MODELOS PREDICTIVOS AUTORREGRESIVOS SOBRE SERIES TEMPORALES

by

Jesús Mariscal Carbón

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalusia

September 2022

MODELOS PREDICTIVOS AUTORREGRESIVOS SOBRE SERIES TEMPORALES

Autor
Jesús Mariscal Carbón

Máster en Economía, Finanzas y Computación

Supervisores
Jared Aurentz y Juan Diego Borrero Sánchez

Universidad de Huelva y Universidad Internacional de Andalucía

Abstract

The development of predictive models is a line that has been studied extensively in recent years due to its great utility for management and improvement in decision-making in sectors of all kinds. Among the models currently used, the autoregressive integrated processes of mobile media (ARIMA) constitute a family of predictive models widely used since their popularization in the seventies, due to their simplicity due to their linear nature. However, these models have serious limitations when it comes to preceding time series whose behavior is far from linear. For this reason, in this work we will describe a type of hybrid model, the integrated moving media autoregressive processes corrected by a support vector regression machine (ARIMA-SVR), which include a non-linear factor that helps to correct the limitations it presents. In addition, we will conclude our study by comparing the two models in two practical examples.

JEL classification: C02;C22; C53

Keywords: Series temporales, Modelos predictivos, ARIMA, Máquinas de vector soporte

Resumen

El desarrollo de modelos predictivos es una línea investigadora que se encuentra en auge en los últimos años debido a su gran utilidad para la gestión y la mejora en la toma de decisiones en sectores de todo tipo. De entre todos los modelos, los procesos integrados autorregresivos de medias móviles (ARIMA) constituyen una familia de modelos predictivos ampliamente utilizados desde su popularización en los años setenta, debido a su simpleza por su carácter lineal. Sin embargo, estos modelos tienen serias limitaciones a la hora de precedir en series temporales cuyo comportamiento se aleja del carácter lineal. Por ello, en este trabajo describiremos un tipo de modelo híbrido, los procesos integrados autorregresivos de medias móviles corregidos por una máquina de vector soporte de regresión (ARIMA-SVR), que incluyen un factor no lineal que ayude a corregir las limitaciones que presenta. Además, concluiremos nuestro estudio comparando los dos modelos en dos ejemplos prácticos.

Acknowledgments

En este breve pero intenso paso por el máster de MECOFIN me gustaría agradecer todo lo aprendido al plantel completo de profesores, en especial a Jared y Juan Diego por sus enseñanzas desinteresadas, tanto dentro como fuera de clase, y su gran paciencia conmigo.

También son mención especial mis compañeros, especialmente Frank, Osbel, Oti, Alex y Marcos por hacer las clases más llevaderas en un año tan duro en lo personal.

No pueden quedar aparte mi familia, mis padres, mi hermano y mi círculo más cercano. Y por último, un agradecimiento especial a mi pareja, mi compañera de vida, por apoyarme en todo momento y aportarme el equilibrio necesario para que todo salga adelante, a mi melodía favorita para que nunca dejes de sonar en mi vida.

Índice

Índice de figuras

Índice de cuadros

1. Introducción

El desarrollo de modelos predictivos es uno de los sectores que cuenta actualmente con mayor atención por parte de los investigadores de todo el mundo. La consecución de modelos matemáticos que sean capaces de modelar y predecir los valores futuros de una serie temporal se convierte en objetivo fundamental en muchos sectores como el económico, el agroalimentario, el sanitario, el sector energético y un largo etcétera, ya que los avances en este área permiten tanto al ámbito público como al privado una mejor gestión de recursos y una mejora significativa en la toma de decisiones. Por tanto, podemos encontrar un sinfín de publicaciones actuales en los que se desarrollan desde los modelos lineales más sencillos a los no lineales más complejos.

Podemos hacer una gran división en los modelos predictivos, considerando por un lado los modelos autorregresivos que se basan sólo en los valores de la serie temporal observada, y por otro los modelos con variables exógenas que ayuden a explicar el comportamiento de la variable observada. En los primeros es en los que se enmarca nuestro trabajo. La principal ventaja de estos modelos con respecto a los modelos que incluyen variables exógenas es que son sencillos de manejar, ya que sólo necesitamos llevar un registro fiable de los valores de la variable observada, sin necesidad de medir otros aspectos del entorno. Además, los modelos autorregresivos nos permiten observar las propiedades de la serie temporal objeto de estudio, permitiéndonos encontrar patrones en su comportamiento, y ayudándonos, en caso de encontrar un sistema fiable, a adelantarnos a los valores futuros de la serie. Concrétamente este trabajo estará basado en la comparación de los procesos integrados autorregresivos de medias móviles (ARIMA) con un modelo híbrido, los procesos integrados autorregresivos de medias móviles corregidos por una máquina de vector soporte de regresión (ARIMA-SVR).

