FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Modelización de Series de Tiempo Macroeconómicas

Ana Pérez Sanromán

Julio, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Modelización de Series de Tiempo Macroeconómicas

Ana Pérez Sanromán

Julio, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística e investigación operativa.
Título: Modelización de series de tiempo macroeconómicas
Breve descrición do contido
El objetivo de este trabajo es aprender a modelar series de tiempo, concretamente en el ámbito económico a gran escala. Se elegirá una serie con datos de Galicia. El trabajo consiste en la descripción de las herramientas necesarias para analizar una serie temporal, construír modelos adecuados, plantear su estimación y diagnosis, para finalmente hacer predicciones de futuros valores en base al modelo seleccionado.

Índice

Resumen

En este TFG se estudian las series de tiempo, una herramienta estadística útil para estudiar la evolución de unos datos en el tiempo. En particular se aplicará el estudio a las series de tiempo macroeconómicas, que se producen con periodicidad mensual o mayor, por su capacidad de predecir valores futuros a partir de valores pasados. El objetivo de este trabajo es modelizar una serie concreta, que en este caso será "Pasajeros totales del transporte aéreo en Galicia", para predecir futuros valores del número de pasajeros. En primer lugar, se introduce el concepto de series de tiempo, aportando las herramientas y técnicas necesarias para su análisis. Se presenta una metodología utilizada para el análisis de series temporales, llamada Box-Jenkins, que clasifica el proceso en tres etapas (identificación, estimación y validación del modelo) y usa los modelos ARIMA para modelizarlas. Estos modelos ARIMA (Autorregresivos Integrados de Medias Móviles) son de los más usados en este tipo de series de tiempo, y son descritos en esta sección. La última parte de este proceso consiste en obtener futuras predicciones de la serie de tiempo, en este caso, el total de pasajeros del tráfico aéreo.

Abstract

This project is about Time Series, a useful tool in statistics to study the evolution in time of some data. The approach given here will be of macroeconomic time series, produced with mensual or greater periodicity, because of the ability of these series to predict future values from past values observed from the data. The aim of this study is to model a particular time series, in this case"Total air traffic passengers in Galicia ", to predict future values of this variable. First of all, the concept of time series will be introduced, together with the necessary tools and techniques for the analysis. A useful method to analyze time series is presented here, it is Box-Jenkins methodology, and it is classified in three stages (detection, estimation and model checking) using ARIMA models in order to model the time series. ARIMA models (Autoregressive Integrated

Moving Average models) are the most used models in time series, and they are described in this section. The last part of this process consists of obtaining future forecasts of time series, in this case, the total number of air traffic passengers.

Introducción

Una serie de tiempo regular es un conjunto de observaciones de una variable tomadas sobre una población a intervalos de tiempo regulares. La serie se considera estadísticamente como un conjunto de variables aleatorias definidas en un mismo espacio de probabilidad, de las que sólo se dispone de una muestra.
El estudio de las series de tiempo surge de la dificultad de aplicar las técnicas clásicas de estadística a un conjunto de datos observados en el tiempo, pues estas técnicas se basan en la suposición de que los datos son independientes e idénticamente distribuidos, y realmente los datos observados en una serie de tiempo presentan autocorrelación temporal, correlación de una variable consigo misma en diferentes momentos en el tiempo. El análisis de series temporales es, por tanto, una herramienta fundamental en Estadística para estudiar fenómenos que evolucionan en el tiempo. Este trabajo surge del interés de estudiar la evolución de estos fenómenos en el ámbito macroeconómico: la economía no es estática, las variables macroeconómicas cambian con el tiempo de una forma compleja que requiere modelos específicos para su estudio. Modelizar su evolución permite entender el comportamiento de una economía y anticipar tendencias futuras, lo que facilita la toma de decisiones políticas y financieras.

El objetivo de este trabajo es modelizar una serie de tiempo macroeconómica concreta, "Pasajeros totales del tráfico aéreo en Galicia de 1980 a 2019": extraer el patrón que describe el comportamiento pasado de la serie y predecir valores futuros. Matemáticamente, el trabajo consiste en seleccionar el modelo que mejor se ajuste a los datos. Realmente existen datos de esta serie hasta mayo de 2025, pero se decide modelizarla hasta 2018 por el impacto que causó la pandemia en los datos del transporte aéreo, que requeriría técnicas avanzadas de intervención que no son el objetivo de este trabajo. Posteriormente, se realiza una predicción de los valores correspondientes al año 2019, lo que permite evaluar la capacidad predictiva del modelo, pues se compara con los valores reales que toma la serie en este período.

Esta memoria se estructurará de la siguiente manera:

En primer lugar, se realiza un análisis descriptivo de la serie con el objetivo de explorar sus características básicas y de entender cómo se comporta la serie, detectar patrones e irregu-

laridades. Es importante calcular estadísticas básicas de las series de tiempo: media, varianza, función de autocorrelación simple y función de autocorrelación parcial, que permiten más adelante elegir y especificar un modelo apropiado. Esto se hace en la Sección 1.1. A continuación se verá que para trabajar con una serie esta debe ser estacionaria, lo cual implicará una serie de transformaciones. Es conveniente representarla gráficamente a lo largo del tiempo mediante un gráfico secuencial, pues permite identificar la presencia de componentes como tendencia, estacionalidad (patrones que se repiten cada cierto tiempo) o cambios en la varianza, y decidir qué transformaciones serán necesarias para conseguir que la serie sea estacionaria, paso previo a su modelización.

Una vez explicadas las características de la serie, en la Sección 1.2 se introduce formalmente el concepto de estacionariedad, concepto clave en el estudio de series de tiempo, y se estudia qué condiciones requiere e implica. Una serie es estacionaria si presenta media y varianza constantes a lo largo del tiempo, independientemente del instante temporal, del valor que tome la serie o del tamaño muestral. Para conseguir estacionariedad en media, se aplican técnicas como la diferencia o el ajuste por regresión. Para conseguir una varianza homocedástica, se llevan a cabo transformaciones Box-Cox, siendo la transformación logarítmica un caso particular. Estas transformaciones se realizan en el Capítulo 2.

Salvo que se asuman ciertas condiciones sobre la serie de tiempo, las estadísticas básicas no son calculables si solo se dispone de una realización muestral. En la Sección 1.3, se introducen por tanto las estimaciones muestrales correspondientes. La media, varianza y funciones de autocorrelación se estiman por sus correspondientes funciones muestrales.

En el Capítulo 2 se desenvuelve el objetivo de este trabajo, la modelización de una serie temporal concreta. Para ello se emplea la metodología Box-Jenkins , un método utilizado para el análisis de series de tiempo que emplea modelos ARIMA. La clase de modelos ARIMA engloba una amplia variedad de modelos. Se explicarán primero los modelos ARIMA regulares, en la Sección 2.1. Estos modelos incluyen componentes autorregresivas, AR (autoregressive), que captan la dependencia temporal entre valores de la serie; componentes de medias móviles, MA (moving averages), que modelizan la dependencia temporal entre errores de la serie que, como son aleatorios, explican parte de la aleatoriedad de la serie. Esta dependencia puede ser regular, estacional o de ambos tipos, lo que amplía la clase de modelos ARIMA regulares a los ARIMA estacionales multiplicativos, explicados en la Sección 2.2.1. Además, estos modelos incluyen una componente de diferenciación, I (integrated), que transforma la serie en estacionaria, pues como se ha visto la estacionariedad es clave para la modelización. La metodología Box-Jenkins se desarrolla en tres etapas: identificación del modelo, en la Sección 2.2; estimación de sus parámetros, en la Sección 2.3, y validación del modelo elegido, en la Sección 2.4. En cada una de estas etapas, se aplican las técnicas necesarias en cada caso a la serie de tiempo, a la vez que se explican teóri-

camente. En el proceso de modelización, se combina el análisis visual con el análisis estadístico mediante tests y contrastes.

Finalmente, en la Sección 2.5 se llevan a cabo las predicciones en base a los datos, que como se ha mencionado, se pueden comparar con los datos reales, y se representan gráficamente.

Capítulo 1

Fundamentos Matemáticos de las Series de Tiempo

Una serie de tiempo regular es una secuencia de observaciones de una variable tomadas a lo largo del tiempo en intervalos regulares: cada día, cada semana, cada mes... Se puede entender como una secuencia de variables aleatorias, en este caso univariantes. Las series de tiempo se supondrán generadas por procesos estocásticos.

Definición 1.1. [Proceso estocástico]: Una colección de variables aleatorias $\{X_t\}_{t\in C}$ indexadas por t y definidas en el mismo espacio de probabilidad se llama proceso estocástico, donde C es un conjunto arbitrario. Una observación del proceso estocástico se conoce como realización o trayectoria del mismo. (Véase [1, Sección 1.2]).

Definición 1.2. [Series de Tiempo]: Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, una serie de tiempo (estocástica) es una realización del proceso estocástico $\{X_t\}_{t\in C}$, donde cada $X_t : \Omega \to \mathbb{R}$ representa el valor del proceso en el instante t. (Esto se define en [1, Sección 1.2]).

Como se puede ver en [1, Sección 1.2], en la práctica se dispone de una única realización concreta, es decir, se observa una única trayectoria temporal, correspondiente a un experimento o sistema específico $\omega_0 \in \Omega$. Es decir, se dispone de una muestra temporal

$$\{x_t\}_{t\in C} = \{X_t(\omega_0)\}_{t\in C}$$

que representa el proceso, y sobre la que se realiza inferencia estadística. Para analizar una serie de tiempo, es conveniente representarla gráficamente. Se representa cada observación x_t en el eje Y frente al valor t en el eje X, dando lugar a una gráfica de puntos del tipo (t, x_t), que se unen mediante segmentos para dar una visión de la evolución de la serie en el tiempo. Esto es lo que se conoce como gráfico secuencial.

La mayor parte de series de tiempo podrían ser observadas en cualquier punto del tiempo, pero la imposibilidad de recolectar todos estos datos temporales hace que se trabajen como series discretas de puntos igualmente espaciados en el tiempo. Por ello, y aunque en general un proceso estocástico puede estar indexado por un conjunto arbitrario C, en series de tiempo regulares es suficiente con suponer $C = \mathbb{Z}$.

Una serie suele descomponerse en las siguientes componentes, que describen su comportamiento a lo largo del tiempo, (veáse [1, Capítulo 1]). La representación gráfica permite visualizar estas características:

- **Nivel:** Valor medio en torno al cual varían las observaciones de la serie. Si el nivel aumenta/decrece con el tiempo, se dice que hay tendencia creciente/decreciente.
- **Tendencia:** Comportamiento general a largo plazo de la serie. Indica si el nivel tiende a aumentar, disminuir o se mantiene constante a lo largo del tiempo.
- **Estacionalidad:** Patrón repetitivo que aparece en la serie a intervalos fijos, por ejemplo mensuales, trimestrales o anuales. Estos patrones se corresponden con efectos cíclicos asociados a meses o estaciones.
- **Ruido:** Variaciones aleatorias sin un patrón claro.

Además, una serie de tiempo puede presentar heterocedasticidad: cuando la variabilidad no es constante a lo largo del tiempo, depende de factores como el nivel de la serie o el número de observaciones.

1.1 Definiciones teóricas.

En este trabajo, se dispone de una serie de tiempo y se quiere conocer el proceso estocástico que la ha generado para así entender la dinámica de la serie y hacer futuras predicciones.
Siendo $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estocástico real, se definen las principales propiedades de los procesos estocásticos, que permitirán caracterizar y modelizar adecuadamente la serie observada.

Definición 1.3. [Función de medias]: La función de medias se define como

$$\mu_t = \mathbb{E}[X_t] = \int_{-\infty}^{\infty} x f_t(x)\, dx \tag{1.1}$$

siempre que exista, donde $f_t(x)$ es la función de densidad de cada variable X_t, para todo $t \in \mathbb{Z}$. Es una medida de posición central de X_t. Esta definición viene dada en [1, Definición 1.1].

Definición 1.4. [Función de varianzas]: Siguiendo a [1, Capítulo 1], la función de varianzas se define como

$$\sigma_t^2 = Var(X_t) = \mathbb{E}[(X_t - \mu_t)^2] \quad \text{para todo} \quad t \in \mathbb{Z} \tag{1.2}$$

Es una medida de dispersión de X_t respecto de la media.

La dependencia lineal entre dos variables de una serie, X_s y X_t, se puede evaluar numéricamente a través de la covarianza y correlación. Asumiendo que la varianza es finita, se tiene la siguiente definición:

Definición 1.5. [Función de autocovarianza]: La función de autocovarianza se define como el momento del producto

$$\gamma(s,t) = Cov(X_s, X_t) = \mathbb{E}[(X_s - \mu_s)(X_t - \mu_t)] \tag{1.3}$$

para todo $s, t \in \mathbb{Z}$, siendo μ_s, μ_t las medias de X_s y X_t, respectivamente. Se tiene que $\gamma(t,t) = \mathbb{E}[(X_t - \mu_t)^2] = Var(X_t)$. Esta función se define en [1, Definición 1.2].

A continuación se define la función de autocorrelación, muy importante en el análisis de series de tiempo, ya que es un tipo de covarianza estandarizada, que tiene sus valores entre el 1 y el -1, cuando la correlación es totalmente lineal con pendiente positiva o negativa, respectivamente.

Definición 1.6. [Funcion de autocorrelación simple (ACF)]: La función de autocorrelación se define como

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} = \frac{\gamma(s,t)}{\sigma_s \sigma_t} \tag{1.4}$$

para todo $s, t \in \mathbb{Z}$. (Véase [1, Definición 1.3]).

La función de autocorrelación mide la asociación lineal que tiene X_t con respecto a X_s. Por la desigualdad de Cauchy-Schwartz, se tiene que $-1 \leq \rho(s,t) \leq 1$.

Se define ahora la función de autocorrelación parcial, que mide la correlación lineal directa entre X_s y X_t, eliminando el efecto lineal de las variables observadas en los instantes de tiempo comprendidos entre s y t. Esto se consigue restando a cada variable su correspondiente regresión sobre las variables intermedias:

Definición 1.7. [Función de autocorrelación parcial (PACF)]:

$$\alpha(s,t) = Corr(X_s - \hat{X}_s, X_t - \hat{X}_t) = \frac{Cov(X_s - \hat{X}_s, X_t - \hat{X}_t)}{\sqrt{Var(X_s - \hat{X}_s) Var(X_t - \hat{X}_t)}} \quad \text{para todo} \quad s, t \in \mathbb{Z} \tag{1.5}$$

siendo $\hat{X}_s$ la regresión de X_s sobre el conjunto de variables intermedias entre X_s y X_t: $\{X_{s+1}, X_{s+2}, \ldots, X_{t-1}\}$, y lo mismo para $\hat{X}_t$. Se asume que $s < t$, el caso contrario se trata de forma análoga. (Véase [1, Sección 2.3]).

Dado que se tiene una única realización del proceso estocástico, no es posible calcular estas medidas características, pues se basan en esperanzas matemáticas, promedios sobre infinitas realizaciones del proceso estocástico. Se necesita suponer que la serie de tiempo es estacionaria, y sobre esta suposición calcular las medidas muestrales de la serie (como se expone en [1, Sección 1.5]).

