Sección de Matemáticas
Universidad de La Laguna

Jorge Guerra Rodríguez

Fundamentos y variantes de los modelos ARIMA para el análisis de series temporales. Aplicación a la estadística universitaria.

Fundamentals and variations of ARIMA models for time series analysis. Applications to university statistics.

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Junio de 2022

Dirigido por
Carlos Pérez González
Andrés Francisco Palenzuela López

Carlos Pérez González
Matemáticas, Estadística e Investigación Operativa
Universidad de La Laguna
38200 La Laguna, Tenerife

Andrés Francisco Palenzuela López
Gabinete de Análisis y Planificación
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

Aunque podría pensarse lo contrario, haber llegado a este punto no se trata de un cúmulo de casualidades. Todo ha sido tan perfecto que me da la impresión de que ya estaba escrito. Al escribir las últimas líneas de este trabajo, llegan a mi mente todas las personas que me han acompañado en este largo recorrido y que, sin saberlo, me han guiado hasta aquí.

Mi padre, mi madre y mi hermano han sido siempre mi mayor apoyo. No recuerdo un momento de debilidad en el que no me hayan reconfortado. Siempre se han mostrado comprensivos a mis cambios de humor, atentos a mis necesidades y han confiado en mí incluso cuando parecía que todo estaba perdido. Gracias por haber contestado siempre al teléfono. Gracias por brindarme la oportunidad del estudio y transmitirme desde pequeño el valor de la constancia.

Este año termina para mí la vida universitaria, pero empieza para mi hermano. Miguel, deseo que disfrutes de estos años: aprende lo que te apasiona, abre los ojos y colecciona experiencias que vivirán en tus recuerdos para el resto de tu vida.

Indudablemente, nada de esto sería posible sin los profesores y profesoras del grado. Gracias a todo el profesorado de vocación que sabe trasmitir la pasión que siente por las matemáticas a los demás. Gracias, también, a mi tutor Carlos.

También, quiero agradecer a Andrés y a toda la familia del Gabinete de Análisis y Planificación de la Universidad de La Laguna el entorno tan agradable del que me han hecho sentir parte estos últimos meses.

A la gente maravillosa que la vida me ha dado la oportunidad de conocer este año: Sole, María, Jesús y todo el grupo "SICUE" de matemáticas y física que me han incluido en la experiencia de vivir un año de intercambio en Cana-

rias sin moverme de casa y han conseguido instalarse en mi corazón por siempre.

Gorka, gracias por haber elegido Canarias como destino de intercambio, haberte cruzado en el camino que es mi vida y convertirte en un apoyo incondicional. Eres la última pieza de este puzle.

No puedo terminar de escribir esta sección de agradecimientos sin acordarme de ustedes: Jonathan y Fran. Gracias por haber sido una condición necesaria en esta aventura. Me siento muy orgulloso de ustedes porque son la viva imagen del esfuerzo y la constancia. Gracias por estar siempre, por ser mis compañeros de batallas y por todos los momentos inolvidables que me han regalado. Todo lo demás termina, pero espero que nuestra amistad se mantenga constante a medida que el tiempo tiende a infinito.

Finalmente, nada de esto podría haber sido posible sin ti, Jorge. Sigues siendo el mismo niño asustado que cruzó las puertas de la facultad en septiembre de 2018, pero con experiencias que permanecerán grabadas en tu corazón hasta el final de los días. No trates de crecer rápido, pero tampoco te aferres a recuerdos del pasado. Vive el presente, nunca te rindas y recuerda que los sueños no se cumplen si no se persiguen.

Si pudiera volver atrás, no cambiaría nada. Si pudiera volver atrás, esperaría a que el tiempo escribiera de nuevo esta misma historia.

Jorge Guerra Rodríguez
La Laguna, 13 de junio de 2022

Resumen · Abstract

Resumen

El análisis de series temporales es un conjunto de técnicas estadísticas que permite describir y prever el comportamiento de una serie temporal y modelizar el proceso estocástico del que estas provienen con el objetivo de hacer predicciones. La familia de modelos ARIMA es ampliamente utilizada y presenta buenos resultados para horizontes de predicción cercanos en el tiempo de series temporales que presentan comportamientos estacionales. Este trabajo introduce la familia de modelos ARIMA y analiza por completo una serie ofrecida por el Gabinete de Análisis y Planificación de La Universidad de La Laguna prestando atención a sus aspectos de estacionariedad, estacionalidad y diagnosis para construir el modelo que mejor se ajuste a la misma. La serie temporal se analiza con el entorno y lenguaje de programación estadístico R.

Palabras clave: *ARIMA – Serie temporal – Programación en R – Predicción.*

Abstract

Time series analysis is a group of statistical techniques used to describe and anticipate the behaviour of a time series, study the stochastic process where the series comes from and making predictions. The ARIMA models are widely used with an outstanding performance in short-term forecasts of series where seasonality is shown. This thesis introduces the ARIMA models and analyses a time series provided by the Gabinete de Análisis y Planificación of La Laguna University considering the hypothesis of the models and trying to build the model that best fit to the data. The time series is analyzed with the R environment for statistical computing.

Keywords: *ARIMA – Time series – R programming – Forecasting.*

Contenido

Introducción

Las predicciones llevan fascinando a la sociedad durante siglos. El ser humano es un animal curioso que se siente intrigado por conocer qué pasará en el futuro. Antiguamente, se asociaban las predicciones a talentos divinos que solo ciertas personas eran afortunadas de desarrollar. Por suerte, hoy en día, cualquier persona con suficientes conocimientos matemáticos podría aventurarse a, no solo hacer predicciones, sino también a estimar el error de las mismas, sin necesitar ninguna virtud divina o celestial más que el poder que le han otorgado las matemáticas.

Las técnicas predictivas se requieren diariamente en diversas circunstancias del ámbito empresarial, económico, social, tecnológico o científico. Pueden hacerse predicciones para acontecimientos que sucederán dentro de varios años, pero también para horizontes mucho más cercanos como meses o minutos. Esto hace ver que hay una infinidad de situaciones que proponen problemas que precisan de predicciones para ser resueltos, y una forma efectiva y eficiente de abordarlos es mediante el análisis de series temporales.

El análisis de series temporales se basa en un conjunto de datos ordenados en el tiempo que recoge medidas de cierto fenómeno de interés. Si bien estas técnicas conforman conocimientos esenciales para el matemático actual, no son abordadas el grado. La temática de este trabajo nace de estas circustancias, al ser la oportunidad perfecta para introducirse en los modelos predictivos y completar la formación de un *cuasigraduado* en matemáticas.

Este trabajo, pues, se plantea como una introducción al análisis de series temporales mediante la familia de modelos ARIMA, cuyo estudio se presenta en el libro de Box y Jenkins "*Time Series Analysis: Forecasting and Control*" publicado en el año 1970. La construcción de los modelos ARIMA tiene, entonces, como propósito predecir y, particularmente, muestra resultados satisfactorios

para predicciones a corto plazo.

Todos los capítulos se sitúan en un nivel de carácter introductorio y tienen como objetivo guiar al lector por un recorrido que, en unas 40 páginas, consigue que este adquiera los conocimientos necesarios para entender la familia de modelos ARIMA y hacer predicciones basadas en ella.

Es verdad que, cuando se habla de hipótesis, los matemáticos somos poco objetivos y no podemos evitar mostrar cierta debilidad hacia ellas, pero no se trata de un capricho: sientan las bases de los modelos y si alguna de ellas falla, se obtendrán predicciones erróneas y conclusiones catastróficas. Por ello, aunque muchos recursos de internet las pasan por alto, este texto no actúa de la misma manera y se muestra cauteloso a los requerimientos de los modelos.

El Capítulo 1 se presenta como una introducción al análisis de series temporales dando una definición formal del concepto "serie temporal", exponiendo definiciones que serán esenciales para comprender los capítulos que le siguen y mostrando herramientas cruciales para el análisis de una serie temporal.

El Capítulo 2, por su parte, está dedicado a la exposición de los modelos que forman la familia ARIMA. En él se distingue si el proceso estocástico sobre el que se construye el modelo es estacionario o presenta estacionalidad. En el final de este capítulo se dedica una sección a la diagnosis del modelo, que no es más que la comprobación de sus hipótesis.

En mi opinión, si bien considero que el estudio teórico de los modelos es esencial para la correcta implementación de los mismos, este debe complementarse con la aplicación práctica a series temporales reales. Es por ello que el Capítulo 3 deja a un lado la retórica teórica y formal de los capítulos que lo preceden y se muestra como una panorámica de todo lo expuesto hasta el momento para analizar una serie temporal real cedida por el Gabinete de Análisis y Planificación de la Universidad de La Laguna.

La elaboración de los modelos ARIMA es una tarea impensable sin la ayuda de un ordenador debido a, no tanto la complejidad, sino la laboriosidad de los cálculos. Particularmente, en este trabajo se ha utilizado el entorno y lenguaje de programación R que es muy útil y ampliamente utilizado en el ámbito de la estadística y el análisis de datos. La mayoría de las figuras expuestas han sido generadas con el lenguaje R, así como todos los resultados presentados en el tercer capítulo.

1

Conceptos básicos sobre series temporales

En este capítulo se introducen algunas ideas y conceptos fundamentales para el análisis de series temporales y procesos estocásticos. Se comienza con una definición formal de proceso estocástico de la que surge la definición de serie temporal. A continuación, se hace un breve recorrido por los conceptos de descomposición clásica, estacionariedad y diferenciación que serán cruciales para comprender la construcción de la familia de modelos $ARIMA$ del Capítulo 2, así como en el análisis de series temporales reales relativas al número de procedimientos de la sede electrónica de La Universidad de La Laguna que se planteará en el Capítulo 3. El capítulo culmina exponiendo algunas medidas de bondad de predicción y el método de la validación cruzada.

1.1. Series temporales

Intuitivamente, una **serie temporal** es una sucesión $\{y_t\}_{t=1}^{T}$ de datos u observaciones medidos en determinados periodos de tiempo equidistantes y ordenados cronológicamente. Las series pueden tener una **periodicidad** anual, trimestral, mensual, semanal, etc. según los periodos de tiempo en los que se han tomado los datos.

El análisis de series temporales se basa en un conjunto de técnicas que permiten no solo estudiar y modelizar el comportamiento general de la misma, sino que también busca hacer predicciones de los valores que se alcanzarán en un futuro.

A continuación, se definen algunos conceptos básicos con el propósito de describir formalmente qué es una serie temporal.

Definición 1.1. *Un **espacio de probabilidad** es una terna $(\Omega, \mathcal{F}, P)$ donde Ω es un conjunto no vacío llamado espacio muestral, $\mathcal{F}$ es una σ-álgebra de*

subconjuntos de Ω y P es una medida de probabilidad definida sobre todos los elementos de $\mathcal{F}$.

Definición 1.2. *Una **variable aleatoria** real en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ es una función $Y : \Omega \longrightarrow \mathbb{R}$ tal que la imagen inversa de cualquier intervalo de la forma $(-\infty, a]$ es un elemento de $\mathcal{F}$.*

Definición 1.3. *Un **proceso estocástico real** $\{Y_t\}$ es una familia de variables aleatorias reales $\mathbf{Y} = \{Y_t(\omega) \mid t \in C\}$ todas definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Si el conjunto índice, C, es tal que $C \subset \mathbb{Z}$, entonces se dice que el proceso estocástico es discreto, mientras que si C es un intervalo de $\mathbb{R}$, el proceso es continuo.*

Una vez se han introducido las Definiciones 1.1-1.3, es posible dar una definición formal del concepto serie temporal que, desde el punto de vista de Uriel [10] (1995), no es más que una **realización** finita de un proceso estocástico real y discreto.

Definición 1.4. *Una **serie temporal** $\{y_t\}_{t=1}^{T}$ es un conjunto de T valores observados de un proceso estocástico real discreto que están ordenados y equiespaciados en el tiempo. Para cierto periodo de tiempo t, la observación y_{t-h} se llama el h-ésimo desfase de y_t.*

Una serie temporal, por tanto, está definida sobre un periodo muestral que es tan solo una porción de la historia del proceso estocástico del que dicha serie procede. Es por ello que, en la práctica, el análisis de series temporales se realiza con un conjunto finito de datos, mientras que los modelos teóricos se construyen sobre su proceso estocástico asociado, pues consideran un número infinito de observaciones.

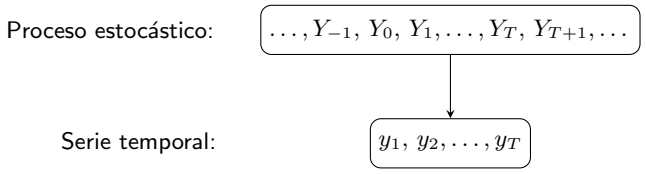

Figura 1.1. Procedencia de una serie temporal observada.

En el análisis clásico de una serie temporal, aparecen un conjunto de factores llamados **componentes** que, cuando interactúan entre sí, provocan fluctuaciones en la misma a lo largo del tiempo.

- Se dice que existe una **tendencia** cuando se aprecia un crecimiento, decrecimiento o estancamiento en los datos por un largo plazo de tiempo. En ocasiones se describe la tendencia como un cambio de dirección.

- La **componente cíclica** refleja comportamientos recurrentes sin una frecuencia fija en los datos a medio plazo. En el estudio de series temporales que se expone en este documento, no se hará distinción entre la tendencia y la componente cíclica.
- La **componente estacional** se aprecia cuando la serie presenta fluctuaciones en un corto periodo de tiempo, generalmente inferior o igual a un año. Manifiesta oscilaciones debidas a la influencia de ciertos fenómenos que se repiten de manera periódica en un año, como las estaciones; en una semana, por ejemplo los fines de semana; o las horas puntas de un día.

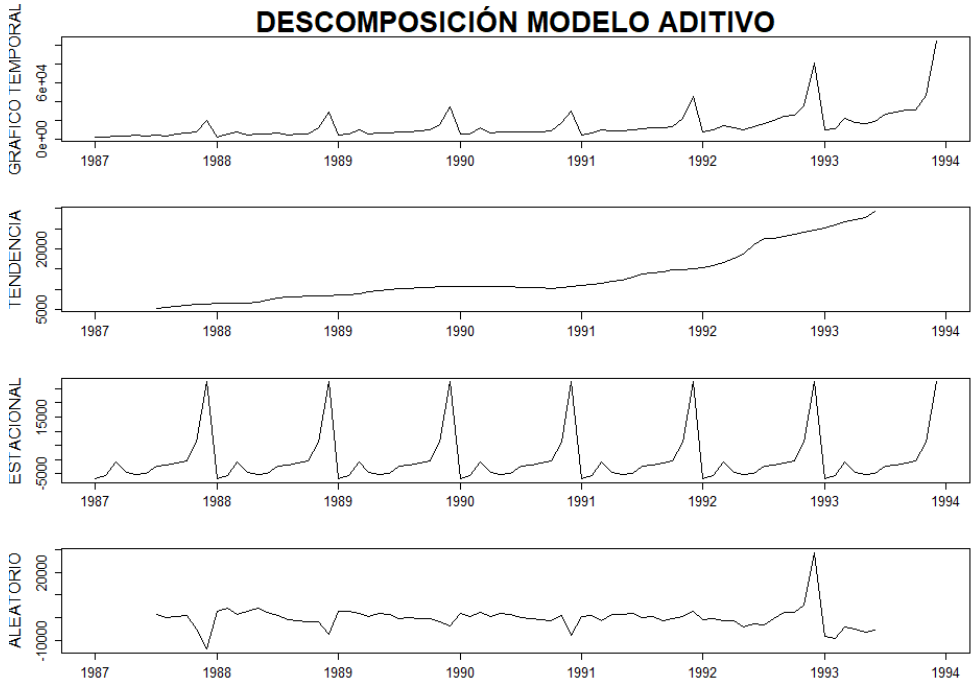

Figura 1.2. Descomposición en sus componentes de la venta de *souvenir* en una tienda de Queensland, Australia, desde enero de 1987 hasta diciembre 1993. Datos de Rob Hyndman.