Los procesos ARIMA son una familia de modelos basados en la relación lineal existente entre las observaciones de una serie temporal y su pasado más cercano y son de frecuente uso en todos los ámbitos científicos debido a su simpleza a la hora de formularlos y a las ventajas de interpretación que ofrecen siendo dando una idea muy intuitiva del comportamiento de la serie temporal que pretende modelar. Aunque tienen su origen en los años 30, se popularizaron aún más después de que George E. P. Box y Gwilym M. Jenkins publicaran su metodología sobre como modelar series temporales con procesos ARIMA (Box and Jenkins, 1976). De hecho, a día de hoy podemos encontrar publicaciones recientes sobre el tema (Mehmood et al., 2019; Hernandez-Matamoros et al., 2020; Jamil, 2020; Yang and O'Connell, 2020).

Sin embargo, aunque el carácter lineal ofrece ventajas con respecto a la simplificación de los modelos, en muchas ocasiones tienen grandes limitaciones para modelar de manera fiable la realidad de la variable que pretende replicar, sobretodo cuando las series temporales presentan muchos picos y cambios de tendencia. Por ello, desde hace unos años están surgiendo modificaciones sobre los procesos ARIMA que ayuden a corregir su comportamiento añadiendo una parte no lineal al modelo que permita ajustarse mejor a la realidad.

En este escenario se presentan los modelos ARIMA-SVR, un modelo híbrido que busca equilibrar la simpleza que aporta la linealidad de los ARIMA con el mejor ajuste que da el carácter no lineal de los modelos SVR. Sobre este concepto podemos encontrar diversas publicaciones (Wang et al., 2018; Xu et al., 2018; Sujjaviriyasup and Pitiruek, 2013) en las que se muestralos modelos ARIMA-SVR como una ventaja significativa con respecto a los procesos ARIMA.

Por tanto, este trabajo estará organizado con un primer capítulo de metodología en que se introducen los conceptos básicos necesarios, así como se describirán de forma exhaustiva ambos modelos, un segundo capítulo donde en el que se hace mención a las series temporales que se emplearán para la comparación de los modelos, un tercer capítulo que recogerá

tanto ejemplos prácticos de cómo se construyen los modelos, como una sección de comparación de resultados, y un último capítulo de conclusiones.

2. Metodología

Esta sección estará dedicada al desarrollo de los diferentes modelos predictivos considerados para este trabajo. En primer lugar haremos un pequeño resumen sobre los conceptos básicos de series temporales que serán necesarios, para continuar describiendo los modelos, pasando desde los modelos lineales más sencillos a los no lineales más complejos.

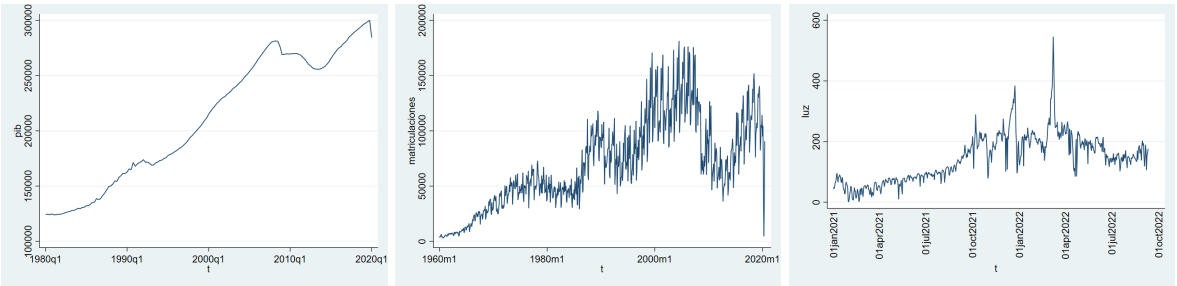
2.1. Introducción a las series temporales

En este apartado se presenta una pequeña introducción acerca de los conceptos previos sobre series temporales necesarios para el desarrollo de los diferentes modelos predictivos que serán tratados a lo largo de este trabajo. Las definiciones dadas a continuación están recopiladas de (González Casimiro, 2009; Brockwell and Davis, 2006).

En primer lugar, empezaremos definiendo que es una serie temporal.

Definición 2.1 *Una serie temporal es una muestra de $T \in \mathbb{N}$ observaciones, ordenadas y distribuidas de forma equidistante a lo largo del tiempo de una (univariante) o varias (multivariante) variables de interés.*

En este trabajo nos centramos en las series temporales univariantes. La definición 2.1 es muy intuitiva y podemos pensar en una serie temporal como un registro en el que vamos almacenando el dato de nuestra variable observada en cada instante t. Es muy importante incidir en el detalle de que se trata de una muestra ordenada en el tiempo, ya que, a diferencia de otros ámbitos en los que se trabaja con muestras poblacionales, el orden en que suceden los acontecimientos afecta directamente a la estructura y las características de la serie. En la figura 1 se muestran tres ejemplos de series temporales con separación temporal trimestral, mensual y diaria respectivamente.



(a) Valor del PIB en España. (b) Número de vehículos matriculados en España. (c) Precio del MWh en España.

