

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

LA SOBREPARAMETRIZACIÓN EN EL ARIMA: UNA APLICACIÓN A DATOS
COSTARRICENSES

Examen de candidatura sometido 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

2022

“Este examen de candidatura fue aceptado 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. Flor Isabel Jiménez Segura
Decano Sistema de Estudios de Posgrado

MSc. Óscar Centeno Mora
Director de Tesis

Ph.D. Gilbert Brenes Camacho
Lector

Ph.D. Shu Wei Chou Chen.
Lector

MSc. Johnny Madrigal Pana
Director Programa de Posgrado en Estadística

César Andrés Gamboa Sanabria
Candidato

Índice

1 INTRODUCCIÓN	1
1.1 Antecedentes	1
1.2 El problema	2
1.3 Objetivos del estudio	2
1.4 Justificación del estudio	3
1.5 Organización del estudio	4
2 MARCO TEÓRICO	5
2.1 Componentes de una serie cronológica	6
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	12
2.3 Identificación del modelo	14
2.4 Modelos Autorregresivos Integrados de Medias Móviles	16
2.4.1 Ecuación de Wold	16
2.4.2 Metodología Box-Jenkins	17
2.4.3 Modelos Autorregresivos	18
2.4.4 Modelos de Medias Móviles	18
2.4.5 Modelos ARIMA	19
2.5 Los autocorrelogramas	20
2.6 La sobreparametrización y el análisis combinatorio	25
3 METODOLOGÍA	27
3.1 Materiales	27
3.1.1 Tasa de mortalidad infantil interanual	27
3.1.2 Mortalidad por causa externa	31
3.1.3 Incentivos salariales del sector público	34
3.1.4 Intereses y comisiones del sector público	36
3.1.5 Herramientas analíticas y procedimiento de simulación	39
3.2 Métodos	42
3.2.1 Análisis exploratorio	42
3.2.2 Partición de los datos	42
3.2.3 Estimación del mejor modelo según la función auto.arima()	42
3.2.4 Estimación del mejor modelo con sobreparametrización	43
3.2.5 Estimación de un modelo ARIMA estándar	45

3.2.6	Análisis de los errores	45
3.2.7	Pronósticos	45
3.2.8	Medidas de bondad de ajuste y de rendimiento	45
3.2.8.1	AIC	45
3.2.8.2	AICc	45
3.2.8.3	BIC	46
3.2.8.4	MAE	46
3.2.8.5	MASE	46
3.2.8.6	RMSE	46
4	RESULTADOS	47
4.1	Análisis exploratorio	48
4.1.1	Datos simulados	48
4.1.1.1	ARIMA(1,0,0)	48
4.1.1.2	ARIMA(1,0,1)	48
4.1.1.3	ARIMA(2,0,3)	48
4.1.1.4	ARIMA(4,0,2)	48
4.1.1.5	ARIMA(0,0,1)(0,1,1)	48
4.1.1.6	ARIMA(2,1,4)(3,0,3)	48
4.1.2	Tasa de mortalidad infantil interanual	48
4.1.3	Mortalidad por causa externa	48
4.1.4	Incentivos salariales	48
4.1.5	Intereses y comisiones del sector público	48
4.2	Partición de los datos	48
4.3	Estimación del mejor modelo según la función auto.arima()	48
4.3.1	Datos simulados	48
4.3.2	Datos reales	48
4.4	Estimación del mejor modelo con sobreparametrización	48
4.4.1	Datos simulados	48
4.4.2	Datos reales	48
4.5	Estimación de un modelo ARIMA estándar	48
4.5.1	Datos simulados	48
4.5.2	Datos reales	48
4.6	Análisis de los errores	48
4.6.1	Datos simulados	48
4.6.1.1	Errores de los modelos estimados con auto.arima()	48
4.6.1.2	Errores de los modelos estimados con sobreparametrización	48

4.6.1.3	Errores de los modelos estimados con un modelo ARIMA estándar	48
4.6.2	Datos reales	48
4.6.2.1	Errores de los modelos estimados con auto.arima()	48
4.6.2.2	Errores de los modelos estimados con sobreparametrización	48
4.6.2.3	Errores de los modelos estimados con un modelo ARIMA estándar	48
4.7	Pronósticos	48
4.7.1	Datos simulados	48
4.7.2	Datos reales	48
4.8	Medidas de bondad de ajuste y de rendimiento	48
4.8.1	Datos simulados	48
4.8.2	Datos reales	48
5	CONCLUSIONES	48
6	ANEXOS	49
7	REFERENCIAS	51

Índice de cuadros

Índice de figuras

1	Número de matrimonios en Costa Rica para el periodo 1978-1983	7
2	Número de turistas en Costa Rica para el periodo 1991-2000	8
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	Componente aleatorio de la serie de matrimonios en Costa Rica para el periodo 1978-1983	11
6	Componente aleatorio de la serie de matrimonios en Costa Rica para el periodo 1978-1983	12
7	Número anual de graduados de la Universidad de Costa Rica para el periodo 1965-2002	21
8	Función de autocorrelación simple de la serie de graduados de la UCR	22
9	Función de autocorrelación parcial de la serie de graduados de la UCR	22
10	Serie diferenciada de graduados de la Universidad de Costa Rica para el periodo 1965-2002	23
11	Función de autocorrelación simple de la serie diferenciada de graduados de la UCR	24
12	Función de autocorrelación parcial de la serie diferenciada de graduados de la UCR	24
13	Tasa de Mortalidad Infantil Interanual 1989 - 2017	29
14	Tasa de Mortalidad Infantil Interanual 1989 - 2017 según periodos	30

15	Descomposición de la TMII en el periodo 2000 - 2017	31
16	Mortalidad por causa externa 2000 - 2017	32
17	Mortalidad por causa externa 2000 - 2017 según mes	33
18	Descomposición de las defunciones por causa externa en el periodo 2000-2017	34
19	Incentivos salariales en el sector público 2007 - 2018	35
20	Incentivos salariales en el sector público 2007 - 2018 según mes	35
21	Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018	36
22	Intereses y comisiones del sector público en el periodo 2007-2018	37
23	Intereses y comisiones del sector público en el periodo 2007-2018 según mes	38
24	Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018	39
25	Valores de referencia para la simulación de series cronológicas	41
26	Series cronológicas simuladas	41

1 INTRODUCCIÓN

1.1 Antecedentes

Estimar los valores futuros en un determinado contexto ha producido un aumento en el análisis de los datos referidos en el tiempo, conocido también como series cronológicas. Este tipo de datos se encuentra en diferentes áreas, tanto en investigación académica como en el análisis de datos para la toma de decisiones. En el campo financiero es común hablar de la devaluación del colón con respecto al dólar, cantidad de exportaciones mensuales de un determinado producto o las ventas, entre otros (Hernández, 2011). Las series cronológicas son particularmente importantes en la investigación de mercados o en las proyecciones demográficas; de manera conjunta apoyan la toma de decisiones para la aprobación presupuestaria en distintas áreas.

En la actualidad, la información temporal es muy relevante: El Banco 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 (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 y

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.

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 et al., 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 despalde esa decisión.

1.3 Objetivos del estudio

El objetivo general de la presente investigación es proponer un algoritmo alternativo más exhaustivo para la selección de modelos ARIMA mediante la sobreparametrización de los términos de la ecuación del ARIMA.

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

Para lograr esto, se pretende:

1. Generar los escenarios de estimación de los distintos modelos ARIMA mediante permutaciones de los términos (p, d, q) y (P, D, Q) para la estimación de los posibles procesos que gobiernan una determinada serie temporal.
2. Aplicar diversos métodos de validación en la estimación de procesos que gobiernan la serie cronológica.
3. Contrastar la precisión de la estimación así como la generación de pronósticos con otros métodos similares, aplicados en datos costarricenses.
4. Integrar el desarrollo de la metodología de análisis de series temporales en una librería del lenguaje estadístico R.

1.4 Justificación del estudio

El uso correcto de las series temporales se puede apreciar en distintos contextos. 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

Las 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 et al., 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. 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. De esta manera, una serie de tiempo puede considerarse una muestra aleatoria, pues para un determinado tiempo t , que se considera el momento actual, la serie tiene tres momentos: el pasado, que son los rezagos denotados como $Y_{t-1}, Y_{t-2}, \dots, Y_{t-1}$, el momento presente, denotado como Y_t , y los pronósticos, denotados como $Y_{t+1}, Y_{t+2}, \dots, Y_{t+h}$; así, una serie temporal univariada, con lapsos equidistantes entre los tiempos, puede representarse como $Y_{t-k}, \dots, Y_{t-2}, Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}, \dots, Y_{t+h}$.

A partir de lo anterior, la serie cronológica se compone de dos partes: la estocástica, que contiene una parte conocida (sistématica) y susceptible de predecir y de una parte totalmente desconocida o aleatoria; y una parte determinística, que representa una ecuación matemática sin error, dado que no posee más que ese componente determinístico, se trata de una variable que está determinada o fija y que no cambia de una muestra a otra. De esta manera, puede concluirse que una serie cronológica cuenta con dos características fundamentales: Los valores se encuentran ordenados cronológicamente y, además, existe una dependencia o correlación entre los valores de dicha serie de tiempo; de no presentarse estas dos condiciones, no se estaría en presencia de una serie cronológica. Así, puede decirse que las series de tiempo se enfocan en tres grandes objetivos que serán detallados en secciones posteriores: la descripción de la serie, la adecuación de un modelo o técnica estocástica, y el pronóstico para hasta un horizonte h determinado; el análisis de la serie debe preguntarse sobre el tipo de serie que se está analizando, el tipo de datos y el periodo de referencia utilizado para ajustar el modelo que servirá para realizar los pronósticos.

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 et al., 2018), y finalmente, los procesos Autorregresivos Integrados de Medias Móviles o ARIMA por sus siglas en inglés (Box et al., 1994), siendo estos últimos el foco de interés en este estudio. Los modelos ARIMA se enfocan en considerar las relaciones pasadas de un serie cronológica asociando los datos de las correlaciones totales y parciales. La forma de abordar una serie de tiempo utilizando los modelos ARIMA consiste, de forma muy general, en hacer una descripción de la serie para corroborar que se trate de una serie estacionaria y, de no serlo, someterla a procesos matemáticos para asegurar esta condición. Posteriormente, se realiza una identificación del posible proceso que gobierna la serie cronológica para luego estimar el modelo del orden seleccionado, sometiendo este a diversas pruebas de bondad de ajuste y rendimiento para finalmente verificar la calidad de los pronósticos obtenidos. El sustento teórico de cada una de estas será discutido a lo largo de este capítulo, que se compone de seis apartados. El primer apartado abarca los cuatro componentes de una serie cronológica. 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 selección de casos.

2.1 Componentes de una serie cronológica

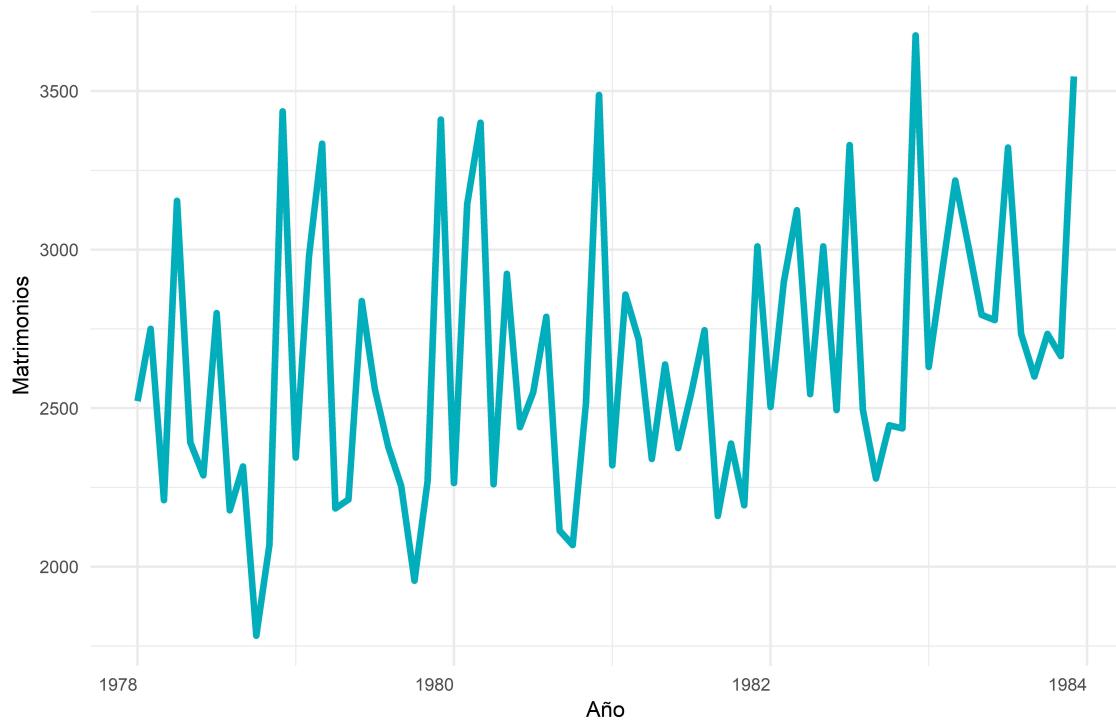
En el análisis de series cronológicas existen dos grandes corrientes de estudio: Los componentes inherentes a la serie cronológica y el estudio de las autocorrelaciones. Según el primer enfoque, de acuerdo con Hernández (2011), las series cronológicas poseen tres componentes principales: Tendencia-ciclos, 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.