1.2. Descomposición clásica de una serie temporal

Las series temporales pueden presentar patrones que se consideran resultantes de la composición de las tres componentes citadas en la Sección 1.1. El análisis, desde un enfoque clásico, de las series temporales trata de descomponer los datos en sus componentes. Este proceso recibe el nombre de **descomposición clásica** que, si bien no es la forma más sofisticada de analizar una serie, es útil para ilustrar la interacción de las componentes y sirve como un primer

contacto con con el conjunto de datos, pues ayuda a entender el comportamiento general del mismo.

La descomposición clásica considera que una serie temporal dada está compuesta por su tendencia, componente estacional y un residuo que recoge todos los efectos casuales de eventos fortuitos que no presentan un carácter periódico reconocible.

Principalmente, existen dos posibles modelos de forma que se asume uno u otro en función del comportamiento de la serie temporal $\{y_t\}_{t=1}^{T}$. Si S_t representa la componente estacional, T_t denota la tendencia y R_t es el residuo, y la serie no presenta fluctuaciones elevadas, se opta por un modelo aditivo que se expresa

$$y_t = S_t + T_t + R_t. \tag{1.1}$$

Si, en cambio, la componente estacional o la tendencia presentan grandes variaciones a lo largo del tiempo, entonces González Sierra [7] (2014) sugiere que será más apropiado asumir un modelo multiplicativo

$$y_t = S_t \times T_t \times R_t. \tag{1.2}$$

La Figura 1.2 muestra el gráfico temporal de una serie acompañada de su tendencia, componente estacional y errores aleatorios. Si se examina el gráfico temporal, puede sospecharse una tendencia creciente así como se aprecia un claro comportamiento estacional, pues cada año se repite el mismo patrón de ventas. Estas sospechas se confirman al prestar atención a los gráficos de la tendencia y componente estacional de la serie.

1.3. Estacionariedad

A grandes rasgos, un proceso estocástico $\{Y_t\} = \{Y_t \mid t \in \mathbb{Z}\}$ se dice estacionario si sus propiedades estadísticas no dependen del periodo de tiempo en el que se observe.

Definición 1.5. *Un proceso estocástico* $\{Y_t\}$ *es* ***estrictamente estacionario*** *si para cualesquier sucesión de periodos* $t_1 < t_2 < \cdots < t_n$ *de su historia, la distribución de probabilidad conjunta de* $(Y_{t_1}, \ldots, Y_{t_n})$ *coincide con la de* $(Y_{t_{1+h}}, \ldots, Y_{t_{n+h}})$ *para cualquier* $h = \pm 1, \pm 2 \ldots$

Las propiedades estadísticas de un proceso estocástico estacionario son muy diferentes a las de un proceso no estacionario. Es por ello que uno de los primeros pasos en el análisis de series temporales es estudiar si la serie es compatible con la hipótesis de estacionariedad, requerimiento esencial para el modelo $ARMA$ expuesto en la Sección 2.2.

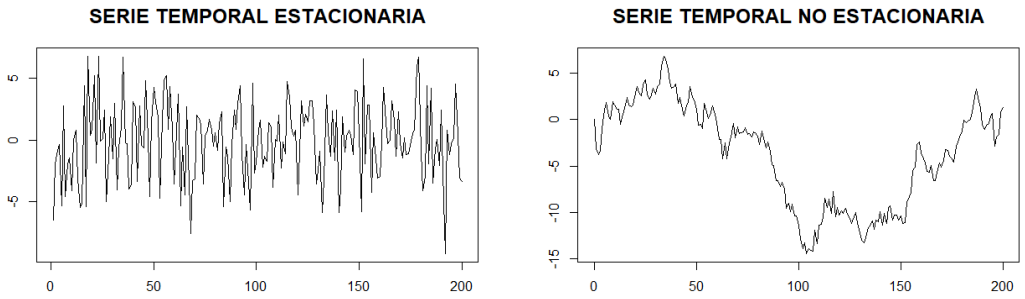

Figura 1.3. A la izquierda se representa una serie temporal estacionaria, mientras que a la derecha una no estacionaria.

Definición 1.6. *Un proceso estocástico real* $\{Y_t\}$ *con* $E[Y_t^2] < \infty$ *para todo* $t = 0, \pm 1, \pm 2, \ldots$ *se dice* ***débilmente estacionario*** *si:*

- $\mu_Y(t) = E[Y_t]$ *es constante para todo* $t = 0, \pm 1, \pm 2, \ldots$. *Es decir, no depende del valor de* t.
- $\gamma_Y(t+h, t) = Cov[Y_t, Y_{t+h}]$ *depende a lo sumo de* h *(entero), pero no del valor de* t *para todo* $t = 0, \pm 1, \pm 2, \ldots$.

En general, se seguirá el criterio de Brockwell [2] (2002) y, cuando se hable de estacionariedad en este texto, se estará haciendo referencia a la estacionariedad débil de la Definición 1.6, pues esta es suficiente como hipótesis de los modelos que serán propuestos en el Capítulo 2.

Como consecuencia de la Definición 1.6, cuando se hace referencia a la función autocovarianza del proceso estacionario $\{Y_t\}$, puede pensarse en la función univaluada en el desfase h

$$\gamma_Y(h) := \gamma_Y(h, 0) = \gamma_Y(t+h, t). \tag{1.3}$$

Este hecho implica que la varianza de un proceso estocástico estacionario es constante ya que para todo $t = 0, \pm 1, \pm 2, \ldots$, sucede que

$$Var[Y_t] = \gamma_Y(t, t) = \gamma_Y(0) = const. \tag{1.4}$$

Los comentarios anteriores hacen ver que, de las tres series que se muestran en la Figura 1.4, la única que puede ser estacionaria es la que se encuentra más a la izquierda pues presenta media y varianza constantes.

Debe notarse, pues, que una serie temporal que presente tendencia o estacionalidad, no procede de un proceso estocástico estacionario. En general, una serie temporal asociada a un proceso estacionario no cumple patrones a largo

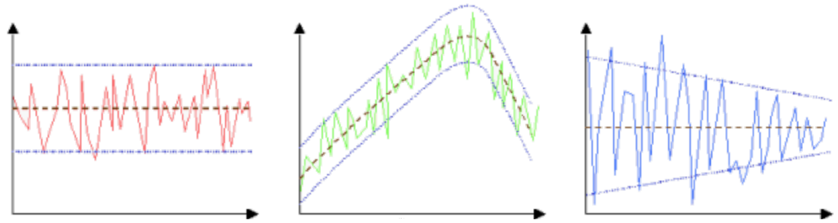

Figura 1.4. De izquierda a derecha: Series temporales con media y varianza constante, media no constante y varianza constante, y media constante y varianza no constante. Fuente: Palachy, S. [14].

plazo, aunque sí podrían permitirse ciertos comportamientos cíclicos que no tienen una duración determinada o fija. Se puede distinguir gráficamente en la Figura 1.3 que una serie temporal es estacionaria gráficamente cuando su forma es estrictamente horizontal y con varianza constante.

1.4. Funciones de autocorrelación y autocorrelación parcial

Definición 1.7. *Sea* $\{Y_t\}$ *un proceso estocástico estacionario. La* ***función de autocovarianza*** *(ACVF, del inglés AutoCoVariance Function) evaluada en el desfase* h *o, simplemente, autocovarianza de orden* h *es*

$$\gamma_Y(h) = Cov[Y_{t+h}, Y_t] = E[(Y_t - \mu)(Y_{t+h} - \mu)]. \tag{1.5}$$

Definición 1.8. *La* ***función de autocorrelación simple*** *(ACF, del inglés AutoCorrelation Function) de un proceso estocástico estacionario* $\{Y_t\}$ *evaluada en el desfase* h *o, simplemente, autocorrelación de orden* h *es*

$$\rho_Y(h) = \frac{\gamma_Y(h)}{\gamma_Y(0)}. \tag{1.6}$$

Definición 1.9. *La* ***función de autocorrelación parcial*** *(PACF, del inglés Partial AutoCorrelation Function) de un proceso estocástico estacionario* $\{Y_t\}$ *evaluada en el desfase* h *se denota* $\alpha_Y(h)$ *y está definida como el coeficiente* ϕ_{hh} *en la regresión*

$$\tilde{Y}_t = \phi_{h1}\tilde{Y}_{t-1} + \phi_{h2}\tilde{Y}_{t-2} + \cdots + \phi_{hh}\tilde{Y}_{t-h} + U_t \tag{1.7}$$

donde $\tilde{Y}_{t-i} = Y_{t-i} - \mu_Y$ $(i = 0, 1, 2, \dots)$ *y* U_t *no depende de* Y_{t-i} *para todo* $i = 0, 1, 2, \dots$.

Estas definiciones son dadas por Mauricio [12] (2007), quien señala que en la práctica, la media, la varianza, la ACF y la PACF no son computables, puesto que solo se cuenta con T observaciones del proceso estocástico. Es por ello que conviene definir sus equivalentes muestrales que sí pueden ser calculados a partir de una serie temporal.

Definición 1.10. *Sea* $\{y_t\}_{t=1}^T$ *una serie temporal que es una realización del proceso estocástico* $\{Y_t\}$. *La* ***media muestral*** *y* ***varianza muestral*** *de la serie son, respectivamente,*

$$\overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \quad y \quad S_y^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2. \tag{1.8}$$

Por su parte, la ***autocovarianza muestral*** *y* ***autocorrelación simple*** *de orden* h *se definen, respectivamente,*

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(y_{t+h} - \overline{y})(y_t - \overline{y}) \quad y \quad \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \tag{1.9}$$

Por último, la ***autocorrelación parcial muestral*** *de orden* h *es el estimador MCO o MV* [1] *de* ϕ_{hh} *del modelo de regresión lineal (1.7) considerado para* $\{y_t\}_{t=1}^T$.

Definición 1.11. *Un* ***gráfico ACF*** *es un diagrama de barras de la autocorrelación simple de la serie temporal en el que una barra situada en el desfase* h *mide* $\hat{\rho}(h)$. *El* ***gráfico PACF*** *se define de forma análoga, pero cada barra mide la autocorrelación parcial muestral de orden* h.

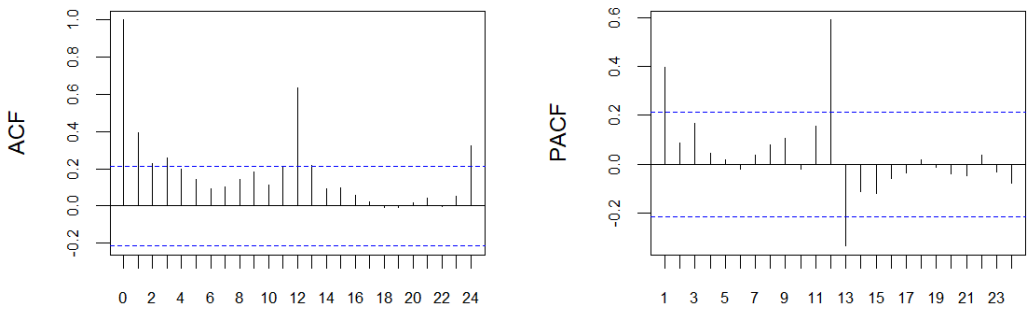

Figura 1.5. Gráficos ACF y PACF de la venta de *souvenir* en una tienda de Queensland, Australia, desde enero de 1987 hasta diciembre 1993. Datos de Rob Hyndman.

Estos gráficos contienen dos líneas discontinuas horizontales equiespaciadas del eje horizontal que representan un intervalo de confianza para cada autocorrelación muestral, usualmente del 95 %. Cuando una barra se sale fuera de estos límites se dice que su desfase asociado es significativo o que presenta una correlación significativa.

[1] MCO denota Mínimos Cuadrados Ordinarios y MV es Máxima Verosimilitud.

El gráfico ACF (también denominado **correlograma**) de una serie temporal es una herramienta visual sumamente útil para identificar si una serie temporal procede de un proceso estocástico estacionario, es decir, es una **serie temporal estacionaria**. En particular, cuando una serie temporal no es estacionaria, su correlograma decae muy lentamente, mientras que si la serie presenta estacionariedad, decaerá bruscamente.

Se comentaba al final de la Sección 1.3, que cuando una serie temporal presenta ciertas tendencias, su varianza no es constante debido a comportamientos estacionales o sucede una combinación de las dos anteriores, la serie no es estacionaria. Sin embargo, esto no debe ser un inconveniente puesto que una serie temporal no estacionaria puede ser transformada adecuadamente en una serie de aspecto estacionario. En la Sección 1.5 se presentan algunas transformaciones sencillas propuestas por Hyndman [8] (2021).

1.5. Diferenciación

Con el objetivo de eliminar la tendencia y la componente estacional de la serie para obtener residuos estacionarios, es necesario efectuar algunas transformaciones. Cuando una serie temporal se observa a lo largo de un periodo dilatado de tiempo, es común que la varianza del mismo se vea afectada por una tendencia. Por ejemplo, si la magnitud de las fluctuaciones parecen aumentar aproximadamente de forma lineal, la serie transformada $\{w_t = \log(y_t)\}_{t=1}^{T}$ logrará disminuir las fluctuaciones de forma que estas sean más constantes en magnitud y se estabilice la varianza en torno a la media de los datos. Otras transformaciones que consiguen estabilizar la varianza es la familia de transformaciones de **Box-Cox** que dependen de un parámetro λ y se define

$$y_t^{(\lambda)} = \begin{cases} \log(y_t) & \text{si } \lambda = 0, \\ \left\{ sign(y_t)|y_t|^{\lambda} - 1 \right\} / \lambda & \text{si } \lambda \neq 0, \end{cases} \tag{1.10}$$

donde

$$sign(y) = \begin{cases} -1 & \text{si } y < 0, \\ 0 & \text{si } y = 0, \\ 1 & \text{si } y > 0. \end{cases} \tag{1.11}$$

Por otro lado, la **diferenciación** consigue estabilizar la media de la serie temporal eliminando (o reduciendo) la tendencia. Con el objeto de introducir el concepto diferenciación, es necesario definir el **operador hacia atrás** o **de retardo** y el **operador diferenciación**.

Definición 1.12. *El **operador hacia atrás**, B (del inglés, Backshift), cuando opera con la variable aleatoria Y_t referida al momento t, tiene el efecto de desfasarla un periodo hacia atrás*

$$BY_t = Y_{t-1}. \tag{1.12}$$

Se puede generalizar el operador hacia atrás aplicándolo sucesivamente sobre la variable. Una simple comprobación prueba que

$$B^d Y_t = B \overset{(d)}{\cdots} BY_t = Y_{t-d}. \tag{1.13}$$

Definición 1.13. *El* ***operador diferenciación****,* ∇*, opera con la variable* Y_t *referida a un momento* t *calculando la diferencia con la referida a su momento inmediatamente anterior*

$$\nabla Y_t = Y_t - Y_{t-1}. \tag{1.14}$$

El operador hacia atrás y el operador diferenciación se encuentran estrechamente relacionados pues el segundo puede ser definido en función del primero. Simplemente, debe notarse que

$$\nabla Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1 - B)Y_t. \tag{1.15}$$

Definición 1.14. *El proceso de* ***diferenciación*** *de un proceso estocástico* $\{Y_t\}$ *consiste en considerar el proceso alternativo*

$$\{W_t = \nabla Y_t\}, \tag{1.16}$$

es decir, el nuevo proceso es la diferencia entre cada dos variables consecutivas del proceso original.

Las Definiciones 1.12-1.16 tienen sus equivalentes muestrales y son adaptadas a una serie temporal en lugar de a un proceso estocástico. En su caso, la serie diferenciada de una serie temporal compuesta por T observaciones estará formada por $T - 1$ datos al no ser posible calcular la diferencia de la primera observación con su inmediata anterior.