Figura 1: Representación gráfica de tres series temporales.

Por lo tanto, todo modelo predictivo se basa en buscar alguna expresión que sea capaz de reproducir las características de la serie temporal que se observa. Esta expresión puede depender de la propia variable observada o de otras variables recogidas a la vez que pudieran tener relación con el valor de la variable dependiente. En nuestro estudio, nos centramos en los modelos autorregresivos.

Definición 2.2 *Sea Y_t una serie temporal de $T \in \mathbb{N}$ elementos. Sea $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ la función empleada para modelar Y_t. Diremos que f representa un modelo autorregresivo si y sólo si se da la relación*

$$\widehat{Y_t} = f\left(Y_{t-1}, Y_{t-2}, ..., Y_{t-k}\right).$$

La definición 2.2 implica que el modelo depende tan sólo del pasado de la variable de estudio. Según la definición de f se pueden diferenciar modelos lineales y no lineales, los cuales serán descritos en secciones venideras.

Así pues, podemos deducir que las características de la serie temporal son las que indican que tipo de función se adapta mejor para reproducirla, por lo que no existe un único método que sea el mejor para todos los casos. Pero, ¿qué determina las características de una serie temporal? Para resolver esta cuestión introducimos el concepto de proceso estocástico.

Definición 2.3 *Un proceso estocástico Y_t es una sucesión de T variables aleatorias relacionadas entre sí, que comparten una ley de distribución conjunta.*

Aunque esta definición no exija que la sucesión esté ordenada temporalmente, nosotros sólo vamos a considerar procesos estocásticos cuyas variables están ordenadas de forma equidistante en el tiempo. Un proceso estocástico está caracterizado por su función de distribución conjunta, $F(Y_1,...,Y_T)$ o por sus momentos. Entre sus momentos, de especial importancia son la esperanza, $E[Y_t] = \mu_t \in \mathbb{R}$, la varianza $VAR[Y_t] = E\left[(Y_t - \mu_t)^2\right]$ y la covarianza, $cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)]$, conceptos ampliamente utilizados en la teoría de la probabilidad y de los procesos estocásticos.

Desde este punto de vista de la definición 2.3, podemos ver una serie temporal como una realización muestral de un proceso estocástico, siendo cada uno de los datos que la componen un resultado puntual de cada una de las variables aleatorias que conforman el proceso estocástico. Por lo tanto, buscar un modelo que se ajuste a la serie temporal y consiga predecir sus valores futuros es equivalente a buscar el proceso estocástico que ha generado la serie temporal. Sin embargo, modelar y predecir una serie temporal es difícil en general, por ello nos centraremos en las series temporales que estén generadas por procesos estocásticos estacionarios.

Definición 2.4 *Sea Y_t un proceso estocástico y F su función de distribución conjunta. Diremos que un proceso estocástico es estacionario en sentido estricto si y sólo si para cualquier subconjunto de variables aleatorias $\left(Y_{t_1},...,Y_{t_n}\right) \in Y_t$ y cualquier $h \in \mathbb{Z}$.*

$$F\left(Y_{t_1}, Y_{t_2},..., Y_{t_n}\right) = F\left(Y_{t_{1+h}}, Y_{t_{2+h}},..., Y_{t_{n+h}}\right).$$

La condición de estacionariedad en sentido estricto es muy fuerte matemáticamente hablando, por lo que nos conformaremos con que los procesos estocásticos sean estacionarios en sentido amplio.

Definición 2.5 *Sea Y_t un proceso estocástico. Diremos que Y_t es estacionario en sentido amplio si y sólo si*

a) $E[Y_t] = \mu < \infty$ *para todo* $t = 1,...,T$.

b) $VAR[Y_t] = \sigma^2 < \infty$ *para todo* $t = 1,...,T$.

c) Sean (Y_s, Y_m), variables aleatorias del proceso estocástico Y_t con distancia temporal $h \in \mathbb{Z}$, entonces

$$cov(Y_s, Y_m) = E[Y_s - \mu][Y_m - \mu] = \gamma_{|s-m|} = \gamma_h < \infty.$$

Es decir, la covarianza entre dos variables aleatorias de Y_t sólo depende de la distancia temporal que existe entre ellos.

En general, los conceptos de covarianza y correlación nos indican la relación existente entre dos variables aleatorias y como afecta el comportamiento de una a la otra. Este concepto se puede aplicar a los procesos estocásticos estacionarios, adaptándolo en este caso como las funciones de autocovarianza y autocorrelación, las cuales nos indican la relación que tienen las variables aleatorias que forman el proceso estocástico entre si.

Definición 2.6 *Sea Y_t un proceso estocástico estacionario con $E[Y_t] = \mu$. Definiremos la función de autocovarianza como*

$$\gamma_h = cov(Y_t, Y_{t+h}) = E[(Y_t - \mu)(Y_{t+h} - \mu)],$$

para $h = 0, 1, 2, 3, ...$ la diferencia temporal entre las variables aleatorias.

