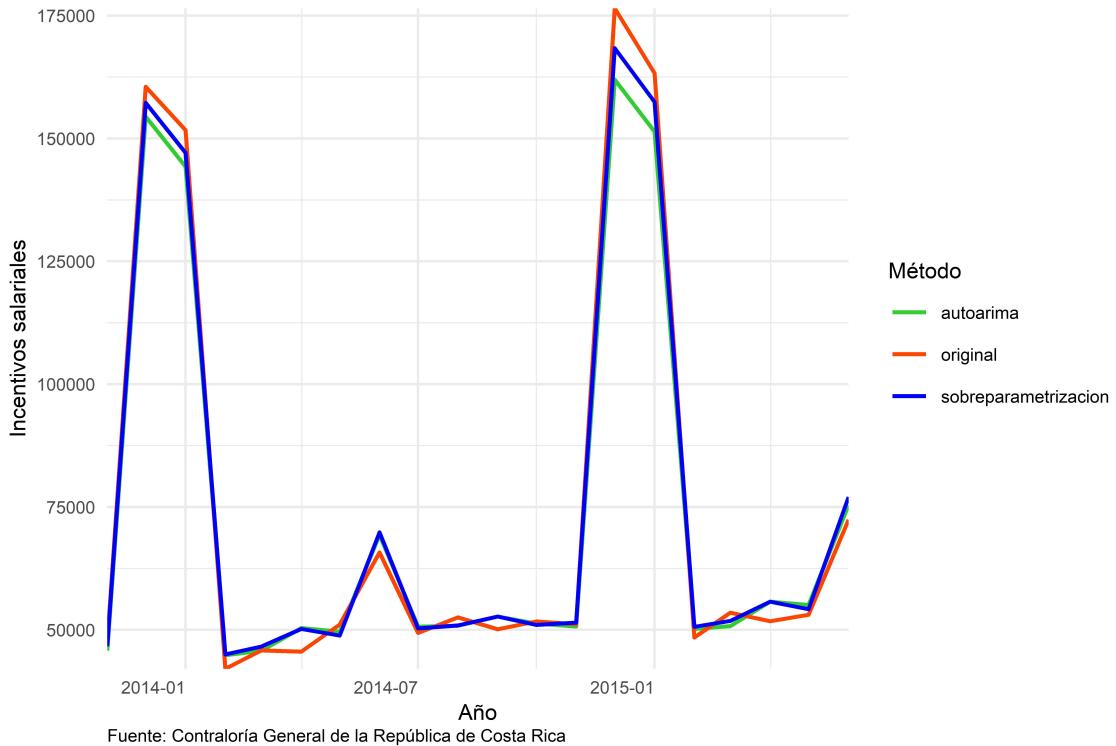


sugerido, utilizando la sobreparametrización, es un $ARIMA(2, 1, 0)(1, 2, 0)_{12}$. Tanto el cuadro 4 como la figura 18 muestran que nuevamente los pronósticos obtenidos son superiores utilizando la sobreparametrización.

Cuadro 4: Medidas de rendimiento según método de estimación para los incentivos salariales

Método	RMSE	MAE	MAPE
autoarima	5212.797	3701.074	4.498137
sobreparametrización	3476.907	2846.975	4.000943

