UNIVERSIDAD DE GRANADA

Escuela Internacional de Posgrado

MÁSTER EN ESTADÍSTICA APLICADA

TRABAJO FIN DE MÁSTER

(MODELIZACIÓN DE SERIES TEMPORALES, MODELOS CLÁSICOS Y SARIMA)

Presentado por:

Dª. NATALIA LAZAREVA

Tutor:

PROF. FRANCISCO JAVIER ALONSO MORALES

CURSO ACADÉMICO 2023/2024

RESUMEN

El objetivo de este trabajo es analizar los conceptos esenciales de las series temporales, diferenciando entre series estacionarias y no estacionarias, así como entre series estacionales y no estacionales. Se aborda la identificación del modelo, la estimación de sus parámetros, los métodos para validar su adecuación y los pasos para predecir valores futuros a partir de la serie original.

En la segunda parte se presenta el conjunto de datos reales de la temperatura media mensual, proporcionados por el Gobierno de la isla Jersey. Se describe detalladamente el proceso de ajuste del modelo $SARIMA(0,1,4)(3,0,3)_{[12]}$ y se comparan las predicciones con las mediciones observadas y con las predicciones del modelo subóptimo $SARIMA(1,0,1)(0,1,1)_{[12]}$.

Con este trabajo se pretende estudiar los conceptos clave de la modelización utilizando modelos $SARIMA$ y demostrar su aplicación con el software estadístico RStudio.

Palabras clave: SARIMA, análisis de series temporales, metodología Box-Jenkins, RStudio

ABSTRACT

The aim of this work is to analyze the essential concepts of time series, differentiating between stationary and non-stationary series, as well as between seasonal and non-seasonal series. The identification of the model, the estimation of its parameters, the methods to validate its adequacy, and the steps to predict future values from the original series are addressed.

In the second part, a real dataset of monthly average temperature, provided by the Government of the island of Jersey, is presented. The process of fitting the $SARIMA(0,1,4)(3,0,3)_{[12]}$ model is described in detail and the predictions are compared with the observed measurements and with the predictions of the suboptimal $SARIMA(1,0,1)(0,1,1)_{[12]}$ model.

The aim of this work is to study the key concepts of modeling using $SARIMA$ models and to demonstrate their application with the statistical software RStudio.

Key words: SARIMA, time series analysis, Box-Jenkins methodology, RStudio

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

Los datos temporales, mejor conocidos como *series temporales*, se utilizan en una gran variedad de campos de la sociedad moderna: economía, ingeniería, medicina, meteorología, etc. Por esta razón, es fundamental poder analizarlos adecuadamente para sacar las conclusiones y predicciones fiables y útiles para el propósito del análisis (por ejemplo, identificación de la tendencia, ciclos y relaciones de los datos, etc). ¿Pero qué es exactamente una serie temporal? En este caso, se refiere a una sucesión de observaciones, ordenada cronológicamente, cada una recolectada en un momento específico.

La metodología propuesta por Box y Jenkins (1970) [2] es el enfoque fundamental para el análisis de los datos cronológicos, que permite **identificar** el modelo, **estimar** sus coeficientes, **validarlo** y **predecir** los valores futuros. Para ello, se utilizan los modelos $ARIMA(p,d,q)$, *AutoRegressive Integrated Moving Average* en inglés, donde $AR(p)$ indica la parte autorregresiva, $MA(q)$ la parte de medias móviles e $I(d)$ indica el órden de diferenciación en el caso de la presencia del comportamiento no estacionario, es decir, sus propiedades estadísticas se varían a lo largo del tiempo. Como el caso particular, se destacan los modelos $SARIMA(p,d,q) \times (P,D,Q)_{[s]}$, que se utilizan para modelizar los datos con el comportamiento estacional (por ejemplo, datos de temperatura).

El objetivo del presente trabajo es estudiar los conceptos fundamentales de las series temporales estacionarias y no estacionarias, estacionales y no estacionales, su modelización, estimación de los parámetros del modelo identificado, los métodos para validar su adecuación y los pasos necesarios para calcular los valores futuros a partir de la serie inicial (Capítulo 2). En el Capítulo 3 se describe la aplicación de la metodología estudiada anteriormente sobre el conjunto de datos reales con el fin de demostrar el proceso de modelización de las series temporales estacionales con el software estadístico RStudio y comparar las predicciones obtenidas con los datos reales.

El conjunto de datos utilizados corresponde a las mediciones de la temperatura media mensual entre enero de 1981 y abril de 2022 en la isla Jersey, proporcionadas por el Gobierno de la propia isla. Para comparar, se utilizan los datos de la temperatura media mensual entre abril de 2022 y abril de 2024, también proporcionados por el Gobierno de Jersey, a pesar de que no formen parte del conjunto de datos inicial.

2. MARCO TEÓRICO

2.1. SERIES TEMPORALES

Una *serie temporal* (*cronológica*, *de tiempo* o simplemente una *serie*) es una sucesión ordenada de observaciones z_t, cada una grabada en un momento específico t [3]. El estudio de una serie temporal tiene como objetivo modelizar la evolución de una variable a lo largo del tiempo, y tiene multitud de aplicaciones en los campos de empresa (por ejemplo, predicción de las ventas), economía (estudio de la inflación), agricultura (seguimiento de plagas y enfermedades), ingeniería (control de calidad), etc.

Las series temporal pueden clasificarse según la *naturaleza* del parámetro temporal:

1. **Discretas**. Es una sucesión de observaciones de una variable que se mide en intervalos de tiempo discretos, por lo general igualmente espaciados.
2. **Continuas**. Es una sucesión de observaciones de una variable que se mide en intervalos de tiempo continuos, es decir, las observaciones se realizan en cualquier momento dentro de un intervalo de tiempo especificado.

En función del *número de variables* incluidas en el modelo:

1. **Univariantes**. Se analiza solo una magnitud a lo largo del tiempo.
2. **Multivariantes**. Se analizan dos o más magnitudes conjuntamente . Uno de los objetivos, en este caso, puede ser buscar una relación entre ellas.

En función de la *estacionariedad*:

1. **Estacionarias**. Las propiedades estadísticas (por ejemplo, momentos, distribuciones, etc.) se mantienen constantes a lo largo del tiempo.
2. **No estacionarias**. La serie muestra distintas medias locales, variabilidades, etc.

En función de la presencia del *comportamiento estacional* [13]:

1. **Estacional**. Una serie temporal que muestra un patrón regular de comportamiento que se repite a lo largo del tiempo.
2. **No estacional**. Una serie temporal que no presenta ningún patrón regular de comportamiento a lo largo del tiempo.

En función de la *linealidad*:

1. **Lineales**. La serie se genera a través de la observación de un proceso lineal, definido matemáticamente por modelos lineales.
2. **No lineales**. Algunas de las series requieren una modelización no lineal, por ejemplo, series temporales bilineares.

Para ilustrar todo lo descrito anteriormente, se propone un ejemplo de la serie temporal que describe la evolución semanal del índice M2 de la Junta de Gobernadores del Sistema de la Reserva Federal (EE.UU.) entre el 1 de enero de 2021 y el 1 de enero de 2024 (Figura 1).

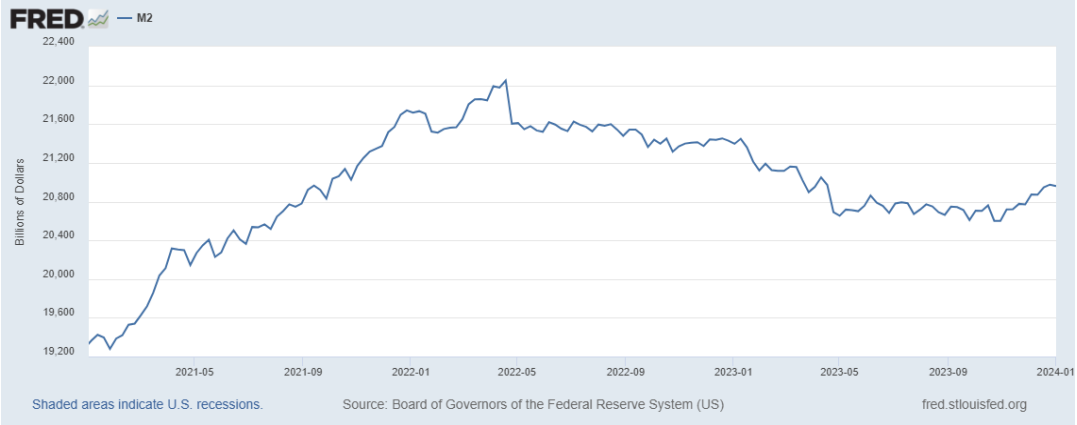

Figura 1: Junta de Gobernadores del Sistema de la Reserva Federal (EE.UU.), M2 [WM2NS], retirado de FRED, Banco de la Reserva Federal de St. Louis; https://fred.stlouisfed.org/series/WM2NS, 10 de febrero de 2024.

La serie del gráfico anterior es *discreta*, ya que ilustra una sucesión de los valores del índice M2, registrados semanalmente. Al observar sólo una serie en la Figura, se afirma que la serie es *univariante*. Aunque es necesario realizar las pruebas correspondientes, ya en este etapa se observa que las propiedades estadísticas no se mantienen constantes a lo largo del tiempo, entonces la serie será *no estacionaria*. Debido que algún comportamiento repititivo tampoco se observa visualmente, la serie será *no estacional*.

2.2. PROCESOS ESTOCÁSTICOS

2.2.1. DEFINICIÓN Y EL CONCEPTO DE ESTACIONARIEDAD

Un *proceso estocástico* es una familia $Z = \{Z_t, t \in T\}$ de las variables aleatorias, indexadas por un conjunto T y definidas en algún espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, donde el conjunto de índices T será un parámetro temporal (un intervalo o un conjunto de valores discretos) y un subconjunto de $\mathbb{R}$ [15]. Para cada $\omega \in \Omega$ la aplicación $Z(\cdot\, \omega)$ se denomina *trayectoria* o *realización* del proceso, y $Z(t,\cdot)$ es una variable aleatoria, así que una serie

temporal será una realización de un proceso estocástico, no obstante, en la práctica el parámetro ω suele suprimirse, y así el proceso estocástico se denota como $\{Z_t, t \in T\}$.

El proceso estocástico Z puede describirse mediante una función de distribución *n*-dimensional, aunque es poco práctico. Se define el conjunto finito de las variables del proceso: $\{Z_{t_1}, \dots, Z_{t_n}\}$, a partir del cual, se considera la función de distribución *n*-dimensional como:

$$F(z_{t_1}, \dots, z_{t_n}) = P[\omega, Z(\omega, t_1) \leq z_{t_1}, \dots, Z(\omega, t_n) \leq z_{t_n}], \forall t_1, \dots, t_n \in T.$$

Tal y como se ha mencionado antes, una serie temporal se llama estrictamente *estacionaria* si sus propiedades estadísticas son invariantes frente al tiempo. De forma matemática, se expresa como:

$$F(z_{t_1}, \dots, x_{t_n}) = F(z_{t_1+k}, \dots, z_{t_n+k}), \forall t_1, \dots, t_n \in T, k \in \mathbb{R},$$

donde $t_1 + k, \dots, t_n + k \in T$. Expresando con distintas palabras, si la distribución conjunta de $(z_{t_1}, \dots, z_{t_n})$ y $(z_{t_1+k}, \dots, z_{t_n+k})$ coincide para todos los $k \in \mathbb{R}$ y $n > 0$. Sin embargo, la condición de estacionariedad estricta es muy difícil de aplicar en la práctica, por lo cual, suele considerarse sólo el cumplimiento de la estacionariedad *débil* o de *segundo orden.* Una serie temporal se considera débilmente estacionaria si la media y la varianza de la serie no cambian con el tiempo y si la covarianza entre dos puntos en la serie depende solo de la distancia entre los puntos y no de su ubicación específica en el tiempo. Un proceso $\{Z_t, t \in T\}$ estrictamente estacionario que tiene los momentos de segundo orden también es un proceso estacionario en el sentido débil, pero no a la inversa. No obstante, existe un caso particular de los procesos *gaussianos* o *normales* que se quedan completamente determinados especificando sus momentos de primer y segundo orden. En este caso específico, la condición de la estacionariedad fuerte es equivalente a la estacionariedad débil.

Para un proceso $\{Z_t, t \in T\}$ se definen las siguientes funciones:

1. *Media*. $\mu_t = E(Z_t), t \in T$.
2. *Varianza*. $\sigma_t^2 = Var(Z_t) = E(Z_t - \mu_t)^2, t \in T$.
3. *Covarianza* entre Z_t y Z_{t+k}. $\gamma_{t,t+k} = Cov(Z_t, Z_{t+k}) = E[(Z_t - \mu_t)(Z_{t+k} - \mu_{t+k})], \forall t, t + k \in T$.
4. *Correlación* entre Z_t y Z_{t+k}. $\rho_{t,t+k} = \frac{\gamma_{t,t+k}}{\sqrt{\sigma_t^2 \sigma_{t+k}^2}}, \forall t, t + k \in T$.

Para un proceso estacionario se verifica que las funciones de media y varianza son constantes ($\mu_t = \mu$ y $\sigma_t^2 = \sigma^2$) y que las funciones de covarianza y correlación dependen únicamente de la diferencia del tiempo, es decir, varía en función de la diferencia entre los instantes *t* y *t+k* y no de los puntos donde se calculan.

Anteriormente se han definido las funciones de covarianza y correlación entre Z_t y Z_{t+k}. Sin embargo, en la práctica suelen reescribirse de la siguiente forma, dado que en el caso estacionario dependen exclusivamente de la diferencia entre dos puntos, z_t y z_{t+k}:

$$\gamma_k = E[(Z_t - \mu)(Z_{t+k} - \mu)],$$

$$\rho_k = \frac{\gamma_{t,t+k}}{\sqrt{\gamma_{t+k,t+k}\gamma_{t,t}}} = \frac{\gamma_k}{\gamma_0} \text{ [14].}$$

Estas funciones suelen llamarse *función de autocovarianza* y *función de autocorrelación simple* y juegan un papel importante en la descripción del comportamiento de las series temporales.

2.2.2. FUNCIONES DE AUTOCORRELACIÓN SIMPLE (F.A.S.), AUTOCORRELACIÓN PARCIAL (F.A.P), AUTOCOVARIANZA (F.A.C.)

Se considera un proceso estocástico estacionario $\{Z_t, t \in T\}$, que tiene asociada la siguiente función de autocovarianza:

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E[(Z_t - \mu)(Z_{t+k} - \mu)], \forall t, t+k \in T.$$

Esta función compara la serie temporal con una réplica de sí misma desplazada en el tiempo. Adopta una forma continua para series de tiempo continuas y una forma discreta para series de tiempo discretas. Para este proceso también se define una función de autocorrelación simple:

$$\rho_k = \frac{\gamma_{t,t+k}}{\sqrt{\gamma_{t+k,t+k}\gamma_{t,t}}} = \frac{\gamma_k}{\gamma_0}, \forall t, t+k \in T\,.$$

Dicha función mide la asociación lineal de la serie en el tiempo *t+k*, usando sólo el valor del tiempo *t*. Varía entre -1 y 1, donde 1 significa que se puede predecir perfectamente el valor $t+k$ basándose en el valor t [14].

Como se ha comentado, la función de autocorrelación simple mide el grado de asociación de la serie en el momento $t+k$ a partir del momento t. Si se desea medir la

correlación entre dos variables separadas por k momentos cuando no se considera la dependencia creada por los retardos intermedios existentes entre ambas, se define la función de autocorrelación parcial (f.a.p.):

$$\phi_{kk} = corr(Z_t - a_1 Z_{t-1} - \cdots - a_{k-1} Z_{t-k+1}, Z_{t-k} - l_1 Z_{t-1} - \cdots - l_{k-1} Z_{t-k+1}).$$

En este caso, ϕ_{kk} se calcula como la correlación entre los residuos de las regresiones lineales sobre $Z_{t-1}, \dots, Z_{t-k+1}$. Sin embargo, en la práctica para su cálculo se utiliza la siguiente expresión:

$$\phi_{kk} = \frac{\begin{vmatrix} \rho_0 & \rho_1 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & \rho_0 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} \rho_0 & \rho_1 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & \rho_0 \end{vmatrix}}, k = 2,3,..$$

Este resultado se obtiene, resolviendo la ecuación de Yule-Walker, que viene dada por:

$$\rho_i = \sum_{j=1}^{k} \phi_{jk}\, \rho_{i-j}, i = 1,2, \dots, k.$$

Resolviendo estas ecuaciones para $k = 1,2,3,$ ..., se obtiene el resultado de ϕ_{kk} anterior. De esta forma, las propiedades más importantes de las tres funciones serán:

1. $\gamma_0 = \sigma^2$ y $\rho_0 = 1$.

2. $|\gamma_k| \leq \gamma_0$ y $|\rho_k| \leq 1$.

3. $\gamma_k = \gamma_{-k}$ y $\rho_k = \rho_{-k}$.

4. γ_k y ρ_k son semidefinidas positivas.

5. $\phi_{11} = \rho_1$.

En la Figura 2 y Figura 3 se presenta un ejemplo de la ilustración gráfica de la función de autocorrelación simple y función de autocorrelación parcial.

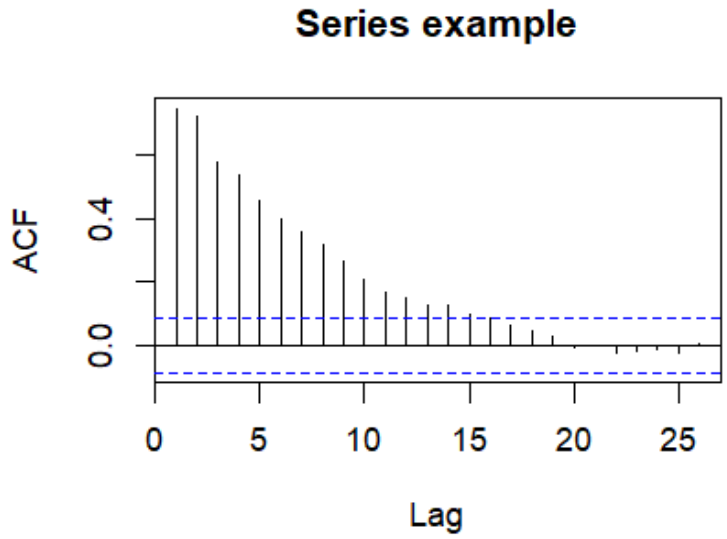

Figura 2: Función de autocorrelación de los datos simulados. Fuente: Elaboración propia.

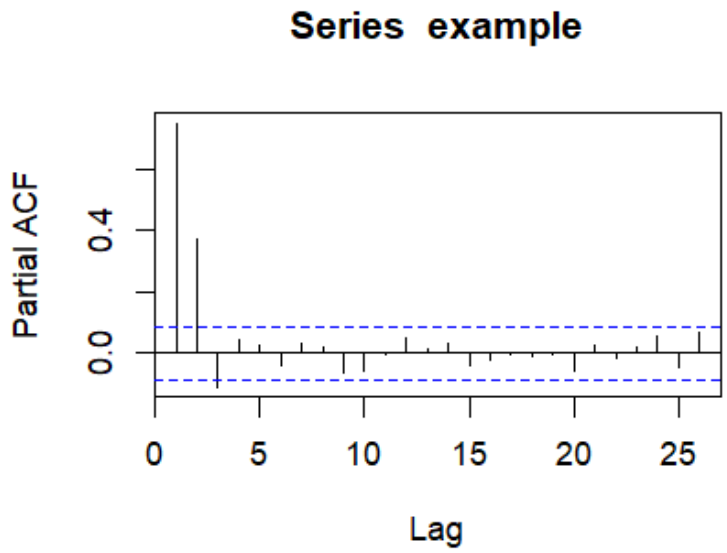

Figura 3: Función de autocorrelación parcial de los datos simulados. Fuente: Elaboración propia.

2.2.3. PROCESO DE RUIDO BLANCO

Un caso particular de los procesos estocásticos es el proceso de *ruido blanco*, una sucesión de variables aleatorias que suele denotarse como a_t y verifica siguientes propiedades:

1. $E(a_t) = 0 = cte., \forall t \in T$.

2. $Var(a_t) = \sigma_t^2 = cte., \forall t \in T$.

3. $Cov(a_t, a_{t+k}) = 0, \forall t \neq t+k \in T$.

4. $\sigma_k = \begin{cases} \sigma_k^2, & k = 0 \\ 0, & k \neq 0 \end{cases}$.

Por lo tanto, su f.a.s. y f.a.p. son iguales a:

$$\begin{cases} \rho_k = 1, k = 0 \\ \rho_k = 0, k \neq 0 \end{cases}$$

$$\begin{cases} \phi_{kk} = 1, k = 0 \\ \phi_{kk} = 0, k \neq 0 \end{cases}$$

Como bien dice el propio nombre, dichos procesos se caracterizan por la ausencia del patrón del comportamiento. Aunque este tipo de procesos es poco frecuente en las series temporales aplicadas, tiene un papel clave a la hora de análisis de las series de tiempo. Por ejemplos, los retornos de los activos financieros se suelen modelizar por estos procesos:

$$log\frac{Z_t}{Z_{t-1}} = logZ_t - logZ_{t-1} = a_t,$$

$$\nabla logZ_t = (1 - B)logZ_t.$$

Un ejemplo de tal proceso aparece en la Figura 4.

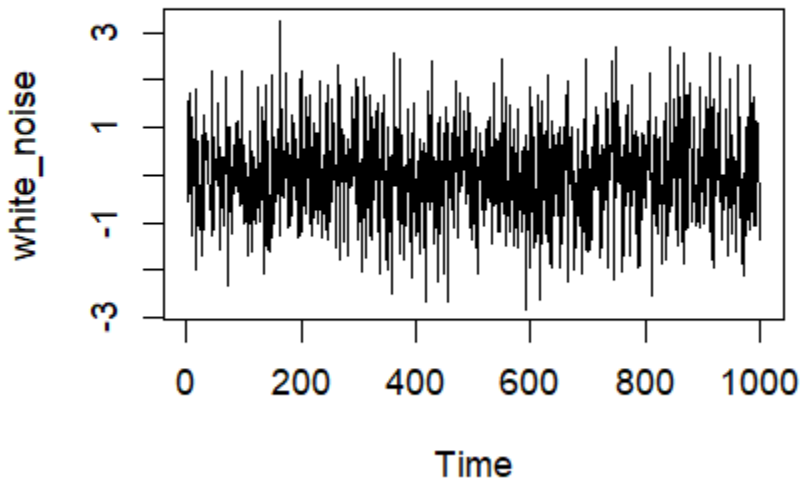

Figura 4: El proceso de ruido blanco simulado. Fuente: Elaboración propia.

2.2.4. METODOLOGÍA BOX-JENKINS

La metodología Box-Jenkins, desarrollada por los estadísticos George E. P. Box y Gwilym M. Jenkins, es un enfoque fundamental para modelar y pronosticar series temporales univariantes. Su objetivo es identificar un modelo $ARIMA$ (Autorregresivo Integrado de Media Móvil) que capture la estructura subyacente de la serie, permitiendo realizar análisis y predicciones precisas. No obstante, la eficacia de esta metodología se encuentra supeditada al cumplimiento de un conjunto de condiciones:

1. Estacionariedad: La serie temporal debe ser estacionaria, lo que implica que sus propiedades estadísticas (media, varianza, autocorrelación) permanecen invariables en el tiempo. De la misma manera, una serie se puede transformar en estacionaria, diferenciandola.

2. **Independencia de los residuos**: Los residuos del modelo, que representan la porción de la serie temporal no explicada por el modelo, deben ser independientes entre sí. La evaluación de la independencia se puede realizar mediante pruebas como, por ejemplo, el test de Ljung-Box.

3. **Normalidad de los residuos**: La distribución de los residuos del modelo debe ajustarse a una normal.

4. **Suficiencia de datos**: Se requiere una cantidad de datos suficiente para la predicción óptima. La mayoría de los resultados son asintóticos, por ello, la cantidad mínima varía en función del modelo que se vaya a utilizar, aunque generalmente se recomienda contar con al menos 50 observaciones.

5. **Ausencia de valores atípicos**: Los valores atípicos, que se definen como puntos de datos que se encuentran significativamente alejados del resto de la serie, pueden afectar la precisión del modelo. Es fundamental identificar y tratar estos valores antes del ajuste del modelo.

En el siguiente capítulo se desarrollan detalladamente los modelos autorregresivos (AR), de medias móviles (MA), autorregresivos de medias móviles ($ARMA$, la combinación de los dos anteriores), autorregresivos de medias móviles estacionales ($SARMA$, se observa el comportamiento estacional de la serie) y además, los modelos integrados (I) en los que aparece una transformación, *la diferenciación*, para obtener la diferenciación.

2.3. MODELOS ESTACIONARIOS

Los modelos estacionarios suponen que la serie temporal es generada por una agregación lineal de perturbaciones aleatorias [2]. Si sólo se busca explicar el valor de una variable observada con dependencia temporal en un momento dado (t) utilizando información sobre su comportamiento en el pasado, tal proceso se denomina *un proceso lineal general.* Dicho proceso demuestra la evolución del proceso estacionario en función de la suma de los valores actuales y pasados del ruido blanco (a_t) ponderado (ψ_t), y se expresa como

$$\dot{Z}_t = a_t + \psi_1 a_{t-1} + \psi_1 a_{t-2} + \cdots = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j},$$

donde $\dot{Z}_t = Z_t - \mu$ es la desviación del procesos de un origen (de media en el caso estacionario). Las propiedades del proceso de ruido blanco a_t ya se han mencionado en la Sección 2.2.3. Con respecto a los coeficientes ψ_j, para el proceso lineal anterior se verifica que $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

Se define el *operador retardo* como $B^j Z_t = Z_{t-j}$. De tal forma, dicho proceso puede expresarse de forma alternativa como

$$\dot{Z}_t = \psi(B) a_t,$$

donde $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$. Esta representación se denomina *medias móviles* (MA).

Puede imponerse la condición más estricta $\sum_{j=0}^{\infty} |\psi_j| < \infty$. En este caso, $\dot{Z}_t$ se define de forma alternativa como la suma ponderada de los $\dot{Z}_t$ pasados y el ruido blanco a_t añadido:

$$\dot{Z}_t = \pi_1 \dot{Z}_{t-1} + \pi_2 \dot{Z}_{t-2} + \cdots + a_t = a_t + \sum_{j=1}^{\infty} \pi_t \dot{Z}_{t-j}.$$

De forma equivalente, se expresa como

$$\pi(B) \dot{Z}_t = a_t,$$

donde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$. Este proceso se denomina *autorregresivo* (AR).

Sin embargo, no sería posible utilizar estos procesos como los modelos de series temporales debido al número infinito de parámetros que contienen. En práctica, se utilizan los

modelos con el número finito de parámetros p para poder estimar el número finito de observaciones disponibles. Entonces, la representación con el número p de parámetros (es decir, de orden p) del proceso AR será:

$$\pi_j = \begin{cases} \phi_j, & j = 1, \dots, p \\ 0, & j > p \end{cases}.$$

El proceso se escribe de forma

$$\dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p} = a_t.$$

El proceso MA de orden q se define de manera similar:

$$\psi_j = \begin{cases} -\theta_j, j = 1, \dots, q \\ 0, \quad j > q \end{cases}.$$

Y puede ser reescrito de la siguiente manera:

$$\dot{Z}_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

En la práctica, suele considerarse el modelo autorregresivo de medias móviles con dos parámetros, p y q. Sin embargo, se aconseja utilizar el modelo con el menor número de parámetros posible para aumentar la precisión del mismo. El modelo $ARMA(p, q)$ se expresa como:

$$\dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

A continuación se describen detalladamente cada uno de los modelos anteriores.