De la definición 2.6 podemos obtener trivialmente dos conclusiones.

a) En primer lugar, para $h = 0$, se tiene $\gamma_0 = VAR[Y_t]$.

b) La función de autocovarianza es simétrica,

$$\gamma_h = E[(Y_t - \mu)(Y_{t+h} - \mu)] = E[(Y_{t-h} - \mu)(Y_t - \mu)] = \gamma_{-h}.$$

A partir de la función de autocovarianza definimos la función de autocorrelación. La función de autocorrelación de un proceso estocástico estacionario mide el grado de correlación lineal que existe entre dos variables aleatorias separadas h periodos. Matemáticamente lo definimos como sigue.

Definición 2.7 *Sea Y_t un proceso estocástico estacionario, sea γ_h su función de autocovarianza. Definimos la función de autocorrelación como*

$$\rho_h = \frac{\gamma_h}{\sqrt{VAR[Y_t]\,VAR[Y_t]}} = \frac{\gamma_h}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_h}{\gamma_0}.$$

Tras las definiciones relacionadas con los procesos estocásticos estacionarios debemos tener en cuenta que tanto los momentos como las funciones de autocovarianza y correlación son conceptos teóricos que a priori no conocemos, ya que, recordamos que vamos a tratar las series temporales como una realización muestral de un proceso estocástico desconocido. Por lo tanto, tenemos que estimar estos momentos con los datos que conocemos de la serie de la siguiente forma:

a) $\widehat{\mu} = \frac{1}{T}\sum_{t=0}^{T} Y_t.$

b) $\widehat{\gamma_0} = \frac{1}{T}\sum_{t=0}^{T} (Y_t - \widehat{\mu})^2.$

c) $\widehat{\gamma_h} = \frac{1}{T}\sum_{t=0}^{T} (Y_t - \widehat{\mu})(Y_{t+h} - \widehat{\mu})$ para $h = 1, 2, ..., T.$

d) $\widehat{\rho} = \frac{\widehat{\gamma_h}}{\widehat{\gamma_h}}$ para $h = 1, 2, ..., T.$

Por último para terminar esta sección, vamos a definir un tipo de proceso estocástico estacionario que tiene una gran relevancia en el campo de las series temporales y los modelos predictivos, el ruido blanco.

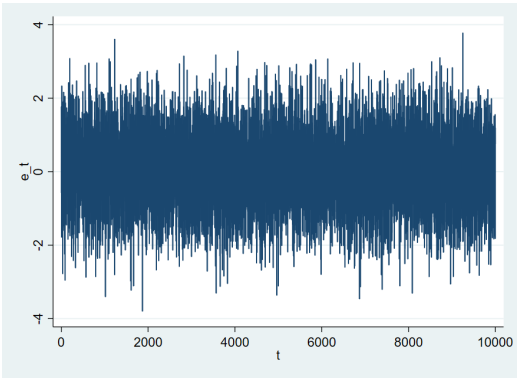
Definición 2.8 *Sea ϵ_t un proceso estocástico. Diremos que ϵ_t es un proceso estocástico si y sólo si se cumple que*

a) $E[\epsilon_t] = 0$ *para todo* t.

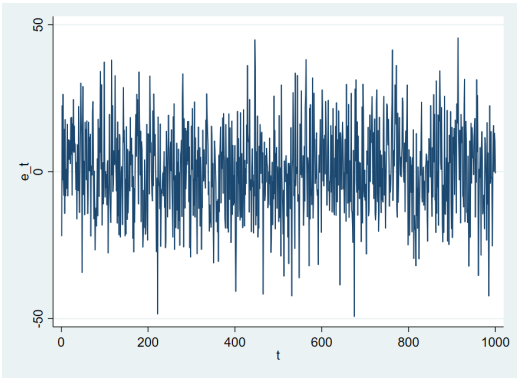
b) $VAR[\epsilon_t] = \sigma^2$ *para todo* t.

c) $cov(\epsilon_s, \epsilon_m) = 0$ *para todo* s, m *con* $s \neq m$.

En lo que sigue, nos referiremos a un proceso de ruido blanco con la notación $\epsilon_t \sim RB\left(0, \sigma_\epsilon^2\right)$ donde σ_ϵ^2 es la varianza del proceso de ruido blanco ϵ_t. En la figura 2 vemos representados dos procesos de ruido blanco de diferente longitud y con distintos valores de σ^2.



(a) Ruido blanco con $T = 100000$ y $\sigma^2 = 1$.



(b) Ruido blanco con $T = 1000$ y $\sigma^2 = 15$.

Figura 2: Representación gráfica de dos procesos de ruido blanco.

