

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

LA SOBREPARAMETRIZACIÓN EN EL ARIMA: UNA APLICACIÓN A
DATOS COSTARRICENSES

Tesis sometida a la consideración de la Comisión del Programa de Estudios de
Posgrado en Estadística para optar por el grado y título de Maestría Académica en
Estadística

CÉSAR ANDRÉS GAMBOA SANABRIA B12672

Ciudad Universitaria Rodrigo Facio, Costa Rica

2022

DEDICATORIA

A mi abuela, un pilar fundamental en mi vida sin cuyo apoyo no hubiese alcanzado muchas de mis metas y de quien aprendí que no desaparece lo que se muere, solo lo que se olvida. A mi madre, que junto con mi abuela me enseñaron en su más amplio sentido que quien no estudia, muere en el charco de la ignorancia.

A mi esposa, quien con su forma de ver la vida me sigue haciendo crecer tanto personal como académica y profesionalmente.

A mi tía, a mi hermana y su esposo, quienes con su comprensión en los momentos más difíciles supieron darme el aliento necesario y la motivación para seguir adelante.

A mis tíos, por velar en distintas etapas de mi vida que tuviera las herramientas necesarias para ser cada día una mejor persona.

AGRADECIMIENTOS

Agradezco profundamente a Óscar Centeno Mora, quien con sus buenos consejos y dedicación ayudó a hacer posible este proyecto, a Gilbert Brenes Camacho y a Shu Wei Chou por sus valiosos aportes técnicos y sustantivos en el contexto de esta tesis, y a Cindy Cárdenas, por su invaluable apoyo administrativo en los momentos justos.

“Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Estadística”

Ph.D. Flor Isabel Jiménez Segura
Decano Sistema de Estudios de Posgrado

M.Sc. Óscar Centeno Mora
Director de Tesis

Ph.D. Gilbert Brenes Camacho
Lector

Ph.D. ShuWei Chou.
Lector

Ph.D. Gilbert Brenes Camacho
Director Programa de Posgrado en Estadística

César Andrés Gamboa Sanabria
Candidato

Índice

DEDICATORIA	I
AGRADECIMIENTOS	II
RESUMEN	VII
ABSTRACT	VIII
1 INTRODUCCIÓN	1
1.1 Antecedentes	1
1.2 El problema	2
1.3 Objetivos del estudio	3
1.4 Justificación del estudio	3
1.5 Organización del estudio	4
2 MARCO TEÓRICO	5
2.1 Definición de una serie de tiempo	5
2.2 Componentes de una serie cronológica	6
2.2.1 La tendencia-ciclo	8
2.2.2 Componentes estacionales	10
2.2.3 Componente irregular	11
2.3 Supuestos en el análisis de series cronológicas	12
2.4 Identificación del modelo	14
2.5 Modelos Autorregresivos Integrados de Medias Móviles (ARIMA)	16
2.5.1 Ecuación de Wold	16
2.5.2 Metodología Box-Jenkins	18
2.5.3 Modelos Autorregresivos	18
2.5.4 Modelos de Medias Móviles	18
2.5.5 Modelos ARIMA	19
2.6 Los autocorrelogramas	20
2.7 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	28
3.1.3 Incentivos salariales del sector público	29
3.1.4 Intereses y comisiones del sector público	29

3.1.5	Herramientas analíticas	30
3.1.6	Procedimiento de simulación	30
3.2	Métodos	31
3.2.1	Análisis exploratorio	31
3.2.2	Partición de los datos	31
3.2.3	Identificación y estimación del mejor modelo según la función auto.arima()	32
3.2.4	Identificación y estimación del mejor modelo con sobreparametrización	32
3.2.5	Estimación de un modelo ARIMA estándar	35
3.2.6	Ánalisis visual de los errores	35
3.2.7	Medidas de bondad de ajuste y de rendimiento	35
3.2.7.1	AIC	35
3.2.7.2	AICc	35
3.2.7.3	BIC	36
3.2.7.4	MAE	36
3.2.7.5	MASE	36
3.2.7.6	RMSE	36
3.2.8	Pronósticos	36
3.2.9	Tiempo de procesamiento	37
3.2.10	Resumen de la forma de análisis	37
4	RESULTADOS	38
4.1	Contraste de métodos de estimación de los ARIMA para las series simuladas	38
4.1.1	Análisis exploratorio	38
4.1.2	Partición de los datos	39
4.1.3	Identificación y estimación	40
4.1.4	Ánalisis de los errores	41
4.1.5	Medidas de bondad de ajuste y de rendimiento	43
4.1.6	Estimación en el periodo de validación	44
4.2	Contraste de métodos de estimación de los ARIMA para las series empíricas	49
4.2.1	Análisis exploratorio	49
4.2.1.1	Tasa de mortalidad infantil interanual	49
4.2.1.2	Incentivos salariales del sector público	52
4.2.1.3	Mortalidad por causa externa	56
4.2.1.4	Intereses y comisiones del sector público	56
4.2.2	Partición de los datos	57
4.2.3	Identificación y estimación	57
4.2.4	Ánalisis de los errores	58

4.2.5	Medidas de bondad de ajuste y de precisión	60
4.2.6	Estimación en el periodo de validación	61
4.3	Resumen de resultados	66
5	CONCLUSIÓN Y DISCUSIONES	67
5.1	Conclusiones	67
5.2	Discusiones	70
6	ANEXOS	72
7	REFERENCIAS	94

RESUMEN

Estimar modelos de series cronológicas es una labor ampliamente extendida en múltiples campos de la investigación y uno de los objetivos es generar pronósticos de la forma más precisa posible dentro de un horizonte determinado. Existe una amplia gama de modelos que puede utilizarse con este fin, entre ellos están los modelos Autorregresivos Integrados de Medias Móviles (*ARIMA*), e incluso existen diversos métodos de estimación automática o semi-automática para esta rama de la estadística.

A pesar de esto, encontrar un modelo que posea un buen ajuste a los datos no es fácil, pues se deben considerar tanto aspectos teóricos como prácticos, y de la temática de estudio para así obtener un modelo adecuado que genere pronósticos realistas y pertinentes para la toma de decisiones dentro de lo posible.

La investigación propone hacer uso de método denominado sobreparametrización en conjunto con el método de permutaciones del análisis combinatorio para someter a prueba una espectro más amplio de posibles modelos *ARIMA*. En la selección de modelos ARIMA los métodos más tradicionales como los correlogramas u otros, no suelen cubrir muchas alternativas para definir la cantidad de coeficientes a estimar en el modelo, lo cual representa un método de estimación que no es óptimo. El presente tesis propone una metodología para obtener pronósticos más precisos en comparación a los métodos tradicionales.

Los resultados se contrastan con datos simulados de series cronológicas y cuatro series reales para ajustar modelos *ARIMA* con la función `auto.arima()`, la sobreparametrización y un modelo *ARIMA* de orden bajo. Para cada una de estas series se realiza una partición del 80 % para entrenar los modelos y el restante 20 % para validación de los pronósticos. En cada una de estas series se realizó un análisis visual del comportamiento de los errores y posteriormente se evalúa la calidad de los resultados de cada modelo obtenido con las tres técnicas descritas mediante medidas de bondad de ajuste (AIC, AICc y BIC) y de precisión (RMSE, MAE y MAPE).

Al tener datos que vienen de un proceso con bajo número de parámetros, la sobreparametrización logra captar de buena manera el comportamiento de la serie en comparación a las otras alternativas, y cuando el proceso que gobierna la serie es de un mayor grado, la metodología propuesta es capaz de capturar de mejor forma el comportamiento de la serie y conseguir pronósticos con una precisión mayor al de los métodos más tradicionales, pues en los resultados de entrenamiento, la sobreparametrización obtuvo el mejor ajuste un 58,33 % de las veces y la mejor precisión el 45,45 % de las veces, mientras que al evaluar los resultados sobre los conjuntos de datos de validación, la sobreparametrización obtuvo el mejor ajuste el 50 % de las veces, mientras que las mejores medidas de precisión se alcanzaron un 67 % del tiempo.

ABSTRACT

Time series analysis is widely extended in several research fields, and one goal is to generate forecasts as accurately as possible within a given horizon. Several models can be used for this purpose, among them are the Autoregressive Integrated Moving Average models (ARIMA). There are even some automatic or semi-automatic estimation methods for their calculation.

Despite this, finding a well-fitted model is not easy, since both theoretical and practical aspects must be considered, as well as particular context to obtain an adequate model that generates realistic and relevant forecasts for decision making as much as possible.

This research proposes the use of overparameterization combinatorial analysis to test a broader spectrum of possible *ARIMA* models. In ARIMA model selection, the most traditional methods, such as correlograms, do not usually cover many alternatives to define the number of the coefficients to be estimated in the model, which represents a non-optimal estimation method. This thesis proposes a methodology to obtain more accurate forecasts compared to traditional methods.

The results are contrasted with simulated time-series data and four real series to fit *ARIMA* models with the `auto.arima()` function, overparameterization, and a low-order *ARIMA* model. For each of these series, an 80 % partition is made to train the models and the remaining 20 % to validate the forecasts.

In each of these series, a visual analysis of the behavior of the errors was carried out, and subsequently, the quality of the results of each model obtained with these three techniques was evaluated through measures of goodness of fit (AIC, AICc, and BIC) and precision (RMSE, MAE, and MAPE).

By having data that comes from a process with a low number of parameters, the overparameterization manages to capture in a good way the behavior of the series compared to the other methods. When the process that rules the time series implies a higher degree, the proposed methodology leads to a well-capture-behavior of the series and achieving forecasts with greater precision than the more traditional methods.

In the training results, the overparameterization obtained the best fit 58,33 % of the time and the best precision the 45,45 % of the time. In comparison, when evaluating the results on the validation data sets, overparameterization obtained the best fit 50 % of the time, while the best precision measures were achieved 67 % of the time.

Índice de cuadros

1	Coeficientes del proceso original y de los métodos de estimación de las series simuladas estacionales	41
2	Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación simulados a partir de datos estacionales simulados	44
3	Coeficientes de las ecuaciones de estimación según método de ajuste	58
4	Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación a partir de las series cronológicas reales	61
5	Distribución porcentual de los métodos de estimación que alcanzaron los mejores resultados según conjunto de datos y tipo de medición	66
6	Tiempos de estimación en minutos para cada modelo según su tipo de estimación.	72
7	Coeficientes del proceso original y de los métodos de estimación de series simuladas no estacionales	72
8	Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación simulados a partir de datos estacionales simulados	75
9	Coeficientes de las ecuaciones de estimación según método de ajuste	88
10	Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación a partir de las series cronológicas reales	89

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	Diagrama de fijo del proceso de simulación de las series cronológicas.	31
14	Comportamiento y tendencia de las series simuladas	39
15	Partición de los datos en los conjuntos de entrenamiento y validación para las series de tiempo simuladas	40
16	Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(0,0,1)(0,1,1)	42
17	Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(2,1,4)(3,0,3)	43
18	Pronóstico de los datos generados mediante un ARIMA(0,0,1)(0,1,1) según el método de estimación	45
19	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(0,0,1)(0,1,1) con la función auto.arima()	45
20	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(0,0,1)(0,1,1) con sobreparametrización	46
21	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(0,0,1)(0,1,1) con el modelo ARIMA estándar	46
22	Pronóstico de los datos generados mediante un ARIMA(2,1,4)(3,0,3) según el método de estimación	47
23	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,1,4)(3,0,3) con la función auto.arima()	47
24	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,1,4)(3,0,3) con sobreparametrización	48
25	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,1,4)(3,0,3) con el modelo ARIMA estándar	48
26	Tasa de Mortalidad Infantil Interanual 1989 - 2017	49
27	Tasa de Mortalidad Infantil Interanual 1989 - 2017 según periodos	50
28	Descomposición de la TMII en el periodo 2000 - 2017	51
29	Autocorrelación de los datos diferenciados de la TMII	51
30	Autocorrelación parcial de los datos diferenciados de la TMII	52
31	Incentivos salariales en el sector público entre los años 2007 y 2018	53
32	Incentivos salariales en el sector público entre los años 2007 y 2018 según mes	53
33	Descomposición de la serie de Incentivos salariales en el periodo 2007 al 2018	54
34	Autocorrelación de los datos diferenciados de la serie de incentivos salariales	55
35	Autocorrelación parcial de los datos diferenciados de la serie de incentivos salariales	55

36	Partición de los datos en los conjuntos de entrenamiento y validación para las series de tiempo reales	57
37	Comportamiento de los errores asociados a los modelos estimados con la serie de la TMII	59
38	Comportamiento de los errores asociados a los modelos estimados con la serie de incentivos salariales	60
39	Pronóstico de la TMII según el método de estimación	62
40	Pronóstico de la serie de incentivos salariales del sector público según el método de estimación	62
41	Errores estándar de los pronósticos obtenidos para la TMII con la función <code>auto.arima()</code>	63
42	Errores estándar de los pronósticos obtenidos para la TMII con sobreparametrización	64
43	Errores estándar de los pronósticos obtenidos para la TMII con el modelo ARIMA estándar	64
44	Errores estándar de los pronósticos obtenidos para la serie de incentivos salariales del sector público con la función <code>auto.arima()</code>	65
45	Errores estándar de los pronósticos obtenidos para la serie de incentivos salariales del sector público con sobreparametrización	65
46	Errores estándar de los pronósticos obtenidos para la serie de incentivos salariales del sector público con el modelo ARIMA estándar	66
47	Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(1,0,0)	73
48	Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(1,0,1)	73
49	Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(2,0,3)	74
50	Comportamiento de los errores de los modelos para los datos generados con un ARIMA(4,0,2)	74
51	Pronóstico de los datos generados mediante un ARIMA(1,0,0) según el método de estimación	76
52	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,0) con la función <code>auto.arima()</code>	76
53	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,0) con sobreparametrización	77
54	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,0) con el modelo ARIMA estándar	77

55	Pronóstico de los datos generados mediante un ARIMA(1,0,1) según el método de estimación	78
56	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,1) con la función auto.arima()	78
57	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,1) con sobreparametrización	79
58	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,1) con el modelo ARIMA estándar	79
59	Pronóstico de los datos generados mediante un ARIMA(2,0,3) según el método de estimación	80
60	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,0,3) con la función auto.arima()	80
61	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,0,3) con sobreparametrización	81
62	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,0,3) con el modelo ARIMA estándar	81
63	Pronóstico de los datos generados mediante un ARIMA(4,0,2) según el método de estimación	82
64	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(4,0,2) con la función auto.arima()	82
65	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(4,0,2) con sobreparametrización	83
66	Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(4,0,2) con el modelo ARIMA estándar	83
67	Mortalidad por causa externa entre los años 2000 y 2017	84
68	Intereses y comisiones del sector público en el periodo 2007 al 2018	84
69	Intereses y comisiones del sector público en el periodo 2007 al 2018 según mes	85
70	Descomposición de la serie de Incentivos salariales en el periodo 2007 al 2018	85
71	Autocorrelación de los datos diferenciados de la mortalidad por causa externa	86
72	Autocorrelación parcial de los datos diferenciados de la mortalidad por causa externa	86
73	Autocorrelación de los datos diferenciados de la serie de intereses y comisiones del sector público	87
74	Autocorrelación parcial de los datos diferenciados de la serie de intereses y comisiones del sector público	87
75	Comportamiento de los errores asociados a los modelos estimados con la serie de mortalidad por causa externa	88

76	Comportamiento de los errores asociados a los modelos estimados con la serie de intereses y comisiones del sector público	89
77	Pronóstico de la Tasa de mortalidad por causa externa (TMCE) según el método de estimación	90
78	Errores estándar de los pronósticos obtenidos para la Tasa de mortalidad por causa externa (TMCE) con la función auto.arima()	90
79	Errores estándar de los pronósticos obtenidos para la Tasa de mortalidad por causa externa (TMCE) con sobreparametrización	91
80	Errores estándar de los pronósticos obtenidos para la Tasa de mortalidad por causa externa (TMCE) con el modelo ARIMA estándar	91
81	Pronóstico de la serie de intereses y comisiones del sector público según el método de estimación	92
82	Errores estándar de los pronósticos obtenidos para la serie de intereses y comisiones del sector público con la función auto.arima()	92
83	Errores estándar de los pronósticos obtenidos para la serie de intereses y comisiones del sector público con sobreparametrización	93
84	Errores estándar de los pronósticos obtenidos para la serie de intereses y comisiones del sector público con el modelo ARIMA estándar	93

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 (que para datos longitudinales son la estimación futura de fechas por presentarse en el tiempo) 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 el método de Box-Jenkins, que refiere al análisis de series de tiempo longitudinales univariadas, una 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

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.

identificación de parámetros mediante autocorrelogramas parciales y totales. Sin embargo, los autocorrelogramas formados no analizan de forma exhaustiva y óptima los posibles coeficientes que podrían contemplarse en la ecuación de Wold, de la cual se desprende que todo proceso estacionario puede definirse de manera específica mediante una ecuación matemática⁴ y que, según su definición matemática, al poseer infinitos coeficientes, dicha expresión debe reducirse a una cantidad finita para ajustar de una mejor manera la ecuación de estimación.

Es debido a lo anterior que 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 ante todo 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 R⁵ (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), seleccionado el orden óptimo para los términos $ARIMA(p, d, q)(P, D, Q)_s$ para una serie cronológica determinada.

Sin embargo, estas pruebas suelen ignorar diversos términos que bien podrían ofrecer mejores pronósticos; no someten a prueba las posibles especificaciones de un modelo en un rango determinado, sino que realizan aproximaciones analíticas para definir el proceso que gobierna la serie cronológica, dejando así un vacío en el cual se corre el riesgo de no seleccionar un modelo que ofrezca mejores pronósticos. Poner a prueba un mayor número de posibilidades para la especificación de los modelos tiene la ventaja de descartar ciertos modelos, y mantener otros con un criterio más científico y una evidencia numérica que respalde esa decisión.

⁴Su notación se define mediante la ecuación 15.

⁵Descarga gratuita en <https://cran.r-project.org/>

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.

Para lograr esto, además de poner a prueba el algoritmo mediante el contraste empírico, 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. Contrastar la precisión de la estimación así como la generación de pronósticos con otros métodos similares, aplicados en datos costarricenses.
3. Aplicar diversos métodos de validación en la estimación de procesos que gobiernan la serie cronológica.
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 basadas principalmente en criterios de bondad de ajuste, las cuales pueden omitir procesos que al momento de pronosticar describirían de una mejor manera el comportamiento futuro de una serie cronológica.

Estimar modelos ARIMA considerando diversas permutaciones de los parámetros de la estructura del modelo, 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 capítulo 1 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 planeó modificar el método de estimación en los modelos ARIMA. Se concluye esta sección con hechos que justifican la importancia de esta investigación.

El capítulo 2 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 capítulo 3 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 4 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 $ARIMA(1, 1, 1)(1, 1, 1)_s$.

En el capítulo 5 se incorpora la conclusión/discusión, que 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 una ecuación de estimación 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](#)).

2.1 Definición de una serie de tiempo

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 $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k}$ donde k es el primer momento de referencia, 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 de selección de casos.

2.2 Componentes de una serie cronológica

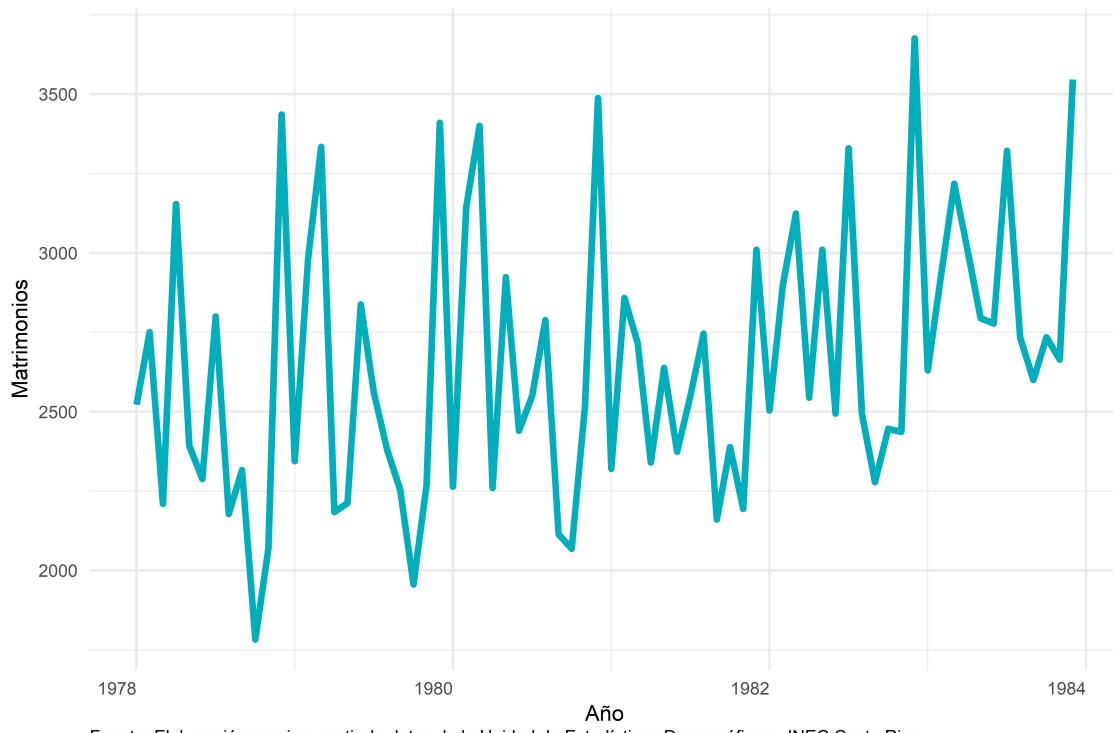
En el análisis de series cronológicas existen dos grandes corrientes de estudio: Los componentes inherentes a la serie cronológica y el estudio de las autocorrelaciones. Según el primer enfoque, de acuerdo con Hernández (2011), las series cronológicas poseen tres componentes principales: Tendencia-ciclos,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

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

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

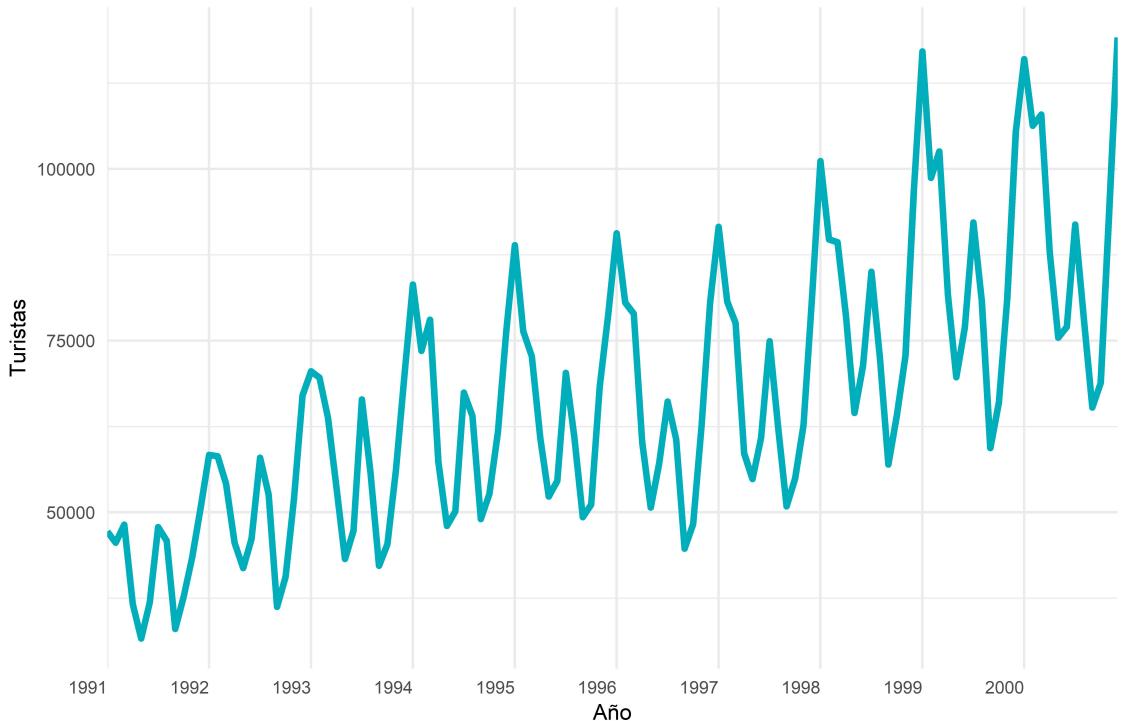
Donde Y es la serie cronológica, T es la tendencia-ciclo, S es la parte estacional, I la parte irregular o aleatoria, y t es el momento en el tiempo. Esta perspectiva clásica del análisis de series de tiempo permite realizar un análisis descriptivo del comportamiento de la serie en cuestión; cada una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

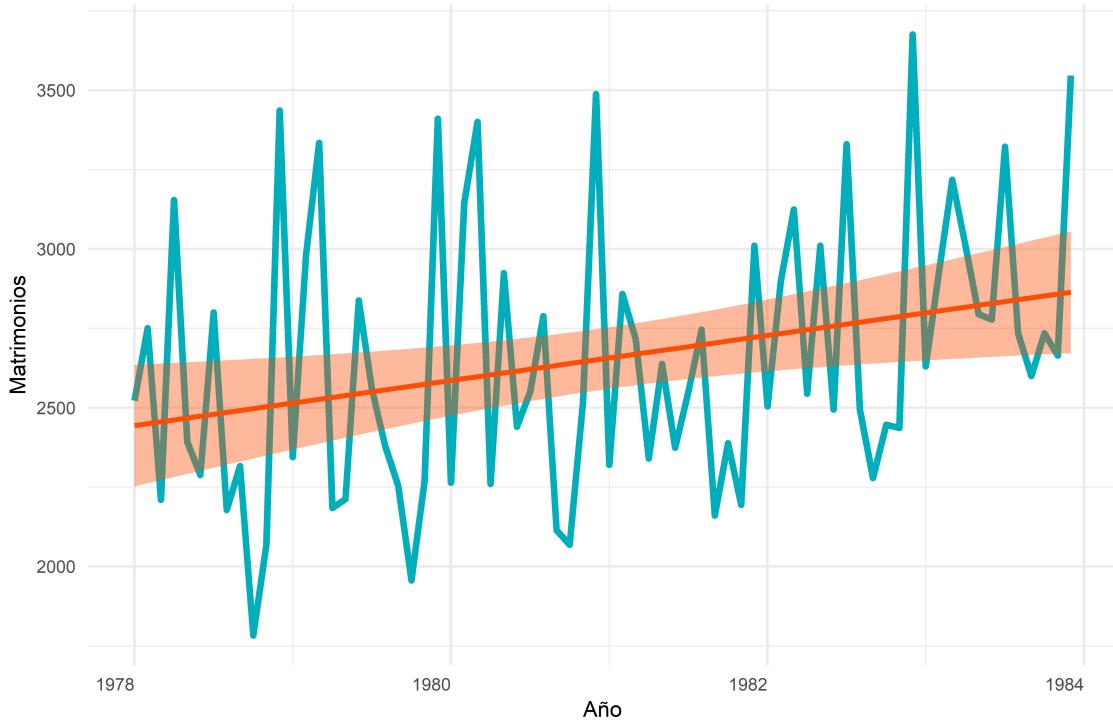
2.2.1 La tendencia-ciclo

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. De esta manera la forma matemática de la tendencia-ciclo para una serie cronológica se muestra en la ecuación (3).

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

Donde $\bar{y}_{t,m}$ representa el promedio móvil de orden m alrededor del momento t , que es un promedio de m observaciones consecutivas de la serie de tiempo y_t . En el caso de una media móvil de 12 periodos su cálculo está dado por $\bar{y}_{t,12} = \frac{y_{t-6} + \dots + y_{t-1} + y_t + y_{t+1} + \dots + y_{t+5}}{12}$, con $t = 7, \dots, n-5$, siendo n el total de observaciones de la serie cronológica. 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.

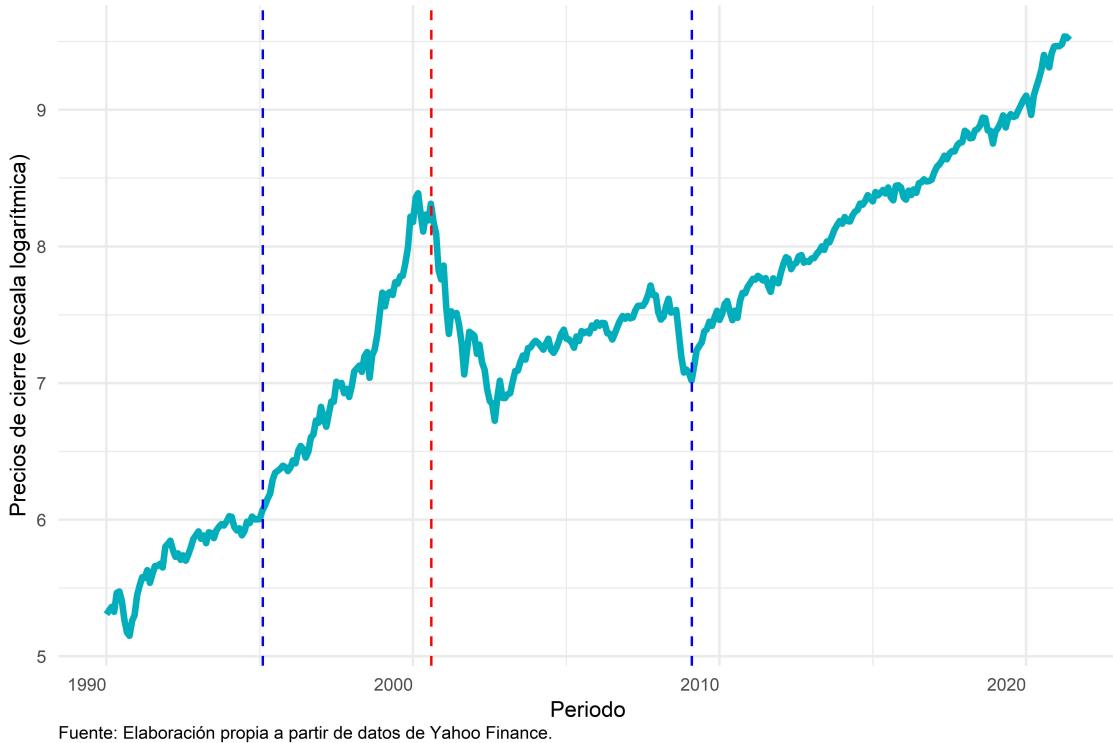
Figura 3: Tendencia del número de matrimonios en Costa Rica para el periodo 1978-1983



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 de los 100 valores de las compañías más importantes del sector de la industria de la tecnología, y sus ciclos suelen presentar un auge, seguido por un descenso que, posteriormente, se vuelve una depresión, y que finalmente se convierte en una recuperación a su estado inicial. La Figura 4 muestra como el índice NASDAQ-100 inicia un auge alrededor de enero de 1995 (primera línea azul punteada), para luego experimentar una fuerte caída a partir de junio del año 2000 (línea roja punteada) y posteriormente iniciar un periodo de recuperación en enero del año 2009 (segunda línea azul punteada).

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



2.2.2 Componentes estacionales

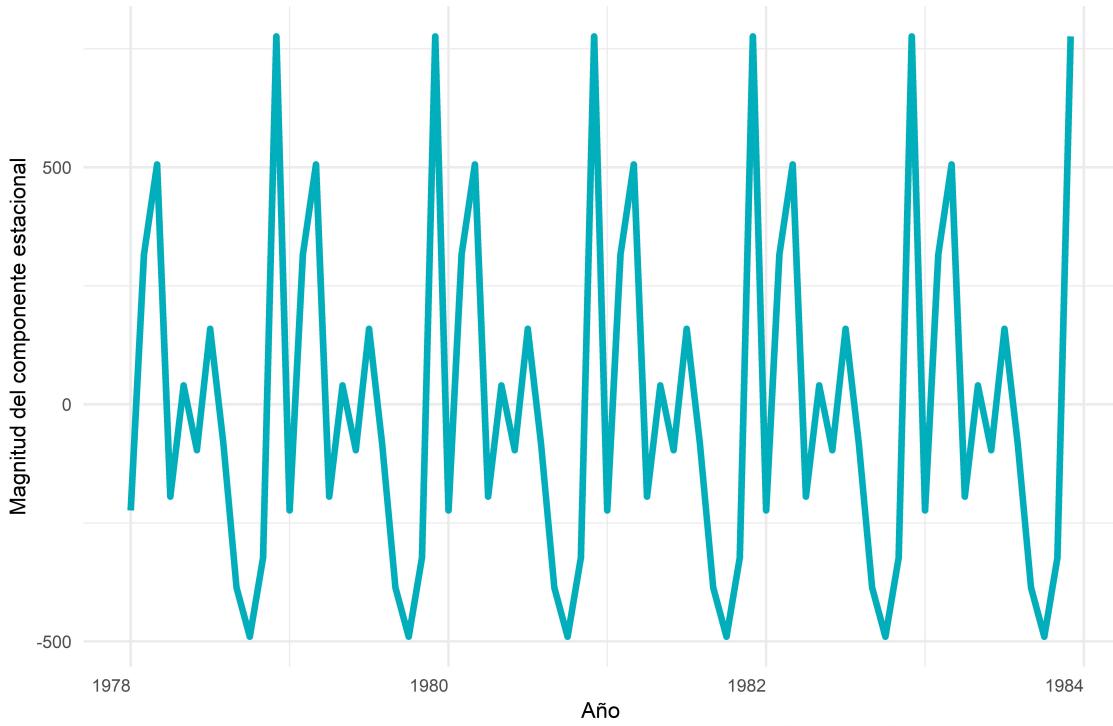
Calderón (2012) también se refiere a los cambios estacionales que se presentan en una serie de tiempo, los cuales se relacionan con las fluctuaciones naturales del fenómeno dentro de una temporada de observaciones. Visualmente los efectos estacionales pueden apreciarse en la Figura 2, en donde los picos más altos de turistas siempre se ubican entre los meses de diciembre y enero. Matemáticamente, el componente estacional puede calcularse utilizando la descomposición clásica como se indica en la ecuación (4):

$$S(t) = \psi_{t,a} - \psi^* \quad \left\{ \begin{array}{l} \psi_{t,a} = \frac{\sum_{i=0}^a \hat{y}_{t+a \cdot i}}{a} \\ \psi^* = \frac{\sum_{i=0}^n \hat{y}_{t+i}}{n} \\ \hat{y}_t = y_t - \bar{y}_{t,S,m} \\ \bar{y}_{t,S,m} = \frac{\sum_{i=1}^m \bar{y}_{t+i-m,S}}{m} \\ \bar{y}_{t,S} = \frac{\sum_{i=1}^S y_{t-i+\frac{S}{2}}}{S} \end{array} \right. \quad (4)$$

donde S representa la frecuencia estacional (por ejemplo cada 12 meses), a es la cantidad de periodos estacionales disponibles (por ejemplo, si se tienen 6 periodos completos de 12 meses, se tienen 6 años), m es la cantidad de periodos que se utiliza para centrar las medias móviles, y_t es la observación de la serie cronológica y en el momento t , $\bar{y}_{t,S}$ es el promedio móvil de S periodos

alrededor del momento t de las serie y_t , $\bar{y}_{t,S,m}$ es el promedio móvil centrado de en m periodos de la serie y_t , \hat{y}_t es la serie de tiempo y_t sin el efecto de la tendencia-ciclo, ψ^* es el promedio de los valores obtenidos para \hat{y}_t , y $\psi_{t,a}$ es el promedio de los a periodos estacionales sin el efecto de la tendencia-ciclo. 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

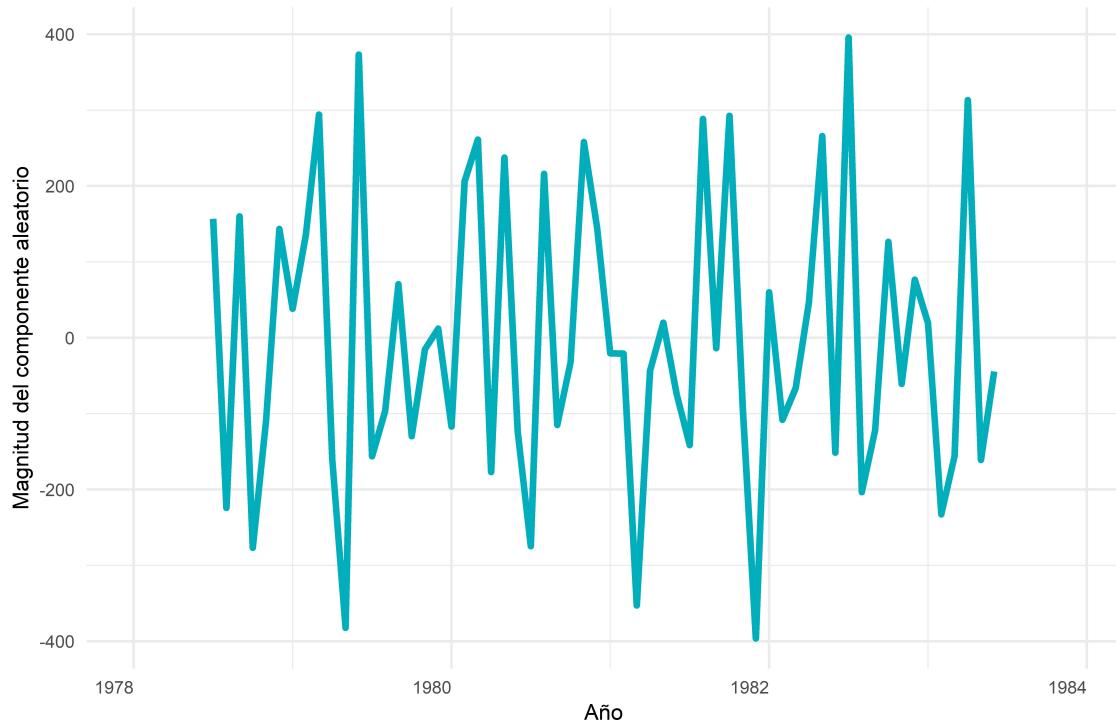


2.2.3 Componente irregular

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



2.3 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 pues carecen de aleatoriedad, razón por la cual el proceso puede definirse fácilmente mediante una ecuación matemática; las series cronológicas, sin embargo, carecen de esta condición. Por el contrario, los procesos no determinísticos son aquellos que no pueden describirse con exactitud mediante una ecuación matemática, sino que deben aproximarse debido al componente aleatorio que poseen de forma intrínseca. Una serie de tiempo puede tratarse de un proceso no determinístico porque usualmente, toda la información que se necesita para describir el proceso de manera determinística es desconocida, o bien porque la naturaleza de la información involucra la aleatoriedad, de esta manera, como los procesos no determinísticos consideran un aspecto aleatorio, pueden estimarse mediante leyes probabilísticas. [Hipel & McLeod \(1994\)](#) sugieren que una serie cronológica puede considerarse como una muestra aleatoria de una serie mucho más grande, es decir, una serie cronológica puede verse como una colección de variables aleatorias ordenadas de manera cronológica. La magnitud de estas variaciones aleatorias pueden variar en función del tiempo, esta condición se

conoce como un proceso estocástico (aleatorio) ([Elmabrouk, 2010](#)).