2.3.1. MODELOS AUTORREGRESIVOS AR(P)

El proceso estacionario autorregresivo AR de orden 1 viene definido como

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + a_t.$$

O bien, de forma compacta:

$$\phi_1(B)\dot{Z}_t = a_t,$$

donde $\phi_1(B) = 1 - \phi_1 B$. El modelo $AR(1)$ siempre es invertible y para que sea estacionario, las raíces de su polinomio característico deben estar fuera del círculo de unidad ($|\phi_1| < 1$). En un proceso autorregresivo, la perturbación en un momento dado afecta directamente al valor en

ese momento y también puede influir en los valores futuros a través de la relación entre los datos pasados y presentes.

Con respecto a las características del modelo, la función de autocovarianza para el modelo $AR(1)$ verifica $\gamma_k = \phi_1 \gamma_{k-1}$. Por lo tanto, la función de autocorrelación simple tiene la siguiente forma: $\rho_k = \phi_1^k$. La f.a.s. va decreciendo conforme va aumentando el k pero nunca alcanza el 0. Los valores de f.a.s. son positivos si $\phi_1 > 0$ y van alterando el signo para los valores negativos de ϕ_1. Por último, la función de autocorrelación parcial se corta tras el primer retardo y se expresa como

$$\phi_{kk} = \begin{cases} \rho_1, & k = 1 \\ 0, & k > 1 \end{cases}.$$

El caso del modelo de primer orden se puede generalizar al caso de p parámetros. Este parámetro indica a cuántos períodos del tiempo se retrocede en el pasado para estudiar el comportamiento de la serie a la hora llevar a cabo el análisis.

El proceso estacionario autorregresivo $AR(p)$ viene dado por:

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \cdots + \phi_p \dot{Z}_{t-p} + a_t.$$

O bien, de forma compacta:

$$\phi_p(B)\dot{Z}_t = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right) Z_t = a_t.$$

Obviamente, tiene las mismas propiedades descritas anteriormente, y sus características vienen dadas como:

f.a.c.: $\gamma_k = \phi_1 \gamma_{k-1} + \cdots + \phi_1 \gamma_{k-p}$,

f.a.s.: $\rho_k = \phi_1 \rho_{k-1} + \cdots + \phi_p \rho_{k-p}$,

f.a.p.: $\phi_{kk} = \begin{cases} \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_j}, & k \leq p \\ 0, & k > p \end{cases}$,

donde $\phi_{kj} = \phi_{k-1,j} - \phi_{kk} \phi_{k-1,k-j}$ para $j = 1,2,\ldots,k-1$. En este caso, la f.a.p., de forma análoga, se corta tras el retardo p, una propiedad que va a servir para identificar el modelo en la Sección 2.6.1.

En la Figura 5 se puede apreciar el gráfico de la evolución en el tiempo de un modelo $AR(1)$ con $\phi_1 = 0.5$ y su correcpondiente f.a.p. (Figura 6)

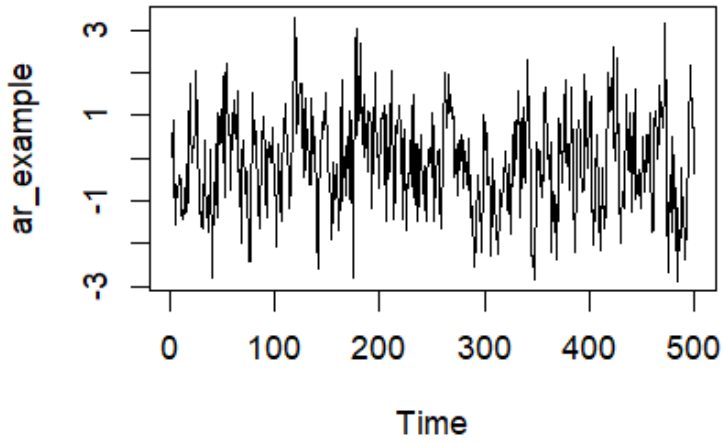

Figura 5: AR(1) con φ1=0.5 simulado. Fuente: Elaboración propia

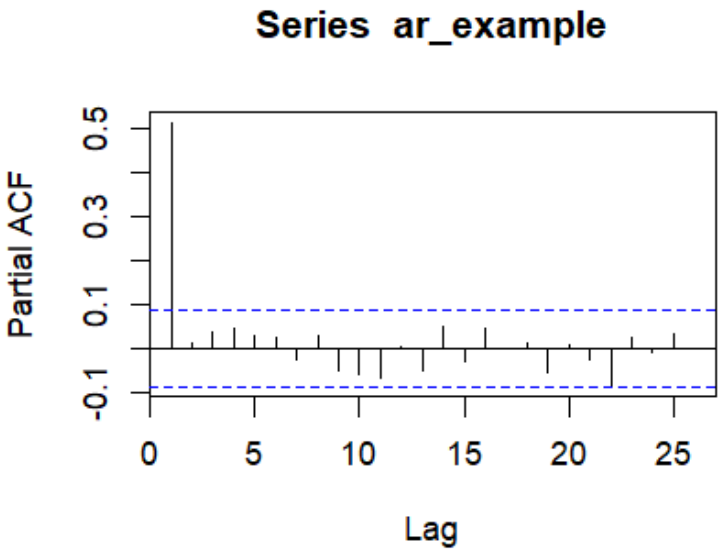

Figura 6: Función de autocorrelación parcial de AR(1) con φ1=0.9. Fuente: Elaboración propia.

2.3.2. MODELOS DE MEDIAS MÓVILES (MA)

El proceso estacionario de medias móviles $MA(1)$ viene dado por

$$\dot{Z}_t = a_t - \theta_1 a_{t-1} = (1 - \theta_1 B)a_t = \theta_1(B)a_t$$

En este caso, el modelo $MA(1)$ (Figura 7) siempre es estacionario y para que sea invertible, los parámetros del polinomio de medias móviles deben estar fuera del círculo de unidad, es decir, $|\theta_1| < 1$. Si se considera el mismo proceso $MA(1)$, significa que solo debe fijarse en un paso anterior en el tiempo para predecir el siguiente paso. La influencia de cualquier perturbación en el sistema se diluye rápidamente con el tiempo. Por ello, son útiles en la descripción de las series donde los sucesos producen efectos durante períodos de tiempo cortos.

En cuanto a las funciones características del proceso, se definen de la siguiente manera:

f.a.c.: $\gamma_k = \begin{cases} \gamma_0 = (1+\theta_1^2)\sigma_a^2, k = 0 \\ \gamma_1 = -\theta_1\sigma_a^2, \quad k = 1, \\ \gamma_k = 0, \quad k > 1 \end{cases}$

f.a.s.: $\rho_k = \begin{cases} \frac{-\theta_1}{1+\theta_1^2}, \ k = 1 \\ 0, \quad k > 1 \end{cases}$,

f.a.p.: $\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1-\sum_{j=1}^{k-1} \phi_{k-1,j}\rho_j}$,

donde $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}$ para $j = 1,2,\dots,k-1$.

La función de autocorrelación simple debe satisfacer que $|\rho_1| = |\theta_1|/(1+\theta_1^2) \leq \frac{1}{2}$ y se corta tras el retardo 1. Se verifica también que $|\phi_{kk}| < 0.5$. Este proceso también puede ser generalizado al caso de q parámetros:

$$\dot{Z}_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

Es una condición necesaria para que sea invertible si $\theta_1 + \theta_2 + \cdots + \theta_q < 1$. Sus funciones de autocovarianza y autocorrelación simple y parcial, que se definen de la siguiente manera:

f.a.c.: $\gamma_k = \begin{cases} \sigma_a^2 \sum_{j=0}^{q} \theta_j\theta_{j+k}, k \leq q \\ 0, \quad k > q \end{cases}$,

f.a.s.: $\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1+\theta_1^2+\dots+\theta_q^2}, \ k \leq q \\ 0, \quad k > q \end{cases}$,

f.a.p.: $\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_j}$

donde $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}$ para $j = 1,2,\dots,k-1$.

Las funciones de autocovarianza y autocorrelación simple se anulan tras el retardo q, (Figura 8) mientras que la función de autocorrelación parcial tiende a 0.

En la Sección 2.3 ya se ha mencionado la combinación de los modelos $AR(p)$ y $MA(q)$, modelo $ARMA(p,q)$, modelo autorregresivo de medias móviles que permite modelar series de tiempo que presentan dependencia temporal tanto en los valores pasados de la propia serie como en los errores pasados.

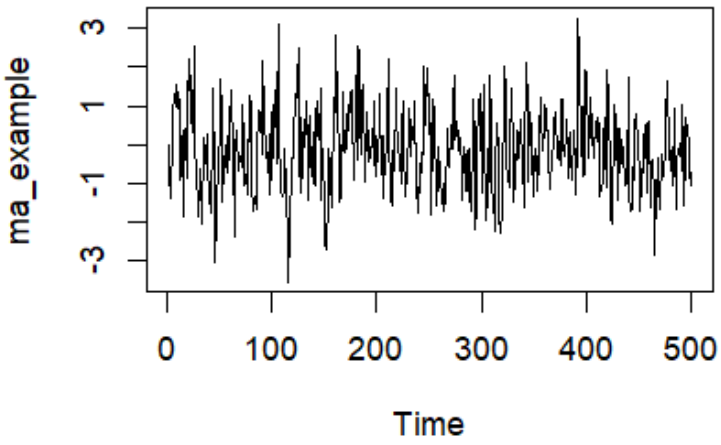

Figura 7: MA(1) con θ=0.4 para los datos simulados. Fuente: Elaboración propia.

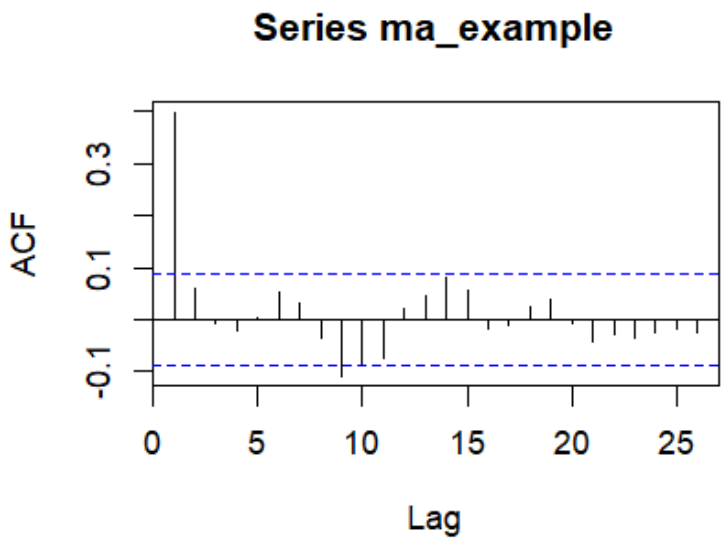

Figura 8: Función de autocorrelación simple del modelo MA(1) para los datos simulados. Fuente: Elaboración propia.

2.3.3. MODELOS AUTORREGRESIVOS DE MEDIAS MÓVILES ARMA(P,Q).

A la hora de llevar a cabo el análisis de series de tiempo, es imprescindible seguir el principio de *parsimonia*, es decir, utilizar el menor número de parámetros posible. No obstante, para alcanzar la parsimonia, en algunos casos resulta necesario considerar el modelo que incluye tanto los parámetros autorregresivos como de medias móviles. El modelo $ARMA(p,q)$ tiene la siguiente estructura:

$$\dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

La forma compacta de este modelo es:

$$\phi_p(B)\dot{Z}_t = \theta_q(B)a_t$$

donde $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ y $\theta_q(B) = 1 - \theta_q B - \cdots - \theta_q B^p$. Obviamente, para que el proceso sea estacionario, todas las raíces de $\phi(B) = 0$ deben estar fuera del círculo unidad, igual que para que el proceso se considere invertible, todas las raíces de $\theta(B) = 0$ también deben estar fuera del círculo unitario.

De forma similar, para el proceso $ARMA(p,q)$ se definen las funciones de autocovarianza y de autocorrelación simple y parcial para los retardos $k \geq q + 1$ se verifica que

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p},$$

$$\rho_k = \phi_1\rho_{k-1} + \cdots + \phi_p\rho_{k-p}.$$

El modelo $ARMA(p, q)$ comparte las características de los dos modelos, por lo cual, su funciones de autocovarianza y de autocorrelación simple no se anulan, sino decrecen sin anularse a partir de un retardo.

2.4. MODELOS NO ESTACIONARIOS

Para analizar una serie temporal con la metodología de Box-Jenkins, es imprescindible que la serie sea estacionaria tanto en media como en varianza. A pesar de eso, en la vida real es habitual enfrentarse con la serie que no es estacionaria por diversas razones (media no es constante, heterocedasticidad de la varianza, etc.). En este caso, se requiere llevar a cabo un estudio de la no estacionariedad para detectar el tipo y tomar unas acciones hacia la falta de estacionariedad para poder investigar la serie correctamente.

2.4.1. PROCESOS NO ESTACIONARIOS. MODELOS ARIMA.

Muchas de la series de tiempo evolucionan de tal manera que tienen la media no constante a lo largo del tiempo, aunque presenten la homocedasticidad de la varianza. Un ejemplo de tal serie se puede encontrar en la Figura 9.

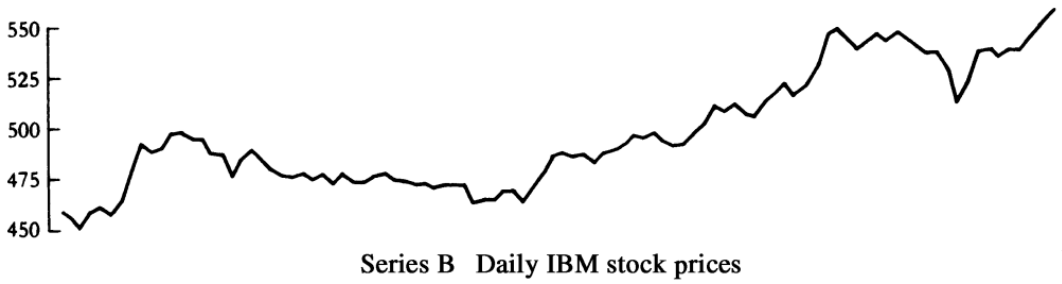

Figura 9: Precios diarios de las acciones de IBM. Retirado de Box, George E. P. el 05.03.2024

La presencia de una cierta tendencia en el comportamiento de la serie origina la no estacionariedad en media de dicha serie. Esta tendencia puede ser creciente o decreciente, exponencial o aproximadamente lineal, y puede clasificarse en dos clases: la tendencia determinística y la tendencia estocástica.

La tendencia *determinística*, a diferencia de la tendencia estocástica, se caracteriza por el comportamiento totalmente predecible, por ejemplo, los datos económicos que muestran un

crecimiento o un declive a lo largo del tiempo. El caso más simple es la tendencia determinista lineal:

$$Z_t = \alpha + \delta t + a_t,$$

aunque existen otros tipos: polinómica, exponencial, logarítmica, etc., dependiendo de la naturaleza de los datos. Como los datos no tienen la media constante, no es posible aplicar directamente la metodología clásica. Para realizar el pronóstico, se requiere una transformación de datos, o bien, utilizar los modelos que tienen en cuenta la tendencia determinista, por ejemplo, la regresión lineal.

Otro tipo de modelos con tendencia son los modelos con tendencia *estocástica*. En este caso, no resulta posible determinar con certeza el comportamiento de la serie, sino la tendencia está influenciada por los factores aleatorios. Un ejemplo de este tipo de tendencia es el comportamiento de la serie de los precios de acciones en el mercado financiero, que son influenciados por la multitud de factores aleatorios: eventos políticos, noticias económicas, etc. Sin embargo, en este caso sí que podemos aplicar la familia de modelos $ARMA(p, q)$, diferenciando la serie número d de veces apropiado para conseguir la estacionariedad en media. La diferenciación se realiza de la siguiente manera:

$$(1 - B)^d Z_t,$$

donde d es el órden de diferenciación regular, en otras palabras, el número de raíces unitarias del proceso, y suele tomar valores 0, 1, 2 (es recomendable diferenciar la serie el menor número de veces posible para evitar la pérdida de datos, el sobreajuste del modelo y el ruido en los datos). Dicho órden de diferenciación se introduce también en el modelo, y a este modelo se le denomina *autorregresivo de medias móviles integrado de órdenes* p, d y q, $ARIMA(p, d, q)$. A parte de corregir la falta de estacionariedad en media, ayuda a estabilizar la varianza de la serie.

Una vez diferenciando la serie tantas veces como sea necesario para alcanzar la estacionariedad en media, se define el modelo $ARIMA$ de órdenes $p,\ d,\ q$ como:

$$\phi(B)(1 - B)^d Z_t = \theta(B) a_t,$$

donde $\phi = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ es el operador AR estacionario y $\theta = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ es el operador MA invertible.

De esta forma, se concluye que cuando $d = 0$ el proceso es estacionario y no se requiere hacer la diferenciación. Los requisitos de estacionariedad e invertibilidad se aplican de manera independiente y, en general, los operadores $\phi(B)$ y $\theta(B)$ tengan el órden distinto. Uno de los métodos mayormente utilizados para la determinación del órden de diferenciación, es el test de Dickey-Fuller (Sección 2.6.1).

Uno de los casos particulares son los modelos $ARIMA(0,1,0)$ que se conocen también como *modelos del paseo aleatorio* ("*random walk*" en inglés) que está graficado en la Figura 10. Su estructura es la siguiente:

$$(1 - B)Z_t = a_t.$$

Su nombre viene de la propia naturaleza de este proceso: el valor actual depende únicamente del valor anterior más el error aleatorio. En el caso de que sea necesario considerar la tendencia determinista, en el modelo anterior se introduce una componente determinista θ_0:

$$Z_t = \theta_0 + Z_{t-1} + a_t.$$

En este caso, el modelo se le denomina *modelo de paseo aleatorio con deriva* ("*random walk with drift*" en inglés).

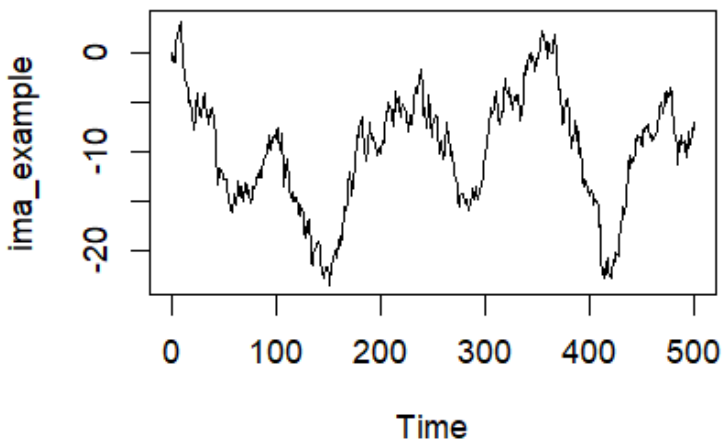

Figura 10: Modelo de paseo aleatorio simulado. Fuente: Elaboración propia.

2.4.2. FUNCIÓN DE AUTOCORRELACIÓN EXTENDIDA

Sin embargo, el proceso de identificación de órdenes de los modelos mixtos mediante f.a.s. y f.a.p. resulta bastante complicado. Por ello, Tsay y Tao (1984) [38] sugieren utilizar la función de autocorrelación extendida (f.a.e.) para mejorar la precisión de identificación de los órdenes del modelo $ARIMA(p,d,q)$:

$$\phi_p(B)\dot{Z}_t = \theta_q(B)a_t,$$

La idea principal detrás de esta función es que en un modelo $ARIMA(p,d,q)$ la parte AR tiene la estructura $MA(q)$, tal y como se nota en la ecuación anterior. Así, si se define $Z_t = \dot{Z}_t - \phi_1\dot{Z}_{t-1} - \cdots - \phi_p\dot{Z}_{t-p}$, su f.a.c. se anula para $k > q$. Este modelo puede expresarse como:

$$Z_t = \sum_{i=1}^{p} \phi_i Z_{t-i} - \sum_{j=1}^{q} \theta_j a_{t-j} + a_t.$$

Para estimar los parámetros de la parte AR de la expresión anterior, se puede aplicar los mínimos cuadrados ordinarios:

$$Z_t = \sum_{i=1}^{p} \hat{\phi}_i^{(0)} Z_{t-i} + \hat{e}_t^{(0)}.$$

En este caso, el superíndice (0) indica que es una regresión ordinaria. No obstante, dicha estimación será consistente únicamente en los procesos $ARIMA(p-d,d,0)$ o $ARIMA(0,d,q)$, lo que demostraron Mann y Wald (1943) [33] para el caso estacionario, mientras que el caso no estacionario fue demostrado por Tsay y Tao (1983) [37]. En el resto de los casos, se obtienen los estimadores inconsistentes, independientemente del tamaño de la muestra, lo que conlleva a una identificación del modelo errónea. Se considera ahora el caso de la regresión iterada primera, añadiendo un residuo retardado, que podría contener alguna información sobre el proceso:

$$Z_t = \sum_{i=1}^{p} \hat{\phi}_i^{(1)} Z_{t-i} + \hat{\beta}_1^{(1)} \hat{e}_{t-1}^{(0)} + \hat{e}_t^{(1)}.$$

Ahora se obtienen los estimadores consistentes para los procesos $ARIMA(p-d,d,1)$ o $ARIMA(0,d,q)$. De forma similar, se considera la segunda regresión iterada:

$$Z_t = \sum_{i=1}^{p} \hat{\phi}_i^{(2)} Z_{t-i} + \hat{\beta}_1^{(2)} \hat{e}_{t-1}^{(1)} + \hat{\beta}_2^{(2)} \hat{e}_{t-2}^{(0)} + \hat{e}_t^{(2)}.$$

En este caso, se obtienen los estimadores consistentes de los procesos $ARIMA(p-d,d,2)$ o $ARIMA(0,d,q)$. Generalizando a j iteracciones de un modelo $AR(k)$:

$$Z_t = \sum_{i=1}^{p} \hat{\phi}_i^{(k)} Z_{t-i} + \sum_{i=1}^{j} \hat{\beta}_i^{(j)} \hat{e}_{t-i}^{(j-i)} + \hat{e}_t^{(k)}.$$

De esta forma, los estimadores serán consistentes para los procesos $ARIMA(p-d,d,q)$ con $q \leq k$ o para los procesos $ARIMA(0,d,q)$ sin ninguna consideración de q.

A partir de las estimaciones consistentes anteriores, se construye la función de autocorrelación extendida (f.a.e. o EACF). Para determinar el órden p de f.a.e., se considera que el proceso Z_t sigue un modelo $ARMA(p,q)$, cuyas raíces de p pueden ser unitarias

Para el caso de $p = 0$, se suprime la parte autorregresiva del modelo $ARMA(p,q)$, y el proceso Z_t sigue ahora el modelo $MA(q)$:

$$\rho_j^{(0)} = 0, \qquad j > q.$$

En este caso, $\rho_j^{(0)}$ indica el coeficiente de autocorrelación ordinario de orden j de la serie Z_t. El superíndice 0 indica que no se ha realizado ninguna transformación sobre la serie Z_t, y $\rho_j^{(0)}$ constituye f.a.e. del orden 0.

Ahora se supone que el verdadero modelo de la serie original es $ARMA(p,q)$. Entonces, se define el coeficiente de autocorrelación de orden k:

$$\rho_j^{(k)} = \rho_j\left(W_k^{(j)}\right), \qquad j = 1,2,3,\ldots$$

donde

$$W_k^{(j)} = \phi_k^{(j)}(B) Z_t.$$

Para $k = p, j \geq q$, $W_k^{(j)}$ sigue un modelo $ARMA(p,q)$, la f.a.e. de orden p será igual a:

$$\rho_j^{(k)} = \begin{cases} = 0, & j > q \; y \; k = p \\ \neq 0, & j = q \; y \; k = p \end{cases}.$$

De esta ecuación se puede concluir que la f.a.e., similar a f.a.s. de un proceso MA, se corta tras el retardo p, indicando los retardos significativos para el modelo.

Al principio se ha asumido que los valores p y q son los verdaderos para el modelo $ARMA$. No obstante, en la práctica se suele conocer de antemano estos órdenes, por lo cual, la igualdad $k = p$ no se cumple en todos los casos. De esta forma, se supone ahora que $k - p > 0$ (modelo sobreajustado, "*overfitting*"), y se calculan los $\rho_j^{(k)}$ correspondientes:

$$\rho_j^{(k)} = \begin{cases} c(k-p, j-q), & 0 \leq j-q \leq k-p \\ 0, & j-q > k-p > 0' \end{cases}$$

donde $c(k-p, j-q)$ es alguna constante distinta del 0, o una variable continua entre -1 y 1.

Tal y como se ha comentado al principio, f.a.e. se utiliza para identificar los órdenes de los modelos mixtos. Para ello, a partir de las funciones f.a.e. de distintos órdenes, se construye la tabla:

AR/MA	0	1	...	q	...	m
0	$\rho_1^{(0)}$	$\rho_2^{(0)}$	...	$\rho_{q+1}^{(0)}$	...	$\rho_{m+1}^{(0)}$
1	$\rho_1^{(1)}$	$\rho_2^{(1)}$	...	$\rho_{q+1}^{(1)}$	...	$\rho_{m+1}^{(1)}$
...	...	...	...	...	...	...
p	$\rho_1^{(p)}$	$\rho_2^{(p)}$	....	$\rho_{q+1}^{(p)}$	...	$\rho_{m+1}^{(p)}$
$p+1$	$\rho_1^{(p+1)}$	$\rho_2^{(p+1)}$	...	$\rho_{q+1}^{(p+1)}$	...	$\rho_{m+1}^{(p)}$

Tabla 1: Tabla de EACF. Extraído de Identificación automática mediante la función

Si el $\rho_j^{(k)}$ es significativo, por la simplicidad, suele marcarse como X. En el caso contrario, en la celda correspondiente se coloca un 0. Para identificar el modelo verdadero, se debe buscar en la tabla un triángulo de ceros. Las coordenadas de su vértice superior izquierdo indicarán los órdenes p y q del modelo.