2.2. Procesos ARIMA

En nuestro recorrido por los diferentes modelos predictivos, comenzamos dedicando esta sección a una familia de modelos lineales conocida como procesos autorregresivos integrados de medias móviles (ARIMA). Para ello, comenzaremos describiendo los modelos más sencillos, procesos autorregresivos (AR) y de medias móviles (MA), para terminar construyendo los procesos autorregresivos de medias móviles (ARMA) y finalmente los procesos ARIMA. Las definiciones y demostraciones de esta sección están basadas en (González Casimiro, 2009; Brockwell and Davis, 2006; Uriel Jímenez, 1991).

2.2.1. Procesos AR(p)

Vamos a comenzar definiendo los procesos autorregresivos de orden p (AR(p)).

Definición 2.9 *Sea Y_t un proceso estocástico con $E[Y_t] = 0$, y $p \in \mathbb{N}$ el orden del proceso autorregresivo. Entonces, definimos un proceso autorregresivo de orden p (AR(p)) como un proceso estocástico que sigue la relación*

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t, \quad donde \quad \epsilon_t \sim RB\left(0, \sigma_\epsilon^2\right).$$

En la definición 2.9 hemos tomado un proceso estocástico de media cero para facilitar la comprensión y las demostraciones que se presentan a continuación. Esta asunción es arbitraria, y podemos definir un proceso $AR(p)$ con media distinta de cero utilizando la expresión

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t.$$

Para comenzar a estudiar las propiedades de los procesos $AR(p)$, vamos a comenzar considerando los procesos $AR(1)$ para tener una primera intuición.

• PROCESOS $AR(1)$

Siguiendo la definición 2.9, tenemos que un proceso $AR(1)$ se define según la expresión

$$Y_t = \phi Y_{t-1} + \epsilon_t, \tag{1}$$

y vamos a comprobar que condiciones debe cumplir el parámetro ϕ para que el proceso $AR(1)$ sea estacionario en sentido amplio. Para ello, tendremos que probar cada uno de los puntos de la definición 2.5.

a) En primer lugar, tenemos que comprobar que $E[Y_t] = \mu < \infty$. Tomemos pues la esperanza de nuestro proceso

$$E[Y_t] = E[\phi Y_{t-1} + \epsilon_t], \tag{2}$$

utilizando la linealidad de la esperanza, y teniendo en cuenta que $\epsilon_t \sim RB(0, \sigma_\epsilon^2)$, obtenemos

$$E[Y_t] = \phi E[Y_{t-1}], \tag{3}$$

si agrupamos ahora en el lado izquierdo y tomamos la hipótesis de que el proceso sea estacionario, obtenemos

$$(1-\phi) E[Y_t] = 0, \tag{4}$$

de donde se obtiene de forma trivial que

$$E[Y_t] = \frac{0}{1-\phi}. \tag{5}$$

Del procedimiento descrito en (2)-(5) obtenemos que para que se cumpla la estacionariedad débil en media, debe ser $\phi \neq 1$.

b) Para ver si se cumple la segunda condición vamos a calcular la varianza del proceso $AR(1)$. Comencemos por tanto tomando

$$VAR[Y_t] = E\left[(Y_t - E[Y_t])^2\right]. \tag{6}$$

Sustituimos el término Y_t por su expresión, teniendo en cuenta que estamos definiendo el proceso $AR(1)$ con espereanza cero

$$VAR[Y_t] = E\left[(\phi Y_{t-1} + \epsilon_t)^2\right], \tag{7}$$

desarrollando los cuadrados y agrupando convenientemente

$$VAR[Y_t] = \phi^2 E\left[Y_t^2\right] + E\left[e_t^2\right] + 2E[Y_t e_t], \tag{8}$$

ahora, teniendo en cuenta que $\epsilon_t \sim RB(0, \sigma_\epsilon^2)$ y que Y_t y ϵ_t son independientes, podemos escribir

$$VAR[Y_t] = \phi^2 E\left[Y_t^2\right] + \sigma_\epsilon^2, \tag{9}$$

para terminar, tomamos la hipótesis de que Y_t es estacionario para ver que condición debemos poner sobre ϕ

$$VAR[Y_t] = \gamma_0 = \frac{\sigma_\epsilon^2}{1-\phi^2}. \tag{10}$$

Por lo tanto, deducimos que para darse la condición $\gamma_0 < \infty$, necesariamente deducimos de (6)-(10) que $|\phi| < 1$.

c) Por último, tenemos que comprobar que la función de autocovarianza también es finita y que su valor sólo depende de la posición entre las variables aleatorias del proceso. Para ello, tomemos la definición de γ_h

$$\gamma_h = E\left[(Y_t - E\left[Y_t\right])(Y_{t+h} - E\left[Y_{t+h}\right])\right]. \tag{11}$$