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

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

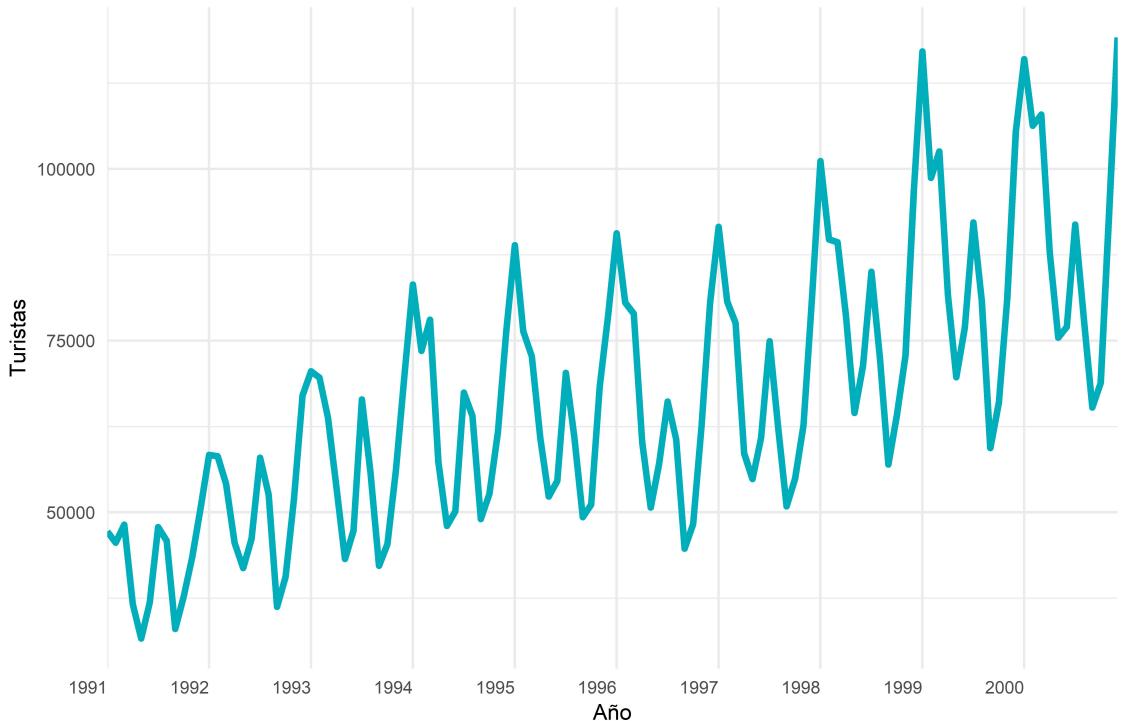
Donde Y es la serie cronológica, T es la tendencia-ciclo, S es la parte estacional, I la parte irregular o aleatoria, y t es el momento en el tiempo. Esta perspectiva clásica del análisis de series de tiempo permite realizar un análisis descriptivo del comportamiento de la serie en cuestión; cada una 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



Fuente: Elaboración propia a partir de datos de la Unidad de Estadísticas Demográficas - INEC Costa Rica.

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



Fuente: Elaboración propia a partir de datos de la Unidad de Estadísticas Demográficas - INEC Costa Rica.

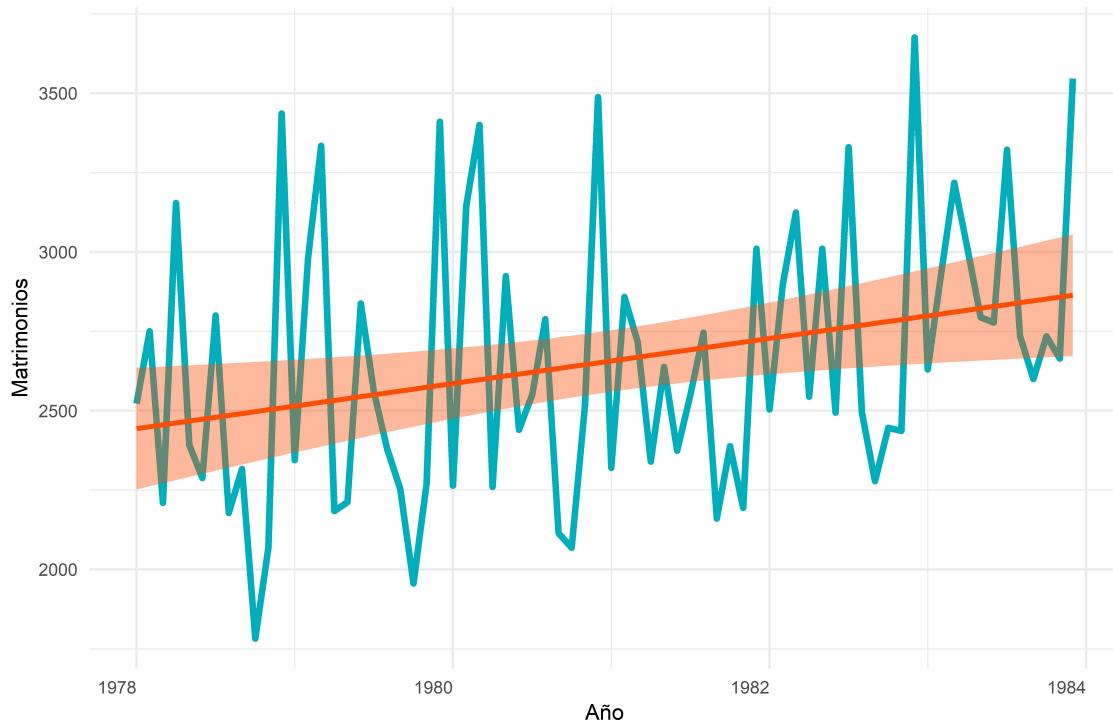
2.1.1 La tendencia-ciclo

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

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

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.

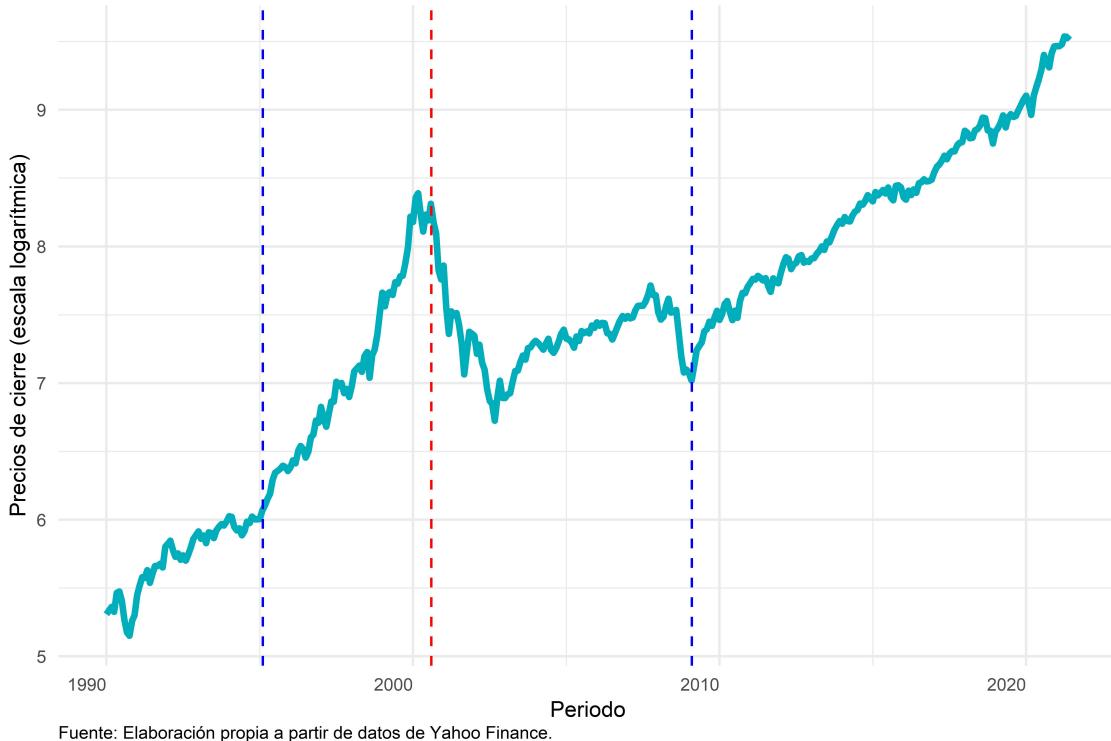
Figura 3: Tendencia del número de matrimonios en Costa Rica para el periodo 1978-1983



Fuente: Elaboración propia a partir de datos de la Unidad de Estadísticas Demográficas - INEC Costa Rica.

Del informe elaborado también por [Calderón \(2012\)](#) se desprende que los periodos 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  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



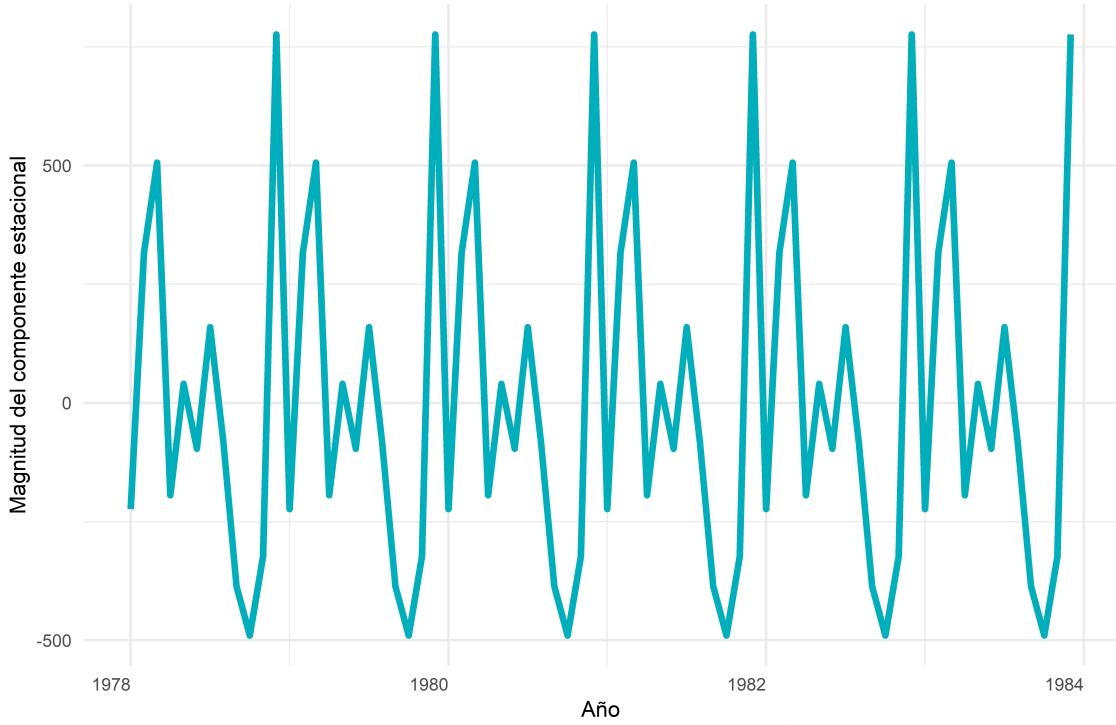
2.1.2 Componentes estacionales

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:

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

donde m representa la cantidad de periodos y s la frecuencia estacional. Gráficamente, el componente estacional se muestra en la figura 5.

Figura 5: Componente aleatorio de la serie de matrimonios en Costa Rica para el periodo 1978-1983

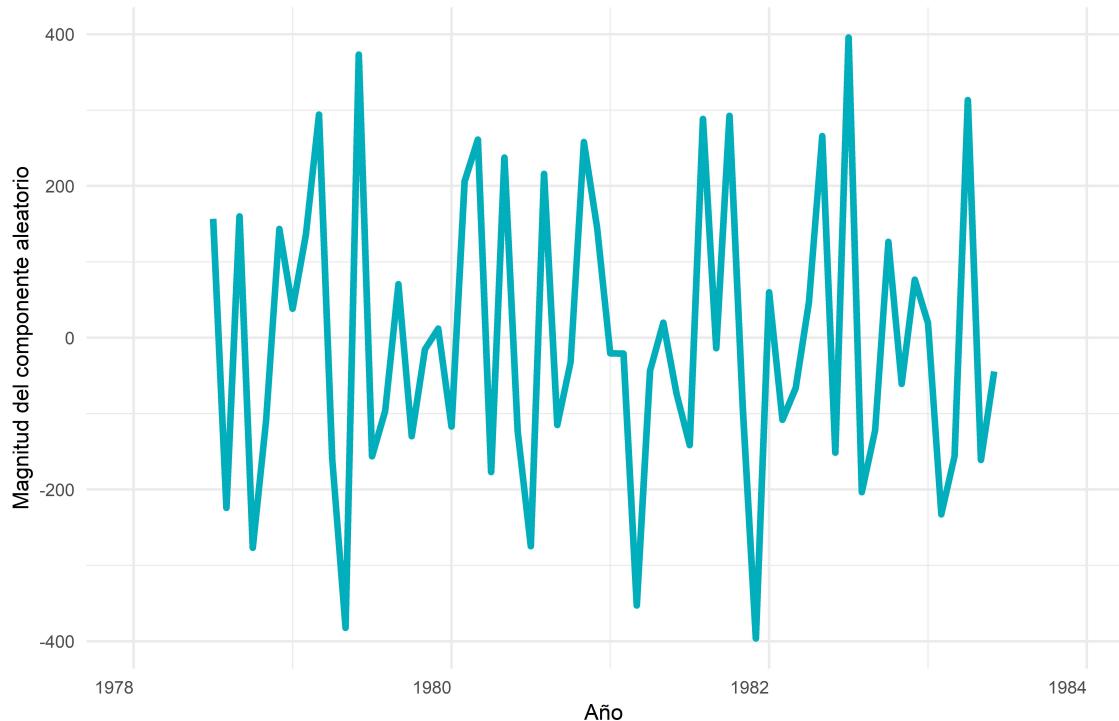


2.1.3 Componente irregular

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. Visualmente, la magnitud del componente aleatorio se muestra en la figura 6

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

Figura 6: Componente aleatorio de la serie de matrimonios en Costa Rica para el periodo 1978-1983



Fuente: Elaboración propia a partir de datos de la Unidad de Estadísticas Demográficas - INEC Costa Rica.

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](#)).

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. Este componente no determinístico es lo que define a un procesos estocástico (aleatorio) como un conjunto de variables aleatorias ordenadas en el tiempo ([Elmabrouk, n.d.](#)). De acuerdo con [Ramírez \(2007\)](#), una forma de definir un proceso estocástico Y_t es mediante los momentos poblacionales de primer y segundo orden tal y como se define en la ecuación 6.



$$Y_t : \begin{cases} E(Y_t) = \mu_t, \forall t \\ V(Y_t) = \sigma_t^2, \forall t \\ COV(T_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)], \forall t, s \end{cases} \quad (6)$$

Lo anterior implica que la media, la varianza y la covarianza dependen del tiempo. De esto se desprende además que existen dos tipos de procesos estocásticos: estacionarios y no estacionarios. 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. De acuerdo con ([Elmabrouk, n.d.](#)), se dice que una serie cronológica Y_t es fuertemente estacionaria si satisface las siguientes tres condiciones:



1. La media de Y_t se mantiene constante e el tiempo.
2. La variabilidad de Y_t se mantiene constante en el tiempo y además es finita.
3. La covarianza entre Y_t y Y_{t-k} únicamente depende de la distancia entre t y $t-k$.

Un proceso estocástico es débilmente estacionario si alguna de las tres condiciones previas no se cumple, en particular la última. 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.