2.4.3. TRANSFORMACIONES BOX-COX.

Otra consideración importante y necesaria para construir el modelo adecuado para una serie de tiempo, es la estacionariedad en varianza. Se refiere a la propiedad de que la variabilidad de una serie temporal permanece constante a lo largo del tiempo. En este caso, el problema no se suele resolver con una diferenciación de la serie, sino en algunas ocasiones sí se requiere llevar a cabo otro tipo de transformación apropiada. Para ello, el proceso Z_t se puede reescribir de la siguiente manera: $Var(Z_t) = f(\mu_t)$, donde f es una función apropiada. Dado

que se supone que la varianza de este proceso no es constante el problema se reduce a encontrar una función adecuada para conseguir la estabilidad de la varianza a lo largo del tiempo. Box y Cox en 1964 [26] propusieron una familia paramétrica de transformaciones más usada para los datos no procesados:

$$Z_t^{(\lambda)} = \begin{cases} \dfrac{Z_t^{(\lambda)} - 1}{\lambda}, & \lambda \neq 0 \\ logZ_t, & \lambda = 0 \end{cases},$$

donde λ determina la transformación considerada, y se denomina el parámetro de transformación. El parámetro se estima por la función de máxima verosimilitud, aunque en la práctica no se requiere la estimación exacta y se recomienda elegir la transformación simple y conocida ($\lambda = \pm 2, \pm 1, \pm 0.5, \pm 0.25, 0$) para facilitar la interpretación del resultado final. Dicha transformación se aplica sólo a las series positivas. En el caso de que haya algunos datos ceros o negativos, se puede añadir una constante positiva a todos los valores para que se conviertan en positivos, y después realizar una transformación Box - Cox. Sin embargo, Yeo y Johnson (2000) [40] propusieron otra familia de transformaciones de potencia, que puede ser aplicada sobre los datos negativos y ceros:

$$Z_t^{(\lambda)} = \begin{cases} \dfrac{(Z_t + 1)^{\lambda} - 1}{\lambda}, & si\ \lambda \neq 0, Z \geq 0 \\ \ln(Z_t + 1), & si\ \lambda = 0, Z \geq 0 \\ -\dfrac{(-Z_t + 1)^{(2-\lambda)} - 1}{2 - \lambda}, & si\ \lambda \neq 2, Z < 0 \\ -\ln(-Z_t + 1), & si\ \lambda = 2, Z < 0 \end{cases}.$$

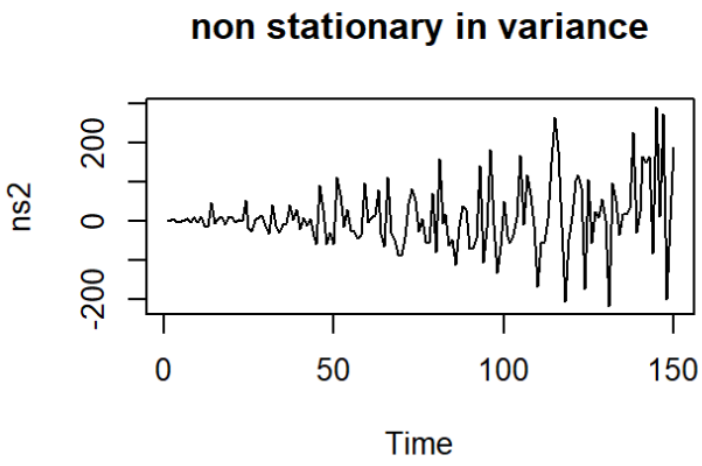

Figura 11: Ejemplo del proceso con la varianza no estable. Extraído desde https://web.vu.lt/mif/a.buteikis/wp-content/uploads/2018/02/TasksR_02.html el 30.03.2024

En el caso de que se necesita aplicar tanto la diferenciación como la transformación de Box-Cox, siempre se realiza primero la de Box-Cox, y luego se realiza la diferenciación de la serie con la varianza estabilizada [21].

2.5. MODELOS ESTACIONALES (SARIMA)

Muchas de las series temporales se observan varias veces a lo largo del año (Figura 12), es decir, tienen el comportamiento repetitivo a lo largo del tiempo. Por ejemplo, datos climatológicos (la temperatura, humedad, precipitaciones, etc.), agrícolas (producción de cultivos depende de la estación), de turismo (mayor cantidad de turistas en la temporada alta), etc. En este caso, existen dos tipos de relaciones: entre las observaciones consecutivas (comportamiento regular de la serie) y entre las observaciones que suceden con algunos desfases estacionales fijos (comportamiento estacional).

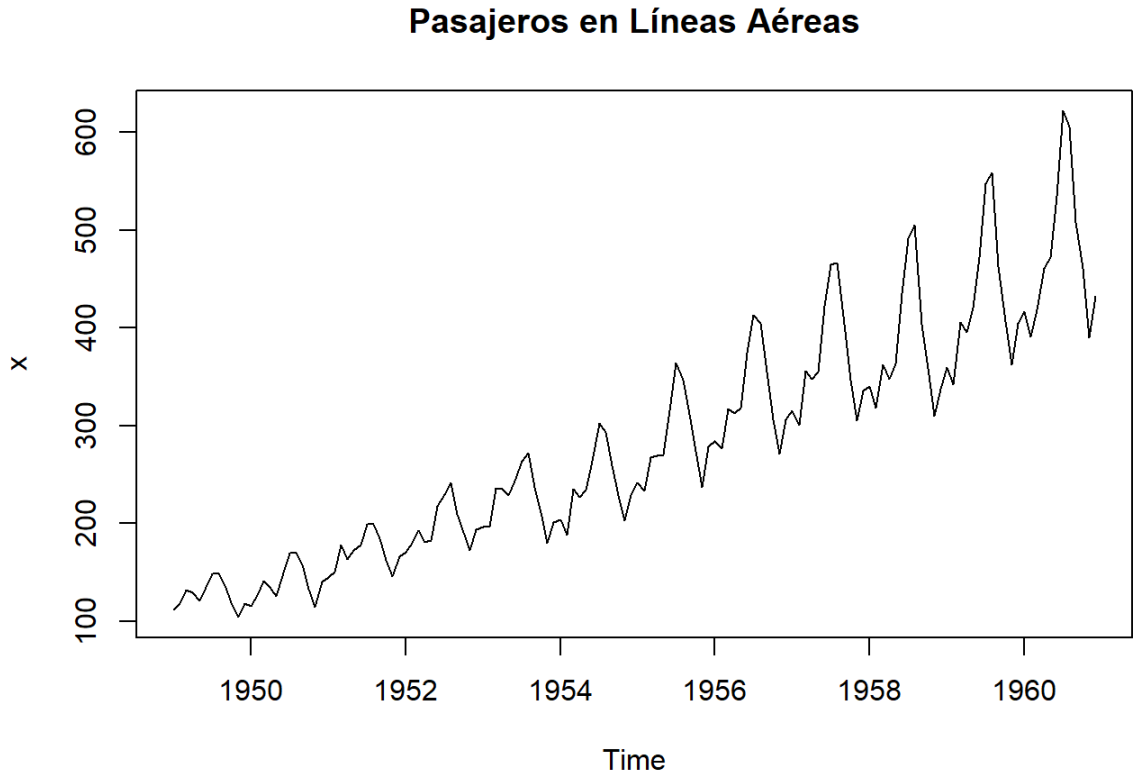

Figura 12: Ejemplo de la serie estacional. Retirado de https://bookdown.org/content/2274/series-temporales.html el 01.04.2024

La *estacionalidad* se define como la variación periódica de la serie que se repite cada s períodos de tiempo, donde s se define como el *período estacional*. En la serie temporal mensual este parámetro es igual a 12, en las trimestrales $s = 4$, etc. De esta forma, se define la extensión de la familia $ARIMA(p, d, q)$, los modelos $SARIMA(p, d, q) \times (P, D, Q)_{[s]}$ ($ARIMA$ estacional multiplicativo de Box - Jenkins), su expresión matemática es:

$$\Phi_p(B^s)\phi_p(B)(1 - B)^d(1 - B^s)^D \dot{Z}_t = \Theta_Q(B^s)\theta_q(B)a_t,$$

donde $\dot{Z}_t = Z_t - \mu$ si $d = D = 0$ y es igual a Z_t en otro caso. La fórmula se compone por:

1. El operador estacionario SAR : $\Phi_p(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP}$ (polinomio autorregresivo estacional)

2. El operador estacionario AR : $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_P B^p$ (polinomio autorregresivo)

3. El operador invertible SMA: $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \cdots - \Theta_Q B^{sQ}$ (polinomio de medias móviles estacional)

4. El operador invertible MA: $\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ (polinomio de medias móviles)

Para identificar el correcto modelo $SARIMA$, el primer paso es encontrar los órdenes de d y D hasta conseguir la estacionariedad de los datos (transformados o no). El siguiente paso consiste en la evaluación de las funciones ACF y PACF. Para encontrar los órdenes estacionales (P, Q) hace falta examinar los retardos que son múltiplos de s. Por ejemplo, para los datos mensuales, con $s = 12$, se examinan los retardos 12, 24, 36. Posteriormente se intenta determinar los órdenes ordinarios (p, q).

2.6. ETAPAS DE MODELIZACIÓN

En las secciones anteriores se ha introducido la familia de los modelos $ARIMA(p, d, q)$ y su extensión, los modelos $SARIMA(p, d, q) \times (P, D, Q)_{[s]}$, que describen las series estacionales. Sin embargo, como el objetivo final de la modelización de una serie temporal es la predicción óptima de su futuro comportamiento, se introducen cuatro etapas de la metodología Box-Jenkins (Figura 13):

1. **Identificación** de los parámetros
2. **Estimación** de los parámetros
3. **Validación** del modelo
4. **Predicción** de los valores futuros

Las herramientas más populares para llevar a cabo el análisis de series de tiempo son: Python (librerías NumPy, Pandas, Stats Models, etc.), R (los paquetes TSA, forecast, timeseries, etc.), Matlab, Statgraphics, SPSS, etc. En este trabajo se utiliza sólo el lenguaje R y su IDE RStudio.

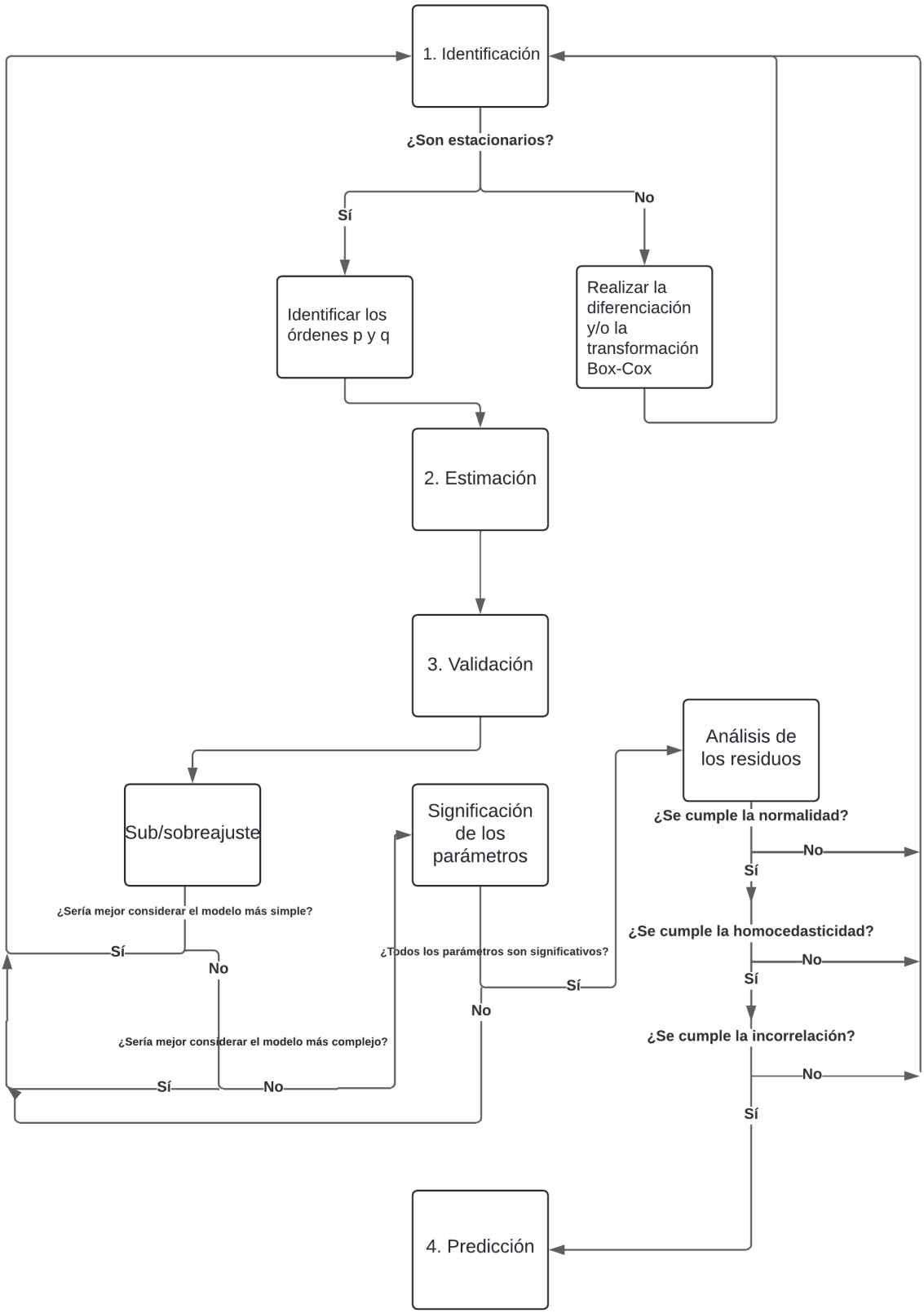

Figura 13: Etapas de modelización. Fuente: Elaboración propia.

2.6.1. IDENTIFICACIÓN

En las Secciones 2.3.1, 2.3.2, 2.3.3 y 2.5 ya se ha mencionado uno de los procedimientos claves para la identificación de los órdenes p y q del modelo. En la Sección 2.4.1 se habla del órden de diferenciación d en el caso de la no estacionariedad de los datos.

Antes de todo, se requiere visualizar la serie temporal con un gráfico simple y llevar a cabo un breve análisis descriptivo. De esta manera, se consigue conocer algunas características importantes de la serie, como, por ejemplo:

- *Valores nulos (NA)*: Los valores ausentes *(missing values* en inglés) pueden complicar significativamente el análisis de la serie. Si se detecta la presencia de los valores nulos, deben ser tratados correctamente.
- *Valores anómalos (Outliers)*: Analizando el gráfico de la serie y los valores mínimo y máximo de la serie, se puede detectar la presencia de los valores que desvían significativamente del resto y que también pueden afectar al resultado final.
- *Comportamiento estacional*. Al hacer la visualización de la serie y de su f.a.s. y f.a.p., es posible detectar si se evoluciona por ciclos o no (Figura 12).
- *Comportamiento estacionario de la serie*. En algunos gráficos la falta de estacionariedad se ve claramente (Figura 1 y Figura 9).
- *Presencia de la tendencia*. La tendencia, un caso particular de no estacionariedad en media, también es una de las cosas que pueden influir en el ajuste del modelo.
- *Intervenciones*. Son cambios significativos en el nivel de la serie, causados por factores externos o internos (por ejemplo, la influencia del COVID-19 sobre el sector de hostelería).
- *Autocorrelación positiva o negativa*. La autocorrelación positiva indica que los valores similares tienden a seguirse en la serie temporal, mientras que la autocorrelación negativa indica una tendencia inversa, donde los valores opuestos tienden a seguirse.

Tras efectuar el análisis descriptivo previo de los datos, es necesario comprobar que los datos cumplen las siguientes hipótesis claves:

1. *Verificar la estacionariedad en media y varianza de los datos*.

Tal y como se ha descrito anteriormente, una serie temporal es estacionaria en media siempre y cuando su media sea constante a lo largo de la evolución de la serie. Uno de los

indicadores de la falta de estacionariedad en media es el decrecimiento lento de la ACF de la serie. Sin embargo, si se desea apoyarse con una prueba más analítica, se realiza el contraste de *Dickey-Fuller Aumentado* (*ADF*). De nuevo se define el proceso $AR(p)$:

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \cdots + \phi_p \dot{Z}_{p-1} + a_t.$$

Este modelo puede ser reescrito como:

$$\Delta\dot{Z}_t = \beta \dot{Z}_{t-1} + \alpha_1 \Delta\dot{Z}_{t-1} + \cdots + \alpha_{p-1}\Delta\dot{Z}_{p-1} + a_t,$$

donde

$$\beta = \sum_{i=1}^{p} \phi_i - 1 \,; \alpha_i = \sum_{j=1}^{i} \phi_{p-i+j}.$$

El proceso $AR(p)$ tiene la raíz unitaria cuando $\sum_{i=1}^{p} \phi_i = 1$, se puede definir la hipótesis nula de existencia de la raíz unitaria como $H_0: \beta = 0$. De forma más práctica, el contraste para la prueba se define como:

$$H_0: \exists\ raíz\ unitaria\ para\ un\ nivel\ de\ confianza \equiv la\ serie\ no\ es\ estacionaria$$
$$H_1: \nexists\ raíz\ unitaria\ para\ un\ nivel\ de\ confianza \equiv la\ serie\ es\ estacionaria$$

La prueba ADF, a diferencia del test de Dickey Fuller original (1979) [29], permite incluir mayor número de retardos en el modelo, se considera el modelo $AR(p)$ en lugar del $AR(1)$, corrigiendo de forma paramétrica la correlación de orden superior.

Como la prueba alternativa al test ADF, se utiliza frecuentemente el test de *Kwiatkowski–Phillips–Schmidt–Shin* (*KPSS-test*) con las siguientes hipótesis:

$$H_0: la\ serie\ es\ estacionaria\ en\ tendencia$$
$$H_1: \exists\ raíz\ unitaria$$

Cada una de las pruebas se enfoca en dos aspectos distintos de la estacionariedad (raíz unitaria y la tendencia). Se puede usar ambos tests para la evaluación más completa y en los casos dudosos: si al menos alguna de las pruebas detecta la falta de la estacionariedad, entonces, se requiere diferenciar la serie (Sección 2.4.1).

La ausencia de la estacionariedad en media puede detectarse igualmente mediante la función de autocorrelación simple: en este caso, se observa su decrecimiento lento.

Una vez diferenciada la serie, hace falta repetir la prueba de estacionariedad para confirmar que ahora la serie es estacionaria. Si tras la primera diferenciación, no se consigue la estacionariedad en media, se debe volver a diferenciar la serie tantas veces cómo sea necesario. Sin embargo, existe el riesgo de la *sobrediferenciación*, es decir, la diferenciación excesiva, lo que conlleva el ajuste inapropiado del modelo y las predicciones inexactas.

La necesidad de aplicar la transformación Box-Cox sobre el modelo puede ser verificada con la función de R correspondiente: `BoxCox.lambda()`.

2. *Identificar los órdenes p y q del modelo regular y los órdenes P, D y Q de la parte estacional, si es necesario.*

Una vez diferenciado la serie *d* veces, es necesario determinar el resto de los parámetros de la serie. Para identificar los órdenes p y q (no suelen ser mayores que 3) de los modelos $MA(q)$ y $AR(p)$, se utilizan las funciones de autocorrelación simple y parcial. Las cuatro situaciones posibles están recogidas en la Tabla 2:

Proceso	f.a.s.	f.a.p.
MA(q)	Se corta tras el retardo q	Decrece rápidamente
AR(p)	Decrece rápidamente	Se corta tras el retardo p

Tabla 2: Relación entre los procesos AR/MA y f.a.s./f.a.p.

En el caso de considerar un modelo mixto $ARMA(p, q)$ o $ARIMA(p, d, q)$, la identificación de los órdenes se puede llevar a cabo tanto con los gráficos de f.a.s. y f.a.p. (más complicado y mayor grado de incertidumbre) (Tabla 2), como con la tabla de la función de autocorrelación extendida. La formulación matemática se describe en la Sección 2.3.3, mientras que el procedimiento de identificación consiste en encontrar el triángulo de "*ceros*", cuyas coordenadas del vértice superior izquierdo indican los órdenes p y q.

En algunas ocasiones, se sabe previamente que la serie tiene el comportamiento estacional debido a su naturaleza, por ejemplo, en los datos climatológicos. En el resto de los casos, se requiere verificar la presencia de la estacionalidad. Para ello, el procedimiento más frecuente es analizar el gráfico de la función de autocorrelación asociada. Por ejemplo, en la Figura 14 se detecta claramente el comportamiento estacional de la serie, puesto que los coeficientes de la f.a.s. para retardos múltiplos del período estacional (es decir, $s, 2s, 3s, etc.$) son significativamente distintos de cero. Dado que este conjunto de datos contiene las

mediciones mensuales, los retardos significativos son 12, 24, 36 y así sucesivamente.

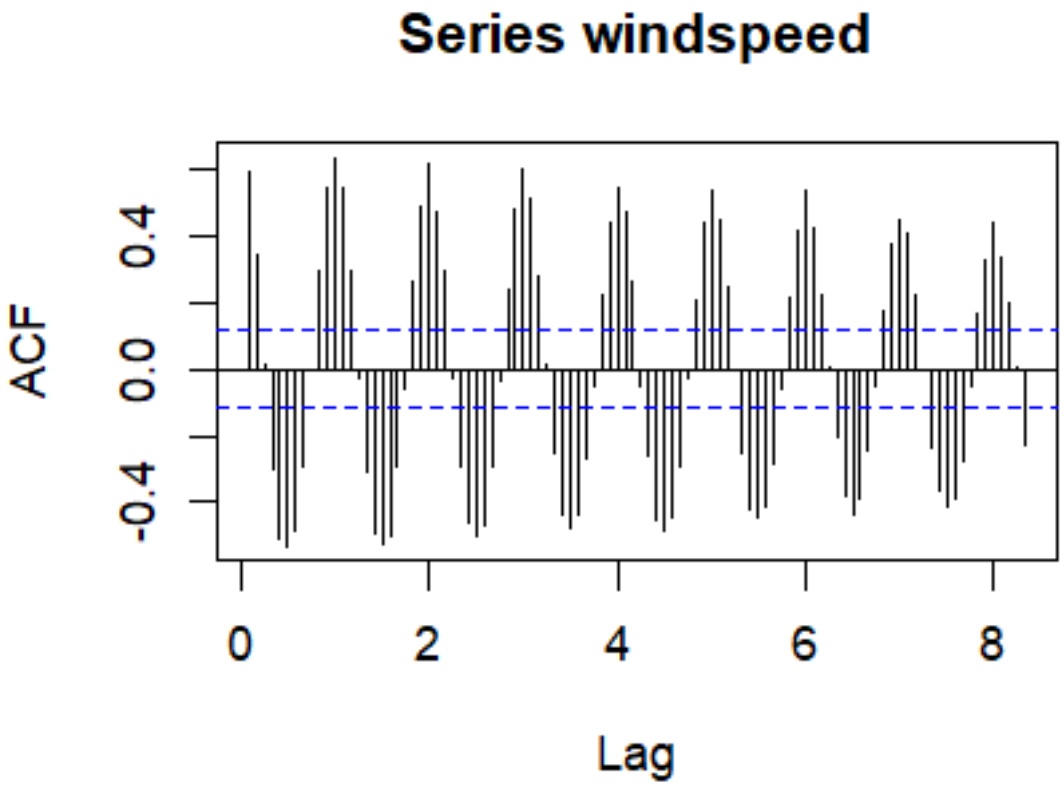

Figura 14: ACF de los datos de la velocidad del viento. Fuente: Elaboración propia

En el caso de haber identificado la presencia de la estacionalidad, se requiere identificar los órdenes *P*, *D* y *Q*. Para saber si habrá que diferenciar la parte estacional o regular, se recomienda probar ambas diferenciaciones y repetir la prueba de estacionariedad. Con respecto a los parámetros autorregresivos y de medias móviles, la identificación se realiza de forma similar a los modelos no estacionales, analizando el comportamiento de las funciones de autocorrelación simple y parcial. En cuanto a la inclusión del término independiente, se debe realizar el contraste con la hipótesis nula de que la media de la serie estacionaria es igual a cero frente a que esto no es cierto:

$$H_0: E(Z_t) = 0$$

$$H_1: E(Z_t) \neq 0.$$

Sin embargo, al ajustar el modelo en R, en el caso de que sea necesario, la constante se incluye automáticamente, si no se indica lo contrario.

3. *Selección del mejor modelo.*

Para seleccionar el modelo que mejor se ajuste a los datos, se suele utilizar dos criterios: *Criterio de Información de Akaike* (AIC) proporcionado por Akaike (1974) [26], *Criterio de Información Bayesiana* (BIC) de Schwarz (1978) [36] y *Criterio de Información de Akaike corregido* (AICc). Si los órdenes *d* y *D* son conocidos, el resto de parámetros se puede elegir mediante dichos criterios:

$$AIC = -2\log(L) + 2(p+q+P+Q+k),$$
$$BIC = -2\log(L) + (p+q+P+Q) + \ln(n),$$
$$AICc = AIC + \frac{2(p+q+P+Q+1)(p+q+P+Q+2)}{n-p-q-P-Q-2},$$

donde $k = 1$ si $c \neq 0$ y 0 en el caso contrario y c es la constante del modelo; L es la verosimilitud maximizada del modelo ajustado a los datos diferenciados; n es el número de observaciones. Sin embargo, los AICs de distintos niveles de diferenciación no son comparables, puesto que la verosimilitud del modelo completo para Z_t no está definido realmente.

El primer término de las fórmulas de AIC y BIC corresponde a $-2/n$ veces del logaritmo de la verosimilitud maximizada, mientras que el segundo término se conoce como "el factor de penalización" por la inclusión de los parámetros adicionales en el modelo. El objetivo es seleccionar el modelo que minimice dichos criterios. Se puede destacar que BIC tiende a elegir los modelos más pequeños debido a que su factor de penalización es más grande en comparación con el AIC. En segundo lugar, cabe destacar que el criterio BIC es más adecuado para las series grandes, mientras que el AICc suele ser más preciso en las series más pequeñas.

Una vez seleccionado el modelo $(S)ARIMA$ adecuado, se requiere estimar sus parámetros.

2.6.2. ESTIMACIÓN

En la segunda etapa del análisis de las series de tiempo, el principal interés se centra en la estimación de los parámetros del modelo. Se considera un modelo $ARMA(p,q)$:

$$\dot{Z}_t - \phi_1\dot{Z}_{t-1} - \cdots - \phi_p\dot{Z}_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

Se asume que la serie tiene n observaciones, se conocen los parámetros p y q, y el objetivo de este paso es estimar $\phi_1, \dots, \phi_p$ $\theta_1, \dots, \theta_q$, μ_t y σ_a^2. A continuación se describen varias técnicas de estimación de estos parámetros.

2.6.2.1. MÉTODO DE LOS MOMENTOS (DE YULE-WALKER)

La idea principal de este método es igualar los momentos poblacionales a los momentos muestrales. La ecuación obtenida se resuelve en términos de los momentos muestrales para los parámetros buscados. Se considera el caso del modelo $AR(p)$, para el cual el método produce los estimadores eficientes (óptimos):

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \cdots + \phi_p \dot{Z}_{p-1} + a_t.$$

Para estimar los parámetros ϕ_p y σ_a^2, se utilizan las ecuaciones de Yule-Walker:

$$\gamma_k = \phi_1 \gamma_{k-1} + \cdots + \phi_1 \gamma_{k-p}, \qquad k = 1,2, \dots, p,$$
$$\sigma_a^2 = \gamma_0 - \phi_1 \gamma_1 - \cdots - \phi_p \gamma_p.$$

De forma matricial: $\Gamma_p \phi = \gamma_p$ y $\sigma_a^2 = \gamma_0 - \phi' \gamma_p$, donde $\Gamma_p = (\gamma_{i-j})_{i,j=1}^{p}$ es la matriz con dimensiones $p \times p$, $\phi = (\phi_1, \dots, \phi_p)'$ es el vector $p \times 1$ y, por último, $\gamma_p = (\gamma_1, \dots, \gamma_p)'$ es el vector $p \times 1$. Usando el método de los momentos, se reemplaza γ_k por $\hat{\gamma}_p$. Factorizando $\hat{\gamma}_0$, los estimadores se reescriben de la siguiente manera:

$$\begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{pmatrix} = \begin{pmatrix} 1 & \hat{\rho}_1 & \hat{\rho}_2 & \dots & \hat{\rho}_{p-2} & \hat{\rho}_{p-1} \\ \hat{\rho}_1 & 1 & \hat{\rho}_1 & \dots & \hat{\rho}_{p-3} & \hat{\rho}_{p-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \hat{\rho}_{p-3} & \dots & \hat{\rho}_1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \vdots \\ \hat{\rho}_p \end{pmatrix},$$

$$\hat{\sigma}_a^2 = \hat{\gamma}_0 \left(1 - \hat{\phi}_1 \hat{\rho}_1 - \hat{\phi}_2 \hat{\rho}_2 - \cdots - \hat{\phi}_p \hat{\rho}_p\right).$$

No obstante, este método da las estimaciones óptimas sólo para los modelos $AR(p)$ puros, con respecto a los procesos mixtos $ARMA(p,q)$ y de medias móviles $MA(q)$, dado que no son lineales en los parámetros, las estimaciones ya no son eficientes. En este caso, se utilizan otras técnicas, descritas a continuación.