Como estamos considerando un proceso estocástico de esperanza cero, tenemos

$$\gamma_h = E\left[Y_t Y_{t+h}\right], \tag{12}$$

empleamos la definición del proceso $AR(1)$ sobre el término Y_{t+h}

$$\gamma_h = E\left[Y_t\left(\phi Y_{t+h-1} + e_{t+h}\right)\right], \tag{13}$$

empleando la linealidad de la esperanza obtenemos

$$\gamma_h = \phi E\left[Y_t Y_{t+h-1}\right] + E\left[Y_t e_{t+h}\right], \tag{14}$$

utilizando la definición de la función de autocovarianza y la independencia entre Y_t y ϵ_t, obtenemos

$$\gamma_h = \phi \gamma_{h-1}. \tag{15}$$

De las ecuaciones (11)-(15) hemos obtenido una expresión recursiva que podemos resumir en

$$\gamma_h = \phi^h \gamma_0, \tag{16}$$

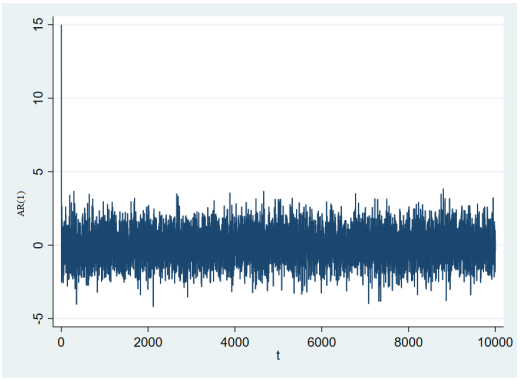
por lo que podemos concluir, en primer lugar que $\gamma_h < \infty$ ya que para que sea un proceso estacionario $|\phi| < 1$, y en segundo lugar, vemos que la autocovarianza entre dos variables aleatorias sólo depende de la distancia temporal entre ellas.

Por lo tanto, concluimos que un proceso $AR(1)$ será estacionario si y sólo si $|\phi| < 1$. En la figura 3 se muestra la representación gráfica de diferentes procesos $AR(1)$ para distintos valores de ϕ. Podemos comprobar como cuando ϕ se acerca a 1, la función de autocorrelación tarda más en decrecer. Por tanto, una primera intuición visual para saber si un proceso $AR(1)$ es estacionario es comprobar si la función de autocorrelación decae rápidamente.

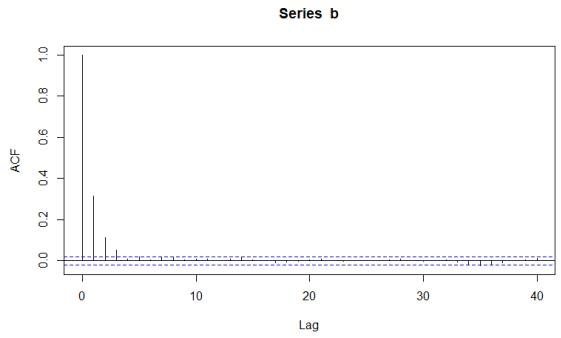
Una vez hemos analizados los procesos $AR(1)$, podemos entender los modelos $AR(p)$ como una generalización de ellos. Cabe destacar, que a veces en lugar de la definición 2.9 se expresan los procesos $AR(p)$ en función del operador de retardos. El operador de retardos, L, se define como un operador tal que $Y_t L = Y_{t-1}$.Por lo tanto, podemos definir un proceso $AR(p)$ como

$$(1 - L - L^2 - ... - L^p) Y_t = \epsilon_t \quad \text{con} \quad \epsilon_t \sim RB(0, \sigma_\epsilon^2). \tag{17}$$

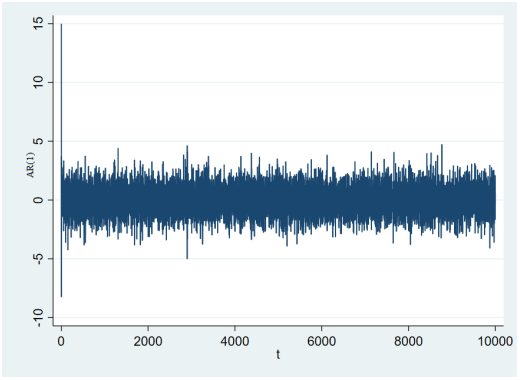
De aquí podemos definir el polinomio autorregresivo de la siguiente forma.



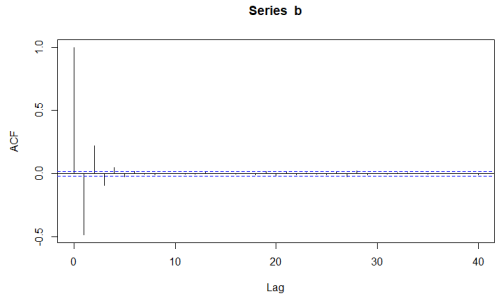
(a) Proceso $AR(1)$, $\phi = 0{,}3$.