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. A partir del texto de [Ramírez \(2007\)](#), un proceso débilmente estacionario se define como se muestra en la ecuación (6).

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

El supuesto de estacionariedad busca simplificar la identificación del proceso con el objetivo de obtener un modelo adecuado para generar los pronósticos. De acuerdo con [Elmabrouk \(2010\)](#), se dice que una serie cronológica Y_t es fuertemente estacionaria cuando la distribución de probabilidad de dicho proceso en cualquier momento t es aproximadamente la misma para cualquier momento en el tiempo, mientras que es débilmente estacionaria si su media y su función de correlación no varía en el tiempo. 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 temporal es una observación o una realización de un proceso estocástico, éstas se encuentran sujetas a múltiples supuestos. En los modelos lineales definidos como procesos estocásticos uno 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; esta condición es deseable, sin embargo, en un modelo de series de tiempo un proceso estocástico puede no ser independiente al tener una estructura que genere un patrón reiterado en el tiempo. 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. Además, los coeficientes del proceso que gobierna la serie cronológica, como mencionan [McLeod \(1999\)](#), son polinomios que no deben compartir raíces comunes.

Además, existe un principio deseable que es quizá el que más debate genera, este 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.

2.4 Identificación del modelo

Los métodos más clásicos para la identificación del proceso que gobierna a una serie cronológica son las funciones de autocorrelación y autocorrelación parcial, las cuales sirven de indicador acerca de qué tan relacionadas están las observaciones unas de otras. Estas funciones ofrecen indicios sobre el orden de los términos para los modelos $AR(p)$, $MA(q)$ y para la diferenciación y, por ende, para la identificación de un modelo $ARIMA$ ([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).

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

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.

La función 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 (*PACF* por sus siglas en inglés), 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 periodos, 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 resultado de la ecuación (8).

$$r_{Y_t, Y_{t-k}} = r_k = \frac{E(Y_t Y_{t-k})}{\sigma_{Y_t} \sigma_{Y_{t-k}}} = \frac{\sum_{t=k+1}^n (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2} \quad (8)$$

De lo anterior se deduce entonces que las k observaciones previas pueden utilizarse para obtener el valor de la serie cronológica en el momento t , como muestra la ecuación (9), por lo que la formulación matemática de la función de autocorrelación parcial de k rezagos de la serie y_t es como se muestra en la ecuación (10).

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

$$PACF(y_t, k) = \frac{Cov((y_t|y_{t-1}, y_{t-2}, \dots, y_{t-k+1}), (y_{t-k}|y_{t-1}, y_{t-2}, \dots, y_{t-k+1}))}{\sigma_{(y_t|y_{t-1}, y_{t-2}, \dots, y_{t-k+1})}\sigma_{(y_{t-k}|y_{t-1}, y_{t-2}, \dots, y_{t-k+1})}} \quad (10)$$

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 (11).

$$\gamma_i = E[\phi_{k1}y_{t-1}y_{t-i} + \phi_{k2}y_{t-2}y_{t-i} + \dots + \phi_{kn}y_{t-n}y_{t-i} + u_ty_{t-i}] = \phi_{k1}\gamma_{i-1} + \phi_{k2}\gamma_{i-2} + \dots + \phi_{kn}\gamma_{n-i} \quad (11)$$

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

$$\rho_i = \phi_{k1}\rho_{i-1} + \phi_{k2}\rho_{i-2} + \dots + \phi_{kn}\rho_{n-i} + \dots \quad (12)$$

Alternando los distintos valores de k a partir de la ecuación (12), se obtiene el sistema de ecuaciones mostrado en (13).

$$\begin{aligned} \rho_1 &= \phi_{k1} + \phi_{k2}\rho_1 + \dots + \phi_{kn}\rho_{n-1} + \dots \\ \rho_2 &= \phi_{k1}\rho_1 + \phi_{k2} + \dots + \phi_{kn}\rho_{n-2} + \dots \\ \rho_3 &= \phi_{k1}\rho_2 + \phi_{k2}\rho_1 + \dots + \phi_{kn}\rho_{n-3} + \dots \\ &\vdots \\ \rho_k &= \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_{k-2} + \dots + \phi_{kn}\rho_{n-k} + \dots \end{aligned} \quad (13)$$

Como resultado del sistema de ecuaciones mostrado en (13), es posible hacer un replanteamiento en forma de un sistema matricial del cuál las autocorelaciones parciales pueden obtenerse a partir del despeje del vector Φ en (14).

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} \quad (14)$$

2.5 Modelos Autorregresivos Integrados de Medias Móviles (ARIMA)

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.

2.5.1 Ecuación de Wold

Según Sargent (1979a), cualquier proceso estacionario puede ser representado mediante la ecuación (15):

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

donde $\forall \psi_j \in \mathbb{R}, \psi_0 = 1, \sum_{j=0}^{\infty} \psi_j^2 < \infty$, y ε_t representa un ruido blanco gaussiano i.i.d., es decir, $\varepsilon_t \sim N(0, \sigma^2)$; además, κ_t es el componente lineal determinístico 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, y como menciona Sargent (1979b), el asumir la estacionariedad de x_t descarta una posible temporalidad, pues la presencia de ésta implicaría que la media de x_t es función de t .

De lo anterior, si se omite la parte determinística κ_t de (15), 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. 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 estacionario: $\kappa_t = m$, donde m es una constante cualquiera. Así, la ecuación (15) puede sustuirse por:

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

y de (16) puede verificarse que,

$$E(x_t) = E \left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + m \right) = \sum_{j=0}^{\infty} \psi_j E(\varepsilon_{t-j}) + m = m \quad (17)$$

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 (1979a), 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:

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

De la ecuación (18) se desprende que el error de previsión asociado está dado por:

$$x_{T+h} - \hat{x}_{T+h} = \sum_{j=1}^{\infty} \psi_j \varepsilon_{T-h+1} \quad (19)$$

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, y es justamente la ecuación de Wold la que deberá ser evaluada por la metodología Box-Jenkins.

2.5.2 Metodología Box-Jenkins

La combinación de un $AR(p)$ y un $MA(q)$, descritos en las ecuaciones (21) y (22) respectivamente, como se mencionó al inicio de esta sección, generan los modelos autorregresivos de medias móviles, $ARMA(p, q)$, representados mediante la ecuación (20).

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

[Cochrane \(1997\)](#) menciona que los modelos $ARMA(p, q)$ suelen manipularse mediante lo que se conoce como operador de rezagos, denotado como $L y_t = y_{t-1}$. Esto significa que en un $AR(p)$ se tiene que $\varepsilon_t = \varphi(L)y_t$, mientras que en $MA(q)$ se tiene que $y_t = \theta(L)\varepsilon_t$, y por consiguiente en un $ARMA(p, q)$ se tiene $\varphi(L)y_t = \theta(L)\varepsilon_t$. Por lo tanto, de lo anterior se desprende que $\varphi(L) = 1 - \sum_{i=1}^p \varphi_i L^i$, y que $\theta(L) = 1 + \sum_{j=1}^q \theta_j L^j$.

Los modelos $ARMA$, sin embargo, solamente pueden ser utilizados en series cronológicas cuyo proceso es estacionario. Esto, en la práctica, es poco común, pues una serie de tiempo a menudo posee tendencias y ciertos patrones estacionales y, además, como menciona [Hamzaçebi \(2008\)](#), presentan procesos no estacionarios por naturaleza. Esta condición hace necesaria la introducción de una generalización de los modelos $ARMA$, la cual se conoce como los modelos $ARIMA$ ([Box et al., 1994](#)).

2.5.3 Modelos Autorregresivos

Un modelo autorregresivo de orden p , denotado como $AR(p)$, considera los valores futuros de una serie cronológica como una combinación lineal de las p observaciones predecesoras, un componente aleatorio y un término constante. [Hipel & McLeod \(1994\)](#) y [Lee \(n.d.\)](#) emplean la notación de la ecuación (21).

$$y_t = c + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t \quad (21)$$

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.

2.5.4 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 (22).

$$y_t = \mu + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (22)$$

Donde μ representa el valor medio de la serie cronológica y cada valor de $\theta_j (j = 1, 2, \dots, q)$ son los parámetros del modelo. Como los $MA(q)$ utilizan los errores pasados de la serie cronológica, se asume que estos son i.i.d. centrados en cero y con una variancia constante, siguiendo una distribución aproximadamente Normal, con lo cual este tipo de modelos pueden considerarse como una regresión lineal entre una observación determinada y los términos de error que le preceden (Agrawal & Adhikari, 2013).

2.5.5 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 (23).

$$\begin{aligned} \varphi(L)(1 - L)^d y_t &= \theta(L)\varepsilon_t \\ (1 - \sum_{i=1}^p \varphi_i L^i)(1 - L)^d y_t &= \left(1 + \sum_{j=1}^q \theta_j L^j\right) \varepsilon_t \end{aligned} \quad (23)$$

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.

2.6 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:

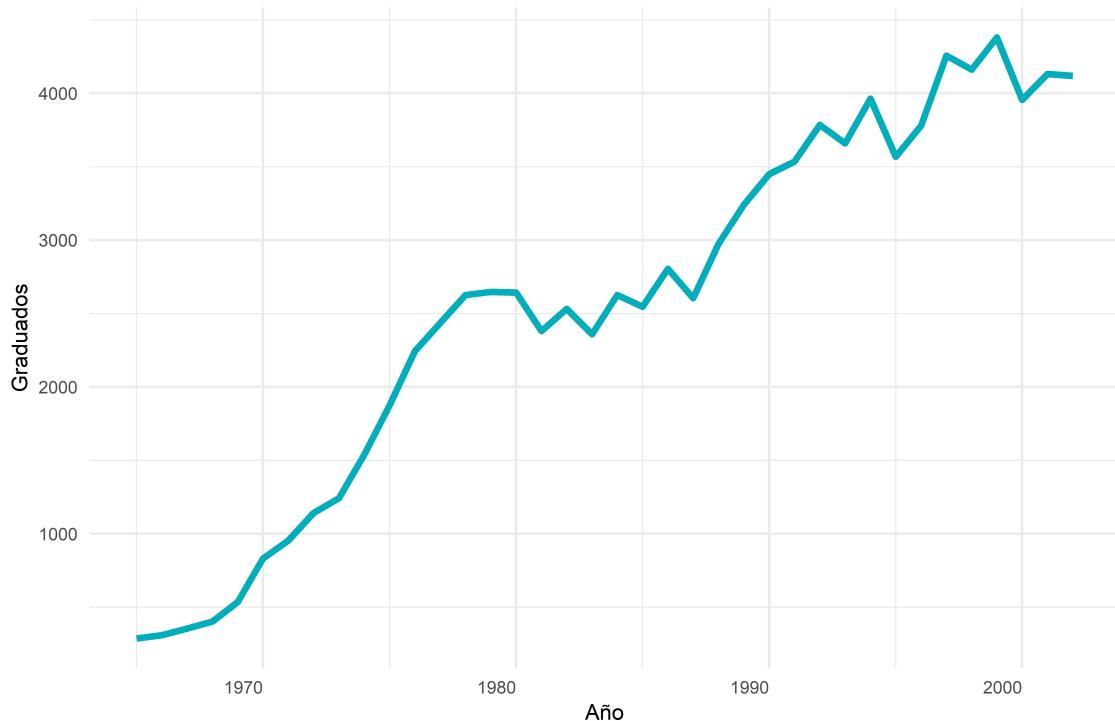
- Si la serie posee autocorrelaciones positivas en un amplio número de rezagos, entonces es posible que se requiera un orden más alto en el valor de d .
- Si la autocorrelación en $t - 1$ es menor o igual a cero, o si las autocorrelaciones resultan ser muy bajas y sin seguir algún patrón en particular,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

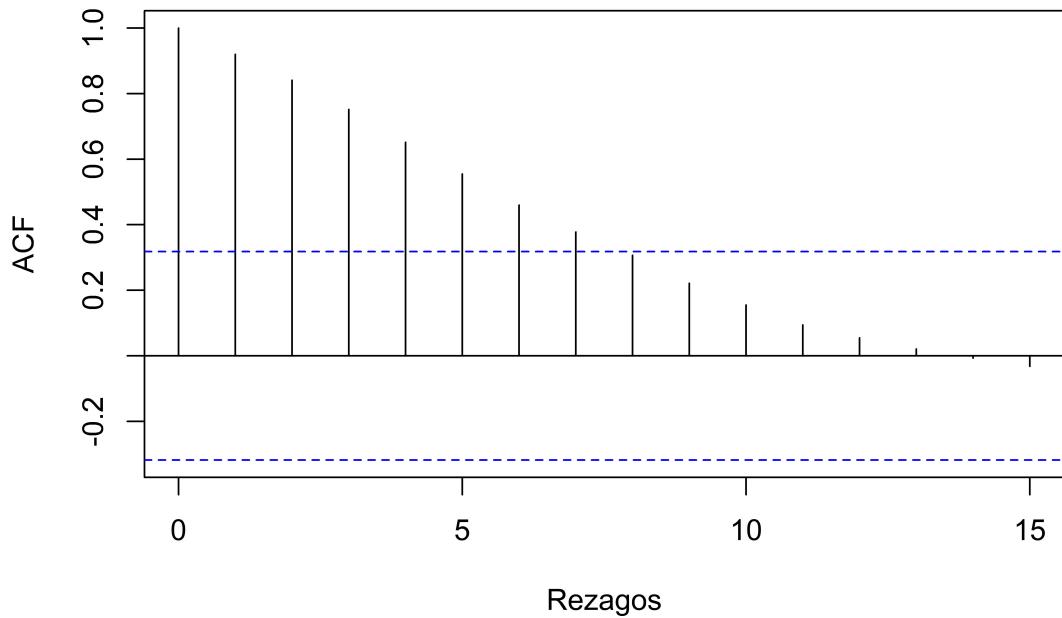
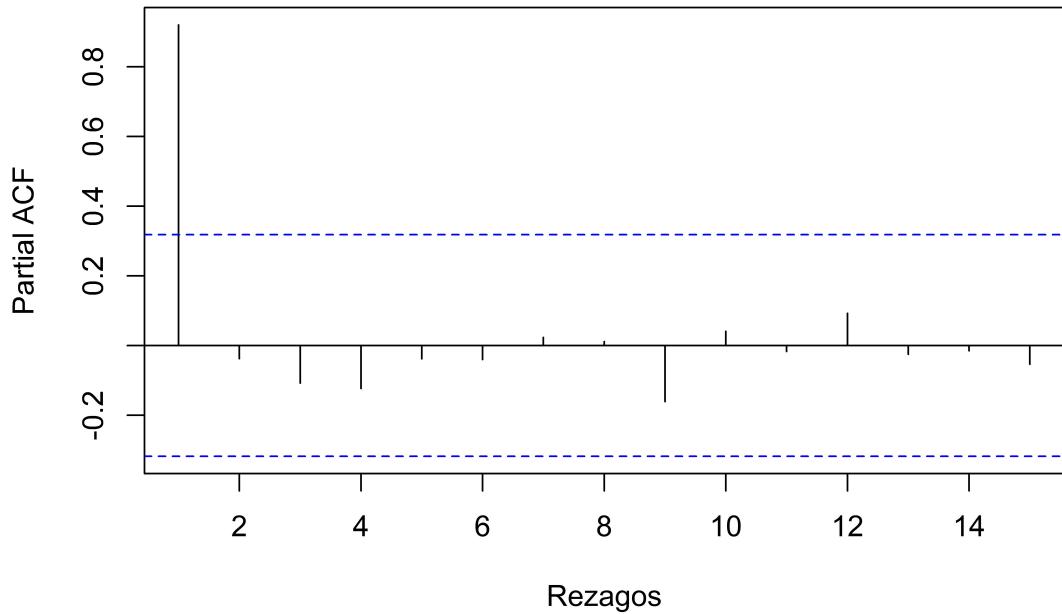


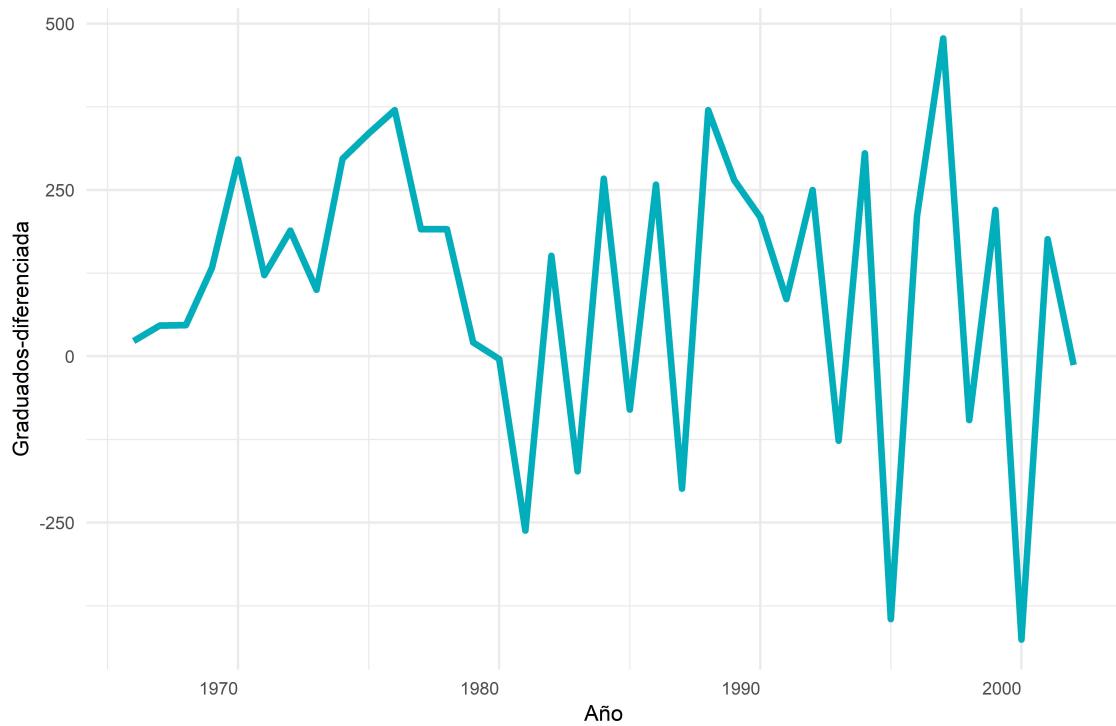
Figura 9: Función de autocorrelación parcial de la serie de graduados de la UCR



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

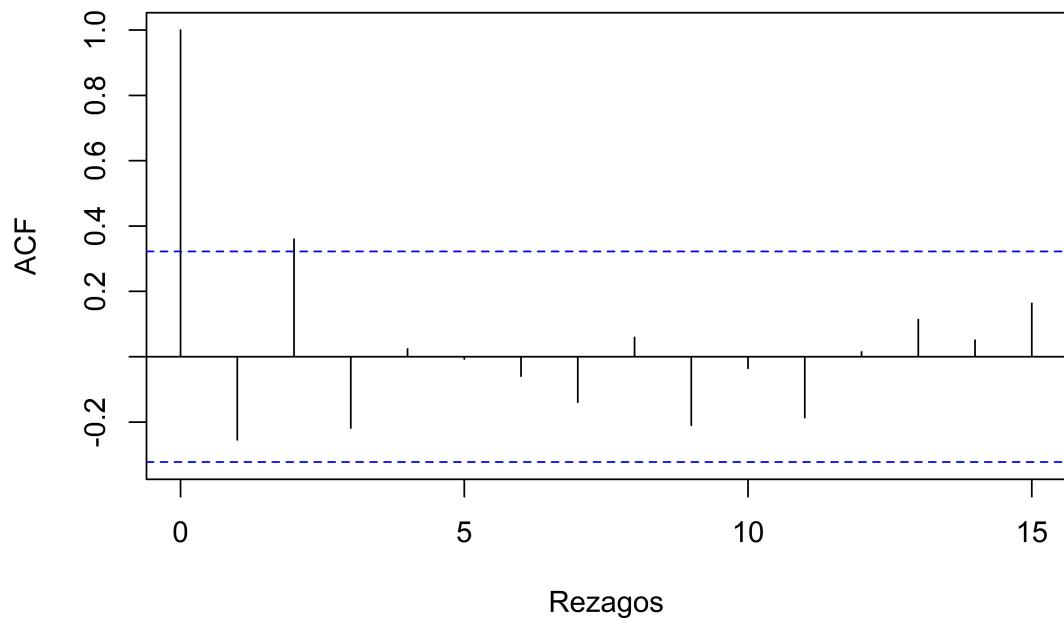
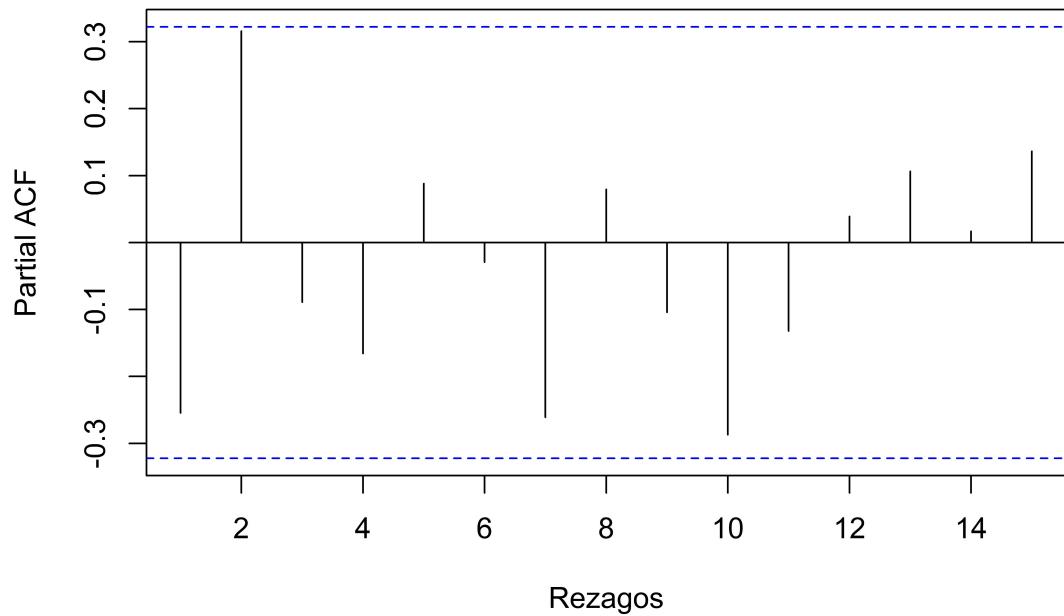


Figura 12: Función de autocorrelación parcial de la serie diferenciada de graduados de la UCR



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.

2.7 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 de 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 (24), 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 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 (25), 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 .

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

$${}_mC_r = \frac{m!}{r!(m-r)!} \quad (25)$$

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 por 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 una ecuación de estimación adecuada de entre varias candidatas, 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.

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 TMI interanual (TMII), que analiza la TMI 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

⁶<http://www.inec.go.cr/>

estudio interanual permite analizar de una mejor manera los cambios entre períodos. Esta forma de definir la TMII genera una autocorrelación natural en los datos, pues la información utilizada en el momento t coincide en alrededor de 92 % con la TMII del mes periodo inmediatamente anterior. Se analizará entonces la TMII desde el periodo Febrero 1989 – Enero 1990 hasta el periodo 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](#)).

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

⁷Técnica de suavizamiento mediante regresión local ponderada.

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 ([INEC, 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.

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, desarraigo, 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; este es un periodo suficientemente extenso para los objetivos de este estudio, por lo que obtener datos más recientes no generaría un aporte sustantivo a esta investigación.

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.

3.1.5 Herramientas analíticas

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).

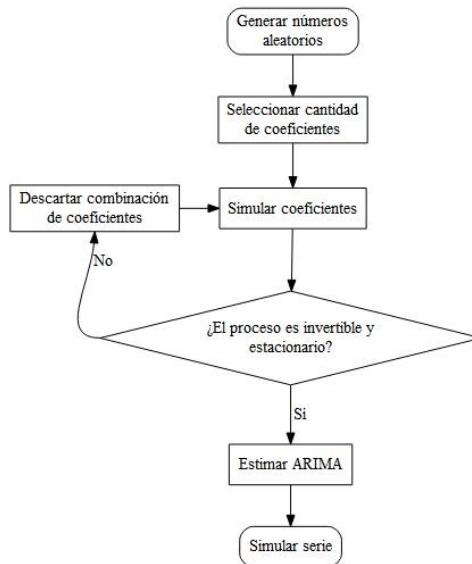
3.1.6 Procedimiento de simulación

Como parte de esta investigación, es necesario validar la identificación de los parámetros de los 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 (también mostrados en la Figura 13):

1. 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: $ARIMA(1, 0, 0)$, $ARIMA(1, 0, 1)$, $ARIMA(2, 0, 3)$, $ARIMA(4, 0, 2)$, $ARIMA(0, 0, 1)(0, 1, 1)_{12}$ y $ARIMA(2, 1, 4)(3, 0, 3)_{12}$.
2. Para cada uno de los procesos seleccionados, se generan valores aleatorios de una distribución uniforme con un valor mínimo de 0 y un valor máximo de 0.99. La cantidad de valores simulados depende de la cantidad de parámetros escogidos en p, q, P, Q . Estos valores aleatorios son transformados de manera tal que los polinomios de la parte autorregresiva y de medias móviles no comparta raíces unitarias, este es un proceso iterativo que se completa hasta que los coeficientes obtenidos tanto para la parte autorregresiva como para la parte de medias móviles generen un proceso invertible y estacionario.
3. Con 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 1..
4. Con cada modelo ajustado, se utiliza la función `simulate.Arima()` para generar 200 observaciones basadas en dichos modelos.

Figura 13: Diagrama de flujo del proceso de simulación de las series cronológicas.



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.

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 transformaciones 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, pues es la proporción más tradicional, sin embargo, esto puede cambiar de acuerdo al interés propio del(la) investigador(a), seleccionando una proporción distinta para las particiones.

3.2.3 Identificación y 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 y realiza la estimación mediante máxima verosimilitud.

3.2.4 Identificación y 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 mediante máxima verosimilitud 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 el 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 para esta investigación todas las permutaciones de parámetros p, d, q hasta tener como máximo un modelo $ARIMA(6, 1, 6)$, aunque el orden del modelo más grande puede permitir un mayor número de diferenciaciones o de parámetros. 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 (26):

$$\begin{array}{cc}
 \overbrace{p, d, q} & \overbrace{P, D, Q} \\
 \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 4 \\ 0 & 0 & 5 \\ 0 & 0 & 6 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \\ \vdots & \vdots & \vdots \\ 6 & 1 & 6 \end{array} \right] & \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{array} \right]
 \end{array} \quad (26)$$

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 (27) posee, en cada línea, una especificación de modelo a estimar:

$$\begin{array}{cc}
 \overbrace{p, d, q} & \overbrace{P, D, Q} \\
 \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 4 & 1 & 4 \end{array} \right] & \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 4 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 4 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 4 \end{array} \right]
 \end{array} \quad (27)$$

4. Debido a que los modelos a estimar cuentan con varias cantidades de parámetros en la parte no estacional y en la parte estacional, resulta pertinente generar una medida de referencia para conocer el nivel de sobreparametrización de los modelos en el espacio de valores paramétricos definido.

A partir de las estimaciones vía sobreparametrización hechas para este estudio, se calcularon los tiempos de estimación para modelos no estacionales y modelos estacionales. Estos tiempos se muestran en el Cuadro 6.

Para calcular el indicador de sobreparametrización (I_{sp}) primero se calcula el tiempo mediano de estimación, tanto para los modelos no estacionales como para los modelos estacionales. Para este estudio se obtuvo que el tiempo mediano para los modelos no estacionales fue de aproximadamente 7,424944 minutos, mientras que para los modelos estacionales el tiempo mediano de estimación fue de 36,0432 minutos, aproximadamente.

Luego, a cada fila de la matriz de valores paramétricos se le calcula un promedio ponderado a la cantidad de coeficientes en la parte no estacional y en la parte estacional, utilizando como

ponderador los tiempos de estimación descritos previamente. Este procedimiento se expresa en la ecuación (28),

$$I_{sp} = \frac{(p+q) \cdot m_n + (P+Q) \cdot m_e}{2 \cdot (m_n + m_e)} \quad (28)$$

donde p, q, P y Q son los términos del modelo ARIMA, m_n es el tiempo mediano expresado en minutos para modelos no estacionales, y m_e es el tiempo mediano expresado también en minutos para modelos estacionales. De esta manera, el valor de I_{sp} para un $ARIMA(3, 1, 1)(2, 0, 1)_S$ estaría dado por $\frac{(3+1) \cdot 7,424944 + (2+1) \cdot 36,0432}{2 \cdot (7,424944 + 36,0432)} \approx 1,58540673$.

Por último, el valor de I_{sp} es reescalado de forma tal que sus valores estén en el intervalo $[0, 100]$ para el espacio de valores paramétricos definido, donde el cero representa el modelo nulo y 100 el modelo más sobreparametrizado posible dentro del espacio de valores paramétricos definido. La razón de esto es poder limitar, si se desea, la cantidad de estimaciones que se deben realizar. Por ejemplo, si se define que la matriz de valores paramétricos llegue como máximo a un $ARIMA(6, 1, 6)(6, 1, 6)_S$, es posible incorporar una restricción para estimar solo aquellas permutaciones de modelos *ARIMA* cuyo indicador de sobreparametrización sea menor o igual a 60.

5. Con la matriz de valores paramétricos, como las mostradas en (26) y (27) y el indicador de sobreparametrización, 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.
6. 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.
7. 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.
8. Tras esto, se aplica un método de consenso 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, 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 consenso es utilizado para obtener un puntaje de cada modelo ARIMA, su cálculo se obtiene de la ecuación (29):

$$\min \left(\sum_i m_i \cdot w_j \right) \quad (29)$$

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 visual de los errores

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

3.2.7 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 precisión, aplicadas a los conjuntos de entrenamiento y validación. Existen múltiples medidas, [Adhikari et al. \(2013\)](#) menciona, entre otras, las siguientes:

3.2.7.1 AIC

Se calcula de la siguiente manera:

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

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

3.2.7.2 AICc

Su forma de cálculo se muestra en la ecuación (31)

$$AICc = -2\log L(\hat{\theta}) + 2k + \frac{2k+1}{n-k-1} \quad (31)$$

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

3.2.7.3 BIC

El último estadístico de bondad de ajuste se calcula como se muestran en la ecuación (32).

$$BIC = -2\log L(\hat{\theta}) + k \cdot \log(n) \quad (32)$$

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

3.2.7.4 MAE

El error absoluto medio se define mediante la ecuación (33)

$$\frac{1}{n} \sum_{t=1}^n |e_t| \quad (33)$$

3.2.7.5 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 (34) y (35).

$$\frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-1} \sum_{t=2}^T |Y_t - Y_{t-1}|} \quad (34)$$

$$\frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-m} \sum_{t=m+1}^T |Y_t - Y_{t-m}|} \quad (35)$$

Donde m es la temporalidad de la serie.

3.2.7.6 RMSE

Es la raíz del error cuadrático medio, como se define en la ecuación (36).

$$\sqrt{\frac{1}{n} \sum_{t=1}^n |e_t|^2} \quad (36)$$

3.2.8 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 precisión.

3.2.9 Tiempo de procesamiento

Para cada uno de los modelos estimados mediante sobreparametrización, se mostrará el tiempo de procesamiento requerido en cada serie cronológica, esto con el fin de evaluar la viabilidad del método propuesto para la obtención de resultados. Esta etapa de importante pues independientemente de los resultados obtenidos con la sobreparametrización, el tiempo de procesamiento debe mantenerse dentro de un margen razonable para quienes apliquen la técnica, ya que si se tardaran varios días en obtener los cálculos, sería un procedimiento poco práctico.

3.2.10 Resumen de la forma de análisis

De esta manera, cada una de las series cronológicas (simuladas y reales) serán sometidas a un análisis exploratorio en donde se analizará más en detalle el comportamiento de cada una. Posteriormente, a partir de la serie completa que sea sujeta a análisis, se ejecuta una partición de los datos con un 80 % para el conjunto de entrenamiento y el restante 20 % para validación, esto con el fin de obtener la mejor ecuación de estimación sugerida por cada una de las tres técnicas.

Una vez que se complete esta etapa, se completará una inspección visual del comportamiento de los errores con el fin de verificar que se cumplan las condiciones descritas en el Marco Teórico de esta investigación para posteriormente calcular las medidas de bondad de ajuste y de precisión para tanto para el periodo de entrenamiento como el de validación. Finalmente, se cierra el análisis de las series cronológicas simuladas y reales con una comparación de los valores predichos con respecto al conjunto de validación, esto con el fin de contrastar la calidad de los resultados.

4 RESULTADOS

En este capítulo se describe el procedimiento a seguir con cada una de las series cronológicas mencionadas en el apartado metodológico, tanto para las series simuladas como para 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 cumplen con las condiciones descritas en el Marco Teórico. Se describe además el proceso de partición de los datos tanto para ajustar los modelos como para validar los pronósticos y, a partir del procedimiento descrito en la metodología, se divide en dos etapas: Contraste de los ARIMA a partir de las series simuladas y contraste de los ARIMA para las series empíricas. Cabe destacar que para las series simuladas se realizó una realización de cada una, pues parte del interés de esta investigación está en poder analizar que tanto se acercan las estimaciones con cada método (ARIMA estándar, `auto.arima()` y sobreparametrización) a la especificación y coeficientes reales de cada proceso, y no en generalizar para cada proceso ARIMA en específico.