2.6.2.2. MÉTODO DE MÁXIMA VEROSIMILITUD CONDICIONAL

Se supone una serie temporal Z_t que está generada por el modelo $ARMA(p,q)$ y se interesa estimar los parámetros $\phi_1, \dots, \phi_p$ $\theta_1, \dots, \theta_q$:

$$\dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q},$$

donde $\dot{Z}_t = Z_t - \mu$. Igual que en el algoritmo anterior, se asume $\mu_t = 0$ para $d > 0$. En el caso contrario, se estima como una media muestral, lo que sería válido para la mayoría de los casos. La distribución de densidad conjunta de a_t, asumiendo que es un ruido blanco Gaussiano, se define como:

$$f(a_1, a_2, \dots, a_n) \propto (\sigma_a^2)^{-\frac{n}{2}} \exp\left[-\left(\sum_{t=1}^{n} \frac{a_t^2}{2\sigma_a^2}\right)\right].$$

La innovación para un modelo $ARMA(p, q)$ se escribe en función de a_t como:

$$a_t = \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + \dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p}.$$

De esta forma, se puede definir la función de verosimilitud condicionada asociada con los parámetros $(\phi, \theta, \sigma_a^2)$. Para ello, sea $Z = (Z_1, Z_2, \dots, Z_n)'$ y se asumen los siguientes valores iniciales: $Z_* = (Z_{1-p}, \dots, Z_{-1}, , Z_0)'$ y $a_* = (a_{1-q}, \dots, a_{-1}, a_0)$. En este caso, el asterisco indica que estas funciones son condicionales sobre la elección de las variables de partida. Entonces, el logaritmo de la función de verosimilitud será:

$$ln\, L_*(\phi, \theta, \sigma_a^2) = \ell_*(\phi, \theta, \sigma_a^2) = -\frac{n}{2}\ln(\sigma_a^2) - \frac{S_*(\phi, \theta)}{2\sigma_a^2},$$

donde *la función condicional de la suma de cuadrados* es

$$S_*(\phi, \theta) = \sum_{t=1}^{n} a_t^2(\phi, \theta | Z_*, a_*, Z),$$

y el σ_a^2 se estima como

$$\widehat{\sigma_a^2} = \frac{S_*(\phi, \theta)}{n}.$$

El objetivo será encontrar los valores $\hat{\phi}, \hat{\theta}$ de tal forma que maximizan el logaritmo de la función de máxima verosimilitud anterior. Dichos valores se denominan los estimadores condicionales de máxima verosimilitud. Esto es equivalente a la minimización de la función condicional de la suma de cuadrados para cualquier valor fijo de σ_a^2 en el espacio $(\phi, \theta, \sigma_a^2)$. De tal manera, suponiendo la normalidad, se puede estudiar el comportamiento de la verosimilitud condicional a través del exámen de la función de la suma de cuadrados. Los

valores obtenidos a través de la minimización de $S_*(\phi, \theta)$, se denominan los estimadores de mínimos cuadrados condicionales.

Para los modelos mixtos $ARMA(p, q)$ y los modelos de medias móviles, dicha aproximación es válida si y sólo si todas las raíces del polinomio $(1 - \phi_1 B - \cdots - \phi_P B^p)$ están fuera del círculo unidad.

2.6.2.3. MÉTODO DE MÁXIMA VEROSIMILITUD INCONDICIONAL. ESTIMADORES DE MÍNIMOS CUADRADOS.

Se asume que existen $N = n + d$ observaciones generadas por el modelo $ARIMA$. Su logaritmo de la función de verosimilitud viene dado como:

$$l(\phi, \theta, \sigma_a^2) = f(\phi, \theta) - \frac{n}{2}\ln(\sigma_a^2) - \frac{S(\phi, \theta)}{2\sigma_a^2},$$

donde $f(\phi, \theta)$ implica el determinante en la densidad conjunta de los Z_t y es la función de ϕ y θ. *La función no condicional de la suma de los cuadrados* será:

$$S(\phi, \theta) = \sum_{t=1}^{n} [a_t|Z, \phi, \theta]^2 + [e_*]'\Omega^{-1}[e_*],$$

donde $[a_t|Z, \phi, \theta] = E[a_t|Z, \phi, \theta]$ indica la esperanza de a_t condicionado por Z, ϕ, θ; e_* denota el vector de las esperanzas condicionales de los valores iniciales, dado $Z, \phi, \theta : e_* = ([\dot{Z}_{1-p}], \dots, [\dot{Z}_0], [a_{1-q}], \dots, [a_0])'$. La función $S(\phi, \theta)$ puede representarse de forma alternativa:

$$S(\phi, \theta) = \sum_{t=-\infty}^{n} [a_t]^2.$$

Habitualmente, $f(\phi, \theta)$ es importante sólo para las series pequeñas. Para las series largas, el logaritmo de la función de verosimilitud incondicional está dominado por $S(\phi, \theta)/2\sigma_a^2$. De esta forma, los contornos de la función incondicional de la suma de los cuadrados en el espacio de los parámetros (ϕ, θ) son muy cercanos a los contornos de la verosimilitud y el logaritmo de la verosimilitud. En concreto, los estimadores de los parámetros obtenidos minimizando la suma de los cuadrados anterior (*estimadores de mínimos cuadrados incondicionales o exactos*), suelen proporcionar las aproximaciones muy cercanas a los estimadores de máxima verosimilitud.

2.6.3. DIAGNOSIS Y VALIDACIÓN

Antes de realizar el pronóstico de los valores futuros de la serie, se precisa comprobar que el modelo ajustado es válido, es decir, asegurarse que las predicciones serán lo más próximas posible a la realidad. El diagnóstico se centra, principalmente, en el estudio de los residuos del modelo, aunque hay otros aspectos a tener en cuenta.

2.6.3.1. SIGNIFICACIÓN DE LOS PARÁMETROS

Otra prueba de validación del modelo pretende comprobar que todos los parámetros del modelo son realmente significativos. Se considera una modelo $ARMA(p, q)$, cuyos coeficientes estimados son:

$$\beta = (c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q).$$

Para estudiar la significación de los coeficientes, se plantean los siguientes contrastes:

$H_0: c = 0$ frente a $H_1: c \neq 0$,

$H_0: \phi_i = 0$ frente a $H_1: \phi_i \neq 0$,

$H_0: \theta_i = 0$ frente a $H_1: \theta_i \neq 0$,

Se afirma que los estimadores siguen una distribución Normal:

$$\hat{\beta}_i \sim N\left(\beta_i, Var(\hat{\beta}_i)\right),$$

donde la varianza viene dada por la inversa de la matriz de información. Para el contraste, se utiliza el estadístico *t*:

$$t = \frac{\hat{\beta}_i - 0}{\sqrt{Var(\hat{\beta}_i)}} \sim N(0,1),$$

La hipótesis nula se rechaza siempre y cuando $|t| > z_{\alpha/2}$ (para el nivel de significación de 5%, es 1.96 aproximadamente). Sin embargo, a parte de comprobar la significación de los parámetros, es preciso verificar también que las raíces de los polinomios autorregresivos y de medias móviles están fuera del círculo de unidad. En el caso contrario, esto podría ser evidencia de falta de estacionariedad de los datos.

2.6.3.2. SOBREAJUSTE

Una de las primeras pruebas de diagnóstico que se puede llevar a cabo, se trata de comprobar que el modelo ajustado no precisa de más parámetros. El procedimiento empieza con la identificación y estimación del modelo de orden inferior y continúa ajustando los modelos más complejos. Si el modelo aumentado tiene sus coeficientes significativos, entonces, el modelo es preferible al anterior. De forma alternativa, se puede empezar con el modelo de orden superior. Si el modelo anterior es sobreajustado, entonces se debe continuar comprobando los modelos de órdenes inferiores, cuyas dimensiones se reducen y se verifican repetidamente contra el sobreajuste.

Para evitar el sobreajuste del modelo, antes de todo, se debe analizar cuidadosamente los gráficos de ACF, PACF y EACF para aumentar la precisión a la hora de identificar el modelo. En segundo lugar, siempre y cuando el modelo más simple se ajuste bien, es preferible a un modelo más complicado. En segundo lugar, es desaconsejable incrementar los órdenes de los componentes AR y MA al mismo tiempo. Por último, se puede utilizar los resultados del análisis residual para saber qué componentes es necesario aumentar.

2.6.3.3. ANÁLISIS RESIDUAL

Otro enfoque muy importante a la hora de diagnosticar el modelo, es el estudio de los residuos del modelo. Para que el modelo sea adecuado, los residuos deben seguir una distribución normal, ser independientes y aleatorios. Sea Z_t una serie temporal que sigue el modelo $ARMA(p, q)$:

$$\phi_p(B)\dot{Z}_t = \theta_q(B)a_t.$$

Los residuos estimados de este modelo se definen como:

$$\hat{a}_t = \hat{\theta}_q^{-1}(B)\hat{\phi}_p(B)\dot{Z}_t,$$

donde $\hat{\phi}_p$ y $\hat{\theta}_q$ son los parámetros estimados por máxima verosimilitud. Los residuos se calculan de forma recursiva a partir del modelo original:

$$\hat{a}_t = \dot{Z}_t - \sum_{j=1}^{p} \hat{\phi}_j \dot{Z}_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{a}_{t-j}\,, \qquad t = 1,2,\dots,n.$$

Usando cero valores iniciales (*método condicional*) o valores iniciales retrospectivos o *back-forecasting* (*método exacto*) para los $\hat{a}_t$ y $\dot{Z}_t$ iniciales. De forma más simple, el residuo se define como la diferencia entre el valor actual y el predicho.

2.6.3.3.1. ESTUDIO DE LA INDEPENDENCIA (AUTOCORRELACIÓN)

Otra consideración importante sobre los residuos es que deben estar incorrelados entre sí, es decir, ser independientes. La falta de la independencia de los residuos podría indicar que el modelo no está capturando completamente la estructura de autocorrelación presente en los datos. Igual al caso anterior, existen varias pruebas estadísticas y gráficas para verificar dicha hipótesis.

Se considera la función de autocorrelación de los residuos, $\hat{r}_k$ de los residuos estimados $\hat{a}_t$. En la práctica, se conoce sólo los a_t estimados, puesto que se conoce sólo los parámetros $\hat{\phi}$ y $\hat{\theta}$ estimados. Se sabe que para el verdadero ruido blanco y *n* grande, las autocorrelaciones están incorreladas y se distribuyen según una normal con media cero y varianza n^{-1}. No obstante, en los retardos pequeños la varianza de $\hat{r}_k$ puede ser inferior a n^{-1}, y los valores $\hat{r}_k(\hat{a})$ pueden estar altamente correlados, incluso en los modelos correctamente definidos. El procedimiento habitual consiste en detección de la cantidad de residuos que salen de los límites de control establecidos, normalmente, $\pm 1.96/\sqrt{n}$. Si los residuos se encuentran dentro de las bandas, no se puede rechazar la ausencia de la autocorrelación.

GRÁFICO DE AUTOCORRELACIONES DE LOS RESIDUOS

Una manera visual de estudiar la autocorrelación de los residuos es mediante el examen del gráfico de residuos. Tal y como se observa en la, todos los retardos, excepto uno, están dentro de las bandas, lo que indica la ausencia de la autocorrelación (bandas al 95% implica 1 de 20 puede estar fuera de las bandas por aleatoriedad).

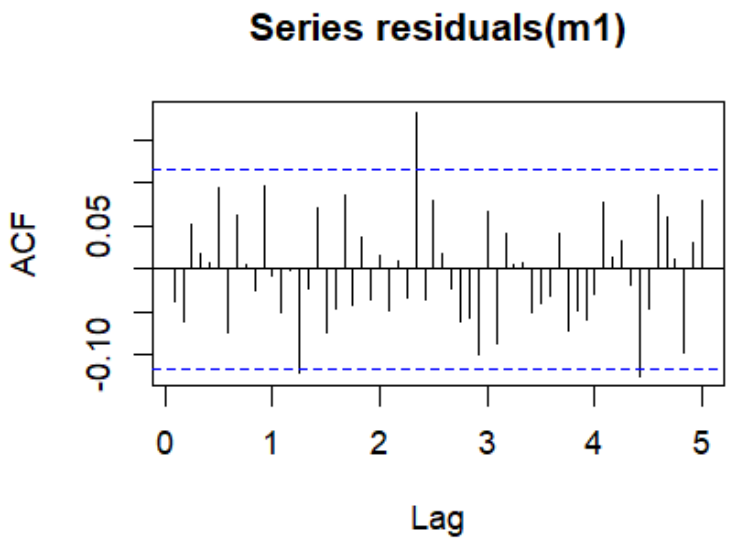

Figura 15: ACF de los residuos de los datos simulados. Fuente: Elaboración propia

TEST DE BOX-PIERCE

Uno de los tests que se utiliza ampliamente utilizados es el de Box-Pierce [2], que toma un grupo de autocorrelaciones de tamaño k y contrasta si es distinto de cero o no:

$$H_0: \rho_1(\hat{a}) = \cdots = \rho_k(\hat{a}) = 0 \Leftrightarrow Los\ residuos\ no\ son\ autocorrelados$$
$$\Leftrightarrow Los\ residuos\ son\ independientes$$
$$H_1: \hat{r}_i(\hat{a}) \neq 0\ para\ al\ menos\ uno\ i = 1,..,k \Leftrightarrow Los\ residuos\ son\ autocorrelados$$
$$\Leftrightarrow Los\ residuos\ no\ son\ independientes$$

El estadística de prueba para un modelo $ARIMA(p,d,q)$ se define como:

$$Q_{BP} = n\sum_{k=1}^{K} \hat{r}_k^2(\hat{a}),$$

que distribuye según una $\chi^2_{(K-p-q)}$ y la hipótesis nula se rechaza para los valores elevados del Q_{BP}. Aunque el k es variable, si es demasiado pequeño, esto lleva al tamaño de la prueba poco fiable. Por otra parte, el número de k muy elevado reduce la potencia de la prueba, porque se incluyen las correlaciones no relevantes. Como regla general, se aconseja tomar como mucho $n/4$ correlaciones. Sin embargo, la aproximación del χ^2 a la hipótesis nula del estadístico de la prueba es bueno sólo para la cantidad de datos muy grande. Por lo cual, Ljung y Box (1978) [32] propusieron la versión modificada del estadístico.

TEST DE LJUNG-BOX

La prueba de Ljung-Box es una versión mejorada del test de Box-Pierce, que es aplicable tanto para las muestras grandes, como para las pequeñas. Esta prueba está implementada en muchos softwares estadísticos y es más popular que el test anterior. Considerando la hipótesis nula definida anteriormente, el estadístico será:

$$Q_{LB} = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r_k^2\,(\hat{a}).$$

El estadístico se distribuye según $\chi^2_{(K-p-q)}$. En este caso, la hipótesis nula también se rechaza para los valores altos de Q_{LB}.

Aunque estas dos pruebas son más utilizadas al día de hoy, existen otras pruebas de autocorrelación de los residuos, como test de Durbin-Watson, test de McLeod, test de McLeod-Li, etc.

2.6.3.3.2. ESTUDIO DE LA NORMALIDAD

A la hora de calcular las estimaciones de los parámetros por máxima verosimilitud, se ha asumido que los a_t siguen una distribución gaussiana o normal. De esta manera, para verificar que las estimaciones de los parámetros son insesgadas y correctas y que el modelo captura correctamente la estructura de la serie temporal, existe una multitud de pruebas (tanto gráficas, como estadísticas) para comprobar el cumplimiento de dicha hipótesis:

$$\begin{cases} H_0: a_t\ siguen\ una\ Normal \\ H_1: a_t\ no\ siguen\ una\ Normal \end{cases}$$

A continuación se describen algunas de las pruebas más populares.

TEST DE JARQUE-BERA

La prueba de Jarque-Bera es uno de los tests más potentes para contrastar la normalidad, especialmente, en las muestras pequeñas. En este caso, la hipótesis nula de normalidad puede definirse de otra manera:

$$\begin{cases} H_0: S = 0\ y\ K = 0 \\ H_1: S \neq 0\ o\ K \neq 0 \end{cases}$$

El estadístico se calcula a partir de la asimetría (*S*) y la kurtosis (*K*):

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right],$$

donde *n* es el número de observaciones de la serie temporal y la asimetría y kurtosis se estiman como:

$$S = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3}{n},$$

$$K = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4}{n}.$$

Con x_i, $\bar{x}$ y s, en este caso, se refiere la observación i-ésima, la media y la desviación típica de la muestra (en este caso, se utilizan los residuos del modelo). El estadístico sigue una distribución χ^2 con 2 grados de libertad. Si la muestra tiene una asimetría y una curtosis cercanas a cero, se espera que el valor del estadístico de Jarque-Bera sea pequeño. Un valor grande del estadístico indica que la distribución difiere significativamente de una distribución normal.

TEST DE KOLMOGOROV-SMIRNOV

Es el test no paramétrico se utiliza principalmente para contrastar la normalidad en las muestras grandes (>50 observaciones). Se calcula la distancia máxima entre la función de distribución empírica de los datos y la función de distribución teórica de la distribución normal:

$$D = max|F^*(x) - S_n(x)|,$$

donde $S_n(x)$ es la distribución acumulada de la muestra y $F^*(x)$ es la función de distribución normal acumulada. Si esta distancia es lo suficientemente pequeña, no se rechaza la hipótesis nula, es decir, si el valor de *D* supera el valor crítico de la tabla $D(n, \alpha)$ obtenida de Haan (1977) [5], entonces se rechaza la hipótesis de la normalidad de los residuos.

TEST DE SHAPIRO-WILK

La prueba de Shapiro-Wilk es frecuentemente aplicada a las muestras de tamaño pequeño (< 50 observaciones), ya que es más potente en esta situación. Se basa en la covarianza entre los datos ordenados y los valores esperados de una muestra de una distribución normal, por lo que su cálculo es más costoso para las muestras grandes. El estadístico se calcula como

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

donde $x_{(i)}$ es el i-ésimo número inferior en la muestra, $\bar{x}$ es la media de la muestra, los coeficientes a_i se definen como $a_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C}$, C es la norma del vector $C = \|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{1/2}$, el vector m de los valores esperados de los estadísticos ordenados $m = (m_1, \ldots, m_n)^T$ y V es la matriz de covarianzas de los estadísticos ordenados. Sin embargo, el test no tiene la distribución de probabilidad asociada, y los p-valores se determinan por la simulación de Monte-Carlo.

QQ-PLOT

Un gráfico cuantil-cuantil (*Q-Q plot*) es una herramienta visual que se utiliza para comparar la distribución de una muestra de datos (en este caso, de los residuos) con una distribución teórica, como la distribución normal. En el eje abscisas del gráfico se encuentran los cuantiles de la distribución normal, y en el eje de ordenadas se encuentran los cuantiles observados de la muestra de datos. Cada punto en el gráfico representa la relación entre un cuantil de la distribución teórica y un cuantil de los datos observados. En un Q-Q plot ideal, los puntos seguirán una línea diagonal perfecta, lo que indicaría que los cuantiles observados y los cuantiles teóricos están en perfecta concordancia. Sin embargo, en la práctica, es común ver desviaciones de esta línea, especialmente en los extremos del gráfico.

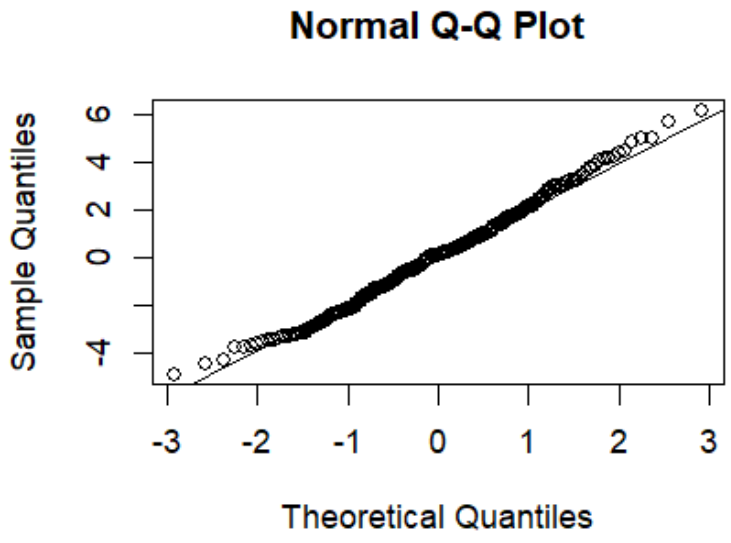

Figura 16: Ejemplo del Q-Q plot de los residuos del modelo SARIMA(0,1,1)*(1,1,1)[12] para los datos simulados. Fuente: Elaboración propia

A parte de las pruebas de diagnóstico de la normalidad mencionadas, existen otros métodos de evaluación: Test de Anderson-Darling, Test de Doornik-Hansen, Test de Chi-Cuadrado, etc.

2.6.3.3.3. ESTUDIO DE LA ALEATORIEDAD

El estudio de los residuos termina con la comprobación de su *aleatoriedad*. A diferencia de la independencia de los residuos, la aleatoriedad asegura que los residuos no siguen un patrón determinado, mientras que la independencia asegura que los residuos en diferentes momentos no están relacionados entre sí.

La forma más simple para investigar la presencia de algún patrón en los residuos, es la inspección visual del gráfico de la evolución de sus autocorrelaciones. Si se desea llevar a cabo una prueba estadística, el test más popular es *el test de rachas*. En esta ocasión, se contrasta la siguiente hipótesis:

$$H_0: Los\ residuos\ son\ aleatorios$$
$$H_1: Los\ residuos\ no\ son\ aleatorios$$

El test examina una secuencia de residuos, que se compone tanto por los valores positivos, como negativos. Si los residuos son aleatorios, entonces la racha tiene el número alternante o equilibrado de los valores positivos y negativos. El estadístico de prueba sería:

$$Z = \frac{R - \bar{R}}{s},$$

donde R es el número observado de rachas, $\bar{R}$ indica el número anticipado de rachas, s es la desviación típica de rachas. El s y el $\bar{R}$ se calculan como

$$\bar{R} = \frac{2n_1 n_2}{n_1 + n_2} + 1,$$
$$s^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)},$$

donde n_1 es la cantidad de signos negativos y n_2 es la cantidad de los negativos. El estadístico se contrasta frente a $Z_{1-\alpha}$ y se concluye que los residuos no son aleatorios si el estadístico es superior que el valor crítico.

2.6.3.4. PODER PREDICTIVO

Una vez descartado los modelos que no cumplen con las hipótesis anteriores, puede ser que habrá que elegir el mejor modelo entre varios propuestos. Existen diversos criterios de selección del modelo, aunque no se aconseja utilizar sólo un criterio (por ejemplo, el *AIC*), sino varios a la vez. Algunas medidas de bondad de ajuste comunes son:

- Error medio: $ME = \frac{1}{n}\sum_{t=1}^{n}\left(Z_t - \hat{Z}_{t-1}(1)\right) = \frac{1}{n}\sum_{t=1}^{n} e_{t-1}(1)$
- Error absoluto medio: $MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Z_t - \hat{Z}_{t-1}(1)\right| = \frac{1}{n}\sum_{t=1}^{n}|e_{t-1}(1)|$
- Suma de cuadrados de errores: $SSE = \sum_{t=1}^{n}\left(Z_t - \hat{Z}_{t-1}(1)\right)^2 = \sum_{t=1}^{n} e_{t-1}^2(1)$
- Error cuadrático medio: $MSE = \frac{SSE}{n-p-q}$.

En todos los casos, se desea encontrar los valores lo más próximos a cero, aunque no existe un umbral de decisión. A partir de los estadísticos anteriores, se puede calcular dos medidas de la varianza total:

$$R^2 = 1 - \frac{SSE}{SST},$$

donde $SST = \sum_{t=1}^{n}(Z_t - \bar{Z})^2$. No obstante, este criterio no tiene en cuenta el número de variables en el modelo, por lo que sería mejor utilizar el R^2 ajustado:

$$\bar{R}^2 = 1 - \left(\left(\frac{n-1}{n-m}\right)(1 - R^2)\right).$$

En ambos casos, se desea maximizar los criterios. Por último, como criterio de selección del modelo, se podría utilizar el *AIC* y *BIC*, introducidos y descritos en el apartado 2.6.1.

2.6.4. PREDICCIÓN

Uno de los principales objetivos del análisis de una serie de tiempo es predecir los valores futuros de manera que sean lo más cercanos posible a la realidad. En general, existen tres enfoques de la predicción:

1. *Predicciones subjetiva*. Se caracterizan por la base subjetiva: juicio, intuición, experiencia, etc.
2. *Predicciones univariantes*. Se basan en el ajuste del modelo unidimensional a los datos y la extrapolación del patrón de la serie.

3. *Predicciones multivariantes*. Se basan en la observación al mismo tiempo de dos o más series y sus modelos.

En la práctica, es normal utilizar múltiples enfoques a la vez, por ejemplo, ajustar la predicción univariante por una predicción subjetiva.

Sea Z_t una serie temporal observada, para $t = 0$ hasta $t = n$, donde n es la última observación de la se dispone. El concepto de la predicción consiste en calcular qué valores tomará la serie en momentos futuros $n + \ell$, donde ℓ es el número de períodos en el futuro considerado. Por lo cual, el valor de la predicción de $Z_{n+\ell}$ a partir de la información conocida hasta el momento *n* para ℓ momentos futuros suele denotarse como $Z_n(\ell)$. Como un criterio de optimalidad, suele utilizarse el error cuadrático medio, para el cual el valor esperado de los errores de la predicción en cuadrado, que viene dado por $E\left[(Z_{n+\ell} - \hat{Z}_n(\ell))^2\right] = E[e_n(\ell)^2]$, es minimizado. Se puede demostrar que, desde este punto de vista, la mejor predicción es la esperanza condicional de $Z_{n+\ell}$, dado las observaciones actuales y pasadas:

$$\widehat{Z_n}(\ell) = E(Z_n(\ell)|Z_1, Z_2, \dots, Z_n),$$

y el *error de predicción* se define como la diferencia entre el valor real y el predicho:

$$e_n(\ell) = Z_{n+\ell} - \hat{Z}_n(\ell).$$

Es obvio que a la hora de elaborar las predicciones se interesa obtener los errores mínimos posibles o nulos, lo que sugiere la predicción de *error cuadrático medio mínimo*.