Figura 18: Pronósticos de los incentivos salariales según método de estimación



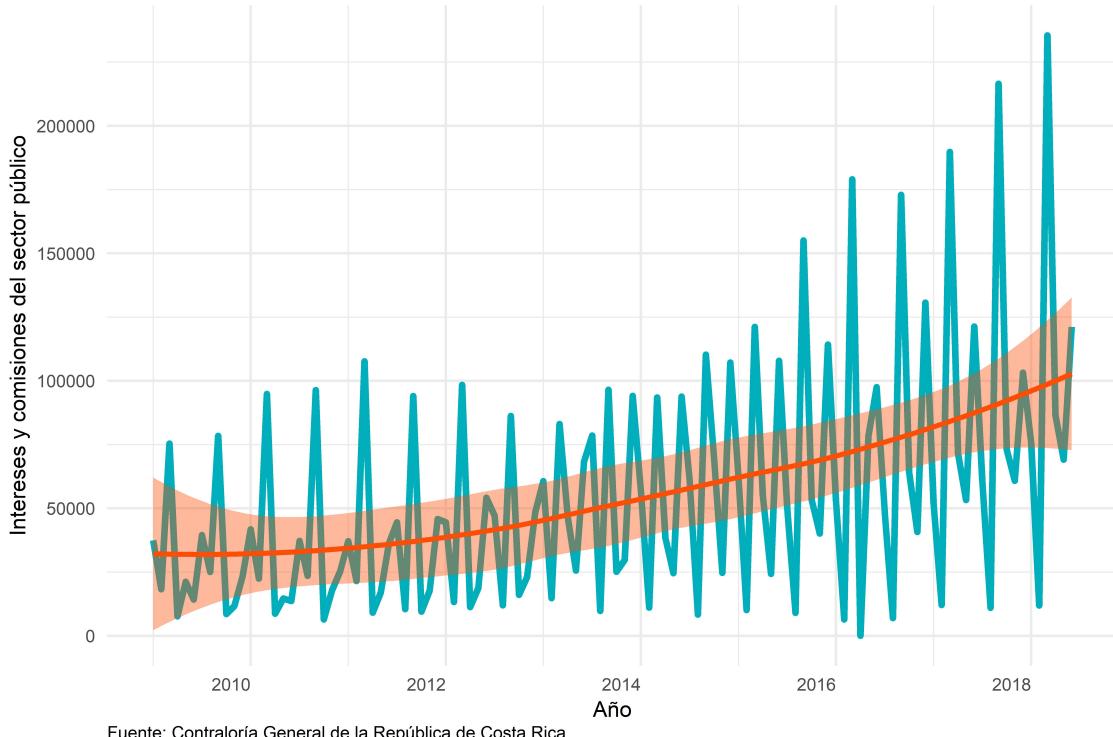
4.3.4 Intereses y comisiones del sector público

Finalmente, se utiliza para este análisis la serie cronológica de los intereses y comisiones del sector público, que comprenden 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 tranzados 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.

Para iniciar el análisis exploratorio de esta serie, la figura 19 muestra que hay un ligero cambio de concavidad a partir de Julio 2010, esto sugiere que a partir de este momento los intereses y

comisiones inician una tendencia al alza, la cual se sostiene hasta Junio del 2018. Por su parte, la figura 20 muestra cómo hay un crecimiento sostenido de los intereses y comisiones del sector público al final de cada trimestre durante todo el periodo, mientras que se mantiene casi constante durante los primeros dos meses de cada trimestre. La caída más pronunciada se dio en abril del 2015 mientras que la tasa de crecimiento más rápida parece darse al final del primer trimestre. Además, en la figura 21 se muestra la descomposición de la serie en sus distintos componentes. Puede observarse, además de un crecimiento a lo largo del tiempo posterior a una disminución, los picos y las caídas en la parte estacional, esto en cuanto a los cierres trimestrales previamente mencionados. El componente aleatorio muestra indicios de que la variabilidad de la serie no es homogénea, sino que cambia conforme pasa el tiempo, pues durante un tiempo se mantuvo relativamente estable pero luego presenta algunos cambios.