4.1 Contraste de métodos de estimación de los ARIMA para las series simuladas

Tras aplicar los pasos descritos en el apartado metodológico mediante la Figura 13, esta sección muestra los resultados obtenidos en sobre las series simuladas.

4.1.1 Análisis exploratorio

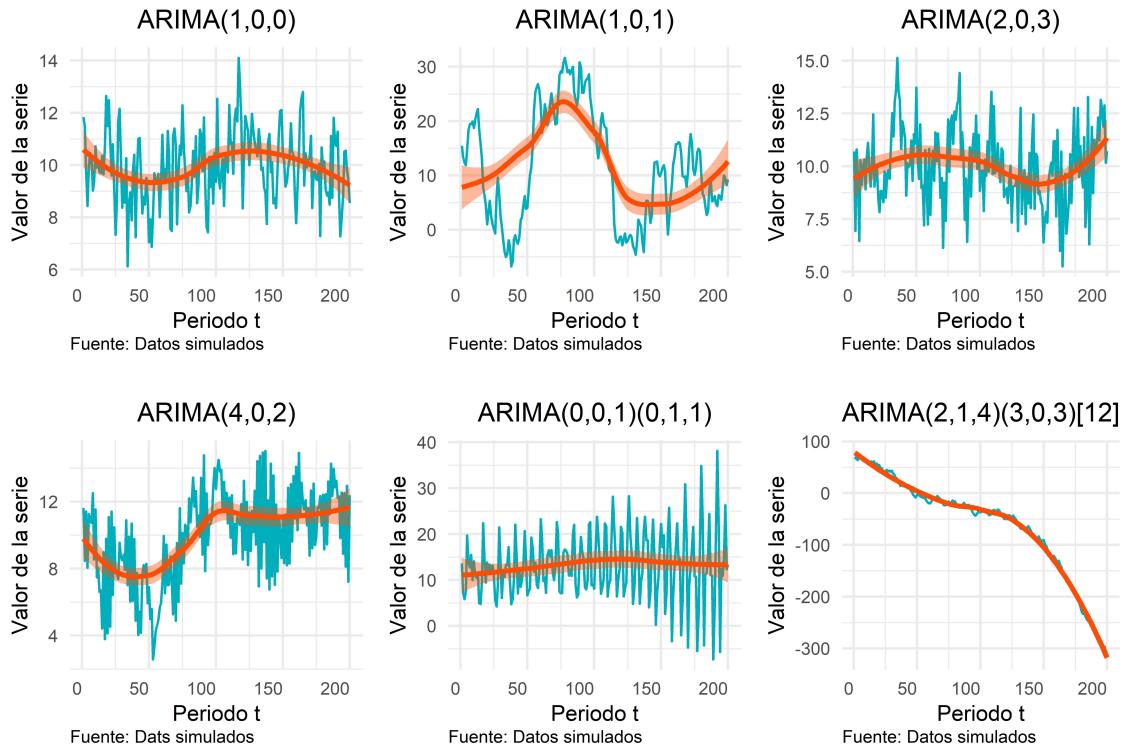
Inicialmente, se analizan las series cronológicas para verificar su comportamiento al ajustar un suavizado de Loess⁸ para buscar señales de tendencia y concavidad en los datos temporales, y buscar indicios de los procesos que gobiernan dichas series, además de verificar que se trate de series cronológicas estacionarias. En la Figura 14 se muestra el comportamiento general de cada una de las series cronológicas simuladas.

Para el *ARIMA*(1, 0, 0) puede verse como los datos simulados bajo este proceso se comportan de manera ligeramente oscilante a lo largo del tiempo sin necesidad de aplicar ninguna transformación a los datos. En el caso de los datos generados bajo un proceso *ARIMA*(1, 0, 1) se observa cómo se comportan de manera más oscilante que el caso anterior a lo largo del tiempo. Para los datos generados mediante un *ARIMA*(2, 0, 3) se observa como los datos simulados no parecen tener ninguna tendencia clara, una alternativa en este caso sería aplicar alguna transformación para volverla aún más estacionaria. En la Figura también se observa como los datos generados mediante un proceso *ARIMA*(4, 0, 2) parecen tener un cierto grado de tendencia, por lo que podría ser necesario aplicar alguna transformación.

⁸Técnica de suavizado mediante regresión local ponderada.

Con respecto a los procesos estacionales simulados, los datos del $ARIMA(0,0,1)(0,1,1)_{12}$ parecen ser aproximadamente estacionarios, pero esto no es del todo claro, una transformación no es estrictamente necesaria, pero puede aplicarse para corregir ligeramente la parte estacional; mientras que en los datos generados mediante un proceso $ARIMA(2,1,4)(3,0,3)_{12}$ se manifiesta una clara tendencia decreciente.

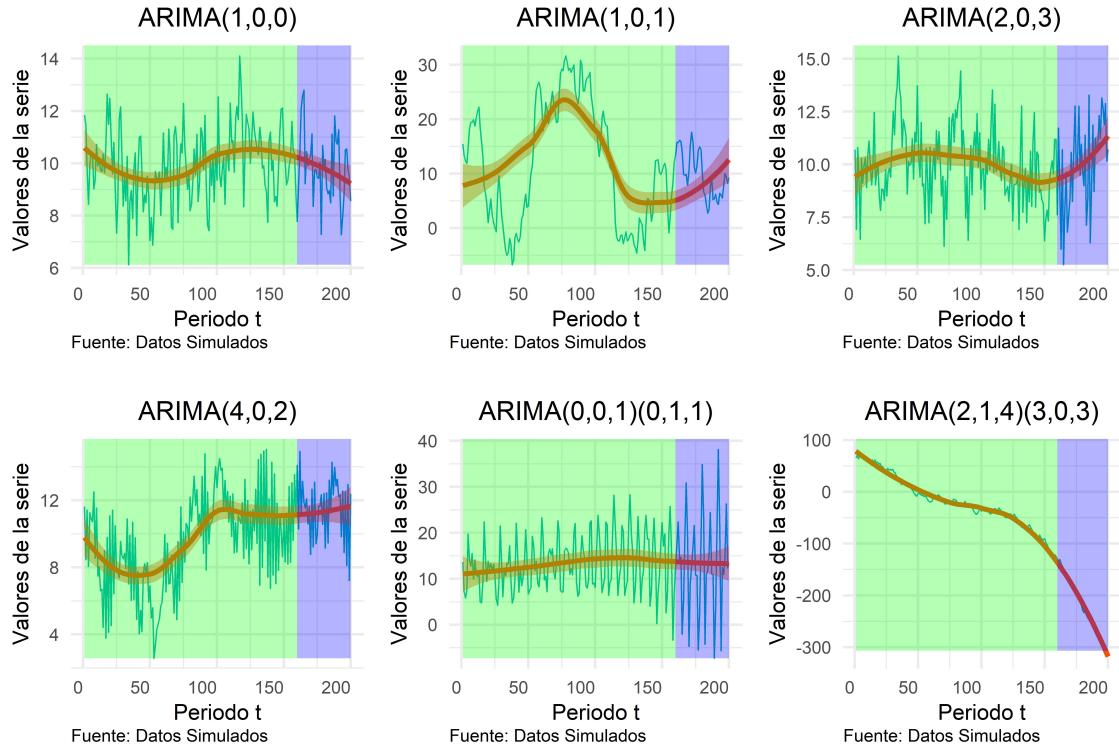
Figura 14: Comportamiento y tendencia de las series simuladas



4.1.2 Partición de los datos

Como se mencionó en la metodología, la partición utilizada consiste en tomar 80 % de las observaciones para generar la serie cronológica de entrenamiento para ajustar los modelos, mientras que el restante 20 % corresponde a la serie cronológica utilizada como validación. En el caso de los datos simulados, la ventana de observación de los datos seleccionados se muestra en la Figura 15.

Figura 15: Partición de los datos en los conjuntos de entrenamiento y validación para las series de tiempo simuladas



4.1.3 Identificación y estimación

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 $ARIMA(1,1,1)(1,1,1)_S$ 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.

Al ajustar modelos con la función `auto.arima()`, mediante un ARIMA estándar y mediante sobreparametrización, al igual que con cualquier otro método para estimar modelos ARIMA, se obtienen estimaciones de los coeficientes. En el caso de los datos simulados, donde se conoce de previo el verdadero proceso que gobierna la serie de tiempo, estos valores pueden compararse con los valores obtenidos.

El cuadro 1 resume los resultados obtenidos en la estimación de los coeficientes para los procesos simulados estacionales, mientras que los resultados de las series no estacionales se resumen en el cuadro 7. En la columna *Proceso original* se indica el proceso a partir del cual se generó la serie cronológica mediante simulación, en la columna *Coeficiente* se indica el coeficiente al cual pertenecen las estimaciones presentes en las columnas *Valor real* (el valor del coeficiente del proceso original), las columnas llamadas *Puntual* muestra la estimación puntual del coeficiente para cada uno de los tres métodos (`auto.arima()`, ARIMA estándar y Sobreparametrización), mientras que

las columnas *L.I.* (Límite Inferior) y *L.S.* (Límite Superior) representan del intervalo de confianza el 95 % para los coeficientes estimados. Como se ha mencionado, la notación más utilizada para los modelos *ARIMA* es $ARIMA(p, d, q)(P, D, Q)_s$, que en cada columna *Puntual* equivale a $ARIMA(ARX, d, MAX)(SARX, D, SMAX)_s$, con $X = \{1, 2, 3, 4\}$, según corresponda.

De esta manera, a modo de ejemplo, para los datos simulados a partir de un $ARIMA(2, 1, 4)(3, 0, 3)_{12}$, los coeficientes desde *MA1* hasta *MA4* representan los cuatro coeficientes de la parte de medias móviles no estacional, mientras que desde *SAR1* hasta *SAR3* son los tres coeficientes del modelo autorregresivo de la parte estacional, la celda del *Valor real* correspondiente al coeficiente *AR1* del proceso original $ARIMA(0, 0, 1)(0, 1, 1)_{12}$ está vacía porque el modelo original solo tenía un coeficiente en el modelo de medias móviles no estacional, es decir, $AR1 = 0$.

Cuadro 1: Coeficientes del proceso original y de los métodos de estimación de las series simuladas estacionales

Proceso original	Coeficiente	Valor real	auto.arima()			ARIMA estándar			Sobreparametrización		
			Puntual	L.I.	L.S.	Puntual	L.I.	L.S.	Puntual	L.I.	L.S.
ARIMA(0,0,1)(0,1,1)	AR1	-	-	-	-	0,37	0,21	0,52	-	-	-
	MA1	0,55	0,59	0,45	0,72	-1	-1,04	-0,96	0,59	0,45	0,72
	SAR1	-	-	-	-	-0,27	-0,69	0,15	-	-	-
	SMA1	0,41	0,32	0,15	0,5	0,63	0,26	1	0,32	0,15	0,5
	AR1	0,01	0,82	0,69	0,95	-0,86	-1,02	-0,71	0,81	0,68	0,95
	AR2	0,56	-	-	-	-	-	-	-	-	-
	MA1	-0,72	-1,07	-1,27	-0,88	0,71	0,51	0,91	-1,07	-1,25	-0,88
	MA2	0,03	0,49	0,3	0,68	-	-	-	0,51	0,33	0,69
	MA3	0,89	-	-	-	-	-	-	-	-	-
	MA4	0,79	-	-	-	-	-	-	-	-	-
ARIMA(2,1,4)(3,0,3)	SAR1	0,37	-0,99	-1,2	-0,78	-0,1	-0,44	0,24	-1,09	-1,27	-0,91
	SAR2	0,38	-0,57	-0,71	-0,44	-	-	-	-0,83	-1,02	-0,64
	SAR3	0,09	-	-	-	-	-	-	-	-	-
	SMA1	0,42	0,68	0,43	0,92	-0,31	-0,62	0	0,84	0,54	1,15
	SMA2	-0,07	-	-	-	-	-	-	0,4	0,05	0,76
	SMA3	0,55	-	-	-	-	-	-	-	-	-

Fuente: Elaboración propia a partir de datos simulados.

4.1.4 Análisis de los errores

Como se discutió en la Metodología, al ver la forma de los errores buscamos que estos se comporten como ruido blanco, es decir, que no tengan ningún patrón temporal en particular. El autocorrelograma de los residuos también puede usarse para este fin, pues el 95 % de los valores debería estar entre las dos líneas azules. Además, un histograma de los residuos es útil para saber si los

residuos siguen una distribución aproximadamente Normal; si no existen grandes desviaciones en este gráfico, entonces puede decirse que el cumplimiento de este supuesto es razonable.

En las Figuras 16 y 17 se visualizan los errores de los modelos estimados a partir de los datos simulados mediante los procesos $ARIMA(0,0,1)(0,1,1)$ y $ARIMA(2,1,4)(3,0,3)$, respectivamente, para los tres métodos de estimación: la función `auto.arima()`, ARIMA estándar y sobreparametrización, mientras que las Figuras 47, 48, 49 y 50 muestran lo mismo pero para las series simuladas no estacionales. En términos generales, los modelos generados mediante la función `auto.arima()` poseen el comportamiento esperado, aunque cuando se incorporan componentes estacionales los errores varían un poco más en los autocorrelogramas y el supuesto de normalidad, mientras que al utilizar un ARIMA estándar los modelos presentan una mayor cantidad de inconvenientes, particularmente en los autocorrelogramas, donde varios de ellos se salen de los márgenes. Por su parte, cuando se realiza la estimación mediante sobreparametrización, los modelos poseen el comportamiento esperado, incluso con menos problemas en la parte estacional.

Figura 16: Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso $ARIMA(0,0,1)(0,1,1)$

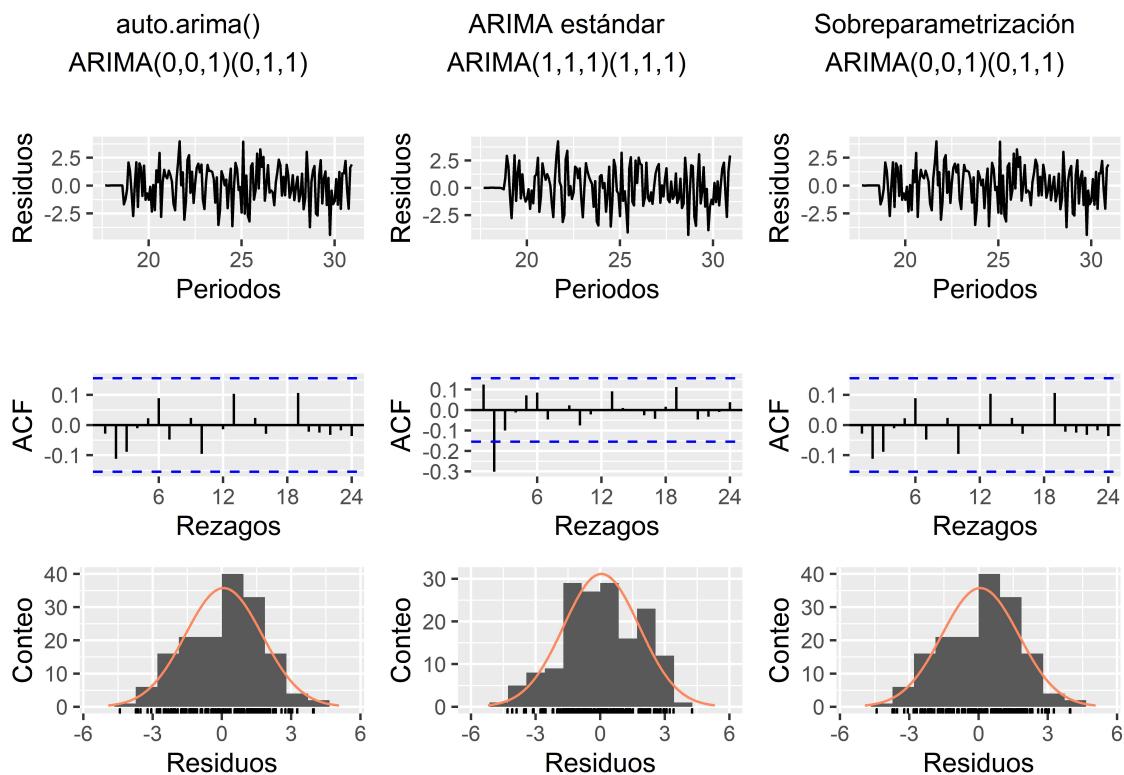
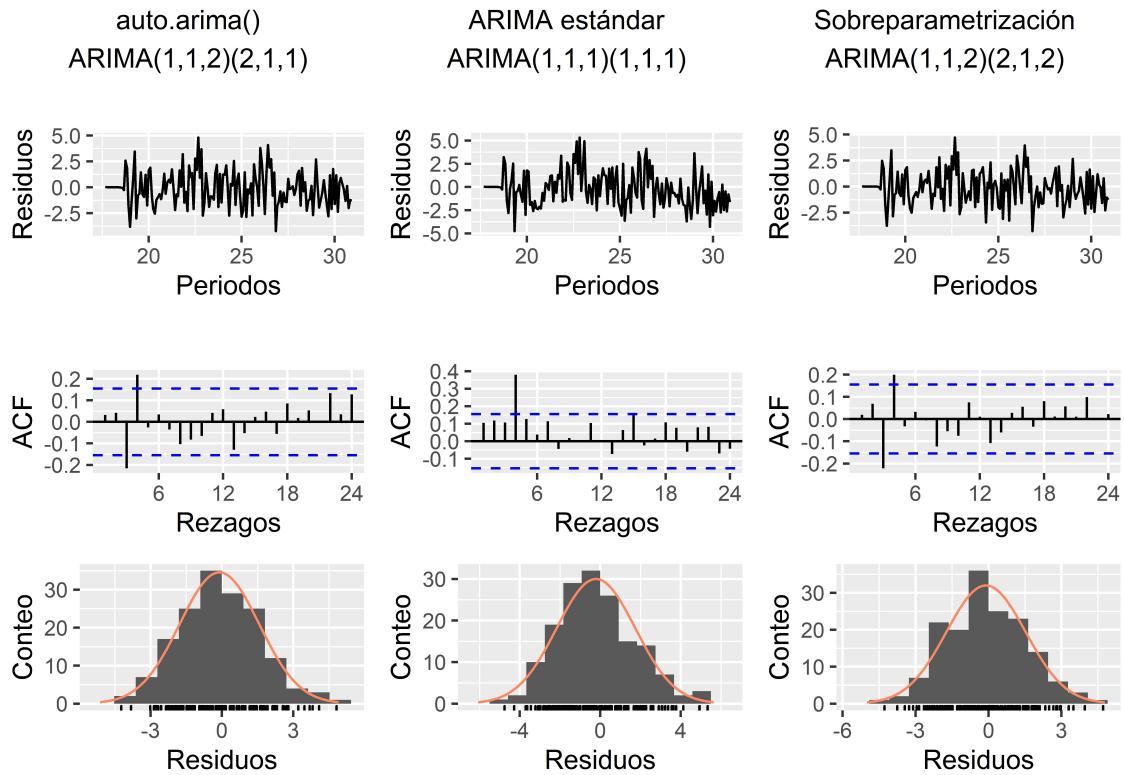


Figura 17: Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(2,1,4)(3,0,3)



4.1.5 Medidas de bondad de ajuste y de rendimiento

Se calculan ahora las medidas de bondad de ajuste y las medidas de precisión descritas en el apartado metodológico. A partir de las series cronológicas simuladas, el cuadro 2 muestra estas medidas para la series de tiempo simuladas a partir de procesos estacionales, mientras que el cuadro 8 hace lo propio para las series generadas a partir de procesos no estacionales. En la mayoría de los casos, las medidas de bondad de ajuste más bajas están presentes al realizar la estimación con sobreparametrización, y sucede lo mismo con respecto a las medidas de precisión.

Cuadro 2: Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación simulados a partir de dtos estacionales simulados

Proceso original	Datos	Estimación	AIC	AICc	BIC	RMSE	MAE	MAPE
ARIMA(0,0,1)(0,1,1)	Entrenamiento	auto.arima()	622,54	622,59	631,53	1,67	1,36	24,71
		ARIMA estándar	643,36	643,44	658,31	1,76	1,42	23,51
	Validación	Sobreparametrización	622,54	622,59	631,53	1,67	1,36	24,71
		auto.arima()	119,17	119,45	123,27	5,6	4,68	49,72
ARIMA(2,1,4)(3,0,3)	Entrenamiento	ARIMA estándar	124,81	125,28	131,64	5,82	4,97	53,61
		Sobreparametrización	119,17	119,45	123,27	5,6	4,68	49,72
	Validación	auto.arima()	632,18	632,28	653,11	1,68	1,34	6,69
		ARIMA estándar	678,92	679	693,87	1,97	1,57	7,84
		Sobreparametrización	626,06	626,19	649,99	1,64	1,29	6,35
		auto.arima()	317,01	317,72	326,58	32,1	30,61	13,34
	Validación	ARIMA estándar	374,03	374,51	380,87	42,99	41,46	18,15
		Sobreparametrización	272,98	273,83	283,92	24,72	23,24	10,08

Fuente: Elaboración propia a partir de datos simulados

4.1.6 Estimación en el periodo de validación

Las Figuras 18, 22, 51, 55 y 59, 63, muestran el ajuste y el pronóstico de cada uno de los modelos estimados con la función `auto.arima()`, con sobreparametrización y con el modelo *ARIMA* estándar. En todos los casos, la línea vertical punteada indica el inicio del periodo de pronóstico.

Además, las Figuras 19, 23, 52, 56, 60 y 64, representan el comportamiento de los errores estándar obtenidos de los pronósticos hechos con el mejor modelo sugerido por la función `auto.arima()`. Las Figuras 21, 25, 54, 58, 62 y 66 representan el comportamiento de los errores estándar obtenidos de los pronósticos hechos con el modelo *ARIMA* estándar. Y finalmente, las Figuras 20, 24, 53, 57, 61 y 65, representan el comportamiento de los errores estándar obtenidos de los pronósticos hechos con el mejor modelo sugerido por la sobreparametrización.

Figura 18: Pronóstico de los datos generados mediante un ARIMA(0,0,1)(0,1,1) según el método de estimación

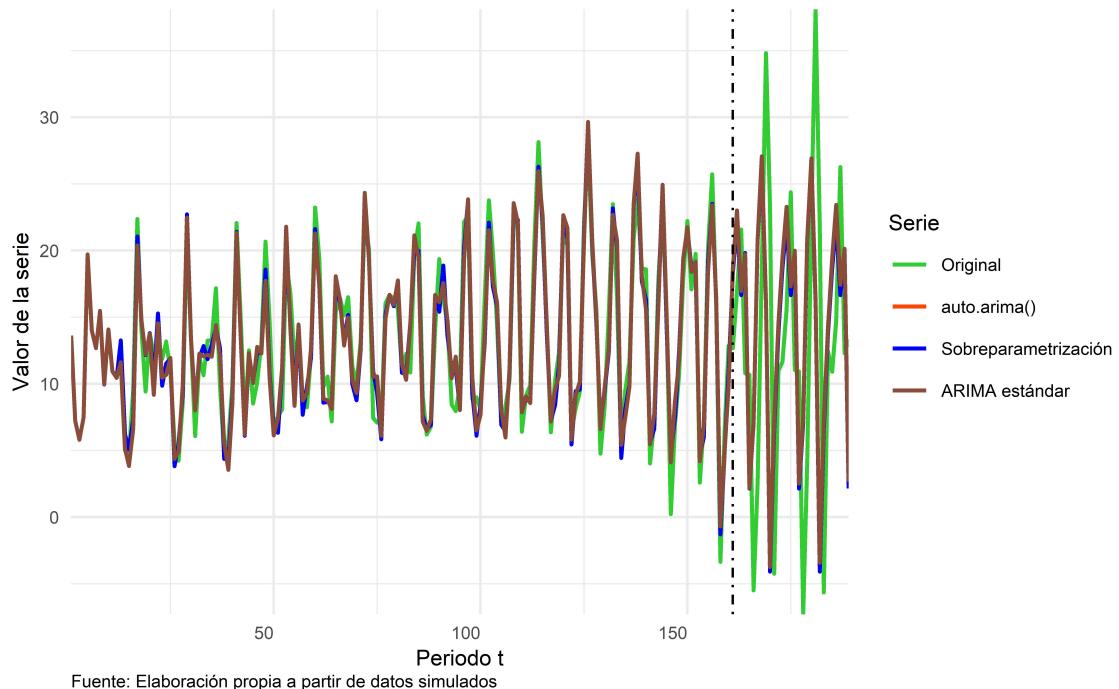
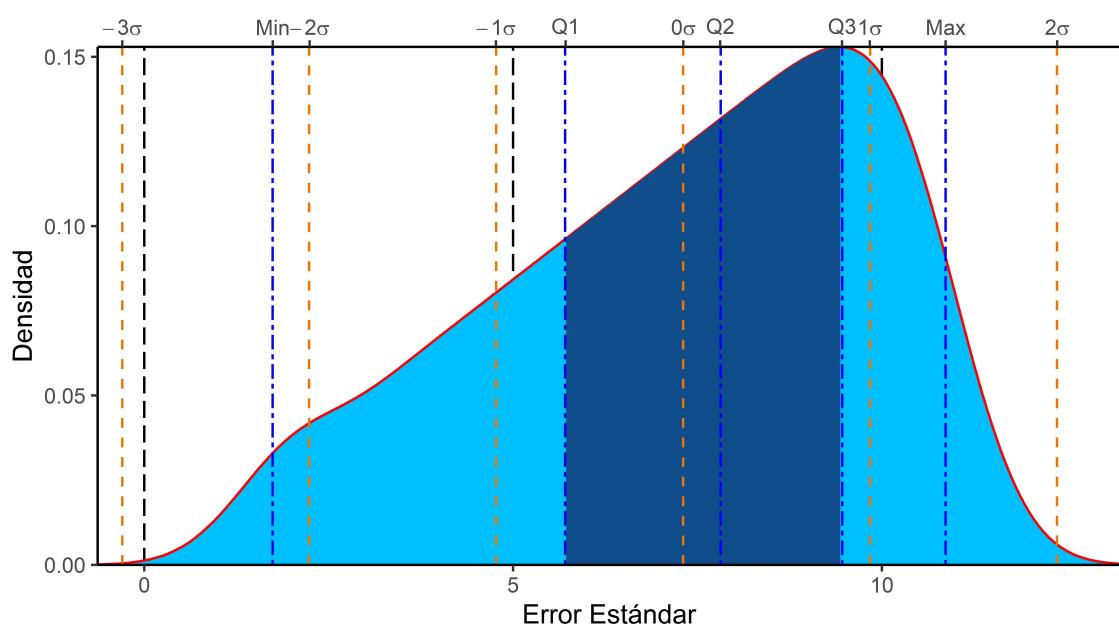


Figura 19: Errores est\'andares de los pron\'osticos obtenidos de los datos generados mediante un ARIMA(0,0,1)(0,1,1) con la funci\'on auto.arima()



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skewness	CV
1.74	5.71	7.81	9.46	10.87	2.53	2.24	-0.51	0.35

Figura 20: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(0,0,1)(0,1,1) con sobreparametrización

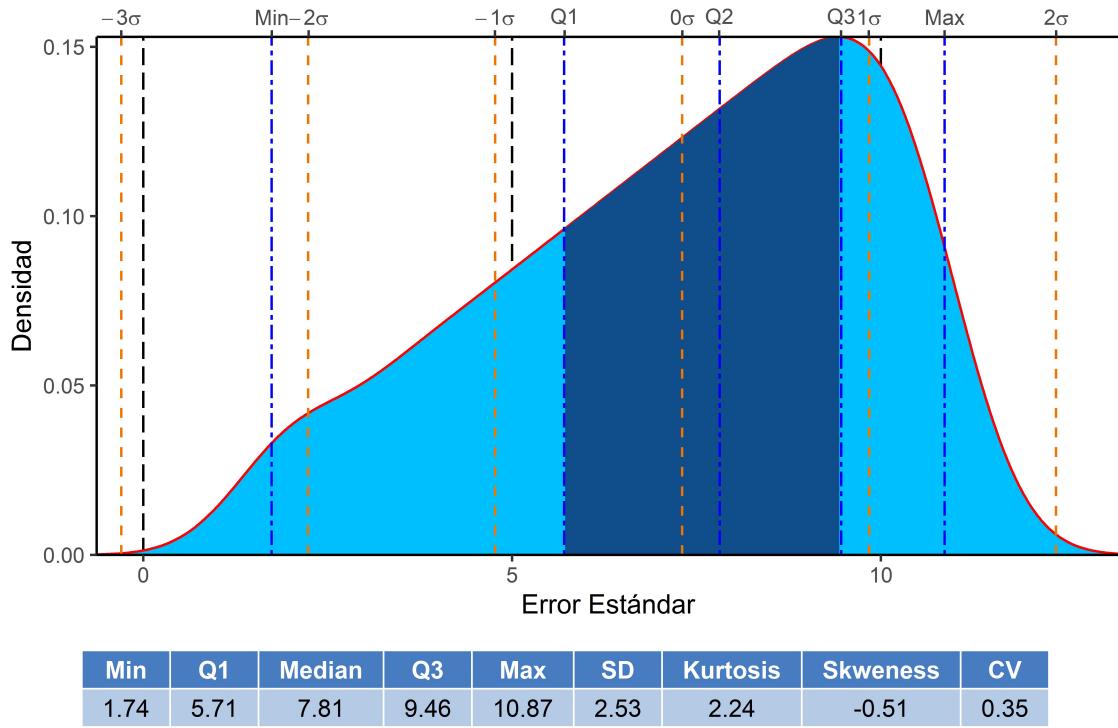


Figura 21: Errores estándares de los pronósticos obtenidos de los datos generados mediante un ARIMA(0,0,1)(0,1,1) con el modelo ARIMA estándar

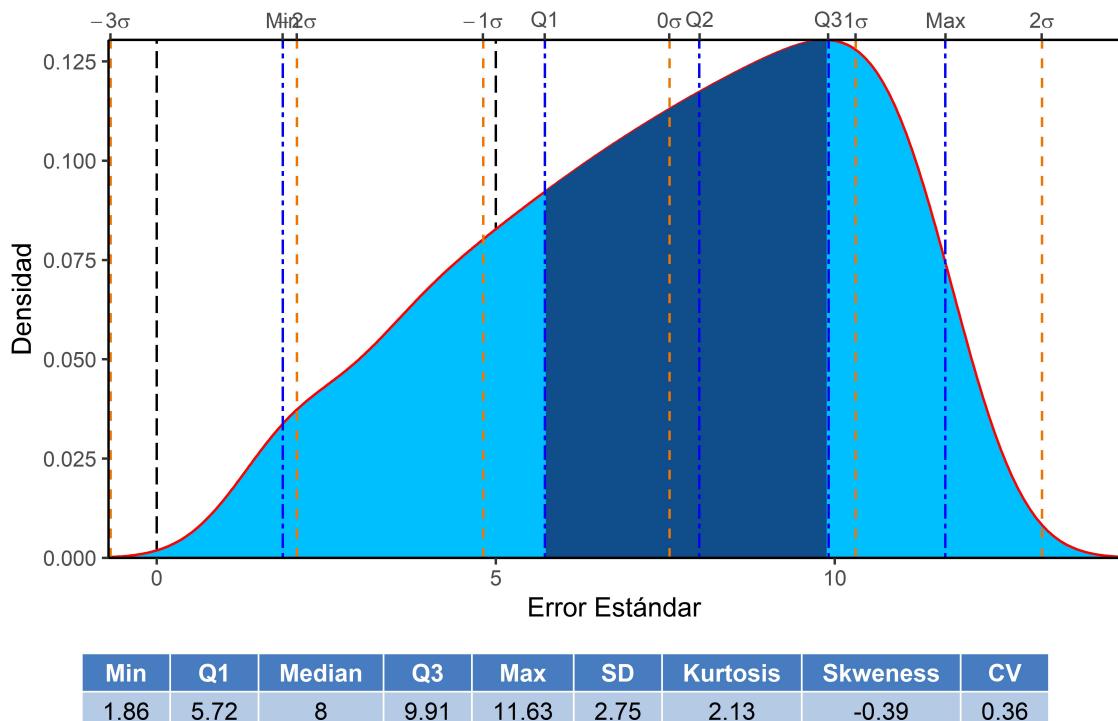


Figura 22: Pronóstico de los datos generados mediante un ARIMA(2,1,4)(3,0,3) según el método de estimación

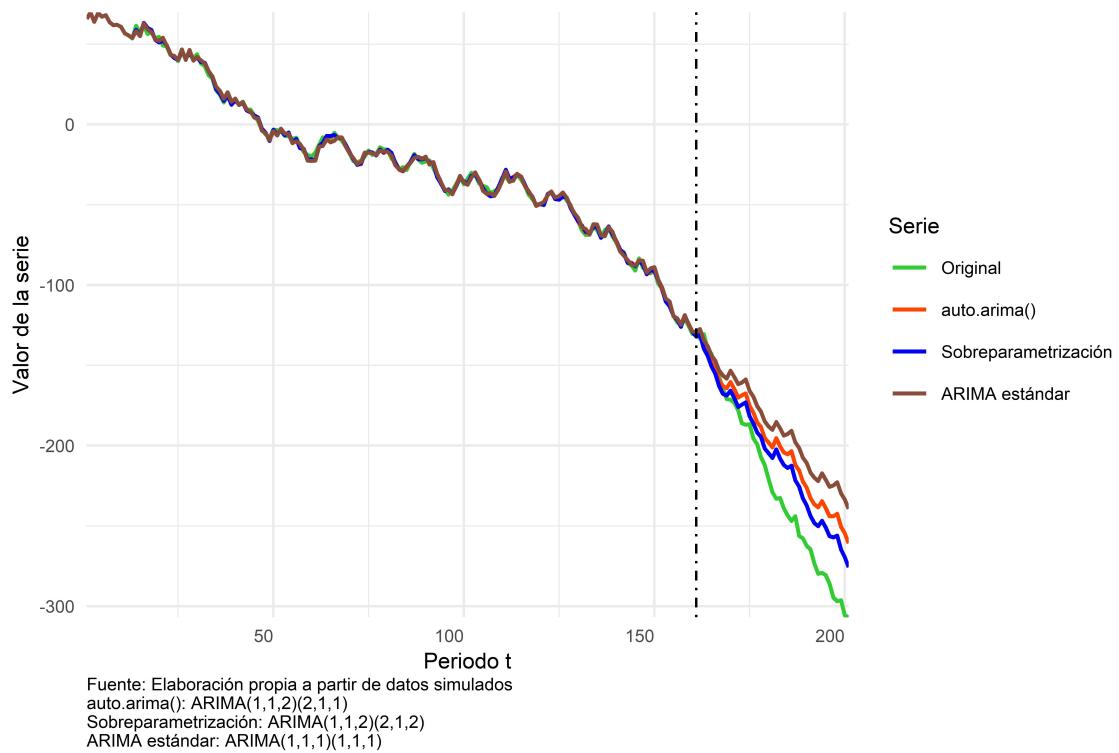


Figura 23: Errores est\'andares de los pron\'osticos obtenidos de los datos generados mediante un ARIMA(2,1,4)(3,0,3) con la funci\'on auto.arima()

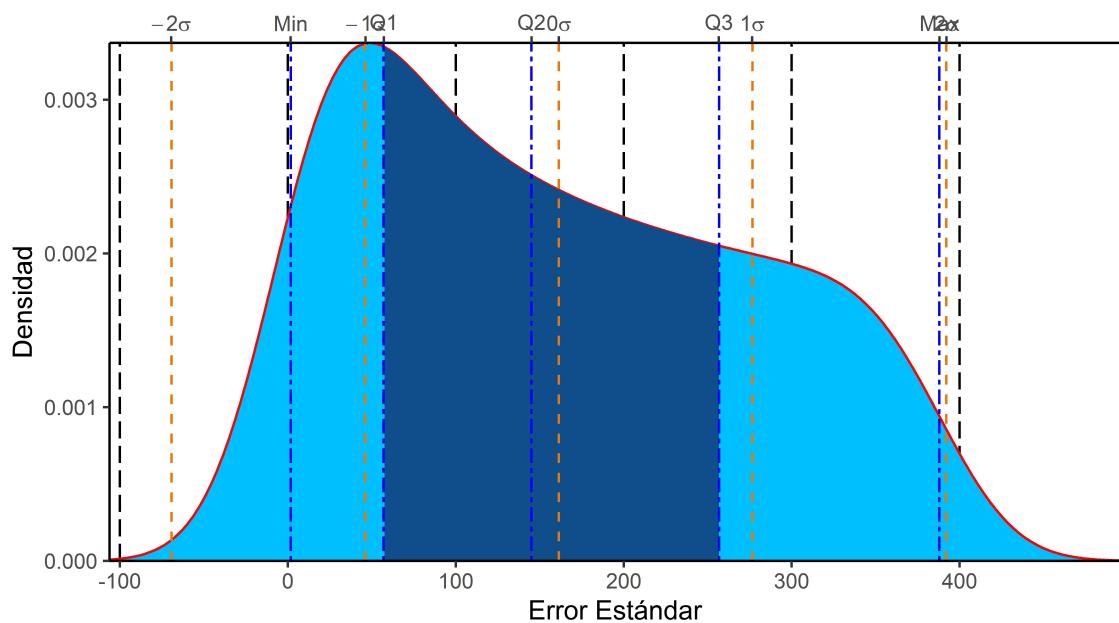


Figura 24: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,1,4)(3,0,3) con sobreparametrización