1.2 Series de tiempo estacionarias

Se introduce en esta sección el concepto de *estacionariedad*. Considerando una serie de tiempo como una colección de T variables aleatorias (T arbitrario pero fijo) en $t_1, t_2, \ldots, t_T$ puntos arbitrarios de tiempo, la función de distribución conjunta viene dada por:

$$F_{t_1,t_2,\ldots,t_T}(c_1, c_2, \ldots, c_T) = \mathbb{P}(X_{t_1} \leq c_1, X_{t_2} \leq c_2, \ldots, X_{t_T} \leq c_T) \tag{1.6}$$

En series de tiempo, se le llama retardo o "lag" al desplazamiento temporal de una variable respecto de su pasado, es decir, X_{t-h} es el valor de X_t desplazado h unidades de tiempo en el pasado. Como en este caso la serie está indexada en $\mathbb{Z}$, el retardo h pertenece a $\mathbb{Z}$.

Definición 1.8. [Estacionariedad estricta]: Un proceso estocástico $\{X_t\}$ es *estrictamente estacionario* si para todo $T \in \mathbb{N}$, para todo conjunto de tiempos $t_1, t_2, \ldots, t_T \in \mathbb{Z}$ y para todo retardo $h \in \mathbb{Z}$, la distribución conjunta no cambia al desplazar los datos en el tiempo, es decir

$$F_{t_1,t_2,\ldots,t_T}(c_1, c_2, \ldots, c_T) = F_{t_1+h,t_2+h,\ldots,t_T+h}(c_1, c_2, \ldots, c_T) \tag{1.7}$$

para todo $c_1, \ldots, c_T$. (Para más detalles, véase [1, Definición 1.6]).

Esta definición implica que las propiedades estadísticas de la serie no dependen del instante de tiempo en concreto, sino de la relación temporal entre las observaciones. Como esta condición debe cumplirse para cualquier número de observaciones T, se estudian casos particulares de los que se extraen las siguientes implicaciones de la estacionariedad estricta:

Caso T = 1:
La condición anterior se reduce a la igualdad entre las funciones de distribución

$$F_t(c) = F_{t+h}(c) \quad \text{para todo} \quad t \in \mathbb{Z}, h \in \mathbb{Z}, c \in \mathbb{R}.$$

Por lo tanto, todas las variables X_t tienen la misma distribución marginal, y por tanto la esperanza matemática es constante a lo largo del tiempo

$$\mathbb{E}[X_t] = \mathbb{E}[X_{t+h}] = \mu \quad \text{para todo} \quad t, h \in \mathbb{Z}.$$

La varianza también es la misma para ambas variables

$$Var(X_t) = Var(X_{t+h})$$

lo que implica que la varianza de la serie $\sigma_t^2 = \sigma^2$ también es constante para todo $t \in \mathbb{Z}$.

Caso T = 2:
La distribución conjunta para cada subconjunto de dos variables es la misma que para dicho conjunto desplazado un retardo h

$$F_{t,s}(c_1, c_2) = F_{t+h,s+h}(c_1, c_2)$$

Por lo tanto, la función de autocovarianza es la misma para los dos conjuntos de variables,

$$\gamma(t, s) = Cov(X_t, X_s) = Cov(X_{t+h}, X_{s+h}) = \gamma(t + h, s + h)$$

Esto significa que la autocovarianza no depende de los puntos temporales t y s, depende únicamente de la separación o retardo h entre ellos, $h = |t - s|$. Para simplificar, se escribe

$$\gamma_h = Cov(X_t, X_{t+h})$$

De igual manera, la función de autocorrelación se puede escribir

$$\rho_h = \frac{\gamma_h}{\gamma_0}$$

Estas implicaciones aparecen detalladas en [1, Sección 1.4].

Dado que es difícil de evaluar la estacionariedad en sentido estricto y manejar la función de distribución conjunta de T variables aleatorias, se da una definición de estacionariedad en un sentido menos fuerte que no requiere conocer la distribución.

Definición 1.9. [Estacionariedad débil]: Según [1, Definición 1.7], un proceso estocástico $\{X_t\}$ se dice estacionario en sentido débil si es un proceso de varianza finita que cumple

i. La función media μ_t es constante, no depende del tiempo t, se escribe μ.

ii. La varianza σ_t^2 es constante para todo t, se escribe σ^2

iii. La función de autocovarianza $\gamma(s, t)$ no depende del tiempo, solo de la diferencia h entre s y t, se escribe γ_h.

Mientras que la estacionariedad estricta implica que la distribución conjunta sea invariante ante desplazamientos de tiempo, la estacionariedad débil se basa en momentos de orden uno y dos.

A partir de ahora, se usará el término *estacionariedad* para la estacionariedad en sentido débil, y si una serie es estrictamente estacionaria, se usará el término estrictamente estacionaria.

Proposición 1.10. *Una serie estacionaria en sentido estricto con varianza finita, es también estacionaria. El recíproco no es cierto, una serie estacionaria no es, por lo general, estrictamente estacionaria. Una condición para que una serie estacionaria lo sea también de forma estricta es que esté generada por un proceso gaussiano (véase [2, Capítulo 2]).*

Definición 1.11. [Proceso Gaussiano]: Un proceso $\{X_t\}$ se dice Gaussiano si para cualquier $T \in \mathbb{N}$ y cualquier conjunto de puntos temporales $t_1, t_2, \ldots, t_T$, el vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_T})$ tiene distribución normal multivariante. (Esta definición puede encontrarse en [1, Definición 1.13]).

1.3 Definiciones muestrales

Las funciones descriptivas de las series de tiempo, como la función de medias, la función de autocovarianza o las funciones de autocorrelación, son herramientas fundamentales cuando se conoce el proceso estocástico que genera la serie, es decir, cuando se dispone de información sobre las distribuciones de las variables aleatorias X_t.
Sin embargo, en la práctica, lo que generalmente se observa es una única realización del proceso, el cual es desconocido. Por tanto, estas propiedades deben estimarse a partir de los datos disponibles.
En este contexto, suponer que la serie es estacionaria resulta fundamental, ya que permite asumir que el comportamiento estadístico de la serie no cambia con el tiempo, y se puedan usar los diferentes instantes de una sola realización como si fueran muestras de este mismo comportamiento (véase [1, Sección 1.5]). Esto explica el uso de la media muestral como estimación de la media poblacional, y la estimación de las funciones de autocovarianza y autocorrelación.
Por tanto, si una serie es estacionaria, la función media μ es constante, y se puede estimar mediante la media muestral:

Definición 1.12. [Media muestral]: Teniendo una realización $\{x_t\}$ del proceso $\{X_t\}$, con T variables (T arbitrario), la media muestral es

$$\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t \tag{1.8}$$

Esta definición se encuentra en [1, Sección 1.5].

Las funciones de autocovarianza y autocorrelación simple y parcial se aproximan por las respectivas funciones muestrales. Bajo la suposición de estacionariedad, estas definiciones no

dependen de los puntos temporales s y t, sino que dependen del retardo h entre variables, por tanto se definen para una variable X_t y su desplazada en el tiempo un retardo h.

Definición 1.13. [Función de autocovarianza muestral]: Sea $\{x_t\}_{t=1}^{T}$ una realización de tamaño T del proceso estocástico $\{X_t\}$. Se define la autocovarianza muestral como:

$$\hat{\gamma}_h = \frac{1}{T}\sum_{t=1}^{T-h}(x_{t+h} - \bar{x})(x_t - \bar{x}) \tag{1.9}$$

donde h puede ser $h = 0, 1, \ldots, T-1$. (Esta definición viene dada en [1, Definición 1.14]).

Definición 1.14. [Función de autocorrelación simple muestral]: Sea $\{x_t\}_{t=1}^{T}$ una realización de tamaño T del proceso estocástico $\{X_t\}$. Se define la autocorrelación muestral como:

$$\hat{\rho}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0} \tag{1.10}$$

(Véase [1, Definición 1.15]).

La función de autocorrelación parcial muestral, en el retardo h, denotada por $\hat{\alpha}_h$, se define como el coeficiente estimado del término X_{t-h} en la regresión lineal de X_t sobre sus h retardos, estimada por mínimos cuadrados:

$$X_t = \beta_{h0} + \beta_{h1}X_{t-1} + \beta_{h2}X_{t-2} + \cdots + \beta_{hh}X_{t-h} + \varepsilon_t, \tag{1.11}$$

Entonces:

Definición 1.15. [Función de autocorrelación parcial muestral]: Sea $\{x_t\}_{t=1}^{T}$ una realización de tamaño T del proceso estocástico $\{X_t\}$. Se define la autocorrelación parcial muestral como:

$$\hat{\alpha}_h = \hat{\beta}_{hh}, \tag{1.12}$$

es decir, el coeficiente del último retardo en la regresión de la Ecuación 1.11. Esta definición se encuentra en [2, Sección 2.3].

1.4 Más definiciones

Aquí se dan algunas definiciones que harán falta más adelante.

Definición 1.16. [Ruido blanco]: El ruido blanco es un conjunto de variables aleatorias indexadas por t, $\{a_t\}_t$, incorrelacionadas entre si, con media 0 y varianza constante σ_a^2. Si a_t es ruido blanco se denota $a_t \sim wn(0, \sigma_a^2)$ ("white noise"). Si además las variables son independientes y tienen la misma distribución, se dice que es ruido blanco independiente e idénticamente distribuido (iid). Esta definición se encuentra en [1, Sección 1.2].

Observación 1.17. Un ruido blanco particular es el ruido blanco gaussiano, que consiste en un conjunto de variables aleatorias independientes e idénticamente distribuidas con distribución normal, se denota $a_t \sim N(0, \sigma_a^2)$

El ruido blanco es importante en series de tiempo porque representa un proceso estocástico sin estructura temporal: sus valores son incorrelacionados, tienen media cero y varianza constante. Muchos modelos, como los ARIMA, tienen como objetivo ajustar la serie observada de forma que los residuos se comporten como ruido blanco. Esto indica que el modelo ha captado toda la dependencia temporal presente en los datos.

Observación 1.18. El ruido blanco es un proceso estacionario, ya que

$$\mathbb{E}[a_t] = 0 \text{ para todo } t$$

$$\gamma_h = \begin{cases} \sigma_a^2, & \text{si } h = 0 \\ 0, & \text{si } h \neq 0 \end{cases}$$

Definición 1.19. [Paseo aleatorio]: Es un proceso con la siguiente estructura

$$X_t = X_{t-1} + a_t$$

donde a_t es ruido blanco y X_0 es un valor dado. También se puede escribir de la siguiente forma

$$X_t = X_0 + \sum_{j=1}^{t} a_j$$

El paseo aleatorio es un proceso no estacionario, ya que su varianza aumenta con el tiempo, $\text{Var}(X_t) = \text{Var}(X_0) + t \cdot \sigma_a^2$. (Véase [1, Sección 1.2]).

Definición 1.20. [Proceso lineal]: Un proceso estocástico X_t se dice que es un proceso lineal si puede escribirse como combinación lineal infinita de términos de ruido blanco. Se representa de la siguiente manera:

$$X_t = c + \sum_{j=-\infty}^{\infty} \psi_j a_{t-j} \qquad \text{con} \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

donde $\{a_t\}$ es un proceso de ruido blanco, ψ_j son coeficientes reales que cumplen la condición de convergencia absoluta y $c \in \mathbb{R}$ es una constante. (Véase[1, Definición 1.12]).

Observación 1.21. Un proceso lineal es estacionario. Su función de autocovarianza, siendo $\{a_t\}_t$ el ruido blanco generador del proceso, es la siguiente:

$$\gamma_h = \sigma_a^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}$$

donde σ_a^2 es la varianza del ruido blanco.

Este tipo de procesos pueden incluir dependencias de valores futuros. Un tipo de proceso lineal que solo depende de los valores pasados y presentes de ruido blanco es el proceso lineal causal.

Definición 1.22. [Proceso lineal causal]: Es un proceso estocástico que se puede expresar como una combinación lineal infinita de términos de ruido blanco con $\psi_j = 0$ para $j < 0$. Se representa como

$$X_t = c + \sum_{j=0}^{\infty} \psi_j a_{t-j} \qquad \text{con} \qquad \sum_{j=0}^{\infty} |\psi_j| < \infty$$

con $c \in \mathbb{R}$ constante y a_t ruido blanco. Esta definición viene dada en [1, Sección 1.2].

Capítulo 2

Modelización de la serie

Una vez establecidas las definiciones teóricas y muestrales de las series de tiempo, en este capítulo se presentan los modelos más habituales para modelizar series macroeconómicas. El objetivo de este capítulo es construir un modelo que explique los datos disponibles y que permita hacer predicciones precisas en base a ellos.
La modelización de esta serie se hará siguiendo la metodología Box-Jenkins, método empleado para el análisis y modelización de series temporales. Este método utiliza modelos de la familia ARIMA, que captan dependencia regular y estacional, y utilizan técnicas llamadas diferencias para transformar la serie en estacionaria. Este proceso se desenvuelve en tres etapas, que son: la identificación, estimación y validación de estos modelos, asegurando que se cumplan las suposiciones necesarias, como se puede ver en [2, Sección 14.4].

- **Identificación:** Se analiza la serie temporal para determinar si es estacionaria, identificar posible tendencia y estacionalidad y transformarla para aplicar uno de los modelos de la familia ARIMA. Especificar el modelo ARIMA elegido que se ajuste a los datos así como determinar sus órdenes, p, d, q, P, D y Q.

- **Estimación:** Se estiman los parámetros del modelo propuesto para que este se ajuste lo mejor posible a los datos observados. Esto se hace generalmente mediante máxima verosimilitud.

- **Validación:** El objetivo principal es evaluar si el modelo ajustado es adecuado. En esta etapa se comprueba si el modelo cumple los supuestos teóricos: que los residuos se comporten como ruido blanco, es decir, que sean incorrelacionados, con media cero y varianza constante, esto significa que el modelo capta toda la dependencia temporal de los datos. También se comprueba si tienen distribución normal, pues esto permite construír intervalos de predicción normales.

Esta metodología funciona de forma iterativa, si en la etapa de validación se observa que no se cumplen las suposiciones o que el modelo no se ajusta a los datos, se vuelve a la etapa de identificación y se especifica otro modelo (véase [2, Capítulo 14]). Una vez se haya validado que el modelo es adecuado, se realizan las predicciones, y se construyen los intervalos de confianza correspondientes.

2.1 Modelos ARIMA regulares

Antes de empezar a modelizar la serie de tiempo, es necesario presentar los modelos que utiliza este método para modelar series temporales. Se trata de la familia de modelos ARIMA (Autoregressive Integraded Moving Average models). Esta familia combina componentes autorregresivos (AR) y de medias móviles (MA) para capturar la dependencia temporal de los datos a través de sus observaciones y errores pasados, esta combinación de modelos son los ARMA, y tanto los ARMA como los AR y MA se utilizan para modelizar series estacionarias. En caso de que la serie no sea estacionaria se incorpora la diferenciación (I), dando lugar a los modelos ARIMA. Cada uno de estos componentes tiene un orden que indica el número de parámetros utilizados: en el modelo AR, el orden p indica el número de componentes autorregresivas; en el modelo MA, el orden q representa el número de términos de media móvil, y en los modelos ARIMA, el parámetro d indica el número de veces que se diferencia la serie para alcanzar estacionariedad. Este desarrollo se detalla en [1, Capítulo 3].

2.1.1 Modelos autorregresivos (AR)

Los modelos autorregresivos (AR) de orden p se utilizan para modelar procesos estacionarios en los que la variable X_t se expresa como combinación lineal de sus p valores pasados:
$X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ más un término de error aleatorio, que se modeliza como ruido blanco.