Figura 19: Intereses y comisiones del sector público en el periodo 2007-2018

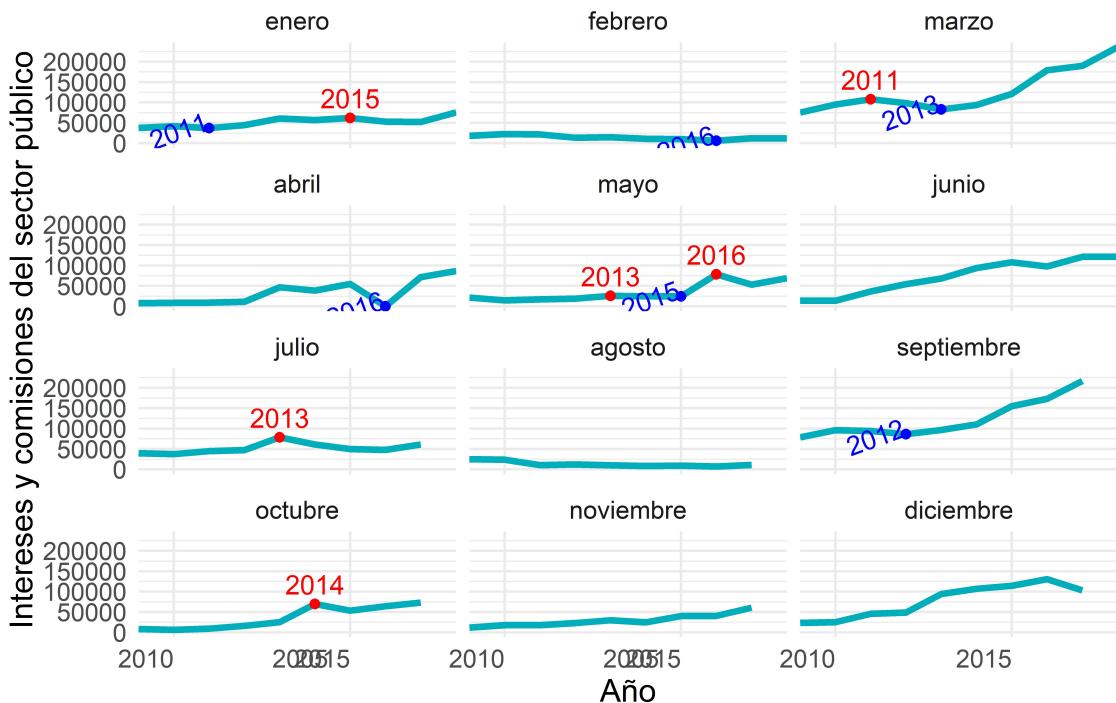


Fuente: Contraloría General de la República de Costa Rica

En la figura 21 se muestra la descomposición de la serie en sus distintos componentes. Pueden observarse, además de un crecimiento a lo largo del tiempo posterior a una disminución, los picos y las caídas en la parte estacional. El componente aleatorio muestra indicios de que la variabilidad de la serie no es homogénea, sino que cambia conforme pasa el tiempo, pues durante un tiempo se mantuvo relativamente estable pero luego presenta algunos cambios.

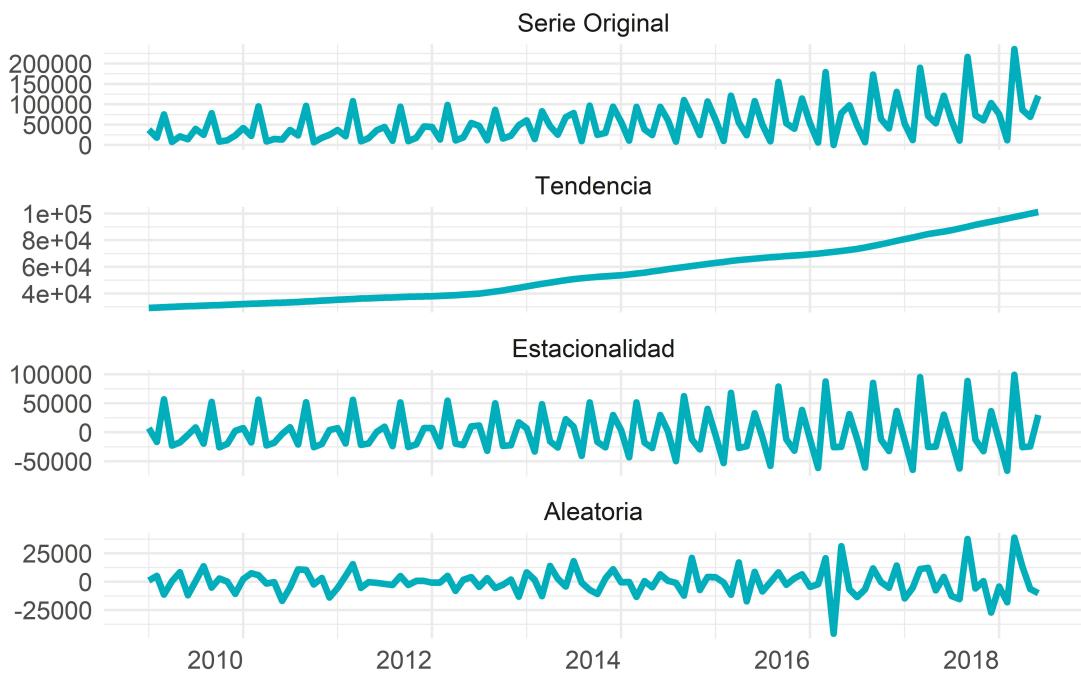
Por último, los modelos ajustados para pronosticar los intereses y comisiones del sector público con un $ARIMA(0,0,1)(0,1,0)_{12}$ para el caso de la función `auto.arima()`; y un $ARIMA(0,1,2)(0,1,0)_{12}$. En este caso, la sobreparametrización es superior en dos de las tres