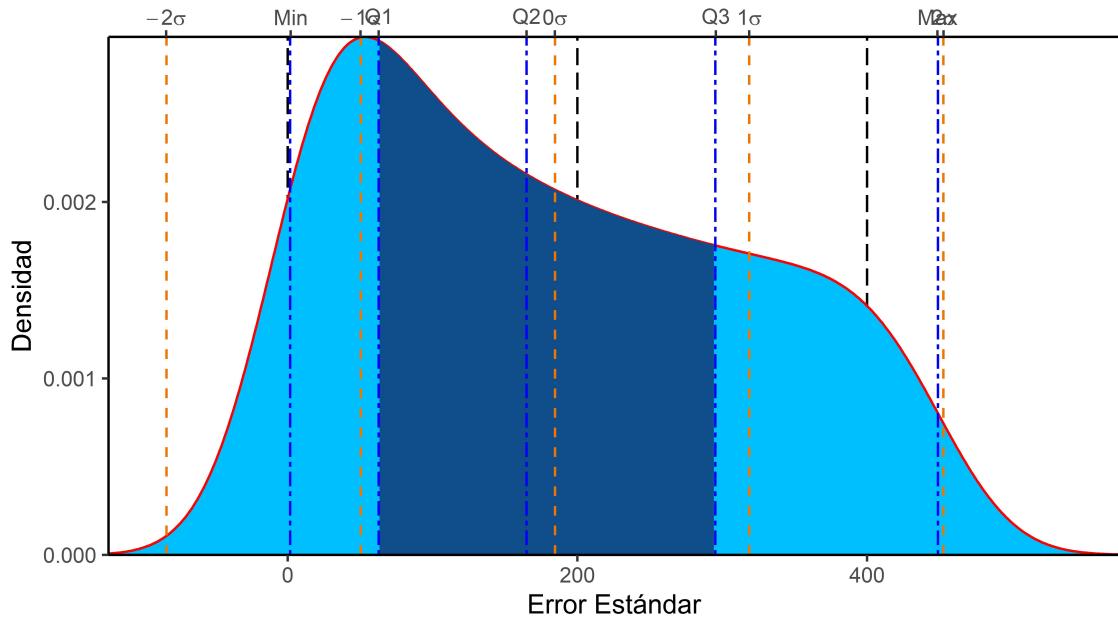
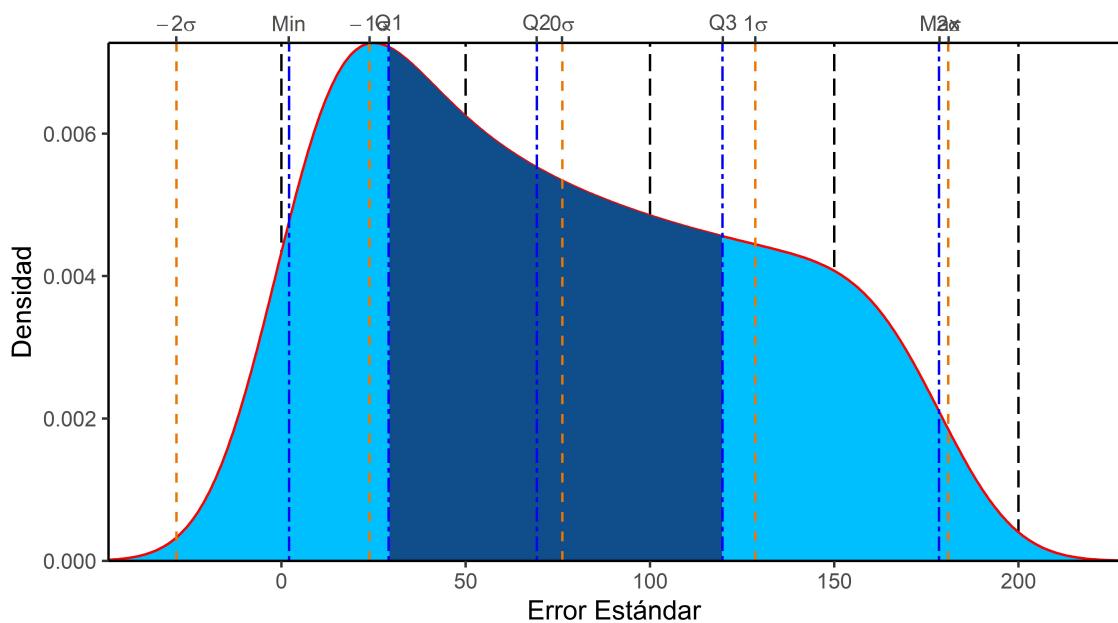


Figura 25: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,1,4)(3,0,3) con el modelo ARIMA estándar



4.2 Contraste de métodos de estimación de los ARIMA para las series empíricas

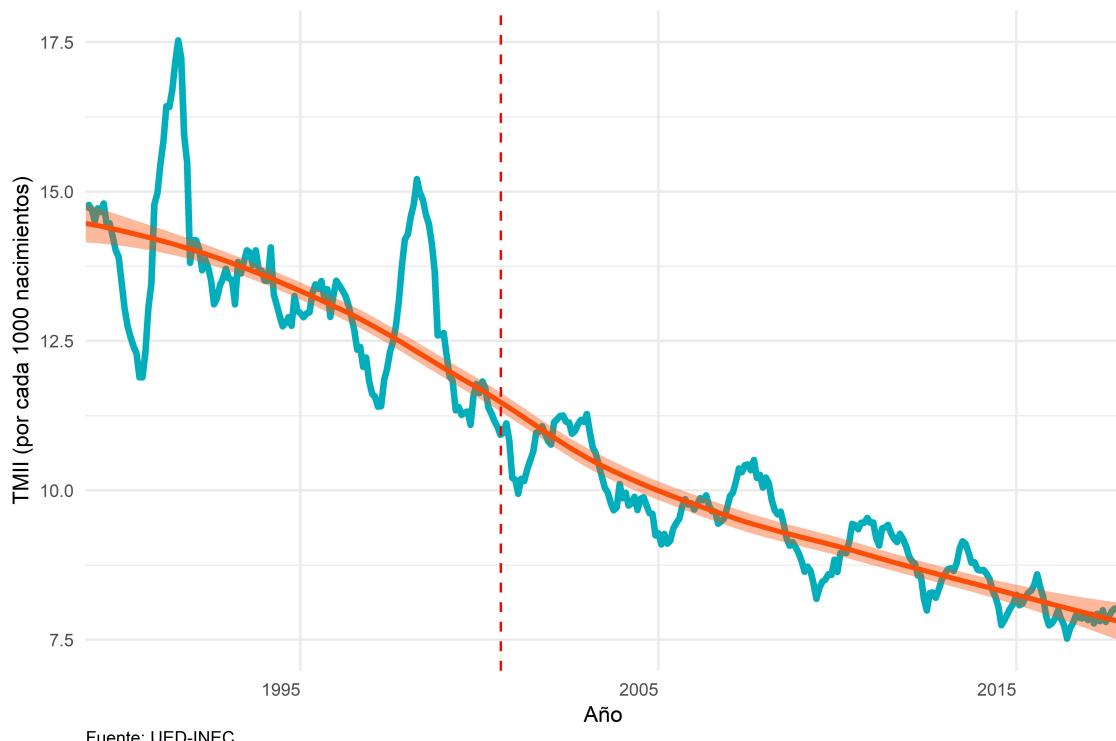
Para complementar la descripción de las series cronológicas reales utilizadas en este estudio, se muestran en este apartado los principales resultados obtenidos con las técnicas descritas en el apartado metodológico

4.2.1 Análisis exploratorio

4.2.1.1 Tasa de mortalidad infantil interanual

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 periodos siguientes se muestran en la Figura 26.

Figura 26: Tasa de Mortalidad Infantil Interanual 1989 - 2017

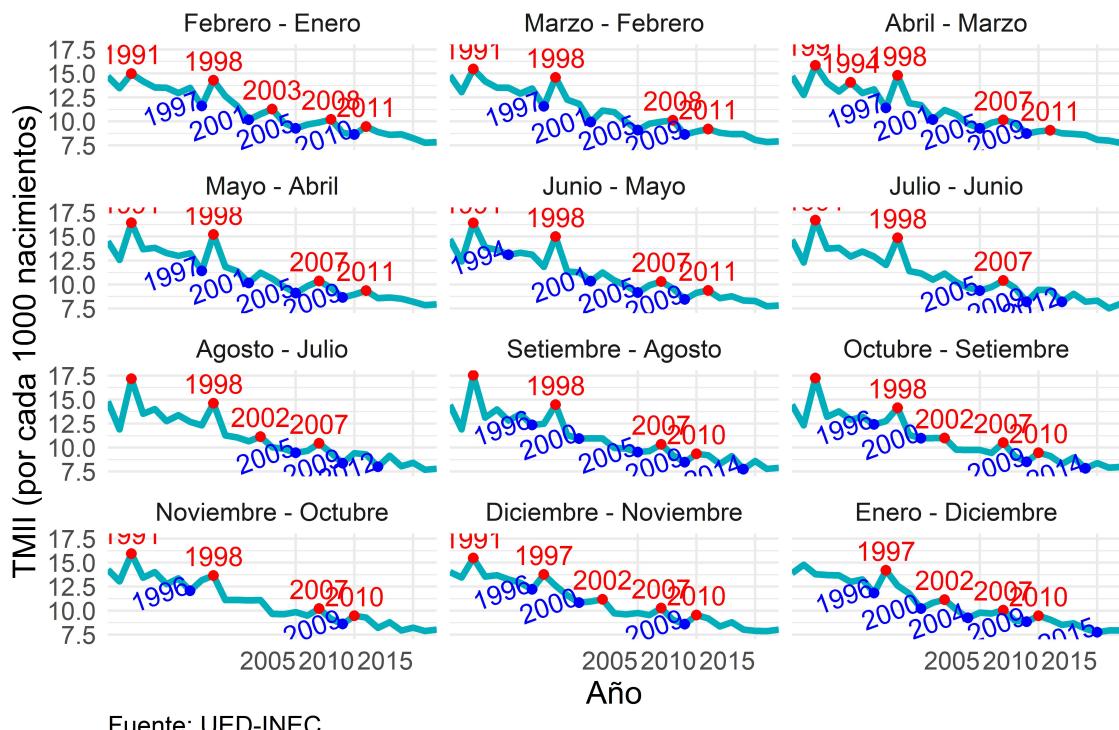


La serie muestra picos y valles pronunciados a lo largo de todo el periodo. A modo de visualización, se ajustó un suavizamiento de Loess para buscar señales de tendencia y concavidad en los datos temporales. La línea roja punteada se ubica aproximadamente en el mes de Julio del año 2000, pues a partir 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 27 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 26.

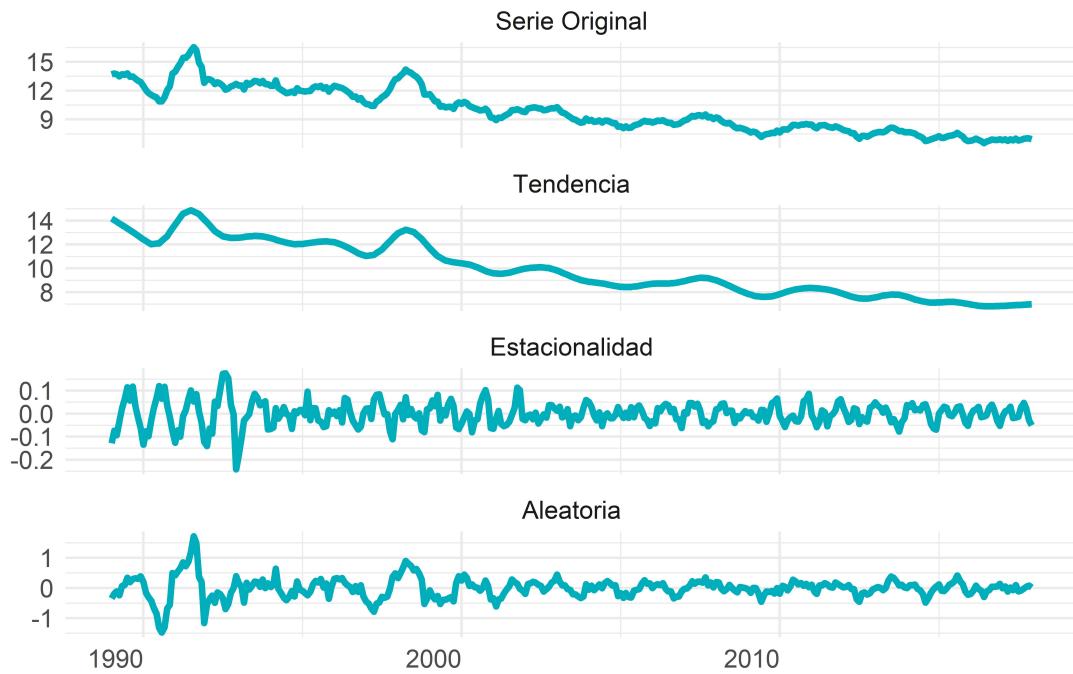
Figura 27: Tasa de Mortalidad Infantil Interanual 1989 - 2017 según períodos



Fuente: UED-INEC

La Figura 28 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 28: Descomposición de la TMII en el periodo 2000 - 2017



Fuente: UED-INEC

Por otro lado, las figuras 29 y 30 sugieren la presencia de procesos estacionales, mientras que en la parte no estacional se observa como la *ACF* va disminuyendo y el *PACF* se corta tras el segundo rezago, lo cual podría ser indicio de un proceso $MA(2)$ o superior.

Figura 29: Autocorrelación de los datos diferenciados de la TMII

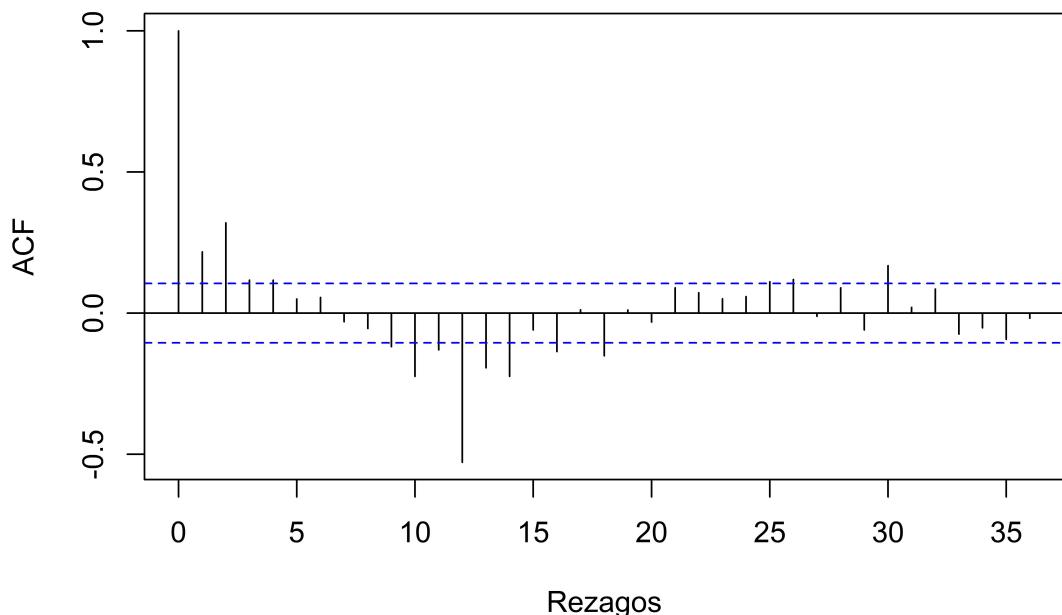
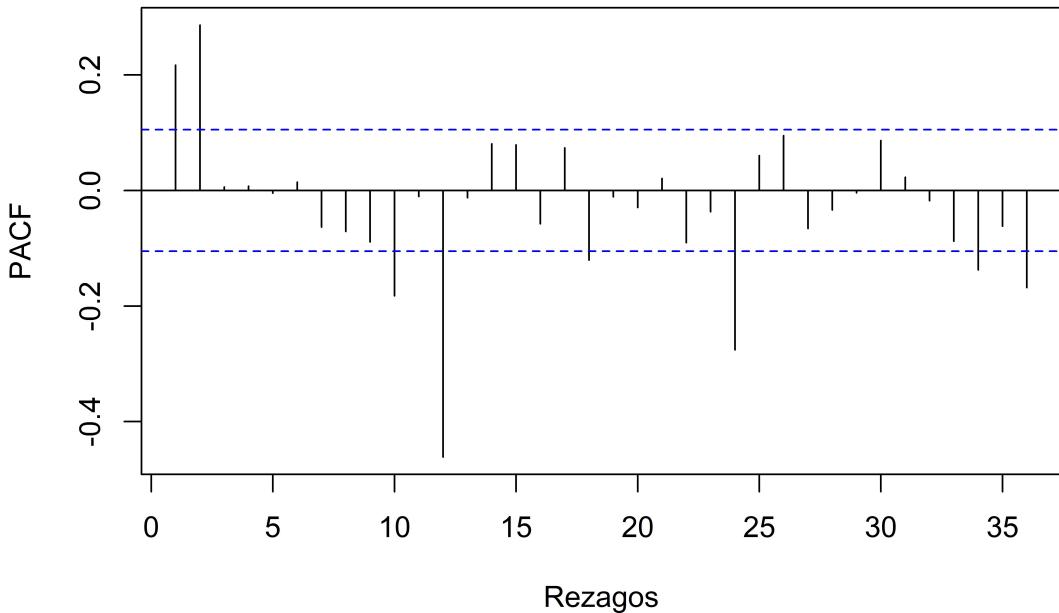


Figura 30: Autocorrelación parcial de los datos diferenciados de la TMII



4.2.1.2 Incentivos salariales del sector público

De manera análoga a lo visto en la TMII, la Figura 31 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 32 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 31: Incentivos salariales en el sector público entre los años 2007 y 2018

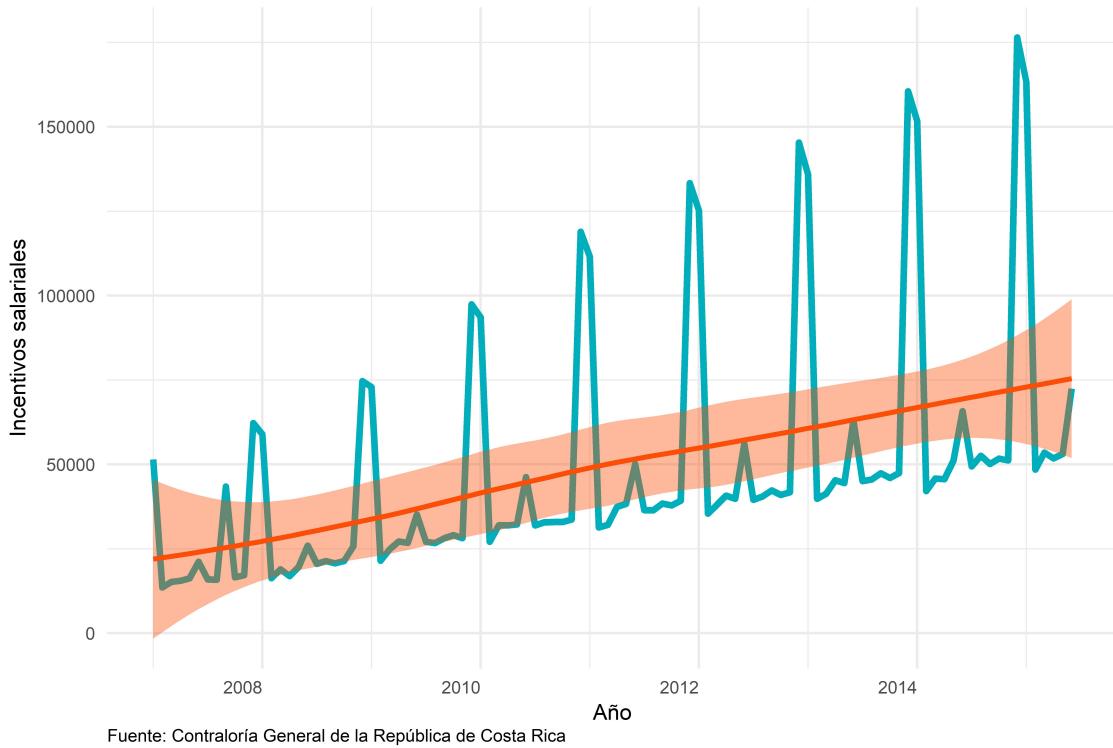
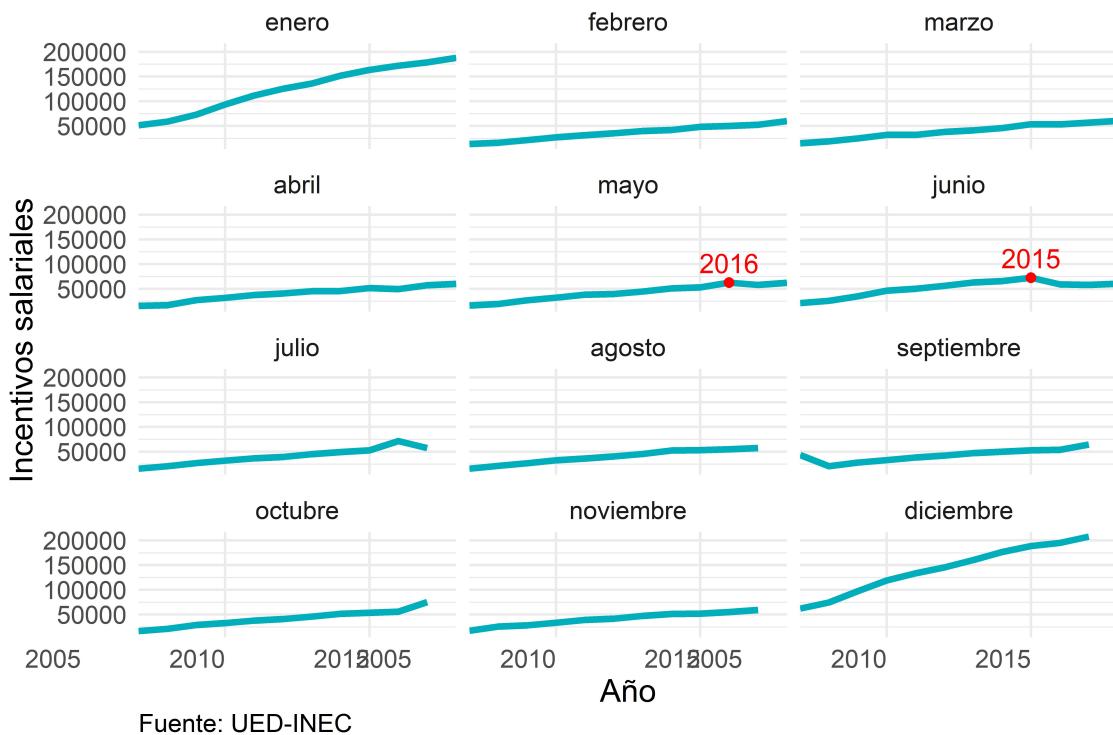


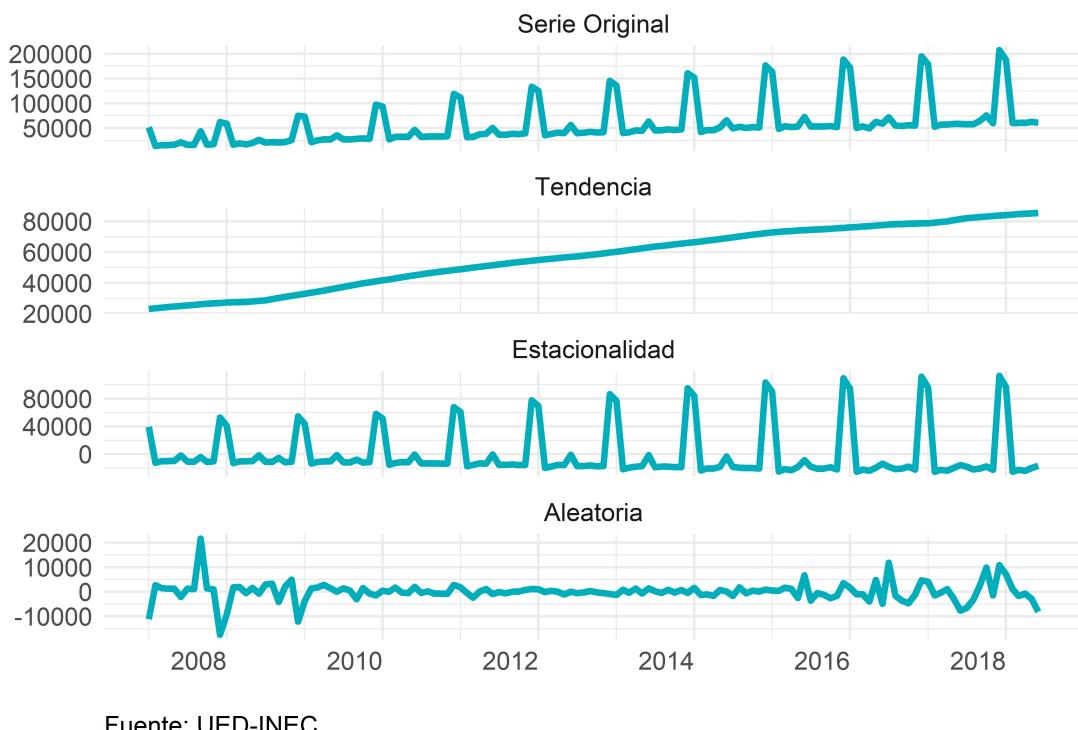
Figura 32: Incentivos salariales en el sector público entre los años 2007 y 2018 según mes



En la Figura 33 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, lo cual 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 33: Descomposición de la serie de Incentivos salariales en el periodo 2007 al 2018



Fuente: UED-INEC

Además, las figuras 34 y 35 no parecen sugerir un proceso en particular tanto para la parte estacional como para la no estacional, a pesar de que las figuras anteriores muestren un efecto periódico.

Figura 34: Autocorrelación de los datos diferenciados de la serie de incentivos salariales

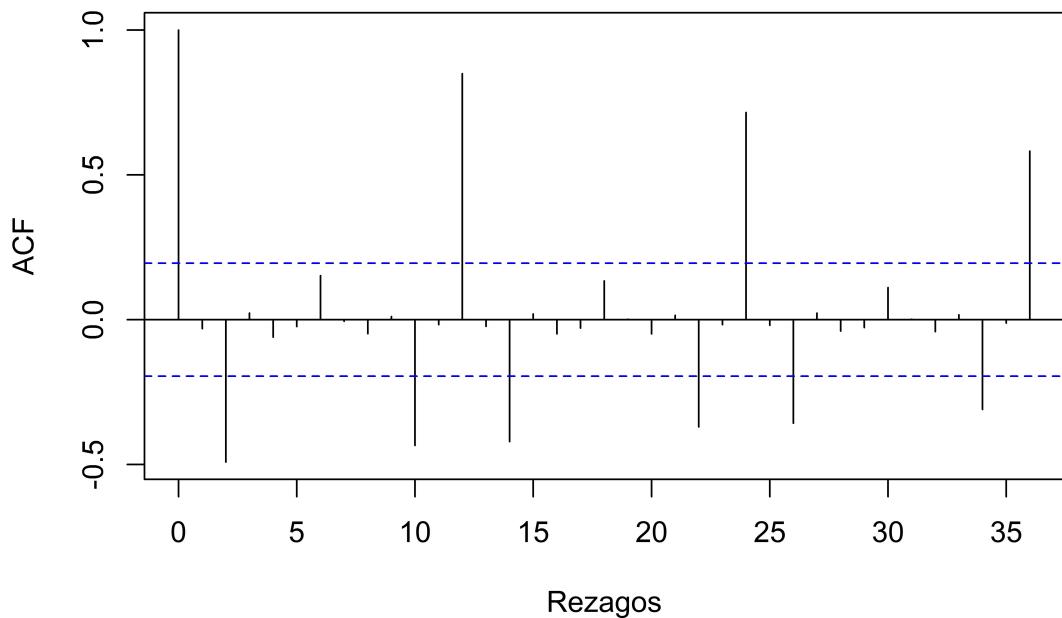
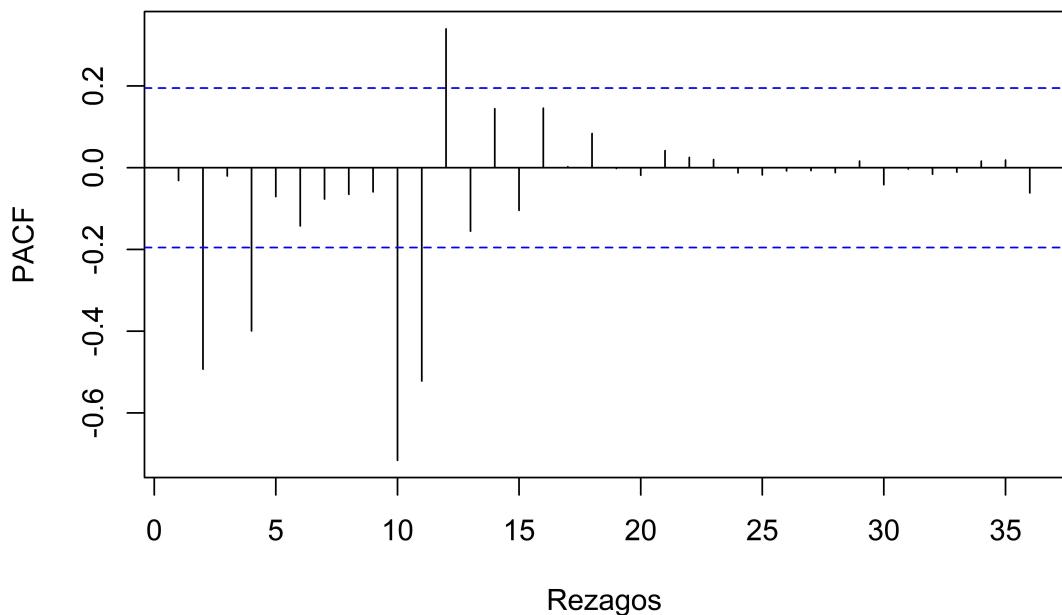


Figura 35: Autocorrelación parcial de los datos diferenciados de la serie de incentivos salariales



4.2.1.3 Mortalidad por causa externa

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 67 en la sección de anexos.

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.

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.

La descomposición de la serie se hará de forma aditiva debido a que no se observan grandes cambios en la variabilidad a lo largo del tiempo. 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.

Además, las figuras 71 y 72 muestran indicios de un proceso bajo en la parte no estacional, y si bien no hay indicios claros de patrones estacionales, es sabido que los meses de Enero, Abril y Diciembre son los que cuentan con un mayor número de defunciones por estas causas.

4.2.1.4 Intereses y comisiones del sector público

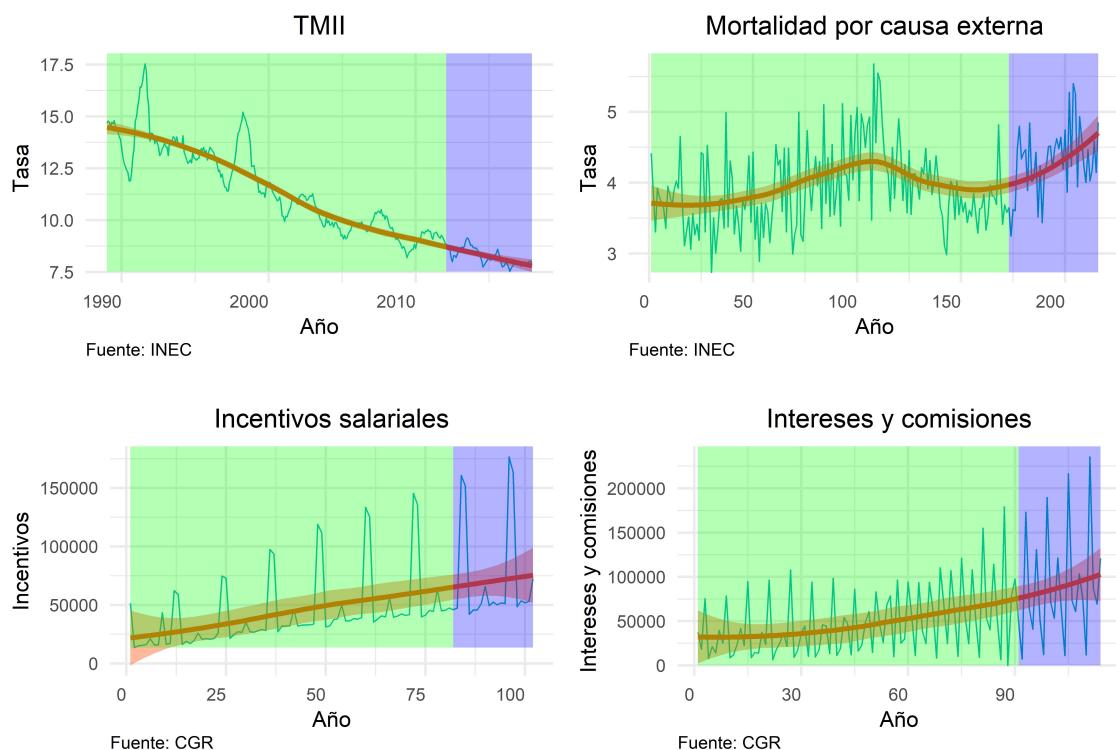
Para iniciar el análisis exploratorio de esta serie, la Figura 68 en la sección de anexos muestra que hay un ligero cambio de concavidad a partir de Julio 2010, esto sugiere que a partir de este momento los intereses y comisiones inician una tendencia al alza, la cual se sostiene hasta Junio del 2018. Por su parte, la Figura 69 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 70 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.

En la Figura 70 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. Las figuras 73 y 74 indican que existen un efecto estacional, pero su identificación no es clara.

4.2.2 Partición de los datos

Para las series cronológicas reales, la Figura 36 muestra la partición realizada.

Figura 36: Partición de los datos en los conjuntos de entrenamiento y validación para las series de tiempo reales



4.2.3 Identificación y estimación

En el caso de las series cronológicas reales, se toma en consideración el mejor modelo sugerido por la función `auto.arima()`, el ARIMA estándar y la sobreparametrización. Los coeficientes obtenidos para la serie de la Tasa de Mortalidad Infantil Interanual y para la de Incentivos salariares se muestran en el cuadro 3, mientras que para la mortalidad por causa externa y los intereses y comisiones del sector público semuestren en la sección de anexos en el cuadro 9.