ERROR CUADRÁTICO MEDIO MÍNIMO PARA ARIMA(P,D,Q)

Se considera el siguiente modelo $ARIMA(p, d, q)$ con $d \neq 0$:

$$\phi(B)(1 - B)^d = \theta(B)a_t,$$

donde $\phi(B) = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$ y $\theta(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$. El pronóstico óptimo del proceso, que minimiza el error cuadrático medio, se obtiene mediante la esperanza condicional $E(Z_{n+\ell}|Z_1, Z_2, \dots, Z_n)$. Dicho modelo $ARIMA$ puede reescribirse en el momento $n + \ell$ como una representación AR ($ARIMA$ es invertible) con el fin de derivar la varianza de la predicción para el modelo original:

$$\pi(B)Z_{t+\ell} = a_{t+\ell},$$

donde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j = \frac{\phi(B)(1-B)^d}{\theta(B)}$. De forma equvalente, la representación anterior se define como

$$Z_{t+\ell} = \sum_{j=1}^{\infty} \pi_j^{(\ell)} Z_{t-j+1} + \sum_{i=0}^{\ell-1} \psi_i a_{t+\ell-i},$$

donde $\pi_j^{(\ell)} = \sum_{i=0}^{\ell-1} \pi_{\ell-1+j-i}\psi_i$. Dado Z_t y considernado que $t \leq n$, las predicciones se calculan como

$$\hat{Z}_n(\ell) = E(Z_{n+\ell}|Z_t, t \leq n) = \sum_{j=1}^{\infty} \pi_j^{(\ell)} Z_{n-j+1},$$

y los errores de predicción correspondientes como

$$e_n(\ell) = \sum_{j=0}^{\ell-1} \psi_j a_{n+\ell-j},$$

donde $\psi_j = \sum_{i=0}^{j-1} \pi_{j-i}\psi_i\,, i = 1, \dots, \ell - 1$. Las predicciones son insesgadas y la varianza del error es

$$Var\big(e_n(\ell)\big) = \sigma_a^2 \sum_{j=0}^{\ell-1} \psi_j^2,$$

y los intervalos de confianza al $(1-\alpha)\%$ se calculan como

$$\hat{Z}_n(\ell) \pm z_{1-\alpha/2} \left[1 + \sum_{j=0}^{\ell-1} \psi_j^2\right]^{1/2} \sigma_a,$$

para los procesos normales, y $N_{\alpha/2}$ es el cuantil de la distribución normal tipificada.

El modelo general $ARIMA(p, d, q)$ se puede expresar análogamente como

$$\big(1 - \Psi_1 B - \cdots - \Psi_{p+d} B^{p+d}\big) Z_t = \big(1 - \theta_1 B - \cdots - \theta_q B^q\big) a_t,$$

donde $\Psi(B) = \phi(B)(1-B)^d = \big(1 - \Psi_1 B - \cdots - \Psi_{p+d} B^{p+d}\big)$. De esta forma, las predicciones se calculan como

$$\hat{Z}_n(\ell) = \Psi_1 \hat{Z}_n(\ell-1) + \cdots + \Psi_{p+d} \hat{Z}_n(\ell-p-d) + \hat{a}_n(\ell) - \theta_1 \hat{a}_n(\ell-1) - \cdots - \theta_q \hat{a}_n(\ell-q),$$

con

$$\hat{Z}_n(j) = E\left(Z_{n+j}\middle|Z_1, Z_2, \dots, Z_n\right), \qquad j \geq 1,$$
$$\hat{Z}_n(j) = Z_{n+j}, \qquad j \leq 0,$$
$$\hat{a}_n(j) = 0, \qquad j \geq 1,$$
$$\hat{a}_n(j) = Z_{n+j} - \hat{Z}_{n+j-1}(1) = a_{n+j}, \qquad j \leq 0.$$

3. MODELIZACIÓN DE LA SERIE DE TEMPERATURA

En esta parte se describe el proceso de elaboración del modelo $SARIMA(p,d,q) \times (P,D,Q)_{[s]}$ para los datos de la temperatura media mensual en la isla Jersey, Reino Unido.

3.1. DESCRIPCIÓN DE LOS DATOS

Para ilustrar la aplicación práctica de los conceptos teóricos definidos en la parte anterior, se ha elegido el conjunto datos climatológicos de la isla Jersey (oficialmente, *Bailía de Jersey*), la dependencia la Corona Británica en el Canal de la Mancha. De todas las islas del canal, es la más grande, y cuenta con 103 267 habitantes (2021) [25].

Figura 17: Localización de la isla Jersey. Retirado de https://www.britannica.com/place/Jersey-island-Channel-Islands-English-Channel el 25.04.2024

El clima es templado marítimo. Esto significa que los veranos son frescos y los inviernos suaves. La temperatura media en verano oscila entre los 15 °C y los 20 °C, mientras que en invierno está entre los 5 °C y los 10 °C. La isla recibe una cantidad significativa de precipitación durante todo el año, con los meses de invierno generalmente más húmedos que los meses de verano. Los vientos son moderados, y la isla rara vez experimenta temperaturas extremas.

El Gobierno de la isla publica en su página web numerosos datos estadísticos, clasificados en distintos temas: tiempo, población, salud, inflación, etc. El conjunto de datos climatológicos contiene los datos promedios mensuales de la *temperatura del aire* (en ºC), *temperatura del mar* (en ºC), *total de precipitaciones* (en milimetros), *total de las horas soleadas* y *la presión del aire promedia* (en milibars) entre el enero de 1981 y abril de 2022. Ya se sabe que los datos relacionados con el tiempo están relacionados directamente con la estación del año, por lo que para hacer la predicción para los meses posteriores se puede utilizar el modelo $SARIMA(p,d,q) \times (P,D,Q)_{[12]}$, con $s = 12$, puesto que los datos anteriores son mensuales.

Antes de todo, se realiza el breve análisis descriptivo de los datos con el fin de detectar los datos anómalos o ausentes:

```
> summary(weather_data)
 Daily air temp (mean)
 Min.   : 1.40
 1st Qu.: 8.50
 Median :11.85
 Mean   :12.22
 3rd Qu.:16.43
 Max.   :21.10

Daily sea temp (mean) Monthly rainfall Monthly sunshine Daily air pressure
(mean)
 Min.   : 7.30         Min.   :  0.20   Min.   : 43.4    Min.   :1001
 1st Qu.: 9.60         1st Qu.: 40.88   1st Qu.:106.2    1st Qu.:1015
 Median :12.80         Median : 65.55   Median :192.4    Median :1017
 Mean   :13.07         Mean   : 74.82   Mean   :185.9    Mean   :1017
 3rd Qu.:16.32         3rd Qu.: 99.25   3rd Qu.:259.8    3rd Qu.:1019
 Max.   :18.70         Max.   :279.80   Max.   :373.9    Max.   :1031
 NA's   :446                            NA's   :370      NA's   :414
```

La serie de la temperatura media del mar tiene casi 90% valores nulos (446 de 496), los datos de las horas del sol tienen el 75% de los valores ausentes (370 de 496), por último, la serie correspondiente a las mediciones de la presión tiene el 83% de los valors perdidos (414 de 496).

Las dos únicas series que no presentan valores nulos son de la temperatura del aire y la cantidad de precipitaciones, tampoco se observan valores anómalos.

Se visualizan las dos series:

```
> temperatura <- ts(data = weather_data$`Daily air temp (mean)`, start = c(1981, 01), frequency = 12)
> plot(temperatura)
```

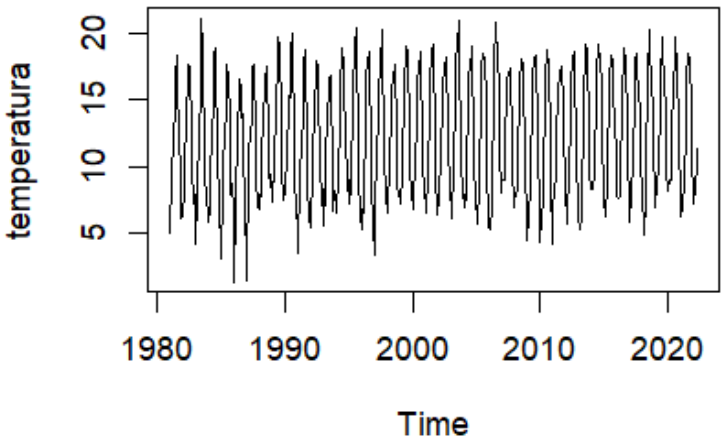

Figura 18: Gráfico de temperatura a lo largo de años. Fuente: Elaboración propia

```
> rainfall <- ts(data = weather_data$`Monthly rainfall`, start = c(1981, 01), frequency = 12)
> plot(rainfall)
```

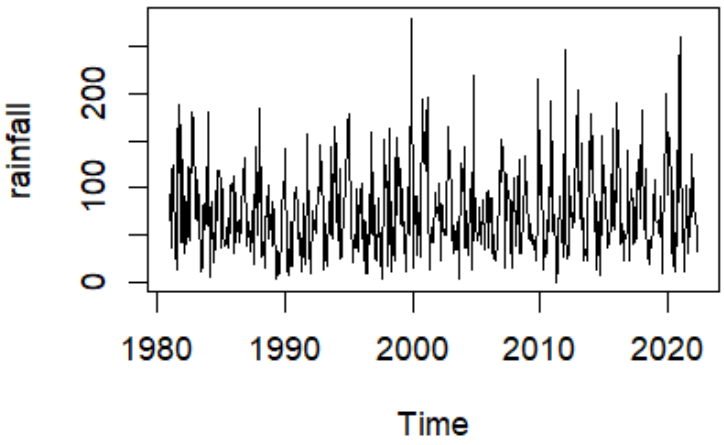

Figura 19: Gráfico de precipitaciones a lo largo de años. Fuente: Elaboración propia

Comparando los dos gráficos, se puede suponer que las dos series tienen la media constante, lo que se comprueba también con el test estadístico en el siguiente apartado. Sin embargo, la serie de precipitaciones presenta violaciones significativas de la estacionariedad de la varianza, lo que se puede comprobar, usando la función `BoxCox.lambda()` para la selección de la transformación lambda:

```
> BoxCox.lambda(rainfall)
[1] 0.2747132
```

La serie de la temperatura media, tras la evaluación visual, no presenta problemas graves con la varianza no constante, por lo que se opta por elegir dicha serie para la predicción.

3.2. IDENTIFICACIÓN

Los pasos necesarios para la correcta identificación del modelo $SARIMA(p,d,q) \times (P,D,Q)_{[s]}$ se describen en la Sección 2.6.1. En primer lugar, se visualiza de la serie (Figura 18) para hacer las conclusiones previas acerca de la misma. Con la función `decompose()` es posible descomponer la serie, utilizando las medias móviles, en tres componentes: parte estacional, de tendencia y aleatoria (Figura 20):

```
> fit <- decompose(temperatura, type='additive')
> autoplot(fit)+
+   labs(title = "Descomposición de la serie de tiempo",
+        x = "Tiempo",
+        y = "Temperatura",
+        colour = "Gears")+
```

```
+    theme_bw()
```

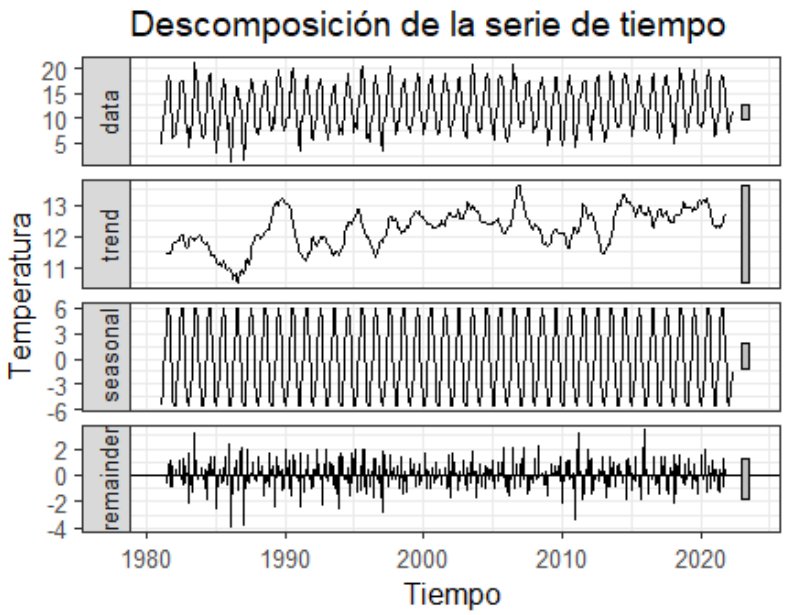

Figura 20: Descomposición de la serie de temperatura. Fuente: Elaboración propia.

En este caso, no se detectan ni los valores nulos o anómalos, ni algún tipo de cambio significativo en el nivel de la serie. Además, se observa el comportamiento estacional anual muy claro. Se nota que desde 1981 las temperaturas medias han surgido un aumento, lo que se refleja en el gráfico de la tendencia y que puede ser originado por el cambio climático global. El último gráfico de residuos proporciona la información sobre la parte de la serie que no se queda explicada por la tendencia o la estacionalidad. El gráfico sugiere que relativamente poca parte de la variación se queda sin explicar.

Tras llevar a cabo el análisis preliminar de la serie, es necesario verificar que la serie es estacionaria y, en el caso contrario, diferenciarla. Para ello, se utiliza el test de Dickey-Fuller aumentado, siendo una de las pruebas más comunes:

```
> adfTest(temperatura, lags = 1)

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    Dickey-Fuller: -3.7205
```

```
  P VALUE:
    0.01

Description:
 Wed May  1 15:08:35 2024 by user: natal
Warning message:
In adfTest(temperatura, lags = 1) : p-value smaller than printed p-value
```

Bajo la hipótesis nula de no estacionariedad de la serie, se puede rechazar la hipótesis nula, al usar sólo un retardo ($p - valor < 0.01$). Una de las ventajas del dicho test es la posibilidad de considerar el mayor número de retardos, por lo que se comprueba si la serie es estacionaria en el retardo 12, dado que los datos son mensuales:

```
> adfTest(temperatura, lags = 12)

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 12
  STATISTIC:
    Dickey-Fuller: 0.2063
  P VALUE:
    0.6821

Description:
 Wed May  1 15:13:50 2024 by user: natal
```

En este caso existe un evidencia fuerte de no rechazar la hipótesis de no estacionariedad ($p - valor = 0.6821$). Al repetir el test tras diferenciar los datos en el retardo 12, se rechaza la hipótesis de no estacionariedad, puesto que el p-valor es inferior a 0.01:

```
> adfTest(diff(temperatura, lag = 12), lags = 12)

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 12
  STATISTIC:
    Dickey-Fuller: -7.8003
  P VALUE:
    0.01
```

```
Description:
 Wed May  1 15:23:42 2024 by user: natal
Warning message:
In adfTest(diff(temperatura, lag = 12), lags = 12) :
  p-value smaller than printed p-value
```

Para identificar los órdenes del modelo, se utilizan los gráficos de autocorrelación simple (Figura 21) y parcial (Figura 22) de la serie diferenciada en el retardo 12:

```
> acf(diff(temperatura, lag = 12),lag.max=108,ci.type='ma')
```

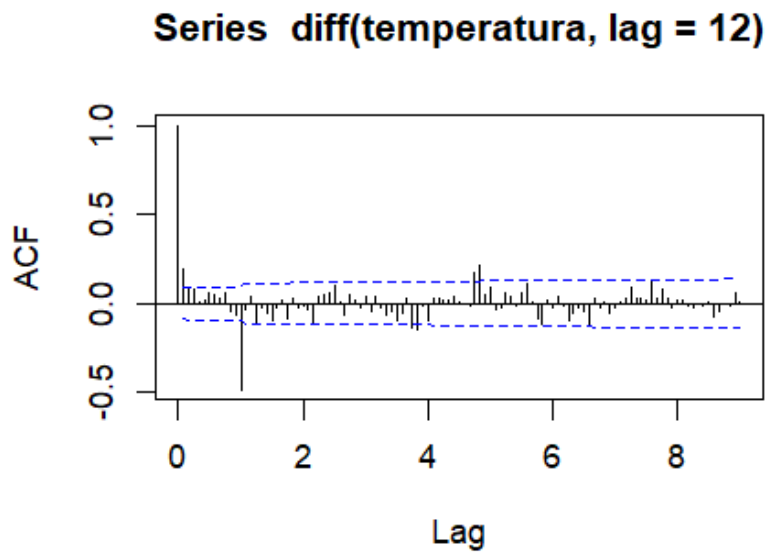

Figura 21: Gráfico de la función de autocorrelación simple de los datos diferenciado en el retardo 12. Fuente: Elaboración propia.

En el gráfico de ACF, los retardos significativos son 1 y 12, mientras que los retardos 1, 12 y 24 son significativos en el gráfico PACF, lo que sugiere el comportamiento estacional anual de la serie. Dado que en ninguno de los dos gráficos se observa el decrecimiento lento, se concluye otra vez que la serie es estacionaria.

```
> pacf(diff(temperatura, lag = 12),lag.max=108)
```

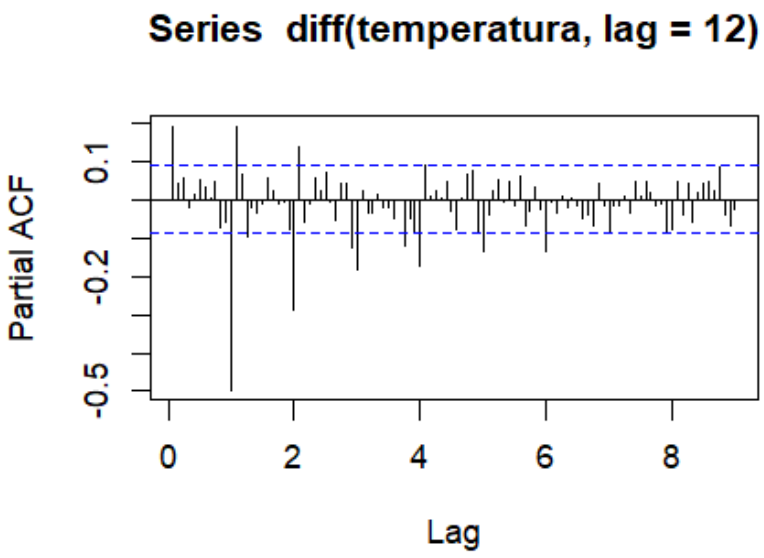

Figura 22: Gráfico de la función de autocorrelación parcial de los datos diferenciado en el retardo 12. Fuente: Elaboración propia.

Otra forma de estimar el número de diferenciaciones necesario para la parte regular y estacional, es utilizando las funciones `ndiffs()` y `nsdiffs()`, respectivamente:

```
> ndiffs(temperatura, test = c("adf"))
[1] 0

> nsdiffs(temperatura, m = 12)
[1] 1
```

Las salidas nos confirman que se tiene que diferenciar una vez la parte estacional, y la regular no. Sin embargo, para la identificación del modelo mixto, se aconseja utilizar la función de autocorrelación extendida de la serie original:

```
> eacf(temperatura, ar.max = 20, ma.max = 20)

AR/MA
   0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
0  x x o x x x x x o x x  x  x  x  o  x  x  x  x  x  o
1  x x o x x x x x o x x  x  x  x  o  x  x  x  x  x  o
2  x x o o o o o o o o o  o  o  o  x  o  o  o  o  o  x
3  x x x o o o o o o o o  o  o  o  x  o  o  o  o  o  x
4  x o x x o o x o o o o  o  o  o  x  o  x  o  o  o  x
5  x x o o o o x o o o o  o  o  o  x  o  o  o  o  o  x
6  x x o o x o o o o o o  o  o  x  o  o  o  o  o  o  o
7  x x o o x x o o o o o  o  o  x  o  o  o  o  o  x  o
8  x x x x x x x o o o o  o  o  o  o  o  o  o  o  o  o
9  x o x x x x o o o o o  o  o  o  x  o  o  o  o  o  o
10 x x x o x o o o o x o  o  o  o  o  o  o  o  o  o  o
```

```
11 x x x o x o x x x o o  o  o  o  o  o  o  o  o  o  o
12 x x x x x o x x x o x  x  o  o  o  o  o  o  o  o  o
13 o x x x o x o x x o x  x  x  o  o  o  o  o  o  o  o
14 o x x x o x o x o o x  x  x  x  o  o  o  o  o  o  o
15 x x x o o x x x o o x  x  x  x  o  o  o  o  o  o  o
16 x o x x o o x x o o x  x  o  x  o  x  o  o  o  o  o
17 x x x o o o x x o o x  x  o  x  o  x  o  o  o  o  o
18 x o x x o o x x o x x  x  x  o  o  o  x  o  o  o  o
19 x x o x o o x x o o x  x  x  o  o  o  x  o  x  o  o
20 o o o x x o x o o o x  o  o  o  o  o  o  o  o  o  o
```

Es importante recordar que se debe buscar un triángulo de "ceros", cuyo vértice superior izquierdo indicará los órdenes $AR(p)$ y $MA(q)$. De esta forma, se empieza con el modelo $SARIMA(0,0,0) \times (0,1,1)_{[12]}$ y se le van añadiendo los órdenes hasta llegar al modelo óptimo.

3.3. ESTIMACIÓN

El segundo paso de la modelización consiste en estimar los coeficientes del modelo propuesto y, si es necesario, modificar los órdenes para conseguir el modelo válido. Por esta razón, con la función `Arima()` se calculan los coeficientes del modelo $SARIMA$ propuesto anteriormente.

```
> modelo <- Arima(temperatura, order = c(0,0,0),
+             seasonal = list(order = c(0,1,1), period = 12), method='ML')
> summary(modelo)
Series: temperatura
ARIMA(0,0,0)(0,1,1)[12]

Coefficients:
         sma1
      -0.8855
s.e.   0.0243

sigma^2 = 1.556:  log likelihood = -802.42
AIC=1608.84   AICc=1608.87   BIC=1617.21

Training set error measures:
                    ME      RMSE       MAE         MPE      MAPE      MASE
  ACF1
Training set 0.1977749 1.230857 0.9552022 -0.5556195 10.59417 0.7288631 0.2
462806
```

Los residuos del modelo con sólo un coeficiente $SMA(1) = -0.8855$ no presentan el comportamiento del ruido blanco, puesto que algunos retardos salen de las bandas en los

gráficos de ACF (Figura 23) y PACF (Figura 24). La función de autocorrelación extendida de los residuos del modelo sugiere estimar el modelo $SARIMA(1,0,1) \times (0,1,1)_{[12]}$:

```
> eacf(residuals(modelo))
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o o o x o o o o o  o  o  o
1 x o o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x x o o o o o o o o o  o  o  o
4 x x x x o o o o o o o  o  o  o
5 x x o x o o o o o o o  o  o  o
6 x x o x x o o o o o o  o  o  o
7 x x x x x o o o o o o  o  o  o
> acf(residuals(modelo), lag.max = 60, ci.type='ma')
```

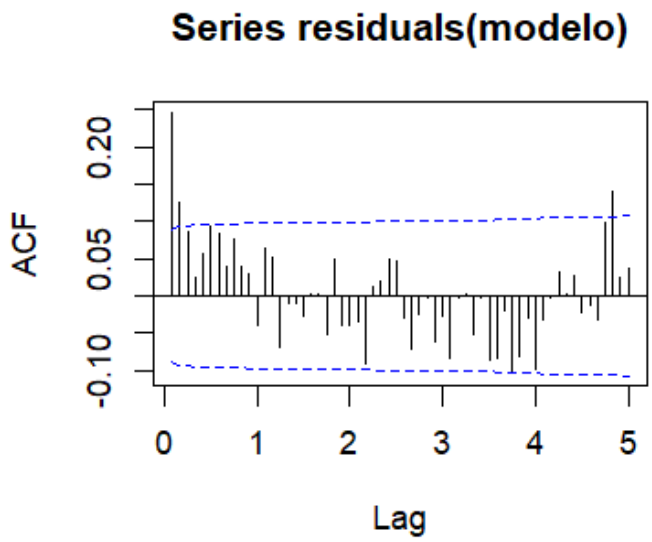

Figura 23: ACF de los residuos del modelo SARIMA(0,0,0)(0,1,1)[12]. Fuente: Elaboración propia.

```
> pacf(residuals(modelo), lag.max = 60)
```

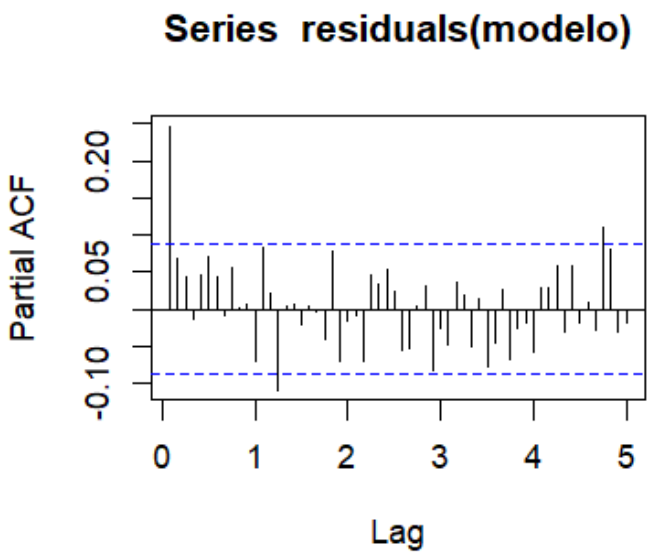

Figura 24: PACF de los residuos del modelo SARIMA(0,0,0)(0,1,1)[12]. Fuente: Elaboración propia.

Se estiman los parámetros del segundo modelo propuesto, $SARIMA(1,0,1) \times (0,1,1)_{[12]}$:

```
> modelo2 <- Arima(temperatura, order = c(1,0,1),
+                  seasonal = list(order = c(0,1,1), period = 12), method='M
L')
> summary(modelo2)
Series: temperatura
ARIMA(1,0,1)(0,1,1)[12]

Coefficients:
         ar1      ma1     sma1
      0.7762  -0.5723  -0.9347
s.e.  0.1906   0.2532   0.0333

sigma^2 = 1.411:  log likelihood = -780.94
AIC=1569.88   AICc=1569.97   BIC=1586.61

Training set error measures:
                    ME     RMSE       MAE         MPE      MAPE      MASE
 ACF1
Training set 0.139921 1.16962 0.9196444 -0.8845387 10.19225 0.7017309 0.023
27738
```

Se observa que los coeficientes del modelo tienen el error estándar bastante elevado, y en el caso del coeficiente $AR(1)$, su intervalo de confianza $\pm 1.96\sqrt{s.e.}$ contiene el valor 1:

```
> confint(modelo2)
```

```
           2.5 %      97.5 %
ar1    0.4026461  1.14980299
ma1   -1.0684882 -0.07607428
sma1 -1.0000507 -0.86933072
```

Se visualizan los gráficos de ACF (Figura 25) y PACF (Figura 26) de los residuos del modelo:

```
> acf(residuals(modelo2), lag.max = 60, ci.type='ma')
```

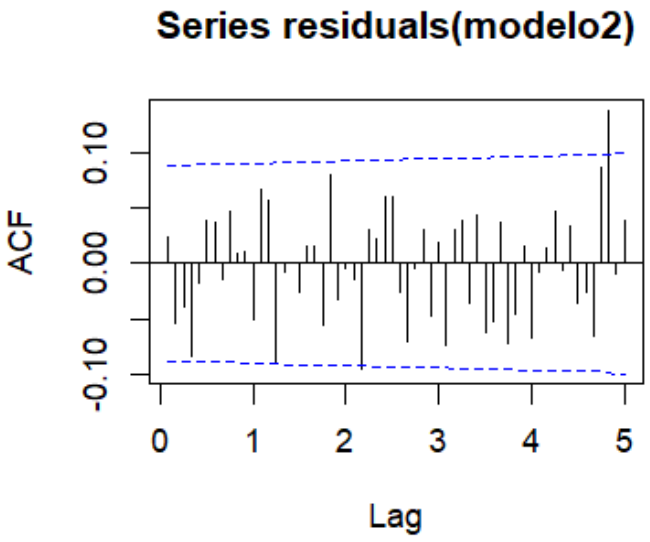

Figura 25: ACF de los residuos del modelo SARIMA(1,0,1)(0,1,1)[12]. Fuente: Elaboración propia.

```
> pacf(residuals(modelo2), lag.max = 60)
```

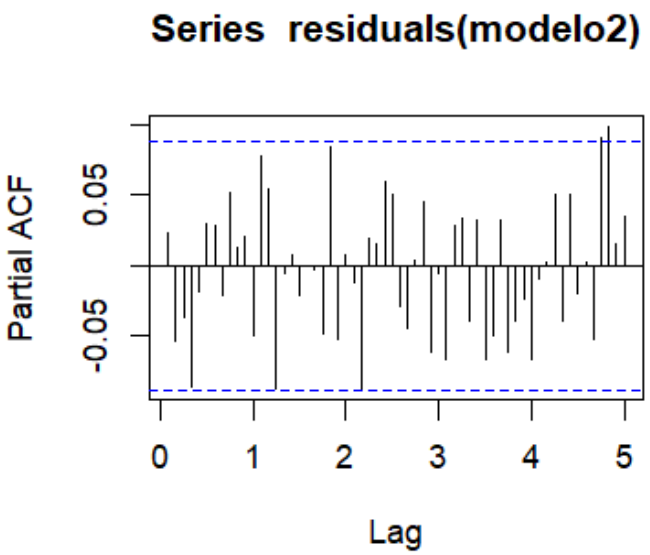

Figura 26: PACF de los residuos del modelo SARIMA(1,0,1)(0,1,1)[12]. Fuente: Elaboración propia.