Figura 20: Intereses y comisiones del sector público en el periodo 2007-2018 según mes



Fuente: Contraloría General de la República de Costa Rica

Figura 21: Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018



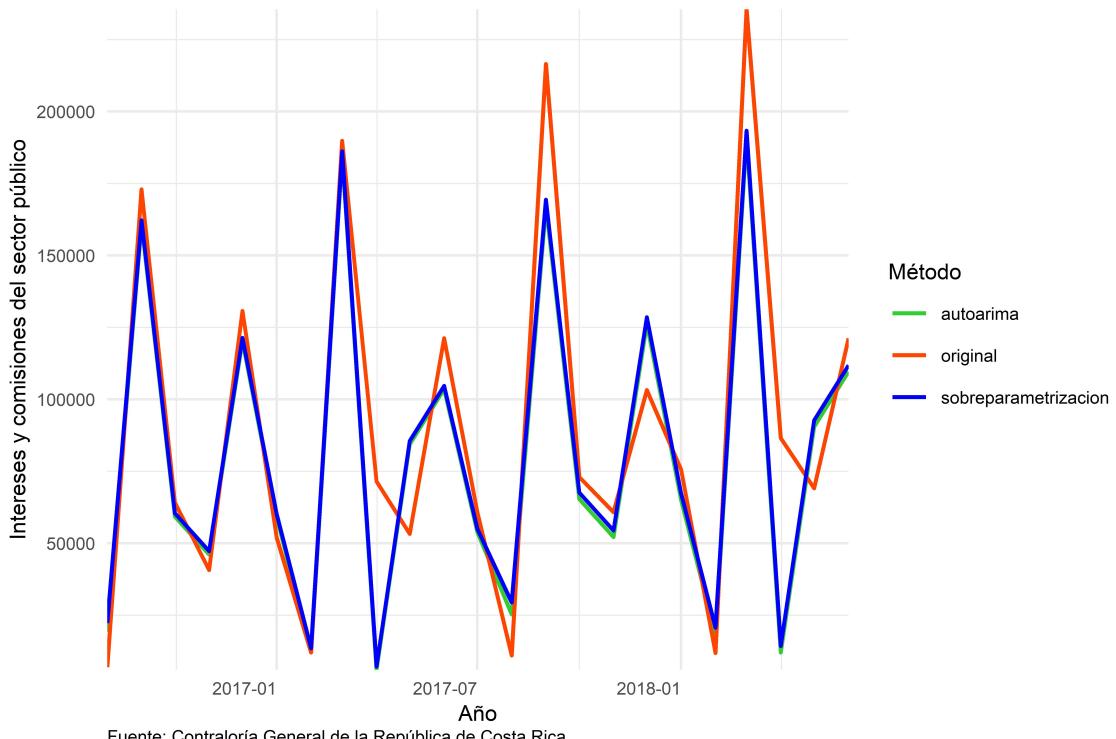
Fuente: Contraloría General de la República de Costa Rica

medidas de rendimiento tal y como muestra el cuadro 5, mientras que la figura 22 muestra de manera gráfica los pronósticos obtenidos, que son bastante similares pues ambos modelos difieren solamente en un parámetro para la parte estacional.

Cuadro 5: Medidas de rendimiento según método de estimación para los intereses y comisiones del sector público

Método	RMSE	MAE	MAPE
autoarima	27737.53	19562.58	35.66424
sobreparametrizacion	27280.46	19346.25	39.89648

Figura 22: Pronósticos de los intereses y comisiones del sector público según método de estimación



5 CONCLUSIONES

La presente investigación ha cubierto las bases fundamentales relacionadas al análisis de series cronológicas. Se ha realizado una recapitulación histórica de esta rama de la estadística y se han puesto en evidencia uno de los principales problemas de las series de tiempo, como lo es la subjetividad del observador en la selección de modelos. Es ante este inconveniente que la investigación buscó siempre proponer un aporte metodológico para la selección de modelos ARIMA y por ende, una mejora en los pronósticos obtenidos a partir de estos modelos.