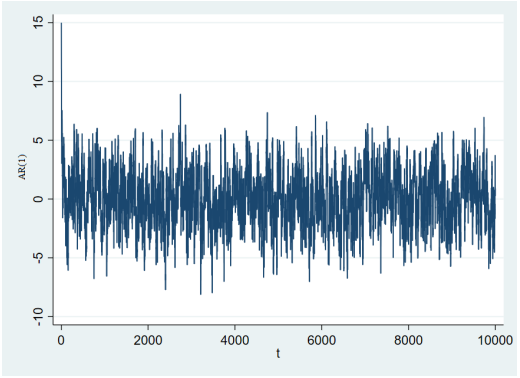
(b) Función autocorrelación para proceso $AR(1)$, $\phi = 0{,}3$.



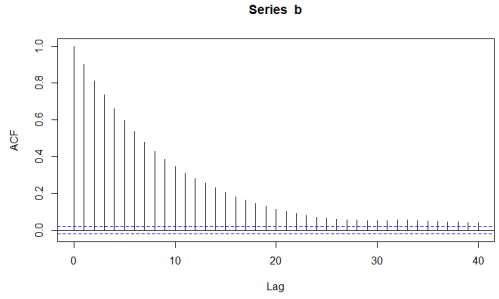
(c) Proceso $AR(1)$, $\phi = -0{,}5$.



(d) Función autocorrelación para proceso $AR(1)$, $\phi = -0{,}5$.



(e) Proceso $AR(1)$, $\phi = 0{,}9$.



(f) Función autocorrelación para proceso $AR(1)$, $\phi = 0{,}9$.

Figura 3: Representación gráfica de diferentes procesos $AR(1)$ y su función de autocorrelación.

Definición 2.10 *Sea Y_t un proceso autorregresivo que sigue la ecuación (17). Llamaremos polinomio autorregresivo al polinomio*

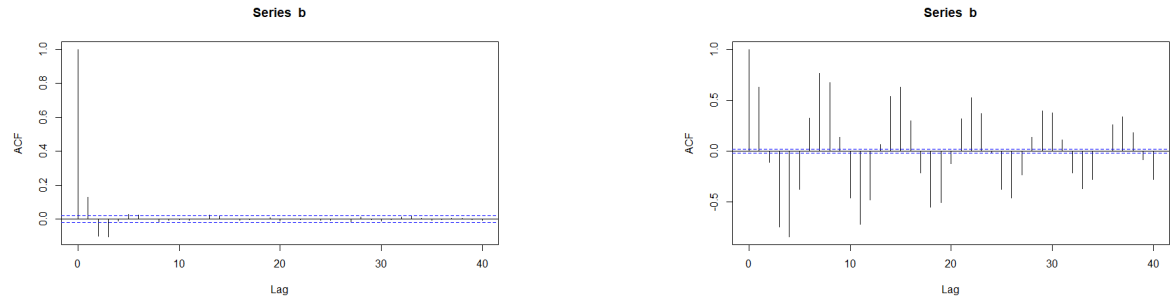
$$\Phi_p(L) = (1 - L - L^2 - ... - L^p),$$

donde el vector $\left(\phi_1, ..., \phi_p\right)$ es conocido como el vector de parámetros autorregresivos.

Aunque en general las demostraciones son más complejas para el caso $p \geq 2$, podemos seguir la misma técnica para obtener tanto la esperanza, como las funciones γ_h y ρ_h. Como principal resultado sobre la estacionariedad de los procesos $AR(p)$ podemos encontrar el siguiente teorema.

Teorema 2.1 *Diremos que un proceso $AR(p)$ con $p \in \mathbb{N}$ es estacionario si y sólo si $\Phi_p(L)$ tiene todas sus raíces fuera del círculo unidad.*

La demostración de este teorema es muy compleja matemáticamente hablando, por lo que la omitiremos en este trabajo. Aún así, visualmente podemos tener la misma intuición, y comprobaremos que la función de autocorrelación desciende rápidamente cuando el proceso $AR(p)$ es estacionario, y desciende muy lentamente cuando no es estacionario. Cuando el polinomio $\Phi_p(L)$ tiene alguna raíz dentro del círculo unidad, diremos que el proceso $AR(p)$ tiene una raíz unitaria. En la figura 4 podemos ver la comparación entre dos procesos $AR(3)$, estacionario y no estacionario respectivamente.



(a) $AR(3)$ estacionario.

(b) $AR(3)$ no estacionario.

Figura 4: Representación gráfica de la función de autocorrelación para dos procesos $AR(3)$.

2.2.2. Procesos MA(q)

Los modelos de medias móviles de orden q ($MA(q)$) se engloban dentro de los modelos lineales que están siendo objeto de estudio en esta sección, y están definidos de la siguiente forma.

Definición 2.11 *Sea Y_t un proceso estocástico y $\epsilon_t \sim RB(0, \sigma^2\epsilon)$. Diremos que Y_t es un proceso de medias móviles si y sólo si se tiene que*

$$Y_t = \sum_{j=i}^{q} \theta_j \epsilon_{t-j} + e_t.$$

Al igual que ocurre en en los procesos $AR(p)$ la pregunta que deberíamos hacernos es que condiciones tenemos que imponer a los valores de θ_j para cumplir la condición de estacionariedad. Nuevamente, vamos a centrarnos primero en describir el caso $MA(1)$ para ver como funcionan este tipo de procesos de forma sencilla.