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.), que según [Evans & Rosenthal \(2005\)](#), un conjunto de variables aleatorias Y_1, \dots, Y_n son independientes e idénticamente distribuidas si el conjunto es independiente y además cada una de las n variables sigue la misma distribución, que usualmente se define como una distribución aproximadamente Normal, con una media y variancia dadas. Este supuesto puede dividirse según el tipo de variable aleatoria:

- 1.: Si el conjunto de variables Y_1, \dots, Y_n pertenecen a una distribución discreta, cada función de probabilidad es idéntica, de manera que $p_{y_1}(y) = p_{y_2}(y) = \dots = p_{y_n}(y) \equiv p(y)$, y además $p_{y_1, \dots, y_n}(y_1, \dots, y_n) = p_{y_1}(y_1)p_{y_2}(y_2) \cdots p_{y_n}(y_n) = p(y_1)p(y_2) \cdots p(y_n)$.
- 2.: Si el conjunto de variables Y_1, \dots, Y_n pertenecen a una distribución continua, cada función de probabilidad es idéntica, de manera que $f_{y_1}(y) = f_{y_2}(y) = \dots = f_{y_n}(y) \equiv f(y)$, y además $f_{y_1, \dots, y_n}(y_1, \dots, y_n) = f_{y_1}(y_1)f_{y_2}(y_2) \cdots f_{y_n}(y_n) = f(y_1)f(y_2) \cdots f(y_n)$.

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 serie de tiempo no es exactamente, i.i.d., sino que siguen un patrón medianamente regular en el largo plazo.

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 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 $ARIMA$ ([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 7.

$$r_{X,Y} = \frac{E(XY)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (7)$$

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 parcial busca medir la asociación lineal entre las observaciones Y_t y Y_{t-k} , es decir, la correlación entre dos observaciones distintas separadas por k períodos, descartando los efectos de los rezagos $1, 2, \dots, k-1$; esta correlación puede obtenerse a partir de la ecuación 7, que al adaptarse a dos observaciones de la misma serie cronológica se obtiene el

⁵ $PACF$ por sus siglas en inglés

resultado de ecuación 8.

$$r_{Y_t, Y_{t-k}} = \frac{E(Y_t Y_{t-k})}{\sigma_{Y_t} \sigma_{Y_{t-k}}} = \frac{\sum_{i=1}^n (Y_{ti} - \bar{Y}_t) (Y_{t-k_i} - \bar{Y}_{t-k})}{\sqrt{\sum_{i=1}^n (Y_{ti} - \bar{Y}_t)^2 \sum_{i=1}^n (Y_{t-k_i} - \bar{Y}_{t-k})^2}} \quad (8)$$

De lo anterior se deduce entonces que unas k observaciones previas pueden utilizarse para obtener el valor de la serie cronológica en el momento t , como muestra la ecuación 9.

$$y_t = \phi_{k1} y_{t-1} + \phi_{k2} y_{t-2} + \cdots + \phi_{kk} y_{t-k} + u_t, k = 1, 2, \dots, K \quad (9)$$

Los valores de cada término ϕ_{kk} , asumiendo que pertenecen a un proceso estacionario, suelen estimarse mediante la ecuación de Yule-Walker (Brockwell & Davis, 2009), cuya forma más general se muestra en la ecuación 10.

$$\gamma_i = E[\phi_{k1} y_{t-1} y_{t-i} + \phi_{k2} y_{t-2} y_{t-i} + \cdots + \phi_{kn} y_{t-n} y_{t-i} + u_t y_{t-i}] = \phi_{k1} \gamma_{i-1} + \phi_{k2} \gamma_{i-2} + \cdots + \phi_{kn} \gamma_{n-i} \quad (10)$$

Al considerar la ecuación en términos de la función de autocorrelación se obtiene lo siguiente:

$$\rho_i = \phi_{k1} \rho_{i-1} + \phi_{k2} \rho_{i-2} + \cdots + \phi_{kn} \rho_{n-i} + \cdots \quad (11)$$

Alternando los distintos valores de k a partir de la ecuación 11, se obtiene el sistema de ecuaciones mostrado en 12.

$$\begin{aligned} \rho_1 &= \phi_{k1} + \phi_{k2} \rho_1 + \cdots + \phi_{kn} \rho_{n-1} + \cdots \\ \rho_2 &= \phi_{k1} \rho_1 + \phi_{k2} + \cdots + \phi_{kn} \rho_{n-2} + \cdots \\ \rho_3 &= \phi_{k1} \rho_2 + \phi_{k2} \rho_1 + \cdots + \phi_{kn} \rho_{n-3} + \cdots \\ &\vdots \\ \rho_k &= \phi_{k1} \rho_{k-1} + \phi_{k2} \rho_{k-2} + \cdots + \phi_{kn} \rho_{n-k} + \cdots \end{aligned} \quad (12)$$

Como resultado del sistema de ecuaciones mostrado en 12, es posible hacer un replanteamiento en forma de un sistema matricial del cual las autocorelaciones parciales pueden obtenerse a partir del despeje del vector Φ en 13.

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} \quad (13)$$

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 et al. \(2013\)](#) menciona entre ellas el *MAE*, *MAPE*, *RMSE*, *MASE*, *AIC*, *AICc* y el *BIC*.

2.4 Modelos Autorregresivos Integrados de Medias Móviles

Hay dos grandes grupos de modelos lineales de series cronológicas: Los modelos Autorregresivos (AR) ([Lee, n.d.](#)) y los modelos de Medias Móviles (MA) ([Box et al., 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 et al., 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.

2.4.1 Ecuación de Wold

Según [Sargent \(1979\)](#), cualquier proceso estacionario puede ser representado mediante la ecuación 14:

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t \quad (14)$$

donde $\forall \psi_j \in \mathbb{R}$, $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, y ε_t representa un ruido  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 14, 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 14 puede sustuirse por:

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + m \quad \blacksquare \quad (15)$$

y de 15 puede verificarse que,

$$E(x_t) = E \left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + m \right) = \sum_{j=0}^{\infty} \psi_j E(\varepsilon_{t-j}) + m = m \quad (16)$$

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:

$$\hat{x}_{T+h} = \sum_{j=1}^{\infty} \psi_j \varepsilon_{T-j+1} \quad \blacksquare \quad (17)$$

De la ecuación 17 se desprende que el error de previsión asociado está dado por:

$$x_{T+h} - \hat{x}_{T+h} = \sum_{j=1}^{\infty} \psi_j \varepsilon_{T-h+1} \quad (18)$$

De esta manera, la ecuación de Wold se convierte en una representación base para representar a una serie cronológica que está gobernada por un determinado proceso, y que al no ser conocido, resulta necesario contar con una herramienta para su aproximación.

2.4.2 Metodología Box-Jenkins

La combinación de un $AR(p)$ y un $MA(q)$, descritos en las ecuaciones 20 y 21 respectivamente, como se mencionó al inicio de esta sección, generan los modelos autorregresivos de medias móviles,

ARMA(p, q), representados mediante la ecuación 19.

$$y_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (19)$$

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 suy 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 et al., 1994).

2.4.3 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 (n.d.) emplean la notación de la ecuación 20.

$$y_t = c + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t \quad (20)$$

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.

2.4.4 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 21.

$$y_t = \mu + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (21)$$

Donde μ representa el valor medio de la serie cronológica y cada valor de θ_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](#)).

2.4.5 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 22.

$$\varphi(L)(1 - L)^d y_t = \theta(L)\varepsilon_t \left(1 - \sum_{i=1}^p \varphi_i L^i\right) (1 - L)^d y_t = \left(1 + \sum_{j=1}^q \theta_j L^j\right) \varepsilon_t \quad (22)$$

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 et al. \(1994\)](#), lo anterior puede generalizarse aún más al considerar los efectos estacionales de la serie cronológica. Si se considera una serie cronológica con observaciones 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 et al. \(1994\)](#) 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 adecue 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.

2.5 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:

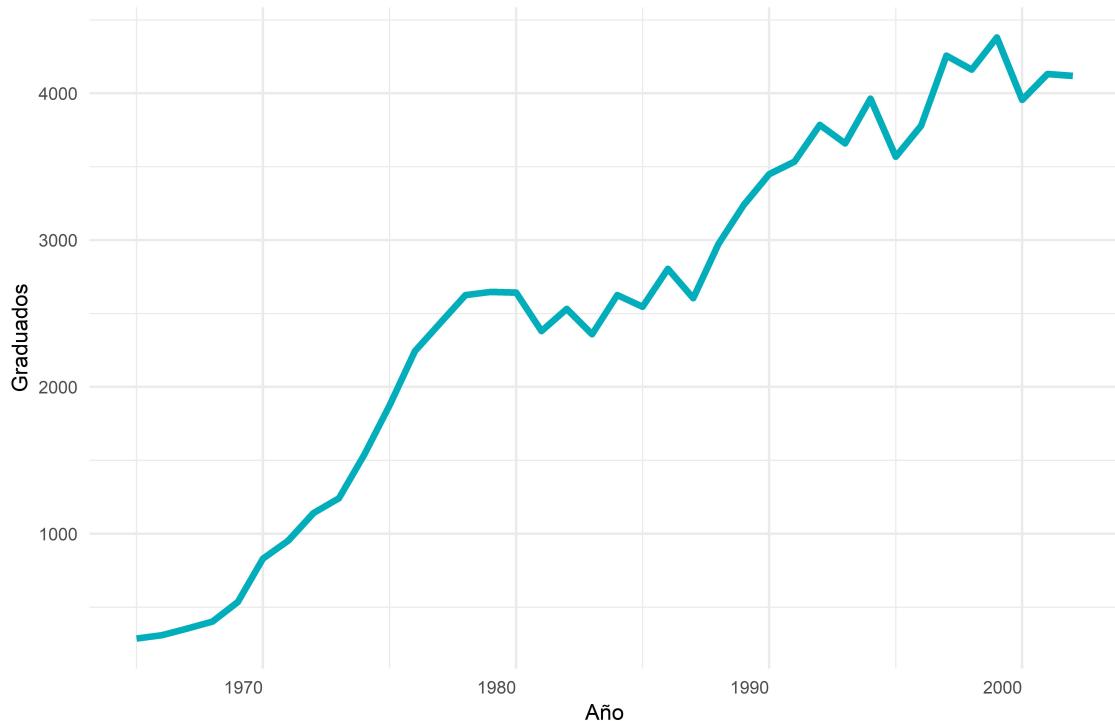
- Si la serie posee autocorrelaciones positivas en un amplio número de rezagos, entonces es posible que se requiera un orden más alto en el valor de d .
- Si la autocorrelación en $t - 1$ es menor o igual a cero, o si las autocorrelaciones resultan ser muy bajas y sin seguir algún patrón en particular, 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.

Para ejemplificar el uso de los autocorrelogramas en la identificación de modelos, se presenta en la figura 7 la serie cronológica expuesta por Hernández (2011) de graduados de la Universidad de Costa Rica (UCR) para el periodo 1965-2002.

Figura 7: Número anual de graduados de la Universidad de Costa Rica para el periodo 1965-2002



Fuente: Introducción a las Series Cronológicas, Óscar Hernández.

Tal y como menciona el autor, la serie cronológica posee una clara tendencia creciente a lo largo del tiempo, lo cual sugiere que no se trata de una serie estacionaria. Esto se confirma al analizar las funciones de autocorrelación simple y parcial de la serie cronológica en las figuras 8 y 9; pues la función de autocorrelación no cae rápidamente a cero, sino que posee un descenso más pausado.

Figura 8: Función de autocorrelación simple de la serie de graduados de la UCR

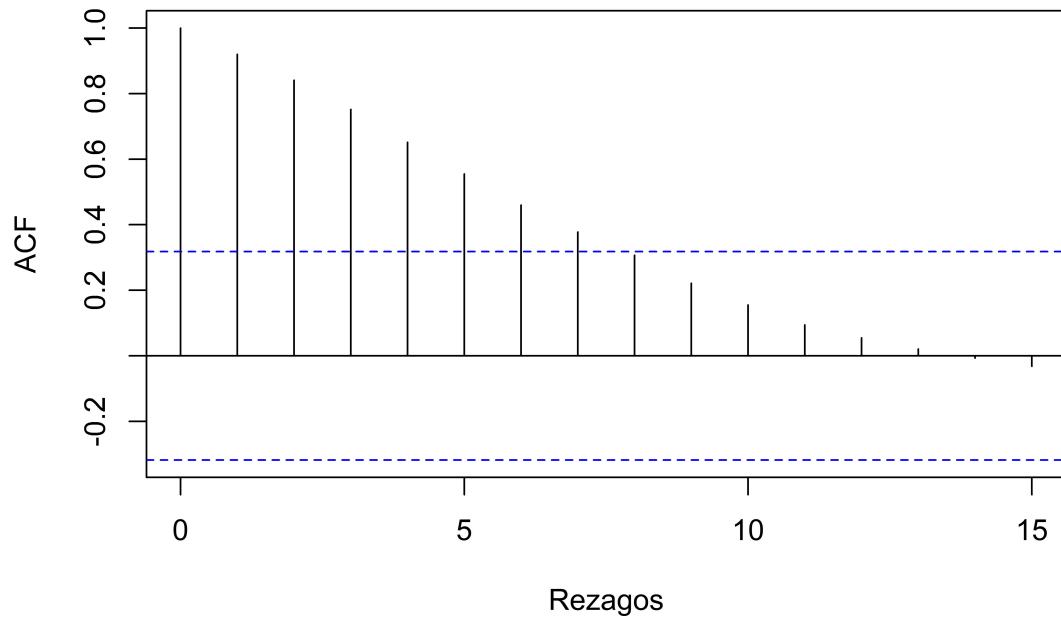
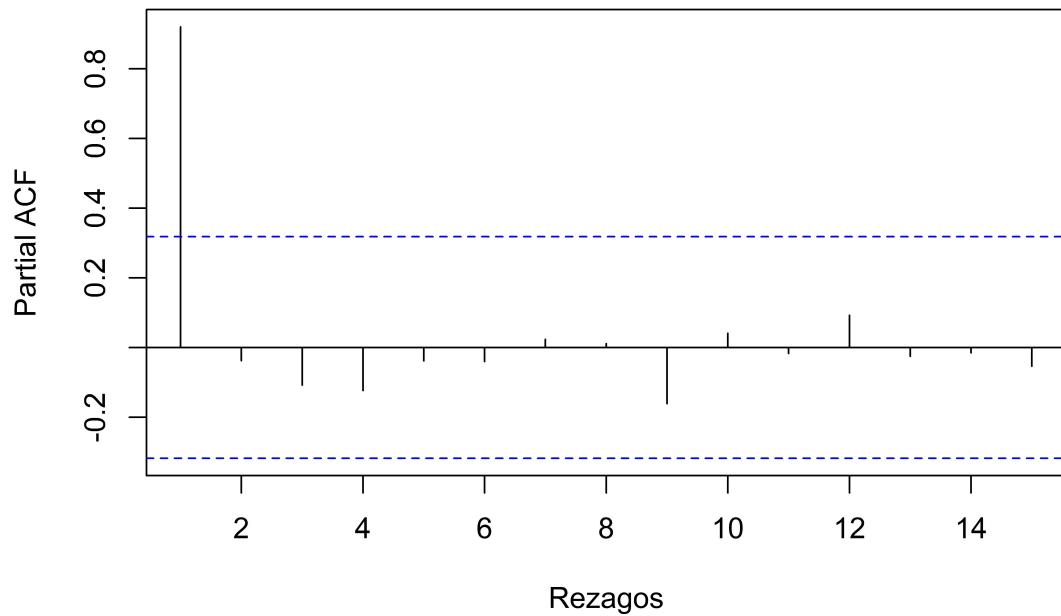