El uso de la sobreparametrización propuesto mediante un algoritmo de selección de modelos ha sido implementado de la forma esperada, pues la evaluación de los resultados al incorporar coeficientes al modelo se puso a prueba tanto en series cronológicas simuladas como en datos reales. El método propuesto implementado en las funciones del lenguaje R permite evaluar una gama más amplia de modelos ARIMA al definir un máximo en la cantidad de parámetros para las partes estacionales y no estacionales de la series cronológicas, pues al definir este máximo se definen todos los posibles escenarios que posteriormente evalúan el aporte de cada nuevo término a los pronósticos. La

incorporación de estos nuevos parámetros en los modelos ARIMA son validados mediante pruebas de significancia estadística, particiones de la serie cronológica, medidas de bondad de ajuste de los modelos y sus correspondientes medidas de rendimiento.

Como parte de la investigación, la series cronológicas utilizadas de forma simulada y generadas a partir de registros administrativos muestran como el uso de la sobreparametrización iguala y en muchos casos mejora la calidad de los pronósticos obtenidos en comparación a métodos ya establecidos, como es el caso de la función `auto.arima()`, o estimación de modelos más genéricos con un bajo número de parámetros, como los modelos estándar $ARIMA(1, 1, 1)$ o $ARIMA(1, 1, 1)(1, 1, 1)_{12}$.

Como puede apreciarse en los resultados, al tener datos que vienen de un proceso con bajo número de parámetros, el uso de la sobreparametrización logra captar de buena manera el comportamiento de la serie y, además, cuando el proceso que gobierna la serie es de un mayor grado, la metodología propuesta, al considerar un mayor espectro paramétrico, es capaz de capturar de buena forma el comportamiento de la serie y conseguir pronósticos con una precisión mayor al de los métodos más tradicionales. Lo anterior representa una mejora en cuanto a la utilización de modelos ARIMA para el pronóstico de series cronológicas, lo cual a su vez aporta herramientas para la toma de decisiones relacionadas a este tipo de análisis.

Una potencial mejora al uso de la sobreparametrización es la inclusión semi-automática de regresores para controlar cambios estructurales de la serie cronológica en estudio, pues estos coeficientes adicionales podrían controlar cambios particulares en la serie y que podrían mejorar la precisión de los pronósticos.

La metodología aquí propuesta se encuentra disponible de manera abierta mediante el paquete de R `popstudy`, el cual fue desarrollado para esta investigación y cuenta con los procedimientos previamente descritos. Se encuentra disponible en un repositorio de Github⁹ y además en el repositorio CRAN¹⁰, que es la fuente oficial de los paquetes del lenguaje R.

⁹<https://github.com/cgamoasanabria/popstudy>

¹⁰Publicación pendiente.