Tal y como se comentó al inicio de esta sección, la Figura 1.6 refleja cómo la diferenciación de una serie hace estable la media de la misma y reduce su tendencia, así como una transformación logarítmica estabiliza la varianza en torno a la media de los datos. Tan solo observando el gráfico temporal de la transformación logarítmica de la serie diferenciada uno podría aventurarse a decir que la serie parece estacionaria.

No debe perderse de vista que el objetivo de diferenciar una serie temporal es hacer que parezca estacionaria. No obstante, es posible que, ocasionalmente, la serie diferenciada no muestre estacionariedad. Ante esta situación es necesario diferenciarla de nuevo o, equivalentemente, aplicar una diferenciación de segundo orden

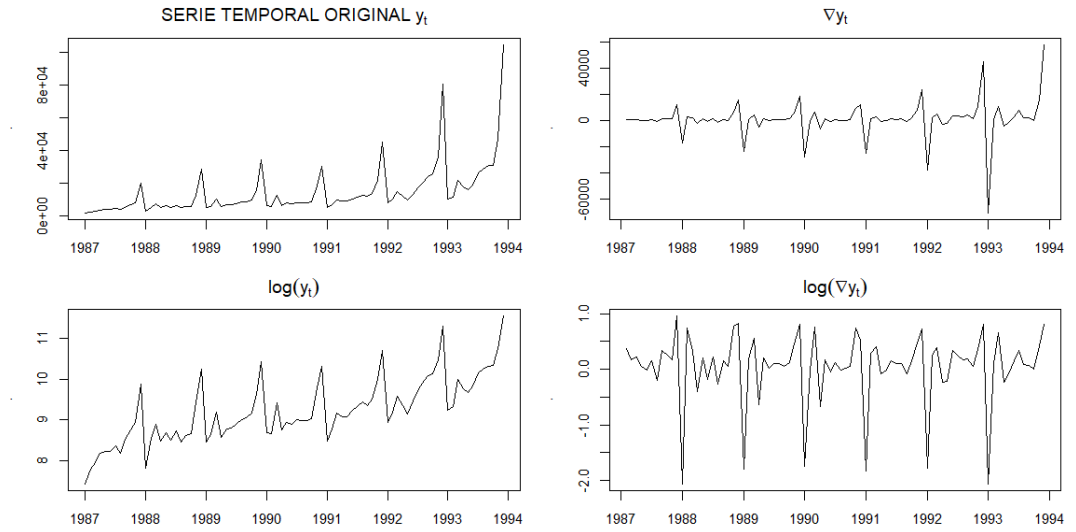

Figura 1.6. Se presenta la serie temporal de ventas de *Souvenir* original, acompañada de la serie temporal diferenciada, transformada por un logaritmo y la transformación logarítmica de la serie diferenciada. Datos de Rob Hyndman.

$$\begin{aligned}\nabla^2 y_t &= \nabla(y_t - y_{t-1}) \\ &= \nabla y_t - \nabla y_{t-1} \\ &= (y_t - y_{t-1}) - (y_{y-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \\ &= (1 - 2B + B^2)y_t \\ &= (1 - B)^2 y_t.\end{aligned} \tag{1.17}$$

Aunque casi nunca es necesario diferenciar la serie más que dos veces, se puede aplicar la **diferenciación de orden superior** d. Es importante apreciar una relación más general que la expuesta en (1.15) entre los dos operadores pues se demuestra que una diferenciación de orden d puede escribirse como

$$\nabla^d Y_t = (1 - B)^d Y_t. \tag{1.18}$$

Definición 1.15. *La* ***desestacionalización*** *o diferenciación estacional de un proceso estocástico* $\{Y_t\}$ *es considerar el proceso transformado de la forma*

$$\nabla_S Y_t = (1 - B^S)Y_t = Y_t - Y_{t-S}. \tag{1.19}$$

Una serie temporal desestacionalizada posee S observaciones menos que la original. En cualquier caso, se profundizará sobre la desestacionalización más adelante, cuando se introduzcan los modelos de series temporales con comportamiento estacional en la Sección 2.4.

En el gráfico temporal de la serie temporal de la Figura 1.2, puede apreciarse un cierto comportamiento estacional pues hay un patrón de ventas de que se repite cada doce meses. Por ello, parece pertinente practicar una desestacionalización de la serie, pero sin olvidar la diferenciación y transformación logarítmica que ya se habían aplicado con anterioridad. Los resultados de estas transformaciones se recogen el la Figura 1.7, donde la serie transformada tiene un comportamiento que encaja mejor en el aspecto estacionario que el de la Figura 1.6. Además, se aprecia que el gráfico ACF decrece rápidamente y esto se trata, como se ha dicho, de un indicativo de estacionariedad.

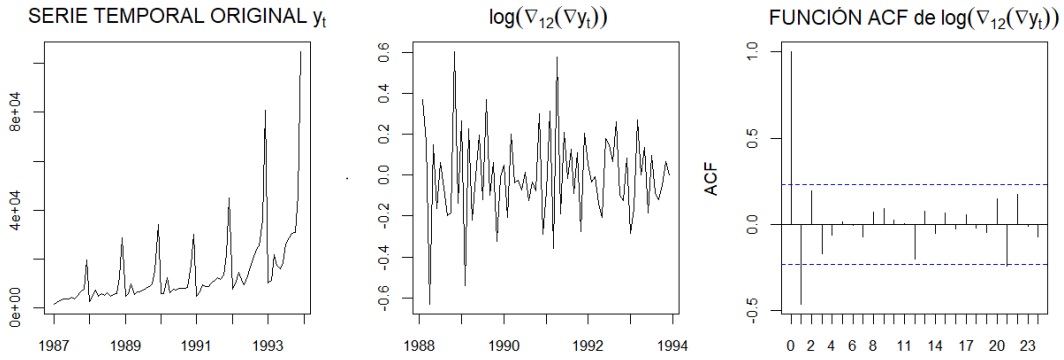

Figura 1.7. Se presenta la serie temporal de ventas de *Souvenir* original, acompañada de la serie temporal transformada y su gráfico ACF. Datos de Rob Hyndman.

1.6. Bondad de predicción

En el Capítulo 2 se introducirán algunos modelos de series temporales que permitirán hacer predicciones de valores futuros de la misma. Para estudiar la bondad de la predicción de un modelo, usualmente se utilizan las medidas MSE (del inglés, *Mean Squared Error*) y RMSE (del inglés, *Root Mean Squared Error*) que propone Calzone [3] (2022).

A pesar de que los modelos estarán pensados para hacer predicciones de valores futuros que son desconocidos, también puede calcularse la predicción del valor conocido de la serie, y_t en cualquier $t = 1, \ldots, T$ y denotarlo $\hat{y}_t$. Esta predicción recibe el nombre de **predicho**. Por tanto, el error que se comete al estimar la observación y_t por su predicho es $e_t = y_t - \hat{y}_t$ para $t = 1, \ldots, T$.

El MSE se define como la media de los errores al cuadrado debido a que, al poder ser los errores positivos o negativos, una media de los mismos tendría sumandos que se anularían unos con otros.

$$MSE = \frac{1}{T}\sum_{t=1}^{T} e_t^2. \tag{1.20}$$

El MSE presenta la limitación de que su unidad de medida es el cuadrado de la unidad de medida en los datos. Por ello, se considera el RMSE que no es más que la raíz cuadrada del MSE.

Otras medidas de bondad de predicción utilizadas son el MAE (del inglés, *Mean Absolute Error*)

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |e_t|, \tag{1.21}$$

y el MAPE (del inglés, Mean Absolute Percentage Error)

$$MAPE = \frac{100}{T}\sum_{t=1}^{T} \left|\frac{e_t}{y_t}\right|, \tag{1.22}$$

pero para evaluar esta última se debe ser cauteloso y no considerar los periodos cuya observaciones sean nulas, pues de lo contrario, se estaría dividiendo por cero.

Conviene comentar que estas medidas sirven para comparar cómo de bien se ajustan varios modelos a los datos de forma que cuanto menor sea el valor de las medidas de bondad de predicción, mejor será el modelo. Sin embargo, se verá en la Sección 1.7 que el hecho de que un modelo actúe bien sobre el conjunto de datos que se usa para estimarlo, no asegura que haga predicciones precisas.

1.7. Validación cruzada

Aunque en la Sección 1.6 se presentaron algunas medidas que sirven para evaluar la bondad de predicción, lo cierto es que, realmente, no estudian la calidad de la predicción, pues son funciones de los errores de datos que se han utilizado para ajustar el modelo. La calidad de una predicción solo puede determinarse midiendo cómo actúa el modelo en datos nuevos, no es los datos que se han utilizado para implementarlo.

Estas medidas expuestas son ampliamente utilizadas, pero carece de sentido medir la veracidad de la predicción en función de los errores que se cometerían si tomara como cada valor de la serie su predicho. Lo ideal, pues, sería contar con valores futuros de la serie que no se han tenido en cuenta para ajustar el

modelo y medir sus errores, tal y como sugiere Hyndman [8] (2021).

Debe tenerse en cuenta que, como uno de los objetivos del análisis de series temporales es predecir valores futuros, no suele contarse con datos más allá de T. Por ello, lo que suele hacerse es dividir el conjunto de datos en dos porciones: el conjunto de entrenamiento, que se utiliza para estimar los parámetros o ajustar el modelo; y el conjunto de validación, en el que se calculan los errores que cometería el modelo haciendo predicciones sobre este conjunto.

Normalmente se propone que el tamaño del conjunto de validación sea un 20 % del de la muestra total, pero esto es meramente orientativo y puede variar según las características de la serie temporal. En cualquier caso, el conjunto de validación debe ser, al menos, tan amplio como el horizonte en el que se pretende predecir la variable de estudio.

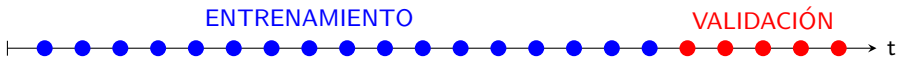

Figura 1.8. Se representa en azul las observaciones que forman parte del conjunto de entrenamiento y en rojo las que conforman conjunto de validación.

Una forma más sofisticada de validar el modelo es el método de la validación cruzada. El método se desarrolla en $T-m+1$ iteraciones. En la primera iteración, el conjunto de validación es la m-ésima observación de la serie temporal y el conjunto de entrenamiento son todas las observaciones que la preceden en el tiempo. A medida que el método avanza, el punto de validación de la iteración anterior pasa a formar parte del conjunto de entrenamiento, y el nuevo punto de validación es su inmediato posterior en el tiempo. Este procedimiento se ilustra en la Figura 1.9, en la que se muestran cuatro iteraciones de la validación cruzada.

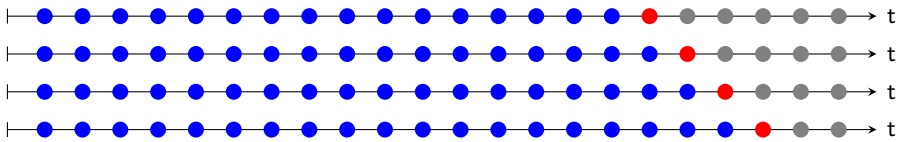

Figura 1.9. Cada punto representa una observación de la serie temporal, y cada fila una iteración en el método de validación cruzada. Las observaciones azules pertenecen al conjunto de entrenamiento, las rojas al de validación y las grises ni validan ni entrenan el modelo.

Como no es posible obtener predicciones fiables basadas en un conjunto de entrenamiento muy pequeño, suele considerarse un valor de m lo suficientemente

grande como para poder entrenar el modelo con $m-1$ datos.

La validación cruzada permite calificar la fiabilidad de predicción con las medidas expuestas en la Sección 1.6 evaluadas sobre los errores que se cometen en cada conjunto de validación o iteración.

Otra posibilidad es seguir considerando conjuntos unipuntuales de validación, pero dejando $k-1$ observaciones entre estos y sus respectivos conjuntos de entrenamientos como se muestra en la Figura 1.10. Este procedimiento recibe el nombre de validación cruzada de k pasos.

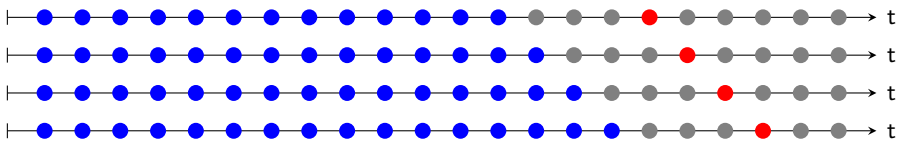

Figura 1.10. Cada punto representa una observación de la serie temporal, y cada fila una iteración en el método de validación cruzada de $k = 3$ pasos. Las observaciones azules pertenecen al conjunto de entrenamiento, las rojas al de validación y las grises ni validan ni entrenan el modelo.

2

Modelos ARIMA

En el Capítulo 1 se presentaron algunas nociones generales y conceptos claves en el análisis de series temporales. Se ha visto que el análisis preliminar de una serie temporal es efectuar su descomposición clásica para entender su comportamiento general y descubrir posibles tendencias o patrones estacionales que aportarían información valiosa para su análisis, pero que descartan de inmediato la estacionariedad de la misma. También, su gráfico temporal y su gráfico ACF serán de gran ayuda para decidir si la serie temporal es estacionaria y, en caso contrario, aplicar una transformación de Box-Cox que estabilice la varianza en torno a la media y tantas diferenciaciones como sean necesarias para convertir la media en constante.

Una vez concluido el análisis preliminar, llega el momento de elaborar un modelo estadístico que describa adecuadamente el proceso estocástico de procedencia de la serie temporal en cuestión, de forma que las implicaciones teóricas del modelo resulten compatibles con las características muestrales que presenta la serie. En este capítulo se define la familia de modelos $ARIMA$ (del inglés, *Autoregressive Integrated Moving Average*), así como algunas técnicas o criterios para decidir qué modelos de esta familia son más apropiados para modelizar una serie temporal dada.

El paso final en el análisis de una serie temporal es la diagnosis del modelo construido que comprueba el cumplimiento de las hipótesis que asume el mismo.

2.1. Ejemplos de modelos

Antes de citar algunos ejemplos clásicos de procesos estocásticos basados en la familia de modelos $ARIMA$, conviene definir el concepto modelo, como lo entiende Mauricio [12] (2007).

Definición 2.1. *Un **modelo** para un proceso estocástico es cualquier conjunto de hipótesis bien definidas sobre las propiedades estadísticas de dicho proceso que usualmente se expresa mediante una ecuación de recurrencia.*

Cuando se nombran las propiedades estadísticas de un proceso estocástico, frecuentemente se hace referencia a ciertas condiciones que debe cumplir la esperanza o covarianza del proceso. En particular, los modelos $ARMA$ asumen que el proceso sea estacionario, por ello se cumplen las propiedades estadísticas de la Definición 1.6.

Ejemplo 2.2. **Ruido blanco**. Un proceso estocástico es ruido blanco y se denota $\{Y_t\} \sim N(0, \sigma^2)$ (del inglés, *Noise*) si se trata de una secuencia de variables aleatorias idéntica e independientemente distribuidas tal que

- $E[Y_t] = 0$ para todo $t = 0, \pm 1, \pm 2, \ldots$
- $Var[Y_t] = \sigma^2$ para todo $t = 0, \pm 1, \pm 2, \ldots$
- $Cov[Y_t, Y_{t+h}] = 0$ para todo $t = 0, \pm 1, \pm 2, \ldots$ y cualquier $h \neq 0$.

Nótese que la segunda condición hace referencia a que la varianza permanece constante en el tiempo, y que las otras dos aseguran la (débil) estacionariedad del proceso.