En este caso, todas las correlaciones se encuentran dentro de las bandas (hay dos retardos al final que están fuera de las bandas, que se puede ignorar), lo que indica que este modelo podría ser óptimo. Sin embargo, teniendo en cuenta los errores de los coeficientes elevados, esto podría reducir la exactitud de las predicciones. Por lo consiguiente, puesto que el intervalo de confianza del coeficiente $AR(1)$ contiene el 1, se propone diferenciar la parte regular de la serie y reducir el número de términos AR en uno, así que el siguiente modelo a estimar sería $SARIMA(0,1,1) \times (0,1,1)_{[12]}$:

```
> modelo3 <- Arima(temperatura, order = c(0,1,1),
+                  seasonal = list(order = c(0,1,1), period = 12), method='
ML')
> summary(modelo3)
Series: temperatura
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
          ma1     sma1
      -0.8811  -1.0000
s.e.   0.0337   0.0401

sigma^2 = 1.391:  log likelihood = -787.34
AIC=1580.67   AICc=1580.72   BIC=1593.21

Training set error measures:
                      ME     RMSE       MAE        MPE     MAPE       MASE
  ACF1
Training set 0.003066256 1.161402 0.918884 -1.953721 10.17764 0.7011507 0.1
380455
```

En el nuevo modelo estimado el coeficiente $SMA(1)$ es igual a -1, lo que indica que el modelo está sobrediferenciado, así que se le elimina el órden D de la parte estacional y se le añade el coeficiente $SAR(1)$. El nuevo modelo propuesto será $SARIMA(0,1,1) \times (1,0,1)_{[12]}$:

```
> modelo4 <- Arima(temperatura, order = c(0,1,1),
+                  seasonal = list(order = c(1,0,1), period = 12), method='
ML')
> summary(modelo4)
Series: temperatura
ARIMA(0,1,1)(1,0,1)[12]

Coefficients:
          ma1  sar1     sma1
      -0.8866     1  -0.9868
s.e.   0.0350     0   0.0137
```

```
sigma^2 = 1.411:  log likelihood = -821.8
AIC=1651.6   AICc=1651.68   BIC=1668.41

Training set error measures:
                     ME      RMSE       MAE        MPE      MAPE       MASE
  ACF1
Training set 0.01330455 1.183037 0.9495345 -2.086338 10.50359 0.7245384 0.1
600526
```

Sin embargo, al intentar ajustar dicho modelo, el coeficiente $SAR(1) = 1$, lo que puede indicar la necesidad de diferenciar de nuevo la parte estacional del modelo y sumar un órden más a la parte SMA. Por lo cual, el quinto modelo hipotético será $SARIMA(0,1,1) \times (1,1,2)_{[12]}$:

```
> modelo5 <- Arima(temperatura, order = c(0,1,1),
+                  seasonal = list(order = c(1,1,2), period = 12), method='
ML')
> summary(modelo5)
Series: temperatura
ARIMA(0,1,1)(1,1,2)[12]

Coefficients:
          ma1    sar1     sma1    sma2
      -0.8803  0.0387  -1.0835  0.0835
s.e.   0.0329  1.0616   1.0620  1.0608

sigma^2 = 1.391:  log likelihood = -786.87
AIC=1583.74   AICc=1583.87   BIC=1604.64

Training set error measures:
                      ME     RMSE       MAE        MPE     MAPE      MASE
Training set 0.003883816 1.158903 0.9169513 -1.923677 10.14218 0.6996759
                   ACF1
Training set 0.1390003
```

Se visualiza el gráfico de la función de autocorrelación simple (Figura 27) y parcial (Figura 28) y se observa que en ambos gráficos hay varios retardos que están fuera de las bandas.

```
> acf(residuals(modelo5), lag.max = 60, ci.type='ma')
```

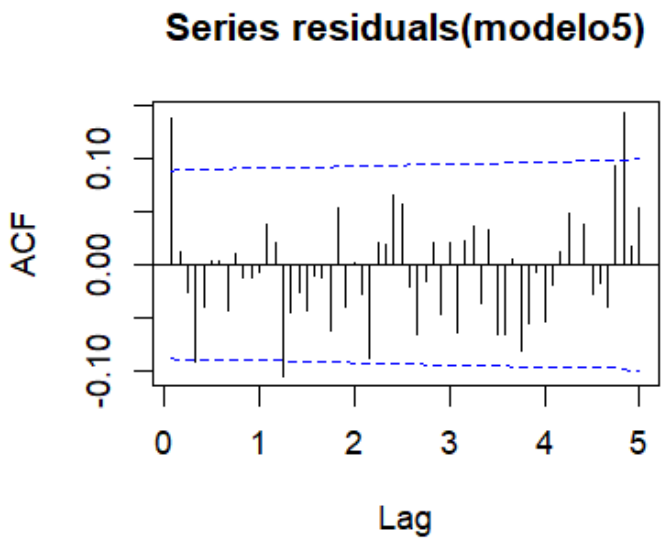

Figura 27: ACF de los residuos del modelo SARIMA(0,1,1)(1,1,2)[12]. Fuente: Elaboración propia.

```
> pacf(residuals(modelo5), lag.max = 60)
```

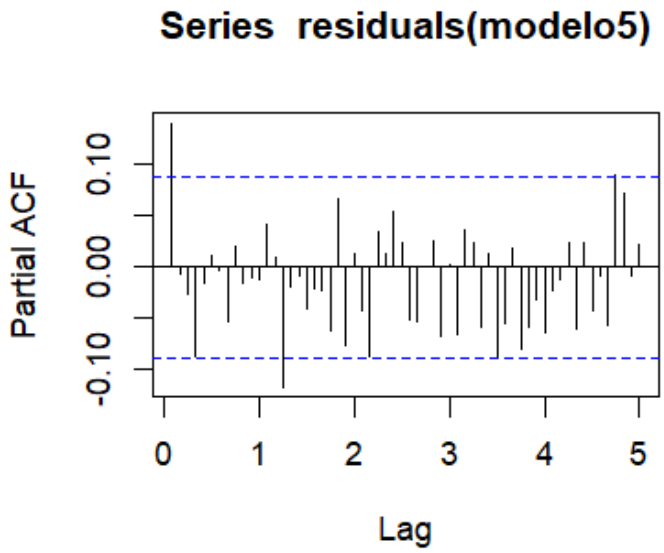

Figura 28: PACF de los residuos del modelo SARIMA(0,1,1)(1,1,2)[12]. Fuente: Elaboración propia.

En particular, se observa que en el gráfico de ACF el retardo 1 está fuera de las bandas, lo que sugiere añadir el coeficiente $MA(2)$ en el modelo actual. De esta manera, se estiman los coeficientes del modelo $SARIMA(0,1,2) \times (1,1,2)_{[12]}$:

```
> modelo6 <- Arima(temperatura, order = c(0,1,2),
+                  seasonal = list(order = c(1,1,2), period = 12), method='
ML')
```

```
> summary(modelo6)
Series: temperatura
ARIMA(0,1,2)(1,1,2)[12]

Coefficients:
          ma1      ma2     sar1     sma1    sma2
      -0.7609  -0.1692  -0.0018  -1.0399  0.0399
s.e.   0.0441   0.0483   0.9021   0.9049  0.9032

sigma^2 = 1.354:  log likelihood = -780.48
AIC=1572.97   AICc=1573.15   BIC=1598.05

Training set error measures:
                     ME      RMSE       MAE       MPE     MAPE      MASE
    ACF1
Training set 0.006536486 1.142411 0.8984064 -1.94175 10.01933 0.6855253 0.0
09281451
```

Sin embargo, al observar que los errores estándares de los coeficientes estacionales $SAR(1), SMA(1), SMA(2)$ son bastante elevados, se puede rechazar la necesidad de realizar la diferenciación estacional, dado que la sobrediferenciación del modelo origina la multicolinealidad. Se ajusta el séptimo modelo con sólo una diferenciación en la parte regular, $SARIMA(0,1,2) \times (2,0,2)_{[12]}$:

```
> modelo7 <- Arima(temperatura, order = c(0,1,2),
+                  seasonal = list(order = c(2,0,2), period = 12), method='
ML')


> summary(modelo7)
Series: temperatura
ARIMA(0,1,2)(2,0,2)[12]

Coefficients:
          ma1      ma2    sar1    sar2    sma1     sma2
      -0.7705  -0.2268  0.0149  0.9850  0.0220  -0.9563
s.e.   0.0413   0.0413  0.0742  0.0742  0.0757   0.0729

sigma^2 = 1.388:  log likelihood = -811.1
AIC=1636.19   AICc=1636.42   BIC=1665.63

Training set error measures:
```

```
                      ME      RMSE       MAE       MPE     MAPE      MASE
   ACF1
Training set 0.02171799 1.169762 0.9269761 -2.196172 10.38163 0.7073253 0.0
2535856
```

Se nota que los errores estándares de los coeficientes estacionales se han reducido drásticamente, pero aún así algunos retardos están fuera de las bandas en los gráficos de ACF (Figura 29) y PACF (Figura 30):

```
> acf(residuals(modelo7), lag.max = 60, ci.type='ma')
```

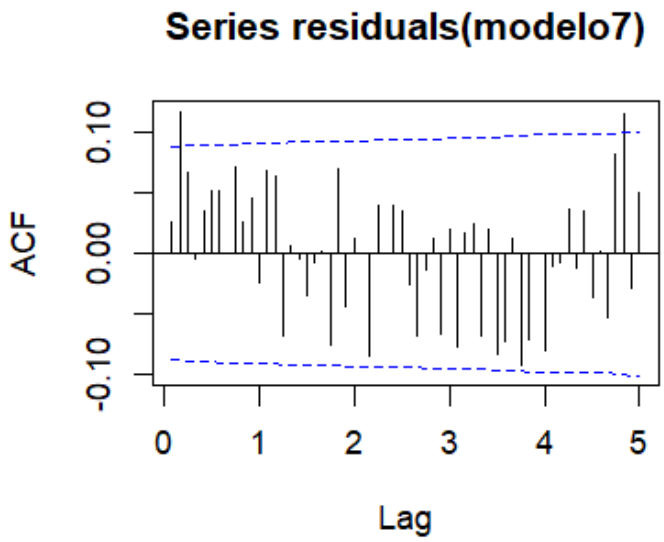

Figura 29: ACF de los residuos del modelo SARIMA(0,1,1)(2,0,2)[12]. Fuente: Elaboración propia.

```
> pacf(residuals(modelo7), lag.max = 60)
```

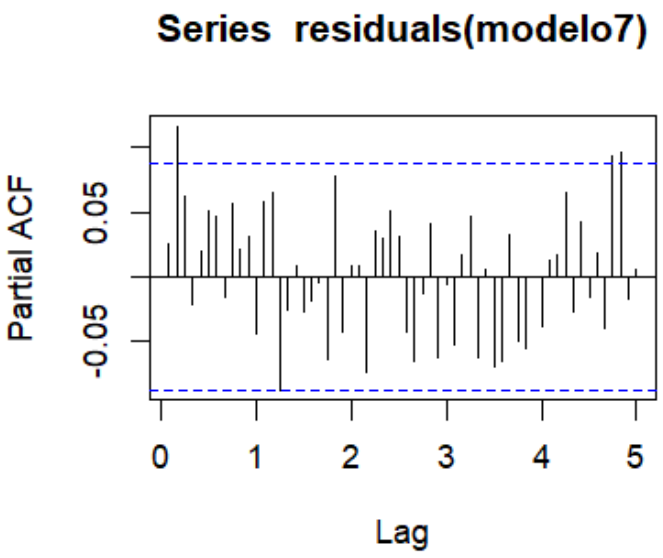

Figura 30: PACF de los residuos del modelo SARIMA(0,1,1)(2,0,2)[12]. Fuente: Elaboración propia.

Con el fin de intentar eliminar la autocorrelacción restante en los residuos, se le suman al modelo los coeficientes $SAR(3)$ y $SMA(3)$, dado que el intervalo de confianza de los coeficientes $SAR(2)$ y $SMA(2)$ incluye el 1 y el -1, respectivamente. Se propone ajustar el modelo $SARIMA(0,1,2) \times (3,0,3)_{[12]}$:

```
> modelo8 <- Arima(temperatura, order = c(0,1,2),
+                  seasonal = list(order = c(3,0,3), period = 12), method='
ML')
> summary(modelo8)
Series: temperatura
ARIMA(0,1,2)(3,0,3)[12]

Coefficients:
          ma1      ma2     sar1    sar2    sar3    sma1     sma2     sma3
      -0.7666  -0.2307  -0.8246  0.8741  0.9504  0.8840  -0.8582  -0.9637
s.e.   0.0415   0.0415   0.0904  0.0469  0.0649  0.1648   0.0654   0.1462

sigma^2 = 1.336:  log likelihood = -807.85
AIC=1633.7   AICc=1634.08   BIC=1671.55

Training set error measures:
                     ME      RMSE       MAE       MPE     MAPE      MASE
   ACF1
Training set 0.02039294 1.145529 0.9044657 -2.111906 10.07863 0.6901488 0.0
2203212
> acf(residuals(modelo8), lag.max = 60, ci.type='ma')
```

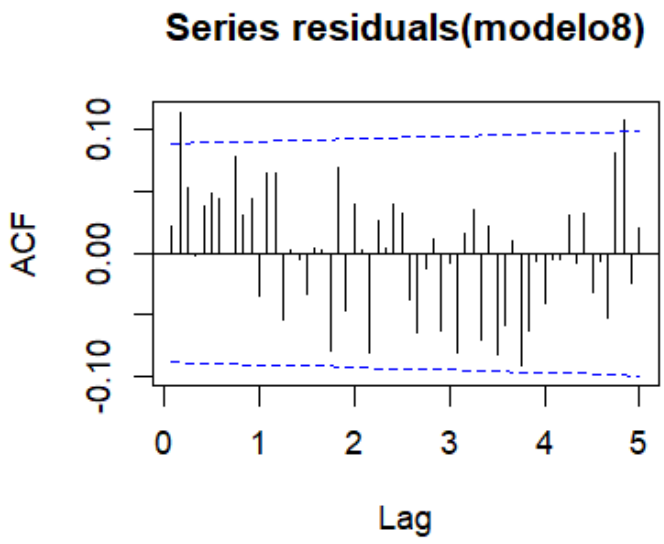

Figura 31: ACF de los residuos del modelo SARIMA(0,1,2)(3,0,3)[12]. Fuente: Elaboración propia

```
> pacf(residuals(modelo8), lag.max = 60)
```

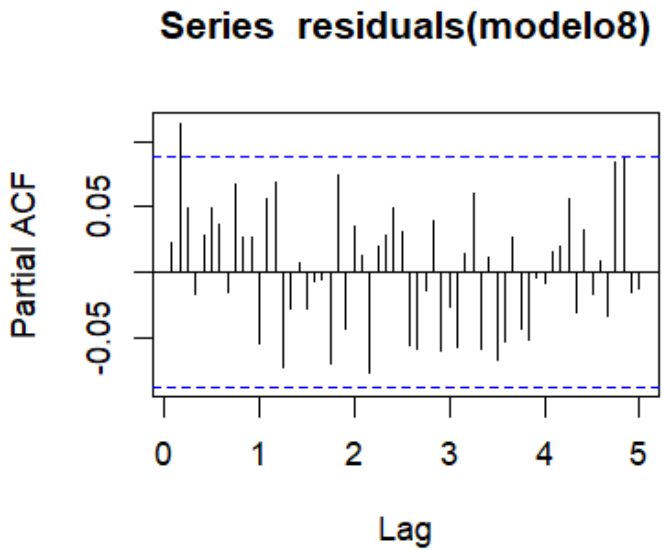

Figura 32: PACF de los residuos del modelo SARIMA(0,1,2)(3,0,3)[12]. Fuente: Elaboración propia

Tras visualizar los gráficos de la función de autocorrelación simple (Figura 31) y parcial (Figura 32), se puede llegar a la conclusión que es necesario sumar al modelo los coeficientes $MA(3)$ y $MA(4)$, dado que en el gráfico de ACF el retardo 2 está fuera de las bandas. Entonces, se estiman los coeficientes del modelo $SARIMA(0,1,4) \times (3,0,3)_{[12]}$:

```
> modelo9 <- Arima(temperatura, order = c(0,1,4),
+                  seasonal = list(order = c(3,0,3), period = 12), method='
ML')
```

```
> summary(modelo9)
Series: temperatura
ARIMA(0,1,4)(3,0,3)[12]

Coefficients:
          ma1      ma2      ma3      ma4     sar1    sar2    sar3
      -0.7406  -0.1315  -0.0357  -0.0896  -0.6343  0.6867  0.9475
s.e.   0.0551   0.0707   0.0589   0.0529   0.3077  0.1891  0.4833
        sma1     sma2     sma3
      0.6270  -0.6698  -0.8951
s.e.  0.3758   0.2400   0.6109

sigma^2 = 1.343:  log likelihood = -803.58
AIC=1629.16   AICc=1629.71   BIC=1675.41

Training set error measures:
                     ME     RMSE       MAE       MPE     MAPE
Training set 0.03280527 1.145779 0.9098314 -1.983164 10.09072
                   MASE         ACF1
Training set 0.6942431 -0.003161792
```

No obstante, el hecho de sumar dos coeficientes a la parte MA del modelo hace que los errores estándares de la parte estacional suben drásticamente y, dado que el coeficiente $MA(3) = -0.0357$ es casi igual a cero, se intenta estimar los coeficientes del mismo modelo, pero fijando $MA(3) = 0$ para reducir dichos errores:

```
> m1 <- Arima(temperatura, order = c(0,1,4),
+             seasonal = list(order = c(3,0,3), period = 12), method='ML',
+             fixed = c(NA, NA, 0, NA, NA, NA, NA, NA, NA, NA))
> summary(m1)
Series: temperatura
ARIMA(0,1,4)(3,0,3)[12]

Coefficients:
          ma1      ma2  ma3      ma4     sar1    sar2    sar3    sma1
      -0.7513  -0.1430    0  -0.1028  -0.8484  0.8975  0.9509  0.9157
s.e.   0.0445   0.0502    0   0.0365   0.0380  0.0563  0.0391  0.0868
         sma2     sma3
      -0.8826  -0.9870
s.e.   0.0641   0.0755

sigma^2 = 1.308:  log likelihood = -803.85
AIC=1627.7   AICc=1628.15   BIC=1669.74

Training set error measures:
                     ME     RMSE       MAE       MPE     MAPE      MASE
```

```
Training set 0.02916225 1.132191 0.8926147 -2.017942 9.964032 0.6811059
                    ACF1
Training set 0.00126695
```

Tal y como se observa, los errores estándares de todos los coeficientes, tanto de la parte regular, como de la estacional, vuelven a ser aceptables, así que el modelo final será modelo $SARIMA(0,1,4) \times (3,0,3)_{[12]}$. En los gráficos de la función de autocorrelación simple (Figura 33) y parcial (Figura 34) se observa que todos los residuos se encuentran dentro de las bandas:

```
> acf(residuals(m1), lag.max = 60)
```

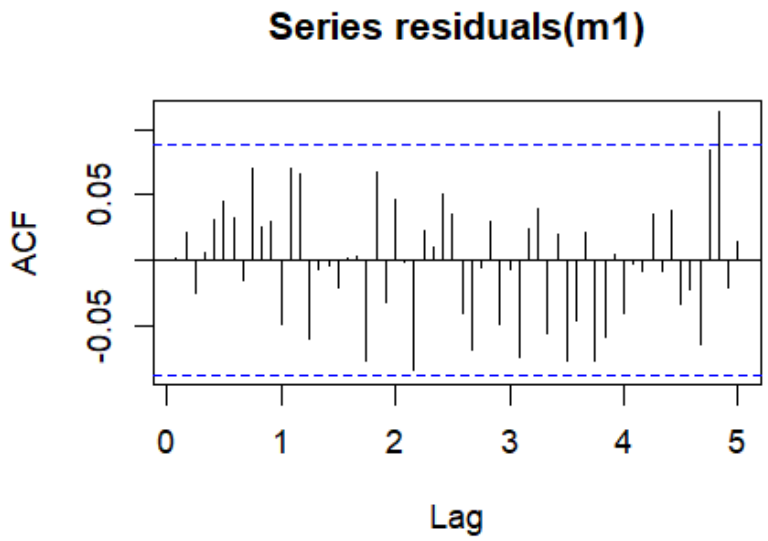

Figura 33: ACF de los residuos del modelo SARIMA(0,1,4)(3,0,3)[12]. Fuente: Elaboración propia.

```
> pacf(residuals(m1), lag.max = 60)
```

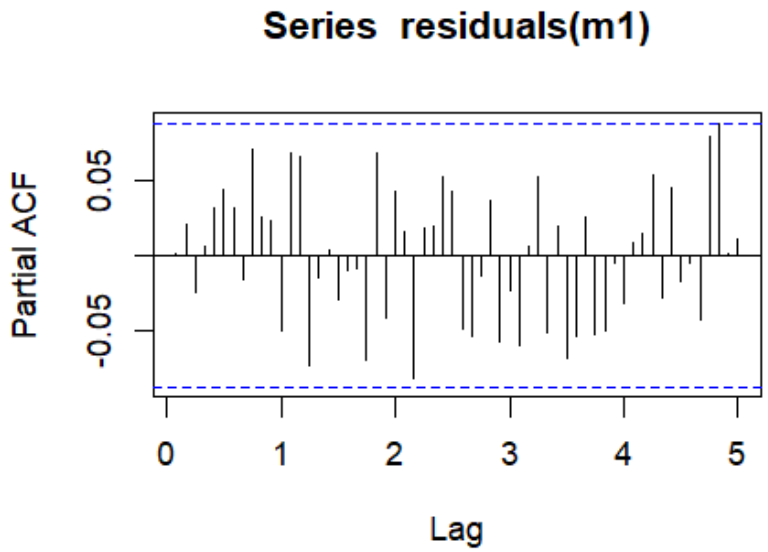

Figura 34: PACF de los residuos del modelo SARIMA(0,1,4)(3,0,3)[12]. Fuente: Elaboración propia.

3.4. VALIDACIÓN Y DIAGNOSIS

Tras ajustar el modelo y calcular sus coeficientes, es necesario verificar que los coeficientes del modelo ajustado son significativos y que los residuos cumplen con la hipótesis de normalidad, aleatoriedad e independencia.

3.4.1. SIGNIFICACIÓN DE LOS PARÁMETROS

Se comprueba que todos los coeficientes son significativos:

```
> coeftest(m1)

z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ma1  -0.751320   0.044532 -16.8714 < 2.2e-16 ***
ma2  -0.143031   0.050175  -2.8506  0.004363 **
ma4  -0.102805   0.036514  -2.8155  0.004870 **
sar1 -0.848415   0.037973 -22.3426 < 2.2e-16 ***
sar2  0.897519   0.056310  15.9388 < 2.2e-16 ***
sar3  0.950863   0.039071  24.3367 < 2.2e-16 ***
sma1  0.915714   0.086762  10.5543 < 2.2e-16 ***
sma2 -0.882564   0.064076 -13.7736 < 2.2e-16 ***
sma3 -0.986982   0.075485 -13.0753 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se verifica que todos los coeficientes son significativos tanto a nivel de confianza 0.05, como a 0.01.

3.4.2. NORMALIDAD DE LOS RESIDUOS

Para comprobar la normalidad de los residuos, se utiliza el test de Kolmogorov-Smirnov, ya que la muestra contiene más de 50 observaciones:

```
> ks.test(residuals(m1), "pnorm", mean = mean(residuals(m1)), sd = sd(resid
uals(m1)))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals(m1)
D = 0.027109, p-value = 0.8593
alternative hypothesis: two-sided
```

Bajo la hipótesis de normalidad, se concluye que existe evidencia fuerte ($p - valor = 0.8593$) de no rechazar la hipótesis de normalidad de los residuos. Tras examinar el QQ-plot de los residuos, se saca la misma conclusión, dado que los puntos están bastante cerca de la línea:

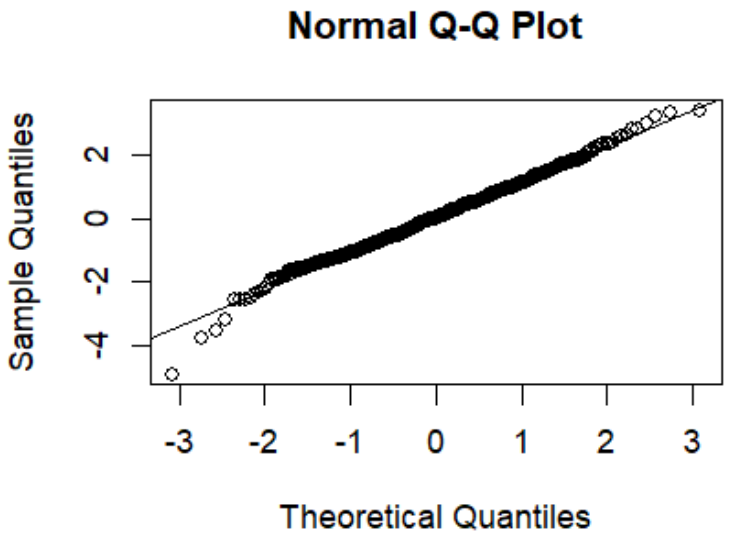

Figura 35: QQ-plot de los residuos del modelo ajustado. Fuente: Elaboración propia.

3.4.3. INDEPENDENCIA DE LOS RESIDUOS

Se lleva a cabo el test Ljung-Box para verificar que los residuos son independientes:

```
> Box.test(residuals(m1), type = "Ljung-Box", lag = 24, fitdf = 9)

        Box-Ljung test

data:  residuals(m1)
X-squared = 21.262, df = 15, p-value = 0.1287
```

Como se observa, no existe evidencia para rechazar la hipótesis de independencia, por lo que los residuos del modelo son independientes entre sí ($p - valor = 0.1287$).

3.4.4. ALEATORIEDAD DE LOS RESIDUOS

Por último, con el test de rachas se comprueba la aleatoriedad de los residuos

```
> runs.test(as.factor(m1$residuals > median(m1$residuals)))

        Runs Test
```

```
data:  as.factor(m1$residuals > median(m1$residuals))
Standard Normal = -0.089893, p-value = 0.9284
alternative hypothesis: two.sided
```

Tal y como se ve, existe una evidencia bastante fuerte para aceptar la hipótesis de aleatoriedad de los residuos ($p - valor = 0.9284$).

3.5. PREDICCIÓN

Tras verificar el cumplimiento de todas las hipótesis necesarias y comprobar que el modelo ajustado es el óptimo, se lleva a cabo el paso clave de todo el proceso del ajuste del modelo, la predicción de los valores futuros. En este caso, la predicción se realiza para dos años siguientes (desde mayo de 2022 hasta abril de 2024).