Cuadro 3: Coeficientes de las ecuaciones de estimación según método de ajuste

Serie real	Coeficiente	auto.arima()			ARIMA estándar			Sobreparametrización		
		Puntual	L.I.	L.S.	Puntual	L.I.	L.S.	Puntual	L.I.	L.S.
TMII	AR1	0,19	0,08	0,31	0,75	0,56	0,94	-0,39	-0,4	-0,38
	AR2	0,32	0,21	0,44	-	-	-	0,49	0,48	0,5
	AR3	-	-	-	-	-	-	0,73	0,73	0,73
	AR4	-	-	-	-	-	-	0,17	0,17	0,17
	MA1	-	-	-	-0,57	-0,8	-0,35	1,51	1,46	1,56
	MA2	-	-	-	-	-	-	1,32	1,3	1,33
	MA3	-	-	-	-	-	-	0,61	0,58	0,64
	MA4	-	-	-	-	-	-	0,21	0,21	0,22
	SAR1	-	-	-	-0,48	-0,58	-0,37	-0,99	-0,99	-0,98
	SAR2	-	-	-	-	-	-	-0,52	-0,53	-0,52
	SAR3	-	-	-	-	-	-	-0,29	-0,29	-0,28
	SAR4	-	-	-	-	-	-	-0,21	-0,22	-0,2
	SMA1	-0,78	-0,87	-0,69	-1	-1,07	-0,93	-0,69	-0,71	-0,68
	SMA2	-	-	-	-	-	-	-0,46	-0,47	-0,45
	SMA3	-	-	-	-	-	-	0,12	0,11	0,13
	SMA4	-	-	-	-	-	-	0,14	0,13	0,15
Incentivos Salariales	AR1	-	-	-	0,06	-0,24	0,35	-0,57	-0,78	-0,37
	AR2	-	-	-	-	-	-	-0,5	-0,71	-0,29
	MA1	0,22	-0,06	0,49	-0,87	-1,03	-0,72	-	-	-
	SAR1	0,54	0,31	0,76	0,79	0,48	1,1	0,43	0,14	0,72
	SAR2	-	-	-	-	-	-	0,35	0,04	0,65
	SMA1	-	-	-	-0,33	-0,9	0,23	-	-	-

Fuente: Elaboración propia a partir de datos simulados.

4.2.4 Análisis de los errores

De manera análoga a lo trabajado con los datos simulados, al analizar los errores buscamos que estos se comporten como ruido blanco, es decir, que no tengan ningún patrón en particular. El autocorrelograma de los residuos también puede usarse para este fin, pues el 95 % de los valores debería estar entre las dos líneas azules. Además, un histograma de los residuos es útil para saber si los residuos siguen una distribución aproximadamente Normal; si no existen grandes desviaciones en este gráfico, entonces puede decirse que el cumplimiento de este supuesto es razonable.

En las Figuras 37 y 38 se visualizan los errores de los modelos estimados a partir de la serie cronológicas de la Tasa de Mortalidad Infantil Interanual y de la serie de Incentivos Salariales, respectivamente, para los tres métodos de estimación: la función `auto.arima()`, ARIMA estándar

y sobreparametrización, mientras que las figuras 75 y 76 muestran lo mismo pero para las series mortalidad por acusa externa y de intereses y comisiones del sector público. Nuevamente los modelos generados poseen a grandes rasgos el comportamiento esperado.

Figura 37: Comportamiento de los errores asociados a los modelos estimados con la serie de la TMII

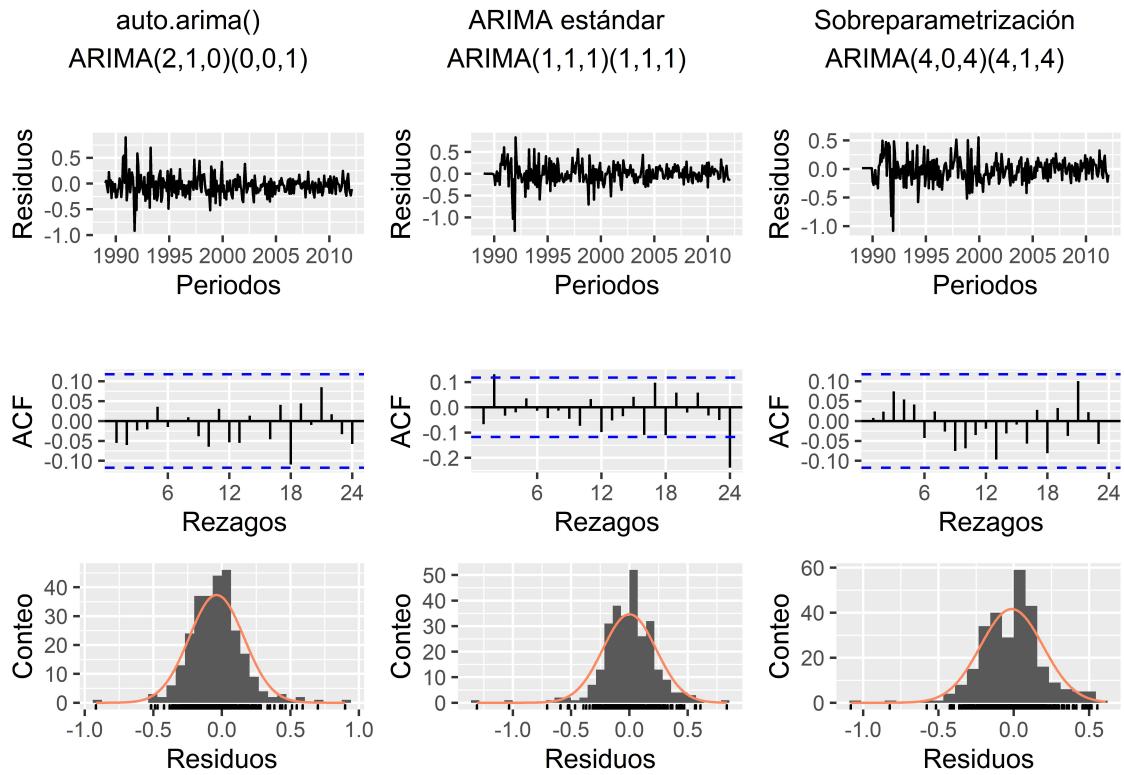
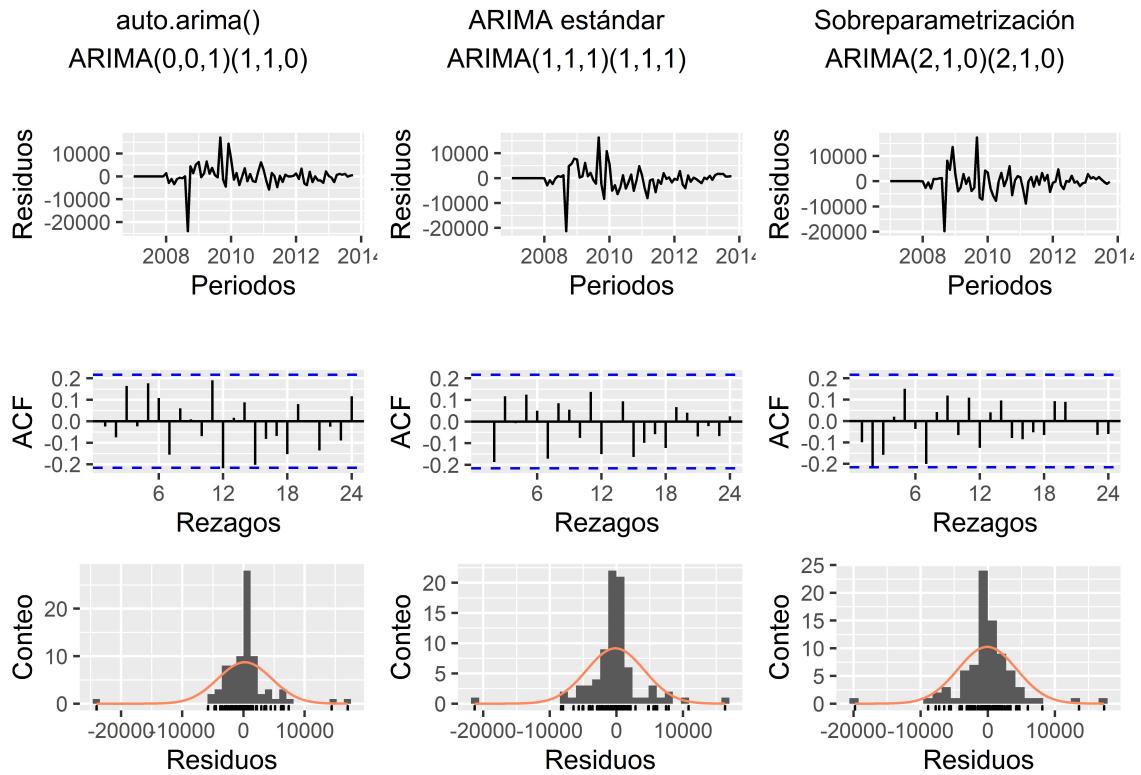


Figura 38: Comportamiento de los errores asociados a los modelos estimados con la serie de incentivos salariales



4.2.5 Medidas de bondad de ajuste y de precisi\'on

En este apartado se muestran las medidas de bondad de ajuste y de precisi\'on para la serie de la tasa de mortalidad infantil interanual y la de incentivos salariales en el cuadro 4, mientras que para las series de la mortalidad por causa externa y la ide intereses y comisiones se muestran en la secci\'on de anexos mediante el cuadro 10.

Cuadro 4: Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación a partir de las series cronológicas reales

Proceso original	Datos	Estimación	AIC	AICc	BIC	RMSE	MAE	MAPE
TMII	Entrenamiento	auto.arima()	-88,34	-88,31	-73,84	0,21	0,16	1,29
		ARIMA estándar	-21,99	-21,94	-4,09	0,24	0,17	1,38
		Sobreparametrización	-50,66	-50,52	10,26	0,22	0,16	1,33
	Validación	auto.arima()	369,38	369,52	378,37	1,16	1,03	12,69
		ARIMA estándar	78,11	78,28	89,35	0,37	0,32	3,7
		Sobreparametrización	79,87	80,54	118,09	0,34	0,27	3,17
Incentivos Salariales	Entrenamiento	auto.arima()	1615,68	1615,81	1624,67	4375,95	2349,42	6,84
		ARIMA estándar	1615,21	1615,39	1626,38	4310,66	2491,66	7,51
		Sobreparametrización	1617,05	1617,22	1628,22	4359,18	2564,76	8,07
	Validación	auto.arima()	407,12	407,72	411,11	5212,81	3701,08	4,51
		ARIMA estándar	397,69	398,48	402,67	3917,41	3011,28	3,94
		Sobreparametrización	392,95	393,73	397,93	3476,92	2846,98	4,01

Fuente: Elaboración propia a partir de datos simulados

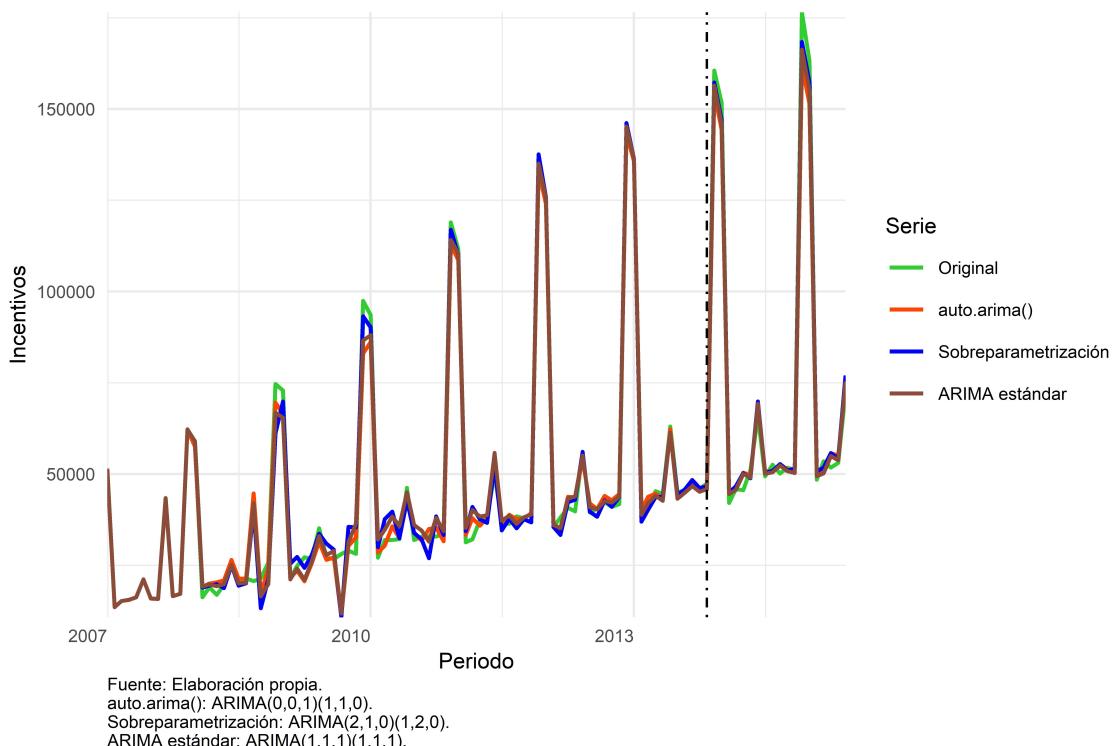
4.2.6 Estimación en el periodo de validación

De la misma manera en que se hizo con los datos simulados, las figuras 39 y 40 representan el pronóstico para los distintos modelos generados con la serie de la tasa de mortalidad infantil interanual y la de incentivos salariales, respectivamente, siendo estas primeras dos donde la tendencia se acerca más al comportamiento real; mientras que las figuras 77 y 81 en la sección de anexos muestran el resultado obtenido para las series de mortalidad por causa externa y de intereses y comisiones del sector público. La línea vertical punteada marca el inicio del pronóstico.

Figura 39: Pronóstico de la TMII según el método de estimación



Figura 40: Pronóstico de la serie de incentivos salariales del sector público según el método de estimación



Además, para las series de la tasa de mortalidad infantil interanual, la de mortalidad por causa externa, la series de incentivos y la de intereses y comisiones, se resumen visualmente los errores

estándar obtenidos con la función `auto.arima()` en las figuras 41, 78, 44 y 82, respectivamente. De manera similar, pero utilizando la sobreparametrización, los errores estándar de las series de la tasa de mortalidad infantil interanual, la de mortalidad por causa externa, la series de incentivos y la de intereses y comisiones, se resumen visualmente en las figuras 42, 79, 45 y 83, respectivamente. Por último, para las mismas series cronológicas, los errores estándar obtenidos con el modelo *ARIMA* estándar se resumen en las figuras 43, 80, 46 y 84.

Figura 41: Errores estándar de los pronósticos obtenidos para la TMII con la función `auto.arima()`

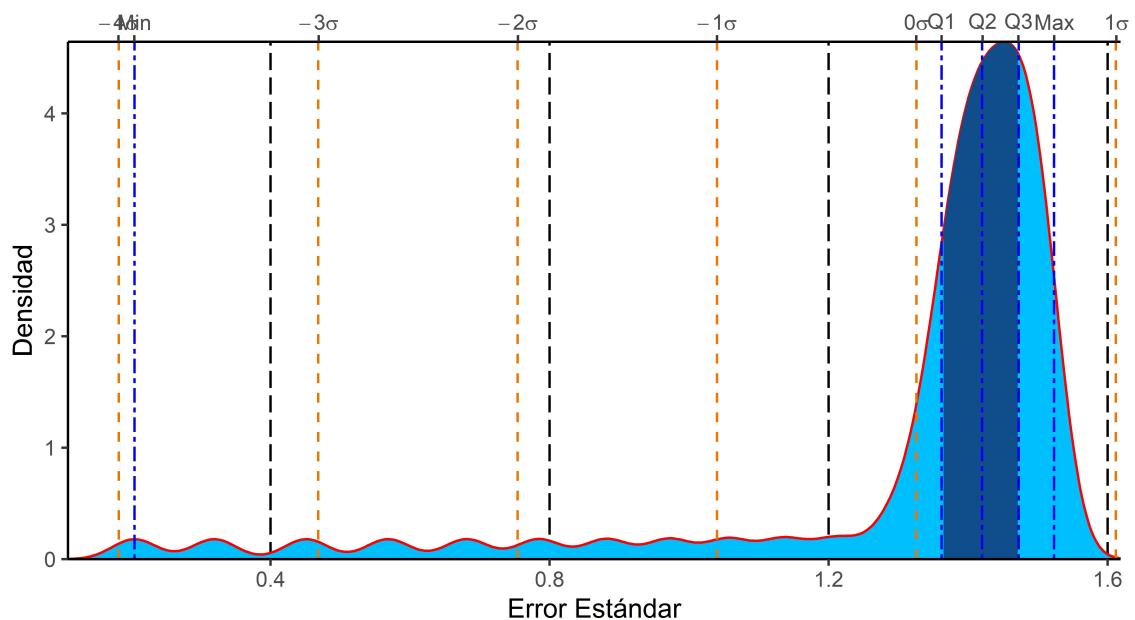


Figura 42: Errores estándar de los pronósticos obtenidos para la TMII con sobreparametrización

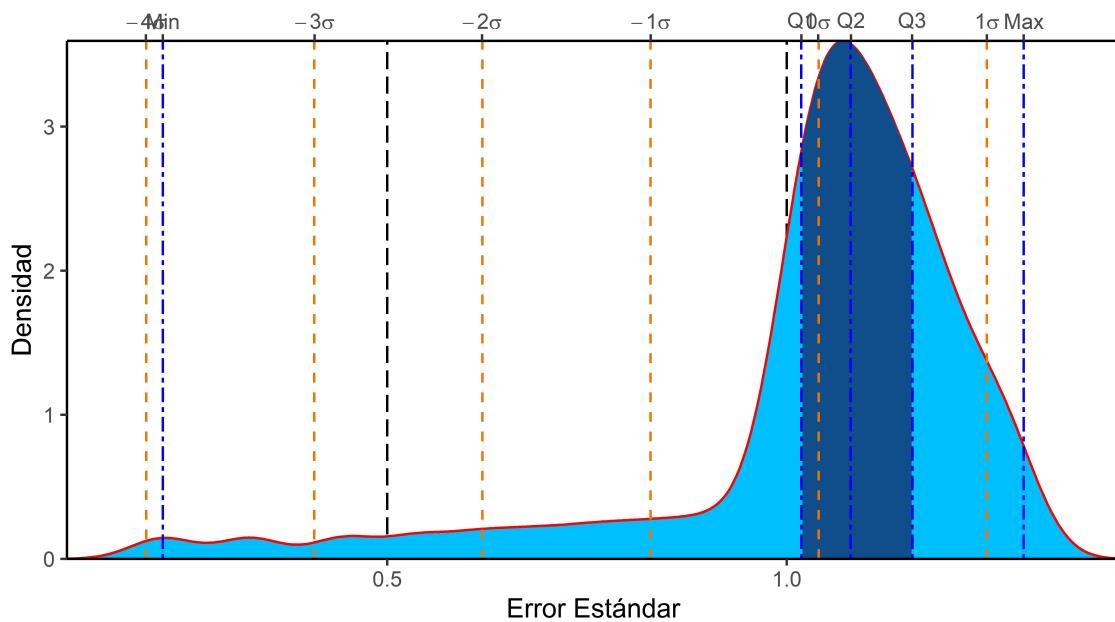


Figura 43: Errores estándares de los pronósticos obtenidos para la TMII con el modelo ARIMA estándar

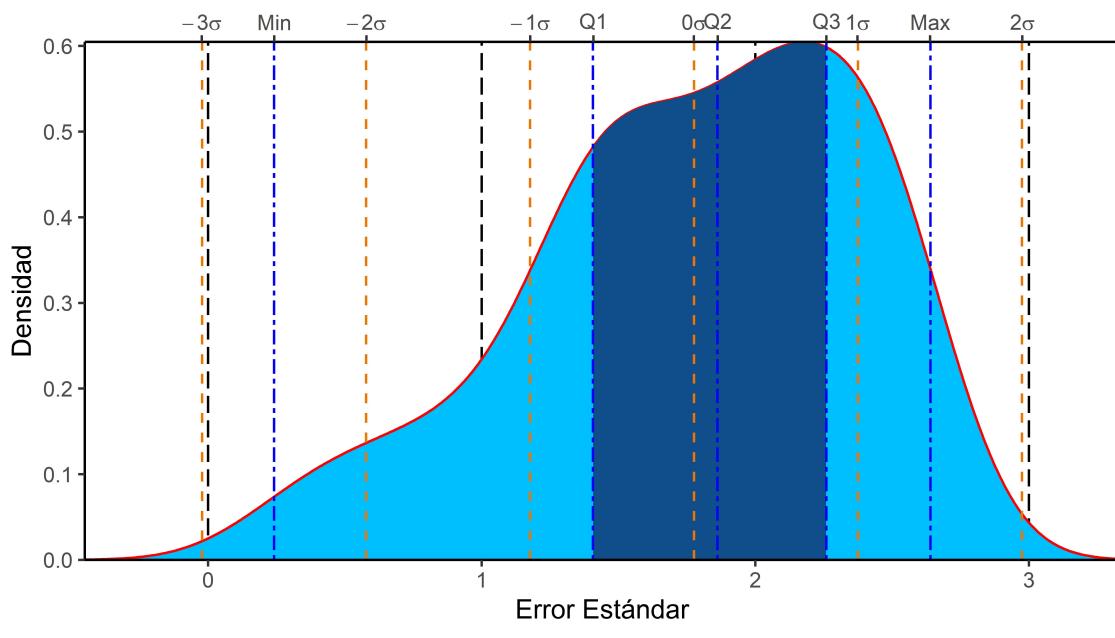


Figura 44: Errores estándar de los pronósticos obtenidos para la serie de incentivos salariales del sector público con la función auto.arima()

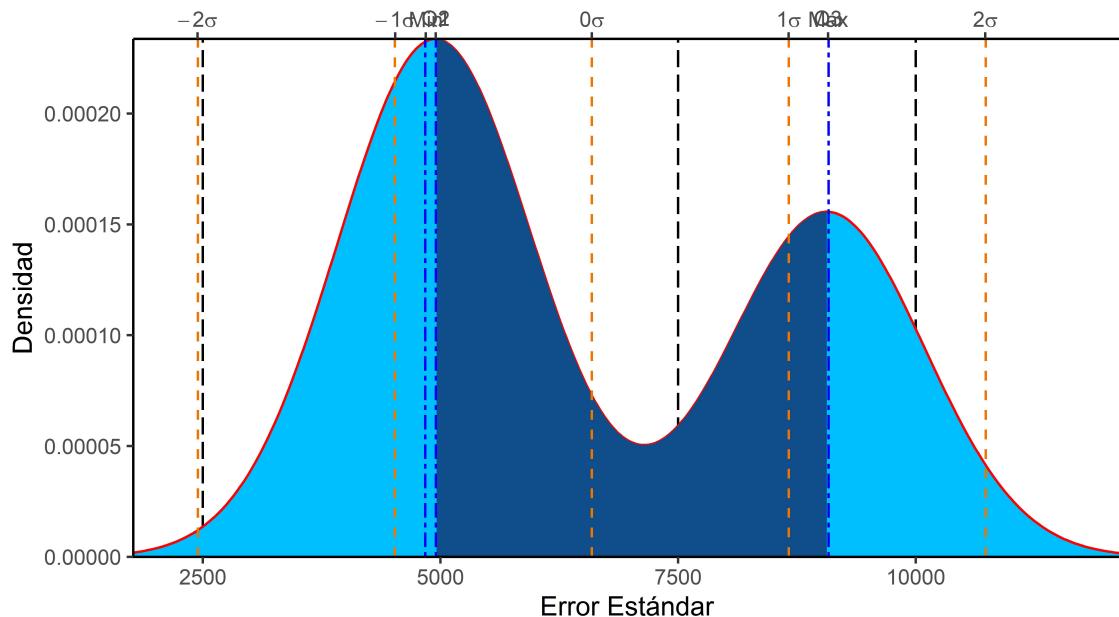


Figura 45: Errores estándar de los pronósticos obtenidos para la serie de incentivos salariales del sector público con sobreparametrización

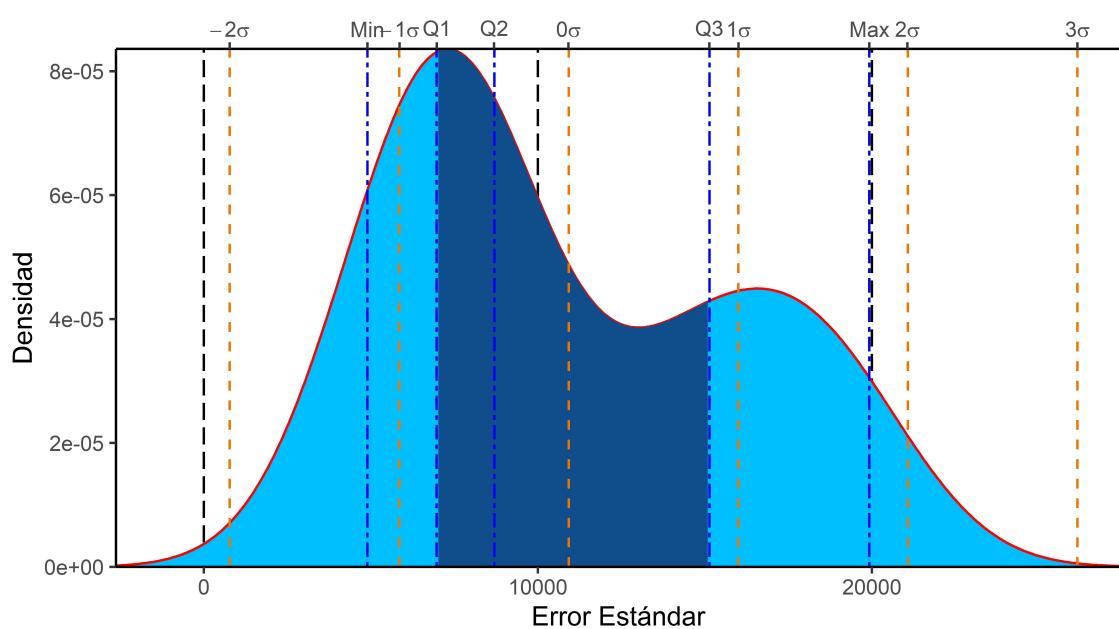
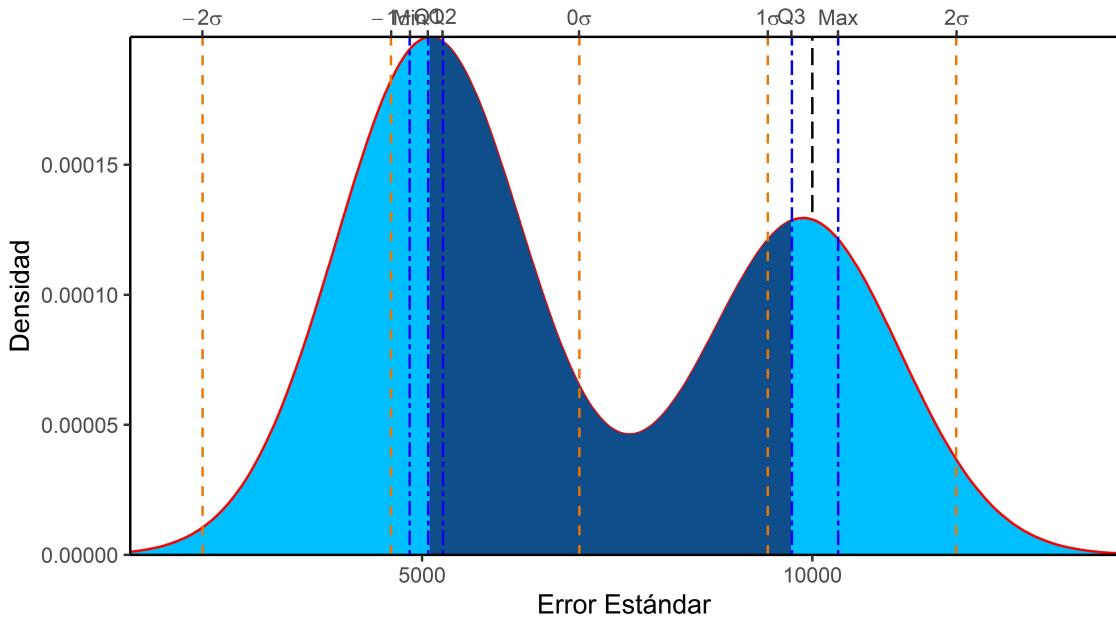


Table 1: Descriptive statistics for the standard errors in Figure 44 and 45

Min	Q1	Median	Q3	Max	SD	Kurtosis	Skewness	CV
4841.06	4953.21	4953.21	9082.39	9082.39	2071.45	1.17	0.41	0.31

Figura 46: Errores estándar de los pronósticos obtenidos para la serie de incentivos salariales del sector público con el modelo ARIMA estándar



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skewness	CV
4841.67	5078.93	5269.09	9738.4	10332.23	2413.92	1.21	0.42	0.34

4.3 Resumen de resultados

A partir de las secciones anteriores, el cuadro 5 muestra de forma porcentual la frecuencia relativa en que cada método de estimación (auto.arima, ARIMA estándar y sobreparametrización) logra alcanzar las mediciones más bajas, tanto para el conjunto de datos de entrenamiento como en el de validación, para las medidas de bondad de ajuste y de precisión.

Cuadro 5: Distribución porcentual de los métodos de estimación que alcanzaron los mejores resultados según conjunto de datos y tipo de medición

Conjunto de datos	Medidas	auto.arima()	ARIMA estándar	Sobreparametrización
Entrenamiento	Bondad de ajuste	33,34	8,34	58,34
	Precisión	45,46	9,1	45,46
Validación	Bondad de ajuste	25,01	25,01	50,01
	Precisión	33,34	0,01	66,68

Fuente: Elaboración propia a partir de los resultados obtenidos

5 CONCLUSIÓN Y DISCUSIONES

5.1 Conclusiones

Ante uno de los principales problemas en el análisis de series de tiempo mediante modelos ARIMA, como lo es la subjetividad del observador en la identificación de procesos, esta investigación ha tenido como objetivo principal siempre proponer un aporte metodológico para la selección de modelos ARIMA y por ende, una mejora en los pronósticos obtenidos a partir de estos modelos.

Las bases de este estudio se presentaron en el segundo capítulo, donde además de describir qué es una serie cronológica se explican los componentes de la misma: tendencia-ciclo, componente estacional y componente irregular. Además, se plantearon los supuestos asociados al análisis de series cronológicas, así como los distintos criterios para la identificación del modelo que gobierna a una determinada serie cronológica. Entre las distintas clases de modelos, el interés de esta investigación se centra en los modelos autorregresivos integrados de medias móviles, razón por la cual se deja claro el fundamento teórico de la ecuación de Wold y la metodología de Box-Jenkins para poder estimar los modelos autorregresivos y los modelos de medias móviles, y de esta manera justificar la unión de ambas clases en los modelos autorregresivos integrados de medias móviles. Aunado a esto, se discutió el uso de los autocorrelogramas como método de identificación de modelos para así introducir los elementos básicos del análisis combinatorio y su uso conjunto con la sobreparametrización para encontrar modelos ARIMA adecuados.

A partir de lo anterior, se describieron los materiales a usar: series cronológicas simuladas y reales (tasa de mortalidad infantil interanual, tasa de mortalidad por causa externa, incentivos salariales del sector público e intereses y comisiones del sector público). Con estos insumos, el proceso de análisis consistió en realizar un análisis exploratorio de cada serie cronológica, aplicarle a las mismas una partición de datos para dividirla en dos conjuntos de datos, entrenamiento y validación, una para ajustar los modelos y otra para evaluar la calidad de los pronósticos, respectivamente. Se realiza el ajuste de los modelos con las series simuladas y luego con las reales para posteriormente hacer un análisis de los residuales de cada modelo y verificar el cumplimiento de los supuestos discutidos previamente. Una vez estimados los modelos, se realizaron los pronósticos de todas las series cronológicas para obtener un análisis visual del comportamiento de los valores ajustados y de los pronósticos, además de evaluar la calidad de estos mediante distintas medidas de bondad de ajuste y de rendimiento.

Al aplicar la función `auto.arima()` y la sobreparametrización sobre una serie cronológica generada a partir de un $ARIMA(1, 0, 0)$, los modelos estimados son un $ARIMA(2, 1, 1)$ y un $ARIMA(1, 0, 0)$, respectivamente. Los errores estándar son menores al utilizar la sobreparametrización y también tiene una menor variabilidad. Las medidas de bondad de ajuste en el conjunto

de datos de entrenamiento, como se muestra en el cuadro 8, son menores en el modelo sugerido por la sobreparametrización, mientras que las mejores medidas de rendimiento las muestra el `auto.arima()`, aunque las diferencias son muy grandes.

Al generar datos a partir de un $ARIMA(1, 0, 1)$ los mejores pronósticos se obtienen tanto con el `auto.arima()` como con la sobreparametrización tal y como se muestra en el cuadro 8, pues ambos son superiores a los obtenidos mediante un $ARIMA(1, 1, 1)$, sin embargo, el comportamiento en la predicción es casi constante. En este escenario, los errores estándar tienen prácticamente el mismo comportamiento, tanto para la sobreparametrización como para el `auto.arima()` y el ARIMA estándar.

Al ir incorporando términos en las series no estacionales, como es el caso de los datos simulados a partir de un $ARIMA(2, 0, 3)$, los mejores pronósticos se obtienen mediante el uso de la sobreparametrización (ver cuadro 8), pues la magnitud de sus errores son siempre menores, sin embargo, el comportamiento de los pronósticos también se vuelve constante, tal y como se aprecia en la Figura 59. Con los tres métodos utilizados, a pesar de que los pronósticos en series de orden bajo no son los mejores, entre las tres opciones la sobreparametrización brinda mejores resultados en términos de rendimiento y bondad de ajuste.

Los últimos modelos generados a partir de los datos simulados se obtuvieron de la serie cronológica proveniente de un proceso $ARIMA(4, 0, 2)$. Como muestra la Figura 63, los pronósticos son bastante competentes en los primeros períodos, pero posteriormente se van acotando; mientras que los errores estándar son prácticamente los mismos con los modelos. Las medidas de rendimiento para estas estimaciones están bastante cercanas entre sí, siendo las del `auto.arima()` ligeramente mejores.

Cuando se consideran series cronológicas estacionales, se presenta un comportamiento similar a lo previamente descrito. Al tener una baja cantidad de parámetros en los datos simulados de un proceso estacional $ARIMA(0, 0, 1)(0, 1, 1)_{12}$, la sobreparametrización iguala los resultados obtenidos mediante la función `auto.arima()` tal y como puede apreciarse en la Figura 18 y el cuadro 2; por lo que los errores estándar también poseen el mismo comportamiento, igualando así los intervalos de confianza.

Al incorporar más parámetros al modelo generador de los datos como en el caso del $ARIMA(2, 1, 4)(3, 0, 3)_{12}$, el cuadro 2 y en la Figura 22 muestran como la sobreparametrización logra captar de mejor manera el comportamiento de los datos, obteniendo así menores mediciones del error, es decir, los pronósticos se acercan más a los datos reales. Los errores estándar son de mayor magnitud en la sobreparametrización, lo cual genera intervalos de confianza más amplios.

En el caso de las series de tiempo generadas a partir de datos reales, y en el caso particular de

la Tasa de mortalidad infantil interanual, la función `auto.arima()` sugiere como mejor modelo un $ARIMA(2, 1, 0)(0, 0, 1)$, mientras que utilizando la sobreparametrización se tiene como mejor modelo un $ARIMA(4, 1, 0)(4, 1, 0)$. El cuadro 4 y la Figura 39 muestran como el uso de la sobreparametrización ajusta los valores pronosticados de una mejor manera a los datos reales con respecto a utilizar `auto.arima()` o un modelo $ARIMA$ estándar. El valor mediano de los errores estándar es inferior al utilizar la sobreparametrización, lo cuál generará intervalos de confianza más acotados. Esto es de particular interés porque al tener intervalos de un mismo nivel de confianza pero más acotados, puede tenerse más certeza del posible rango que alcanzaría esta serie de tiempo e incluso otras que presenten comportamientos similares en el sentido de su correlación natural con sus períodos previos.

De manera similar, tras ajustar un modelo utilizando la función `auto.arima()` a la tasa de mortalidad por causa externa el mejor modelo sugerido es un $ARIMA(1, 1, 1)$, mientras que la sobreparametrización propone como mejor modelo un $ARIMA(2, 0, 1)(0, 1, 1)_{12}$. Los resultados de la Figura 77 y el cuadro 10 muestran que la sobreparametrización supera a los resultados del $ARIMA$ estándar, y este a su vez supera al `auto.arima()`. Aunque no se replican a la perfección los saltos de la serie, el comportamiento general es similar, y las medidas de rendimiento y bondad de ajuste en la sobreparametrización indican que este modelo es más adecuado en comparación a los demás.