6 ANEXOS

6.1 Función de sobreparametrización

Código 1: Función op.arima

```
op.arima <- function(arima_process = c(p = 1, d = 1, q = 1,
                                         P = 1, D = 1, Q = 1),
                       seasonal_periodicity,
                       time_serie, reg = NULL, horiz = 12,
                       prop=.8, training_weight=.2, testing_weight=.8,
                       parallelize=FALSE,
                       clusters=detectCores(logical = FALSE),...){

  data_partition <- round(length(time_serie)*prop, 0)
  train <- subset(time_serie, end=data_partition)
  test <- subset(time_serie, start=data_partition+1)

  arima_model <- function(time_serie, non_seasonal, seasonal, periodic,
                         regr = NULL,...){

    if(is.list(non_seasonal)){
      non_seasonal <- unlist(non_seasonal)
    }

    if(is.list(seasonal)){
      seasonal <- unlist(seasonal)
    }

    seasonal_part <- list(order=seasonal, period=periodic)
    if(is.null(regr)){
      arima_model <- tryCatch({
        Arima(time_serie,
              order = non_seasonal,
              seasonal = seasonal_part,...)
      },
      error = function(e) NULL)
    }
  }
}
```

```

}

if(!is.null(regr)){
  arima_model <-tryCatch({
    Arima(time_serie,
          order = non_seasonal,
          seasonal = seasonal_part,
          xreg = regr,...)
  },
  error = function(e) NULL)
}

if(!is.null(arima_model)){
  degrees_of_freedom <- arima_model$nobs - length(arima_model$coef)
  t_value <- arima_model$coef/sqrt(diag(arima_model$var.coef))
  prob <- stats::pf(t_value^2, df1 = 1, df2 = degrees_of_freedom,
                     lower.tail = FALSE)
  ifelse(sum(1*prob>0.05)<1, return(arima_model), 1)
}

}

arima_measures <- function(arima_model, testing, horizon, regr = NULL){

  model_spec <- capture.output(arima_model)
  model_spec <- substr(model_spec[2],1, 23)

  data <- capture.output(summary(arima_model))
  data <- data[grep("AIC", data) == T]
  model_info <- strsplit(data, " ")
  pos <- which(sapply(model_info, nchar)>0)
  model_info <- model_info[[1]][pos]
  model_info <- do.call("rbind", strsplit(model_info, "=")) %>%
    data.frame()
  colnames(model_info) <- c("Medida", "Valor")
}

```

```

model_info <- model_info %>%
  mutate(Valor = as.numeric(as.character(Valor))) %>%
  spread(Medida, Valor) %>%
  data.frame(arima_model = model_spec)

model_performance <- data.frame(arima_model = c(model_spec,
                                                 paste(model_spec,
                                                       "Validacion")),
                                 accuracy(forecast(arima_model, horizon,
                                                       xreg = regr),
                                                       testing))

merge(model_info, model_performance, by="arima_model", all = TRUE) %>%
  select(arima_model, AIC, AICc, BIC, MAE, RMSE, MASE)
}

arima_selected <- function(model_table, Wtrain=training_weight,
                           Wtest=testing_weight){

  model_table <- model_table %>%
    distinct(arima_model, .keep_all = TRUE)

  model_table <- model_table %>%
    mutate(mod = as.character(c(0, rep(1:(nrow(model_table)-1) %/% 2)))))

  tabla2 <- model_table %>%
    mutate_at(vars(contains("C")), function(x){x-min(x, na.rm=TRUE)}) %>%
    mutate_if(is.numeric, function(x) ifelse(is.na(x), 0, x)) %>%
    mutate(puntaje = AIC+AICc+BIC+MAE+RMSE+MASE,
          donde = ifelse(grepl("Validacion", arima_model)==TRUE,
                         Wtest, Wtrain),
          puntaje = puntaje*donde)

  suppressMessages({
    minimal_score <- tabla2 %>%
      group_by(mod) %>%
      summarise(puntaje=sum(puntaje)) %>%
  })
}

```

```

ungroup
})

pos <- minimal_score$mod[which(
  minimal_score$puntaje==min(minimal_score$puntaje))]

model_table %>%
  filter(mod %in% pos) %>%
  dplyr::select(arima_model:MASE)
}

suppressWarnings({
  valores <- expand.grid(p = 0:arima_process[1],
                         d = 0:arima_process[2],
                         q = 0:arima_process[3],
                         P = 0:arima_process[4],
                         D = 0:arima_process[5],
                         Q = 0:arima_process[6])

  non_seasonal_values <- split(as.matrix(valores[, 1:3]),
                                 row(valores[, 1:3]))
  seasonal_values <- split(as.matrix(valores[, 4:6]),
                            row(valores[, 4:6]))

  if(parallelize==FALSE){
    arima_models <- mapply(arima_model,
                           non_seasonal=non_seasonal_values,
                           seasonal=seasonal_values,
                           MoreArgs = list(time_serie=train,
                                           regr=reg,
                                           periodic=seasonal_periodicity),
                           SIMPLIFY = FALSE)
  }else{
    clp <- makeCluster(clusters, type = "SOCK", useXDR=FALSE)
  }
})