Definición 2.1. [Modelo autorregresivo de orden p]: Un modelo autorregresivo de orden p, AR(p), se define como:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t$$

donde $a_t \sim wn(0, \sigma_a^2)$, y $c, \phi_1, \phi_2, \ldots, \phi_p$ son constantes ($\phi_p \neq 0$).

Con el fin de expresar el modelo de forma más concisa, se introduce el operador retardo u operador *backward*:

Definición 2.2. [Operador backward]: El operador retardo o backward se define como

$$BX_t = X_{t-1}$$

Si se aplica dos veces, da lugar a $B^2X_t = B(BX_t) = BX_{t-1} = X_{t-2}$, lo que se generaliza a un número k de veces como

$$B^k X_t = X_{t-k}$$

Usando esta notación, el modelo AR(p) se escribe como:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots \phi_p B^p) X_t = c + a_t$$

o, de forma más general:

$$\phi_p(B) X_t = c + a_t \tag{2.1}$$

donde $\phi_p(B)$ es el operador autorregresivo de orden p, que se define:

Definición 2.3. [Operador autorregresivo $\phi_p(B)$]:

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

Además del operador, se define también el polinomio autorregresivo, útil para el análisis de las propiedades del proceso estocástico:

Definición 2.4. [Polinomio AR]: El polinomio autorregresivo se definen como

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \tag{2.2}$$

con $\phi_p \neq 0$ y $z \in \mathbb{C}$.

Este polinomio es clave en el estudio de la estacionariedad de los procesos autorregresivos, así como en el análisis de modelos ARMA.

Observación 2.5. Los modelos AR(p) se emplean para modelizar procesos estacionarios. No obstante, un proceso generado por un modelo autorregresivo es estacionario si y solo si las raíces del polinomio $\phi(z)$ están fuera del círculo unidad, es decir, tienen módulo mayor que uno.

2.1.2 Modelos de medias móviles (MA)

Los modelos de medias móviles MA de orden q modelan procesos estacionarios en los que la variable X_t se expresa como combinación lineal de q términos de ruido blanco pasado.

Definición 2.6. [Modelo de medias móviles de orden q]: El modelo de medias móviles de orden q, MA(q), es el siguiente

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q}$$

donde $a_t \sim wn(0, \sigma_a^2)$ y $c, \theta_1, \theta_2, \dots, \theta_q$ son parámetros ($\theta_q \neq 0$).

Al igual que el modelo autorregresivo, el modelo de medias móviles de orden q tiene una representación general haciendo uso del operador backward

$$X_t = c + (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)a_t \tag{2.3}$$

donde, definiendo el operador de medias móviles de orden q, $\theta_q(B)$, de la siguiente manera

Definición 2.7. [Operador de medias móviles]: El operador de medias móviles de orden q es una expresión de la siguiente forma:

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

se puede escribir de manera compacta y general como

$$X_t = c + \theta_q(B)a_t.$$

Al igual que ocurre en el caso autorregresivo, se define el polinomio de medias móviles como función polinómica algebraica con raíces complejas.

Definición 2.8. [Polinomio MA]: El polinomio de medias móviles se define como

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \tag{2.4}$$

con $\theta_q \neq 0$ y $z \in \mathbb{C}$.

Observación 2.9. Los procesos generados por un modelo de medias móviles siempre son estacionarios.

2.1.3 Modelos autorregresivos de medias móviles (ARMA)

Un modelo ARMA(p, q) es la combinación un modelo autorregresivo de orden p, AR(p), y un modelo de medias móviles de orden q, MA(q); y modeliza la dependencia temporal regular entre una variable X_t y sus observaciones y errores pasados.

Definición 2.10. [Modelo ARMA(p, q)]: El modelo autorregresivo de medias móviles, ARMA, representa el siguiente proceso estocástico

$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} \tag{2.5}$$

siendo $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ constantes, con $\phi_p \neq 0$, $\theta_q \neq 0$ y $a_t \sim wn(0, \sigma_a^2)$.

El modelo ARMA (p, q) se puede representar de forma más compacta usando los operadores autorregresivo y de medias móviles

$$\phi_p(B)X_t = c + \theta_q(B)a_t \tag{2.6}$$

Observación 2.11. Aparte de la definición general, los polinomios $\phi(z)$ y $\theta(z)$ no deben tener factores comunes, porque de tenerlos, unos términos podrían cancelar a otros y el modelo podría ser más simple de lo que aparenta en principio, dando lugar a una interpretación incorrecta del modelo.

Observación 2.12. Los modelos ARMA representan procesos estacionarios cuando el polinomio autorregresivo, $\phi(z)$ tiene sus raíces fuera del círculo unidad, es decir, si $\phi(z) \neq 0$ para $|z| \leq 1$.

Si el proceso es estacionario, entonces tiene media constante μ, y se cumple que

$$\mathbb{E}[X_t] = \mu \quad \text{para todo} \quad t.$$

Aplicando la esperanza al modelo, se tiene

$$\mu = \mathbb{E}[X_t] = c + \phi_1\mu + \cdots + \phi_p\mu$$

y por tanto, en este caso la relación entre la constante c del modelo y la media μ es la siguiente

$$c = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p). \tag{2.7}$$

Una propiedad importante de los modelo ARMA es la *causalidad*, definida a continuación.

Definición 2.13. [ARMA causal]: Un modelo ARMA se dice causal si el proceso estocástico que define se puede escribir como un proceso lineal causal (combinación lineal infinita de ruido blanco pasado)

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} = \psi(B)a_t \quad \text{con la condición} \quad \sum_{j=0}^{\infty} |\psi_j| < \infty \tag{2.8}$$

siendo $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ y $\psi_0 = 1$. A esta representación se le llama MA(∞).

En otras palabras, un modelo ARMA(p, q) se dice causal si admite una representación MA(∞).

Para saber si un proceso es causal, se tiene la siguiente proposición:

Proposición 2.14. ***Condición para la causalidad de un proceso ARMA*** (p, q)***.*** *Un modelo ARMA*(p, q) *es causal si y solo si* $\phi(z) \neq 0$ *para* $|z| \leq 1$*. Los coeficientes del proceso lineal dado en* (2.8) *se pueden determinar resolviendo*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \qquad |z| \leq 1. \tag{2.9}$$

En otras palabras, un proceso ARMA es causal cuando las raíces del polinomio autorregresivo $\phi(z)$ están fuera del círculo unidad, es decir, si $\phi(z) = 0$ solo cuando $|z| > 1$.

Observación 2.15. En estos modelos, estacionariedad equivale a causalidad.

En los modelos ARMA, la componente de medias móviles puede admitir múltiples representaciones equivalentes, es decir, diferentes conjuntos de coeficientes θ_j que generen el mismo proceso estocástico. Para garantizar la unicidad de la representación del modelo, se impone una condición conocida como *invertibilidad*, que se define a continuación.

Definición 2.16. [ARMA invertible]: Un modelo ARMA es invertible si el proceso estocástico que lo define se puede representar como

$$\pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = a_t \quad \text{con la condición} \quad \sum_{j=0}^{\infty} |\pi_j| < \infty \tag{2.10}$$

siendo $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$. Esto se llama representación invertible o AR(∞). Por tanto, un proceso es invertible si admite una representación AR(∞).

Para saber si un proceso es invertible, se tiene la siguiente proposición:

Proposición 2.17. ***Condición de invertibilidad de un proceso ARMA*** (p, q)***.*** *Un modelo ARMA es invertible si y solo si* $\theta(z) \neq 0$ *para* $|z| \leq 1$. *Los coeficientes* π_j *de* $\pi(B)$ *dados en* (2.10) *se determinan resolviendo*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \qquad |z| \leq 1. \tag{2.11}$$

Es decir, un proceso ARMA es invertible cuando las raíces de $\theta(z)$ están fuera del círculo unitario, esto es, cuando $\theta(z) = 0$ solo si $|z| > 1$.

Observación 2.18. Un modelo ARMA se dice *gaussiano* si el ruido blanco del proceso es gaussiano, es decir, si $a_t \sim \mathcal{N}(0, \sigma_a^2)$.

2.1.4 Modelos ARIMA(p, d, q)

Como se ha visto, los modelos ARMA tienen gran capacidad para modelizar procesos estacionarios. Cuando en la práctica se tiene un proceso no estacionario, se aplica un modelo ARIMA regular, que consiste en aplicar diferencias regulares para transformar la serie en estacionaria. La diferencia regular es una técnica llevada a cabo mediante el operador diferencia:

Definición 2.19. [Operador diferencia de orden uno]: Se define como sigue

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}$$

La primera diferencia elimina la tendencia lineal. Si no es suficiente, se aplica una segunda diferencia, que elimina la tendencia cuadrática. Siguiendo esta notación la diferencia de orden dos sería

$$\nabla^2 X_t = (1-B)^2 X_t = X_t - 2X_{t-1} + X_{t-2}.$$

Si se diferencia la serie un número d de veces

Definición 2.20. [Operador diferencia de orden d]: La diferencia de orden d se define como

$$\nabla^d X_t = (1-B)^d X_t$$

con $d \in \mathbb{Z}$. Diferenciar una serie d veces neutraliza tendencias polinómicas de grado d.

El número de veces que se aplica la operación diferencia hasta alcanzar la estacionariedad se denota por d en el modelo ARIMA(p, d, q).

Por tanto los modelos ARIMA se definen de la siguiente manera

Definición 2.21. [Modelos ARIMA (p, d, q)]: El modelo ARIMA (Autoregressive Integrated Moving Average) representa un proceso estocástico de la forma

$$\phi_p(B)(1-B)^d X_t = c + \theta_q(B) a_t \tag{2.12}$$

Observación 2.22. Existe un inverso para el operador backward, es el operador de avance o *forward*.

Definición 2.23. [Operador forward]: El operador forward se define como

$$F X_t = X_{t+1}$$

Normalmente se aplica un ARIMA regular para modelar una serie temporal cuando esta presenta tendencia. Para otros tipos de no estacionariedad, la aplicación de diferencias regulares puede no ser suficiente o requerir adaptaciones. La tendencia que los modelos ARIMA son capaces de capturar suele ser de alguno de estos dos tipos:

- **Tendencia determinista:** puede ser lineal

  $$X_t = \beta_0 + \beta_1 t + Y_t$$

  siendo $\{Y_t\}_t$ un proceso estacionario; o polinómica

  $$X_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k + Y_t$$

  En estos casos, la tendencia puede eliminarse mediante la aplicación de diferencias o mediante regresión.

- **Tendencia estocástica:** la tendencia evoluciona aleatoriamente con el tiempo. Un ejemplo es el paseo aleatorio

$$X_t = X_{t-1} + a_t$$

donde $a_t \sim wn(0, \sigma_a^2)$. Aunque la media $\mathbb{E}[X_t]$ es constante, la varianza aumenta con el tiempo, dando lugar a un comportamiento no estacionario y una tendencia estocástica que se elimina aplicando diferencias.

Los modelos ARIMA permiten modelar una gran variedad de series temporales. En las siguientes secciones se desarrollan las tres etapas de la metodología Box-Jenkins, orientadas a seleccionar, ajustar y validar el modelo adecuado, a la vez que se aplican a la serie de tiempo elegida.

2.2 Identificación

Esta etapa consiste en determinar la estructura del modelo adecuado para la serie temporal. Para ello, se determina si la serie es estacionaria o hay que aplicar técnicas para que lo sea. Una vez la serie es estacionaria, se analiza la estructura de correlación mediante las funciones de autocorrelación parcial y simple para determinar los órdenes del modelo.
En este proceso se combina el análisis visual, representando el gráfico secuencial de la serie y los gráficos de sus funciones de autocorrelación simple y parcial muestrales (ACF y PACF, respectivamente); con el análisis estadístico a través de contrastes de hipótesis y tests estadísticos.
Los datos originales fueron descargados en formato Excel desde el Instituto Gallego de Estadística (IGE). Estos datos contienen información acerca del año, mes y número de pasajeros por mes, con datos desde enero de 1980 a mayo de 2025, aunque, como se ha indicado anteriormente, se usarán los datos hasta 2018 por el impacto que tuvo la pandemia en los datos del transporte aéreo. Se transforman los datos originales en RStudio para obtener la variable de interés, que es el número de pasajeros.

Se representan gráficamente los datos en el tiempo en la figura Figura 2.1. La variable del número de pasajeros mensuales se convierte a formato `ts` para su análisis como serie de tiempo.

```
pasajeros_ts = ts(pasajeros, start = c(1980, 1), frequency = 12)
```

La serie de interés es $\{X_t\}_t$, donde X_t es el número de pasajeros del transporte aéreo en el mes t, con $t = 1, \ldots, 468$.

Se puede observar en el gráfico secuencial que no es estacionaria: la varianza no es constante, crece según aumenta el nivel de la serie. Además, tiene tendencia creciente y una clara estaciona-

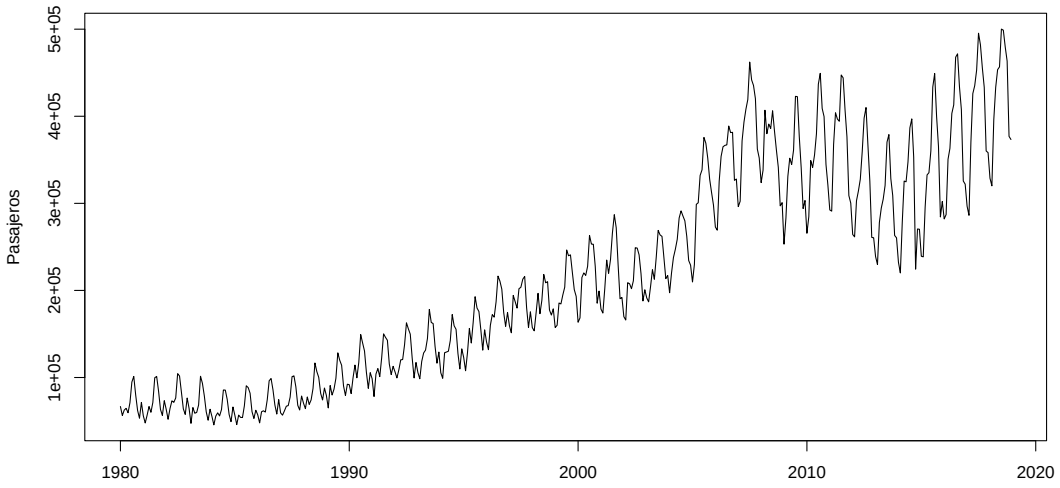

Figura 2.1: Pasajeros totales del tráfico aéreo (1980-2018)

lidad. Con el objetivo de eliminar la variabilidad no constante, se aplica a la serie una transformación logarítmica (como se puede ver en [1, Sección 2.2]).

$$Y_t = \log(X_t) \tag{2.13}$$

Se representa gráficamente la serie transformada en la Figura 2.2.

La serie logarítmica de pasajeros $\{Y_t\}$ presenta una varianza ya estabilizada, aunque sigue presentando tendencia y estacionalidad. Se representa a continuación de su gráfica secuencial su gráfica de autocorrelación simple para comprobar la presencia de tendencia.