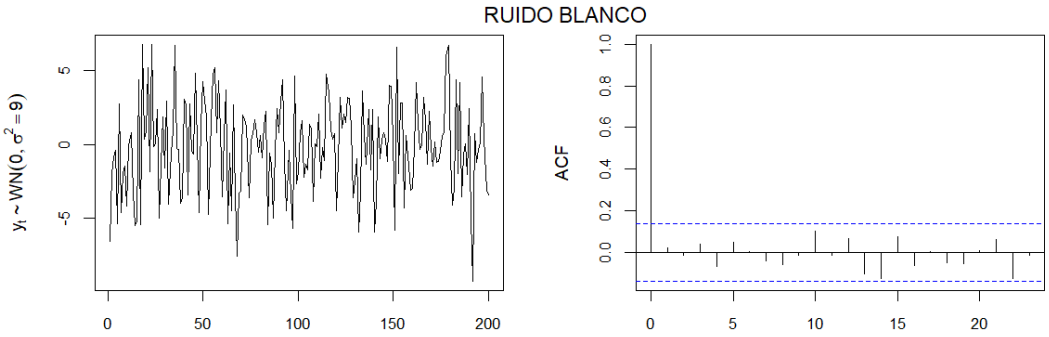

Figura 2.1. Gráfico temporal de ruido blanco $N(0, \sigma^2 = 9)$ simulado en R, acompañado de su gráfico ACF en el que se puede apreciar la estacionariedad del proceso.

No es fortuito que el ruido blanco sea el primer ejemplo introducido en este capítulo ya que será un proceso protagonista en los modelos ARIMA.

Ejemplo 2.3. **Paseo aleatorio**. Un proceso estocástico no estacionario $\{Y_t\}$ es un paseo aleatorio cuando

$$Y_t = \mu + Y_{t-1} + \varepsilon_t, \tag{2.1}$$

para todo $t = 0, \pm 1, \pm 2, \ldots$ siendo μ un parámetro y $\{\varepsilon_t\} \sim N(0, \sigma^2)$. El interés del paseo aleatorio radica en que si se considera su proceso estocástico diferenciado, entonces se obtiene ruido blanco.

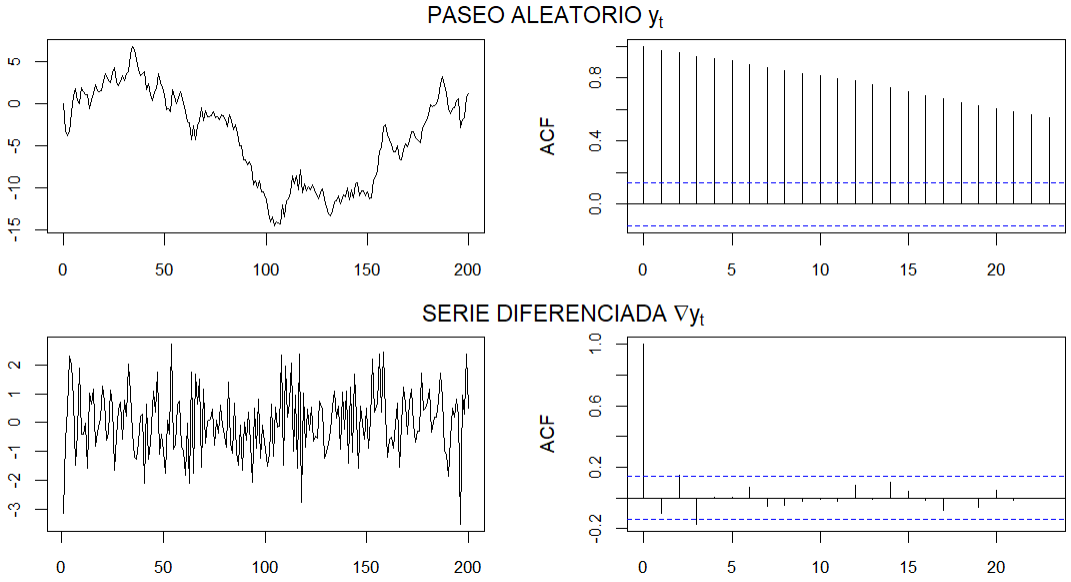

Figura 2.2. Gráfico temporal de de un paseo aleatorio simulado en R, acompañado de su gráfico ACF. Los gráficos de la segunda fila corresponden al paseo aleatorio diferenciado en los que se aprecia que, en efecto, la serie diferenciada luce como un ruido blanco.

2.2. Modelos ARMA

En en análisis de series temporales, el modelo de medias móviles es un proceso común para modelizar datos univariables que establece una relación lineal entre la variable de estudio y los errores de predicciones pasadas.

Definición 2.4. *Se dice que un proceso estocástico* $\{Y_t\}$ *es un proceso* $MA(q)$ ***media móvil*** *de orden* q *si*

$$Y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \tag{2.2}$$

para todo $t = 0, \pm1, \pm2, \ldots$ *y* $\{\varepsilon_t\} \sim N(0, \sigma^2)$.

Ejemplo 2.5. **Modelo media móvil de orden 1**. Cuando un proceso estocástico sigue un modelo $MA(1)$, se escribe

$$Y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1}, \tag{2.3}$$

con $\{\varepsilon_t\} \sim N(0, \sigma^2)$.

Su media puede expresarse

$$\begin{aligned} E[Y_t] &= E[c + \varepsilon_t + \theta_1\varepsilon_{t-1}] \\ &= E[c] + E[\varepsilon_t] + \theta_1 E[\varepsilon_{t-1}] \\ &= E[c] = c \end{aligned} \tag{2.4}$$

teniendo en cuenta la linealidad de la esperanza matemática y que la esperanza de una constante es la propia constante. Por su parte, la varianza del proceso puede describirse como

$$\begin{aligned} Var[Y_t] &= Var[c + \varepsilon_t + \theta_1 \varepsilon_{t-1}] \\ &= Var[c] + Var[\varepsilon_t] + \theta_1^2 Var[\varepsilon_{t-1}] \\ &= \sigma^2(1 + \theta_1^2) \end{aligned} \tag{2.5}$$

en virtud de la independencia del ruido blanco. En cuanto a la autocovarianza de orden h, se obtiene la expresión

$$\begin{aligned} \gamma_Y(h) &= E[(Y_t - \mu)(Y_{t-h} - \mu)] \\ &= E[(\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_{t-h} + \theta_1 \varepsilon_{t-h-1})] \\ &= E[\varepsilon_t \varepsilon_{t-h} + \theta_1 \varepsilon_{t-1} \varepsilon_{t-h} + \theta_1 \varepsilon_t \varepsilon_{t-h-1} + \theta_1^2 \varepsilon_{t-1} \varepsilon_{t-h-1}] \end{aligned} \tag{2.6}$$

Ahora bien, si se toma $h = 1$, entonces sucede que

$$\gamma_Y(1) = E[\varepsilon_t \varepsilon_{t-1}] + \theta_1 E[\varepsilon_{t-1} \varepsilon_{t-1}] + \theta_1 E[\varepsilon_t \varepsilon_{t-2}] + \theta_1^2 E[\varepsilon_{t-1} \varepsilon_{t-2}] = \theta_1 \sigma^2, \tag{2.7}$$

mientras que si $h > 1$, la autocovarianza es $\gamma_Y(h) = 0$. De esto se deduce que la autocorrelación del proceso es

$$\rho_Y(h) = \begin{cases} \theta_1/(1 + \theta_1^2) & \text{si } h = 1, \\ 0 & \text{si } h > 1. \end{cases} \tag{2.8}$$

Se ha probado que un proceso $MA(1)$ es débilmente estacionario. Se demuestra que un modelo media móvil $MA(q)$ es estacionario para cualquier valor de q.

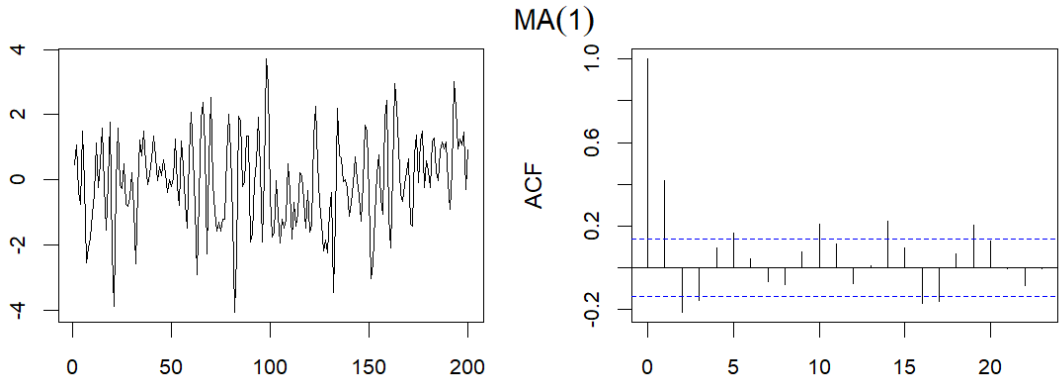

Figura 2.3. Gráfico temporal de de un proceso $MA(1)$ simulado con el lenguaje R para $\theta_1 = 0.9$, acompañado de su gráfico ACF.

Debe apreciarse en la Figura 2.3 el comportamiento estacionario de la serie. Además, aunque la autocorrelación calculada en (2.8) es teórica y, evidentemente

es ligeramente distinta a la muestral, se observa en el gráfico ACF que $\hat{\rho}_Y(1) \approx 0.9/(1+0.9^2) = 0.497$ y que la autocorrelación es muy próxima a cero para el resto de desfases.

Los modelos autoregresivos, por su parte, tratan de describir la variable de interés referida al momento t como una combinación lineal de los valores pasados de esa misma variable. El término autoregresión hace referencia a que el modelo es una regresión lineal de la variable de estudio sobre sí misma.

Definición 2.6. *Un proceso estocástico* $\{Y_t\}$ *describe un modelo* $AR(p)$ ***autoregresivo*** *de orden* p *si*

$$Y_t = c + \phi_1 Y_{t-1} + \phi Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \tag{2.9}$$

para todo $t = 0, \pm 1, \pm 2, \ldots$ *y* $\{\varepsilon_t\} \sim N(0, \sigma^2)$.

Ejemplo 2.7. **Modelo autorregresivo de orden 1**. Un proceso estocástico sigue un modelo $AR(1)$ si

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t, \tag{2.10}$$

con $\{\varepsilon_t\} \sim N(0, \sigma^2)$.

Si se asume que el proceso es débilmente estacionario, sucede que $E[Y_t] = \mu$ para todo $t = 0, \pm 1, \pm 2 \ldots$ de forma que

$$\begin{aligned} E[Y_t] &= E[c + \phi_1 Y_{t-1} + \varepsilon_t] \\ &= E[c] + \phi_1 E[Y_{t-1}] + E[\varepsilon_t] \\ &= c + \phi_1 \mu, \end{aligned} \tag{2.11}$$

y, al despejar μ de (2.11), se concluye que

$$E[Y_t] = \frac{c}{1 - \phi_1}. \tag{2.12}$$

En virtud de la independencia entre los errores y el proceso estocástico, la varianza cumple la relación

$$\begin{aligned} Var[Y_t] &= Var[c + \phi_1 Y_{t-1} + \varepsilon_t] \\ &= Var[c] + \phi_1^2 Var[Y_{t-1}] + Var[\varepsilon_t] \\ &= \phi_1^2 Var[Y_{t-1}] + \sigma^2. \end{aligned} \tag{2.13}$$

Resulta realmente interesarte cómo la restricción $|\phi_1| < 1$ es una condición necesaria para asegurar la estacionariedad del proceso que se había asumido al comienzo de este ejemplo. En efecto, si el proceso es estacionario, entones $Var[Y_t] = Var[Y_{t-1}]$ para todo $t = 0, \pm 1, \pm 2, \ldots$ y de (2.13) se desprende que

$$Var[Y_t](1-\phi_1^2)=\sigma^2. \tag{2.14}$$

Debido a que tanto la varianza del proceso como la del ruido blanco son positivas, es necesario que $1-\phi_1^2>0$ o, equivalentemente, $|\phi_1|<1$ pues, de lo contrario, la igualdad (2.14) resultaría en un absurdo.

Para dar la expresión de la autocovarianza de orden h, se multiplica (2.10) por Y_{t-h} y se toman esperanzas a ambos lados del igual

$$E[Y_{t-h}Y_t]=\phi_1 E[Y_{t-h}Y_{t-1}]+E[Y_{t-h}\varepsilon_t], \tag{2.15}$$

de donde sigue la fórmula recursiva

$$\gamma_Y(h)=\phi_1\gamma_Y(h-1),\ \gamma_Y(1)=\phi_1\gamma_Y(0). \tag{2.16}$$

Todo esto permite afirmar que

$$\gamma_Y(h)=\phi_1^h\gamma_Y(0) \tag{2.17}$$

y, consecuentemente, la autocorrelación de orden h es

$$\rho_Y(h)=\frac{\gamma_Y(h)}{\gamma_Y(0)}=\frac{\phi_1^h\gamma_Y(0)}{\gamma_Y(0)}=\phi_1^h. \tag{2.18}$$

Tanto en el gráfico temporal como en el gráfico ACF de la Figura 2.4 se aprecia que la serie temporal es estacionaria, comportamiento posibilitado por el hecho de que $-1<\phi_1=0.5<1$, que también asegura que la función autocorrelación sea decreciente tal y como se ha definido en (2.18).

Definición 2.8. *Un proceso estocástico* $\{Y_t\}$ *es un proceso* $ARMA(p,q)$ *si es débilmente estacionario y para cualquier* $t=0,\pm1,\pm2,\ldots$

$$Y_t-\phi_1Y_{t-1}-\cdots-\phi_pY_{t-p}=c+\varepsilon_t+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q}, \tag{2.19}$$

donde

- $\{\varepsilon_t\}\sim N(0,\sigma^2)$ *es ruido blanco,*
- $\phi(z)=1-\phi_1z-\cdots-\phi_pz^p$ *es el polinomio autorregresivo,*
- $\theta(z)=1+\theta_1z+\cdots+\theta_qz^q$ *es el polinomio media móvil,*

y los polinomios autorregresivo y media móvil no tienen factores en común.

Es conveniente escribir este proceso con una notación más concisa haciendo uso del operador hacia atrás:

$$\phi(B)Y_t=c+\theta(B)\varepsilon_t. \tag{2.20}$$

Si los polinomios de la Definición 2.8 tuvieran raíces en común, entonces el modelo estaría sobreparametrizado y sería recomendable modelar el proceso con un $ARMA(p-1,q-1)$.

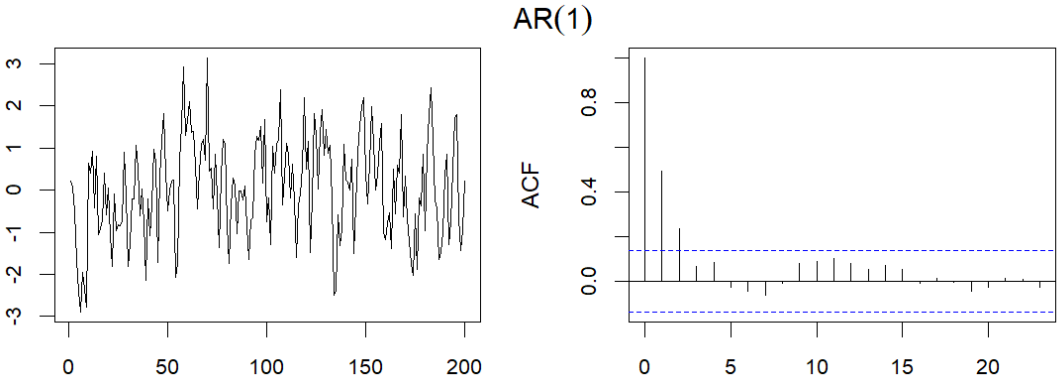

Figura 2.4. Gráfico temporal de de un proceso $AR(1)$ simulado con el lenguaje R para $\phi_1 = 0.5$, acompañado de su gráfico ACF.