```
> (pred <- forecast(m1, h = 24))
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
May 2022      14.339151 12.845943 15.832358 12.055487 16.622815
Jun 2022      16.935709 15.395564 18.475854 14.580261 19.291157
Jul 2022      19.025695 17.476736 20.574655 16.656766 21.394624
Aug 2022      18.920977 17.363194 20.478759 16.538554 21.303399
Sep 2022      16.967686 15.409898 18.525475 14.585254 19.350118
Oct 2022      14.268632 12.710829 15.826435 11.886178 16.651086
Nov 2022      10.623858  9.066040 12.181675  8.241381 13.006334
Dec 2022       8.221880  6.664037  9.779723  5.839364 10.604396
Jan 2023       7.434967  5.877068  8.992866  5.052366  9.817567
Feb 2023       7.440378  5.882474  8.998283  5.057769  9.822988
Mar 2023       9.074741  7.516830 10.632651  6.692122 11.457359
Apr 2023      11.572420 10.014504 13.130336  9.189793 13.955047
May 2023      14.185110 12.623562 15.746659 11.796929 16.573292
Jun 2023      16.703162 15.141231 18.265093 14.314395 19.091929
Jul 2023      18.608584 17.046530 20.170638 16.219629 20.997539
Aug 2023      18.928522 17.366346 20.490699 16.539380 21.317665
Sep 2023      17.422614 15.860430 18.984797 15.033461 19.811766
Oct 2023      14.206016 12.643785 15.768246 11.816791 16.595241
Nov 2023      10.862953  9.300675 12.425231  8.473655 13.252251
Dec 2023       8.560593  6.998252 10.122935  6.171198 10.949988
Jan 2024       7.496035  5.934339  9.057731  5.107627  9.884443
Feb 2024       7.337072  5.775369  8.898774  4.948654  9.725490
Mar 2024       9.459948  7.898239 11.021657  7.071520 11.848375
Apr 2024      11.007851  9.446136 12.569566  8.619414 13.396289

> plot(pred)
```

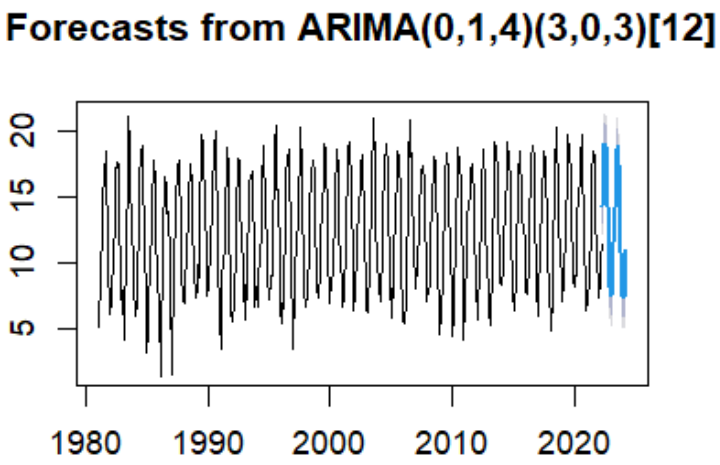

Figura 36: Predicción de la temperatura para dos años siguientes con el modelo SARIMA(0,1,4)(3,0,3)[12]. Fuente: Elaboración propia.

La función `pred()` devuelve la predicción puntual y las bandas superiores e inferiores para los intervalos de confianza de 80% y 95%. En la Tabla 3 y en la Figura 37 se puede apreciar la comparación entre los valores predichos por el modelo y los valores reales de temperatura en este mes [23].

Período (*mes-año*)	**Temperatura media observada (*en ºC*)**	**Temperatura media estimada (*en ºC*)**
05-2022	14.8	14.3
06-2022	17.4	16.9
07-2022	20.6	19.0
08-2022	20.9	18.9
09-2022	17.4	17.0
10-2022	15.7	14.3
11-2022	11.9	10.6
12-2022	6.8	8.2
01-2023	7.8	7.4
02-2023	7.9	7.4

03-2023	9.3	9.1
04-2023	10.8	11.6
05-2023	14	14.2
06-2023	18.3	16.7
07-2023	18	19.0
08-2023	18.1	18.9
09-2023	19.6	17.0
10-2023	15.6	14.3
11-2023	10.9	10.6
12-2023	9.8	8.2
01-2024	6.6	7.5
02-2024	9.6	7.4
03-2024	9.7	9.1
04-2024	11.3	11.0

Tabla 3: Comparación de la temperatura observada y predicha. Fuente: Elaboración propia.

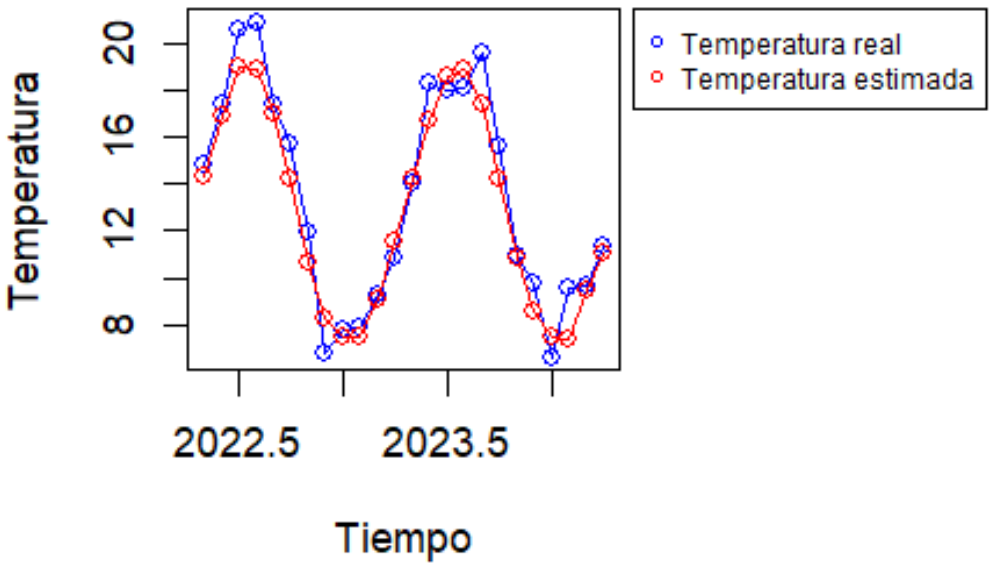

Figura 37: Comparación visual de la temperatura media observada y predicha. Fuente: Elaboración propia.

No obstante, en la Sección 3.2 se ha propuesto también el modelo $SARIMA(1,0,1) \times (0,1,1)_{[12]}$, como uno de los posibles, aunque el intervalo de confianza del coeficiente $AR(1)$ contiene 1. Se realiza la predicción con dicho modelo para el mismo período del tiempo:

```
> (pred <- forecast(modelo2, h = 24))
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
May 2022      14.010435 12.487936 15.532935 11.681973 16.338898
Jun 2022      16.692104 15.138264 18.245944 14.315711 19.068497
Jul 2022      18.616697 17.044276 20.189118 16.211886 21.021507
Aug 2022      18.509336 16.925825 20.092848 16.087565 20.931108
Sep 2022      16.927127 15.336971 18.517283 14.495193 19.359060
Oct 2022      13.894442 12.300296 15.488587 11.456406 16.332477
Nov 2022      10.383300  8.786756 11.979843  7.941597 12.825003
Dec 2022       8.075434  6.477448  9.673420  5.631526 10.519342
Jan 2023       7.061712  5.462904  8.660521  4.616545  9.506879
Feb 2023       7.130945  5.531613  8.730277  4.684977  9.576913
Mar 2023       8.846566  7.246918 10.446213  6.400116 11.293016
Apr 2023      11.003855  9.404017 12.603693  8.557114 13.450595
May 2023      13.688804 12.084526 15.293082 11.235273 16.142336
Jun 2023      16.442446 14.837780 18.047111 13.988321 18.896570
Jul 2023      18.422906 16.818007 20.027805 15.968424 20.877388
Aug 2023      18.358911 16.753871 19.963951 15.904215 20.813608
Sep 2023      16.810363 15.205239 18.415487 14.355538 19.265189
Oct 2023      13.803807 12.198632 15.408981 11.348904 16.258709
Nov 2023      10.312947  8.707743 11.918151  7.857999 12.767894
Dec 2023       8.020824  6.415604  9.626045  5.565852 10.475797
Jan 2024       7.019323  5.414138  8.624508  4.564405  9.474241
Feb 2024       7.098041  5.492850  8.703233  4.643112  9.552970
Mar 2024       8.821025  7.215829 10.426221  6.366090 11.275960
Apr 2024      10.984030  9.378832 12.589228  8.529091 13.438968
> plot(pred)
```

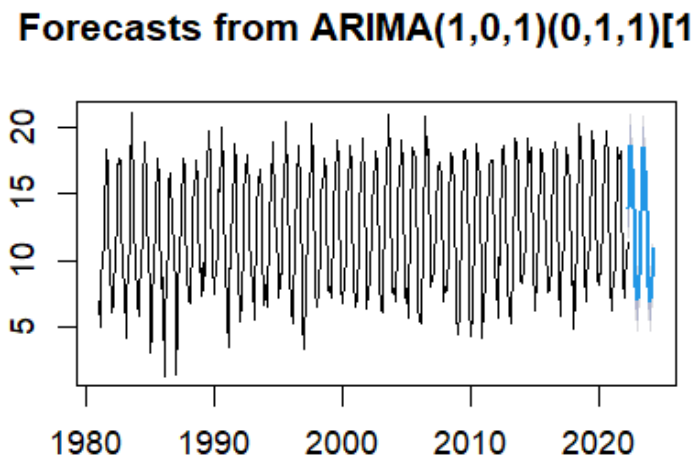

Figura 38: Predicción de la temperatura para dos años siguientes con el modelo SARIMA(1,0,1)(0,1,1)[12]. Fuente: Elaboración propia.

Se elabora la siguiente tabla para comparar ambas predicciones con las temperaturas observadas:

Período (*mes-año*)	**Temperatura media observada (*en ºC*)**	**Modelo** $SARIMA(0,1,4) \times (3,0,3)_{[12]}$	**Modelo** $SARIMA(1,0,1) \times (0,1,1)_{[12]}$
05-2022	14.8	14.3	14.01
06-2022	17.4	16.9	16.7
07-2022	20.6	19.0	18.6
08-2022	20.9	18.9	18.5
09-2022	17.4	17.0	16.9
10-2022	15.7	14.3	13.9
11-2022	11.9	10.6	10.4
12-2022	6.8	8.2	8.1
01-2023	7.8	7.4	7.1
02-2023	7.9	7.4	7.1
03-2023	9.3	9.1	8.9

04-2023	10.8	11.6	11.0
05-2023	14	14.2	13.7
06-2023	18.3	16.7	16.4
07-2023	18	19.0	18.4
08-2023	18.1	18.9	18.4
09-2023	19.6	17.0	16.8
10-2023	15.6	14.3	13.8
11-2023	10.9	10.6	10.3
12-2023	9.8	8.2	8.0
01-2024	6.6	7.5	7.0
02-2024	9.6	7.4	7.1
03-2024	9.7	9.1	8.8
04-2024	11.3	11.0	10.1

Tabla 4: Comparación de las predicciones realizadas por ambos modelos con los datos reales. Fuente: Elaboración propia.

Al final, se comparan los valores predichos por ambos modelos con los datos reales en la Figura 39:

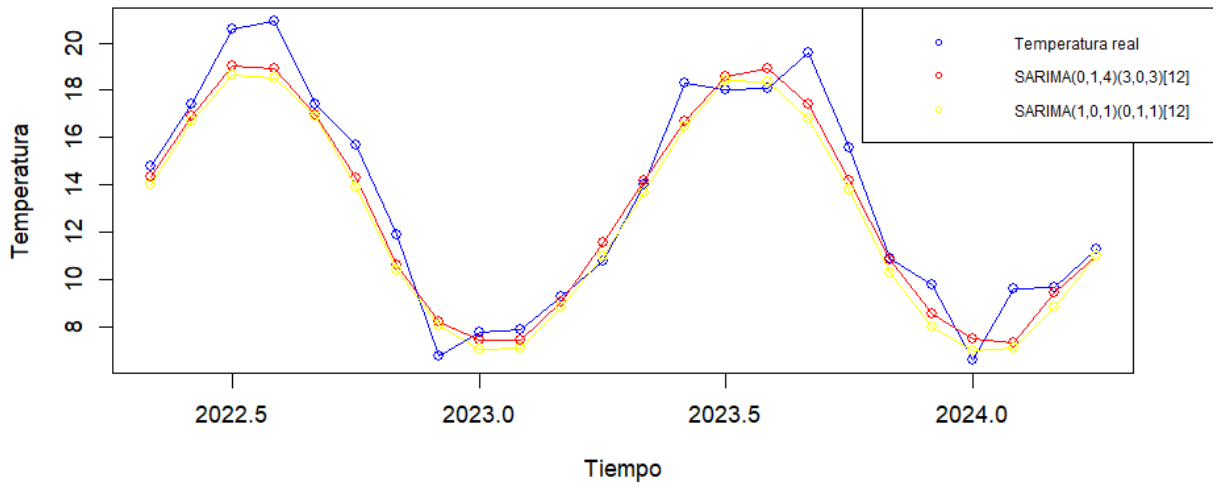

Figura 39: Comparación de las predicciones de los modelos ajustados con los datos reales. Fuente: Elaboración propia.

4. CONCLUSIONES

En el Capítulo 2 del presente trabajo se han estudiado y descrito los principios básicos de los datos temporales, los procesos estocásticos. En segundo lugar, se han descrito los modelos autorregresivos de medias móviles $ARIMA(p,d,q)$ y su caso particular, los modelos estacionales $SARIMA(p,d,q) \times (P,D,Q)_{[12]}$, propuestos por Box y Jenkins (1970). Se han detallado los métodos de *identificación* de los modelos, utilizando la metodología Box-Jenkins. También se han descrito distintos métodos de *estimación* de los coeficientes del modelo propuesto, junto con las pruebas de *diagnóstico* y *validación* del modelo resultante. Por último, se ha descrito el método de *predicción* de los valores futuros de la serie, basado en mínimo error cuadrático medio

En el Capítulo 3 se ha introducido el conjunto de datos elegido para la modelización, que son los datos de la temperatura media mensual en la isla Jersey. Se han detallado los pasos necesarios para la identificación del modelo $SARIMA(0,1,4) \times (3,0,3)_{[12]}$ con el software estadístico *RStudio*. El objetivo principal de la modelización ha sido realizar el pronóstico de la media mensual de la temperatura desde mayo de 2022 hasta el abril de 2024. Mirando las medidas de exactitud del modelo, se puede concluir que las predicciones se aproximan bastante a la realidad ($ME = 0.02916225, RMSE = 1.132191, MAE = 0.8926147$). La buena precisión de las predicciones se confirma también comparándolas con las mediciones de la temperatura media smensual, proporcionadas también por el Gobierno de la isla, y con las predicciones realizadas con el modelo subóptimo $SARIMA(1,0,1) \times (0,1,1)_{[12]}$. Los valores estimados con el modelo final están más cerca de las mediciones reales que las predicciones realizadas con el modelo subóptimo, aunque los dos tienden a subestimar la temperatura real en mayoría de los casos (Figura 39).

En conclusión, el presente trabajo demuestra perfectamente con qué dificultades puede enfrentarse el investigador a la hora de trabajar con el conjunto de datos reales y la importancia que tiene la base teórica para el correcto ajuste del modelo. Se ha demostrado también que no siempre se puede apoyarse en las pruebas preliminares básicas y, en algunos casos, es necesario emplear el análisis más profundo para encontrar el modelo adecuado, aunque no sea simple.

5. BIBLIOGRAFIA

LIBROS

1. Borodin, A. N. (2017). Stochastic processes. En Probability and its applications. https://doi.org/10.1007/978-3-319-62310-8
2. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series analysis: Forecasting and Control. John Wiley & Sons.
3. Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting. En Springer texts in statistics. https://doi.org/10.1007/978-3-319-29854-2
4. Cryer, J. D., & Chan, K. (2008a). Time Series analysis: With Applications in R. Springer Science & Business Media.
5. Haan, C. (1977). Statistical methods in hydrology. http://openlibrary.org/books/OL9354416M/Statistical_Methods_in_Hydrology
6. Hamilton, J. D. (1994). Time Series analysis. Taylor & Francis US.
7. Hitchcock, D. B. (s. f.). *STAT 520: Forecasting and Time Series* (University of South Carolina). https://people.stat.sc.edu/hitchcock/stat520ch1slides.pdf
8. Kirchgässner, G., Wolters, J., & Hassler, U. (2012). Introduction to Modern Time Series Analysis. Springer Science & Business Media.
9. Lütkepohl, H., & Krätzig, M. (2004). Applied Time Series Econometrics. Cambridge University Press.
10. Montes Suay, F. (2005). Procesos Estocásticos para Ingenieros: Teoría y Aplicaciones (Universidad de Valencia). https://www.uv.es/montes/SPE/manual.pdf
11. Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2011). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons.
12. Nielsen, H. B. (2004). Non-Stationary Time Series and Unit Root Tests. https://web.archive.org/web/20161130133316/http://www.econ.ku.dk/metrics/Econometrics2_05_II/Slides/08_unitroottests_2pp.pdf
13. Palit, A. K., & Popovic, D. (2006). Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications. Springer Science & Business Media.
14. Shumway, R. H., & Stoffer, D. S. (2000). Time Series Analysis and Its Applications. En Springer texts in statistics. https://doi.org/10.1007/978-1-4757-3261-0

15. Uribe Bravo, G. (s. f.). *Procesos estocásticos i: Capacitación técnica especializada en el nuevo marco de solvencia* (Universidad Nacional Autónoma de México). https://www.matem.unam.mx/~geronimo/162cnsf/notas162cnsf.pdf
16. Wei, W. W. S. (2006). Time Series analysis: Univariate and Multivariate Methods. Addison-Wesley Longman.

PÁGINAS WEB

17. *5.5 Selecting predictors*. (s. f.). Forecasting: Principles And Practice. https://otexts.com/fpp2/selecting-predictors.html
18. Fernandez-Casal, R. (2017, 23 octubre). *Diagnosis de la independencia | R Machinery*. https://rubenfcasal.github.io/post/diagnosis-de-la-independencia/#test-de-rachas
19. Kumar, R. (2023, 19 junio). *Top 10 time series analysis tools - DevOpsSchool.com*. DevOpsSchool.com. https://www.devopsschool.com/blog/top-10-time-series-analysis-tools/
20. Rodó, P. (2022, 24 noviembre). *Modelo autorregresivo (AR)*. Economipedia. https://economipedia.com/definiciones/modelo-autorregresivo-ar.html
21. *SPSS Modeler Subscription*. (s. f.). https://www.ibm.com/docs/es/spss-modeler/saas?topic=data-series-transformations
22. Wikipedia contributors. (2023, 26 mayo). *Jarque–Bera test*. Wikipedia. https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

CONJUNTOS DE DATOS

23. Goverment of Jersey. (s. f.). Jersey climate [Conjunto de datos]. En Jersey Met. https://www.gov.je/weather/jerseyclimate/
24. Goverment of Jersey. (2022). *Monthly temperature, rainfall, sunshine and air pressure figures - 1981 onwards* [Conjunto de datos]. https://opendata.gov.je/dataset/monthly-weather-data/resource/c98a3ffe-58aa-4225-a8ad-c19e01edbc5b?inner_span=True.
25. Goverment of Jersey. (2022b). *Population over time 1821 - 2021* [Conjunto de datos]. https://opendata.gov.je/dataset/population-over-time-1821-2011/resource/ea99f1f5-ef9a-474d-a382-20f089da73b0

PUBLICACIONES

26. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions On Automatic Control, 19(6), 716-723. https://doi.org/10.1109/tac.1974.1100705

27. Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. Journal Of The Royal Statistical Society. Series B. Methodological, 26(2), 211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x
28. Cardoso, F. C., Berri, R. A., Lucca, G., Borges, E. N., & De Mattos, V. L. D. (2023). Normality tests: a study of residuals obtained on time series tendency modeling. *Exacta*. https://doi.org/10.5585/2023.22928
29. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal Of The American Statistical Association*, *74*(366a), 427-431. https://doi.org/10.1080/01621459.1979.10482531
30. Gea, J., & Uriel, E. (1986). Identificación automática mediante la función de autocorrelación extendida. *Estadística Española*, *111*, 87-103. https://www.uv.es/=uriel/publicaciones/44%201986%20EE.pdf
31. Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3), 1–22. https://doi.org/10.18637/jss.v027.i0318
32. Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297-303. https://doi.org/10.1093/biomet/65.2.297
33. Mann, H. B., & Wald, A. (1943). On the Statistical Treatment of Linear Stochastic Difference Equations. Econometrica, 11(3/4), 173. https://doi.org/10.2307/1905674
34. Newbold, P. (1975). The Principles of the Box-Jenkins Approach. *Journal Of The Operational Research Society*, *26*(2), 397-412. https://doi.org/10.1057/jors.1975.88
35. Noor, T. H., Almars, A. M., Alwateer, M., Almaliki, M., Gad, I., & Atlam, E. (2022). SARIMA: A Seasonal Autoregressive Integrated Moving Average Model for Crime Analysis in Saudi Arabia. *Electronics*, *11*(23), 3986. https://doi.org/10.3390/electronics11233986
36. Schwarz, G. (1978). Estimating the Dimension of a Model. Annals Of Statistics, 6(2). https://doi.org/10.1214/aos/1176344136
37. Tiao, G. C., & Tsay, R. S. (1983). Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models. *Annals Of Statistics*, *11*(3). https://doi.org/10.1214/aos/1176346252

38. Tsay, R. S., & Tiao, G. C. (1984). Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models. Journal Of The American Statistical Association, 79(385), 84-96. https://doi.org/10.1080/01621459.1984.10477068
39. Uba, G., & Yusuf, M. (2022). Test of the Randomness of Residuals and Detection of Potential Outliers for the Modified Logistics Used in the Fitting of the Growth Curve of Immobilized Pseudomonas putida on Phenol. Journal Of Environmental Microbiology And Toxicology, 10(1), 14-18. https://doi.org/10.54987/jemat.v10i1.692
40. Yeo, I., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959. https://doi.org/10.1093/biomet/87.4.954

ANEXO I: CÓDIGO EN R

```
# Librerias utilizadas
library(readr)
library(lubridate)
library(fUnitRoots)
library(tseries)
library(forecast)
library(lmtest)
library(MASS)
library(TSA)
library(ggplot2)

# Se cargan los datos
weather_data <- read_csv("C:/username/weather-data.csv")

# El resumen de los datos
summary(weather_data)

rainfall <- ts(data = weather_data$`Monthly rainfall`, start =
 c(1981, 01), frequency = 12)
plot(rainfall)
BoxCox.lambda(rainfall)

# Se crea el objeto ts() con la fecha de partida de 1981-01
para la serie de tempratura
temperatura <- ts(data = weather_data$`Daily air temp (mean)`,
 start = c(1981, 01), frequency = 12)
plot(temperatura)

# Se visualiza la serie
fit <- decompose(temperatura, type='additive')
autoplot(fit)+
   labs(title = "Descomposición de la serie de tiempo",
        x = "Tiempo",
        y = "Temperatura",
        colour = "Gears")+
   theme_bw()

# Se realiza el test aumentado de Dickey-Fuller en el retardo
1 y 12
adfTest(temperatura, lags = 1)
adfTest(temperatura, lags = 12)
adfTest(diff(temperatura), lags = 12)
nsdiffs(temperatura, m = 12)
diff_temp = diff(temperatura, lag = 12)

acf(diff_temp, lag.max = 60, ci.type='ma')
pacf(diff_temp, lag.max = 60)
eacf(temperatura, ar.max = 20, ma.max = 20)
```

```
# Estimación
modelo <- Arima(temperatura, order = c(0,0,0),
            seasonal = list(order = c(0,1,1), period = 12), met
hod='ML')
summary(modelo)
coeftest(modelo)
acf(residuals(modelo), lag.max = 60, ci.type='ma')
pacf(residuals(modelo), lag.max = 60)
eacf(residuals(modelo))

modelo2 <- Arima(temperatura, order = c(1,0,1),

                 seasonal = list(order = c(0,1,1), period =
12), method='ML')

summary(modelo2)

coeftest(modelo2)

confint(modelo2)

acf(residuals(modelo2), lag.max = 60, ci.type='ma')

pacf(residuals(modelo2), lag.max = 60)


modelo3 <- Arima(temperatura, order = c(0,1,1),

                  seasonal = list(order = c(0,1,1), period =
12), method='ML')

summary(modelo3)


modelo4 <- Arima(temperatura, order = c(0,1,1),

                  seasonal = list(order = c(1,0,1), period =
12), method='ML')

summary(modelo4)


modelo5 <- Arima(temperatura, order = c(0,1,1),

                  seasonal = list(order = c(1,1,2), period =
12), method='ML')

summary(modelo5)

acf(residuals(modelo5), lag.max = 60, ci.type='ma')
```

```
pacf(residuals(modelo5), lag.max = 60)


modelo6 <- Arima(temperatura, order = c(0,1,2),
                 seasonal = list(order = c(1,1,2), period =
12), method='ML')
summary(modelo6)


modelo7 <- Arima(temperatura, order = c(0,1,2),
                 seasonal = list(order = c(2,0,2), period =
12), method='ML')
summary(modelo7)
acf(residuals(modelo7), lag.max = 60, ci.type='ma')
pacf(residuals(modelo7), lag.max = 60)


modelo8 <- Arima(temperatura, order = c(0,1,2),
                 seasonal = list(order = c(3,0,3), period =
12), method='ML')
summary(modelo8)
acf(residuals(modelo8), lag.max = 60, ci.type='ma') # sale lag
2
pacf(residuals(modelo8), lag.max = 60)


modelo9 <- Arima(temperatura, order = c(0,1,4),
                 seasonal = list(order = c(3,0,3), period =
12), method='ML')
summary(modelo9)


m1 <- Arima(temperatura, order = c(0,1,4),
            seasonal = list(order = c(3,0,3), period = 12),
method='ML',
            fixed = c(NA, NA, 0, NA, NA, NA, NA, NA, NA, NA))
summary(m1)
```

```
acf(residuals(m1), lag.max = 60)

pacf(residuals(m1), lag.max = 60)


# Validación y diagnosis
coeftest(m1)

Box.test(residuals(m1), type = "Ljung-Box", lag = 24, fitdf =
9)


ks.test(residuals(m1), "pnorm", mean = mean(residuals(m1)), sd
= sd(residuals(m1))) # pasa

qqnorm(residuals(m1))

qqline(m1$residuals)


runs.test(as.factor(m1$residuals > median(m1$residuals)))


# Predicción
temperatura_real <- ts(data = datos_reales$Temperature, start
= c(2022, 05), frequency = 12)


(pred1 <- forecast(m1, h = 24))

plot(pred1)


(pred <- forecast(modelo2, h = 24))

plot(pred)


par(mar=c(5.1, 4.1, 4.1, 8.1), xpd=TRUE)

plot(temperatura_real, type = "o", col = "blue", xlab =
"Tiempo", ylab = "Temperatura")

lines(pred1$mean, type = "o", col = "red")

lines(pred$mean, type = "o", col = "yellow")

legend("topright", legend = c("Temperatura real",
"SARIMA(0,1,4)(3,0,3)[12]", "SARIMA(1,0,1)(0,1,1)[12]"),
```