La ACF, Figura 2.3, decae lentamente con valores próximos a 1, lo cual puede indicar no estacionariedad, aunque no siempre.
Se tiene por tanto una serie homocedástica, con tendencia y con una clara componente estacional, como se puede observar en el gráfico secuencial de la serie logarítmica.
Esta componente estacional puede ser dos tipos, que se definen a continuación.

- **Estacionalidad determinista:** Es un patrón repetitivo y fijo en el tiempo, no aleatorio, que aparece directamente en el nivel de la serie. Suele deberse a factores externos (como el calendario o las estaciones) y puede eliminarse aplicando una diferencia estacional. Corresponde a una tendencia estacional (para más detalle véase [2, Capítulo 8]).
- **Estacionalidad estocástica:** Es un patrón repetitivo pero aleatorio, que no afecta di-

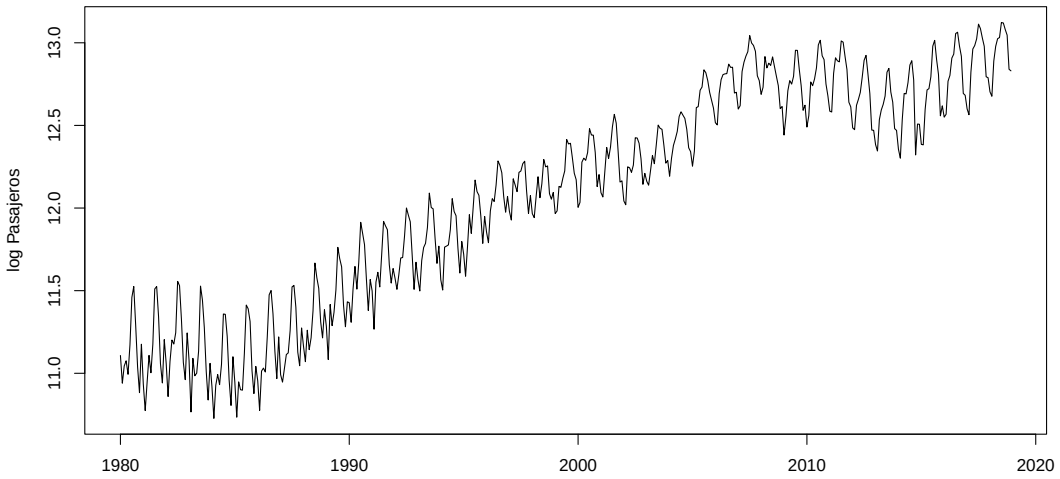

Figura 2.2: Pasajeros totales (1980-2018). Serie logarítmica

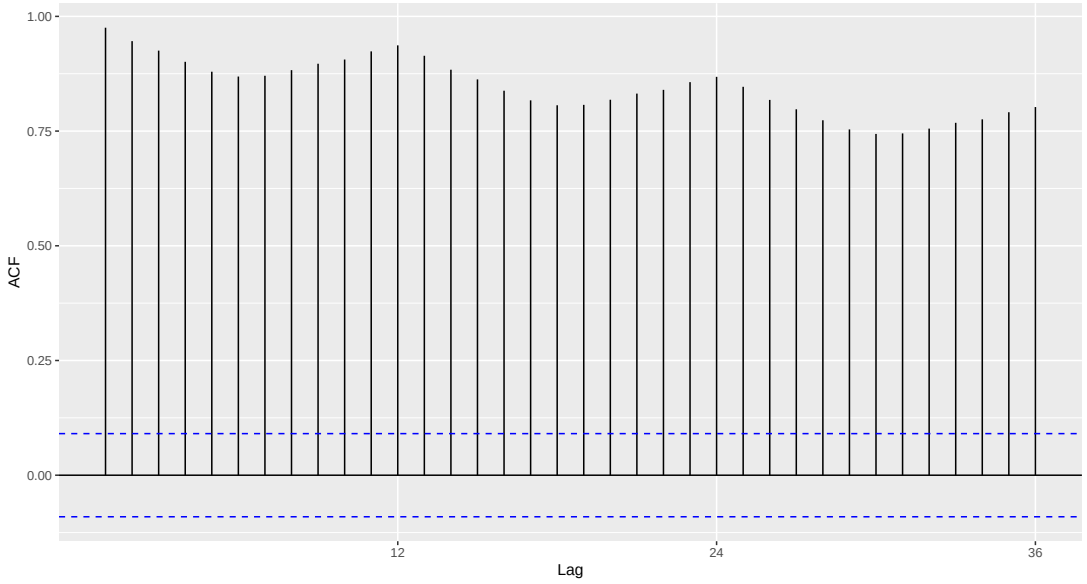

Figura 2.3: Gráfica ACF de la serie logarítmica

rectamente al nivel de la serie, sino a su estructura de dependencia temporal. Por ejemplo, el valor de la serie en un mes concreto puede estar fuertemente correlacionado con el valor del mismo mes en años anteriores. Este tipo de estacionalidad se modeliza mediante componentes autorregresivas o de media móvil estacionales (como viene dado en [2, Capítulo 8]).

Hasta ahora se han visto los modelos ARIMA regulares, que resultan útiles para modelizar series temporales con dependencia temporal regular y no estacionariedad provocada por la presencia de tendencia (determinista o estocástica). Sin embargo, estos modelos no captan la dependencia estacional entre observaciones de la serie ni la no estacionariedad provocada por esta componente estacional. Por ello, se introduce una nueva clase de modelos que extiende a la familia ARIMA, los modelos ARIMA estacionales. Estos modelos se desarrollan en [1, Sección 3.9].

2.2.1 Modelos ARMA$(P,Q)_s$

Los modelos ARMA estacionales, denotados como ARMA$(P,Q)_s$, permiten modelizar la dependencia estacional de tipo estocástico. En este caso, la variable X_t se relaciona con sus propios valores pasados y errores anteriores separados por múltiplos de un período fijo s, conocido como período estacional.

Definición 2.24. [Modelo ARMA$(P,Q)_s$]: El modelo ARMA estacional representa el siguiente proceso

$$\begin{aligned} X_t = c + \Phi_1 X_{t-s} + \Phi_2 X_{t-2s} + \cdots + \Phi_P X_{t-Ps} + a_t \\ + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \cdots + \Theta_Q a_{t-Qs} \end{aligned} \tag{2.14}$$

donde $c, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots, \Theta_Q$ son constantes, con $(\Phi_P, \Theta_Q \neq 0)$ y $a_t \sim wn(0, \sigma_a^2)$.

El modelo ARMA$(P,Q)_s$ se puede escribir de forma compacta usando el operador retardo estacional

Definición 2.25. [Operador retardo estacional]:

$$B^s X_t = X_{t-s} \tag{2.15}$$

y definiendo los siguientes operadores:

Definición 2.26. [Operador autorregresivo estacional de orden P y período s]:

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \tag{2.16}$$

Definición 2.27. [Operador de medias móviles estacional de orden Q y período s]:

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \tag{2.17}$$

Por tanto, escrito de forma compacta:

$$\Phi_P(B^s) X_t = c + \Theta_Q(B^s) a_t \tag{2.18}$$

Observación 2.28. Un ARMA$(P,Q)_s$ es un ARMA(sP, sQ) con muchos coeficientes nulos.

2.2.2 Modelos ARMA$(p,q) \times (P,Q)_s$

Se acaban de definir los ARMA estacionales, que modelizan la dependencia estacional de un proceso. En la realidad, la serie temporal suele mostrar tanto dependencia regular como estacional, por lo que se necesitan ambos modelos ya descritos: ARMA(p,q) y ARMA$(P,Q)_s$. A esta combinación de modelos ARMA regular y estacional se le llama modelo ARMA estacional multiplicativo, y se denota por $ARMA(p,q) \times (P,Q)_s$.

Observación 2.29. Por simplicidad de notación, los operadores autorregresivos regular y estacional de órdenes p y P se denotarán por $\phi(B)$ y $\Phi(B^s)$, y los polinomios de medias móviles de órdenes q y Q por $\theta(B)$ y $\Theta(B^s)$, respectivamente.

Definición 2.30. [Modelo ARMA$(p,q)\times(P,Q)_s$]: El modelo ARMA estacional multiplicativo representa el proceso estocástico

$$\phi(B)\Phi(B^s) X_t = c + \theta(B)\Theta(B^s) a_t \tag{2.19}$$

donde $\phi(B)$ y $\theta(B)$ son los operadores regulares definidos en la sección de los modelos ARIMA.

Observación 2.31. Un ARMA$(p,q)\times(P,Q)_s$ es un ARMA$(p+sP, q+sQ)$ con muchos coeficientes nulos.

2.2.3 Modelos ARIMA$(p,d,q) \times (P,D,Q)_s$

A partir de los modelos ARMA$(p,q) \times (P,Q)_s$, al incorporar operadores de diferencia, se obtiene la clase más general de modelos de series temporales: los modelos ARIMA estacionales multiplicativos, denotados como ARIMA$(p,d,q) \times (P,D,Q)_s$. Estos modelizan series que presentan tanto tendencia como estacionalidad. Aplican diferencias regulares de orden d para eliminar la tendencia determinista y diferencias estacionales de orden D para eliminar la componente estacional determinista. Una vez diferenciada la serie, el modelo resultante incluye componentes autorregresivas y de medias móviles tanto en su parte regular como estacional, que permiten capturar la dependencia temporal y estacional de tipo estocástico. Al igual que los ARIMA clásicos, las componentes no estacionarias se abordan así:

- **Tendencia:** se elimina aplicando d diferencias regulares $(1-B)^d$.
- **Estacionalidad:** se elimina aplicando D diferencias estacionales $(1-B^s)^D$.

La diferencia estacional se define como sigue

Definición 2.32. [Diferencia estacional de orden D]: Una diferencia estacional de orden D es de la forma

$$\nabla_s^D X_t = (1-B^s)^D X_t \tag{2.20}$$

siendo s el período estacional.

El número de veces que se aplica la diferencia estacional se denota por D en el modelo ARIMA$(p,d,q)\times(P,D,Q)_s$. Normalmente, con una diferencia estacional es suficiente:

$$\nabla_s X_t = (1-B^s) = X_t - X_{t-s}$$

Una vez que se haya diferenciado (regular y/o estacionalmente) la serie ya es estacionaria, y se puede modelizar su estructura mediante

- un modelo ARMA(p,q): si presenta únicamente dependencia regular.
- un modelo ARMA$(P,Q)_s$: si presenta únicamente dependencia estacional.
- un modelo ARMA$(p,q)\times(P,Q)_s$: si presenta ambos tipos de dependencia.

Por tanto, ya se puede definir formalmente el modelo ARIMA estacional multiplicativo:

Definición 2.33. [Modelo ARIMA$(p,d,q)\times(P,D,Q)_s$]: Representa un proceso de la siguiente forma

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = c + \theta(B)\Theta(B^s)a_t \tag{2.21}$$

donde los polinomios asociados $\phi(z)\Phi(z^s)$ y $\theta(z)\Theta(z^s)$ no tienen raíces de módulo < 1, para que el proceso sea estacionario e invertible, respectivamente, c representa la constante del modelo, s es el período estacional y a_t es ruido blanco.

La clase de modelos ARIMA$(p,d,q)\times(P,D,Q)_s$ generaliza todos los modelos que se han visto hasta ahora, y son muy usados en la práctica, especialmente en el análisis y modelización de series de tiempo macroeconómicas.

Aplicando los conceptos previamente descritos, se continúa con la modelización de la serie. Como ya se ha visto, la diferencia estacional es una técnica utilizada para eliminar la componente

estacional determinista de una serie. Por lo tanto se aplicará esta operación a la serie logarítmica $\{Y_t\}$, considerando un periodo estacional de $s = 12$, pues el gráfico de autocorrelación (ACF) en la Figura 2.3 muestra picos significativos en los múltiplos de este valor, algo coherente con el tipo de serie que se está tratando, pues el tráfico aéreo en un mes determinado suele estar relacionado con el tráfico aéreo de ese mismo mes en años anteriores. La serie transformada mediante diferencia estacional de orden 1 se define

$$Z_t = \nabla_{12} Y_t = Y_t - Y_{t-12} \tag{2.22}$$

y se representa gráficamente a continuación:

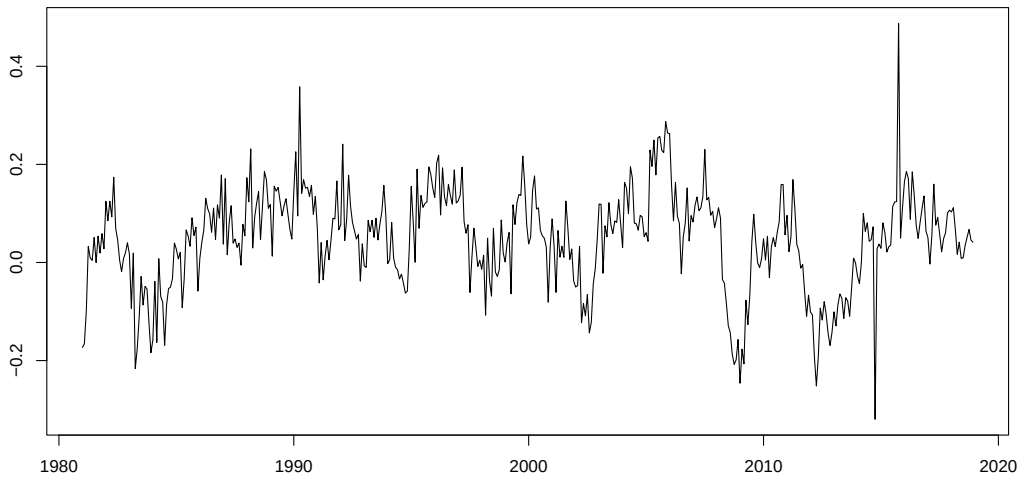

Figura 2.4: Pasajeros totales del tráfico aéreo (1980-2018). Serie logarítmica diferenciada estacionalmente

Tras aplicar la diferencia estacional, se elimina la tendencia estacional en el nivel, de modo que la serie ya no presenta un crecimiento sistemático anual. Sigue presentando cierta estructura de dependencia estacional, que se abordará en la siguiente sección.
Con todo, se ha llegado a una serie aparentemente estacionaria. Esto se comprueba mediante los tests estadísticos ADF (Augmented Dickey-Fuller) y el KPSS.

Test ADF (Augmented Dickey-Fuller)

Este test contrasta la hipótesis de raíz unitaria. Si la hipótesis nula es cierta y la serie tiene raíz de módulo 1, no es estacionaria. Esta condición se explica para modelos autorregresivos de orden 1, AR(1), aunque se generaliza para el resto de modelos. Una raíz unitaria implica que el

coeficiente del término autorregresivo es igual a uno: el polinomio autorregresivo de un AR(1) es $1 - \phi z$, y tiene raíz $z = \frac{1}{\phi}$. Por tanto, si la raíz es unitaria, el coeficiente ϕ tiene módulo 1, y la serie no es estacionaria, pues la expresión de la serie $X_t = c + X_{t-1} + a_t$ es un paseo aleatorio, y esta serie no tiene varianza constante. Por tanto, este test consiste en la hipótesis

$$H_0 : \text{la serie tiene raíz unitaria (no es estacionaria)}$$
$$H_1 : \text{la serie es estacionaria}$$

```
adf.test(dif_est_log_pasajeros_ts)


>
>  Augmented Dickey-Fuller Test
>
> data:  dif_est_log_pasajeros_ts
> Dickey-Fuller = -3.9657, Lag order = 7, p-value = 0.0108
> alternative hypothesis: stationary
```

El p-valor es 0.0108, por lo que se puede rechazar la hipótesis de raíz unitaria al 5%, e incluso al 2%, por lo que existen pruebas significativas de que la serie es estacionaria. Este test se presenta en la publicación [3, Páginas 427–431], donde se describe su construcción detallada.