2.2.1. Estacionariedad e Invertibilidad

Aunque un proceso media móvil de orden q es siempre estacionario sin requerir ningún tipo de condición sobre los coeficientes de su polinomio media móvil, no ocurre lo mismo con los procesos autorregresivos de orden p o con los procesos $ARMA(p,q)$. En efecto, el Ejemplo 2.7 ilustró esta situación probando que la condición $|\phi_1| < 1$ en un proceso $AR(1)$ es necesaria para que este sea estacionario. Por ello, a continuación, se introducen dos definiciones con sus respectivas caracterizaciones aportadas por Brockwell [2] (2002).

Definición 2.9. *Sea $\{Y_t\}$ un proceso $ARMA(p,q)$. Se dice que $\{Y_t\}$ es **estacionario** si existen las constantes $\{\psi_j\}$ tales que $\sum_{j=0}^{\infty} |\psi_j| < \infty$ y para cualquier t se cumple*

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

La estacionariedad del proceso es equivalente a la condición $\phi(z) \neq 0$ para cualquier $|z| \leq 1$. Se demuestra que los coeficientes $\{\psi_j\}$ están determinados por el desarrollo en series de potencias

$$\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \frac{\theta(z)}{\phi(z)}, \ |z| \leq 0. \tag{2.21}$$

Ejemplo 2.10. Sea $\{Y_t\}$ un proceso $ARMA(1,0)$ estacionario. La condición de invertibilidad exige que $|\phi_1| < 1$ y efectuando sustituciones reiteradamente en la expresión del modelo

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \varepsilon_t \\ &= \phi_1(\phi_1 Y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 Y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^3 Y_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &\vdots \\ &= \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}, \end{aligned} \tag{2.22}$$

se llega a un proceso $MA(\infty)$. Nótese la importancia de la condición de estacionariedad. Si el coeficiente $|\phi_1| > 1$, entonces el valor ϕ_1^k aumenta a medida que lo hace k y los errores menos recientes tienen una ponderación muy alta en el periodo actual, lo que carece de sentido.

Un proceso $ARMA(p, q)$ es, por tanto, estacionario si puede ser expresado en términos de su error actual y errores pasados en forma de un $MA(\infty)$. Un concepto fuertemente relacionado con la estacionariedad es la invertibilidad.

Definición 2.11. *Sea $\{Y_t\}$ un proceso $ARMA(p, q)$. Se dice que $\{Y_t\}$ es* ***invertible*** *si existen las constantes $\{\pi_j\}$ tales que $\sum_{j=0}^{\infty} |\pi_j| < \infty$ y para cualquier t se cumple*

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}.$$

La invertibilidad del proceso es equivalente a la condición $\theta(z) \neq 0$ para todo $|z| \leq 1$. Se demuestra que los coeficientes $\{\pi_j\}$ están determinados por el desarrollo en series de potencias

$$\sum_{j=0}^{\infty} \pi_j Y_{t-j} = \frac{\phi(z)}{\theta(z)}, \ |z| \leq 0. \tag{2.23}$$

Ejemplo 2.12. Sea $\{Y_t\}$ un proceso $ARMA(0, 1)$ invertible. La condición de invertibilidad exige que $|\theta_1| < 1$ y sustituyendo reiteradamente en la expresión del modelo

$$\begin{aligned} \varepsilon_t &= Y_t - \theta_1 \varepsilon_{t-1} \\ &= Y_t - \theta_1(Y_{t-1} - \theta_1 \varepsilon_{t-2}) \\ &= Y_t - \theta_1 Y_{t-1} + \theta_1^2 \varepsilon_{t-2} \\ &= Y_t - \theta_1 Y_{t-1} + \theta_1^2 Y_{t-2} - \theta_1^3 \varepsilon_{t-3} \\ &\vdots \\ &= \sum_{j=0}^{\infty} (-\theta_1)^j Y_{t-j}, \end{aligned} \tag{2.24}$$

se obtiene un proceso $AR(\infty)$. Nuevamente, es necesario que $|\theta_1| < 1$ puesto que, de lo contrario, las observaciones más antiguas tendrían una mayor influencia en el error actual.

2.3. Modelos ARIMA

Podría ocurrir que la serie temporal analizada no proceda de un proceso estocástico estacionario y, por tanto, no satisfaga las hipótesis de los modelos expuestos en la Sección 2.2. Si la serie temporal presenta estacionariedad, entonces se trata de ajustar un modelo $ARMA$. De lo contrario, se busca una serie transformada estacionaria que, frecuentemente, se consigue diferenciando la serie original. Nace de aquí la necesidad de definir la clase de modelos $ARIMA$, que no es más que una generalización de los modelos $ARMA$ que considera series no estacionarias.

Definición 2.13. *Sea d un entero no negativo. Se dice que $\{Y_t\}$ es un proceso $ARIMA(p,d,q)$ si la serie $X_t = (1-B)^d Y_t$ es un proceso $ARMA(p,q)$ estacionario.*

Esta definición implica que la serie $\{Y_t\}$ satisface la relación

$$\phi(B)(1-B)^d Y_t = c + \theta(B)\varepsilon_t$$

donde $\phi(z)$ y $\theta(z)$ son los polinomios autorregresivo y media móvil de grado p y q, respectivamente, tales que no tienen raíces en común y $\{\varepsilon_t\} \sim N(0, \sigma^2)$. Además, debe notarse que el proceso es estacionario si y solo si $d = 0$, en cuyo caso se trataría de un modelo $ARMA(p,q)$.

2.4. Modelos ARIMA estacionales

Se llama estacionalidad a un patrón que se aprecia en una serie temporal y se repite regularmente cada S periodos de tiempo. Por ejemplo, existe estacionalidad en los conjuntos de datos recogidos mensualmente en los que los valores altos tienden a ocurrir siempre en un mes en particular. En este caso, $S = 12$ definiría el comportamiento estacional.

Como se expuso en la Sección 1.5, la diferenciación es un proceso que sirve para eliminar la tendencia de una serie, mientras que la diferenciación estacional o desestacionalización se usa para reducir su comportamiento estacional. Cuando en una serie temporal están presentes tanto la tendencia como la estacionalidad, es posible que sea necesario efectuar una diferencia no estacional seguida de una diferencia estacional. De estas circunstancias surge una clase de

modelos $ARIMA$ que considera los comportamientos estacionales de un proceso estocástico y mejora notablemente la precisión de las predicciones: la familia de modelos $SARIMA$ o $ARIMA$ estacionales.

Definición 2.14. *Si d y D son enteros no negativos, entonces se dice que $\{Y_t\}$ es un modelo $ARIMA(p,d,q) \times (P,D,Q)_S$ estacional de periodo S si la serie $X_t = (1-B)^d(1-B^s)^D Y_t$ es un proceso $ARMA$ estacionario definido por*

$$\phi(B)\Phi(B^s)X_t = c + \theta(B)\Theta(B^s)\varepsilon_t, \tag{2.25}$$

donde

- $\{\varepsilon_t\} \sim N(0,\sigma^2)$ *es ruido blanco,*
- $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ *es el polinomio autorregresivo,*
- $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ *es el polinomio media móvil,*
- $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$ *es el polinomio autorregresivo estacional,*
- $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$ *es el polinomio media móvil estacional.*

2.5. Selección de órdenes de los modelos ARMA, ARIMA y ARIMA estacional

Una vez se ha transformado la serie original combinando la diferenciación y las transformaciones de Box-Cox que resultan en una nueva serie transformada estacionaria, se trata de ajustar los datos a un modelo $ARMA$. Surge entonces la cuestión de seleccionar adecuadamente los órdenes p y q, así como de elegir el modelo que mejor se ajuste a las características de la serie. Para cumplir esta tarea exitosamente, suelen consultarse los gráficos ACF y PACF siguiendo varios criterios de decisión según el comportamiento de los mismos.

A la hora de decidir qué modelo $AR(p)$, $MA(q)$ o $ARMA(p,q)$ se ajusta mejor al conjunto de datos, así como la selección de sus órdenes, se tienen en cuenta los comentarios de Nau [13] (2020) que se exponen a continuación.

1. Si el gráfico PACF de la serie diferenciada muestra un decaimiento brusco y/o su primera barra es positiva, sucede que la serie parece estar subdiferenciada y debe considerarse un modelo en el que intervenga la parte autorregresiva. El valor del último desfase significativo será tomado como orden p del término autorregresivo.

2. Si el gráfico ACF de la serie diferenciada posee un decaimiento brusco y/o su primera barra es negativa, se aprecia que la serie está sobrediferenciada y sería conveniente considerar un modelo en el que intervenga la parte media

móvil. El valor del último desfase significativo será considerado como orden q del término media móvil.

	ACF	**PACF**
$AR(p)$	Decrecimiento rápido de tipo geométrico puro, y geométrico con alternancia de signos, sinusoidal o mezcla de varios tipos.	Se anula para retardos superiores a p.
$MA(q)$	Se anula para retardos superiores a q.	Decrecimiento rápido de tipo exponencial y/o sinusoidal.
$ARMA(p,q)$	Los primeros valores iniciales no tienen patrón fijo y van seguidos de una mezcla de oscilaciones sinusoidales y/o exponenciales amortiguadas.	Los primeros valores iniciales no tienen patrón fijo y van seguidos de una mezcla de oscilaciones sinusoidales y/o exponenciales amortiguadas.

Tabla 2.1. Comportamiento de las funciones de autocorrelación y autocorrelación parcial en los procesos $AR(p)$, $MA(q)$ y $ARMA(p,q)$ que ayudan a identificar el proceso del que procede una serie temporal. Fuente: Uriel, E. [10].

2.5.1. Criterios de información de Akaike y Bayesiano

Es posible que la inspección de los gráficos dé lugar a distintas interpretaciones y se puedan proponer distintos órdenes del modelo $ARMA(p,q)$. Ante esta situación Hyndman [9] (2008) sugiere tomar diferentes combinaciones de p y q, y comprobar qué combinación aporta mejor Criterio de Información de Akaike (AIC) o Criterio de Información Bayesiano (BIC), que asumen que los errores del modelo siguen una distribución normal.

El Criterio de Información de Akaike evalúa cómo de bien se ajusta el modelo a los datos sin tener en cuenta demasiada información. Intuitivamente, el criterio mide la bondad del ajuste y penaliza que el modelo considere muchos datos antiguos sin ser necesarios. Este criterio es útil para comparar dos modelos de forma que el que menor AIC presente, mejor balance entre ajuste y complejidad tendrá. Se define

$$AIC = -2 \cdot \frac{\ell}{T} + 2 \cdot \frac{k}{T}, \tag{2.26}$$

siendo k el número de parámetros que se estiman, y

$$\ell = -\frac{T}{2}\left(1 + \log(2\pi) + \log\left(MSE\right)\right). \tag{2.27}$$

Por otro lado, el Criterio de Información Bayesiano está fuertemente relacionado con el AIC, pero en él, a diferencia del anterior, interviene el número

de observaciones que posee la serie. De la misma manera, un menor BIC indica una selección de órdenes más apropiada. Su expresión viene dada por

$$BIC = -2 \cdot \frac{\ell}{T} + k \cdot \frac{\log(T)}{T}. \tag{2.28}$$

Si la serie temporal no es estacionaria, el parámetro d de los modelos $ARIMA$ representa el número de diferenciaciones no estacionales que requiere la serie para ser estacionaria.

Si la serie presenta cierta estacionalidad, entonces debe implementarse un modelo $ARIMA$ estacional o $SARIMA$ que recoja los comportamientos estacionales de la misma. Estos comportamientos pueden ser descubiertos en los gráficos temporales, pero también se manifiestan en los gráficos ACF y PACF en los que se aprecia un patrón que se repite cada S periodos de tiempo.

Para determinar los órdenes de los términos autorregresivo y media móvil estacionales, se examina el gráfico PACF y el gráfico ACF de la de la serie, respectivamente, prestando atención tan solo a las barras que se encuentran en los múltiplos de S. Sin embargo, esta tarea puede ser complicada y, por ello, se suele dejar esta labor a una máquina que compara varios modelos decidiéndose según los AIC y BIC.

2.6. Raíces unitarias

Los test de raíces unitarias están orientados a indicar si la serie necesita una diferenciación adicional para ser estacionaria. Ya se ha explicado que la existencia de tendencia en un proceso estocástico es incompatible con que el mismo sea estacionario. Por ello, el contraste de hipótesis de Dickey-Fuller permite saber si hay presencia significativa de tendencia en cierta serie temporal y descartar su estacionariedad.

El contraste asume como hipótesis nula que una serie temporal no es estacionaria y como hipótesis alternativa que es estacionaria. Por ello, según lo expuesto en este trabajo, interesaría rechazar la hipótesis nula. En el Ejemplo 2.7 se expone que si el primer coeficiente de un modelo $AR(1)$ es en módulo uno, entonces el proceso no es estacionario y existe cierta tendencia. Por ello, una forma de descartar la estacionariedad de una serie es en función del valor del primer regresor de la autoregresión.

En particular, siguiendo la notación de la Definición 2.6, si $\phi_1 = 1$ entonces la serie no es estacionaria, pero si $\phi_1 < 1$, la serie será estacionaria. Se considera el modelo $AR(1)$

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t, \tag{2.29}$$

se resta Y_{t-1} a ambos lados de la igualdad y se denota $\delta = 1 - \phi_1$

$$\nabla Y_t = c + \delta Y_{t-1} + \varepsilon_t. \tag{2.30}$$

Ahora, la hipótesis nula de que la serie no es estacionaria ($\phi_1 = 1$) es equivalente a que $\delta = 0$, y la hipótesis alternativa que $\delta < 0$. En cualquier caso, la forma más simple de describir el contraste es

$$\begin{cases} H_0 : \text{La serie temporal no es estacionaria.} \\ H_1 : \text{La serie temporal es estacionaria.} \end{cases} \tag{2.31}$$

Ahora bien, el test de Dickey-Fuller tan solo analiza la estacionariedad de un proceso $AR(1)$, pero puede generalizarse para un proceso $ARMA(p, q)$, en cuyo caso recibe el nombre de contraste de Dickey-Fuller Aumentado, que es algo más complejo de construir, pero se basa en las ideas del no aumentado y se formula como en (2.31).

Yaffee [17] (2000) señala la importancia de que la serie temporal cumpla las hipótesis del contraste: que los residuos sean ruido blanco y que el proceso diferenciado sea invertible.

2.7. Diagnosis del modelo

El objetivo al construir un modelo $ARIMA$ es encontrar un modelo que sea lo más adecuado posible para representar el comportamiento general del proceso estocástico del que proviene la serie temporal analizada. Un modelo ideal, según Uriel [10] (1995), sería el que cumple las siguientes propiedades:

1. Los residuos del modelo estimado se aproximan al comportamiento de un ruido blanco.
2. El modelo estimado es estacionario e invertible.

Es esencial que los residuos del modelo estimado se comporten como un ruido blanco pues es la hipótesis fundamental de la familia de modelos $ARIMA$ y, en caso contrario estos contendrían información relevante para la predicción.

Se demuestra que la sucesión de autocorrelaciones muestrales procedentes de un proceso de ruido blanco siguen, en muestras grandes, una distribución normal $N(0, 1/T)$. Por ello, cabe esperar que, si un proceso es ruido blanco, entonces el 95 % de las barras de su gráfico ACF tomen valores en $\pm 1.96/\sqrt{T}$. Entonces, examinar el gráfico ACF es el primer paso que debe darse para estudiar si los residuos son ruido blanco.

Una vez analizado el gráfico ACF de la serie de los residuos, se observa su histograma y gráfico cuantil-cuantil para estudiar de forma estimada su supuesta normalidad. A continuación, es usual utilizar el contraste de Jarque-Bera que resulta eficiente para muestras grandes y mide si el conjunto de datos posee una asimetría y curtosis compatibles con la distribución normal. La hipótesis nula asume la normalidad de los residuos, mientras que la hipótesis alternativa es que no siguen una distribución normal. Su estadístico viene definido por

$$JB = \frac{T}{6}\left(SK^2 + \frac{1}{4}(KU-3)^2\right), \tag{2.32}$$

donde SK y KU son los coeficientes muestrales de asimetría y curtosis, respectivamente. El estadístico se distribuye, bajo la hipótesis nula de normalidad del ruido blanco que como una χ^2 con dos grados de libertad.