```

```

clusterEvalQ(clp, expr = {
  library(forecast)
})

arima_models <- clusterMap(cl=clp, fun = arima_model,
                           non_seasonal=non_seasonal_values,
                           seasonal=seasonal_values,
                           MoreArgs = list(time_serie=train, regr=reg, periodic=sea-
                           SIMPLIFY = FALSE, .scheduling = "dynamic"))

stopCluster(clp)

})

pos <- which(sapply(lapply(arima_models, class), length)>1)

final_measures <- do.call("rbind", lapply(arima_models[pos],
                                             arima_measures,
                                             testing = test,
                                             horizon= horiz,
                                             regr = reg)) %>%
  mutate_if(is.numeric, round, 2)

final_list <- list(arima_models=arima_models[pos],
                    final_measures=final_measures,
                    bests=arima_selected(final_measures, Wtrain = training_weight, Wtest =
mod_index <- final_list$bests %>%
  row.names %>%
  as.numeric %>%
  floor %>%
  unique %>%
  as.character

final_list$best_model <- eval(parse(text = paste0(
  "final_list$arima_models$", "``", mod_index, "``")))

```

```

    final_list
  })
}

```

6.2 Función de simulación de series cronológicas

Código 2: Función ts.sim

```

ts.sim <- function(data, n, temporalidad,
                     no.estacional, estacional=c(0,0,0),
                     p=NULL, q=NULL, P=NULL, Q=NULL){

  require(forecast)
  tryCatch({
    coeficientes <- list(p, q, P, Q)
    coeficientes.simulados <- lapply(c(no.estacional[c(1,3)],
                                         estacional[c(1,3)]),
                                       function(x) sample(seq(-1,1,.1), x))

    pos <- which(sapply(coeficientes, is.null)==TRUE)
    pos2 <- which(sapply(coeficientes.simulados, length)>0)

    coeficientes[pos] <- coeficientes.simulados[pos]

    names(coeficientes) <- c("p", "q", "P", "Q")
    coeficientes <- coeficientes[pos2]

    if(TRUE %in% (c("P", "Q") %in% names(coeficientes))){
      modelo <- Arima(ts(data=data, freq=temporalidad),
                      order = no.estacional,
                      seasonal = estacional,
                      fixed=c(unlist(coeficientes)))
    }else{
      modelo <- Arima(ts(data=data, freq=temporalidad),
                      order = no.estacional,
                      seasonal = estacional,
                      fixed=c(unlist(coeficientes), NA))
    }
  })
}

```

```
})

datos <- simulate(modelo, nsim=(n+length(data)))

datos <- subset(datos, start=length(data)+1)

list(modelo=modelo, datos=datos)},
      error = function(e) ts.sim(data, n, temporalidad,
                                no.estacional, estacional,
                                p, q, P, Q))
}
```

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>
- Agrawal, R., & Adhikari, R. (2013). An introductory study on time series modeling and forecasting. *Nova York: CoRR.*
- Aphalo, P. J. (2021). *ggpmisc: Miscellaneous Extensions to 'ggplot2'*. Recuperado de <https://CRAN.R-project.org/package=ggpmisc>
- Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid"Graphics*. Recuperado de <https://CRAN.R-project.org/package=gridExtra>
- Benesty, J., & Chen, Y. and 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 en 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>
- Calderón, C. E. (2012). Estadística para Estudiantes de Administración de Empresas de la Universidad Nacional del Callao. *Editorial San Marcos, 2da Edición, Lima Perú*. Recuperado de https://unac.edu.pe/documentos/organizacion/vri/cdcitra/Informes_Finales_Investigacion/IF_JUNIO_2012/IF_CALDERON%20OTOYA_FCA/capitulo%208.pdf
- 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>
- Cardona, G. ;. F., D.; Escané. (2013). Mortalidad por causas externas: Un problema de salud pública. Argentina, Chile y Colombia. 2000-2008. *Revista electrónica semestral*, 10(2). Recuperado de https://www.researchgate.net/publication/274885475_Mortalidad_por_causa_externas_un_problema_de_salud_publica_Argentina_Chile_y_Colombia_2000-2008
- Cochrane, J. H. (1997). *Time Series for Macroeconomics and Finance*. Recuperado de <http://econ.lse.ac.uk/staff/wdenhaan/teach/cochrane.pdf>
- De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3), 475-484. Recuperado de <https://www.sciencedirect.com/science/article/pii/S0883542506000222>