Test KPSS

Este test evalúa si la serie es estacionaria en torno a una media constante. Si los residuos tienen varianza creciente, se interpreta como una señal de que no es estacionaria. Este test tiene como hipótesis nula que la serie es estacionaria, al contrario que el test ADF. Por tanto, usar ambos tests es útil para confirmar los resultados.

$$H_0 : \text{la serie es estacionaria}$$
$$H_1 : \text{la serie no es estacionaria}$$

```
kpss.test(dif_est_log_pasajeros_ts)


>
>  KPSS Test for Level Stationarity
>
> data:  dif_est_log_pasajeros_ts
> KPSS Level = 0.26027, Truncation lag parameter = 5, p-value = 0.1
```

El p-valor es 0.1, por lo que no hay pruebas suficientes para rechazar la hipótesis nula, por lo que se acepta que la serie es estacionaria. La construcción de este test viene dada en [4, Páginas 159–178].

Hasta ahora se tiene un serie a la que se le ha aplicado una transformación logarítmica para estabilizar la varianza. Dado que la serie presentaba estacionalidad anual, picos repetidos cada 12 meses, se le aplicó una diferencia estacional con el período $s = 12$, que ha transformado la serie en estacionaria, eliminando la estacionalidad determinista(en nivel) y la tendencia, y se ha comprobado mediante dos tests estadísticos. El modelo que se tiene es por tanto, un

$$\text{ARIMA}(p, 0, q) \times (P, 1, Q)_{12}$$

Ahora se tratará de encontrar los órdenes de dependencia regular, p y q, y los de dependencia estacional P y Q. Es decir, ver qué tipo de modelo ARMA estacional multiplicativo se adapta a los datos actuales.

2.2.4 Representación de la ACF y la PACF

Una vez que la serie ha sido transformada y diferenciada el número de veces necesario para alcanzar la estacionariedad, se procede al análisis de la función de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie resultante, con el fin de identificar los órdenes del modelo correspondiente.
Las gráficas de autocorrelación de los modelos ARMA$(p, q) \times (P, Q)_s$ sugieren los órdenes de los componentes autorregresivos (AR) y de medias móviles (MA), tanto en su parte regular como estacional (que posteriormente se comprueban estadísticamente)
Las gráficas ACF representan, en retardos bajos, el comportamiento de la parte regular (ARMA(p, q)). Por tanto, si la ACF presenta un corte brusco a partir de un retardo q, esto sugiere que se está ante un modelo MA(q). En cambio, si decae de forma progresiva, sugiere la presencia de una componente AR. Por otro lado, en los retardos estacionales $(s, 2s, \dots)$, la ACF revela la posible presencia de componentes estacionales, donde un corte a partir del retardo Qs indicaría un componente MA estacional de orden Q.
La PACF se interpreta de forma análoga. En retardos bajos, se observa la autocorrelación parcial de la parte regular. Al igual que sucede con la ACF, la PACF tiene un comportamiento claro para los modelos con dependencia regular: si se corta a partir de un retardo p, se está ante un modelo AR(p), mientras que decae lentamente para un MA(q). En los retardos estacionales de la PACF, $(s, 2s, \dots)$, se observa la autocorrelación parcial de la parte estacional. Por tanto, si hay picos significativos en los retardos estacionales hasta el retardo Ps, esto indica que se está ante un modelo con componente estacional autorregresiva de orden P. La interpretación de estas gráficas se explica en [1, Capítulo 3].
Representando las gráficas ACF y PACF en la Figura 2.5 se procede a su análisis visual.

En la ACF, se observa un lento descenso en los primeros retardos, lo que sugiere que puede haber una componente regular autorregresiva. En este gráfico, aunque se haya diferenciado esta-

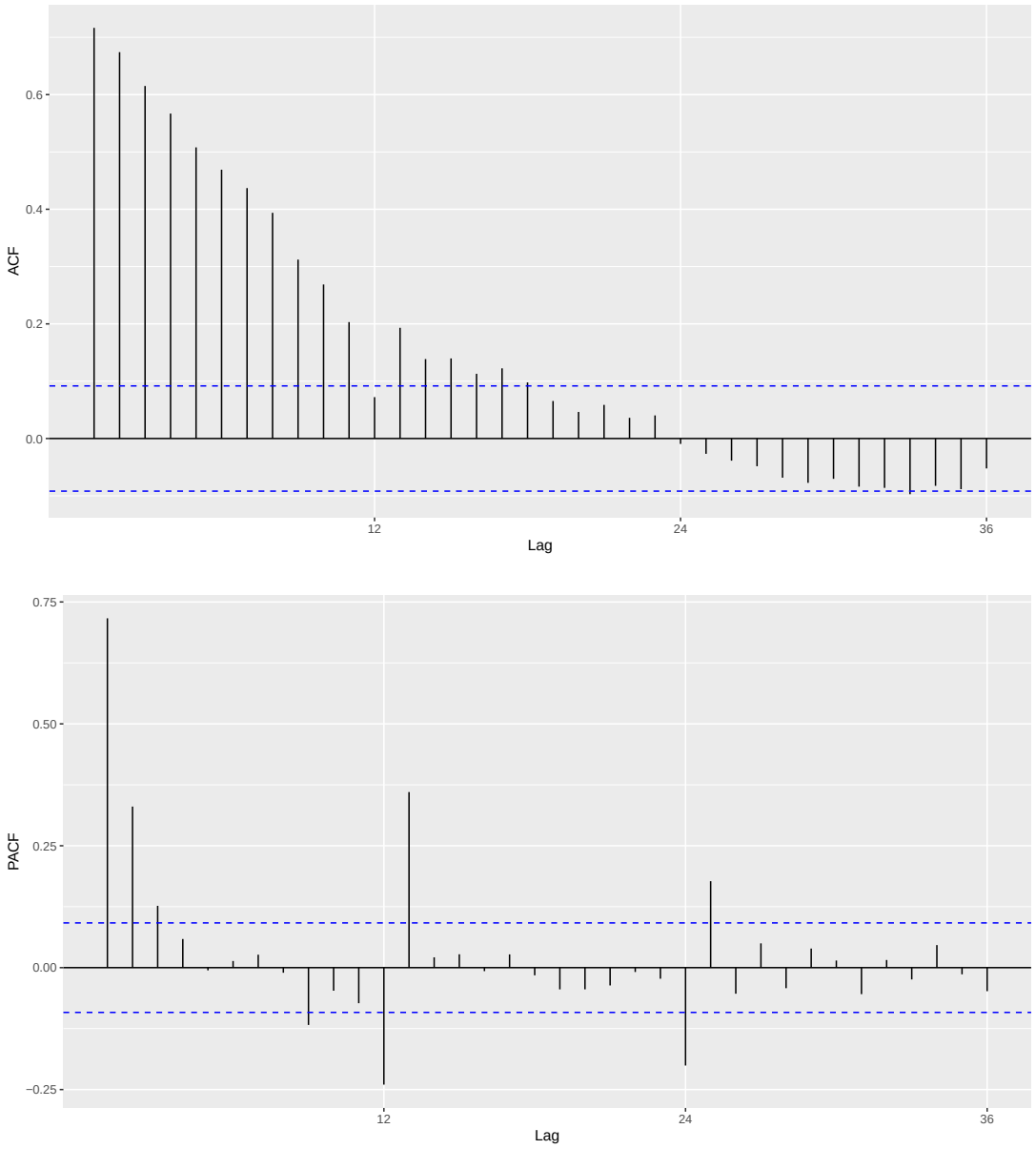

Figura 2.5: Gráficas ACF y PACF de la serie transformada Z_t

cionalmente y los picos estacionales se hayan reducido, aún se puede observar que existen picos en los retardos 12 y 24, que luego se cortan, lo cual puede indicar una componente estacional MA de orden 2.
Por otro lado, en la PACF, se observan dos retardos significativos, lo que puede indicar una componente de medias móviles de orden 2. A su vez, los picos en los retardos 12 y 24 hacen pensar que se está ante una componente autorregresiva estacional de orden 2.
Por tanto, el modelo propuesto para la serie estacionaria es un ARMA$(2,2)\times(2,2)_{12}$. El modelo sobre la serie logarítmica es, por tanto

$$\text{ARIMA}(2,0,2)\times(2,1,2)_{12} \tag{2.23}$$

2.3 Estimación

En esta sección se lleva a cabo el ajuste del modelo propuesto: se estiman sus coeficientes. Aunque la formulación teórica se expone a continuación, la estimación práctica se realizará mediante comandos en RStudio.
La expresión matemática del modelo ARIMA$(2,0,2)\times(2,1,2)_{12}$ es:

$$\begin{aligned}&(1-\phi_1B-\phi_2B^2)(1-\Phi_1B^{12}-\Phi_2B^{24})(1-B^{12})Y_t=\\&c+(1+\theta_1B+\theta_2B^2)(1+\Theta_1B^{12}+\Theta_2B^{24})a_t\end{aligned} \tag{2.24}$$

donde X_t es la serie original, $Y_t=\log(X_t)$ la serie transformada logarítmicamente y a_t el error del modelo. Los parámetros ϕ_1,ϕ_2,θ_1 y θ_2 corresponden a los coeficientes AR(2) y MA(2) de la parte regular del modelo, mientras que Φ_1,Φ_2,Θ_1 y Θ_2 representan los coeficientes de las componentes AR(2) y MA(2) estacionales. En este modelo con diferencia estacional, la constante c representa la media del cambio logarítmico estacional, es decir, el crecimiento relativo medio anual de la serie X_t.
Los parámetros ϕ_1, ϕ_2, Φ_1, Φ_2, θ_1, θ_2, Θ_1, Θ_2 y c, así como la varianza del error, σ_a^2, son desconocidos, y se estiman mediante máxima verosimilitud (véase [2, Sección 9.1]).

Estimación por máxima verosimilitud
Sea el modelo (2.24), la estimación por máxima verosimilitud de los parámetros ϕ_1, ϕ_2, Φ_1, Φ_2, θ_1, θ_2, Θ_1, Θ_2, c y σ_a^2 consiste en encontrar los valores de estos parámetros que maximicen la probabilidad de encontrar los valores observados $y_1,y_2,\ldots,y_T$ bajo el modelo asumido, siendo $y_t=\log(x_t)$, con $t=1,\ldots,T(=468)$, T es el número de observaciones de la serie.
Esta estimación se lleva a cabo mediante la función de máxima verosimilitud:

Definición 2.34. [Función de verosimilitud (teórica)]: Fijada una realización muestral $y_1,\ldots,y_T$, la función

$$g_{\phi_1,\phi_2,\ldots,\Theta_1,\Theta_2,c,\sigma_a^2}(y_1,\ldots,y_T)$$

como función de los parámetros ϕ_1, ϕ_2, Φ_1, Φ_2, θ_1, θ_2, Θ_1, Θ_2, c y σ_a^2 recibe el nombre de función de verosimilitud, y es la función de probabilidad conjunta de un vector aleatorio $(Y_1, \ldots, Y_T)$ obtenido del proceso estocástico (2.24).

La función de verosimilitud se denota por

$$L_{y_1,\ldots,y_T}(\phi_1, \phi_2, \ldots, \Theta_1, \Theta_2, \sigma_a^2)$$

Por tanto, la estimación por máxima verosimilitud de estos parámetros son los valores $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\Phi}_1$, $\hat{\Phi}_2$, $\hat{\Theta}_1$, $\hat{\Theta}_2$, $\hat{c}$ y $\hat{\sigma}_a^2$ que maximizan la función de verosimilitud L, es decir:

$$\left(\hat{\phi}_1, \hat{\phi}_2, \hat{\theta}_1, \hat{\theta}_2, \hat{\Phi}_1, \hat{\Phi}_2, \hat{\Theta}_1, \hat{\Theta}_2, \hat{c}, \hat{\sigma}_a^2\right) = \max_{(\phi_1,\phi_2,\ldots,\Theta_1,\Theta_2,\sigma_a^2)} L\left(\phi_1, \phi_2, \ldots, \Theta_1, \Theta_2, \sigma_a^2\right) \tag{2.25}$$

Observación 2.35. A menudo, en vez de usar la función de verosimilitud como se ha definido, se le aplica un logaritmo, que la hace más manejable en muchos casos. A esta nueva función se le llama función de log-verosimilitud, y la estimación se lleva a cabo maximizando esta función. La log-verosimilitud también se usa para medir qué tan bien se ajusta el modelo a los datos.

En la práctica, el proceso de estimación de los coeficientes se lleva a cabo mediante la implementación `Arima` del paquete `forecast` de RStudio, que también usa la máxima verosimilitud para la estimación de los coeficientes. Esta implementación proporciona los valores estimados de los parámetros del modelo junto con sus errores estándar, con los que se pueden llevar a cabo contrastes de significación (explicados en 2.6.2).

```
modelo_inicial <- Arima(log_pasajeros_ts, order = c(2, 0, 2), seasonal = list(order = c(2,
    1, 2), period = 12), include.constant = TRUE)
summary(modelo_inicial)
```

```
> Series: log_pasajeros_ts
> ARIMA(2,0,2)(2,1,2)[12] with drift
>
> Coefficients:
>          ar1     ar2     ma1      ma2     sar1     sar2     sma1     sma2
>       0.0052  0.9412  0.5281  -0.4719  -0.5703  -0.0458  -0.0336  -0.3558
> s.e.  0.0187  0.0181  0.0477   0.0469   0.2849   0.0750   0.2820   0.1894
>        drift
>       0.0036
> s.e.  0.0016
>
> sigma^2 = 0.003108:  log likelihood = 669.35
> AIC=-1318.7   AICc=-1318.2   BIC=-1277.47
>
> Training set error measures:
>                       ME       RMSE      MAE        MPE      MAPE      MASE
> Training set 0.001249692 0.05448275 0.039725 0.01022599 0.3305423 0.4383678
>                     ACF1
> Training set 0.002976422
```

El modelo ARIMA$(2,0,2) \times (2,1,2)_{12}$ fue ajustado por máxima verosimilitud sobre la serie logarítmica. Sin embargo, los coeficientes ϕ_1, Φ_2 y Θ_1 han resultado no significativos. Se ha comprobado individualmente que su exclusión no empeoraba el ajuste del modelo: tanto los criterios de información AIC, AICc y BIC, como las medidas de error predictivo se mantenían similares. Por tanto, se llevó a cabo la exclusión de los tres coeficientes, dando lugar a un modelo equivalente en términos de ajuste y más simple en cuanto al número de parámetros. El nuevo modelo ajustado se obtiene de la siguiente forma

```
modelo <- Arima(log_pasajeros_ts, order = c(2, 0, 2), seasonal = list(order = c(1,
    1, 2), period = 12), include.constant = TRUE, fixed = c(0, NA, NA, NA, NA, 0,
    NA, NA))
summary(modelo)
```

```
> Series: log_pasajeros_ts
> ARIMA(2,0,2)(1,1,2)[12] with drift
>
> Coefficients:
>       ar1     ar2     ma1      ma2     sar1  sma1     sma2   drift
>         0  0.9413  0.5305  -0.4695  -0.5938     0  -0.4116  0.0037
> s.e.    0  0.0182  0.0440   0.0433   0.0449     0   0.0495  0.0014
>
> sigma^2 = 0.003091:  log likelihood = 669.14
> AIC=-1324.28   AICc=-1324.03   BIC=-1295.42
>
> Training set error measures:
>                      ME       RMSE        MAE         MPE      MAPE      MASE
> Training set 0.0010228 0.05451881 0.03974351 0.008286539 0.3307068 0.4385721
>                      ACF1
> Training set 0.002842384
```

donde los coeficientes estimados fueron $\hat{\phi}_2 = 0.9413$, $\hat{\theta}_1 = 0.5305$, $\hat{\theta}_2 = -0.4695$, $\hat{\Phi}_1 = -0.5938$, $\hat{\Theta}_2 = -0.4116$ y $\hat{c} = 0.0037$. La estimación de la varianza del ruido blanco es $\hat{\sigma}_a^2 = 0.003091$.