En la práctica, además de examinar su gráfico ACF y gráfico temporal, los residuos suelen analizarse mediante el contraste de Ljung-Box (2.33) que estudia si las autocorrelaciones simples del modelo son todas iguales a cero hasta cierto desfase M, lo que sería un indicador de la independencia de los residuos.

$$\begin{cases} H_0 : \rho_Y(1) = \rho_Y(2) = \cdots = \rho_Y(M) = 0 \\ H_1 : c.c. \end{cases} \tag{2.33}$$

El valor de M, es decir, el número de coeficientes de autocorrelacion que se incluyen en el contraste es arbitrario. Si se toma M muy grande, se tiene la ventaja de que se pueden captar valores significativos correspondientes a desfases elevados, pero a cambio, a medida que aumenta M disminuye la potencia del contraste.

El estadístico Q del contraste es

$$Q_{LB} = T(T+2)\sum_{k=1}^{M}\frac{\hat{\rho}_Y(k)^2}{T-k}, \tag{2.34}$$

que se distribuye como una χ^2 con $M-p-q$ grados de libertad bajo la hipótesis nula de que los residuos son independientes, donde p y q son el orden de los términos autorregresivo y media móvil, respectivamente.

Una cuestión importante en el análisis de los coeficientes estimados es el examen de si se cumplen las condiciones de estacionariedad y de invertibilidad. Esto se hace factorizando el polinomio autorregresivo y media móvil prestando atención a sus raíces según las caracterizaciones dadas en la Subsección 2.2.1. Si el polinomio autorregresivo tiene raíces en el círculo unidad, es un indicio de que el modelo no es estacionario y es aconsejable tomar una diferenciación

adicional. Si el polinomio media móvil tiene alguna raíz en el círculo unidad, sería indicativo de que el modelo está sobrediferenciado. Cuando se propone un modelo $SARIMA$ la invertibilidad y estacionariedad se estudian comprobando si los polinomios $\theta(z)\Theta(z)$ y $\phi(z)\Phi(z)$ tienen raíces en el círculo unidad, respectivamente.

3

Análisis de un conjunto de datos real

En Capítulo 2 se ha expuesto la familia de modelos $ARMA$ que asume la estacionariedad de la serie temporal a la que se pretenden ajustar. Si una serie no es estacionaria, no debe perderse la calma, pues las transformaciones del Capítulo 1 pueden conseguir el comportamiento estacionario de la misma. Otra posibilidad es considerar la familia de modelos $ARIMA$ que no asumen estacionariedad sobre los datos, pero sí sobre la serie diferenciada d veces. Si, además, la serie presenta un comportamiento estacional, será pertinente practicar una desestacionalización antes de tratar ajustar un modelo $ARIMA$ o, en su defecto, considerar la familia de modelos $SARIMA$. En cualquier caso, siempre debe realizarse la diagnosis del modelo.

Este capítulo se enfoca como la aplicación práctica de todos los conceptos expuestos hasta el momento sobre un conjunto de datos reales. La serie temporal que se estudiará ha sido cedida por el Gabinete de Análisis y Planificación de La Universidad de La Laguna y será analizada en el lenguaje y entorno de programación R que está enfocado al análisis estadístico.

3.1. Descripción del conjunto de datos

La Sede Electrónica de La Universidad de La Laguna se creó en 2011 para fomentar el uso de los trámites electrónicos de forma segura. La sede ofrece un catálogo de procedimientos o trámites a disposición de la comunidad universitaria y la ciudadanía general.

El Gabinete de Análisis y Planificación de La Universidad de La Laguna (GAP), por su parte, se encarga de gestionar la sede electrónica de la universidad y registra todos los procedimientos que son iniciados en ella diariamente desde el año 2014. La serie temporal que se analiza en este capítulo ha sido generosamente cedida por el GAP y recoge el número de procedimientos del catálogo iniciados mensualmente en la sede electrónica desde enero de 2016 hasta marzo

de 2022.

Así, en las próximas secciones se construirá un modelo $ARIMA$ sobre la serie temporal ofrecida por el GAP con el objetivo de predecir las futuras demandas en la sede.

3.2. Análisis de la serie temporal

3.2.1. Identificación

El análisis de una serie temporal comienza examinando su gráfico temporal, del que se pueden extraer algunas conclusiones sobre la tendencia, el comportamiento estacionario y la estacionalidad de la misma.

La Figura 3.1 sugiere una tendencia creciente, así como un cierto comportamiento estacional que se repite cada 12 meses y se manifiesta como dos grandes picos en los meses de julio y septiembre. Este comportamiento descarta por completo la estacionariedad de la serie, tal y como se comentó en la Sección 1.3. Por ello, sería necesario practicar una diferenciación y desestacionalización para hacer constante la media, acompañada de una transformación de Box-Cox que consiga estabilizar la varianza de la serie en torno a su media.

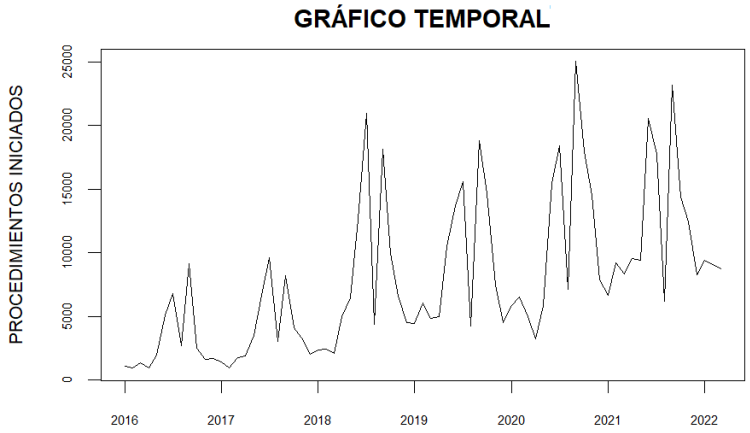

Figura 3.1. Serie temporal del número de procedimientos mensuales iniciados en la sede electrónica.

Para comprender el comportamiento general de la serie, resulta útil estudiar su descomposición clásica. Siguiendo los comentarios de la Sección 1.2, como la componente estacional y la tendencia parecen crecer en el tiempo, es más apropiado asumir un modelo multiplicativo (1.2).

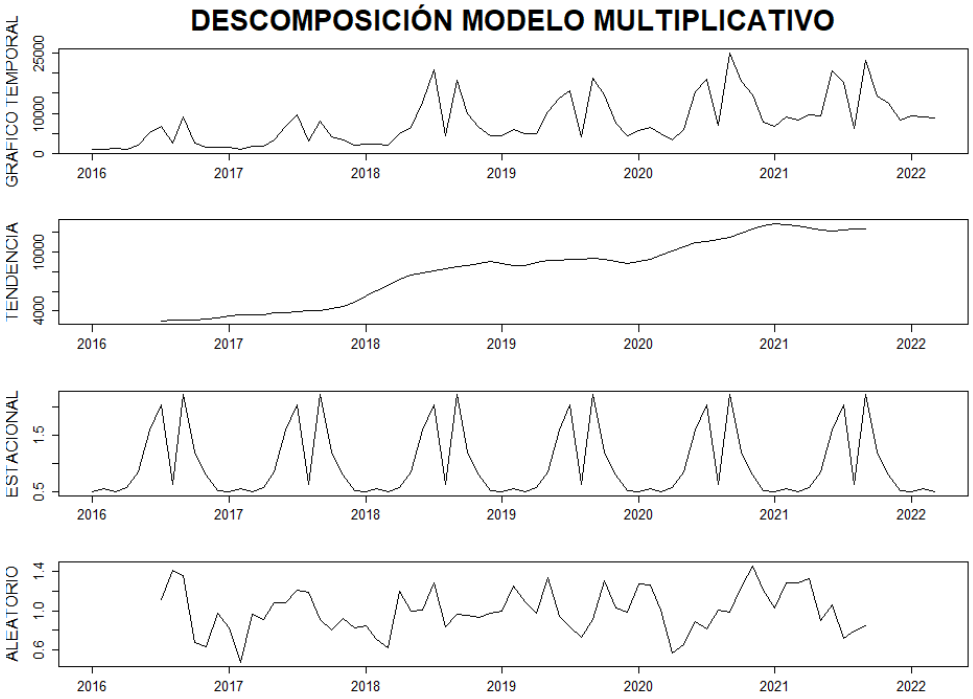

Figura 3.2. Descomposición clásica de la serie temporal asumiendo un modelo multiplicativo.

Las sospechas de tendencia y estacionalidad formuladas más arriba, se resuelven examinando el gráfico de la Figura 3.2 que representa las componentes de la serie: se aprecia una clara tendencia creciente, así como un comportamiento estacional.

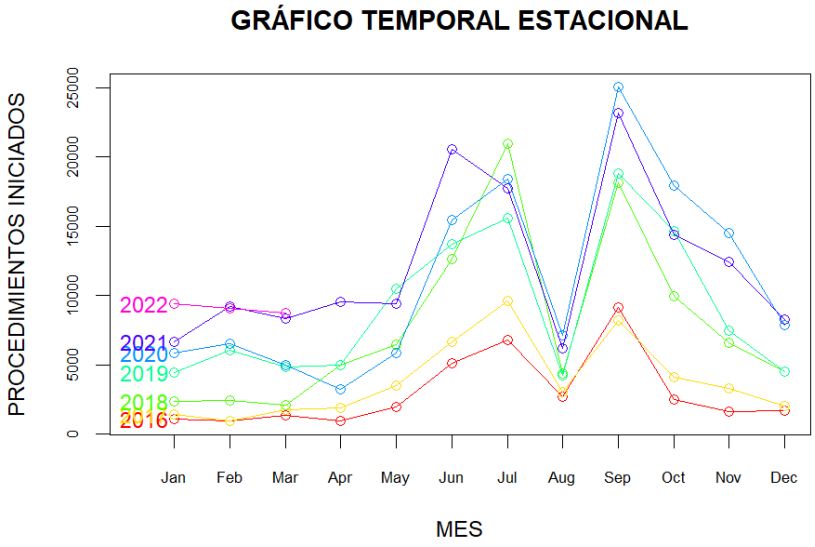

Figura 3.3. Gráfico estacional.

Para salir de dudas, puede generarse un gráfico estacional como el de la Figura 3.3, que recoge la evolución de la serie temporal mensualmente mostrando de un color diferente cada año. Este revela que la serie muestra dos picos en los meses de julio y septiembre, y que, en efecto, el comportamiento estacional es un patrón que se repite cada 12 meses.

En la Sección 1.4 se indicó que el gráfico ACF es una herramienta útil para detectar estacionariedad de cierta serie temporal. En particular, como el correlograma de la Figura 3.4 no decae rápidamente, sino que muestra un decrecimiento lento de carácter oscilatorio, la serie temporal analizada parece no estacionaria. Una herramienta más fiable para analizar el comportamiento estacionario de una serie temporal es el contraste de hipótesis de Dickey-Fuller Aumentado (ADF) que se formula como (2.31) y que, aplicado a la serie analizada, devuelve un p-valor de 0.1279 que es mayor que los niveles de significación usuales 0.01 y 0.05. Esto es que no hay evidencias suficientes para rechazar la hipótesis nula de no estacionariedad y se concluye que, en efecto, la serie temporal es no estacionaria.

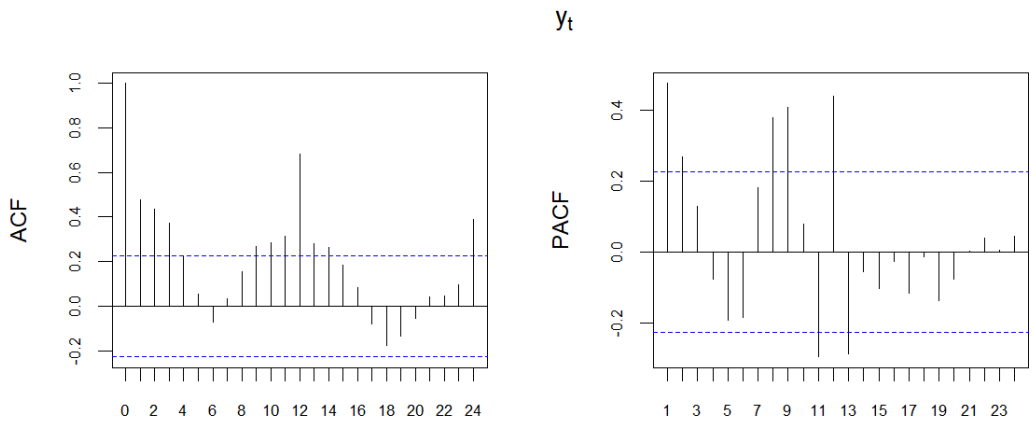

Figura 3.4. Gráfico ACF y gráfico PACF de la serie temporal.

Sea $\{y_t\}_{t=1}^{75}$ la serie original no estacionaria que se está analizando. El comportamiento estacional descubierto más arriba sugiere un modelo $SARIMA$ (2.25) para $S = 12$, luego deben identificarse los órdenes p, d, q, P, D y Q del modelo. Para ello, se considera la serie transformada $\{\nabla_{12}(\nabla y_t^{(\lambda)})\}_{t=1}^{75}$ a la que se ha aplicado una diferenciación (1.14), acompañada de una desestacionalización (1.19) de $S = 12$ y una transformación de Box-Cox (1.10) para $\lambda = 0.318$, cuyo valor ha sido aportado por el lenguaje R. Esta serie es estacionaria pues, además de que tanto el gráfico temporal como el correlograma de la Figura 3.5 parecen indicarlo, el contraste ADF devuelve un p-valor inferior a 0.01, lo que hace rechazar su hipótesis nula.

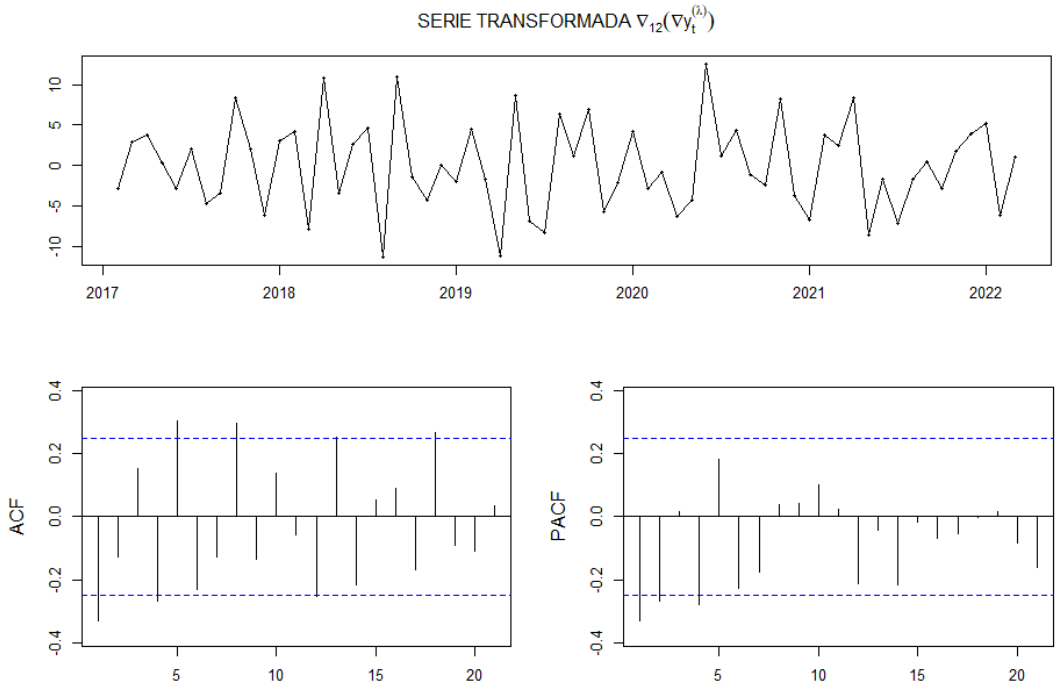

Figura 3.5. Gráfico temporal, gráfico ACF y gráfico PACF de la serie trasformada.