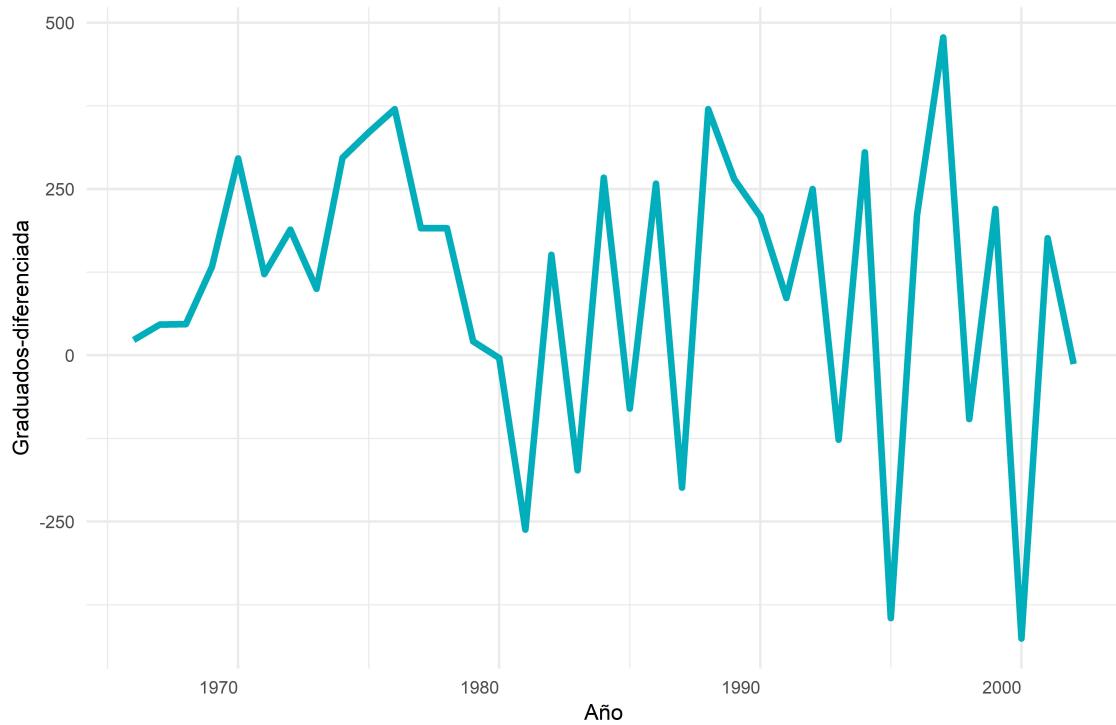
Figura 9: Función de autocorrelación parcial de la serie de graduados de la UCR



Dado que la serie mostrada no es estacionaria, es posible aplicar una diferenciación para hacerla cumplir esta condición, tal y como se muestra en la figura 10. Al analizar la figura 11 se observa

cómo la función de autocorrelación cae rápidamente a cero, lo cual confirma que se posee una serie estacionaria. Posteriormente, para intentar identificar el proceso que gobierna la serie cronológica, puede verse que hay dos barras en la figura 12 y que además la función de autocorrelación de la figura 11 cae rápidamente hacia cero, lo cual sugiere que se está en presencia de un modelo autorregresivo de orden 2.

Figura 10: Serie diferenciada de graduados de la Universidad de Costa Rica para el periodo 1965-2002



Fuente: Introducción a las Series Cronológicas, Óscar Hernández.

Figura 11: Función de autocorrelación simple de la serie diferenciada de graduados de la UCR

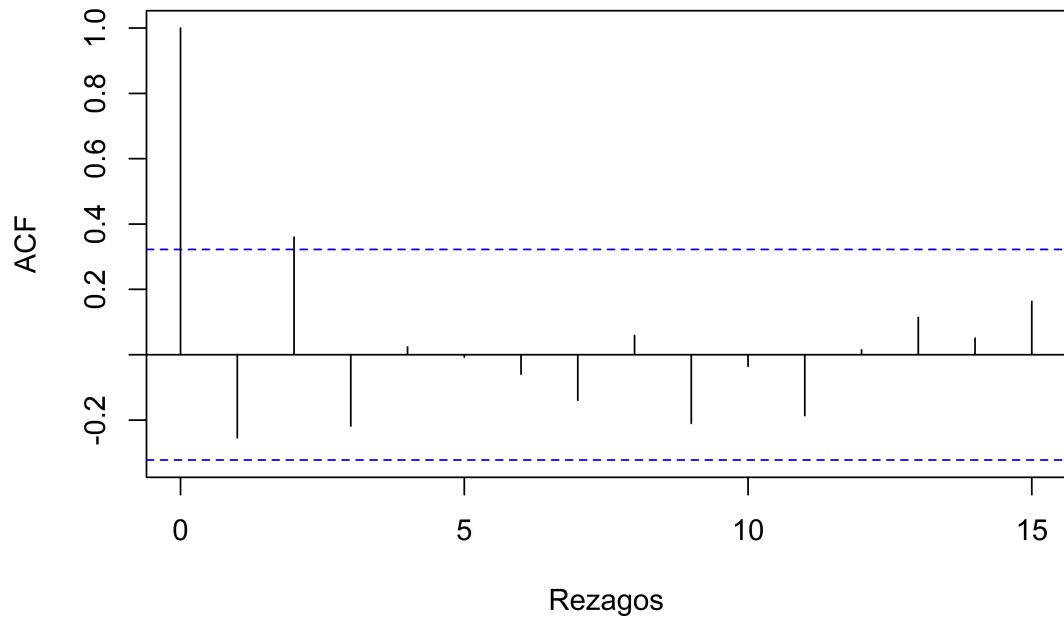
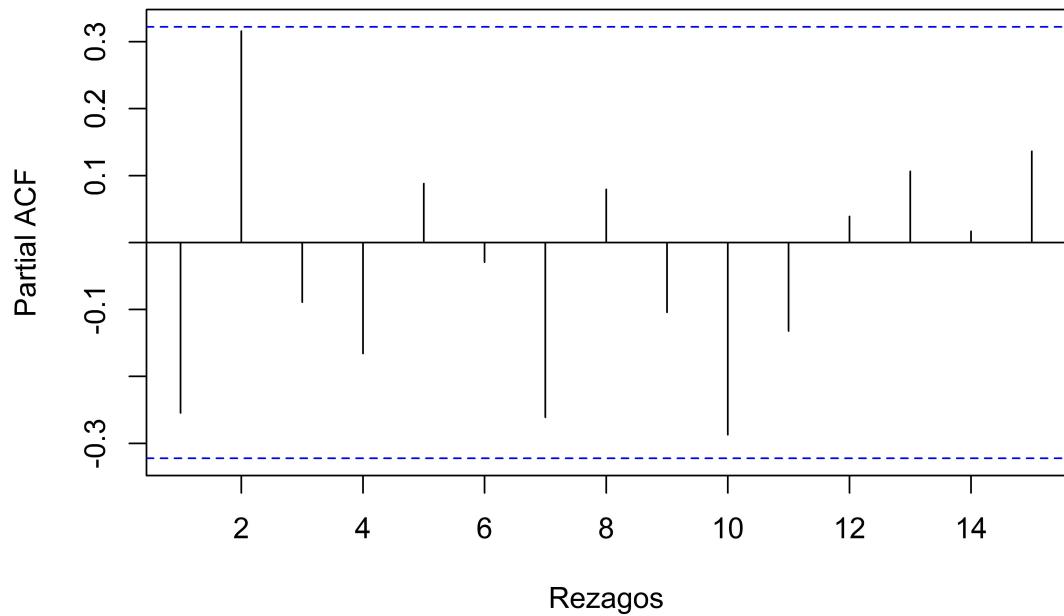


Figura 12: Función de autocorrelación parcial de la serie diferenciada de graduados de la UCR



Tener en consideración estos y otros posibles criterios para la identificación del proceso que gobierna la serie cronológica puede fácilmente volverse 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 para contrastar una gran cantidad de modelos, es decir, utilizar 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.

Es aquí donde entra en escena el análisis combinatorio, pues a partir de sus procedimientos es posible conocer la cantidad de modelos que deben ser probados. Resulta pertinente discutir dos principios fundamentales del análisis combinatorio mencionados por [Hernández \(2008\)](#): Uno es el principio de adición, el cual indica que si se tienen dos procedimientos A y B , los cuales pueden realizarse de k_A y k_B maneras, respectivamente, entonces la cantidad de maneras que se puede realizar uno u otro procedimiento es $k_A + k_B$. Por otro lado se tiene el principio de multiplicación, con el cual,

 si el procedimiento A se puede realizar de k_A formas distintas, seguido de otro procedimiento B que puede realizarse de k_B formas, entonces si a cada forma de realizar el procedimiento A se puede asociar a cualquiera de las k_B maneras de realizar el procedimiento B , entonces ambos procedimientos pueden realizarse de $k_A \cdot k_B$ formas distintas.

Es a partir de estos dos principios que pueden obtenerse la cantidad de formas distintas que pueden ordenarse m elementos tomando r elementos a la vez. Uno de ellos son las permutaciones, descritos en la ecuación 23, la cual describe la forma de calcular la cantidad de formas distintas que puede ordenarse m elementos tomando r a la vez, donde el orden sí importa, a modo de ejemplo, si se quiere saber la cantidad  de formas que pueden ordenarse las letras A, B y C tomando dos letras a la vez, se tendría que existen $\frac{3!}{(3-1)!} = 6$ formas distintas, que son AB, AC, BC, BA, CA y CB . De manera similar, se tienen las combinaciones, cuya fórmula se describe en la ecuación 24, que brinda la cantidad de maneras distintas en que pueden ordenarse m elementos tomando r a la vez donde

el orden no importa; es decir, si se desean ordenar las letras A, B y C tomando dos a la vez, se tendrían $\frac{3!}{2!(3-1)!} = 3$ formas distintas, las cuales son AB, AC y BC .

$${}_mP_r = \frac{m!}{(m-r)!} \quad (23)$$

$${}_mC_r = \frac{m!}{r!(m-r)!} \quad (24)$$

Es a partir de esto que la sobreparametrización se utiliza en conjunto con el análisis combinatorio y en particular con el método de permutaciones, pues el orden de la cantidad de coeficientes a estimar en un modelo $ARIMA(p, d, q)$ sí importa, debido a que no es lo mismo estimar un modelo $ARIMA(2, 1, 3)$ que un modelo $ARIMA(3, 1, 2)$. En las elección de modelos ARIMA normalmente los métodos tradicionales como los correlogramas u otros, no suelen abarcar un espectro más amplio de coeficientes, y esto podría representar un método de estimación que no es el mejor, por esto, la presente tesis propone una metodología que mezcla la sobreparametrización con las permutaciones con el objetivo de lograr estimar el mejor modelo $ARIMA$ de una amplia cantidad de posibles candidatos para conseguir pronósticos más precisos en comparación a los métodos tradicionales.

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.

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 los materiales a utilizar, así como los métodos, incluyendo el proceso de estimación, el procedimiento de simulación empleado para la verificación del método propuesto, y las medidas de bondad de ajuste y de precisión a utilizar. Se describe en detalle el uso de la sobreparametrización como herramienta para la generación de pronósticos de series cronológicas con temporalidades mensuales, bimestrales, 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 modelos 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 Materiales

Se describen a continuación las series cronológicas reales que servirán de insumo para poner a prueba el método propuesto.

3.1.1 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 Censos⁶ (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.

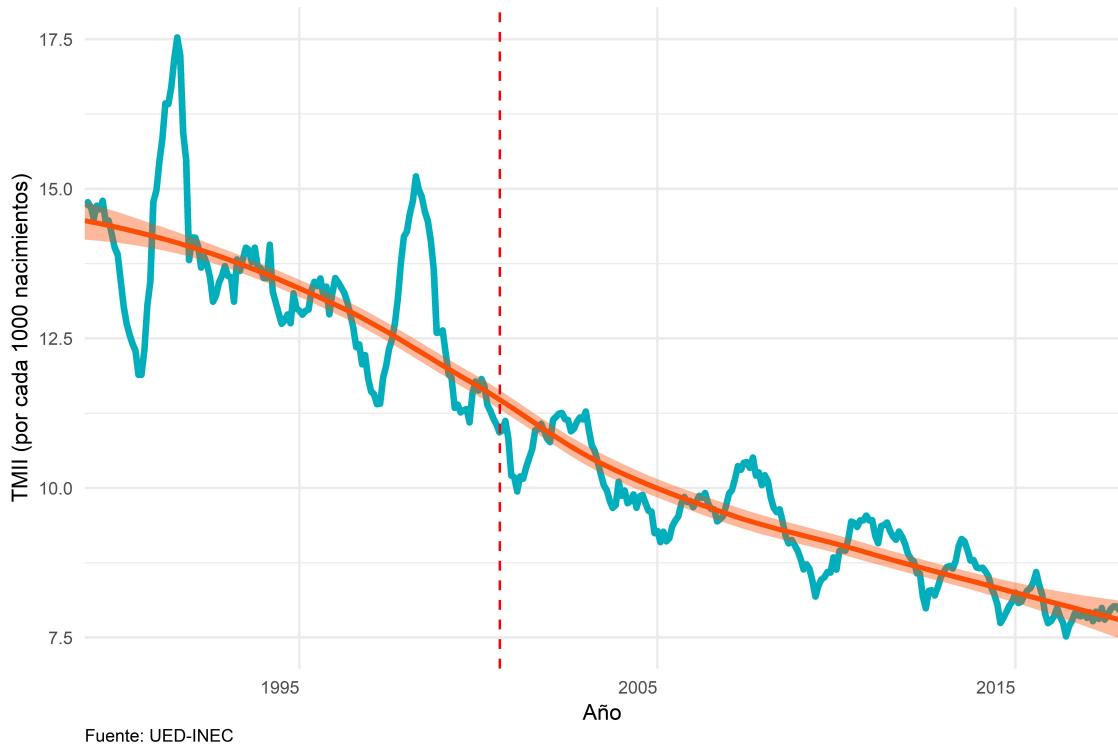
En el INEC, cada mes se publica el boletín de la TMII interanual (TMII), que analiza la TMII de un mes y los 11 meses previos para comparar los períodos correspondientes (INEC, 2004). Este apartado busca hacer un análisis de la TMII para los 12 períodos desde el año 1989 y hasta 2017, y no de manera 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 13.