El modelo resultante se representa como

$$(1 - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B^{12})Y_t = c + (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta_2 B^{24})a_t \tag{2.26}$$

sustituyendo los parámetros por sus estimaciones. Este modelo tiene una estructura ARIMA $(2,0,2) \times (1,1,2)_{12}$, aunque con los coeficientes ϕ_1 y Θ_1 nulos por no ser estadísticamente significativos.

2.4 Validación

Una vez el modelo ha sido ajustado, la siguiente fase es comprobar si se cumplen las hipótesis supuestas sobre el modelo. Esta es la etapa de validación o diagnosis del modelo, y consiste en comprobar si los errores del modelo $\{a_t\}$ (también llamados innovaciones) son ruido blanco, pues esto significa que el modelo ha captado toda la dependencia temporal y no ha dejado información sin modelizar (para más detalle, véase [2, Capítulo 14]). Los errores son ruido blanco si:

- Están incorrelacionados entre si.
- Tienen media cero.
- Tienen varianza constante σ_a^2.

Si el modelo no cumple estos supuestos, no es un modelo adecuado como generador de la serie, pues no ha captado toda la dependencia temporal. También se comprueba si las innovaciones tienen distribución gaussiana, pues es una hipótesis importante en la etapa de predicción de futuros valores: la construcción de intervalos para estas predicciones se basa en la distribución normal de los errores del modelo.
Como los errores a_t del modelo no han sido observados, se trabaja con los residuos. Los residuos se obtienen estimando $\hat{a}_t$ al aplicar el modelo con coeficientes estimados, y se consideran una aproximación de los errores a_t.

En RStudio, los residuos se calculan con el comando `residuals`:

```
# Cálculo de los residuos
residuos <- residuals(modelo)
```

A continuación, con el comando `checkresiduals` se representan las gráficas secuencial, de autocorerlación simple y de comparación con la normal de los residuos del modelo.

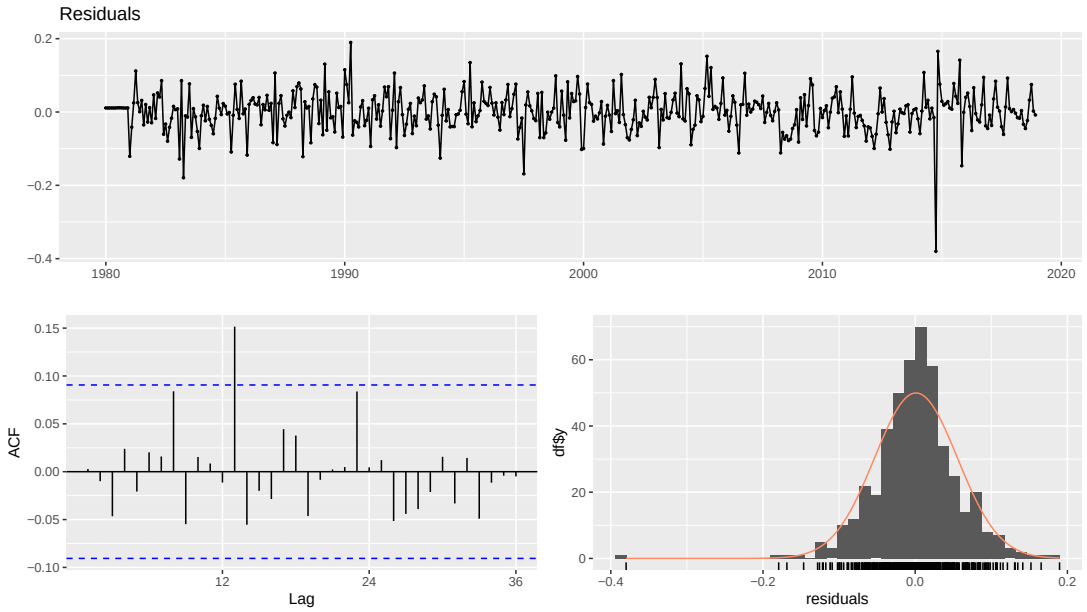

Figura 2.6: Gráficas de los residuos del modelo

En la gráfica secuencial, se observa que la media de los residuos está en torno a cero y que su varianza es constante. En la ACF, se ve que hay correlación en algún retardo; y en la gráfica de comparación con la distribución normal estándar podría parecer que los residuos se comportan como esta distribución (aunque con colas largas). No obstante, estos aspectos se comprueban mediante tests estadísticos, explicados y llevados a cabo a continuación.

2.4.1 Contrastes estadísticos

El vector de residuos es $(\hat{a}_1, \ldots, \hat{a}_T)$, donde T es el tamaño de la muestra. En esta sección, ρ_k denota la correlación entre innovaciones a_t de la serie, $\rho_k = \frac{\mathrm{Cov}(a_t, a_{t+k})}{\mathrm{Var}(a_t)}$. Por tanto, $\hat{\rho}_k$ denotará la correlación simple muestral de los residuos, $\hat{\rho}_k = \frac{\sum_{t=1}^{T-k}(\hat{a}_t - \bar{\hat{a}})(\hat{a}_{t+k} - \bar{\hat{a}})}{\sum_{t=1}^{T}(\hat{a}_t - \bar{\hat{a}})^2}$, siendo $\bar{\hat{a}}$ la media muestral de los residuos.

Contraste de incorrelación: test de Ljung-Box

El contraste de incorrelación que se llevará a cabo es el de Ljung-Box, pues permite contrastar si la autocorrelación es nula hasta determinado retardo H. Este test se basa en el contraste

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_H = 0$$
$$H_1 : \rho_j \neq 0 \quad \text{para algún} \quad j \in 1, 2, \ldots, H$$

El estadístico es

$$Q_H = T(T+2)\sum_{k=1}^{H}\frac{\hat{\rho}_k^2}{T-k} \tag{2.27}$$

y bajo el supuesto de que el tamaño muestral T es grande, tiene distribución $\chi^2_{H-(p+q+P+Q)}$, donde los grados de libertad vienen dados por el retardo H menos el número de parámetros estimados por el modelo, que en este caso es $p+q+P+Q=7$.
Por tanto, la hipótesis nula de incorrelación se rechaza cuando el valor observado del estadístico excede el cuantil de distribución, para un determinado nivel de significación.
En RStudio, este contraste con $H=24$ se lleva a cabo de la siguente forma

```
# Contraste de incorrelación de los residuos
Box.test(residuos, lag = 24, fitdf = 7, type = "Ljung-Box")


>
>  Box-Ljung test
>
> data:  residuos
> X-squared = 26.308, df = 17, p-value = 0.06902
```

Como el p-valor es mayor que 0.05, no hay pruebas suficientes para rechazar la hipótesis de que los residuos están incorrelacionados, por tanto se acepta que el modelo cumple la suposición de que los residuos no presentan correlación. Este test viene propuesto en [5, Páginas 297–303].

Contraste de media cero

En esta sección, se llamará μ_a a la media teórica de las innovaciones a_t del modelo. $\bar{\hat{a}}$ es la media muestral y $\hat{s}^2_{\hat{a}}$ la cuasivarianza de los residuos $\hat{a}_t$.
El test que se lleva a cabo se basa en las hipótesis

$$H_0 : \mu_a = 0$$
$$H_1 : \mu_a \neq 0$$

El estadístico, con su distribución, es el siguiente

$$\frac{\bar{\hat{a}}}{\hat{s}_{\hat{a}}/\sqrt{T}} \approx \mathcal{N}(0,1) \tag{2.28}$$

Por tanto, se rechaza la hipótesis nula cuando el valor observado del estadístico excede el cuantil de la distribución, al nivel de significación determinado o cuando el p-valor es muy bajo.
En este caso

```
# Contraste de media cero de los residuos
t.test(residuos, mu = 0)


>
>  One Sample t-test
>
> data:  residuos
> t = 0.40549, df = 467, p-value = 0.6853
> alternative hypothesis: true mean is not equal to 0
> 95 percent confidence interval:
>  -0.003933831  0.005979431
> sample estimates:
> mean of x
> 0.0010228
```

No existen pruebas significativas para rechazar la hipótesis, por lo que se acepta que el modelo cumple el supuesto de que los residuos tienen media nula.

Contraste de varianza constante: Test ARCH-LM

Este test se usa para comprobar si los residuos de un modelo presentan varianza heterocedástica dependiente del tiempo t. El test se basa en el contraste

$$H_0 : \text{la varianza es constante}$$
$$H_1 : \text{la varianza depende del tiempo pasado } t$$

El test ARCH-LM emplea los residuos al cuadrado $\hat{a}_t^2$ y hace una regresión sobre sus $q \in \mathbb{Z}$ residuos pasados $\hat{a}_{t-1}, \ldots, \hat{a}_{t-q}$. El estadístico se calcula como

$$LM = TR^2 \tag{2.29}$$

donde T es el número de observaciones de la serie y R^2 el coeficiente de determinación de la regresión sobre los q residuos pasados. Para elegir q se suele tomar el número de retardos en los que los residuos muestran correlación significativa. El estadístico tiene una distribución ji-cuadrado con q grados de libertad

$$LM \sim \chi_q^2$$

Por tanto, se rechaza la hipótesis nula de varianza constante con cierta significación cuando el valor del estadístico obtenido es mayor que el cuantil correspondiente o cuando el p-valor es pequeño. Se lleva a cabo este contraste en RStudio con el retardo 13, pues es el que más correlación presenta en el gráfico ACF de los residuos

```
ArchTest(residuos, lag = 13)


>
>  ARCH LM-test; Null hypothesis: no ARCH effects
>
> data:  residuos
> Chi-squared = 15.584, df = 13, p-value = 0.2723
```

Se tiene un p-valor alto, por lo que se acepta la hipótesis nula de varianza constante, y por tanto el modelo cumple este supuesto. Este contraste se formula en [6, Páginas 987–1007].

Contraste de normalidad: Test de Shapiro-Wilk

Se contrasta la hipótesis nula H_0 : el vector $(a_1, \ldots, a_T)$ tiene distribución gaussiana, a través de un estadístico ω que utiliza inversas de valores de la normal estándar. La hipótesis nula de normalidad se rechaza cuando el valor obtenido del estadístico es pequeño.
Efectuando este contraste en R:

```
# Contraste de normalidad de los residuos
shapiro.test(residuos)


>
>  Shapiro-Wilk normality test
>
> data:  residuos
> W = 0.95761, p-value = 2.391e-10
```

El p-valor es de orden muy bajo, por tanto el modelo no cumple el supuesto de normalidad de los residuos. La construcción de este test se detalla en [7, Páginas 591–611].

Como se acaba de ver, el modelo propuesto ARIMA$(2,0,2) \times (1,1,2)_{12}$ para los datos logarítmicos del total de pasajeros del transporte aéreo es adecuado como generador de esta serie: aunque no cumple el supuesto de que los errores tienen distribución normal, sí cumple las hipótesis de incorrelación, varianza constante y media cero de los errores.

2.5 Predicción

Tras identificar, estimar y validar un modelo que es adecuado como generador de los datos de la serie de tiempo, la última fase consiste en la predicción de futuros valores y la construcción

de intervalos de confianza que determinen cuánto de precisas son estas predicciones. Aunque el modelo propuesto no cumple la suposición de normalidad de los residuos, por el Teorema Central del Límite, cuando hay suficientes datos, los intervalos de predicción son aproximadamente normales. Aun así, se empleará la técnica de bootstrap para obtener predicciones e intervalos empíricos sin necesidad de asumir normalidad. Para más detalles sobre la fase de predicción, véase [1, Sección 3.4].

2.5.1 Predicciones bootstrap

Sea $\{Y_t\}_{t=1}^{T}$ la serie transformada logarítmicamente, sobre la que se ha ajustado un modelo ARIMA$(2,0,1)\times(1,1,2)_{12}$, puede representarse de forma compacta como

$$\phi(B)\,\Phi(B^{12})\,\nabla_{12}Y_t = \hat{c} + \theta(B)\,\Theta(B^{12})\,\hat{a}_t \tag{2.30}$$

donde $\phi(B)$ y $\theta(B)$ son los polinomios autorregresivo y de medias móviles de la parte regular, mientras que $\Phi(B^{12})$ y $\Theta(B^{12})$ son los correspondientes polinomios estacionales (con los coeficientes estimados). ∇_{12} es la diferencia estacional con periodo $s=12$ y $\hat{c}$ la constante estimada del modelo. Los términos $\hat{a}_t$ representan los residuos del modelo. Esta expresión corresponde al modelo con los parámetros estimados por máxima verosimilitud. Desarrollando esta expresión, se puede obtener Y_t en función de valores y errores pasados de la siguiente forma

$$\begin{aligned} Y_t = {} & \hat{\phi}_2 Y_{t-2} + (1+\hat{\Phi}_1)Y_{t-12} - (\hat{\phi}_2 + \hat{\phi}_2\hat{\Phi}_1)Y_{t-14} - \hat{\Phi}_1 Y_{t-24} + \hat{\phi}_2\hat{\Phi}_1 Y_{t-26} \\ & + \hat{c} + \hat{a}_t + \hat{\theta}_1\hat{a}_{t-1} + \hat{\theta}_2\hat{a}_{t-2} + \hat{\Theta}_2\hat{a}_{t-24} + \hat{\theta}_1\hat{\Theta}_2\hat{a}_{t-25} + \hat{\theta}_2\hat{\Theta}_2\hat{a}_{t-26} \end{aligned} \tag{2.31}$$

donde se le llamará $\hat{f}(Y_{t-1}, Y_{t-2}, \dots; \hat{a}_t, \hat{a}_{t-1}, \dots)$ a esta función de predicción inducida por el ARIMA estimado.

Sea Y_T la última observación de la serie logarítmica, la predicción puntual de Y_{T+1} se obtiene sustituyendo en el modelo los valores pasados observados de Y_T y de los residuos estimados $\hat{a}_t$ mediante la función $\hat{f}$ recién definida.

$$\hat{Y}_{T+1} = \hat{f}(Y_{T-1}, Y_{T-2}, \dots; \hat{a}_T, \hat{a}_{T-1}, \dots)$$

Para realizar un número h de predicciones por bootstrap sobre la observación T, se simulan B trayectorias futuras de $Y_{T+1}, Y_{T+2}, \dots, Y_{T+h}$ a partir del modelo estimado:

$$\overset{(j)}{Y}_{T+k} = \hat{f}\left(\overset{(j)}{Y}_{T+k-1}, \overset{(j)}{Y}_{T+k-2}, \dots; \overset{(j)}{\hat{a}}_{T+k}, \overset{(j)}{\hat{a}}_{T+k-1}, \dots\right)$$

con $j = 1, \dots, B$ y $k \in \{1, \dots, h\}$.