Por el momento, se conoce que se aplicará un modelo

$$SARIMA(p,1,q) \times (P,1,Q)_{12} \tag{3.1}$$

a la serie $\{y_t^{(\lambda)}\}_{t=1}^{75}$ y falta por determinar el resto de los órdenes. Las directrices expuestas en la Sección 2.5 aplicadas a los gráficos ACF y PACF de la Figura 3.5, sugieren los valores $p = 2$, $q = 1$, $P = 0$ y $Q = 1$, que describen por completo el modelo.

3.2.2. Ajuste

Una vez concluye la fase de identificación y se decide que el modelo a utilizar es

$$SARIMA(2,1,1) \times (0,1,1)_{12}, \tag{3.2}$$

llega el momento de estimar sus coeficientes. Esta labor se deja al lenguaje R que devuelve el modelo

$$\phi(B)(1-B)(1-B^{12})\ Y_t^{(\lambda)} = \theta(B)\Theta(B)\ \varepsilon_t, \tag{3.3}$$

siendo

- $\phi(z) = 1 + 0.4099z + 0.0316z^2$,

- $\theta(z) = 1 - 0.8986z$,
- $\Theta(z) = 1 - 0.5472z^{12}$,

cuyos AIC (2.26) y BIC (2.28) son, respectivamente, 375.28 y 385.92. Además las medidas de bondad de predicción expuestas en la Sección 1.6, se pueden consultar el la Tabla 3.1.

RMSE	MAE	MAPE
2172.613	1490.413	19.001

Tabla 3.1. Medidas de bondad de predicción sobre la serie original

No debe perderse de vista que el modelo está estimado para la serie $\{y_t^{(\lambda)}\}_{t=1}^{75}$ transformada por Box-Cox. Esto implica que, a la hora de hacer predicciones en la Sección 3.2.4, habrá que deshacer la transformación.

3.2.3. Diagnosis

Cuando el modelo está estimado, es importante realizar la diagnosis del mismo. Los dos aspectos fundamentales a comprobar son que el modelo sea estacionario e invertible, y que los residuos se comporten como ruido blanco.

- El modelo es estacionario e invertible

Este aspecto es realmente sencillo de comprobar pues basta observar que los polinomios autorregresivo, media móvil y media móvil estacional son, respectivamente

$$\begin{aligned} \phi(z) &= 1 + 0.4099z + 0.0316z^2, \\ \theta(z) &= 1 - 0.8986z, \\ \Theta(z) &= 1 - 0.5472z^{12}. \end{aligned} \tag{3.4}$$

Las raíces del polinomio autorregresivo son $z_1 = -9.7137$ y $z_2 = -3.2578$, valores mayores que la unidad en módulo, luego el proceso es estacionario. En cuanto al polinomio media móvil, su raíz $z = 1.1128$ cae, también, fuera del círculo unidad. Además, como todas las raíces del polinomio media móvil estacional son de módulo 1.0515, puede concluirse que el proceso es invertible.

- Los residuos del modelo son ruido blanco

En primer lugar, tal y como se recomendaba en la Sección 2.7, se examina el gráfico temporal y correlograma de los residuos que ofrece la Figura 3.6. En

efecto, puede apreciarse un comportamiento ideal de los mismos: parecen estacionarios y las barras de su gráfico ACF caen dentro de las bandas discontinuas azules horizontales.

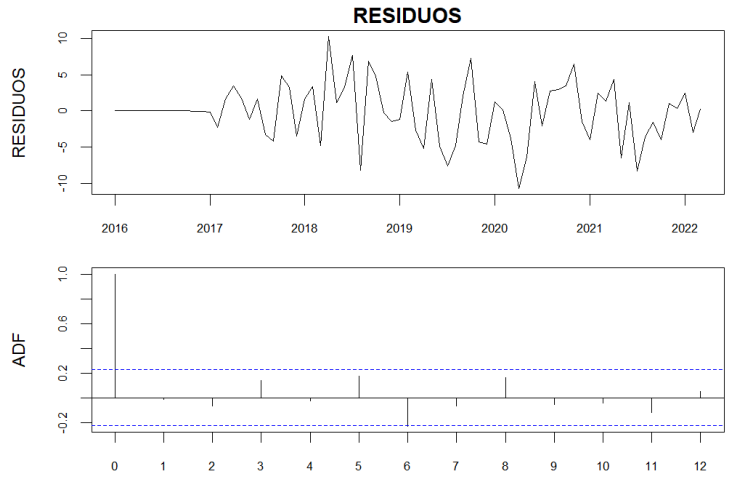

Figura 3.6. Residuos del modelo estimado.

En cuanto a la normalidad de los residuos, el gráfico cuantil-cuantil (QQ, del inglés *Quantile-Quantile*) y el histograma que muestra la Figura 3.7, parecen indicar que siguen la distribución normal deseada puesto que los puntos del gráfico QQ caen mayoritariamente sobre la linea roja de 45° y el histograma dibuja una gaussiana. Además, el p-valor del contraste de hipótesis para la normalidad de Jarque-Bera es 0.9348, un valor mucho mayor que los niveles de significación usuales que no aporta evidencias suficientes para rechazar la hipótesis nula que asume la normalidad de los residuos.

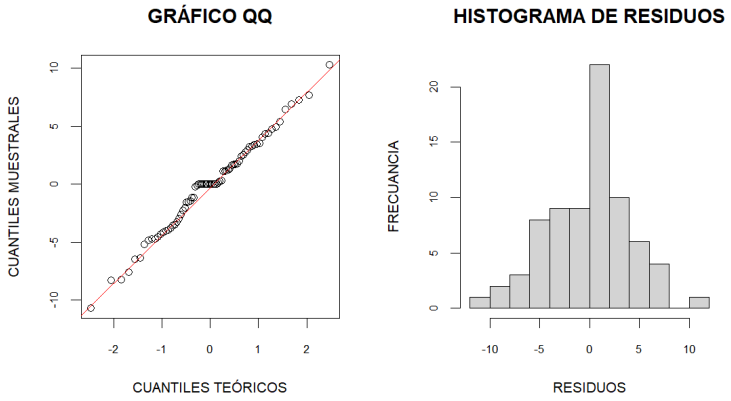

Figura 3.7. Gráfico cuantil-cuantil e histograma de los residuos.

Por último, es esencial comprobar la independencia de los residuos mediante el contraste de hipótesis de Ljung-Box (2.33). Debe recordarse que este contraste compara tan solo los M primeros coeficientes de autocorrelación y no es sencillo decantarse por un valor de M en particular. Por ello, la Figura 3.8 recoge los p-valores para distintos tamaños de M y se aprecia que estos son mucho mayores que los niveles de significación usuales. Estos resultados no aportan evidencias para rechazar la hipótesis que asume independencia de los residuos y, con todo esto, se concluye que los residuos se comportan como ruido blanco, finalizando así la etapa de diagnosis del modelo.

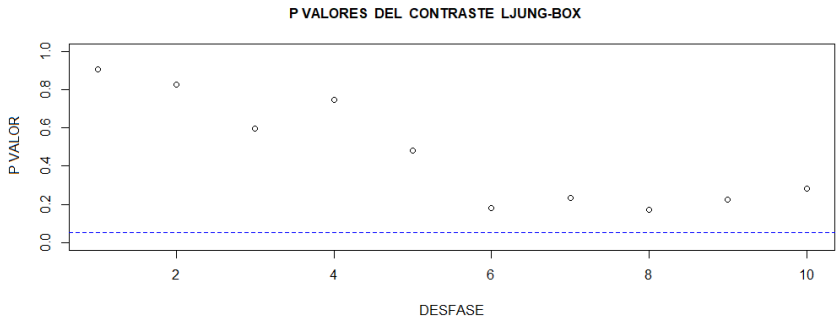

Figura 3.8. p-valores del contraste de Ljung-Box para distintos valores de M.

3.2.4. Predicción

Se ha estimado el modelo y llega el momento de hacer predicciones. El lenguaje R ha construido el modelo para la serie transformada $\{y_t^{(\lambda)}\}_{t=1}^{75}$, pero interesa conocer las predicciones para la serie original. Por ello, es importante deshacer la transformación de Box-Cox.

La Tabla 3.2 ofrece la predicción de procedimientos iniciados que aporta el modelo (columna "*Forecast*") en los diferentes puntos del horizonte de predicción seleccionado de 12 meses. Junto a esta información, se encuentra los límites inferiores y superiores de confianza al 95 % (columnas *"Lo 95"* y *"Hi 95"*, respectivamente) de cada predicción.

Todo esta información es representada gráficamente en la Figura 3.9, en la que se encuentra la serie original representada, los predichos del modelo y la predicción en el horizonte junto a su intervalo de confianza al 95 %.

Point	Forecast	Lo 95	Hi 95
Apr 2022	9079.774	5315.100	14 346.240
May 2022	11 727.073	6723.634	18 810.240
Jun 2022	22 674.778	14 243.277	33 996.010
Jul 2022	23 756.261	14 856.874	35 736.760
Aug 2022	9029.378	4624.268	15 672.930
Sep 2022	28 941.111	18 403.248	42 989.040
Oct 2022	19 468.357	11 511.278	30 535.530
Nov 2022	15 457.519	8688.199	25 149.470
Dec 2022	10 303.342	5260.835	17 920.790
Jan 2023	10 821.751	5548.136	18 770.450
Feb 2023	11 724.458	6080.810	20 176.830
Mar 2023	10 879.554	5503.971	19 041.460

Tabla 3.2. Predicciones del lenguaje R del modelo $SARIMA(2,1,1) \times (0,1,1)_{12}$.

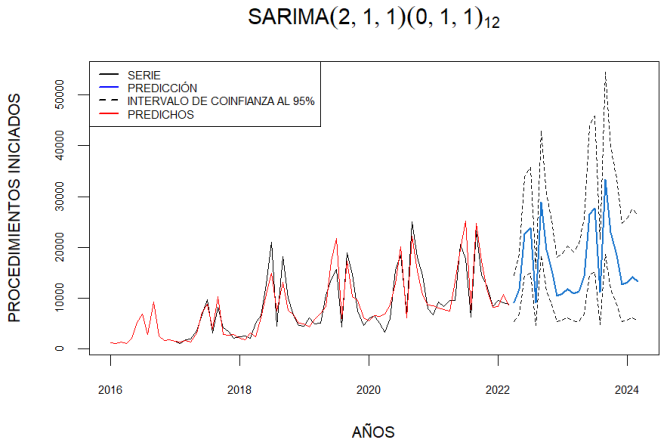

Figura 3.9. Predicción del modelo $SARIMA(2,1,1) \times (0,1,1)_{12}$ en un horizonte de 24 meses.

3.3. Modelización automática de la serie temporal

El paquete *"forecast"* del lenguaje R tiene una función que compara varios modelos $ARIMA$ y propone el más adecuado según el algoritmo descrito en la Subsección 3.3.1. Esta función transforma automáticamente la serie original con una transformación de Box-Cox para $\lambda = 0.318$ y propone el modelo

$$SARIMA(1,0,0) \times (0,1,1)_{12} \tag{3.5}$$

formulado como en (2.25) donde

- $c = 0.4020$,
- $\phi(z) = 1 + 0.4951z$,
- $\Theta(z) = 1 - 0.6005z^{12}$.

Este modelo es estacionario e invertible pues la función nunca devuelve un modelo que no lo sea. Además, se puede comprobar que el modelo propuesto pasa la fase de la diagnosis procediendo como en la Sección 3.2.3.

3.3.1. Algoritmo de la modelización automática

El algoritmo que selecciona el mejor modelo es descrito por Hyndman [9] (2008) y se desarrolla en 2 grandes pasos.

PASO 1. Se implementan cuatro posibles modelos:

- $ARIMA(2,d,2)$ si $S=1$, o $SARIMA(2,d,2)\times(1,D,1)_S$ si $S>1$.
- $ARIMA(0,d,0)$ si $S=1$, o $SARIMA(0,d,0)\times(0,D,0)_S$ si $S>1$.
- $ARIMA(1,d,0)$ si $S=1$, o $SARIMA(1,d,0)\times(1,D,1)_S$ si $S>1$.
- $ARIMA(0,d,1)$ si $S=1$, o $SARIMA(0,d,1)\times(0,D,1)_S$ si $S>1$.

Todos estos modelos se construyen con $c\neq 0$ si $d+D\leq 1$ y, en el caso contrario, $c=0$. Se calcula el AIC de cada uno de estos cuatro modelos y se selecciona como modelo "actual" el que menor AIC ofrezca.

PASO 2. Se consideran las variaciones del modelo actual:

- Se permite que solo uno de los órdenes p, q, P o Q del modelo actual varíe en ± 1.
- Los dos órdenes p y q del modelo varían en ± 1 simultáneamente.
- Los dos órdenes P y Q del modelo varían en ± 1 simultáneamente.
- Si el modelo actual no incluye la constante c, se considera el modelo con la constante. En el caso contrario, si el modelo actual incluye la constante c, entonces se considera el modelo que la excluye.

Cada vez que se efectúa una de estas modificaciones, se computa el AIC y, si se reduce, se selecciona dicho modelo como el actual y se vuelve al **PASO 2.** El algoritmo se detiene cuando no se encuentra un modelo que reduzca el AIC del modelo actual.

Este algoritmo tiene en cuenta algunas restricciones sobre los órdenes y parámetros que pueden ser consultadas en [9].

3.3.2. Resultados de la modelización automática

Llegados a este punto, habría que decidir qué modelo se ajusta mejor a la serie temporal analizada. Para ello, debe prestarse atención a la Tabla 3.3.2 que ofrece los valores de AIC, BIC y de bondad de predicción de los modelos (3.2) y (3.5). Sucede que el modelo propuesto por el modelizador automático del lenguaje R parece obtener mejores resultados.

Modelo	AIC	BIC	RMSE	MAE	MAPE
$SARIMA(2,1,1)\times(0,1,1)_{12}$	375.28	385.92	2172.613	1490.413	19.001
$SARIMA(1,0,0)\times(0,1,1)_{12}$	375.20	383.77	2133.135	1453.391	17.963

Tabla 3.3. Comparativa de los dos modelos

Por este motivo, habría que deshacerse del modelo asumido en la Subsección 3.2.1 y hacer predicciones con el modelo que propone el modelizador automático (3.5). Estas predicciones se recogen en la Tabla 3.4 y la representación gráfica de esta información se ofrece en la Figura 3.10.

Point	Forecast	Lo 95	Hi 95
Apr 2022	9667.481	5856.622	14 893.950
May 2022	12 997.066	7807.032	20 150.240
Jun 2022	24 554.215	16 126.572	35 575.730
Jul 2022	26 369.589	17 451.063	37 980.240
Aug 2022	10 413.417	5894.506	16 857.530
Sep 2022	31 636.697	21 445.765	44 713.980
Oct 2022	21 440.650	13 756.138	31 630.920
Nov 2022	17 006.900	10 518.782	25 798.220
Dec 2022	11 606.059	6708.176	18 507.920
Jan 2023	12 125.220	7067.820	19 217.380
Feb 2023	13 109.961	7752.726	20 560.590
Mar 2023	12 230.296	7140.576	19 361.170

Tabla 3.4. Predicciones del lenguaje R del modelo $SARIMA(1,0,0)\times(0,1,1)_{12}$.