⁶<http://www.inec.go.cr/>

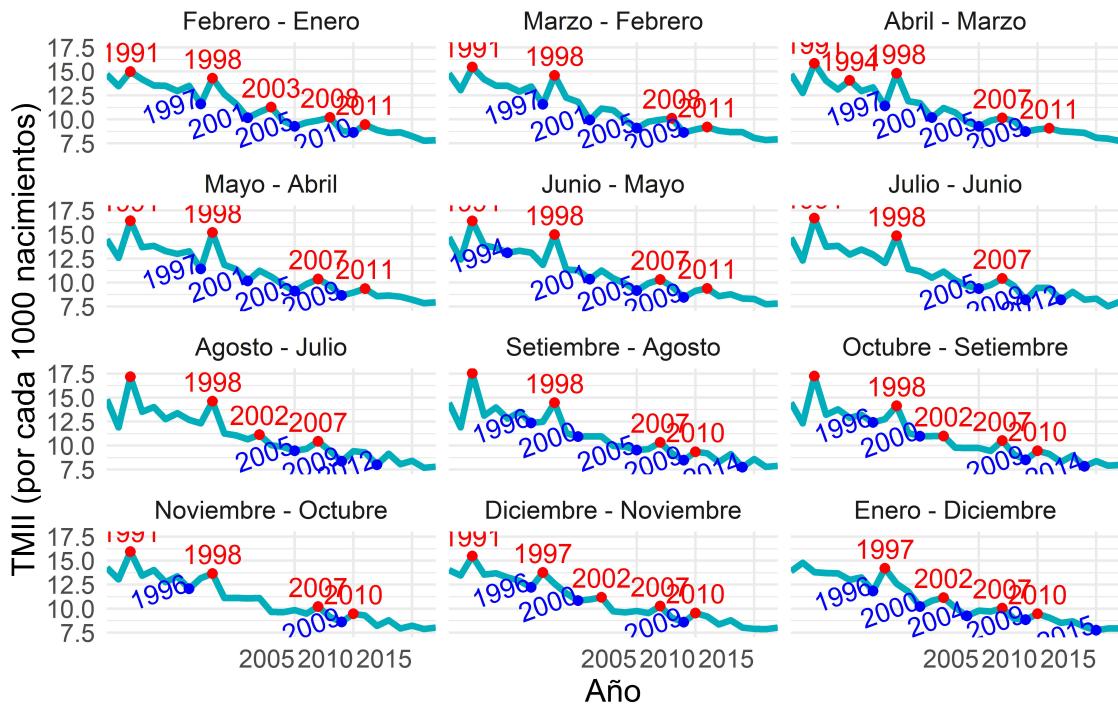
Figura 13: Tasa de Mortalidad Infantil Interanual 1989 - 2017



La serie muestra picos y valles pronunciados a lo largo de todo el periodo. 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 14 parece respaldar el supuesto de que la mortalidad no posee 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 13.

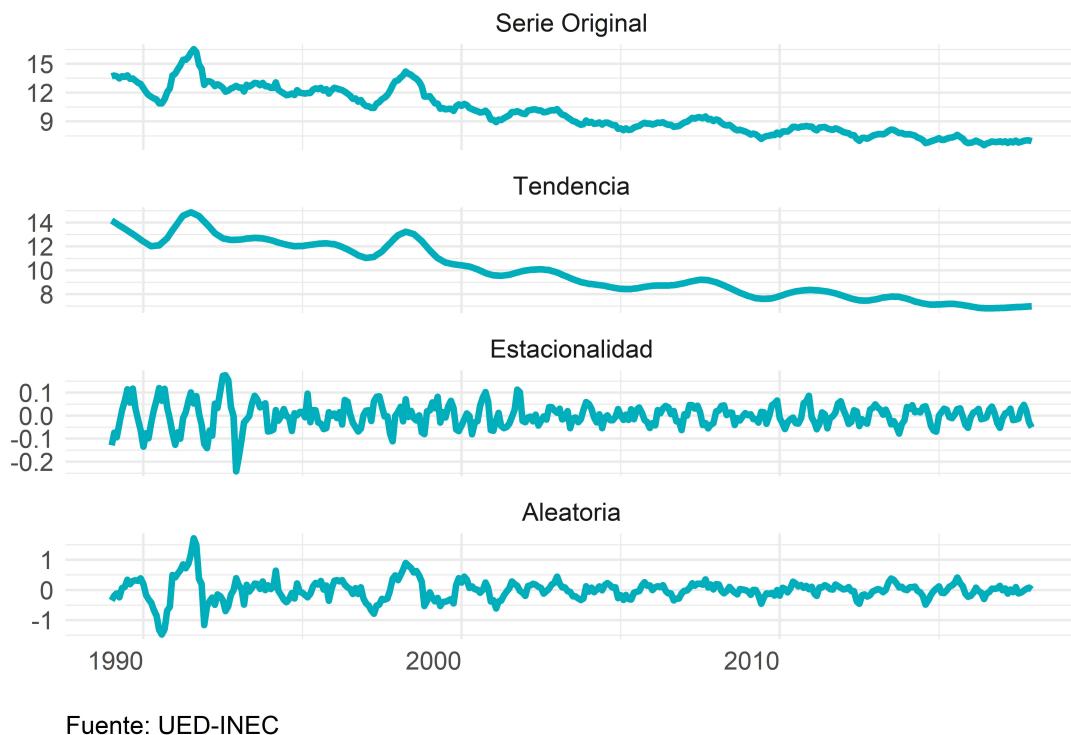
Figura 14: Tasa de Mortalidad Infantil Interanual 1989 - 2017 según períodos



Fuente: UED-INEC

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 13 pueden observarse cambios considerables en la variabilidad de la serie a lo largo del tiempo. La figura 15 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.

Figura 15: Descomposición de la TMII en el periodo 2000 - 2017



Fuente: UED-INPEC

3.1.2 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 un problema de salud pública. En base a la clasificación Internacional de Enfermedades (OPS, 2016) de la Organización Mundial de la Salud⁷, 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 categoría pertenece o aún está en investigación, tal es el caso de una persona que fallece debido a

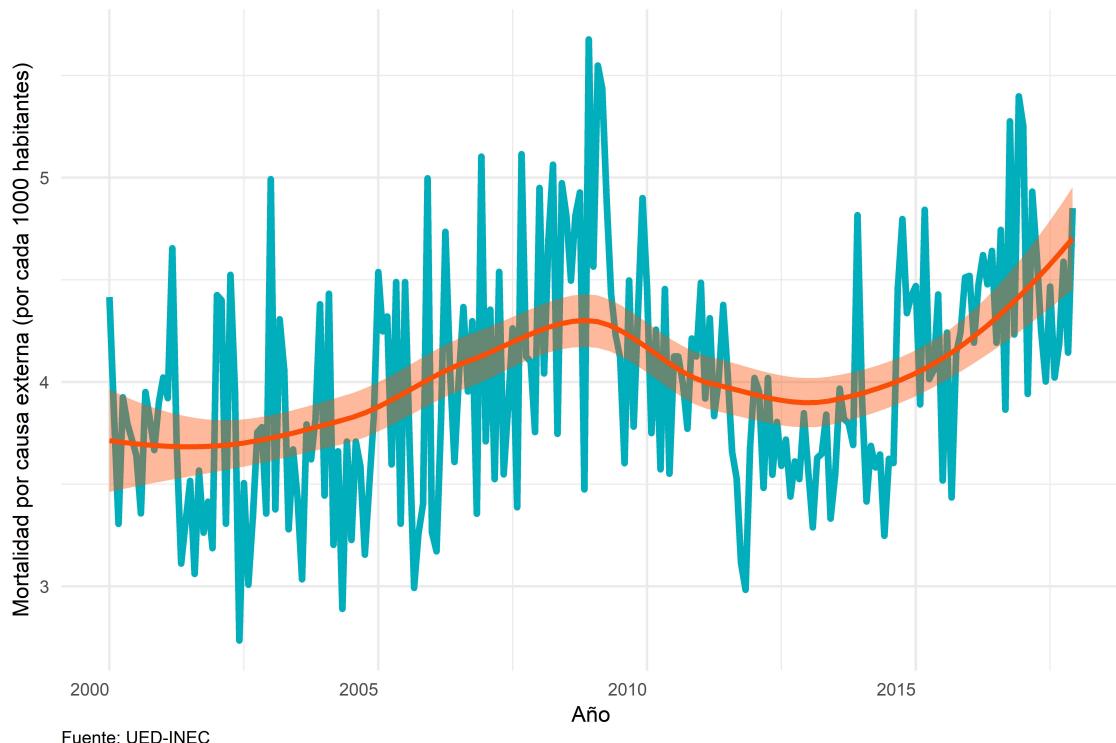
⁷<https://www.paho.org/salud-en-las-americas-2017/?lang=es>

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 16.

Figura 16: Mortalidad por causa externa 2000 - 2017

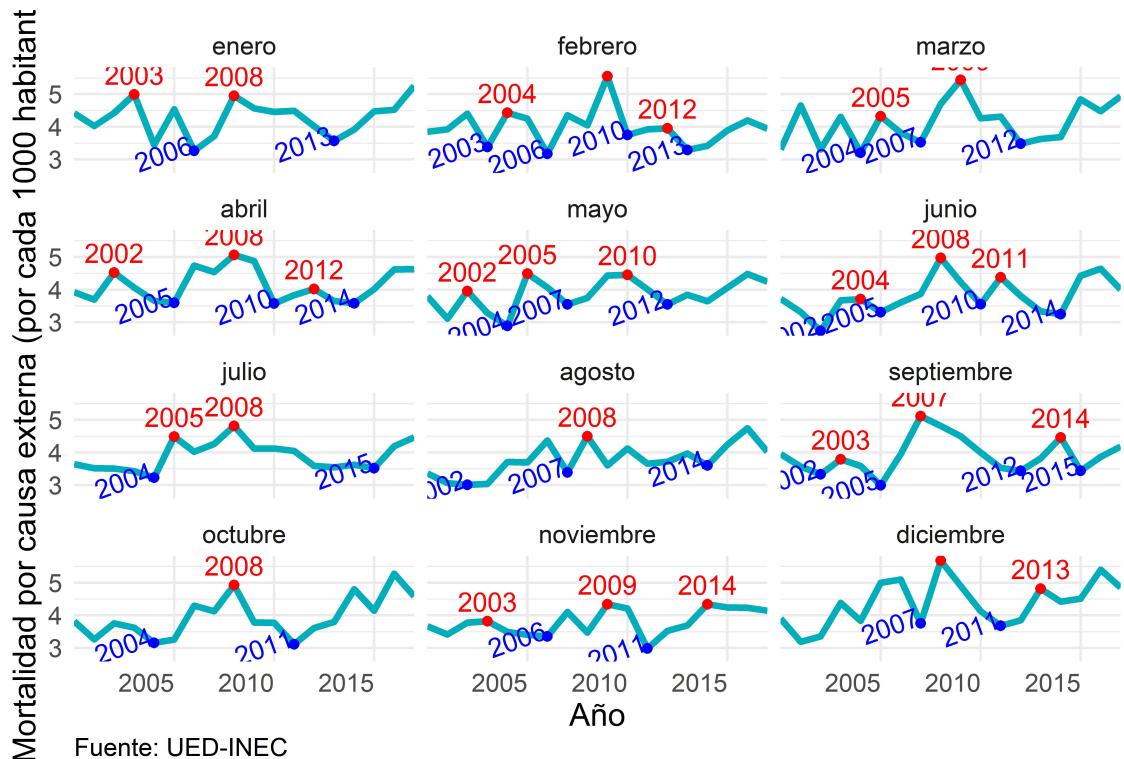


Es importante recalcar que, entre Junio del año 2012 y Diciembre del año 2017, el aumento en la tasa de cambio de la cantidad de defunciones debido a causas externas coincide con el aumento de la flotilla de motocicletas, pues en un período de cinco años esta cifra creció en un 189 % ([Vázquez, 2017](#)).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

De la figura 17 puede notarse que cada mes tiene sus picos y valles durante cada mes a lo largo del

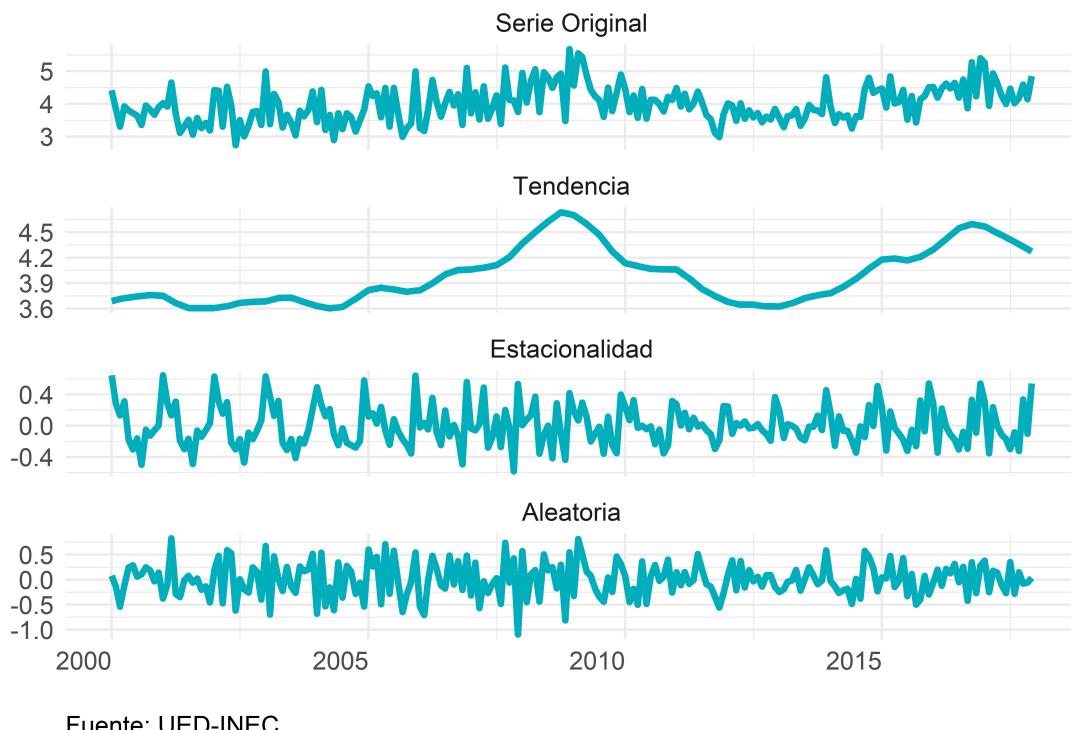
periodo, siendo los meses de Enero, Abril y Diciembre los que presentaron valores ligeramente más altos entre los años 2000 y 2017.