```
       col = c("blue", "red", "yellow"), inset=c(-0.85,0),
pch=c(1,1), cex = 0.7)
```

ANEXO II: DATOS UTILIZADOS

Month Start Date	Year	Month	Daily air temp (mean)	Daily sea temp (mean)	Monthly rainfall	Monthly sunshine	Daily air pressure (mean)
1981-01-01	1981	1	7		69.2		
1981-02-01	1981	2	5.1		65.6		
1981-03-01	1981	3	9.2		119.5		
1981-04-01	1981	4	9.5		36.8		
1981-05-01	1981	5	12.1		124.9		
1981-06-01	1981	6	14.7		25.7		
1981-07-01	1981	7	16.7		45.3		
1981-08-01	1981	8	18.4		13.9		
1981-09-01	1981	9	16.9		137.3		
1981-10-01	1981	10	11.8		187.8		
1981-11-01	1981	11	10.1		42.5		
1981-12-01	1981	12	6.1		165.9		
1982-01-01	1982	1	6.3		94.2		
1982-02-01	1982	2	6.7		42		
1982-03-01	1982	3	7.9		91.8		
1982-04-01	1982	4	10.1		30.5		
1982-05-01	1982	5	13.5		49.2		
1982-06-01	1982	6	17		121.8		

1982-07-01	1982	7	17.6		52.7		
1982-08-01	1982	8	17.2		45.1		
1982-09-01	1982	9	17.5		60.1		
1982-10-01	1982	10	12.5		179.7		
1982-11-01	1982	11	10.3		117		
1982-12-01	1982	12	7.2		174.1		
1983-01-01	1983	1	7.9		71		
1983-02-01	1983	2	4.2		67.5		
1983-03-01	1983	3	7.5		66		
1983-04-01	1983	4	8.9		108		
1983-05-01	1983	5	11.9		81.7		
1983-06-01	1983	6	16.1		11.3		
1983-07-01	1983	7	21.1		23		
1983-08-01	1983	8	19.1		14.9		
1983-09-01	1983	9	16.2		81.1		
1983-10-01	1983	10	12.9		82.8		
1983-11-01	1983	11	9.9		56.9		
1983-12-01	1983	12	7.4		65.6		
1984-01-01	1984	1	6.9		179.7		
1984-02-01	1984	2	5.9		60.7		
1984-03-01	1984	3	6.9		79.2		

1984-04-01	1984	4	10.5		5.2		
1984-05-01	1984	5	11.1		84.4		
1984-06-01	1984	6	16.1		20.9		
1984-07-01	1984	7	18.5		21.3		
1984-08-01	1984	8	18.9		44.1		
1984-09-01	1984	9	15.7		97.2		
1984-10-01	1984	10	13.1		65.2		
1984-11-01	1984	11	10.7		118.3		
1984-12-01	1984	12	7.7		118.4		
1985-01-01	1985	1	3.1		107.5		
1985-02-01	1985	2	4.7		36.9		
1985-03-01	1985	3	6.5		99.6		
1985-04-01	1985	4	10		54.3		
1985-05-01	1985	5	12.5		38.3		
1985-06-01	1985	6	14.3		41.2		
1985-07-01	1985	7	17.7		49.3		
1985-08-01	1985	8	16.8		67.8		
1985-09-01	1985	9	16.9		36.8		
1985-10-01	1985	10	13.5		66		
1985-11-01	1985	11	7.9		102.6		
1985-12-01	1985	12	8.8		96.8		

1986-01-01	1986	1	6.7		112.3		
1986-02-01	1986	2	1.4		30.2		
1986-03-01	1986	3	6.8		95.8		
1986-04-01	1986	4	7.5		42.6		
1986-05-01	1986	5	12.5		61		
1986-06-01	1986	6	15.9		66.4		
1986-07-01	1986	7	16.5		42.7		
1986-08-01	1986	8	15.9		60.3		
1986-09-01	1986	9	13.7		79.3		
1986-10-01	1986	10	13.9		104.1		
1986-11-01	1986	11	10.1		110.6		
1986-12-01	1986	12	8.3		132.5		
1987-01-01	1987	1	1.5		37.8		
1987-02-01	1987	2	5.2		60.8		
1987-03-01	1987	3	6.4		58.5		
1987-04-01	1987	4	11.6		63.6		
1987-05-01	1987	5	11.9		33.5		
1987-06-01	1987	6	14.3		75.6		
1987-07-01	1987	7	17.4		65.5		
1987-08-01	1987	8	17.7		18.3		
1987-09-01	1987	9	17.1		44		

1987-10-01	1987	10	13		144		
1987-11-01	1987	11	9.4		108.9		
1987-12-01	1987	12	7		50.1		
1988-01-01	1988	1	8		183.9		
1988-02-01	1988	2	6.9		85.5		
1988-03-01	1988	3	8.5		122.5		
1988-04-01	1988	4	10.6		26.6		
1988-05-01	1988	5	13.9		41.7		
1988-06-01	1988	6	15.8		14.8		
1988-07-01	1988	7	16.4		47.8		
1988-08-01	1988	8	17.5		91.4		
1988-09-01	1988	9	16		69.1		
1988-10-01	1988	10	13.6		103.1		
1988-11-01	1988	11	9.1		46		
1988-12-01	1988	12	9.4		70.8		
1989-01-01	1989	1	7.4		39.3		
1989-02-01	1989	2	7.9		84.8		
1989-03-01	1989	3	9.9		71.1		
1989-04-01	1989	4	8.9		75.1		
1989-05-01	1989	5	15.7		3.6		
1989-06-01	1989	6	16.7		32.5		

1989-07-01	1989	7	19.7		8.1		
1989-08-01	1989	8	19.1		8.6		
1989-09-01	1989	9	17.8		22.3		
1989-10-01	1989	10	15		38.9		
1989-11-01	1989	11	9.8		75.9		
1989-12-01	1989	12	7.5		101.2		
1990-01-01	1990	1	8.2		112.5		
1990-02-01	1990	2	9.5		140.7		
1990-03-01	1990	3	9.9		11.1		
1990-04-01	1990	4	10.2		63		
1990-05-01	1990	5	15.3		7.3		
1990-06-01	1990	6	15.2		64.2		
1990-07-01	1990	7	18.7		18		
1990-08-01	1990	8	20		17.5		
1990-09-01	1990	9	16.5		36		
1990-10-01	1990	10	14.5		91.4		
1990-11-01	1990	11	9.5		90.8		
1990-12-01	1990	12	6.6		100.7		
1991-01-01	1991	1	5.7		75.9		
1991-02-01	1991	2	3.5		46		
1991-03-01	1991	3	9.4		52.9		

1991-04-01	1991	4	9.5		43.5		
1991-05-01	1991	5	11.7		10.9		
1991-06-01	1991	6	13.8		82.5		
1991-07-01	1991	7	17.6		34.2		
1991-08-01	1991	8	18.7		18.2		
1991-09-01	1991	9	17.4		49.2		
1991-10-01	1991	10	12.4		90.8		
1991-11-01	1991	11	8.9		156.2		
1991-12-01	1991	12	6.2		37.6		
1992-01-01	1992	1	5.5		9.2		
1992-02-01	1992	2	6.9		35.2		
1992-03-01	1992	3	8.5		50.8		
1992-04-01	1992	4	10		76.3		
1992-05-01	1992	5	14.9		60.7		
1992-06-01	1992	6	16.5		52.7		
1992-07-01	1992	7	17.9		63.3		
1992-08-01	1992	8	17.7		100.1		
1992-09-01	1992	9	15.3		88.9		
1992-10-01	1992	10	11		101.5		
1992-11-01	1992	11	10.6		145.9		
1992-12-01	1992	12	7.1		114		

1993-01-01	1993	1	8.3		68.9		
1993-02-01	1993	2	5.6		13.2		
1993-03-01	1993	3	8.4		31.1		
1993-04-01	1993	4	11		61.3		
1993-05-01	1993	5	13.6		18		
1993-06-01	1993	6	16.1		85.4		
1993-07-01	1993	7	16.5		59.9		
1993-08-01	1993	8	16.9		51		
1993-09-01	1993	9	14.8		144.3		
1993-10-01	1993	10	11.3		87.3		
1993-11-01	1993	11	6.7		46.8		
1993-12-01	1993	12	8.2		164.3		
1994-01-01	1994	1	7.7		133.6		
1994-02-01	1994	2	6.6		83.7		
1994-03-01	1994	3	9.4		43.5		
1994-04-01	1994	4	9.7		64.9		
1994-05-01	1994	5	13		120.3		
1994-06-01	1994	6	15.8		25		
1994-07-01	1994	7	18.9		27.8		
1994-08-01	1994	8	17.8		88		
1994-09-01	1994	9	15.2		79.8		

1994-10-01	1994	10	13.1		99.4		
1994-11-01	1994	11	12		90.9		
1994-12-01	1994	12	9.2		168		
1995-01-01	1995	1	7.2		178.4		
1995-02-01	1995	2	9		139		
1995-03-01	1995	3	8		77.6		
1995-04-01	1995	4	9.9		46.7		
1995-05-01	1995	5	14.1		44.3		
1995-06-01	1995	6	15.4		20.4		
1995-07-01	1995	7	19.2		49.4		
1995-08-01	1995	8	20.4		36.4		
1995-09-01	1995	9	15.6		98.8		
1995-10-01	1995	10	15.5		33.4		
1995-11-01	1995	11	9.9		63		
1995-12-01	1995	12	5.9		99.2		
1996-01-01	1996	1	6.9		46.8		
1996-02-01	1996	2	5.4		105.3		
1996-03-01	1996	3	7.5		30		
1996-04-01	1996	4	10.3		22.5		
1996-05-01	1996	5	11		64		
1996-06-01	1996	6	16.7		10.3		

1996-07-01	1996	7	17.9		8.9		
1996-08-01	1996	8	18.6		93.3		
1996-09-01	1996	9	15.5		40.8		
1996-10-01	1996	10	13.7		60.7		
1996-11-01	1996	11	9		158.8		
1996-12-01	1996	12	5.5		75.7		
1997-01-01	1997	1	3.4		24		
1997-02-01	1997	2	8.1		81.7		
1997-03-01	1997	3	9.8		22.1		
1997-04-01	1997	4	10.9		30.4		
1997-05-01	1997	5	13.8		61.3		
1997-06-01	1997	6	15.6		88.5		
1997-07-01	1997	7	17.5		16.9		
1997-08-01	1997	8	20.3		58.8		
1997-09-01	1997	9	17.2		4.3		
1997-10-01	1997	10	13.5		113.7		
1997-11-01	1997	11	10.9		151.2		
1997-12-01	1997	12	7.9		100.6		
1998-01-01	1998	1	6.6		105.7		
1998-02-01	1998	2	7.6		17.7		
1998-03-01	1998	3	9.3		59.1		

1998-04-01	1998	4	9.5		162.4		
1998-05-01	1998	5	14.9		11.9		
1998-06-01	1998	6	15.8		84.3		
1998-07-01	1998	7	16.6		35.5		
1998-08-01	1998	8	17.7		22.4		
1998-09-01	1998	9	16.8		113.2		
1998-10-01	1998	10	13.5		152.4		
1998-11-01	1998	11	8.8		70.6		
1998-12-01	1998	12	7.8		131.9		
1999-01-01	1999	1	8.1		111.8		
1999-02-01	1999	2	7.3		63.9		
1999-03-01	1999	3	9.4		59.5		
1999-04-01	1999	4	11		98.8		
1999-05-01	1999	5	14.8		30.8		
1999-06-01	1999	6	15.8		51.8		
1999-07-01	1999	7	19		10.7		
1999-08-01	1999	8	18.8		89.3		
1999-09-01	1999	9	18		100.6		
1999-10-01	1999	10	13.2		71		
1999-11-01	1999	11	10		49.9		
1999-12-01	1999	12	8.1		279.8		

2000-01-01	2000	1	6.9		14.7		
2000-02-01	2000	2	8.6		75.6		
2000-03-01	2000	3	9.2		61		
2000-04-01	2000	4	9.7		91		
2000-05-01	2000	5	13.9		90.2		
2000-06-01	2000	6	16.5		28.7		
2000-07-01	2000	7	17.2		73.3		
2000-08-01	2000	8	18.6		26.7		
2000-09-01	2000	9	17.2		67.1		
2000-10-01	2000	10	12.9		185.7		
2000-11-01	2000	11	9.7		193.6		
2000-12-01	2000	12	8.7		125.6		
2001-01-01	2001	1	6.6		181.9		
2001-02-01	2001	2	7.3		119		
2001-03-01	2001	3	8.8		196.2		
2001-04-01	2001	4	10		90.8		
2001-05-01	2001	5	13.8		13.4		
2001-06-01	2001	6	16.3		23.3		
2001-07-01	2001	7	18.6		57.7		
2001-08-01	2001	8	19.1		44.2		
2001-09-01	2001	9	15.8		42.9		

2001-10-01	2001	10	15.5		95		
2001-11-01	2001	11	9.6		72.7		
2001-12-01	2001	12	6.4		78.8		
2002-01-01	2002	1	7.5		65.3		
2002-02-01	2002	2	8.8		104.2		
2002-03-01	2002	3	9.9		63.9		
2002-04-01	2002	4	11.6		22.4		
2002-05-01	2002	5	14.1		81.1		
2002-06-01	2002	6	15.9		64.7		
2002-07-01	2002	7	17.5		49.3		
2002-08-01	2002	8	18.2		63.8		
2002-09-01	2002	9	17.2		50.2		
2002-10-01	2002	10	13.8		164.1		
2002-11-01	2002	11	11.5		154.3		
2002-12-01	2002	12	8.6		126.3		
2003-01-01	2003	1	6.4		93.4		
2003-02-01	2003	2	6.2		44.4		
2003-03-01	2003	3	10		60.1		
2003-04-01	2003	4	11.6		30		
2003-05-01	2003	5	13.5		53.3		
2003-06-01	2003	6	17.8		34.3		

2003-07-01	2003	7	19.2		64.2		
2003-08-01	2003	8	20.9		3.2		
2003-09-01	2003	9	17.6		41.9		
2003-10-01	2003	10	12.4		97.1		
2003-11-01	2003	11	10.9		126.5		
2003-12-01	2003	12	7.6		97.8		
2004-01-01	2004	1	8		143.6		
2004-02-01	2004	2	7		37.1		
2004-03-01	2004	3	7.8		35.8		
2004-04-01	2004	4	10.7		75.4		
2004-05-01	2004	5	13.9		29.9		
2004-06-01	2004	6	17.1		29.3		
2004-07-01	2004	7	17.3		94		
2004-08-01	2004	8	19		101.8		
2004-09-01	2004	9	17.2		14.2		
2004-10-01	2004	10	13.3		218.9		
2004-11-01	2004	11	10.7		43.7		
2004-12-01	2004	12	7.2		88.1		
2005-01-01	2005	1	8		63.4		
2005-02-01	2005	2	5.8		43.6		
2005-03-01	2005	3	8.6		51.8		

2005-04-01	2005	4	10.8		79.6		
2005-05-01	2005	5	13.3		52.6		
2005-06-01	2005	6	17.4		39.6		
2005-07-01	2005	7	18.5		86.8		
2005-08-01	2005	8	18.2		34.2		
2005-09-01	2005	9	17.5		36.6		
2005-10-01	2005	10	15.5		88.4		
2005-11-01	2005	11	9.6		93.6		
2005-12-01	2005	12	6.5		97.6		
2006-01-01	2006	1	5.6		37.4		
2006-02-01	2006	2	5.4		90		
2006-03-01	2006	3	7		89.6		
2006-04-01	2006	4	10.3		31.2		
2006-05-01	2006	5	13.6		49.8		
2006-06-01	2006	6	17.5		26		
2006-07-01	2006	7	20.8		23		
2006-08-01	2006	8	18		44.2		
2006-09-01	2006	9	18.3		54		
2006-10-01	2006	10	15.7		87.2		
2006-11-01	2006	11	11.3		87.6		
2006-12-01	2006	12	8.1		150.7		

2007-01-01	2007	1	8.8		106.5		
2007-02-01	2007	2	9		143.2		
2007-03-01	2007	3	9		93.4		
2007-04-01	2007	4	13.4		15.7		
2007-05-01	2007	5	13.9		113.9		
2007-06-01	2007	6	16.5		105.3		
2007-07-01	2007	7	17		93.9		
2007-08-01	2007	8	17.4		82.5		
2007-09-01	2007	9	16.1		46.4		
2007-10-01	2007	10	13.2		15.4		
2007-11-01	2007	11	10.1		70.9		
2007-12-01	2007	12	7		98.4		
2008-01-01	2008	1	8.1		111.2		
2008-02-01	2008	2	7.8		37.8		
2008-03-01	2008	3	8.3		111.9		
2008-04-01	2008	4	10.3		61.7		
2008-05-01	2008	5	15.8		129.5		
2008-06-01	2008	6	16.1		29.8		
2008-07-01	2008	7	18.1		30.4		
2008-08-01	2008	8	17.7		75.5		
2008-09-01	2008	9	15.3		66.7		

2008-10-01	2008	10	12.3		118.1		
2008-11-01	2008	11	9.9		133.9		
2008-12-01	2008	12	6.2		84		
2009-01-01	2009	1	4.5		99.4		
2009-02-01	2009	2	6.1		46.6		
2009-03-01	2009	3	8.8		45		
2009-04-01	2009	4	11.2		62.7		
2009-05-01	2009	5	13.5		40.5		
2009-06-01	2009	6	16.8		46.3		
2009-07-01	2009	7	18.1		44.2		
2009-08-01	2009	8	18.3		22.7		
2009-09-01	2009	9	16.8		23.3		
2009-10-01	2009	10	14.2		73.4		
2009-11-01	2009	11	11.4		216.3		
2009-12-01	2009	12	6.9		107.9		
2010-01-01	2010	1	4.4		93.9		
2010-02-01	2010	2	6		123.3		
2010-03-01	2010	3	7.8		30.9		
2010-04-01	2010	4	11.3		14		
2010-05-01	2010	5	13		39.4		
2010-06-01	2010	6	16.7		44.6		

2010-07-01	2010	7	18.7		30.8		
2010-08-01	2010	8	17.5		101		
2010-09-01	2010	9	16.1		53.5		
2010-10-01	2010	10	13.6		124.3		
2010-11-01	2010	11	9		192.4		
2010-12-01	2010	12	4.2		109.4		
2011-01-01	2011	1	6.6		53.8	97.6	1019.3
2011-02-01	2011	2	8.5		72.2	94.9	1016.5
2011-03-01	2011	3	9		27.6	192.9	1021.7
2011-04-01	2011	4	13.8		0.2	281.4	1019.3
2011-05-01	2011	5	14		18	326.2	1019.8
2011-06-01	2011	6	15.6		44.8	256	1017.3
2011-07-01	2011	7	17		91.8	273.5	1014.7
2011-08-01	2011	8	17.5		43	234.1	1014.9
2011-09-01	2011	9	17.5		66	215.1	1016.1
2011-10-01	2011	10	14.6		40.2	117	1019.1
2011-11-01	2011	11	12.4		27	97.7	1016.7
2011-12-01	2011	12	9		247.4	51.5	1016.8
2012-01-01	2012	1	8.3		56.6	78.4	1004.1
2012-02-01	2012	2	5.7		24	104.6	1030.9
2012-03-01	2012	3	10.3		30.4	228.5	1028

2012-04-01	2012	4	9.7		112.8	253.5	1005
2012-05-01	2012	5	13.3		69.8	267.5	1017
2012-06-01	2012	6	16.1		91.6	211.2	1013.1
2012-07-01	2012	7	18.1		58.2	260	1015.8
2012-08-01	2012	8	18.6		67	271	1015.7
2012-09-01	2012	9	15.7		63.8	217.7	1016.9
2012-10-01	2012	10	13.3		176.1	106.3	1010.7
2012-11-01	2012	11	9.4		132	104.4	1011
2012-12-01	2012	12	8.3		202.8	75.5	1011.3
2013-01-01	2013	1	6.3		123.4	55.2	1014.8
2013-02-01	2013	2	5.3		67.9	109.5	1017.9
2013-03-01	2013	3	6.1		147		
2013-04-01	2013	4	9.2		55.8		
2013-05-01	2013	5	12		67.1		
2013-06-01	2013	6	14.7		23.5		
2013-07-01	2013	7	19.2		34.4		
2013-08-01	2013	8	18.7		23.6		
2013-09-01	2013	9	16.6		46.9		
2013-10-01	2013	10	14.9		134.6		
2013-11-01	2013	11	9.6		148.7		
2013-12-01	2013	12	8.5		119.4		

2014-01-01	2014	1	8.3		177.7	70.5	1004.1
2014-02-01	2014	2	8.3		133.3	129.2	1001.9
2014-03-01	2014	3	9.5		54.3	199.8	1016.7
2014-04-01	2014	4	11.8		67.8	225.4	1015
2014-05-01	2014	5	13.8		84.9	262	1016.3
2014-06-01	2014	6	17.1		12.6	322.9	1019.2
2014-07-01	2014	7	19.1		42.1	336.1	1016.6
2014-08-01	2014	8	17.4		84.8	257.7	1014.1
2014-09-01	2014	9	18.5		6.8	274.4	1018.5
2014-10-01	2014	10	15.7		91.1	156.7	1014.3
2014-11-01	2014	11	11.7		155.8	78.6	1006.2
2014-12-01	2014	12	8.8		94.4	65.2	1022.5
2015-01-01	2015	1	7.7		106.6	70.4	1017.9
2015-02-01	2015	2	6.3		99.6	130.4	1018
2015-03-01	2015	3	8.8		37.4	157.9	1022
2015-04-01	2015	4	12.4		39.7	258.9	1022
2015-05-01	2015	5	13.4		40.5	249.9	1015.9
2015-06-01	2015	6	16.6		40.9	315	1021
2015-07-01	2015	7	18.4		65.3	247.6	1016.7
2015-08-01	2015	8	18		163.4	192.3	1015.1
2015-09-01	2015	9	15.4		61.4	251.2	1017.4

2015-10-01	2015	10	13.5		56.1	139.8	1017.5
2015-11-01	2015	11	12.6		97.1	77.2	1019.6
2015-12-01	2015	12	11.6		85.8	76.1	1021.8
2016-01-01	2016	1	7.8		190.9	60.9	1011.2
2016-02-01	2016	2	7.6		134.1	106.2	1012.3
2016-03-01	2016	3	7.8		110.9	177.3	1015.4
2016-04-01	2016	4	9.8		39.5	238	1013.5
2016-05-01	2016	5	13.9		42.4	262.7	1014.7
2016-06-01	2016	6	16		62.7	180.7	1016.1
2016-07-01	2016	7	17.8		23.2	268	1019.7
2016-08-01	2016	8	18.9		53.9	293.6	1020.3
2016-09-01	2016	9	18.1		47	192.5	1018.5
2016-10-01	2016	10	13.1		56.6	177.6	1020.6
2016-11-01	2016	11	9.8		139.2	96.1	1014.8
2016-12-01	2016	12	8.2		22.6	98.7	1028.2
2017-01-01	2017	1	5.9		72.4	107.2	1024.1
2017-02-01	2017	2	8		98.4	86.6	1015.2
2017-03-01	2017	3	10.4		80.8	147.6	1016.6
2017-04-01	2017	4	11.1		40.8	285	1023.3
2017-05-01	2017	5	14.7		79.4	259	1016.5
2017-06-01	2017	6	18		42	295.7	1015.4

2017-07-01	2017	7	18.5		47.6	265.9	1015.9
2017-08-01	2017	8	17.7		104.6	244.6	1018.1
2017-09-01	2017	9	15.4		129.4	175.1	1015.6
2017-10-01	2017	10	14.4		67	134.5	1021.2
2017-11-01	2017	11	10.3		111	108.9	1018.9
2017-12-01	2017	12	8.2		181.8	56.6	1017
2018-01-01	2018	1	8.5		149.8	69.9	1014.2
2018-02-01	2018	2	4.9		56.4	149.4	1017.2
2018-03-01	2018	3	7.5	7.3	119.2	131.5	1001.4
2018-04-01	2018	4	11.9	9.5	66.4	191.2	1011
2018-05-01	2018	5	14.6	12.3	25.8	323.7	1017.7
2018-06-01	2018	6	16.9	14.8	41	268	1019.3
2018-07-01	2018	7	20.2	17.4	20	351.8	1017.3
2018-08-01	2018	8	18.5	18.5	43.6	221.7	1019
2018-09-01	2018	9	16.8	18.1	35.6	243.8	
2018-10-01	2018	10	13.7	16.1	59.8	191.8	
2018-11-01	2018	11	10.1	12.9	77.4	121	
2018-12-01	2018	12	9.2	11	109.1	50.4	
2019-01-01	2019	1	7	9.5	66.6	43.4	
2019-02-01	2019	2	8.6	8.5	63.4	162.3	
2019-03-01	2019	3	10	9.6	65.1	185.2	

2019-04-01	2019	4	11.5	10.8	54.4	227.5	
2019-05-01	2019	5	13.1	12.9	47.6	285.4	
2019-06-01	2019	6	16.5	15	91.6	264.5	
2019-07-01	2019	7	19.7	17.5	10.2	373.9	
2019-08-01	2019	8	18.6	18.4	69.6	268.8	
2019-09-01	2019	9	16.7	17.8	77.2	201.4	
2019-10-01	2019	10	13.8	16.1	194	116.5	
2019-11-01	2019	11	9.4	12.7	200	76.8	
2019-12-01	2019	12	8.6	10.5	123	92.6	
2020-01-01	2020	1	8.2	9.6	93.4	81	
2020-02-01	2020	2	9	9.1	152.4	110.2	
2020-03-01	2020	3	9	9.3	92.4	187.9	
2020-04-01	2020	4	13.5	10.9	41.8	278.7	
2020-05-01	2020	5	14.9	13.1	17.2	360.2	
2020-06-01	2020	6	16.7	15.3	97.6	251	
2020-07-01	2020	7	17.7	17.2	10.8	294.2	
2020-08-01	2020	8	19.7	18.7	48.6	293.1	
2020-09-01	2020	9	17.4	18.5	38.7	228.7	
2020-10-01	2020	10	13	15.3	241.6	118.7	
2020-11-01	2020	11	11.4	13.4	84.5	108.1	
2020-12-01	2020	12	8.2	11	259.4	74.5	

2021-01-01	2021	1	6.3	8.9	156.9	68.5	
2021-02-01	2021	2	6.9	8.1	44.8	114	
2021-03-01	2021	3	9	8.6	39.8	201	
2021-04-01	2021	4	9.4	10.1	12	316.9	
2021-05-01	2021	5	12.3	12.1	102.4	286.7	
2021-06-01	2021	6	16.6	14.8	92.2	235.4	
2021-07-01	2021	7	18.5	16.7	50.8	273.1	
2021-08-01	2021	8	17.6	17.8	30.4	239.7	
2021-09-01	2021	9	18.2	18.1	61.6	245.5	
2021-10-01	2021	10	14.4	16.4	135.5	197.3	
2021-11-01	2021	11	10	13.7	69.1	95.4	
2021-12-01	2021	12	8.5	10.5	110.2	58.1	
2022-01-01	2022	1	7.3	9.5	78.9	59.1	
2022-02-01	2022	2	8.4	9.2	70.7	106.8	
2022-03-01	2022	3	10.2	9.6	49.8	206	
2022-04-01	2022	4	11.3	11	33.4	270.4	