- nal of Forecasting*, 22(3), 443-473. <https://doi.org/10.1016/j.ijforecast.2006.01.001>
- Donoso, E. (2004). Desigualdad en mortalidad infantil entre las comunas de la provincia de Santia-go. *Revista médica de Chile*, 132, 461-466. Recuperado de https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0034-98872004000400008&nrm=iso
- Ellis, P. (2018). *ggseas: 'stats' for Seasonal Adjustment on the Fly with 'ggplot2'*. Recuperado de <https://CRAN.R-project.org/package=ggseas>
- 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>
- Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), 1-25. Recuperado de <https://www.jstatsoft.org/v40/i03/>
- 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. (2018a). *Forecasting: principles and practice*. Recuperado de https://books.google.co.cr/books?id=_bBhDwAAQBAJ
- Hyndman, R. J., & Athanasopoulos, G. (2018b). *Forecasting: principles and practice*. Recuperado de https://books.google.co.cr/books?id=_bBhDwAAQBAJ
- Hyndman, Rob J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3), 1-22. Recuperado de <https://www.jstatsoft.org/article/view/v027i03>
- 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. (2004). *Documento Metodológico de Defunciones Infantiles*. INEC.
- Jammalamadaka, S. R., Qiu, J., & Ning, N. (2018). *Multivariate Bayesian Structural Time Series*

- Model.* Recuperado de <https://arxiv.org/pdf/1801.03222.pdf>
- Kassambara, A. (2020). *ggsnubr: 'ggplot2' Based Publication Ready Plots.* Recuperado de <https://CRAN.R-project.org/package=ggsnubr>
- 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.*
- León, B. ; E., R.; Gallegos. (1998). Mortalidad infantil: Análisis de un decenio. *Revista Cubana de Medicina General Integral, 14*, 606-610. Recuperado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21251998000600017&nrm=iso
- Nación. (2013). Morbilidad y mortalidad en Costa Rica. *La Nacion.* Recuperado de <https://bit.ly/2xWUeXU>
- OPS. (2016). *Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud.* OMS.
- 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
- R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* Recuperado de <https://www.R-project.org/>
- Rezaee, Z., Aliabadi, S., Dorestani, A., & Rezaee, N. J. (2020). Application of Time Series Models in Business Research: Correlation, Association, Causation. *Sustainability, 12*(12), 4833.
- Rincon, M. (2000). *Métodos para proyecciones demográficas.*
- 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>
- Sargent, T. J. (1979). *Macroeconomic Theory.* Recuperado de <https://books.google.co.cr/books?id=X6u7AAAAIAAJ>
- Stoffer, D. (2020). *astsa: Applied Statistical Time Series Analysis.* Recuperado de <https://CRAN.R-project.org/package=astsa>
- Surhone, L. M., Timpledon, M. T., & Marseken, S. F. (2010). *Wold Decomposition.* Recuperado de <https://books.google.co.cr/books?id=7cSqcQAAQAAJ>
- Tadayon, M., & Iwashita, Y. (2020). *Comprehensive Analysis of Time Series Forecasting Using Neural Networks.* Recuperado de <https://arxiv.org/pdf/2001.09547.pdf>
- Vázquez, J. (2017). En 5 años flotilla de motos se disparó en un 189 por ciento. *CR Hoy.* Recuperado de <https://bit.ly/2QmQQfE>
- Villalón, S. ; O., G.; Vera. (2006). *Tabla de vida por método de mortalidad óptima.* INE.

-
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Recuperado de <https://ggplot2.tidyverse.org>
- Wickham, H., & Bryan, J. (2019). *readxl: Read Excel Files*. Recuperado de <https://CRAN.R-project.org/package=readxl>
- Wickham, H., François, R., Henry, L., & Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. Recuperado de <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., & Henry, L. (2019). *tidyR: Tidy Messy Data*. Recuperado de <https://CRAN.R-project.org/package=tidyr>
- 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>
- Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in R. En V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Computational Research*. Recuperado de <http://www.crcpress.com/product/isbn/9781466561595>
- Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.
- Zhu, H. (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. Recuperado de <https://CRAN.R-project.org/package=kableExtra>