Esta información junta puede servir de insumo para construir los diferentes índices, tasas y otros indicadores que revelan la situación demográfica del país, información de gran relevancia para la planificación nacional, regional y local en diversos campos. Uno de estos principales campos de acción es la salud pública, para la cual la tasa de mortalidad infantil se considera uno de los indicadores prioritarios dado que refleja no solo las condiciones de salud de la población infante, sino también los niveles de desarrollo del país, pues depende de la calidad de la atención de la salud, principalmente de la prenatal y perinatal, así como de las condiciones de saneamiento. Por tanto, su continuo monitoreo es fundamental para diseñar, implementar y evaluar políticas de salud pública orientadas a disminuir y erradicar aquellas que son prevenibles.

Para la serie de incentivos salariales del sector público, la función `auto.arima()` indica que el mejor modelo es un $ARIMA(0, 0, 1)(1, 1, 0)_{12}$, mientras que la sobreparametrización sugiere un $ARIMA(2, 1, 0)(1, 2, 0)_{12}$ como el más adecuado. Los resultados del cuadro 4 y la Figura 40 muestran que nuevamente los pronósticos obtenidos son superiores utilizando la sobreparametrización al analizar los pronósticos, y aunque no presenta las mejores medidas de ajuste, la diferencia con los otros modelos no es tan grande.

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

medidas de rendimiento y en las medidas de bondad de ajuste, tal y como muestra el cuadro 10, mientras que la Figura 81 muestra de manera gráfica los pronósticos obtenidos, que son bastante similares pues ambos modelos difieren solamente en un parámetro para la parte no estacional.

5.2 Discusiones

El uso de la sobreparametrización propuesto mediante un algoritmo de selección de modelos implementado en esta investigación permite evaluar una gama más amplia de modelos ARIMA al definir un máximo en la cantidad de parámetros para las partes estacionales y no estacionales de la series cronológicas, pues al definir este máximo se definen todos los posibles escenarios que posteriormente evalúan el aporte de cada nuevo término a los pronósticos. La incorporación de estos nuevos parámetros en los modelos ARIMA son validados mediante pruebas de significancia estadística, particiones de la serie cronológica, medidas de bondad de ajuste de los modelos y sus correspondientes medidas de rendimiento.

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

Al tener datos que vienen de un proceso con bajo número de parámetros los pronósticos obtenidos se terminan volviendo constantes en el caso de series no estacionales, sin embargo el uso de la sobreparametrización logra captar de buena manera el comportamiento de la serie en comparación a las otras alternativas más utilizadas y, además, cuando el proceso que gobierna la serie es de un mayor grado, la metodología propuesta, al considerar un mayor espectro paramétrico, es capaz de capturar de mejor forma el comportamiento de la serie y conseguir pronósticos con una precisión mayor al de los métodos más tradicionales. Lo anterior representa una mejora en cuanto a la utilización de modelos ARIMA para el pronóstico de series cronológicas, lo cual a su vez aporta herramientas para la toma de decisiones relacionadas a este tipo de análisis.

Además, en términos computacionales, ejecutar la sobreparametrización sobre equipos con 8Gb de memoria RAM o superiores no es algo que acortará el tiempo de estimación, pues simplemente es algo relacionado al almacenamiento de los objetos de R que se van creando. El uso de servicios gratuitos como Google Colab son una excelente alternativa para este tipo de procesos, pues tienen habilitado una capacidad más que aceptable de memoria y son capaces de ejecutar los procesos no solo mediante CPU's, sino también mediante GPU's, las cuales sí podrían reducir en cierta medida el tiempo de procesamiento.

Una potencial mejora al uso de la sobreparametrización es la inclusión semi-automática de regre-

sores para controlar cambios estructurales de la serie cronológica en estudio, pues estos coeficientes adicionales podrían controlar cambios particulares en la serie y que podrían mejorar la precisión de los pronósticos.

Lo anterior también abre una futura ventana de investigación, y es la posibilidad de validar el uso de la sobreparametrización sobre las demás alternativas disponibles haciendo un mayor número de realizaciones para cada proceso simulado, esto con la finalidad de poder generalizar el comportamiento del método propuesto ante distintos escenarios que podrían presentarse para un mismo proceso $ARIMA(p, d, q)(P, D, Q)_s$.

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

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

¹⁰<https://cran.r-project.org/web/packages/popstudy/index.html>.

6 ANEXOS

Cuadro 6: Tiempos de estimación en minutos para cada modelo según su tipo de estimación.

Serie	autoarima	Sobreparametrización	ARIMA estándar
ARIMA(1,0,0)	0,1056	8,236	0,0064
ARIMA(1,0,1)	0,0425	8,1266	0,0045
ARIMA(2,0,3)	0,0783	5,2904	0,0044
ARIMA(4,0,2)	0,1097	6,7233	0,0047
ARIMA(0,0,1)(0,1,1)[12]	0,1625	39,5444	0,2976
ARIMA(1,1,0)(1,1,0)[12]	0,0794	23,4079	0,3153
ARIMA(2,1,4)(4,1,4)[12]	0,1296	16,04	0,1951
ARIMA(2,1,4)(3,0,3)[12]	5,088	26,8193	0,1979
TMII	3,919	53,1779	0,3798
EXTERNA	0,1412	46,0911	0,2142
INCENTIVOS	2,8172	21,4405	0,1049
INTERESES	0,6145	32,542	0,108

Fuente: Elaboración propia

Cuadro 7: Coeficientes del proceso original y de los métodos de estimación de series simuladas no estacionales

Proceso original	Coeficiente	Valor real	auto.arima()			ARIMA estándar			Sobreparametrización		
			Puntual	L.I.	L.S.	Puntual	L.I.	L.S.	Puntual	L.I.	L.S.
ARIMA(1,0,0)	AR1	0,49	0,39	0,22	0,57	0,39	0,21	0,57	0,44	0,3	0,58
	AR2	-	-0,11	-0,28	0,06	-	-	-	-	-	-
	MA1	-	-0,92	-1,02	-0,82	-0,95	-1,03	-0,86	-	-	-
ARIMA(1,0,1)	AR1	0,94	-	-	-	-0,05	-0,24	0,15	-	-	-
	MA1	0,55	0,73	0,63	0,82	0,74	0,63	0,86	0,73	0,63	0,82
ARIMA(2,0,3)	AR1	0,34	0,46	0,27	0,65	-0,31	-0,53	-0,09	-	-	-
	AR2	0,19	-	-	-	-	-	-	-	-	-
	MA1	-0,41	-0,3	-0,5	-0,11	-0,39	-0,59	-0,19	0,15	0,01	0,3
	MA2	0,42	0,15	-0,03	0,32	-	-	-	0,35	0,21	0,49
	MA3	0,97	0,66	0,46	0,85	-	-	-	0,74	0,62	0,85
	MA4	-	-	-	-	-	-	-	0,26	0,09	0,42
	MA5	-	-	-	-	-	-	-	0,36	0,2	0,51
	AR1	-0,02	-0,81	-0,94	-0,69	-0,91	-0,98	-0,84	-0,91	-0,97	-0,85
	AR2	0,29	-	-	-	-	-	-	-	-	-
	AR3	0,04	-	-	-	-	-	-	-	-	-
ARIMA(4,0,2)	AR4	0,57	-	-	-	-	-	-	-	-	-
	MA1	0,04	-0,09	-0,3	0,11	0	-0,15	0,14	-	-	-
	MA2	0,8	0,29	0,12	0,47	-	-	-	-	-	-
	MA3	-	-0,21	-0,41	-0,01	-	-	-	-	-	-

Fuente: Elaboración propia a apartir de datos simulados.

Figura 47: Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(1,0,0)

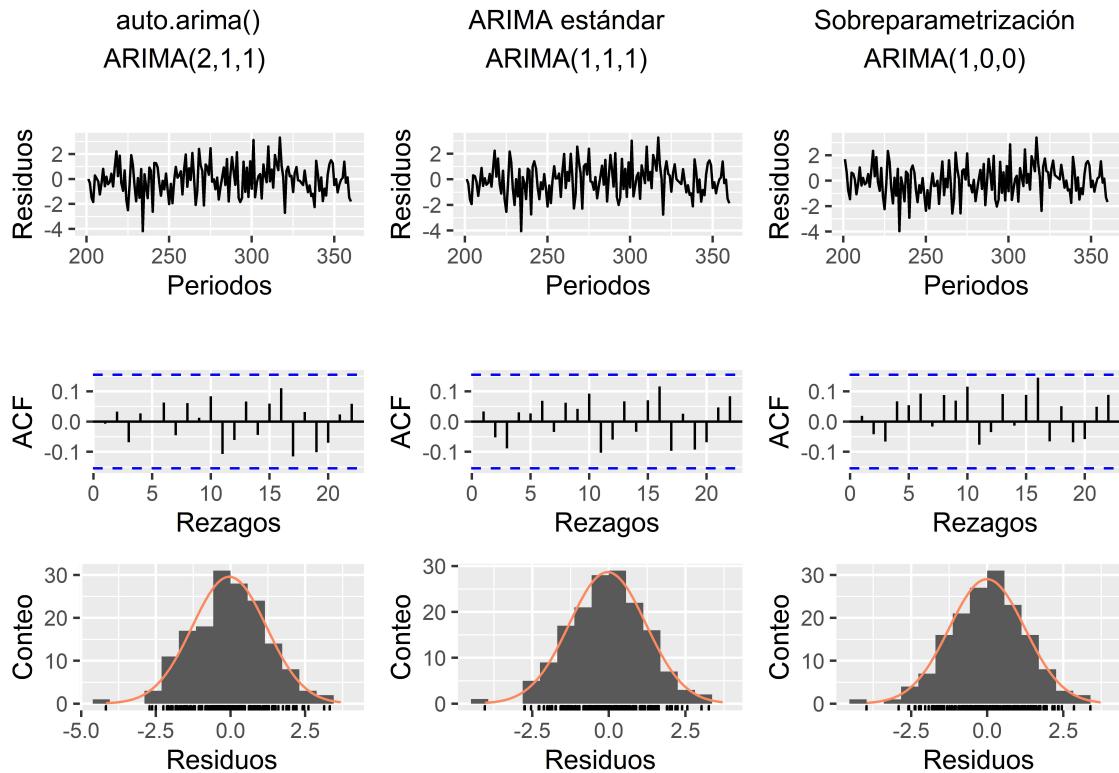


Figura 48: Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(1,0,1)

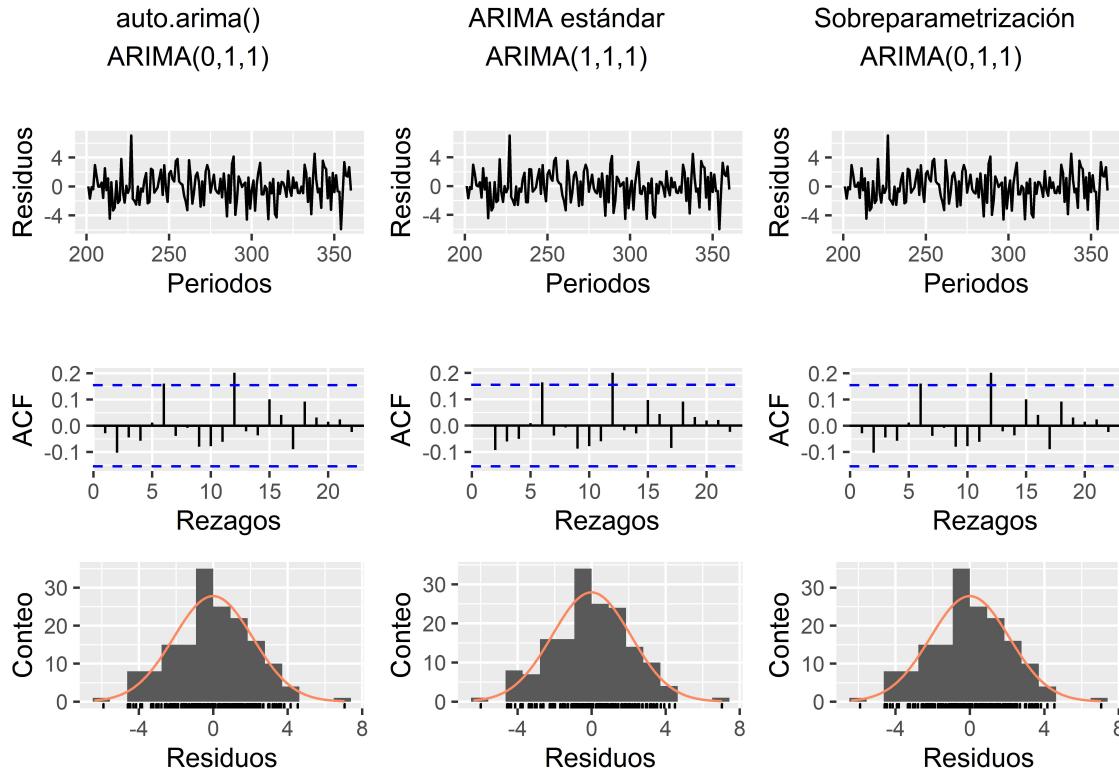


Figura 49: Comportamiento de los errores asociados a los modelos estimados con datos generados a partir de un proceso ARIMA(2,0,3)

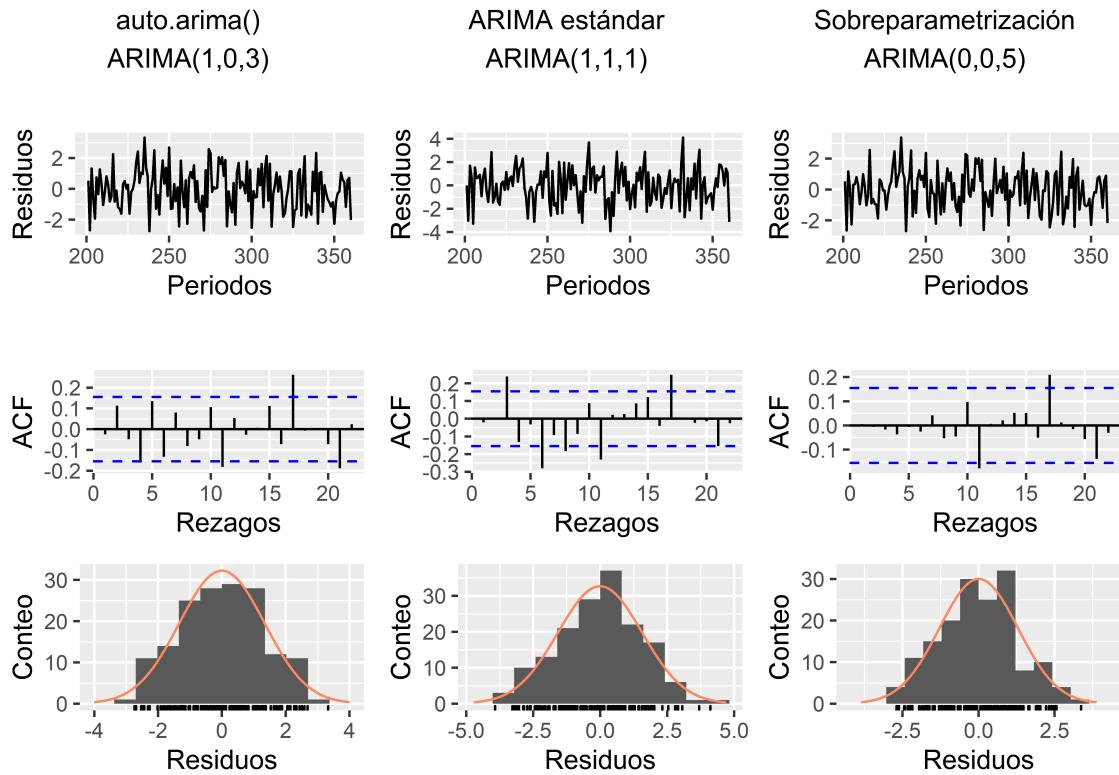
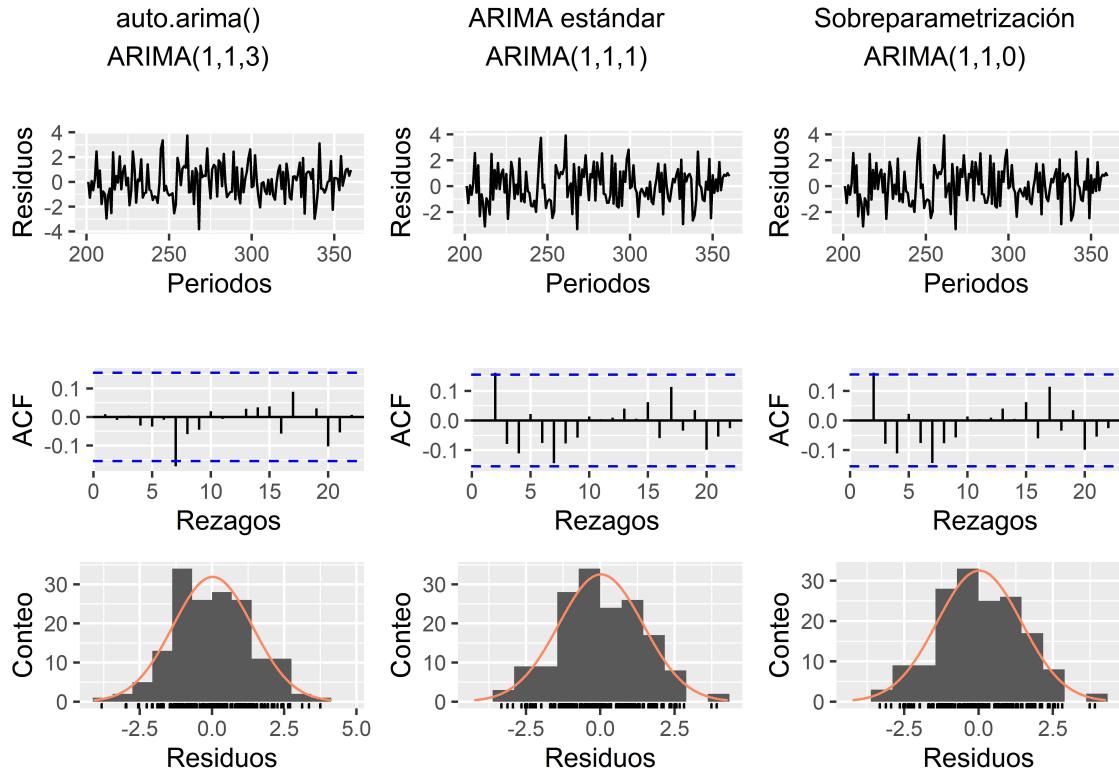


Figura 50: Comportamiento de los errores de los modelos para los datos generados con un ARIMA(4,0,2)



Cuadro 8: Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación simulados a partir de datos estacionales simulados

Proceso original	Datos	Estimación	AIC	AICc	BIC	RMSE	MAE	MAPE
ARIMA(1,0,0)	Entrenamiento	auto.arima()	531,36	531,42	543,64	1,25	0,99	10,33
		ARIMA estándar	530,81	530,86	540,02	1,26	0,99	10,37
		Sobreparametrización	529,31	529,36	538,54	1,25	1	10,39
	Validación	auto.arima()	140,17	140,42	146,92	1,27	1	10,37
		ARIMA estándar	139,54	139,73	144,6	1,29	1,02	10,64
		Sobreparametrización	139,8	139,99	144,86	1,3	1,03	10,63
ARIMA(1,0,1)	Entrenamiento	auto.arima()	698,53	698,56	704,67	2,13	1,68	41
		ARIMA estándar	700,32	700,37	709,53	2,13	1,68	40,71
		Sobreparametrización	698,53	698,56	704,67	2,13	1,68	41
	Validación	auto.arima()	254,9	255,03	258,27	5,24	4,59	74,29
		ARIMA estándar	256,94	257,14	262,01	5,25	4,6	74,33
		Sobreparametrización	254,9	255,03	258,27	5,24	4,59	74,29
ARIMA(2,0,3)	Entrenamiento	auto.arima()	557,49	557,57	575,94	1,34	1,1	11,39
		ARIMA estándar	603,38	603,43	612,59	1,57	1,28	13,32
		Sobreparametrización	548,48	548,58	570,01	1,3	1,05	10,86
	Validación	auto.arima()	181,35	181,74	191,48	2,02	1,7	18,83
		ARIMA estándar	180,65	180,84	185,72	2,16	1,81	19,15
		Sobreparametrización	182,05	182,52	193,87	1,99	1,67	18,41
ARIMA(4,0,2)	Entrenamiento	auto.arima()	565,19	565,27	580,54	1,38	1,12	13,53
		ARIMA estándar	571,64	571,69	580,85	1,43	1,17	14,05
		Sobreparametrización	569,64	569,67	575,78	1,43	1,17	14,05
	Validación	auto.arima()	182,01	182,34	190,46	2,03	1,44	14,74
		ARIMA estándar	179,15	179,35	184,22	2,06	1,47	15,05
		Sobreparametrización	177,17	177,31	180,55	2,06	1,47	15,06

Fuente: Elaboración propia a partir de datos simulados

Figura 51: Pronóstico de los datos generados mediante un ARIMA(1,0,0) según el método de estimación

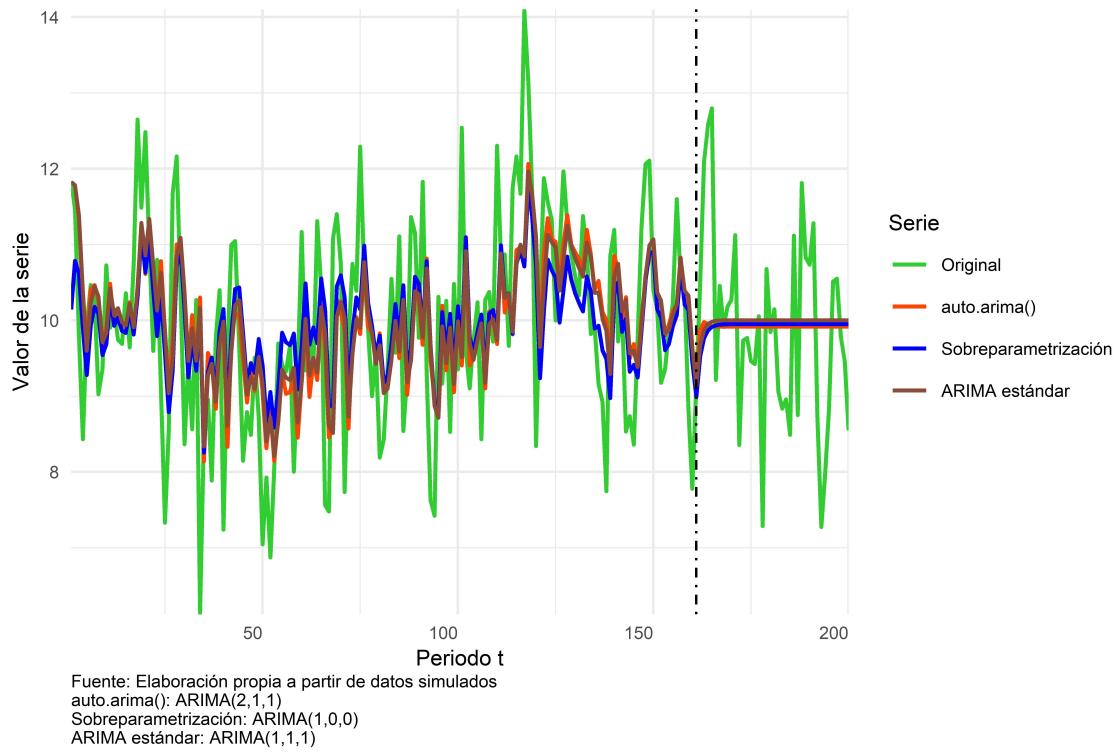
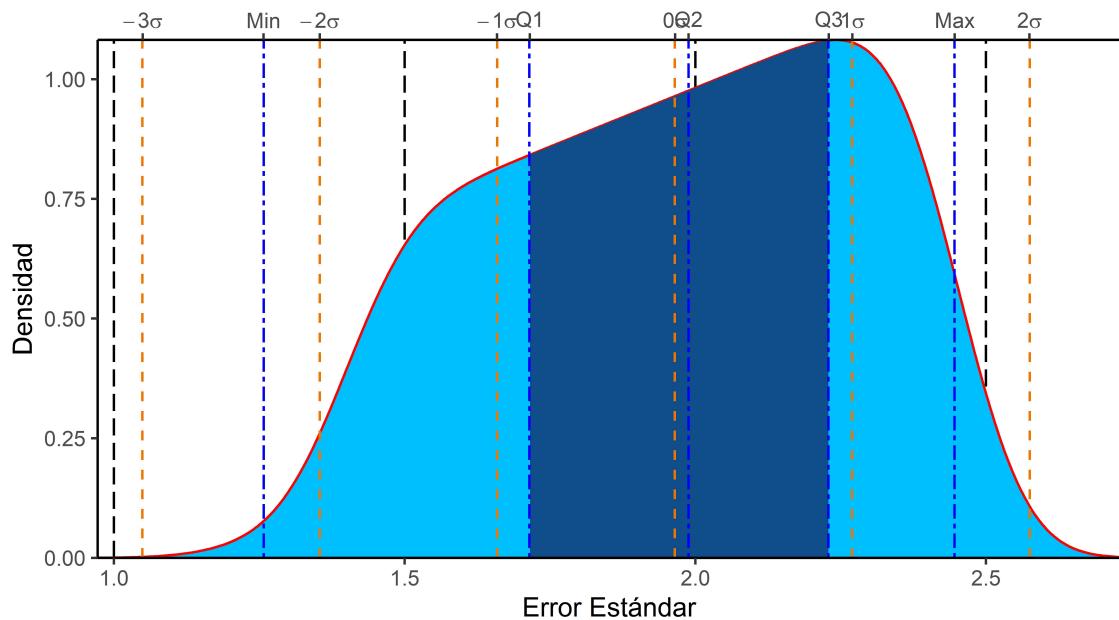


Figura 52: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,0) con la función auto.arima()



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
1.26	1.71	1.99	2.23	2.45	0.31	1.91	-0.21	0.16

Figura 53: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,0) con sobreparametrización

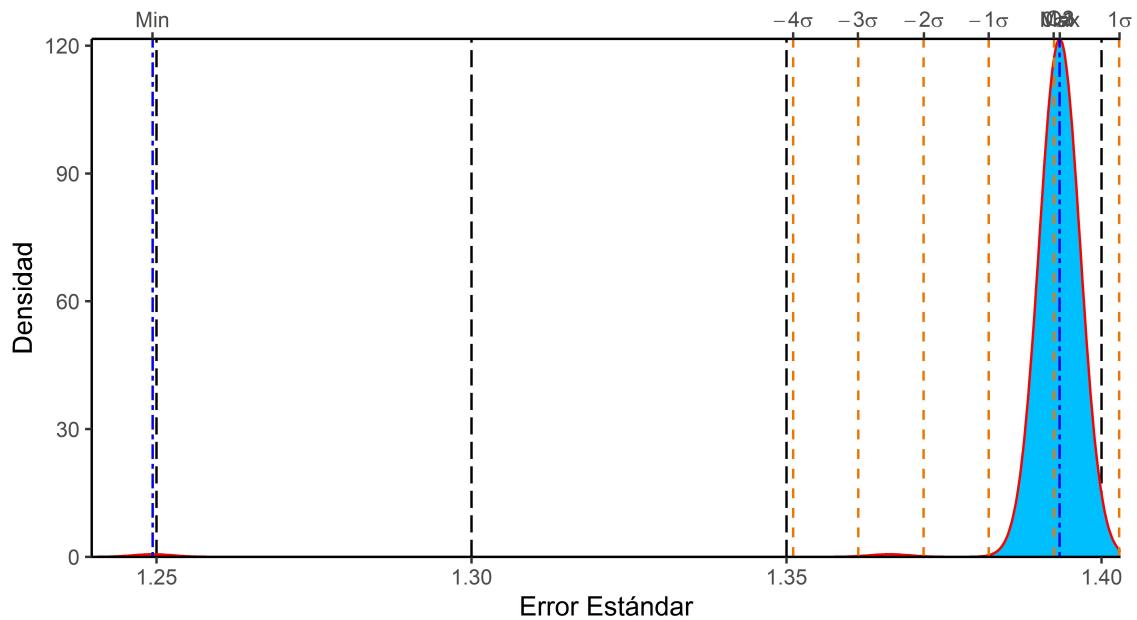


Figura 54: Errores estándares de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,0) con el modelo ARIMA estándar

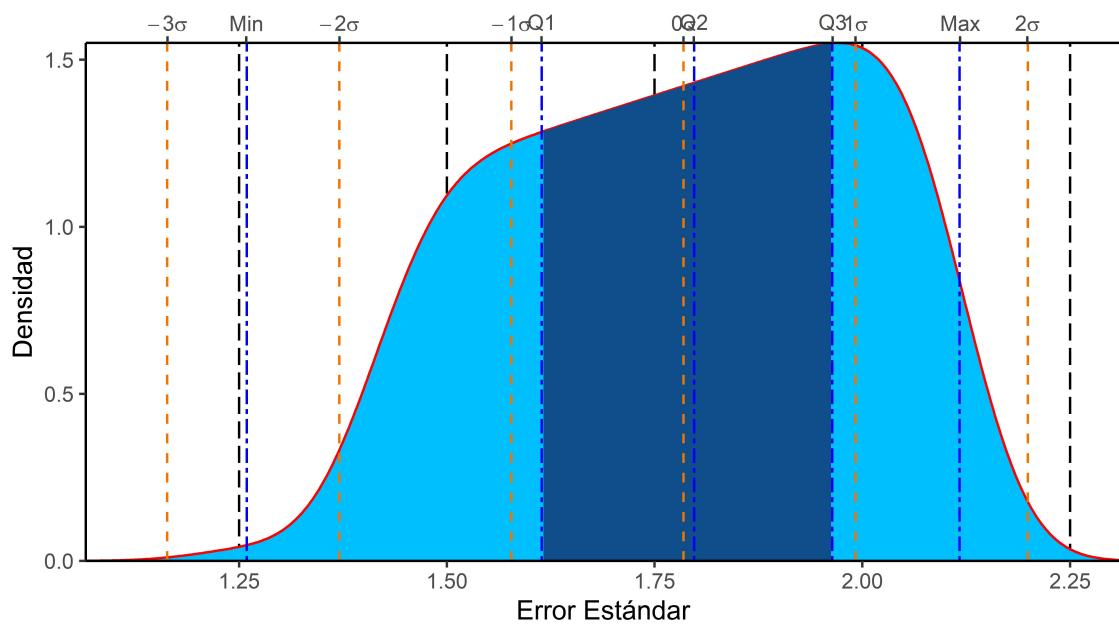


Figura 55: Pronóstico de los datos generados mediante un ARIMA(1,0,1) según el método de estimación

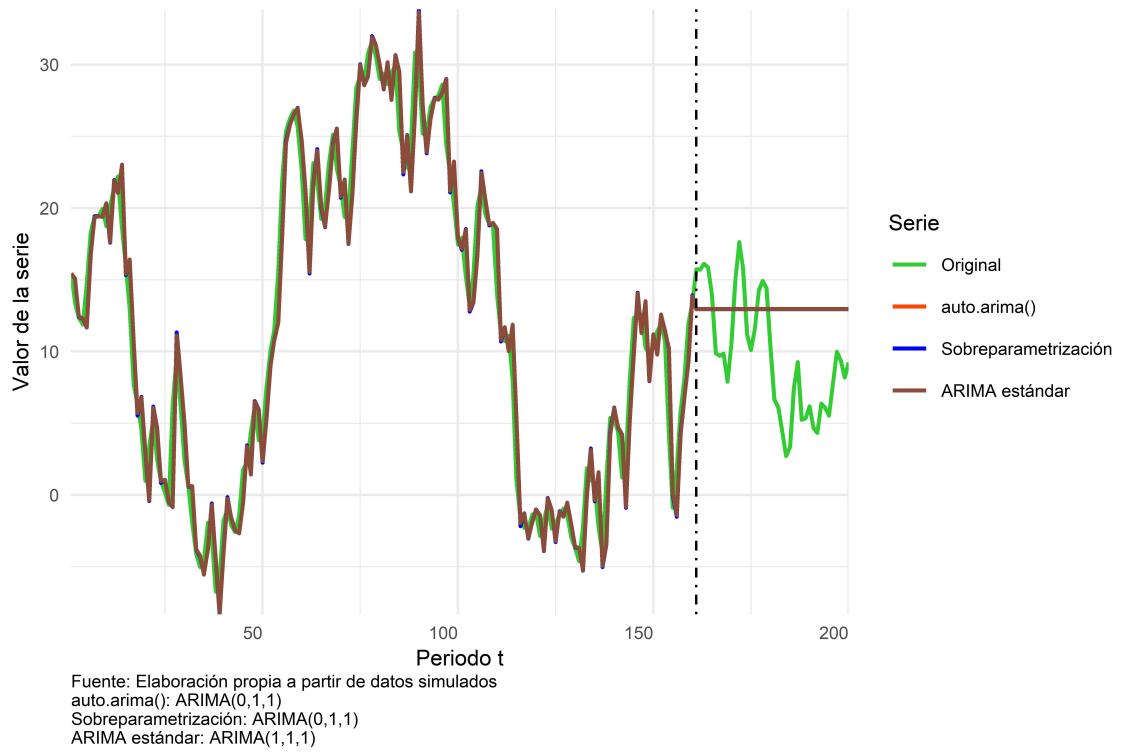
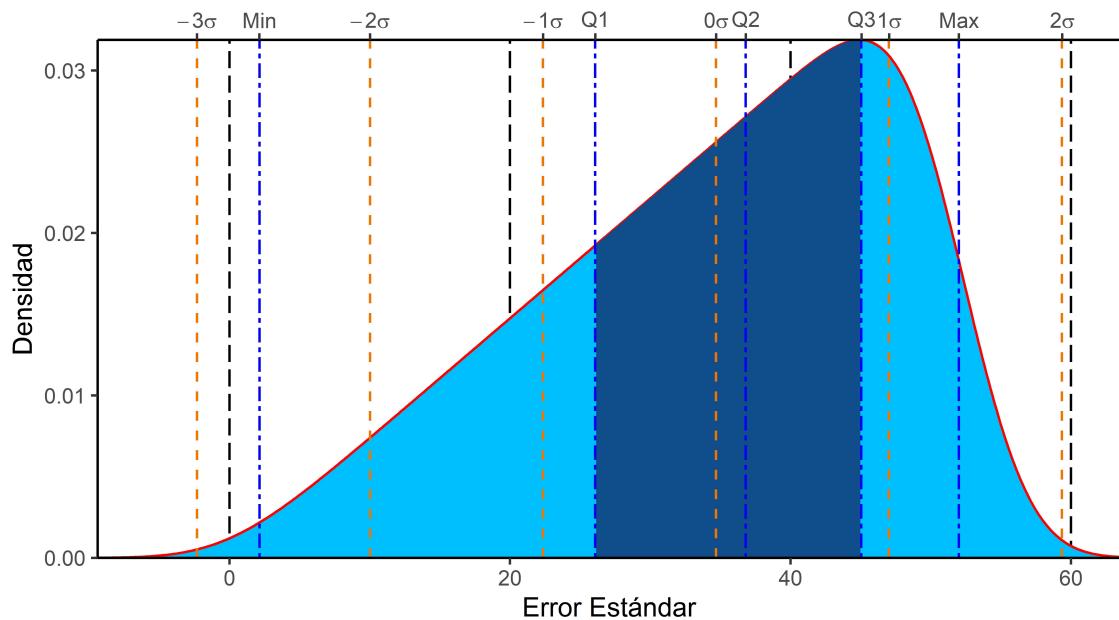
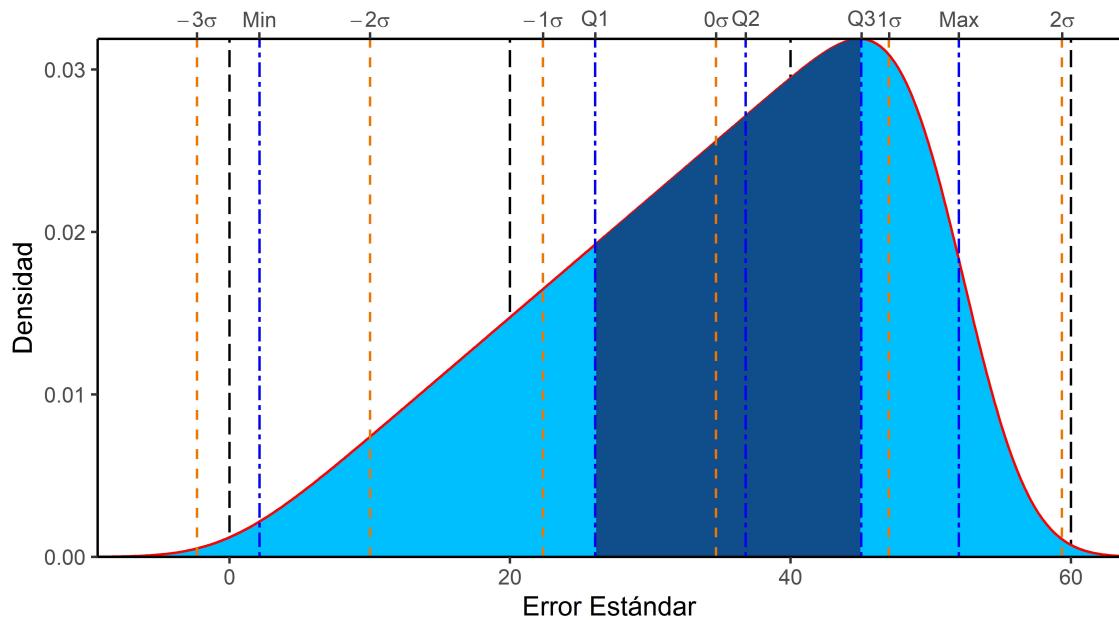