Figura 17: Mortalidad por causa externa 2000 - 2017 según mes



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 18 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.

Figura 18: Descomposición de las defunciones por causa externa en el periodo 2000-2017



Fuente: UED-INEC

3.1.3 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 19 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 20 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 19: Incentivos salariales en el sector público 2007 - 2018

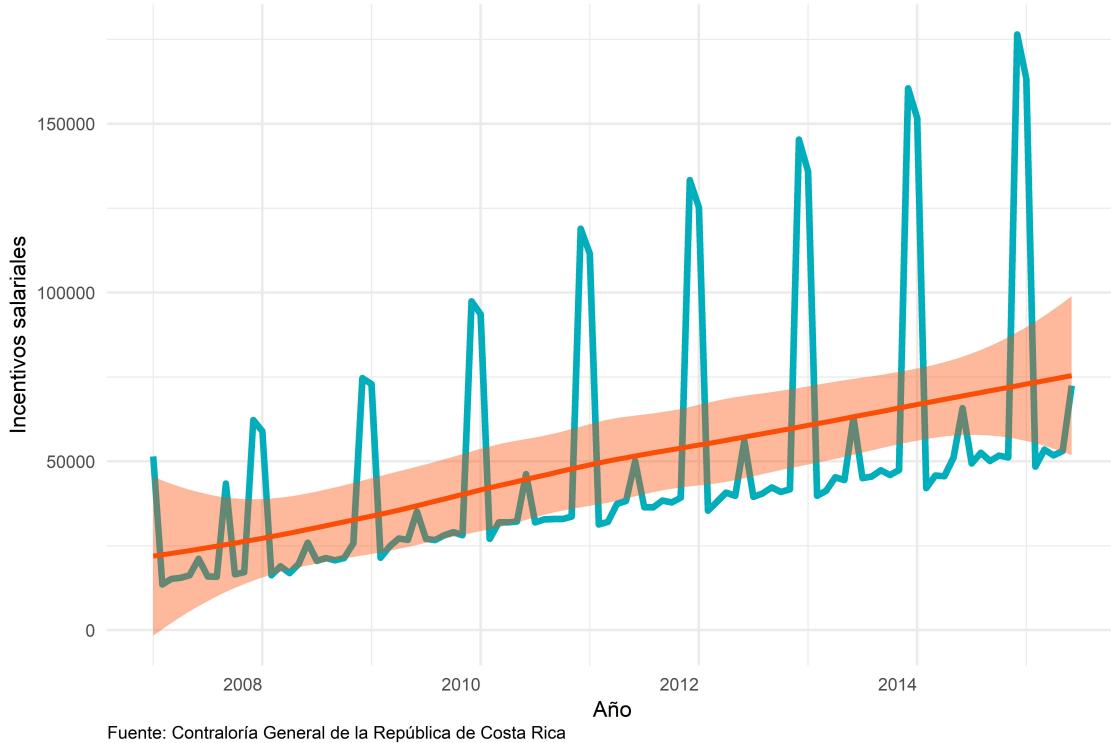
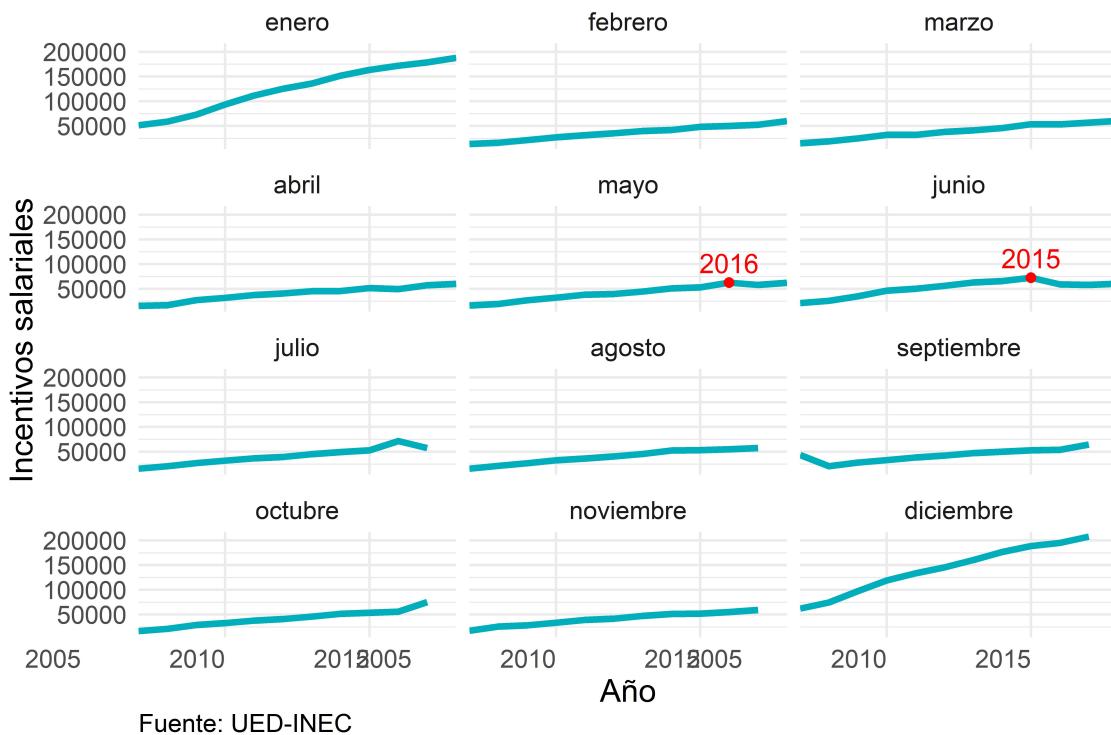


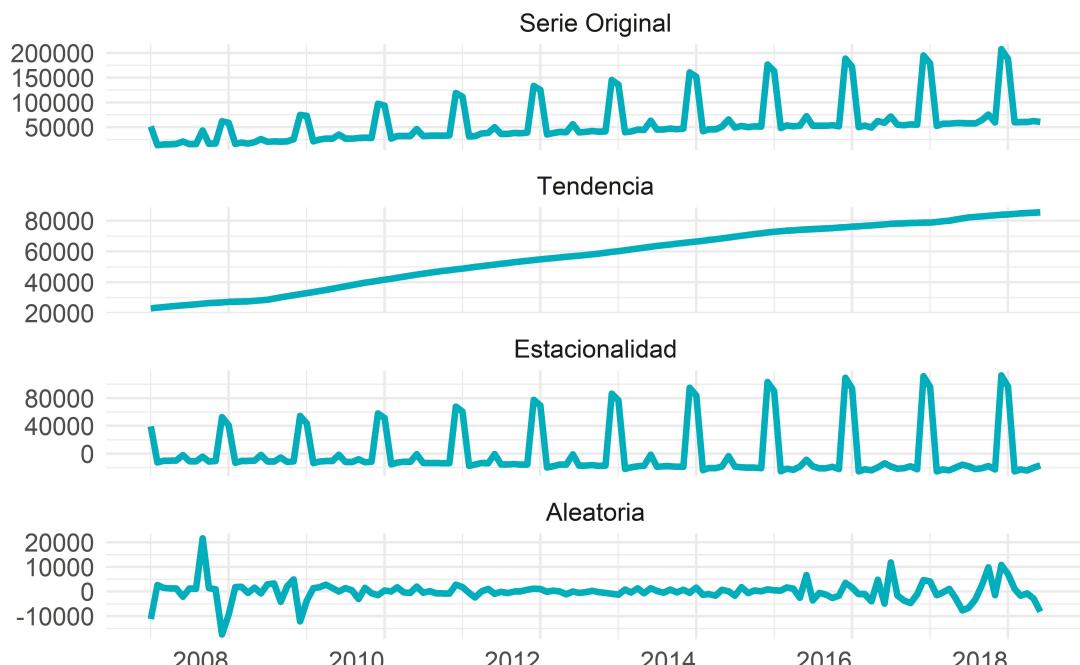
Figura 20: Incentivos salariales en el sector público 2007 - 2018 según mes



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, 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.

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



Fuente: UED-INEC

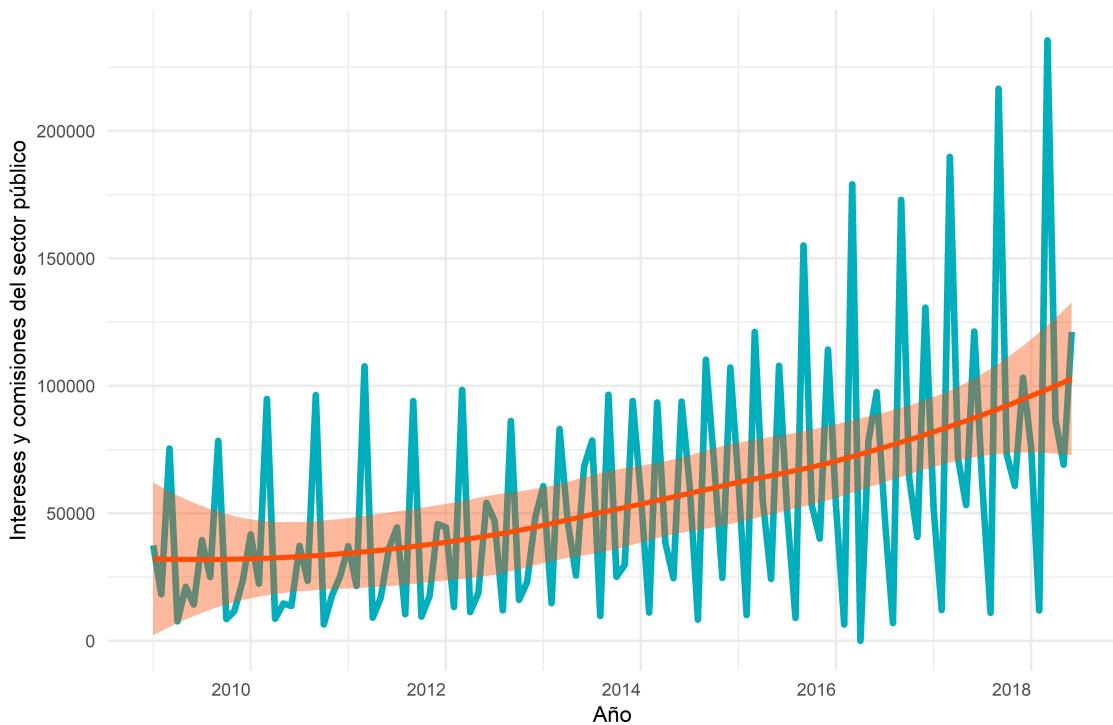
3.1.4 Intereses y comisiones del sector público

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 22 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,

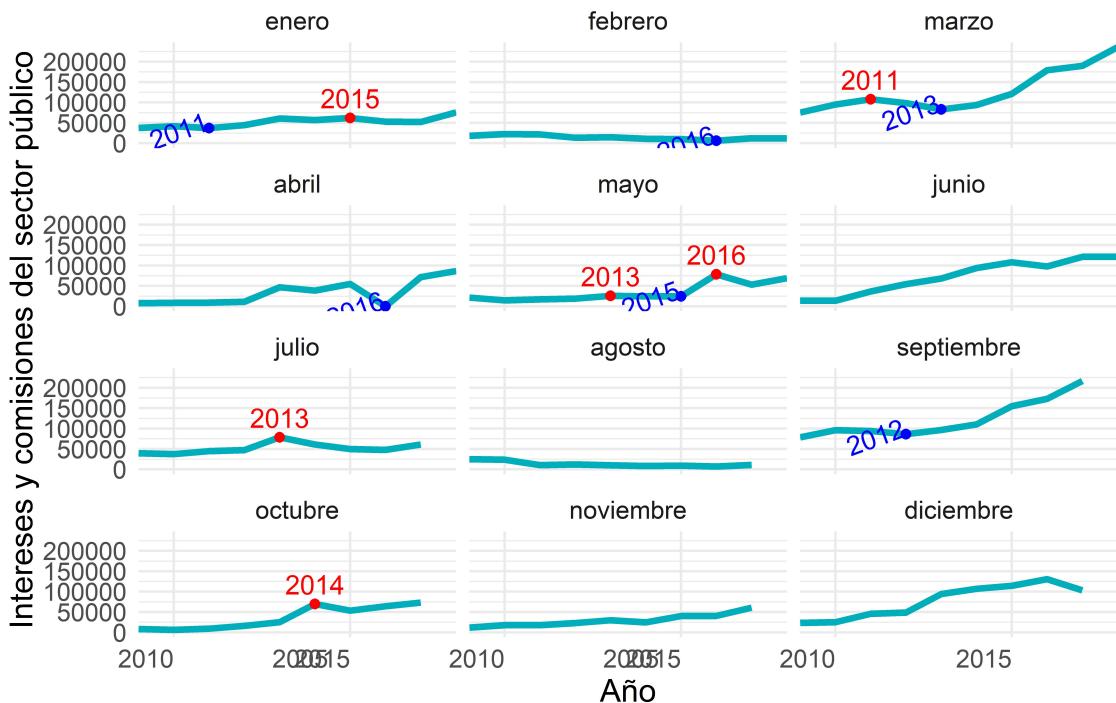
la figura 23 muestra cómo hay un crecimiento sostenido de los intereses y comisiones del sector público al final de cada trimestre durante todo el periodo, mientras que se mantiene casi constante durante los primeros dos meses de cada trimestre. La caída más pronunciada se dio en abril del 2015 mientras que la tasa de crecimiento más rápida parece darse al final del primer trimestre. Además, en la figura 24 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 22: Intereses y comisiones del sector público en el periodo 2007-2018