Dado que los residuos $\hat{a}_{T+1}, \ldots, \hat{a}_{T+k}$ no son observados, se generan aleatoriamente mediante remuestreo con reemplazo de los residuos $\hat{a}_t$. Es decir, en cada trayectoria j, se toma al azar (con reemplazo) uno de los residuos $(\hat{a}_1, \ldots, \hat{a}_T)$.

Para cada instante futuro $T+k$, con $k = 1, \ldots, h$, se generan B trayectorias bootstrap:

$$\{\overset{(1)}{Y}_{T+k}, \overset{(2)}{Y}_{T+k}, \ldots, \overset{(B)}{Y}_{T+k}\}$$

La predicción bootstrap $\hat{Y}_{T+k}$ se define como la media de estas trayectorias:

$$\hat{Y}_{T+k} = \frac{1}{B} \sum_{j=1}^{B} \overset{(j)}{Y}_{T+k} \tag{2.32}$$

Se dispone de una serie transformada logarítmicamente $y_1, \ldots, y_T$, con $T = 468$ observaciones mensuales correspondientes al periodo comprendido entre enero de 1980 y diciembre de 2018. El objetivo es predecir los valores futuros de la serie transformada para los 12 meses de 2019, es decir, realizar $h = 12$ predicciones bootstrap de y_{T+k}, con $k = 1, \ldots, 12$. Se realizan en RStudio estas 12 predicciones un número de veces $B = 1000$. Para calcular las predicciones en base al modelo ajustado, se usa el comando `simulate`.

```
# Número de valores a predecir
h <- 12
# Número de réplicas de bootstrap
B <- 1000
# Almacenar todas las predicciones bootstrap
predicciones <- matrix(NA, nrow = B, ncol = h)
set.seed(123)
for (i in 1:B) {
    # Estimacion de residuos
    err <- sample(residuos, size = h, replace = TRUE)
    # Predicciones
    pred <- simulate(modelo, nsim = h, future = TRUE, innov = err)
    # Guardar predicciones
    predicciones[i, ] <- pred
}
# Calcular predicción media
pred_media <- colMeans(predicciones)
```

Las predicciones para la serie logarítmica del total de pasajeros del transporte aéreo para el 2019 son, por tanto, las mostradas a continuación

```
pred_media


>  [1] 12.74613 12.72171 12.94262 13.03101 13.07772 13.10403 13.21350 13.21365
>  [9] 13.14872 13.06262 12.88384 12.87891
```

2.5.2 Intervalos de confianza para las predicciones

Para saber qué precisas son las predicciones bootstrap, se calculan intervalos de confianza para cada predicción. Los extremos del intervalo de predicción para cada $\hat{Y}_{T+k}$ vendrán dados por los percentiles muestrales de las B trayectorias:

$$\{\overset{(1)}{Y}_{T+k}, \overset{(2)}{Y}_{T+k}, \ldots, \overset{(B)}{Y}_{T+k}\}$$

para cada $k \in \{1, 2, \ldots, h\}$.

Por ejemplo, el intervalo de confianza al 95% para $\hat{Y}_{T+k}$ viene dado por

$$\left[\overset{(k)}{\hat{q}}_{0.025}, \overset{(k)}{\hat{q}}_{0.975}\right]$$

donde $\overset{(k)}{\hat{q}}_{0.025}$ y $\overset{(k)}{\hat{q}}_{0.975}$ son los percentiles muestrales 2.5% y 97.5% de las B trayectorias correspondientes a la predicción k.

Para las 12 predicciones obtenidas anteriormente, se calculan los intervalos al 95%

```
# Intervalos de predicción
extremo_inferior <- apply(predicciones, 2, quantile, probs = 0.025)
extremo_superior <- apply(predicciones, 2, quantile, probs = 0.975)
# Visualización de las predicciones junto a sus intervalos (serie transformada)
data.frame(Mes = 1:h, Predicciones = pred_media, Extremo_Inferior = extremo_inferior,
    Extremo_Superior = extremo_superior)
```

```
>    Mes Predicciones Extremo_Inferior Extremo_Superior
> 1    1     12.74613         12.63730         12.85276
> 2    2     12.72171         12.59564         12.84446
> 3    3     12.94262         12.80770         13.06693
> 4    4     13.03101         12.88078         13.16830
> 5    5     13.07772         12.90609         13.22870
> 6    6     13.10403         12.94241         13.26190
> 7    7     13.21350         13.04222         13.37780
> 8    8     13.21365         13.03635         13.38773
> 9    9     13.14872         12.97417         13.33264
> 10  10     13.06262         12.87998         13.23927
> 11  11     12.88384         12.69119         13.06860
> 12  12     12.87891         12.69491         13.05846
```

Para más detalles sobre la técnica de bootstrap, véase [1, Sección 6.7].

2.5.3 Obtención de predicciones sobre la serie original

Finalmente, como las predicciones se han obtenido para la serie logarítmica $\{Y_t\}$, es necesario deshacer esta transformación: aplicar la inversa del logaritmo para obtener predicciones e inter-

valos en la serie original del tráfico de pasajeros.

Para deshacer el logaritmo $\log(X_t) = Y_t$, se aplica la función exponencial: $X_t = \exp(Y_t)$. Por tanto, para recuperar las trayectorias de la serie original, $\overset{(1)}{X}_{T+k}, \overset{(2)}{X}_{T+k}, \ldots, \overset{(B)}{X}_{T+k}$, se aplica esta función:

$$\overset{(j)}{X}_{T+k} = \exp\left(\overset{(j)}{Y}_{T+k}\right)$$

a cada trayectoria $j = 1, \ldots, B$, y luego se calcula la predicción de la serie original

$$\hat{X}_{T+k} = \frac{1}{B}\sum_{j=1}^{B} \overset{(j)}{X}_{T+k}$$

Esto se hace para $k = 1, \ldots, h$, lo que permite obtener las predicciones de la serie original de pasajeros. En este caso, $h = 12$ para predecir los valores de los 12 meses de 2019.

```
# Deshacer transformación logarítmica
predicciones_x <- matrix(NA, nrow = B, ncol = 12)
for (i in 1:B) {
    predicciones_x[i, ] <- exp(predicciones[i, ])
}
pred_media_x <- colMeans(predicciones_x)
```

Las predicciones de la serie original para el año 2019 son las siguientes

```
pred_media_x


>  [1] 343762.2 335673.0 418627.1 457595.6 479597.4 492576.6 549686.6 549868.9
>  [9] 515394.2 472964.0 395642.9 393675.8
```

Se calculan sus intervalos de confianza igual que para la serie logarítmica, al 95%, y se muestran a continuación

```
data.frame(Mes = 1:h, Predicción = pred_media_x, Extremo_Inferior = ic_lower_x,
    Extremo_superior = ic_upper_x)


>    Mes Predicción Extremo_Inferior Extremo_superior
> 1    1   343762.2         307829.3         381839.5
> 2    2   335673.0         295269.4         378683.9
> 3    3   418627.1         365019.0         473038.4
> 4    4   457595.6         392690.4         523506.2
> 5    5   479597.4         402758.8         556099.7
> 6    6   492576.6         417655.0         574870.4
> 7    7   549686.6         461490.0         645514.5
> 8    8   549868.9         458789.4         651954.2
> 9    9   515394.2         431132.2         617007.8
> 10  10   472964.0         392376.1         562005.8
> 11  11   395642.9         324873.1         473827.2
> 12  12   393675.8         326084.4         469049.1
```

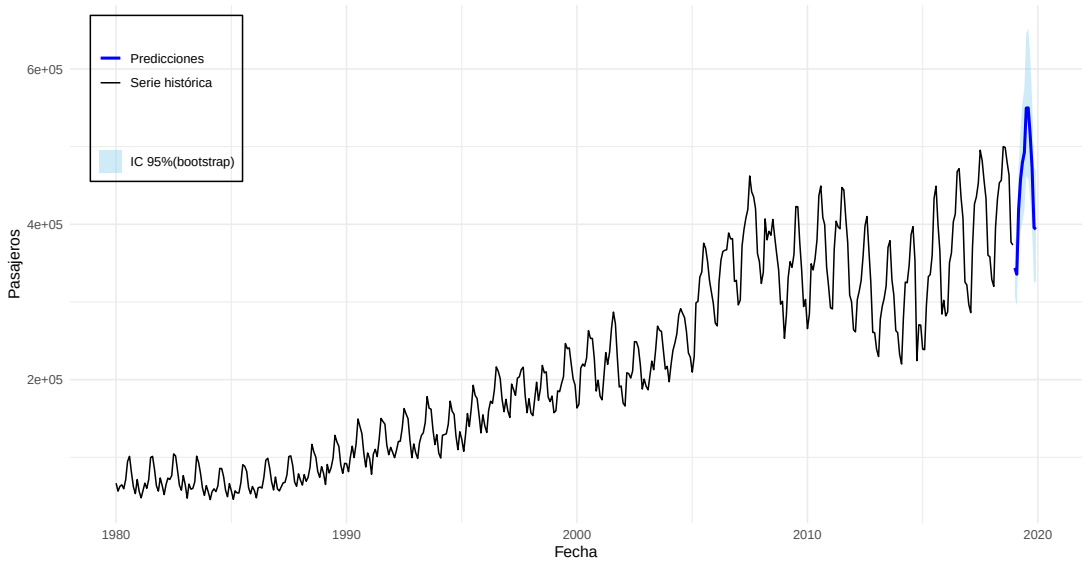

Figura 2.7: Predicciones bootstrap junto a sus intervalos. Serie completa (1980-2020)

Se pueden representar simultáneamente las predicciones junto con sus intervalos de confianza obtenidos por bootstrap y la serie original, en la Figura 2.7.

Como en la base de datos originales se dispone de los datos hasta 2025, se comparan las predicciones obtenidas mediante bootstrap para los 12 meses de 2019 con los valores que tomó la serie real en este año. Se representa la serie desde el año 2010 para una visualización más clara de esta comparación en la Figura 2.8.

2.6 Limitaciones del modelo

A veces, siguiendo el método iterativo Box-Jenkins, el modelo propuesto no es adecuado para capturar correctamente la dinámica de la serie, pues identificar visualmente un modelo mediante las funciones de autocorrelación y autocorrelación parcial no garantiza una buena especificación. Por ello, existen otras maneras de seleccionar modelos que se adapten a los datos de una serie temporal. Esto se hace mediante los criterios de información.

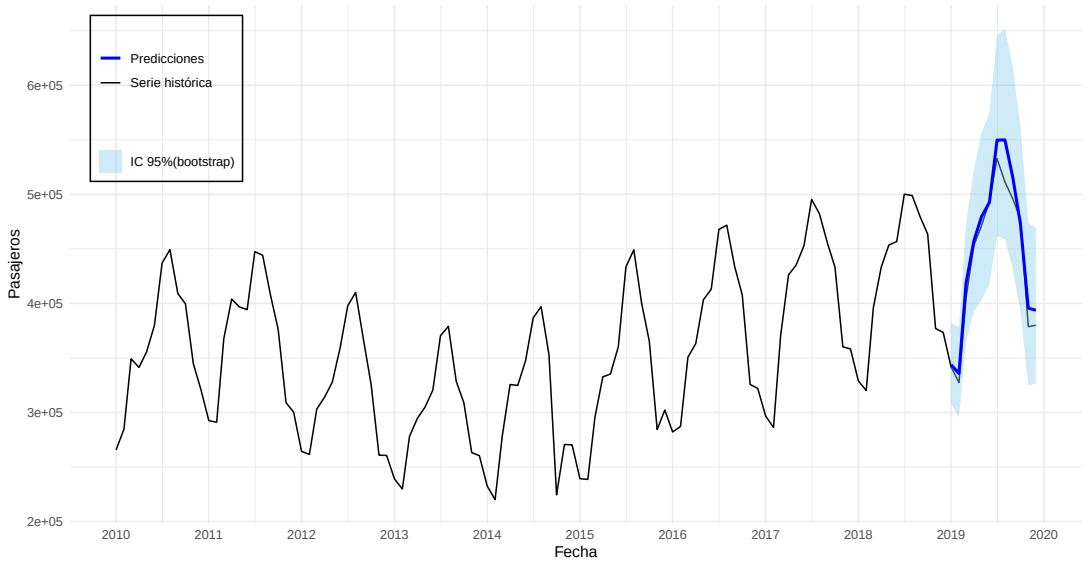

Figura 2.8: Serie original vs. Predicciones (2010-2020)

2.6.1 Criterios de información

Representar las funciones de autocorrelación simple y parcial no siempre es suficiente para elegir los órdenes del modelo: puede haber modelos que ajusten bien los datos pero viendo la ACF y la PACF sean pasados por alto, o puede ocurrir que estas gráficas no sugieran ningún modelo. Por tanto, se dispone de las siguientes funciones, llamadas criterios de información, que miden la calidad de ajuste del modelo, penalizando el exceso de parámetros. Siendo k el número de parámetros en el modelo y L su función de máxima log-verosimilitud, se definen a continuación:

- **Criterio de información de Akaike (AIC):** Penaliza el exceso de parámetros a través del segundo sumando.
  $\text{AIC} = -2\log L + 2k$
- **Criterio de información de Akaike corregido (AICc):** Es una versión corregida del AIC para muestras pequeñas.

  $\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}$, siendo n el número de observaciones de la serie.
- **Criterio de información Bayesiano:** Penaliza la complejidad del modelo más fuertemente que el AIC, sobre todo en muestras grandes.
  $\text{BIC} = -2\log L + k\log n$

Estos criterios se utilizan para comparar modelos ajustados sobre la misma serie, y miden

equilibrio entre buen ajuste del modelo y su complejidad. Valores más bajos de estas funciones indican mejor equilibrio: un buen ajuste sin demasiados parámetros.

El modelo que minimiza la función AICc se obtiene automáticamente mediante la función `auto.arima` del paquete `forecast` de RStudio, que selecciona el modelo evaluando múltiples combinaciones posibles de órdenes ARIMA. Este comando también estima los parámetros del modelo que selecciona.

Sin embargo, el modelo que minimice los criterios de información como el AICc, no tiene por qué ser el más adecuado: puede presentar un buen equilibrio entre el ajuste a los datos y el número de parámetros pero no verificar los supuestos comprobados en la etapa de validación. Estas definiciones vienen dadas en [1, Sección 2.1].