Figura 56: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,1) con la función auto.arima()



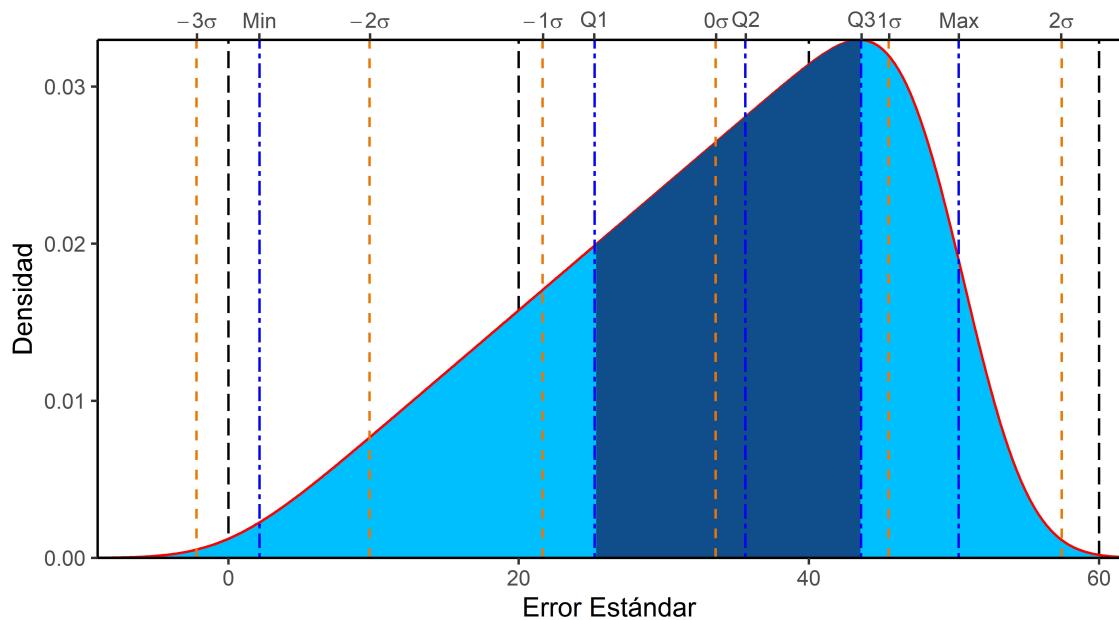
Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
2.13	26.07	36.8	45.05	52	12.33	2.41	-0.57	0.36

Figura 57: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,1) con sobreparametrización



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
2.13	26.07	36.8	45.05	52	12.33	2.41	-0.57	0.36

Figura 58: Errores estándares de los pronósticos obtenidos de los datos generados mediante un ARIMA(1,0,1) con el modelo ARIMA estándar



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
2.14	25.24	35.62	43.59	50.32	11.92	2.41	-0.57	0.36

Figura 59: Pronóstico de los datos generados mediante un ARIMA(2,0,3) según el método de estimación

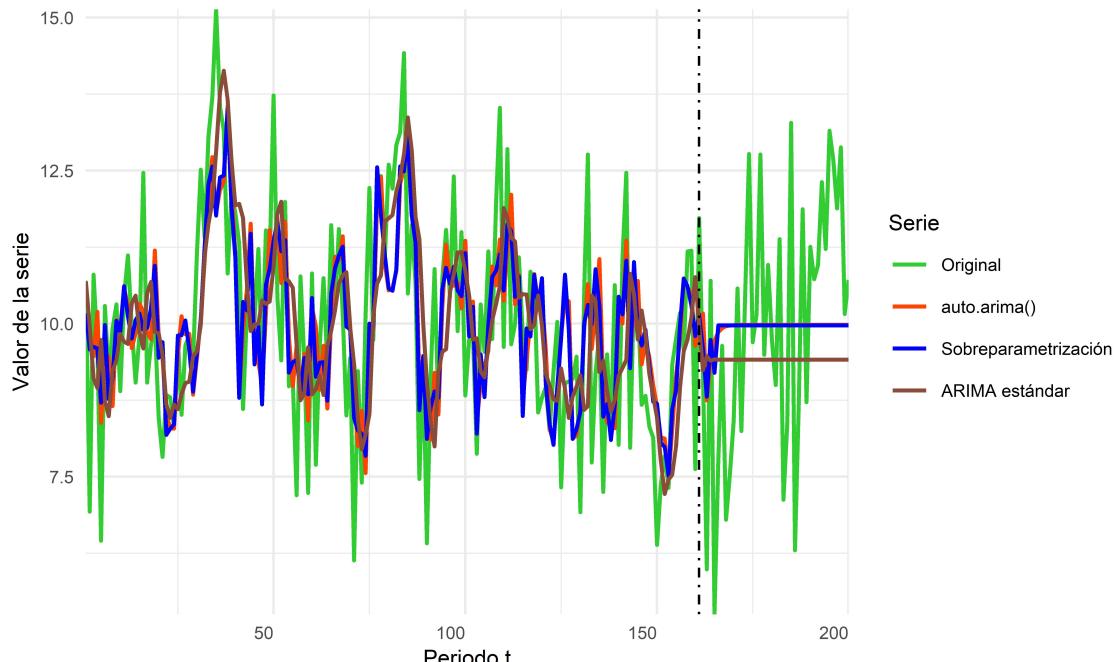
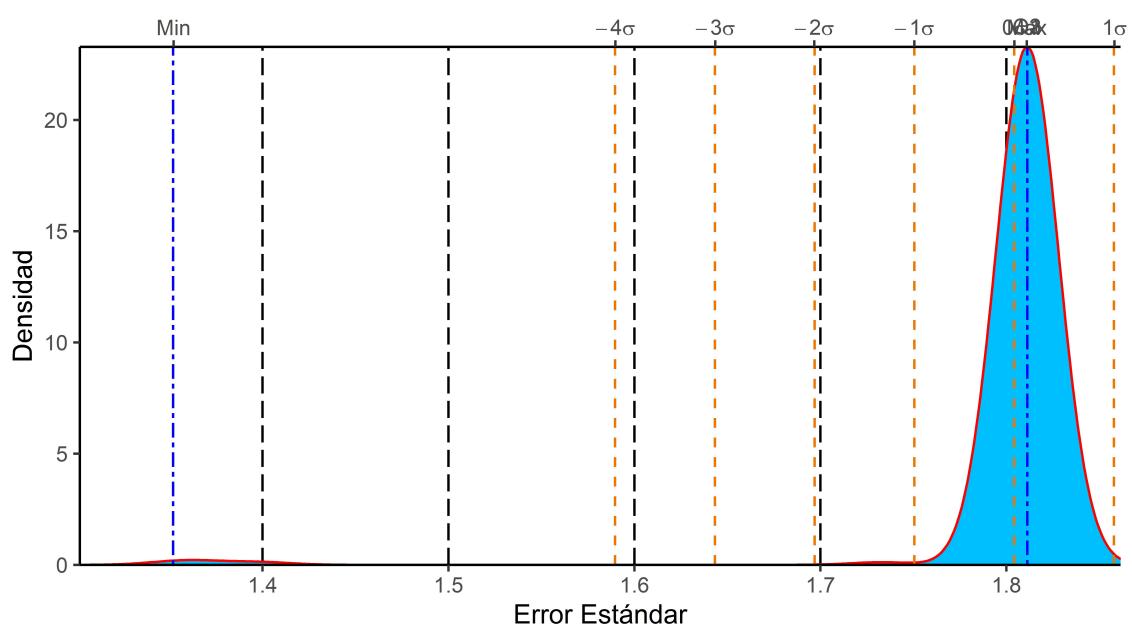


Figura 60: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,0,3) con la función auto.arima()



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skewness	CV
1.35	1.81	1.81	1.81	1.81	0.05	63.79	-7.88	0.03

Figura 61: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,0,3) con sobreparametrización

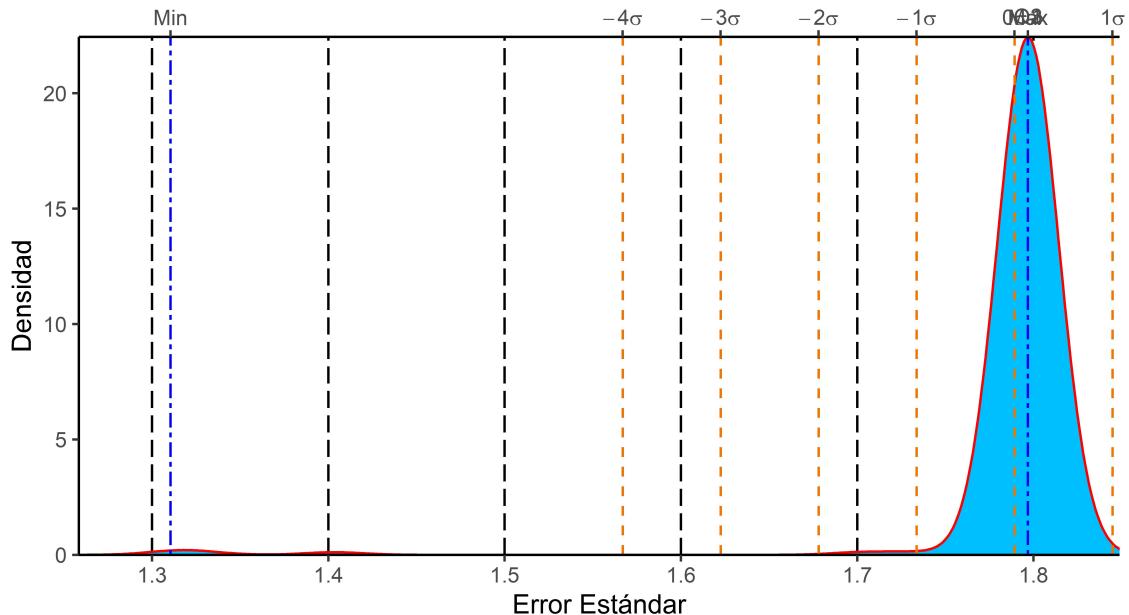


Figura 62: Errores estándares de los pronósticos obtenidos de los datos generados mediante un ARIMA(2,0,3) con el modelo ARIMA estándar

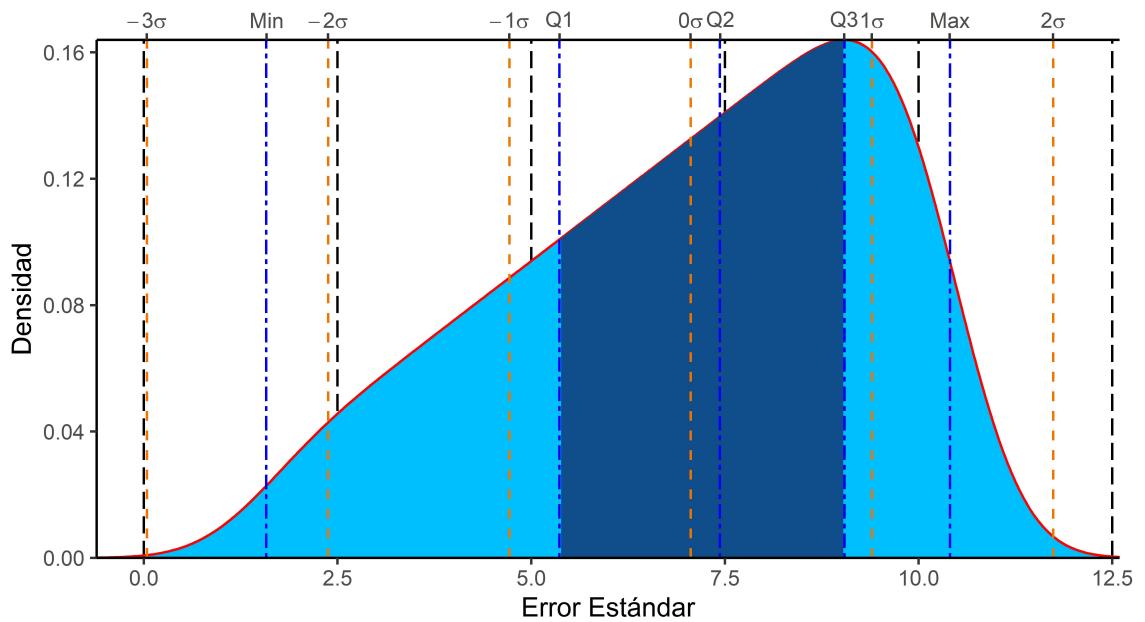


Figura 63: Pronóstico de los datos generados mediante un ARIMA(4,0,2) según el método de estimación

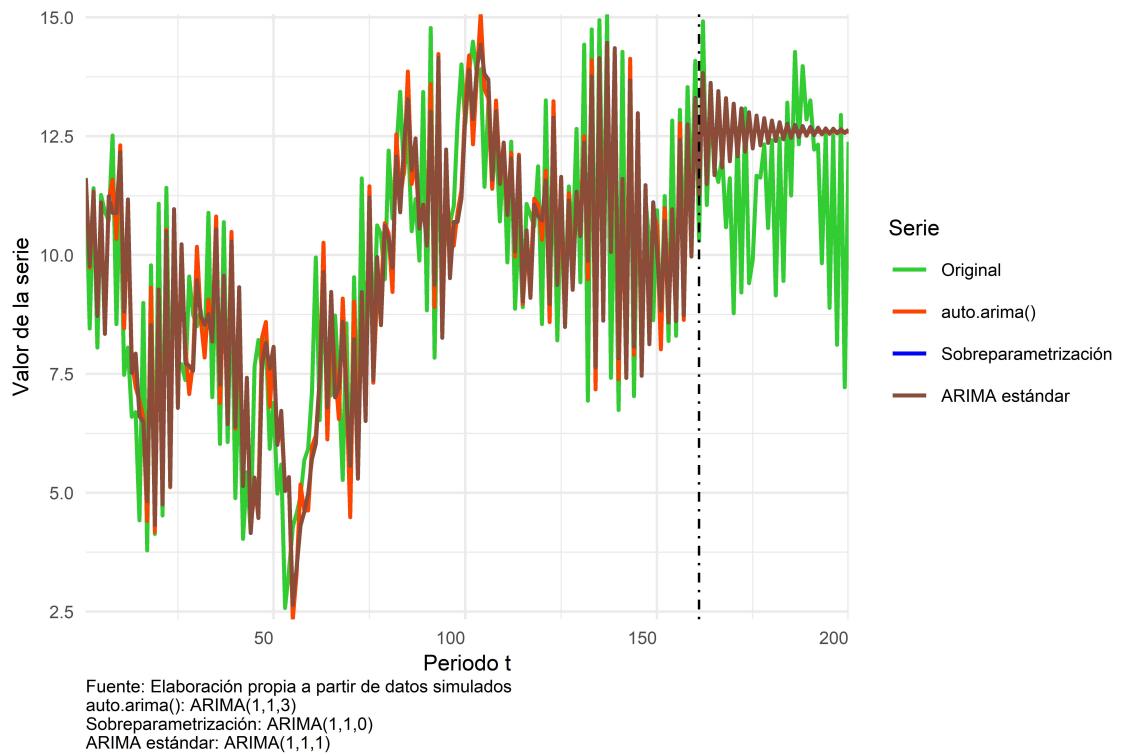
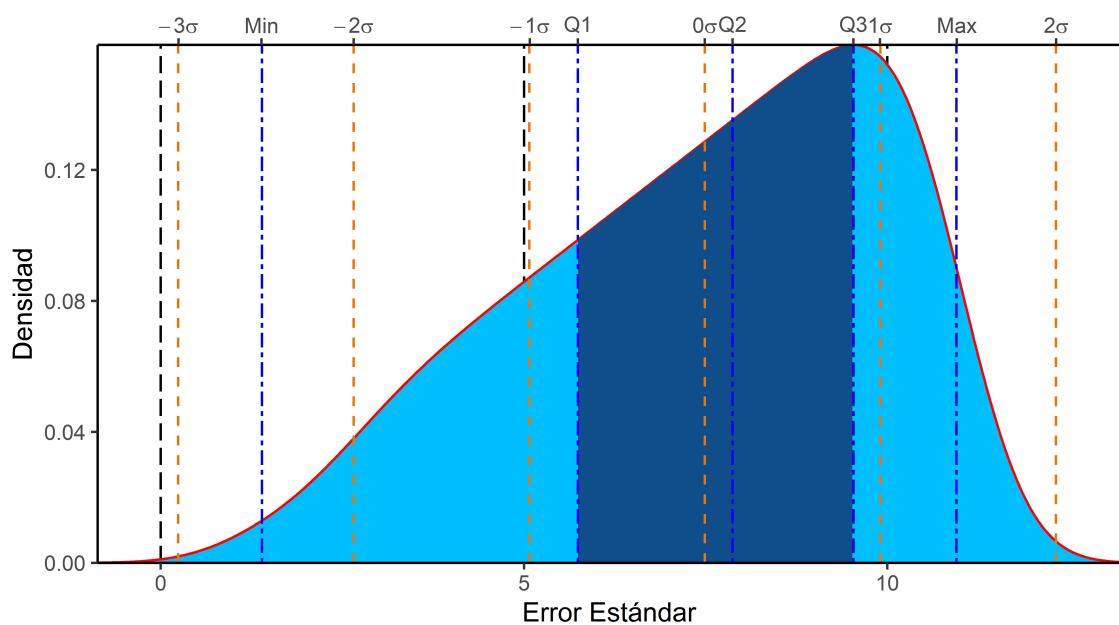
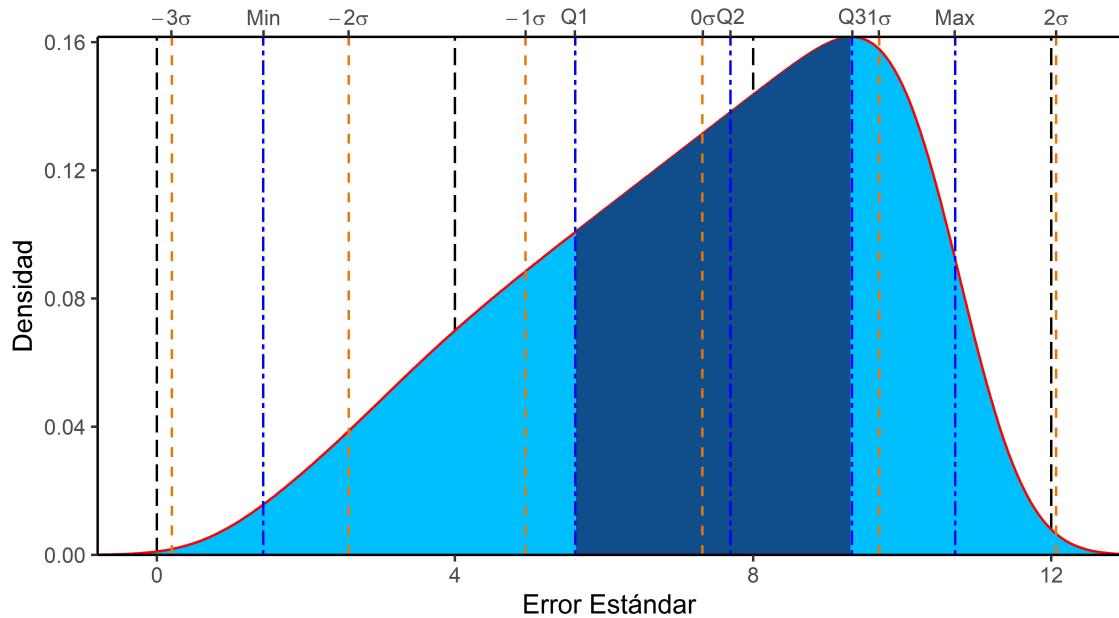


Figura 64: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(4,0,2) con la función auto.arima()



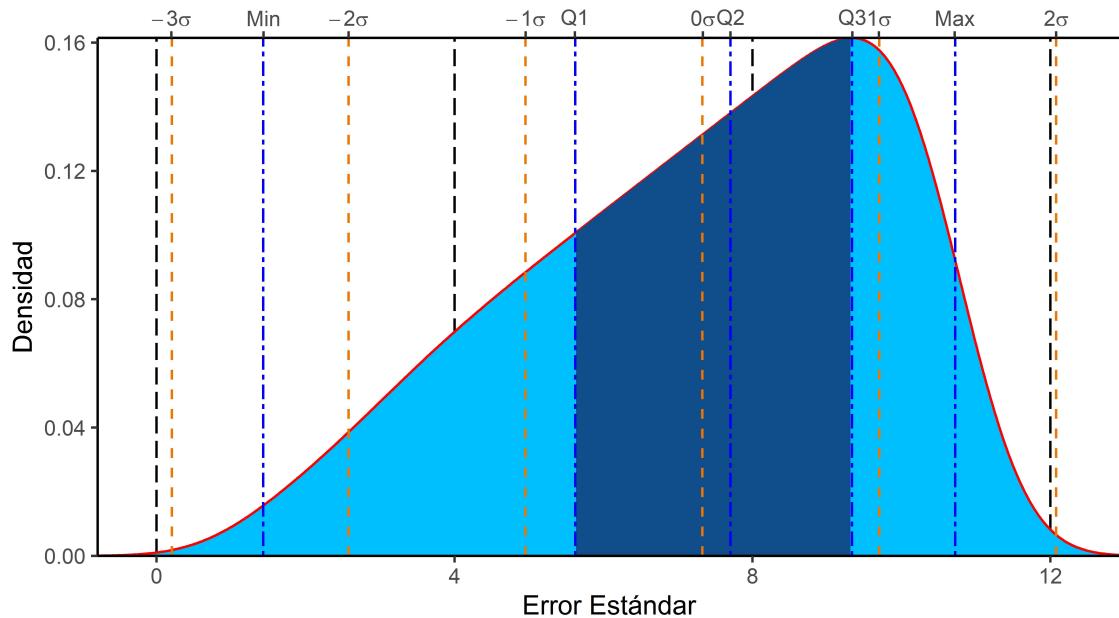
Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
1.39	5.74	7.87	9.53	10.95	2.42	2.3	-0.5	0.32

Figura 65: Errores estándar de los pronósticos obtenidos de los datos generados mediante un ARIMA(4,0,2) con sobreparametrización



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
1.43	5.61	7.7	9.33	10.71	2.37	2.31	-0.51	0.32

Figura 66: Errores estándares de los pronósticos obtenidos de los datos generados mediante un ARIMA(4,0,2) con el modelo ARIMA estándar



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
1.43	5.62	7.71	9.34	10.72	2.37	2.31	-0.51	0.32

Figura 67: Mortalidad por causa externa entre los años 2000 y 2017

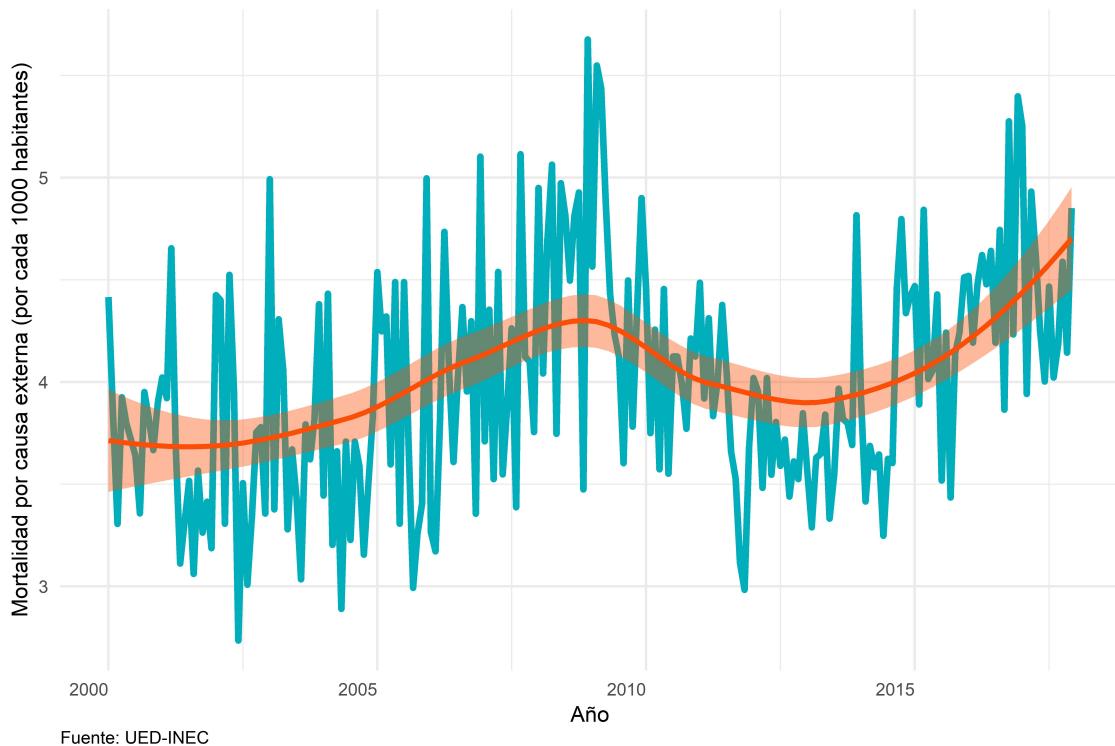


Figura 68: Intereses y comisiones del sector público en el periodo 2007 al 2018

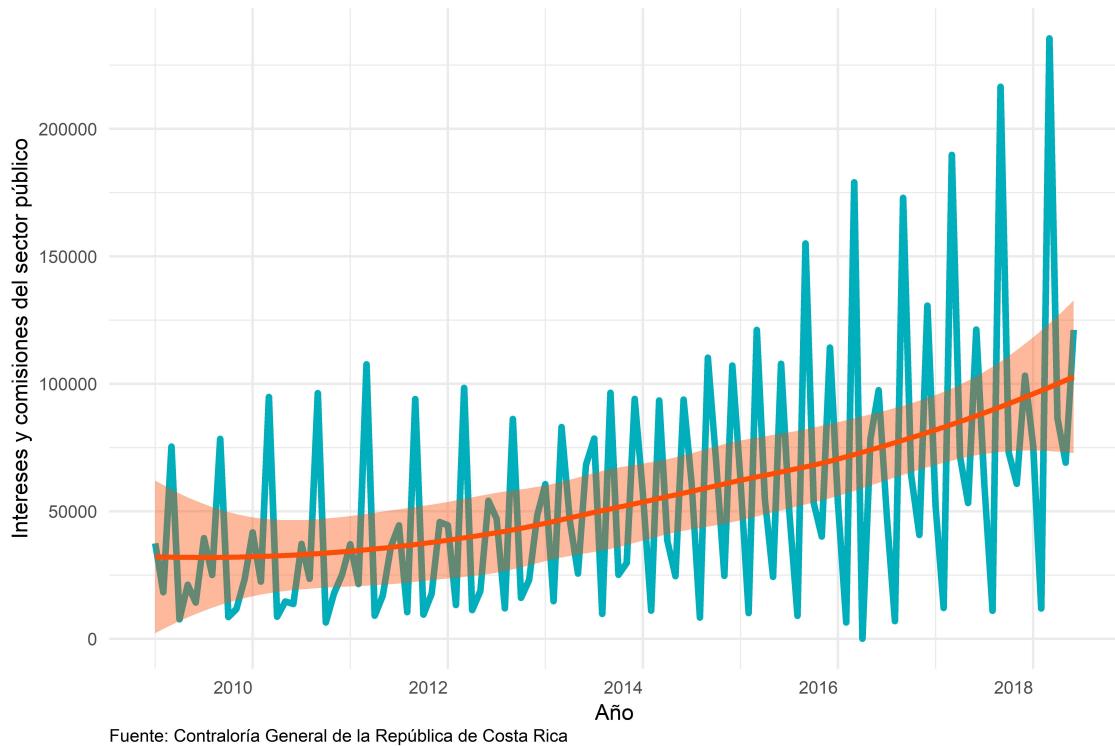


Figura 69: Intereses y comisiones del sector público en el periodo 2007 al 2018 según mes

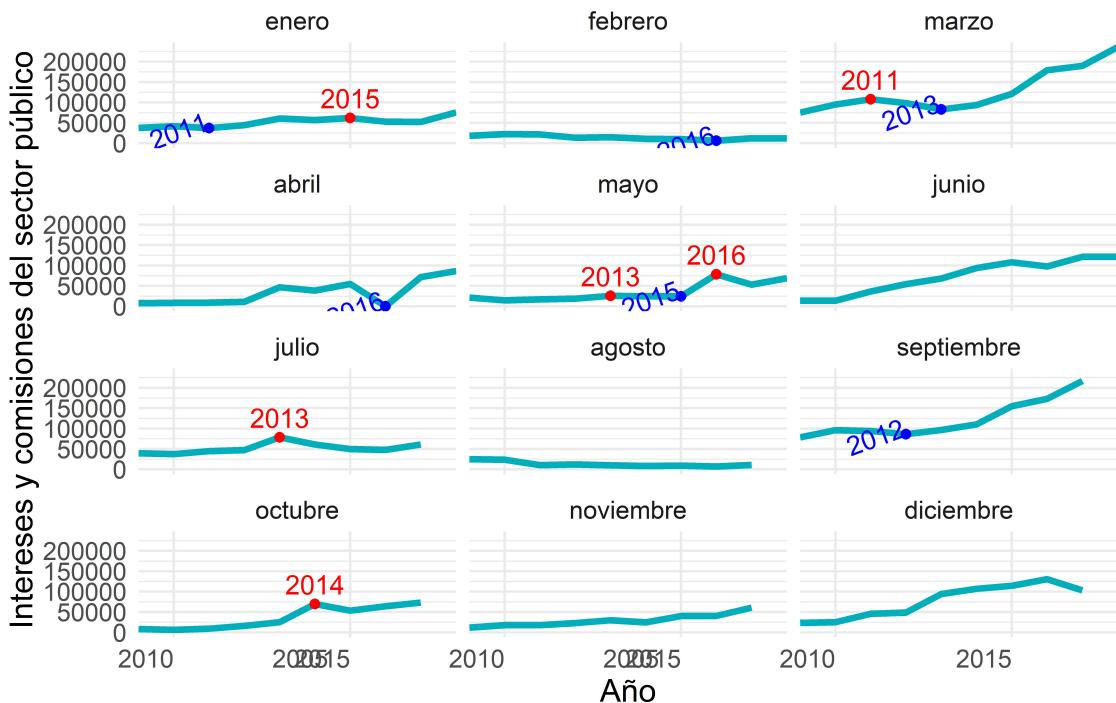


Figura 70: Descomposición de la serie de Incentivos salariales en el periodo 2007 al 2018

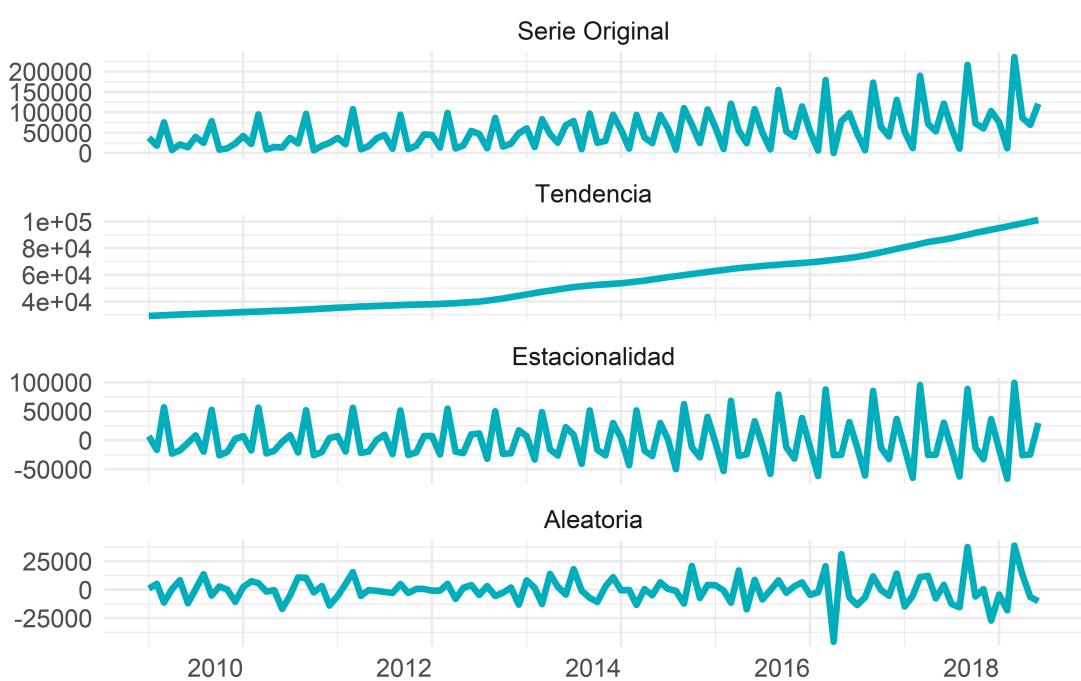


Figura 71: Autocorrelación de los datos diferenciados de la mortalidad por causa externa

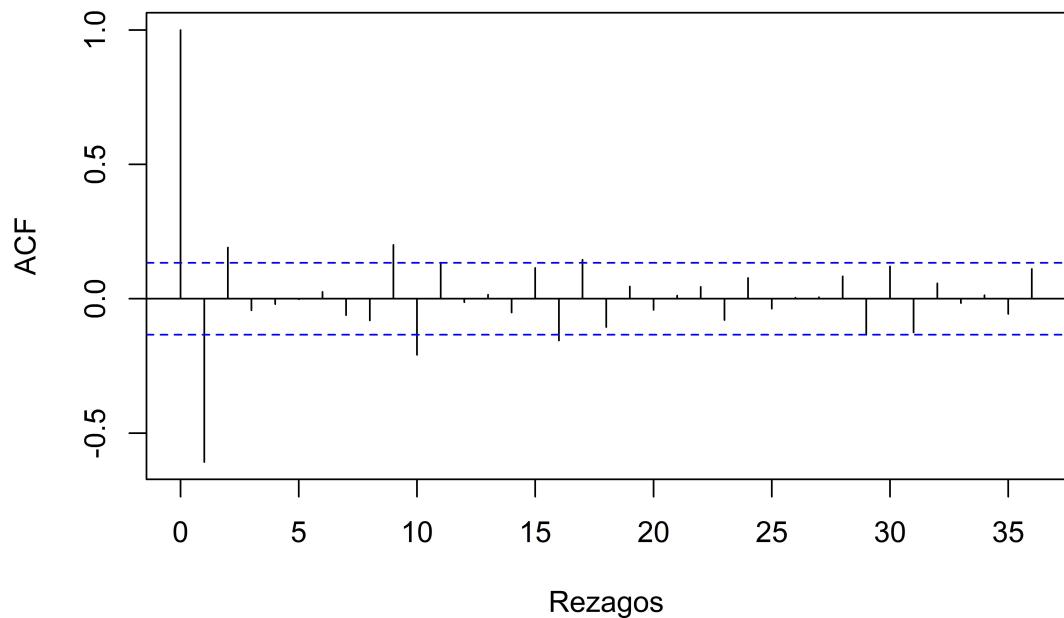


Figura 72: Autocorrelación parcial de los datos diferenciados de la mortalidad por causa externa