Fuente: Contraloría General de la República de Costa Rica

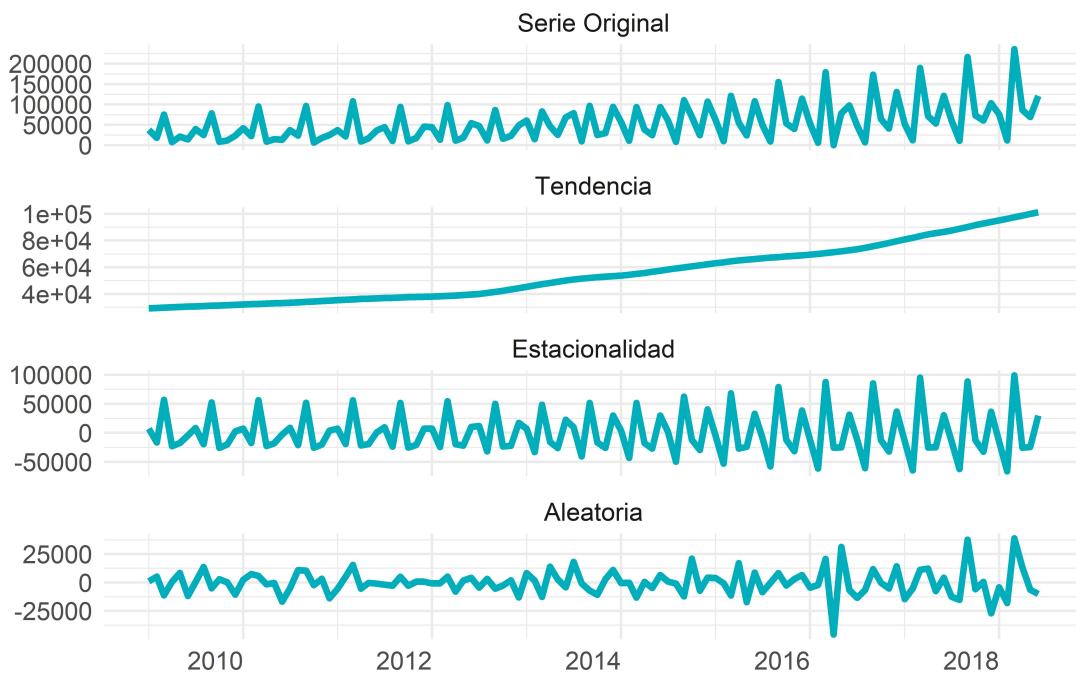
Figura 23: Intereses y comisiones del sector público en el periodo 2007-2018 según mes



Fuente: Contraloría General de la República de Costa Rica

En la figura 24 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.

Figura 24: Descomposición de la serie de Incentivos salariales en el periodo 2007 - 2018



Fuente: Contraloría General de la República de Costa Rica

3.1.5 Herramientas analíticas y procedimiento de simulación

Como se ha mencionado en este documento, el lenguaje de programación R ha sido utilizado para los análisis. Específicamente, 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` (Golemund & Wickham, 2011), `ggseas` (Ellis, 2018), `ggpmisc` (Aphalo, 2021) y `forecast` (Hyndman & Khandakar, 2008).

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, se comparan los resultados obtenidos mediante la sobreparametrización con los obtenidos utilizando la función `auto.arima()` de R, aplicado a las distintas series cronológicas.

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 generar series cronológicas que son gobernadas por un proceso determinado y previamente conocido para poder compararlo con los modelos identificados tanto con la sobreparametrización,

como con la función `auto.arima()` y el correspondiente modelo *ARIMA* estándar.

Con este fin, se programó una función que sigue los siguientes pasos:

1. Se generan valores aleatorios de alguna distribución de probabilidad. Para esta investigación se escogen 100 valores de una distribución Normal con media 10 y varianza 1. Estos valores se resumen en la figura 25; 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.
2. Se seleccionan mediante un muestreo simple al azar la cantidad de coeficientes a utilizar en los términos del modelo $ARIMA(p, d, q)(P, D, Q)_S$ que gobierna la serie. Para esta investigación fueron seleccionados los siguientes procesos: $ARIMA(1, 0, 0)$, $ARIMA(1, 0, 1)$, $ARIMA(2, 0, 3)$, $ARIMA(4, 0, 2)$, $ARIMA(0, 0, 1)(0, 1, 1)$, $ARIMA(2, 1, 4)(3, 0, 3)$.
3. Para cada uno de los procesos seleccionados, se genera una secuencia de valores en el intervalo $[-1, 1]$ con saltos de 0.1 para posteriormente seleccionar de manera aleatoria el valor que tomarán los coeficientes de cada uno de los procesos mencionados en el punto anterior.
4. Con los valores simulados, la cantidad de parámetros y sus respectivos valores definidos en los puntos anteriores, se ajusta cada uno de los modelos *ARIMA* descritos en el inciso 2..
5. Con cada modelo ajustado, se utiliza la función `simulate.Arima()` para generar 200 observaciones basadas en dichos modelos.

Tras aplicar los pasos anteriores y obtener las correspondientes series cronológicas simuladas, el comportamiento y proceso gobernante de cada una se muestra en la figura 26.

Figura 25: Valores de referencia para la simulación de series cronológicas

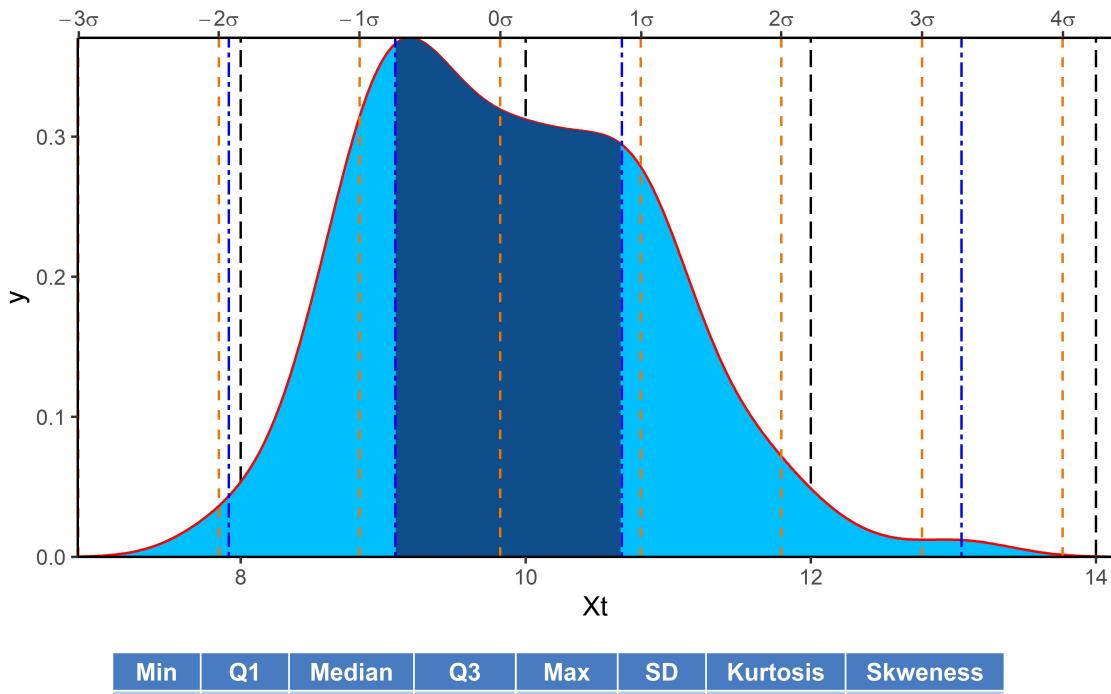
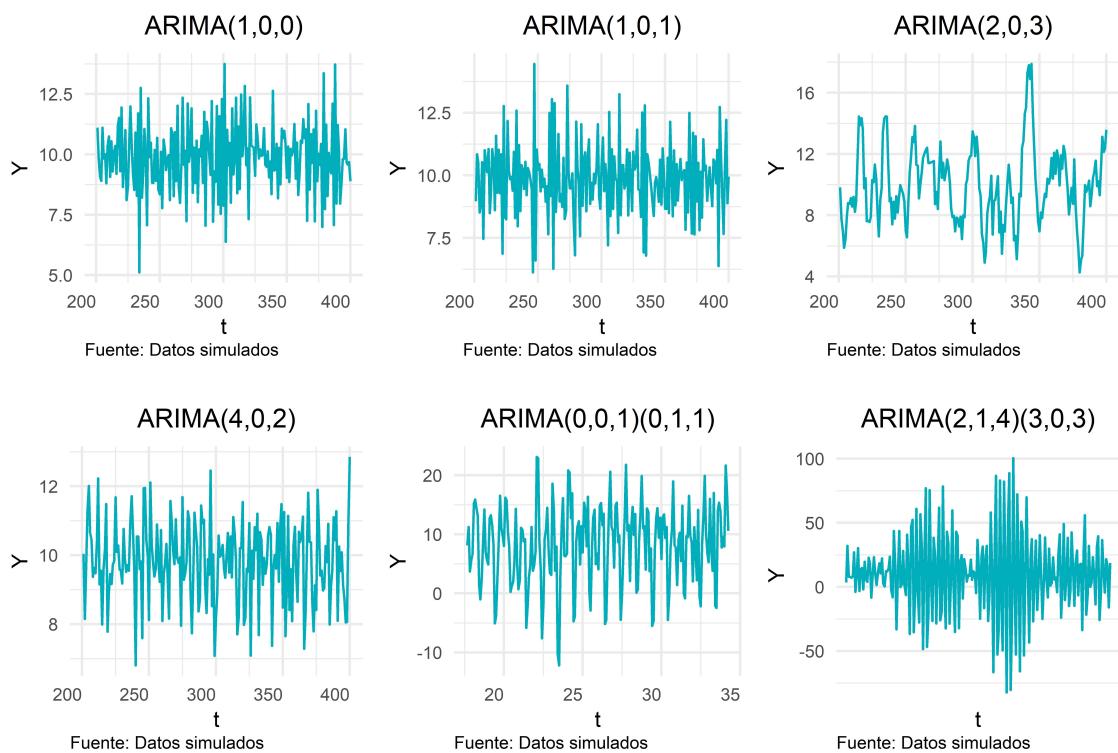


Figura 26: Series cronológicas simuladas



3.2 Métodos

En este apartado se describe el procedimiento a seguir con cada una de las series cronológicas mencionadas previamente, tanto las series simuladas como las reales. Para comprobar el poder predictivo del método propuesto se realiza inicialmente un análisis exploratorio para verificar si las series temporales sujetas a análisis son o no estacionarias, y en caso de no serlo, si requieren algún proceso de diferenciación. Se describe además el proceso de partición de los datos tanto para ajustar los modelos como para validar los pronósticos. Para cada serie de tiempo, se estima el mejor modelo utilizando la función `auto.arima()`, la sobreparametrización y un modelo ARIMA estándar, que  puede ser un $ARIMA(1, 1, 1)$ para las series no estacionales, o un $ARIMA(1, 1, 1)(1, 1, 1)$  en el caso de las series cronológicas estacionales. Una vez obtenidos estos modelos, se analizan los residuales obtenidos. En última instancia, se obtienen los pronósticos y sus medidas de bondad de ajuste y de rendimiento.

3.2.1 Análisis exploratorio

Como fue mencionado en el Marco Teórico, debe corroborarse que la serie cronológica a trabajar posea un comportamiento estacionario y, de no serlo, someterla a procesos  matemáticos para asegurar esta condición, estando entre los más comunes la diferenciación o la aplicación del logaritmo natural. Posteriormente, se realiza una identificación del posible proceso que gobierna la serie cronológica al graficar las funciones de autocorrelación y autocorrelación parcial, las cuales también sirven para verificar si la serie (transformada o no) es estacionaria.

3.2.2 Partición de los datos

A partir de la serie cronológica que se someterá a análisis, se realiza una partición de los datos para tener dos conjuntos distintos: entrenamiento y validación. El primero servirá precisamente para entrenar y estimar los distintos modelos; mientras que el segundo servirá para validar los pronósticos obtenidos. De manera predeterminada, se utilizará 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(la) investigador(a).

3.2.3 Estimación del mejor modelo según la función `auto.arima()`

Con el correspondiente conjunto de datos de entrenamiento, se utiliza la función `auto.arima()` para encontrar el mejor modelo ARIMA sugerido con este método, que como fue mencionado en la introducción de esta tesis, usa como criterio la minimización del AICc.

3.2.4 Estimación del mejor modelo con sobreparametrización

A partir del mismo conjunto de datos de entrenamiento de la correspondiente serie cronológica, se utiliza la sobreparametrización para encontrar el mejor modelo a partir de distintas permutaciones de la cantidad de coeficientes de los términos p, d, q, P, D, Q , según sea el caso.

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 R, el cuál fue construido haciendo uso de los paquetes de R `tidyR` (Wickham & Henry, 2019), `dplyr` (Wickham et al., 2019) y `parallel` (R Core Team, 2019), los procesos internos de esta función son descritos a continuación:

1. 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.
2. Para el caso de series cronológicas no estacionales, los valores P, D y Q son iguales a cero (porque precisamente, no se estiman coeficientes para la parte estacional). Se estiman todas las permutaciones de parámetros p, d, q hasta tener como máximo un modelo $ARIMA(6, 1, 6)$. Para ello se genera una matriz con cada una de estas permutaciones, denominada matriz de valores paramétricos, en donde cada fila representa la especificación del modelo $ARIMA(p, d, q)$ que se va a estimar, tal y como se muestra en 25:

$$\begin{array}{c|ccc} & \overbrace{p, d, q} & \overbrace{P, D, Q} \\ \hline & 0 & 0 & 1 & 0 & 0 & 0 \\ & 0 & 0 & 2 & 0 & 0 & 0 \\ & 0 & 0 & 3 & 0 & 0 & 0 \\ & 0 & 0 & 4 & 0 & 0 & 0 \\ & 0 & 0 & 5 & 0 & 0 & 0 \\ & 0 & 0 & 6 & 0 & 0 & 0 \\ & 0 & 1 & 0 & 0 & 0 & 0 \\ & 0 & 1 & 1 & 0 & 0 & 0 \\ & 0 & 1 & 2 & 0 & 0 & 0 \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 6 & 1 & 6 & 0 & 0 & 0 \end{array} \quad (25)$$

3. De manera análoga, al trabajar con series cronológicas estacionales, se decide trabajar (para una temporalidad determinada, como mensual) hasta un modelo máximo de $ARIMA(4, 1, 4)(4, 1, 4)_{12}$. Así, la matriz de valores paramétricos mostrada en 26 posee, en cada línea, una especificación de modelo a estimar:

$$\begin{array}{cc}
 \overbrace{p, d, q} & \overbrace{P, D, Q} \\
 \left[\begin{array}{cccccc}
 0 & 0 & 1 & 0 & 0 & 1 \\
 0 & 0 & 1 & 0 & 0 & 2 \\
 0 & 0 & 1 & 0 & 0 & 3 \\
 0 & 0 & 1 & 0 & 0 & 4 \\
 0 & 0 & 1 & 0 & 1 & 1 \\
 0 & 0 & 1 & 0 & 1 & 2 \\
 0 & 0 & 1 & 0 & 1 & 3 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 4 & 1 & 4 & 4 & 1 & 4
 \end{array} \right] & (26)
 \end{array}$$

4. Con la matriz de valores paramétricos, como las mostradas en 25 y 26, se inicia la estimación los modelos en orden ascendente, es decir, 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 que el nuevo término incorporado al modelo es significativamente distinto de cero, es decir, el nuevo parámetro está generando un impacto en el modelo.
5. Al tratarse de un proceso iterativo, el cálculo puede 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.
6. Cuando se han realizado las pruebas de significancia estadística a los modelos, son calculadas las medidas de bondad de ajuste y de rendimiento que se mencionarán más adelante.
7. Tras esto, se aplica un método de concenso para seleccionar el modelo más adecuado. Este criterio consiste en darle una 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 seleccionado. El método de concenso es utilizado para obtener un puntaje de cada modelo ARIMA, su cálculo se obtiene de la ecuación 27:

$$\min \left(\sum_i m_i \cdot w_j \right) \quad (27)$$

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.