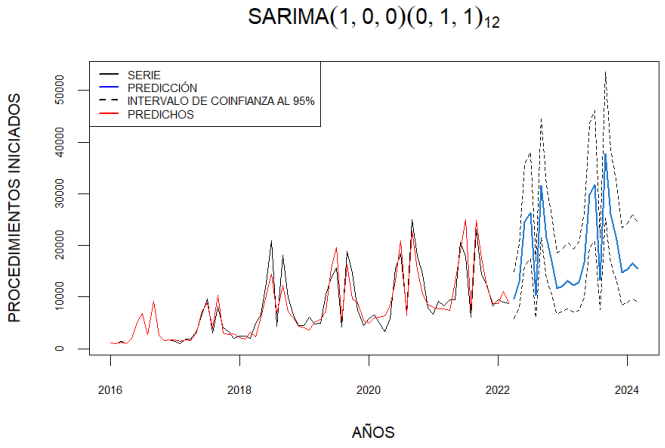

Figura 3.10. Predicción del modelo $SARIMA(1,0,0)\times(0,1,1)_{12}$ en un horizonte de 24 meses.

Conclusiones

En esta memoria se ha presentado la familia de modelos ARIMA para el análisis de series temporales. Se ha hecho un estudio de estos modelos tanto teórico como práctico, con el objetivo de entender en profundidad los mismos.

Este texto comienza definiendo el concepto de proceso estocástico y entendiendo una serie temporal como una realización del mismo. Seguidamente, se introducen conceptos fundamentales para el análisis de series temporales, así como elementos que serán esenciales para construir la familia de modelos ARIMA y analizar cualquier conjunto de datos.

Sin lugar a dudas, la parte más atractiva de este trabajo es el análisis de la serie temporal cedida por el GAP que se presenta en el último capítulo. El interés de esta última parte reside en que se aplican todos los conceptos teóricos expuestos hasta el momento a una base de datos real.

Se ha elegido el modelo de la familia ARIMA que parece ajustarse a la serie siguiendo los criterios de los primeros capítulos. Una vez estimado el modelo, se comprobaron los requerimientos en la etapa de diagnosis y se obtuvieron los resultados deseados. También, se ofrecieron las predicciones que aporta el modelo junto a un gráfico de las mismas en el que se aprecia que la serie continúa con una tendencia creciente y con su comportamiento estacional característico.

Por último, se dejó la modelización de la serie a una librería del lenguaje R que sigue un algoritmo que pretende conseguir el modelo que menor AIC aporte. Los resultados de este modelo automático se incluyeron en la memoria tanto en forma de tabla de datos, como en su formato gráfico. Este modelo se comparó con el propuesto anteriormente mediante las medidas de bondad de predicción y, al ser estas medidas menores en el modelo automático, se consideraron las predicciones del modelo propuesto por el modelizador automático como más

certeras.

Se recomienda como futuras lineas de trabajo o formación la introducción del análisis espectral de datos al análisis de series temporales que se presenatan en este trabajo, así como el estudio de los modelos ARIMAX, que consideran una o varias variables regresoras; los modelos GARCH, para series temporales que presentan heterocedasticidad condicionada; o las técnicas de suavizado exponencial. El estudio de estos nuevos modelos es interesante para poder comparar los resultados de las predicciones, mediante las medidas expuestas en esta memoria. Otra línea de trabajo muy amplia es el estudio de diferentes librerías del lenguaje R para el análisis de series temporales.

A

Apéndice

A.1. Código en el lenguaje R para el análisis de series temporales

Se presentan, a continuación, los comandos del lenguaje R utilizados para analizar la serie temporal almacenada en el objeto de clase `ts` llamado `tsdata`. Previamente, deben cargarse las librerías `dplyr`, `xts`, `lattice`, `forecast`, `tseries`, `astsa`, `fUnitRoots`, `latex2exp`, `ggpubr`, `moments` y `nortest`.

```
####### 1. IDENTIFICACIÓN
# GRÁFICO TEMPORAL
plot(tsdata, xlab = "", ylab = "PROCEDIMIENTOS INICIADOS", main = "GRÁFICO TEMPORAL")
# GRÁFICO ADF PACF
mx=12*2
par(mfrow=c(1,2))
tsdata %>% acf(lag.max=mx, xaxt="n", main = TeX("$y_t$"), xlab = "", ylab = "ACF")
axis(1, at=0:mx/12, labels=0:mx)
tsdata %>% pacf(lag.max=mx, xaxt="n", main = "", xlab = "", ylab = "PACF")
axis(1, at=0:mx/12, labels=0:mx)
# DESCOMPOSICIÓN CLÁSICA
tsdata  %>% decompose(type = "multiplicative") %>% plot()
# GRÁFICO ESTACIONAL
seasonplot(tsdata, col = rainbow(7),  year.labels=FALSE, year.labels.left=TRUE,
    ylab = "PROCEDIMIENTOS INICIADOS", xlab = "MES", main = "GRÁFICO TEMPORAL
    ESTACIONAL")
# ADF TEST
adfTest(tsdata) # RESULTADO: NO ESTACIONARIO
# NÚMERO DE DIFERENCIAS (SIMPLES O ESTACIONALES, RESPECTIVAMENTE) SUGERIDAS
ndiffs(tsdata)  # RESULTADO d = 1
nsdiffs(tsdata) # RESULTADO D = 1
# TRANSFORMACIÓN DE BOX-COX
lambda <- BoxCox.lambda(tsdata) #SUGIERE lambda = 0.318
tsdata.transformed <- tsdata %>% BoxCox(lambda)
# SERIE ESTACIONARIA
stationary <- tsdata.transformed %>% diff(differences =  1) %>% diff(differences = 1,
              lag = 12)
stationary %>% plot(xlab ="", ylab = TeX("$\\nabla_{12}(\\nabla y_t^{(\\lambda)})$"),
                    main = "GRÁFICO TEMPORAL")
par(mfrow=c(1,2))
stationary %>% acf(lag.max=mx, xaxt="n", xlab = "", ylab = "ACF",
                   main = TeX(" $\\nabla_{12}(\\nabla y_t^{(\\lambda)})$"))
axis(1, at=0:mx/12, labels=0:mx)
stationary %>% pacf(lag.max=mx, xaxt="n", main = "", xlab = "", ylab = "PACF")
```

```
axis(1, at=0:mx/12, labels=0:mx)

####### 2. AJUSTE
# AFC Y PACF PLOT SUGIERE UN ARIMA(2,1,2)(0,1,1)[12]
t1 <- tsdata.transformed %>% Arima(order = c(2,1,2), seasonal = list(order = c(0,1,1),
                                    period=12))

####### 3. DIAGNOSIS
# LOS RESIDUOS SON ESTACIONARIOS
plot(t1$residuals, xlab = "", ylab = "", main = "RESIDUOS")

mx = 12
par(mfrow = c(2,1))
plot(t1$residuals, xlab = "", ylab = "RESIDUOS", main = "RESIDUOS")
t1$residuals %>% acf(lag.max=mx, xaxt="n", xlab = "", ylab = "ADF", main = "")
axis(1, at=0:mx/12, labels=0:mx)
# LOS RESIDUOS SON INDEPENDIENTES
t1$residuals %>% Box.test(type = "Ljung-Box")
tsdiag(t1)
# LOS RESIDUOS SIGUEN UNA DISTRIBUCIÓN NORMAL
par(mfrow=c(1,2))
qqnorm(t1$residuals, main = "GRÁFICO QQ", xlab = "CUANTILES TEÓRICOS", ylab =
        "CUANTILES MUESTRALES")
qqline(t1$residuals, col = "red")
hist(t1$residuals, main = "HISTOGRAMA DE RESIDUOS", xlab = "RESIDUOS", ylab =
     "FRECUANCIA")
t1$residuals %>% jarque.bera.test() #RESULTADO: p-valor 0.9348

####### 4. PREDICCIÓN
horizonte <- 12*2
pred <- t1 %>% forecast(h = horizonte,  level = c(95))
# DESHACER TRANSFORMACIÓN
pred$mean <- pred$mean %>% InvBoxCox(lambda)
pred$lower <- pred$lower %>% InvBoxCox(lambda)
pred$upper <- pred$upper %>% InvBoxCox(lambda)
pred$x <- pred$x %>% InvBoxCox(lambda)
pred$fitted <- pred$fitted %>% InvBoxCox(lambda)
pred$residuals <- pred$residuals %>% InvBoxCox(lambda)
#MEDIDAS DE BONDAD DE PREDICCIÓN
accuracy(pred)
# PREDICCIÓN
pred %>% plot(shaded = FALSE, xlab = "AÑOS", ylab = "PROCEDIMIENTOS INICIADOS",
    main = TeX("$SARIMA(2,1,2)(0,1,1)_{12}$"))
lines(pred$fitted, col = "red")
legend("topleft", legend=c("SERIE", "PREDICCIÓN", "INTERVALO DE COINFIANZA AL 95%",
       "PREDICHOS"), col=c("black", "blue", "black", "red"), lty=c(1,1,2,1), lwd = 2,
       cex = 0.6)

### MODELIZACIÓN AUTOMÁTICA
t1 <- tsdata %>% auto.arima()
horizonte <- 12*2
pred <- t1 %>% forecast(h = horizonte,  level = c(95))
```

Bibliografía

[1] Banerjee, P. (2020, 23 octubre). *ARIMA Model for Time Series Forecasting.* Kaggle. https://www.kaggle.com/code/prashant111/arima-model-for-time-series-forecasting/notebook#ARIMA-Model-for-Time-Series-Forecasting

[2] Brockwell, P. J., Davis, R. J., Rose, C., Richard A. Davis, P. J. B., Smith, M. D., Calder, M. V., & Springer (Firm). (2002). *Introduction to Time Series and Forecasting.* Springer Publishing.

[3] Calzone, O. (2022, 7 abril). *MAE, MSE, RMSE, and F1 score in Time Series Forecasting.* Medium. https://medium.com/@ottaviocalzone/mae-mse-rmse-and-f1-score-in-time-series-forecasting-d04021ffa7ce

[4] Date, S. (2019, 9 noviembre). *The Akaike Information Criterion - Towards Data Science.* Medium. https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2

[5] F. (2021, 1 diciembre). *How to Interpret ACF and PACF plots for Identifying AR, MA, ARMA, or ARIMA Models.* Medium. https://medium.com/@ooemma83/how-to-interpret-acf-and-pacf-plots-for-identifying-ar-ma-arma-or-arima-models-498717e815b6

[6] Gençay, R., Selçuk, F., Whitcher, B. J. (2001). *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics* (1.a ed.). Academic Press.

[7] González Sierra, M. A. (2014). *Lecciones de Estadística Descriptiva.* Fotocopias Campus.

[8] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). Otexts.

[9] Hyndman, R. J., Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3), 1–22. https://doi.org/10.18637/jss.v027.i03

[10] Jiménez, E. U., Uriel, E. (1995). *Análisis de series temporales.* Paraninfo.

[11] Masum, M., PhD. (2020, 13 agosto). *Time Series Analysis: Identifying AR and MA using ACF and PACF Plots*. Medium. https://towardsdatascience.com/identifying-ar-and-ma-terms-using-acf-and-pacf-plots-in-time-series-forecasting-ccb9fd073db8

[12] Mauricio, J. A. (2007). *Introducción al Análisis de Series Temporales*. CERSA. https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf

[13] Nau, R. (2020) *Statistical forecasting: notes on regression and time series analysis* Fuqua School of Business. Duke University. https://people.duke.edu/~rnau/411home.htm

[14] Palachy, S. (2019, 8 abril). *Stationarity in time series analysis - Towards Data Science.* Medium. https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322

[15] Pennsylvania State University, Eberly College of Science. (2021). *Penn State University Applied Time Series Analysis*. PennState: Statistics Online Courses. https://online.stat.psu.edu/stat510/

[16] Salvi, J. (2019, 27 marzo). *Significance of ACF and PACF Plots In Time Series Analysis.* Medium. https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8

[17] Yaffee, R. A., McGee, M. (2000). *Introduction to Time Series Analysis and Forecasting.* Elsevier Gezondheidszorg.

Fundamentals and variations of ARIMA models for time series analysis. Applications to university statistics.

Jorge Guerra Rodríguez
Facultad de Ciencias • Sección de Matemáticas
Universidad de La Laguna
alu0101225814@ull.edu.es

Abstract

TIME SERIES ANALYSIS *is a group of statistical techniques used to describe and anticipate the behaviour of a time series, study the stochastic process where the series comes from and forecasting. The ARIMA models are widely used and they are characterized by an outstanding performance in short-term forecasts of series where seasonality is shown.*

This thesis introduce the ARIMA models and analyses a real time series considering the hypothesis of the models and trying to build the model that best fit to the data. The time series is analyzed with the R environment for statistical computing.

1. ARIMA Models

TIME SERIES are sequences of measurements of the same variable made at regular time intervals over time. Seasonal $ARIMA$ models or $SARIMA(p,d,q) \times (P,D,Q)_S$ are widely used for modeling time series that show seasonality. Seasonality arises while checking a time plot of a series as a repeating behaviour that occurs at specific time intervals.

2. Time Series Analysis

TIME SERIES ANALYSIS is usually made in three steps: identification, estimation and diagnostic.

The time series that is being analysed count the monthly procedures in the electronic platform of the ULL since January 2016 to March 2022.

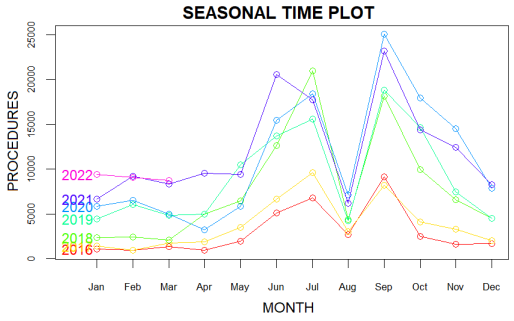

Figure 1: *Procedures on the electronic headquarters every month of each year.*

The Figure 1 reveals that the time series has a seasonal behaviour that arise as two peaks at July and September. This suggests that a seasonal $ARIMA$ model would fit the data well. The identification step points out the model

$$SARIMA(2,1,1) \times (0,1,1)_{12}, \tag{1}$$

which passes the diagnostic step and could be used for forecasting. Anyways, one could let an R's library do all the work and the identification step would give the model

$$SARIMA(1,0,0) \times (0,1,1)_{12}, \tag{2}$$

which also passes the diagnostic steps and performs better over the data as shown by the goodness of fit measurements given in Table 1. Hence, (2) will be used for forecasting.

Model	RMSE	MAE	MAPE
(1)	2172.613	1490.413	19.001
(2)	2133.135	1453.391	17.963

Table 1: *Goodness of fit measurements.*

What is done by R language is a step-wise algorithm that gets the seasonal $ARIMA$ model with lowest Akaike Information Criterion (AIC). (1) gives an AIC of 375.28 while (2) 375.20.

3. Forecasting

THE PREDICTIONS and its 95% confidence interval made by (2) are shown in Figure 2. The forecasts have the increasing trend and seasonal behaviour that one could expect by examining the older data.

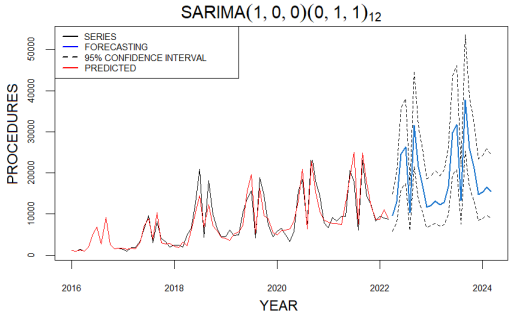

Figure 2: *Procedures on the electronic headquarters every month of each year.*

Table 2 gives the forecasts and its 95% confidence interval for a few future months.

Point	Forecast	lower-bound	upper-bound
April 2022	9667.481	5856.622	14893.950
May 2022	12997.066	7807.032	20150.240
June 2022	24554.215	16126.572	35575.730
July 2022	26369.589	17451.063	37980.240
August 2022	10413.417	5894.506	16857.530

Table 2: *Forecasts.*

References

[1] HYNDMAN, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). Otexts.

[2] BROCKWELL, P. J., Davis, R. J., Rose, C., Richard A. Davis, P. J. B., Smith, M. D., Calder, M. V., & Springer (Firm). (2002). *Introduction to Time Series and Forecasting*. Springer Publishing.