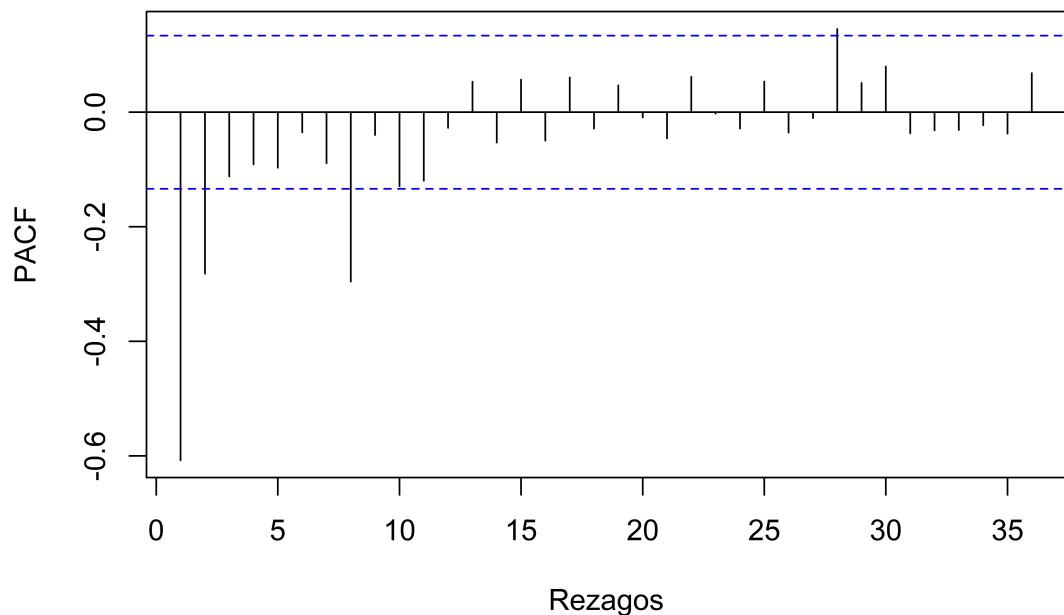


Figura 73: Autocorrelación de los datos diferenciados de la serie de intereses y comisiones del sector público

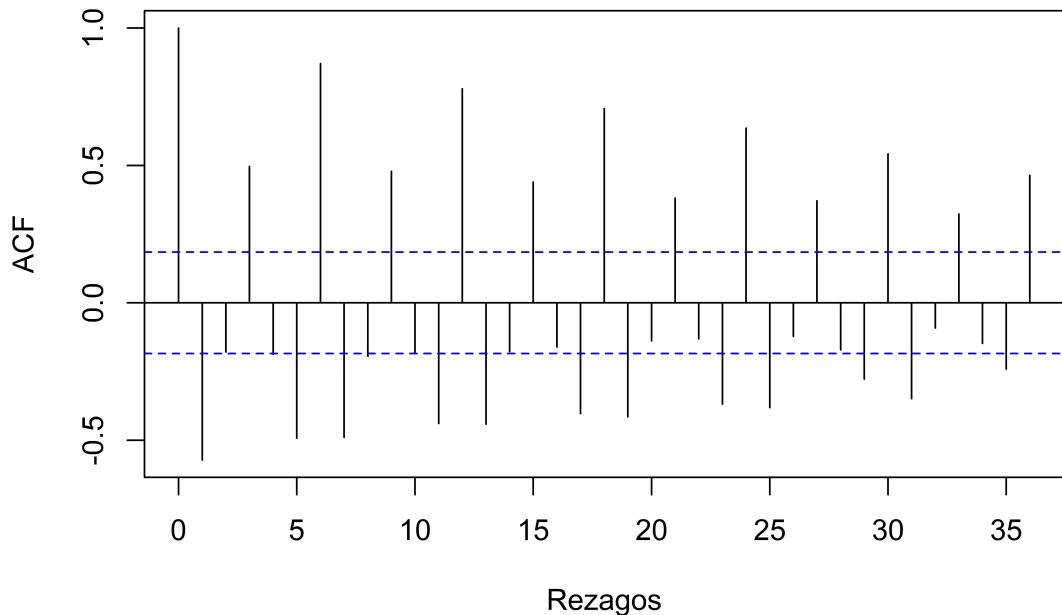
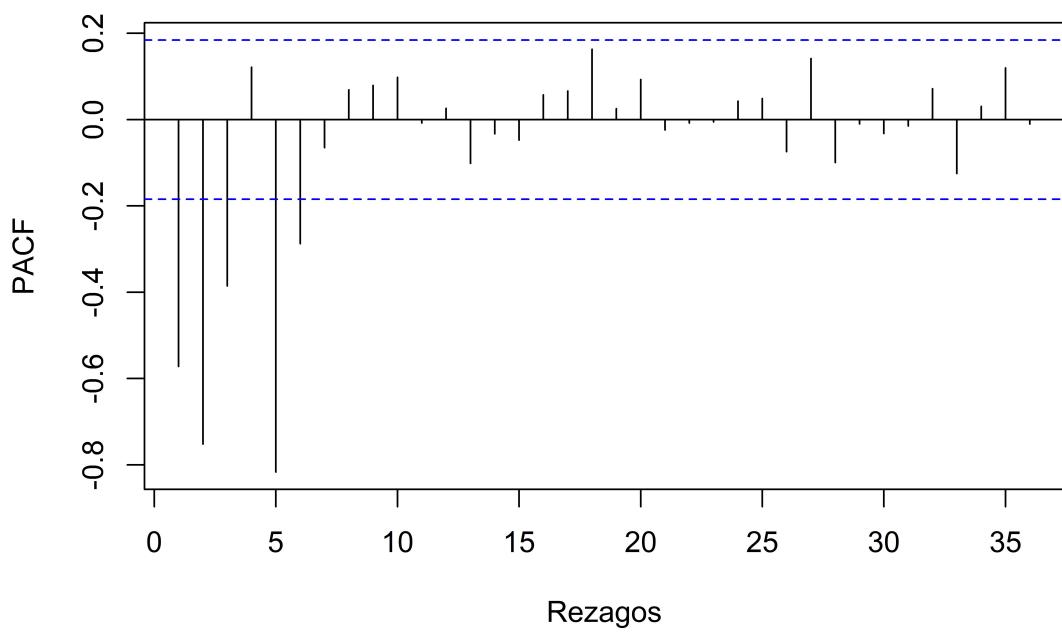


Figura 74: Autocorrelación parcial de los datos diferenciados de la serie de intereses y comisiones del sector público



Cuadro 9: Coeficientes de las ecuaciones de estimación según método de ajuste

Serie real	Coeficiente	auto.arima()			ARIMA estándar			Sobreparametrización		
		Puntual	L.I.	L.S.	Puntual	L.I.	L.S.	Puntual	L.I.	L.S.
	AR1	-0,11	-0,29	0,08	-0,22	-0,42	-0,03	0,71	0,45	0,97
	AR2	-	-	-	-	-	-	0,24	0,04	0,44
	MA1	-0,82	-0,94	-0,7	-0,76	-0,91	-0,61	-0,72	-0,95	-0,48
Mortalidad por causa externa	SAR1	-	-	-	-0,03	-0,2	0,13	-	-	-
	SMA1	-	-	-	-1	-1,57	-0,43	-1	-1,29	-0,71
	AR1	-	-	-	-0,44	-0,64	-0,24	-	-	-
	MA1	-0,44	-0,62	-0,27	-0,97	-1,05	-0,88	-1,45	-1,63	-1,27
	MA2	-	-	-	-	-	-	0,49	0,31	0,67
Intereses y comisiones	SAR1	-	-	-	0,2	-0,92	1,31	-	-	-
	SMA1	-	-	-	-0,11	-1,21	0,99	-	-	-

Fuente: Elaboración propia a partir de datos simulados.

Figura 75: Comportamiento de los errores asociados a los modelos estimados con la serie de mortalidad por causa externa

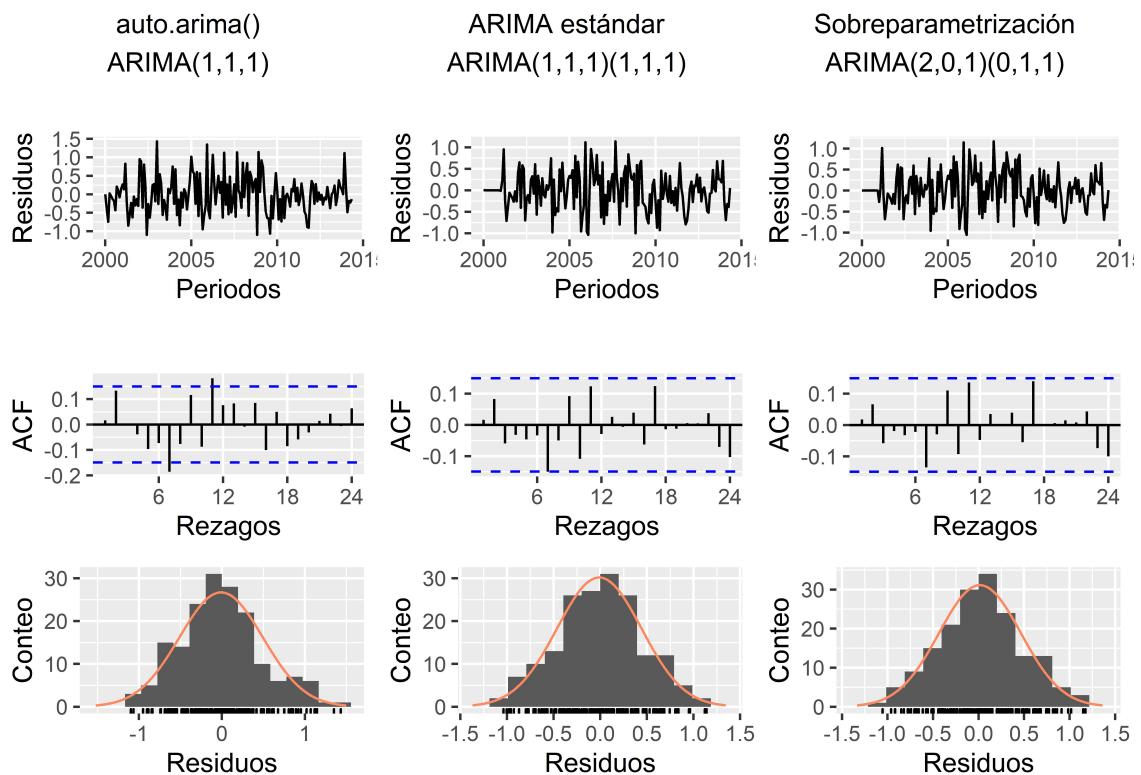
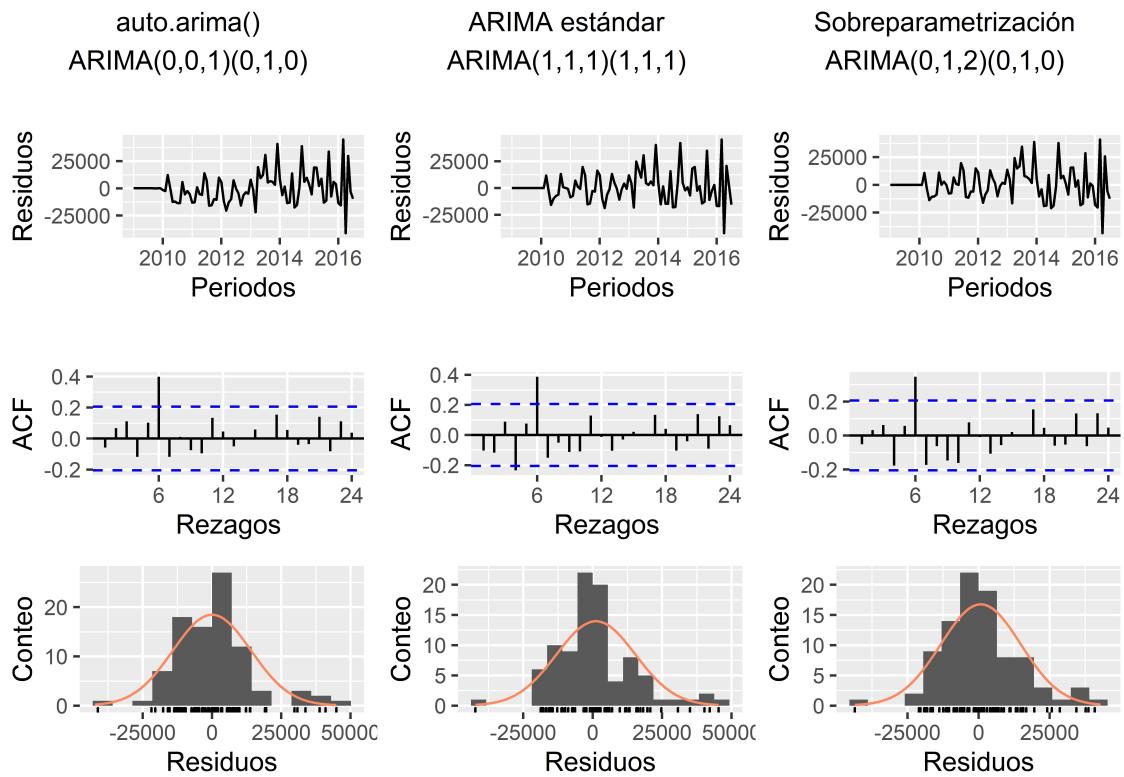


Figura 76: Comportamiento de los errores asociados a los modelos estimados con la serie de intereses y comisiones del sector público



Cuadro 10: Medidas de bondad de ajuste y de rendimiento según el método de estimación para los conjuntos de entrenamiento y validación a partir de las series cronológicas reales

Proceso original	Datos	Estimación	AIC	AICc	BIC	RMSE	MAE	MAPE
Mortalidad por causa externa	Entrenamiento	auto.arima()	258,4	258,44	267,84	0,51	0,4	10,05
		ARIMA estandar	225,76	225,83	241,14	0,46	0,36	9,14
	Validación	Sobreparametrización	223,41	223,48	238,81	0,46	0,36	9
Intereses y comisiones	Entrenamiento	auto.arima()	129,77	129,95	135,06	0,77	0,66	14,36
		ARIMA estandar	166,96	167,25	175,76	0,84	0,73	16,04
	Validación	Sobreparametrización	132,51	132,81	141,32	0,72	0,61	13,29
	Entrenamiento	auto.arima()	2002,15	2002,25	2009,26	14028,52	10046,55	4341,75
		ARIMA estandar	2008,17	2008,32	2019,95	14184,93	9681,33	4443,19
	Validación	Sobreparametrización	2001,82	2001,91	2008,89	14003,13	9853,75	4586,75

Fuente: Elaboración propia a partir de datos simulados

Figura 77: Pronóstico de la Tasa de mortalidad por causa externa (TMCE) según el método de estimación

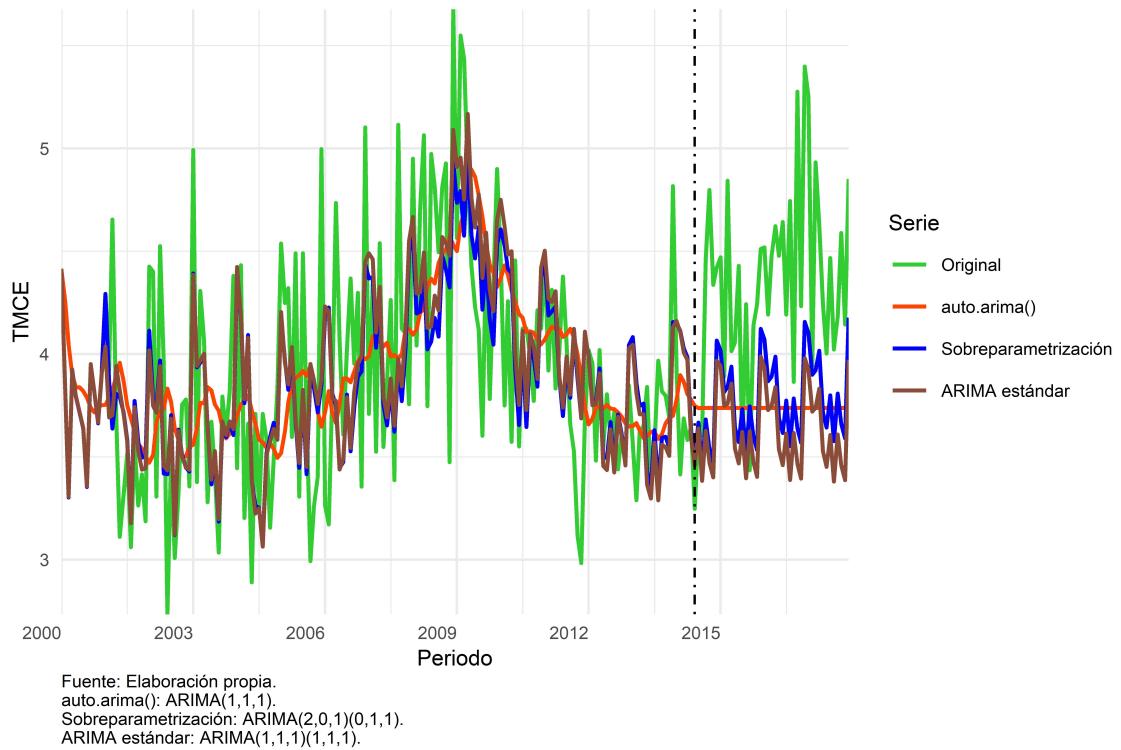
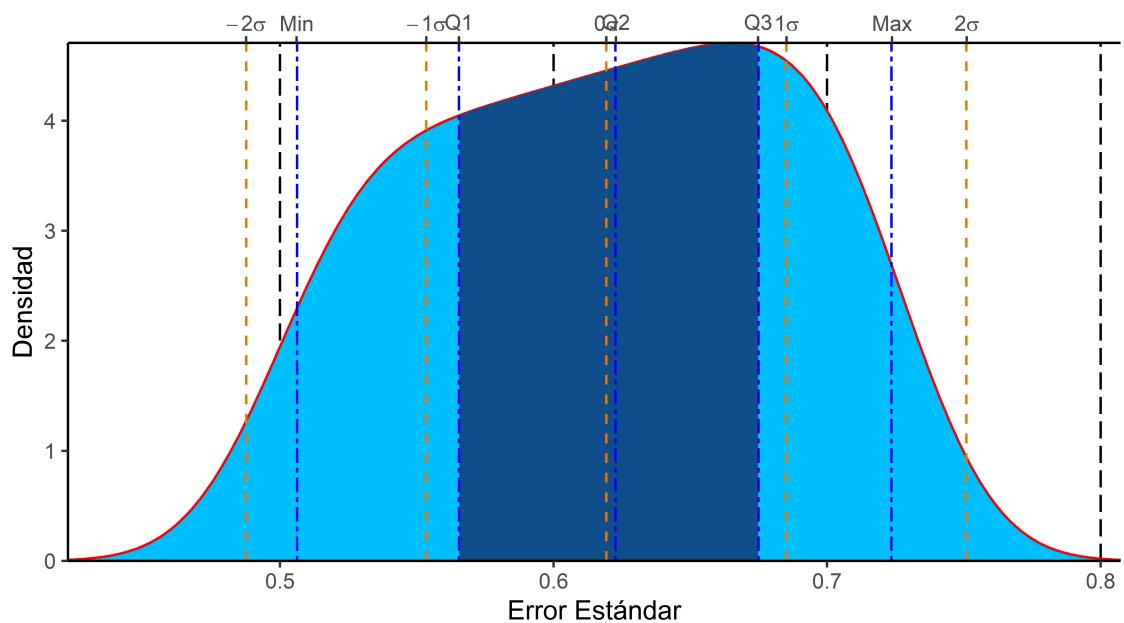


Figura 78: Errores estándar de los pronósticos obtenidos para la Tasa de mortalidad por causa externa (TMCE) con la función auto.arima()



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	CV
0.51	0.57	0.62	0.67	0.72	0.07	1.81	-0.12	0.11

Figura 79: Errores estándar de los pronósticos obtenidos para la Tasa de mortalidad por causa externa (TMCE) con sobreparametrización

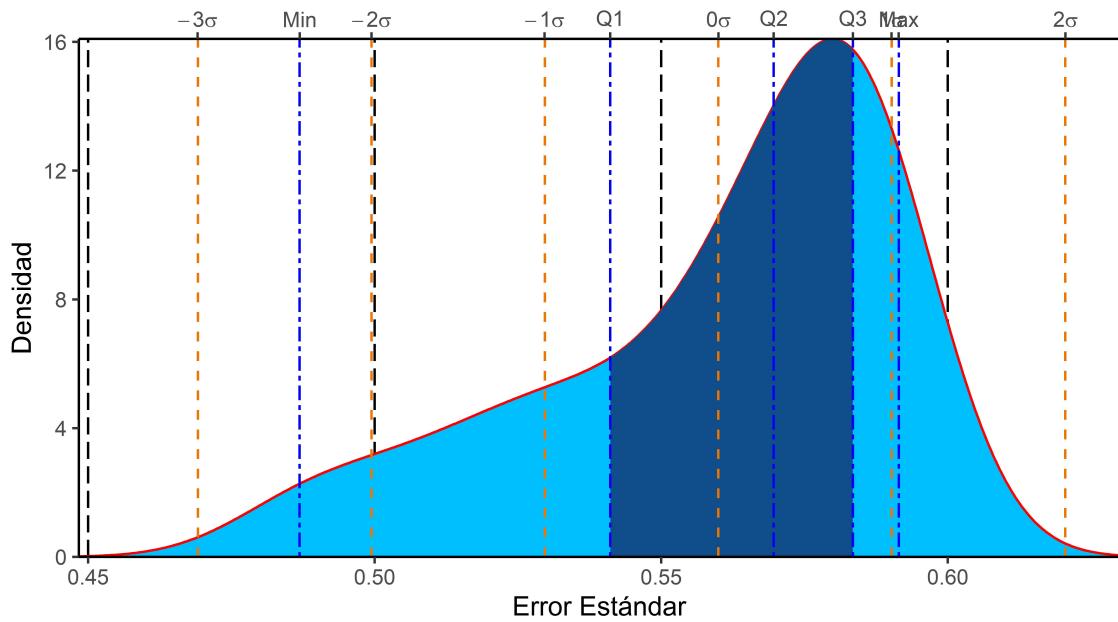


Figura 80: Errores estándares de los pronósticos obtenidos para la Tasa de mortalidad por causa externa (TMCE) con el modelo ARIMA estándar

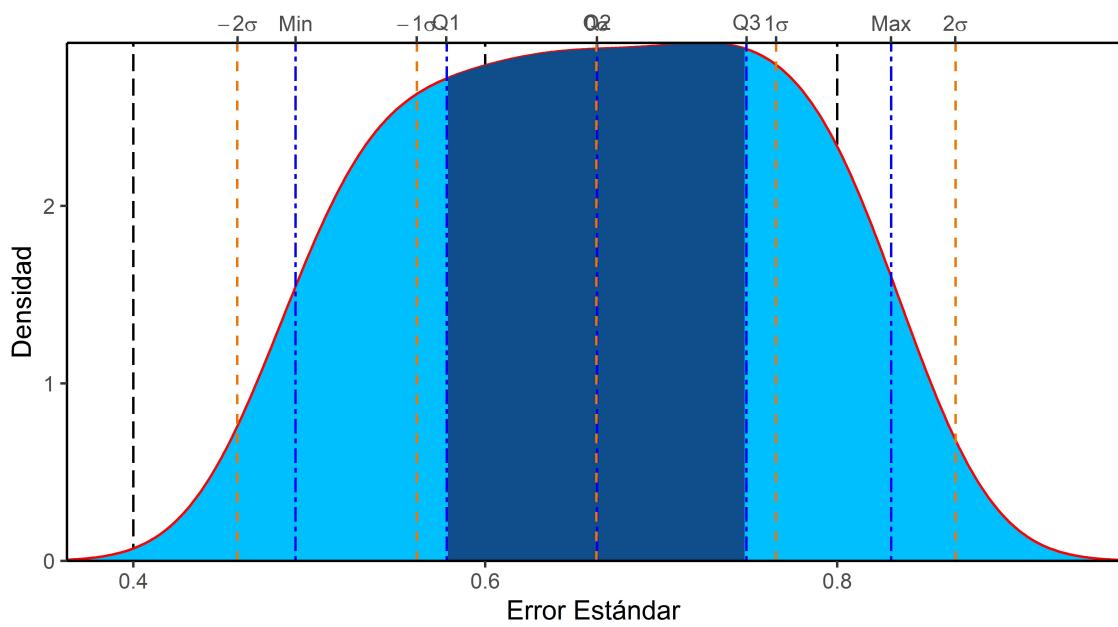


Figura 81: Pronóstico de la serie de intereses y comisiones del sector público según el método de estimación

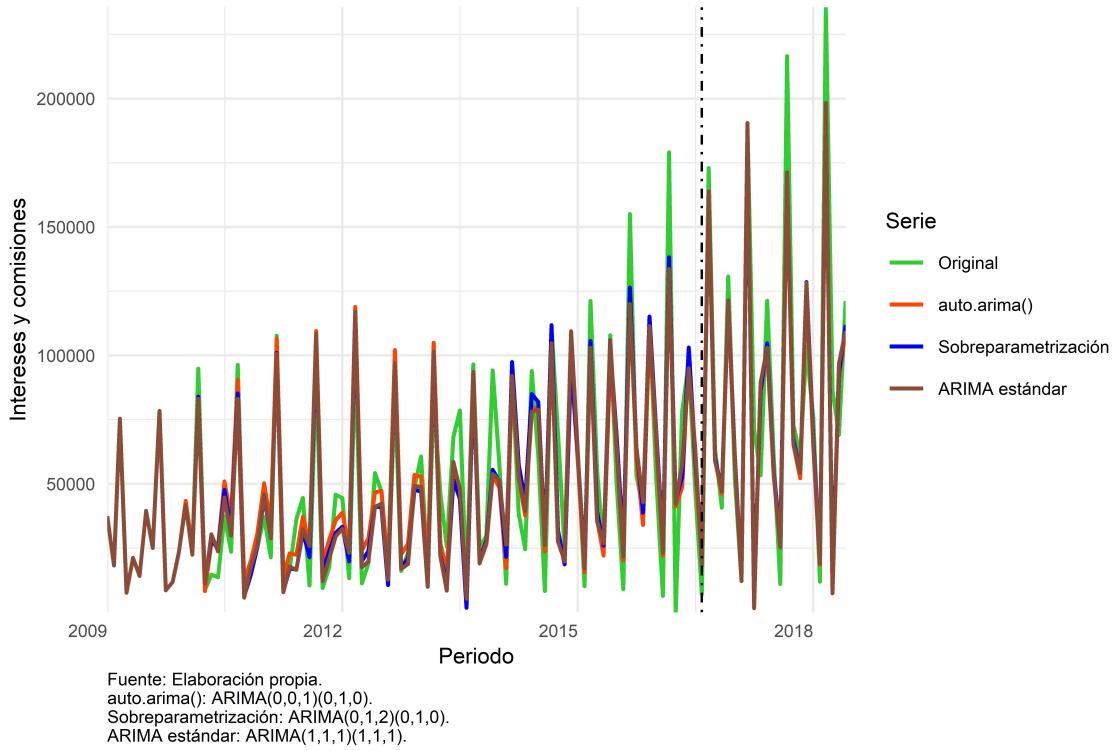
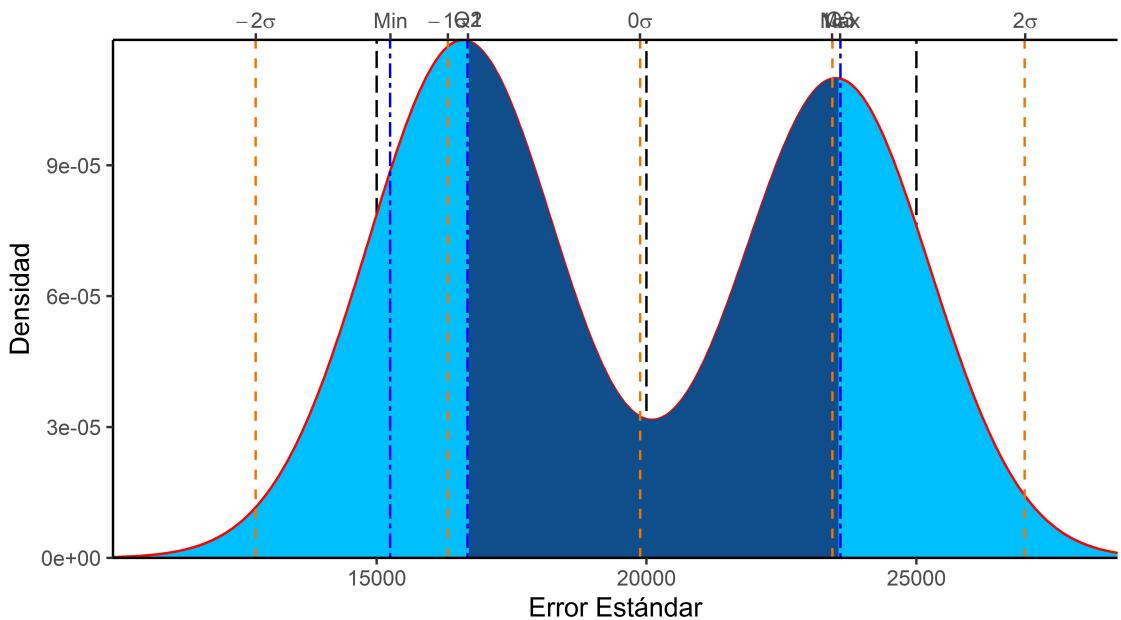


Figura 82: Errores estándar de los pronósticos obtenidos para la serie de intereses y comisiones del sector público con la función auto.arima()



Min	Q1	Median	Q3	Max	SD	Kurtosis	Skweness	C
250.59	16682.48	16682.48	23592.59	23592.59	3562.07	1.05	0.07	0.

Figura 83: Errores estándar de los pronósticos obtenidos para la serie de intereses y comisiones del sector público con sobreparametrización

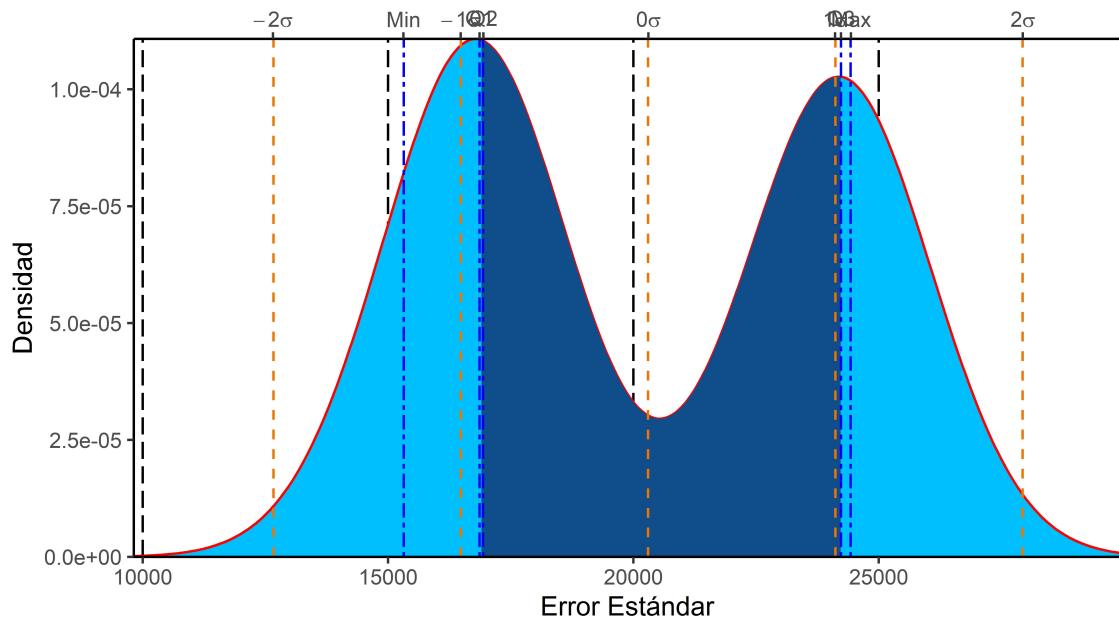
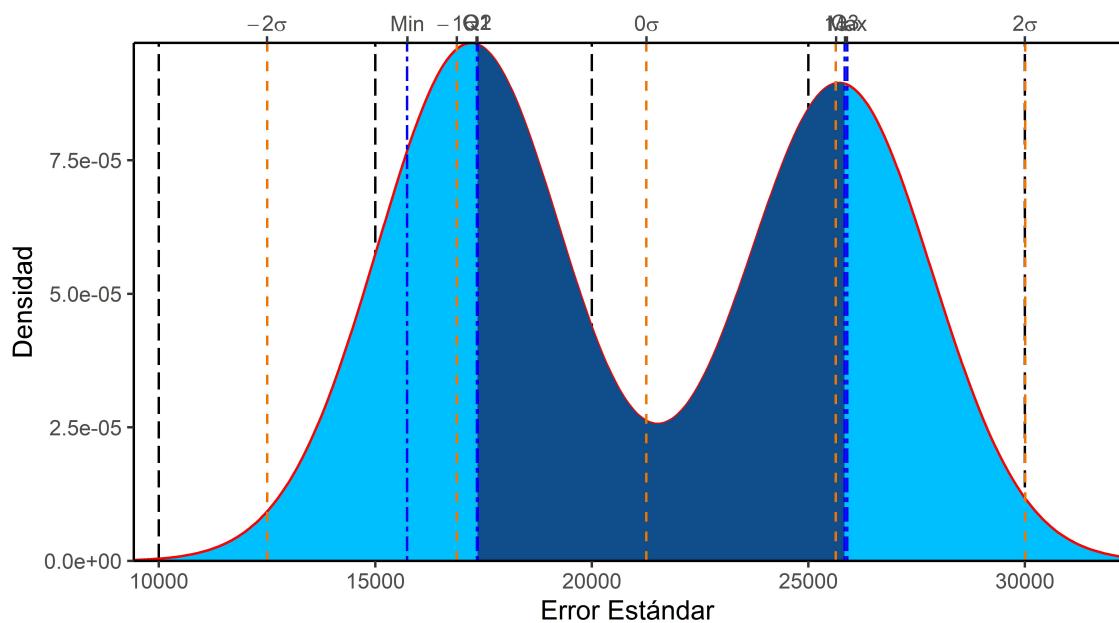


Figura 84: Errores estándares de los pronósticos obtenidos para la serie de intereses y comisiones del sector público con el modelo ARIMA estándar



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/ftp/arxiv/papers/1302/1302.6613.pdf>
- Agrawal, R., & Adhikari, R. (2013). An introductory study on time series modeling and forecasting. *Nova York: CoRR.*
- Aphalo, P. J. (2021). *Ggpmisc: Miscellaneous extensions to 'ggplot2'*. <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=_DeYu/_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%20OTOA_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*

- nal of Forecasting*, 22(3), 443–473. <https://doi.org/10.1016/j.ijforecast.2006.01.001>
- Donoso, E. (2004). Desigualdad en mortalidad infantil entre las comunas de la provincia de 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. (2010). *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. (2013). Morbilidad y mortalidad en costa rica. *La Nacion*. <https://bit.ly/2xWUeXU>
- 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
- McLeod, A. I. (1999). Necessary and sufficient condition for nonsingular fisher information matrix in ARMA and fractional ARIMA models. *The American Statistician*, 53(1), 71–72.
- 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. (1979a). *Macroeconomic theory*. Academic Press. <https://books.google.co.cr/books?id=X6u7AAAAIAAJ>
- Sargent, T. J. (1979b). *Macroeconomic theory* (pp. 286–290). Academic Press. <https://faculty.wcas.northwestern.edu/lchrist/finc520/wold.pdf>

- 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>