3.2.5 Estimación de un modelo ARIMA estándar

Para contrastar los dos métodos de selección de modelos anteriores (`auto.arima()` y sobreparametrización), se ajusta también un modelo ARIMA más tradicional o estándar. En el caso de las series cronológicas no estacionales se ajusta un modelo $ARIMA(1, 1, 1)$ y en el caso de las series estacionales se ajusta un modelo $ARIMA(1, 1, 1), (1, 1, 1)_S$.

3.2.6 Análisis de los errores

Una vez que se selecciona un modelo de cada tipo (`auto.arima()`, sobreparametrización y ARIMA estándar), se realiza un análisis de los residuos estandarizados, la autorrelación y el supuesto de normalidad de normalidad de los residuales.

3.2.7 Pronósticos

Para cada modelo estimado, se realiza un pronóstico de h periodos hacia el futuro (donde el valor de h es el tamaño de los conjuntos de validación creados para cada serie) para realizar una inspección visual de los resultados previo a hacer una comparación numérica mediante dos formas distintas pero complementarias: las medidas de bondad de ajuste y de rendimiento.

3.2.8 Medidas de bondad de ajuste y de rendimiento

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

3.2.8.1 AIC

Se calcula de la siguiente manera:

$$AIC = -2\log L(\hat{\theta}) + 2k \quad (28)$$

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

3.2.8.2 AICc

Su forma de cálculo se muestra en la ecuación 29

$$AICc = -2\log L(\hat{\theta}) + 2k + \frac{2k+1}{n-k-1} \quad (29)$$

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

3.2.8.3 BIC

El último estadístico de bondad de ajuste se calcula como se muestran en la ecuación 30.

$$BIC = -2\log L(\hat{\theta}) + k \cdot \log(n) \quad (30)$$

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

3.2.8.4 MAE

El error absoluto medio se define mediante la ecuación 31

$$\frac{1}{n} \sum_{t=1}^n |e_t| \quad (31)$$

3.2.8.5 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 32 y 33.

$$\frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-1} \sum_{t=2}^T |Y_t - Y_{t-1}|} \quad (32)$$

$$\frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-m} \sum_{t=m+1}^T |Y_t - Y_{t-m}|} \quad (33)$$

Donde m es la temporalidad de la serie.

3.2.8.6 RMSE

Es la raíz del error cuadrático medio, como se define en la ecuación 34.

$$\sqrt{\frac{1}{n} \sum_{t=1}^n |e_t|^2} \quad (34)$$

De esta manera, cada una de las series cronológicas (simuladas y reales) serán sometidas a un análisis exploratorio y a una partición de los datos con un 80 % para el conjunto de entrenamiento y el restante 20 % para validación, esto con el fin de estimar modelos mediante la función `auto.arima()`, la sobreparametrización y un ARIMA estándar para evaluar sus correspondientes y la calidad de los pronósticos obtenidos.

4 RESULTADOS

Este capítulo abordará los principales resultados a partir del procedimiento descrito en la metodología. Cada sección de este capítulo tendrá una subsección donde se aplica cada etapa de análisis a las respectivas series cronológicas, en el cuál se espera que los resultados obtenidos mediante la sobreparametrización igualen o mejoren los obtenidos con los métodos tradicionales. La estructura de este capítulo se muestra en el índice de este documento.



4.1 Análisis exploratorio

4.1.1 Datos simulados

4.1.1.1 ARIMA(1,0,0)

4.1.1.2 ARIMA(1,0,1)

4.1.1.3 ARIMA(2,0,3)

4.1.1.4 ARIMA(4,0,2)

4.1.1.5 ARIMA(0,0,1)(0,1,1)

4.1.1.6 ARIMA(2,1,4)(3,0,3)

4.1.2 Tasa de mortalidad infantil interanual

4.1.3 Mortalidad por causa externa

4.1.4 Incentivos salariales

4.1.5 Intereses y comisiones del sector público

4.2 Partición de los datos

4.3 Estimación del mejor modelo según la función auto.arima()

4.3.1 Datos simulados

4.3.2 Datos reales

4.4 Estimación del mejor modelo con sobreparametrización

4.4.1 Datos simulados

4.4.2 Datos reales

4.5 Estimación de un modelo ARIMA estándar

4.5.1 Datos simulados

4.5.2 Datos reales

4.6 Análisis de los errores

4.6.1 Datos simulados

4.6.1.1 Errores de los modelos estimados con auto.arima()

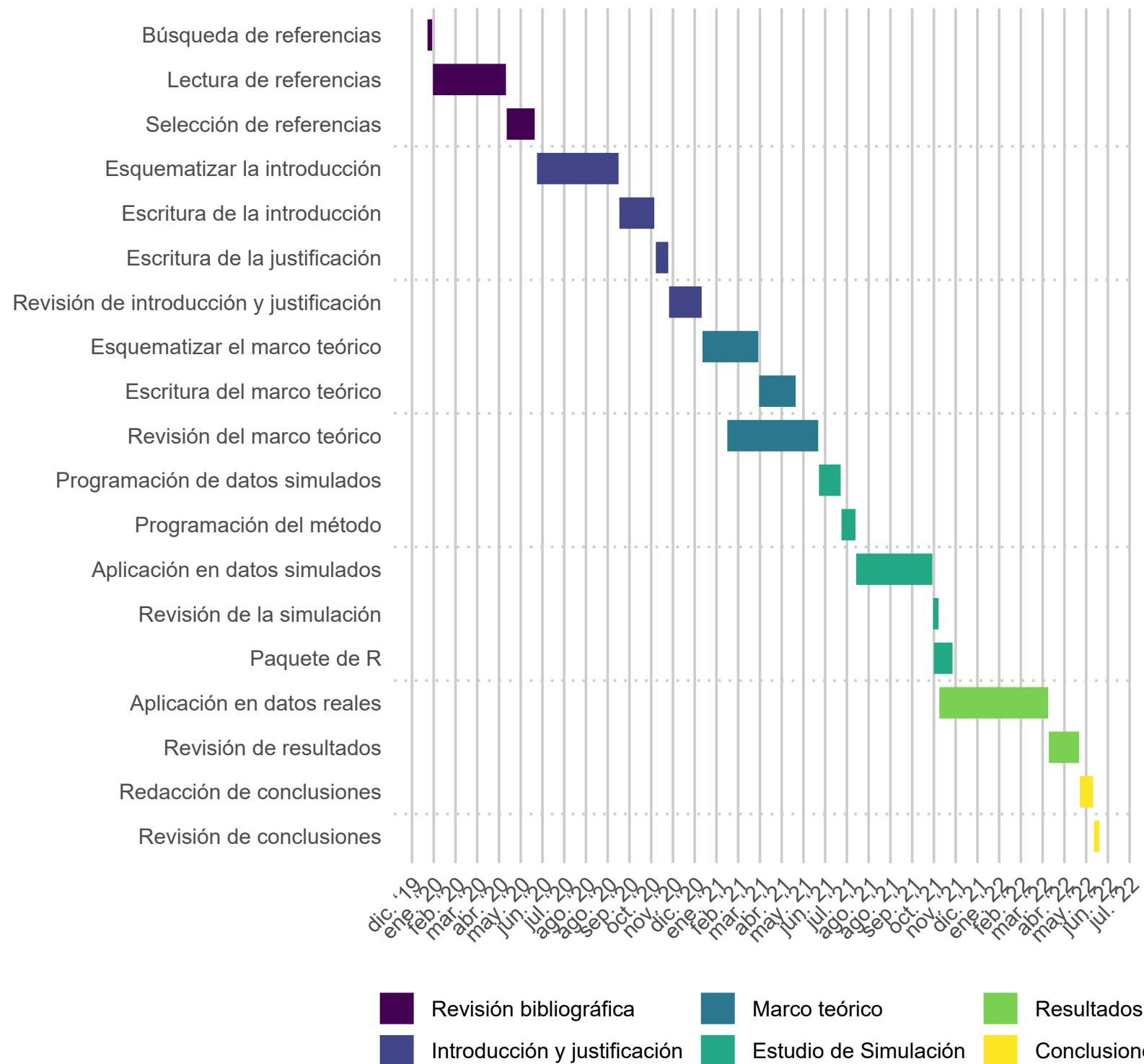
4.6.1.2 Errores de los modelos estimados con sobreparametrización

4.6.1.3 Errores de los modelos estimados con un modelo ARIMA estándar

4.6.2 Datos reales

6 ANEXOS

Cronograma tentativo



7 REFERENCIAS

- Adhikari, R., K, A. R., & Agrawal, R. K. (2013). *An introductory study on time series modeling and forecasting* (pp. 42–45). Lap Lambert Academic Publishing GmbH KG. <https://arxiv.org/pdf/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'*. <https://CRAN.R-project.org/package=ggpmisc>
- Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid"graphics*. <https://CRAN.R-project.org/package=gridExtra>
- Benesty, J., & Chen, Y. and C., J.and Huang. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 37–38). Springer Berlin Heidelberg. 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*. Prentice Hall. <https://books.google.co.cr/books?id=sRzvAAAAMAAJ>
- Brockwell, P. J., & Davis, R. A. (2009). *Time series: Theory and methods* (p. 239). Springer. https://books.google.co.cr/books?id=_DcYu/_EhVzUC
- Brown, R. (1956). *Exponential smoothing for predicting demand*. A.D.Little. <https://www.industrydocuments.ucsf.edu/docs/jzlc0130>
- Burnham, K. P., & Anderson, D. R. (2007). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer New York. <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*. 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. <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). https://www.researchgate.net/publication/274885475_Mortalidad_por_causaExternas_un_problema_de_salud_publica_Argentina_Chile_y_Colombia_2000-2008
- Cochrane, J. H. (1997). *Time series for macroeconomics and finance*. Graduate School of Business, University of Chicago. <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), 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 santiago. *Revista Médica de Chile*, 132, 461–466. 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'*. <https://CRAN.R-project.org/package=ggseas>
- Elmabrouk, O. M. (n.d.). *Measuring reliability of stationary stochastic processes*. https://www.academia.edu/7140606/Measuring_Reliability_of_Stationary_Stochastic_Processes?auto=download
- Evans, M. J., & Rosenthal, J. S. (2005). *Probabilidad y estadística* (p. 121). Reverte. <https://books.google.co.cr/books?id=ZU3MEKZFgsMC>
- Flaherty, J., & Lombardo, R. (2000, January). *Modelling private new housing starts in australia*. http://www.prres.net/papers/Flaherty_Modelling_Private_New_Housing_Starts_In_Australia.pdf
- Fuller, W. A. (1995). *Introduction to statistical time series*. Wiley. <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. <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. (2008). *Modelos probabilísticos discretos* (1st ed.). Editorial Universidad de Costa Rica. <http://www.editorial.ucr.ac.cr/ciencias-naturales-y-exactas/item/2168-modelos-probabilisticos-discretos.html>
- Hernández, O. (2011). *Introducción a las series cronológicas* (1st ed.). Editorial Universidad de Costa Rica. <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*. Elsevier Science. <https://books.google.co.cr/books?id=t1zG8OUbgdgC>
- Hyndman, R. J., & Athanasopoulos, G. (2018a). *Forecasting: Principles and practice*. OTexts. https://books.google.co.cr/books?id=_bBhDwAAQBAJ
- Hyndman, R. J., & Athanasopoulos, G. (2018b). *Forecasting: Principles and practice*. OTexts. https://books.google.co.cr/books?id=_bBhDwAAQBAJ
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 26(3), 1–22. <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*. <https://arxiv.org/pdf/1801.03222.pdf>
- Kassambara, A. (2020). *Ggpubr: 'ggplot2' based publication ready plots*. <https://CRAN.R-project.org/package=ggpubr>
- Kedem, B., & Fokianos, K. (2005). *Regression models for time series analysis*. Wiley. <https://books.google.co.cr/books?id=8r0qE35wt44C>
- Lee, J. (n.d.). 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. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21251998000600017&nrm=iso
- Nación. (2013). Morbilidad y mortalidad en costa rica. *La Nacion*. <https://bit.ly/2xWUeXU>
- OPS. (2016). *Clasificación estadística internacional de enfermedades y problemas relacionados con la salud* (2015th ed., Vol. 2). OMS.
- Osborn, D. R., Chui, A. P. L., Smith, J., & Birchenhall, C. (2009). *Seasonality and the order of integration for consumption*. 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*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ramírez, F. (2007). *Introducción a las series de tiempo. Métodos paramétricos*. Sello Editorial, Universidad de Medellín. <https://books.google.es/books?id=KvLhxFPwvsUC>
- 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*. CCP-UCR. <http://srv-website.cloudapp.net/documents/10179/999061/Nota+t%C3%A9cnica+tablas+de+vida+segunda+parte>
- Sargent, T. J. (1979). *Macroeconomic theory*. Academic Press. <https://books.google.co.cr/books?id=X6u7AAAAIAAJ>
- Stoffer, D. (2020). *Astsa: Applied statistical time series analysis*. <https://CRAN.R-project.org/package=astsa>
- Surhone, L. M., Timpledon, M. T., & Marseken, S. F. (2010). *Wold decomposition*. VDM Publishing. <https://books.google.co.cr/books?id=7cSqcQAACAAJ>
- Tadayon, M., & Iwashita, Y. (2020). *Comprehensive analysis of time series forecasting using neural networks*. <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*. <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*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files*. <https://CRAN.R-project.org/package=readxl>
- Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data*. <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. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC. <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*. <https://CRAN.R-project.org/package=kableExtra>