2.6.2 Significación de los coeficientes estimados

Otra observación que se puede hacer del modelo seleccionado es la significación de los coeficientes estimados. En la sección de estimación, se calcularon los coeficientes mediante máxima verosimilitud. No obstante, se tiene que bajo ciertas condiciones de regularidad, los estimadores de máxima verosimilitud son asintóticamente insesgados y tienen distribución normal asintótica, lo cual permite realizar contrastes de hipótesis que determinan si el estimador es significativamente distinto de cero. Si $\hat{\gamma}$ es el estimador y $\hat{\sigma}_{\hat{\gamma}}$ su error estándar, la distribución del estimador bajo estas condiciones es $\hat{\gamma} \approx \mathcal{N}(\gamma, \hat{\sigma}^2_{\hat{\gamma}})$. Esto se trata en [1, Sección 3.5]. Por tanto, para llevar a cabo el contraste con hipótesis nula $H_0 : \gamma = 0$, se usa el siguiente estadístico

$$\frac{\hat{\gamma}}{\hat{\sigma}_{\hat{\gamma}}} \sim \mathcal{N}(0,1)$$

Se realiza el contraste para el nivel de confianza del 95%:
Bajo hipótesis nula, este estadístico cae en la región de aceptación el 95% de veces si

$$\frac{|\hat{\gamma}|}{\hat{\sigma}_{\hat{\gamma}}} < z_{0.025}$$

siendo $z_{0.025} = 1.96$ el cuantil de la distribución normal estándar que deja una probabilidad de 0.025 a su derecha. Por tanto, se rechaza la hipótesis nula de que el parámetro estimado $\hat{\gamma} = 0$ al nivel de significación del 5% si

$$|\hat{\gamma}| \geq 1.96 \cdot \hat{\sigma}_{\hat{\gamma}} \tag{2.33}$$

Llevando a cabo este contraste para cada parámetro estimado, resultó que tres coeficientes del modelo ajustado no resultaron significativos al 5% (uno resultó casi significativo y se decidió mantenerlo en el modelo). Por ello, se procedió a eliminarlos, obteniéndose un modelo más simple que seguía proporcionando un buen ajuste a los datos y cuyos residuos mantenían validez como ruido blanco incorrelacionado.

2.6.3 Medidas del error de predicción

En las gráficas anteriores, se comparan las predicciones en base al modelo propuesto con los valores reales que tomó la serie en 2019. Para medir la capacidad de predicción de forma numérica, se definen a continuación las siguientes medidas del error. El error viene dado por

$$e_T(k) = X_{T+k} - \hat{X}_{T+h}$$

siendo X_{T+k} el valor real que toma la serie en el momento $T+k$, y $\hat{X}_{T+k}$ la predicción obtenida, con $k = 1, \ldots, h$. En este caso $h = 12$.
Las medidas aquí citadas vienen detalladas en [1, Sección 6].

Error absoluto medio (MAE):
Sea h el total de predicciones

$$\text{MAE} = \frac{1}{h}\sum_{k=1}^{h} |e_T(k)| \tag{2.34}$$

Es útil para cuantificar el tamaño medio del error. Se puede usar para comparar distintos métodos de predicción aplicados a una serie.

Raíz del error cuadrático medio (RMSE):
Sea h el total de predicciones

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{k=1}^{h} e_T(k)^2} \tag{2.35}$$

También es útil para comparar distintos métodos de predicción. El RMSE penaliza más los errores grandes.

Error porcentual absoluto medio (MAPE):
Sea h el total de predicciones

$$\text{MAPE} = \frac{1}{h}\sum_{k=1}^{h} |p_T(k)| \tag{2.36}$$

donde $p_T(k) = \frac{100 e_T(k)}{x_{T+k}}$. Solo se puede usar cuando $t \neq 0$ para todo t.

Se calcula a continuación el error absoluto medio en R:

```
pasajeros_2019 = pasajeros_2010_ts[109:120]  #Serie real en 2019
MAE <- sum(abs(pasajeros_2019 - pred_media_x))/h
MAE


> [1] 12326.67
```

Se tiene un error medio absoluto MAE = 12326.67. Esto indica que de media, las predicciones difieren de las observaciones reales en unos 12327 pasajeros. Dado que la media mensual de

pasajeros desde 1980 a 2019 está en torno a 210000, entonces $\frac{12326.67}{210000} \approx 5.9\%$, es decir, el error representa aproximadamente un 5.9% del valor observado.

Anexo I

Código complementario en R

Aquí se muestran códigos que se han omitido en el texto principal.

I.1 Lectura y preprocesado de los datos

Se muestra el código R utilizado para la importación de datos desde Excel y su transformación antes del análisis.

```
trafico <- read_delim("trafico.csv", delim = ";", show_col_types = FALSE)
trafico <- trafico[-1, ]  #Se quita la primera fila, que contiene datos no numéricos
trafico_2010 = trafico[361:480, ]  #Valores de 1980 a 2019 incluido, para la gráfica comparativa con mejor
# interpretación visual
names(trafico_2010) = c("año", "mes", "pasajeros_2010")
trafico <- trafico[1:468, ]  #Se seleccionan los datos de enero de 1980 a diciembre de 2018
names(trafico) <- c("año", "mes", "pasajeros")
# Se convierten las columnas en variables
año = trafico$año
mes = trafico$mes
pasajeros = trafico$pasajeros
pasajeros = as.numeric(pasajeros)
```

I.2 Representaciones gráficas

En esta sección se muestran los códigos de las representaciones gráficas, así como las transformaciones necesarias para obtener dichas gráficas.

- Gráfica de la serie original, en la Página 19

```
grafico_secuencial = plot(pasajeros_ts, main = "", ylab = "Pasajeros", xlab = "")
```

- Gráfica de la serie transformada logarítmicamente, en la Página 20

```
log_pasajeros = log(pasajeros)
# Se convierte en serie
log_pasajeros_ts = ts(log_pasajeros, start = c(1980, 1), frequency = 12)
grafico_secuencial_log = plot(log_pasajeros_ts, main = "", ylab = "log Pasajeros",
    xlab = "")
```

- Gráfica de la función de autocorrelación de la serie logarítmica, en la Página 20

```
ggAcf(log_pasajeros_ts, lag.max = 36, main = "")
```

- La aplicación de una diferencia estacional sobre la serie logarítmica y su representación gráfica (Página 24) fueron generadas por el código siguiente:

```
dif_est_log_pasajeros = diff(log_pasajeros_ts, lag = 12)
# Se convierte en serie temporal (desde 1981 por la aplicación de la diferencia
# estacional)
dif_est_log_pasajeros_ts = ts(dif_est_log_pasajeros, start = c(1981, 1), frequency = 12)
grafico_secuencial_dif_est_log = plot(dif_est_log_pasajeros_ts, main = "", xlab = "",
    ylab = "")
```

- Gráficas secuencial, de autocorrelación y de aproximación a la normal de los residuos, en la Página 32.

```
checkresiduals(residuos)
```

I.3 Ajuste del modelo eliminando parámetros no significativos

En la sección de estimación, se ajustó el modelo quitando los coeficientes no significativos. Antes de esto, se estimó el modelo eliminando estos coeficientes uno a uno. Se muestra aquí el resultado de cada ajuste. Primero, se ajusta el modelo con $\phi_1 = 0$:

```
modelo_1 <- Arima(log_pasajeros_ts, order = c(2, 0, 2), seasonal = list(order = c(2,
    1, 2), period = 12), include.constant = TRUE, fixed = c(0, NA, NA, NA, NA, NA,
    NA, NA, NA))
summary(modelo_1)


> Series: log_pasajeros_ts
> ARIMA(2,0,2)(2,1,2)[12] with drift
>
> Coefficients:
```

```
>       ar1     ar2     ma1      ma2     sar1     sar2     sma1     sma2   drift
>         0  0.9417  0.5332  -0.4668  -0.5806  -0.0455  -0.0237  -0.3623  0.0037
> s.e.    0  0.0182  0.0443   0.0437   0.2781   0.0756   0.2753   0.1861  0.0015
>
> sigma^2 = 0.003102:  log likelihood = 669.31
> AIC=-1320.63   AICc=-1320.23   BIC=-1283.53
>
> Training set error measures:
>                       ME       RMSE        MAE         MPE     MAPE      MASE
> Training set 0.001041856 0.05449538 0.03969637 0.008453303 0.330315 0.4380519
>                    ACF1
> Training set 0.00290914
```

Se tiene que tanto los criterios de información como las medidas de error predictivo de este nuevo modelo con $\phi_1 = 0$ son parecidas al modelo original, por tanto se obtiene un modelo más simple que funciona igual de bien.

Se ajusta el modelo para $\Phi_2 = 0$:

```
modelo_2 <- Arima(log_pasajeros_ts, order = c(2, 0, 2), seasonal = list(order = c(1,
    1, 2), period = 12), include.constant = TRUE)
summary(modelo_2)
```

```
> Series: log_pasajeros_ts
> ARIMA(2,0,2)(1,1,2)[12] with drift
>
> Coefficients:
>          ar1     ar2     ma1      ma2     sar1    sma1     sma2   drift
>       0.0051  0.9406  0.5261  -0.4739  -0.6185  0.0246  -0.4256  0.0037
> s.e.  0.0188  0.0181  0.0477   0.0467   0.2864  0.2783   0.1681  0.0015
>
> sigma^2 = 0.003103:  log likelihood = 669.18
> AIC=-1320.36   AICc=-1319.96   BIC=-1283.26
>
> Training set error measures:
>                       ME       RMSE       MAE        MPE      MAPE      MASE
> Training set 0.001058632 0.05450551 0.0397706 0.00862899 0.3309178 0.4388711
>                     ACF1
> Training set 0.002454392
```

Análogo al caso anterior, se obtiene un modelo más simple con la misma calidad de ajuste.

Se ajusta el modelo para $\Theta_1 = 0$:

```
modelo_3 <- Arima(log_pasajeros_ts, order = c(2, 0, 2), seasonal = list(order = c(2,
    1, 2), period = 12), include.constant = TRUE, fixed = c(NA, NA, NA, NA, NA, NA,
    0, NA, NA))
summary(modelo_3)
```

```
> Series: log_pasajeros_ts
```

```
> ARIMA(2,0,2)(2,1,2)[12] with drift
>
> Coefficients:


> Warning in sqrt(diag(x$var.coef)): Se han producido NaNs


>          ar1     ar2     ma1      ma2     sar1     sar2  sma1     sma2   drift
>       0.1575  0.7928  0.3551  -0.3775  -0.6136  -0.0470     0  -0.3847  0.0037
> s.e.     NaN     NaN     NaN      NaN   0.0482   0.0743     0   0.0782  0.0015
>
> sigma^2 = 0.003151:  log likelihood = 667.03
> AIC=-1316.05   AICc=-1315.65   BIC=-1278.95
>
> Training set error measures:
>                       ME       RMSE        MAE         MPE      MAPE      MASE
> Training set 0.001106987 0.05492365 0.03978985 0.009044177 0.3310559 0.4390835
>                     ACF1
> Training set 0.009847451
```

El modelo con solo $\Theta_1 = 0$ presenta problemas de identificación, los errores estándar de otros coeficientes no se pueden calcular, pero quitando los otros coeficientes no significativos se estabiliza.

I.4 Cálculo y representación de intervalos de predicción

El cálculo de los intervalos de predicción para la serie original de pasajeros, X_t, es análogo a los de la serie logarítmica, por eso se omiten en el texto principal. Aquí se muestra el código que genera estos intervalos.

```
ic_lower_x <- apply(predicciones_x, 2, quantile, probs = 0.025)
ic_upper_x <- apply(predicciones_x, 2, quantile, probs = 0.975)
```

La representación gráfica de la serie original junto con las predicciones para el año 2019 y sus correspondientes intervalos de predicción (Figura 2.7) se ha realizado mediante la función `ggplot()` del paquete `ggplot2` en el entorno de RStudio:

```
start_pred <- time(pasajeros_ts)[length(pasajeros_ts)] + 1/12
fechas_pred<-seq(from=as.yearmon(start_pred),by=1/12, length.out=h)
fechas_pred=as.Date(fechas_pred)
df_pred <- data.frame(fecha = fechas_pred, media = pred_media_x, li = ic_lower_x,ls = ic_upper_x)
df_hist <- data.frame(fecha = as.Date(as.yearmon(time(pasajeros_ts))),
valor = as.numeric(pasajeros_ts))

ggplot() +
geom_line(data = df_hist, aes(x = fecha, y = valor,color="Serie histórica"))+
```

```
geom_ribbon(data = df_pred, aes(x = fecha, ymin = li, ymax = ls, fill="IC 95%(bootstrap)"), alpha = 0.4) +
geom_line(data = df_pred, aes(x = fecha, y = media,color="Predicciones"), linewidth = 1) +
scale_color_manual(values=c("Serie histórica"="black","Predicciones"= "blue"))+
scale_fill_manual( values=c("IC 95%(bootstrap)"="skyblue"))+
labs(title = "",x = "Fecha", y = "Pasajeros",color="",fill="",
caption = "") +
theme_minimal()+
theme(legend.position = c(0.02, 0.98),
legend.justification = c(0, 1),   # Ancla la esquina superior izquierda de la leyenda
legend.box.background = element_rect(colour = "black", linewidth = 0.5))
```

Como se describe en la sección correspondiente, se incluye una representación gráfica con el objetivo de facilitar la interpretación visual de la comparación entre las predicciones obtenidas mediante bootstrap y los valores reales que tomó la serie en 2019. Para ello, se construyó un vector adicional con valores de la serie original a partir de 2010. A continuación, se muestra el código empleado tanto para la creación de dicho vector como para la generación del gráfico correspondiente, que aparece en la Página 41

```
pasajeros_2010 = trafico_2010$pasajeros_2010
pasajeros_2010 = as.numeric(pasajeros_2010)
pasajeros_2010_ts = ts(pasajeros_2010, start = c(2010, 1), frequency = 12)
```

```
# Serie histórica (convertir a data frame para ggplot)
df_hist_2010 <-data.frame(fecha=as.Date(as.yearmon(time(pasajeros_2010_ts))),
valor = as.numeric(pasajeros_2010_ts))

ggplot() +
geom_line(data = df_hist_2010, aes(x = fecha, y = valor, color = "Serie histórica")) +
geom_ribbon(data = df_pred, aes(x = fecha, ymin = li, ymax = ls, fill = "IC 95%(bootstrap)"), alpha = 0.4) +
geom_line(data = df_pred, aes(x = fecha, y = media, color = "Predicciones"), linewidth = 1) +
scale_color_manual(values=c("Serie histórica"="black","Predicciones"= "blue"))+
scale_fill_manual( values=c("IC 95%(bootstrap)"="skyblue"))+
scale_x_date(date_breaks="1 year", date_labels="%Y")+
labs(title = "",x = "Fecha", y = "Pasajeros",color="",fill="",
caption = "") +
theme_minimal()+
theme(legend.position = c(0.02, 0.98),
legend.justification = c(0, 1),   # Ancla la esquina superior izquierda de la leyenda
legend.box.background = element_rect(colour = "black", linewidth = 0.5))
```

Bibliografía

[1] Shumway, R. H. and Stoffer, D. S. (2016). *Time Series Analysis and Its Applications: With R Examples*. 4th ed. Springer, Cham.

[2] Peña, D., Tiao, G. C. y Tsay, R. S. (eds.) (2001). *A Course in Time Series Analysis*. Wiley-Interscience, John Wiley & Sons, Inc., New York.

[3] Dickey, D. A. and Fuller, W. A. (1979). *Distribution of the Estimators for Autoregressive Time Series With a Unit Root*. Taylor & Francis.

[4] Kwiatkowski, D. and Phillips, P. C. B. and Schmidt, P. and Shin, Y. (1992) *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?* Elsevier.

[5] Ljung, Greta M. and Box, George E. P. (1978). *On a Measure of Lack of Fit in Time Series Models*. Oxford University Press.

[6] Engle, Robert F. (1982). *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation*. Wiley.

[7] Shapiro, S. S. and Wilk, M. B. (1965). *An Analysis of Variance Test for Normality (Complete Samples)*. Oxford University Press.