• PROCESOS $MA(1)$

Siguiendo la definición 2.11,un $MA(1)$ está definido como $Y_t = \theta\epsilon_{t-1} + \epsilon_t$. Vamos a comprobar las condiciones de θ para que un proceso $MA(1)$ sea estacionario.

a) Empecemos comprobando la condición sobre la esperanza.

$$E[Y_t] = E[\theta\epsilon_{t-1} + \epsilon_t]. \tag{18}$$

Empleando la linealidad de la esperanza, obtenemos

$$E[Y_t] = \theta E[\epsilon_{t-1}] + E[\epsilon_t], \tag{19}$$

y como $\epsilon_t \sim RB(0, \sigma^2\epsilon)$, tenemos que

$$E[Y_t] = 0 \quad \text{para todo} \quad \theta \in \mathbb{R}. \tag{20}$$

Por lo tanto de (18)-(20) obtenemos que la esperanza es constante y finita para cualquier valor de θ.

b) Pasemos ahora a comprobar si la varianza es finita y constante.

$$VAR(Y_t) = E\Big[(Y_t - E[Y_t])^2\Big], \tag{21}$$

sabemos que $E[Y_t] = 0$, por lo tanto, tenemos

$$VAR(Y_t) = E[Y_t^2], \tag{22}$$

y utilizando la definición del proceso $MA(1)$,

$$VAR(Y_t) = E\Big[(\theta\epsilon_{t-1} + \epsilon_t)^2\Big], \tag{23}$$

desarrollamos el cuadrado de la esperanza y separamos por linealidad

$$VAR(Y_t) = \theta^2 E\Big[(\epsilon_{t-1})^2\Big] + E\Big[(\epsilon_t)^2\Big] + 2\theta E\left[\epsilon_t\epsilon_{t-1}\right], \tag{24}$$

por último, dada la definición de ruido blanco

$$VAR(Y_t) = \sigma_\epsilon^2\Big(\theta^2 + 1\Big). \tag{25}$$

De las ecuaciones (21)-(25) deducimos que la varianza será constante y finita para cualquier $\theta \in \mathbb{R}$.

c) Una vez hemos comprobado el valor de θ para la varianza (γ_0), tenemos que comprobar el valor para γ_h con $h \geq 1$. Vamos a proceder de forma recursiva, comenzando con γ_1.

$$\gamma_1 = E\left[Y_t Y_{t-1}\right]. \tag{26}$$

Sustituimos Y_t por su expresión,

$$\gamma_1 = E\left[(\theta\epsilon_{t-1} + \epsilon_t)(\theta\epsilon_{t-2} + \epsilon_{t-1})\right], \tag{27}$$

desarrollando el producto, y empleando la linealidad de la esperanza

$$\gamma_1 = E\left[\epsilon_t\epsilon_{t-1}\right] + \theta^2 E\left[\epsilon_{t-1}\epsilon_{t-2}\right] + \theta E\Big[\epsilon_{t-1}^2\Big] + \theta E\left[\epsilon_t\epsilon_{t-2}\right]. \tag{28}$$

Aplicando las propiedades del ruido blanco, tenemos

$$\gamma_1 = \theta\sigma_\epsilon^2. \tag{29}$$

Ahora, seguimos el mismo procedimiento con γ_2.

$$\gamma_2 = E\left[(\theta\epsilon_{t-1} + \epsilon_t)(\theta\epsilon_{t-3} + \epsilon_{t-2})\right], \tag{30}$$

y desarrollando igual que en el caso anterior

$$\gamma_2 = E[\epsilon_t \epsilon_{t-2}] + \theta E[\epsilon_t \epsilon_{t-3}] + \theta E[\epsilon_{t-1} \epsilon_{t-2}] + \theta^2 E[\epsilon_{t-1} \epsilon_{t-3}]. \tag{31}$$

Debido a las propiedades del ruido blanco, en el que $cov(\epsilon_s, \epsilon_m) = 0$ para $s \neq m$, concluimos

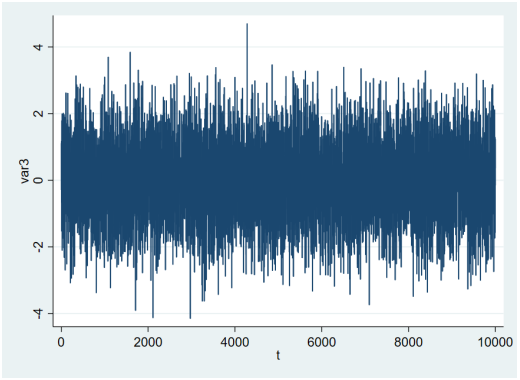
$$\gamma_2 = 0. \tag{32}$$

Ahora bien, si continuamos con la progresión veremos que siempre nos quedan covarianzas cruzadas, por lo que

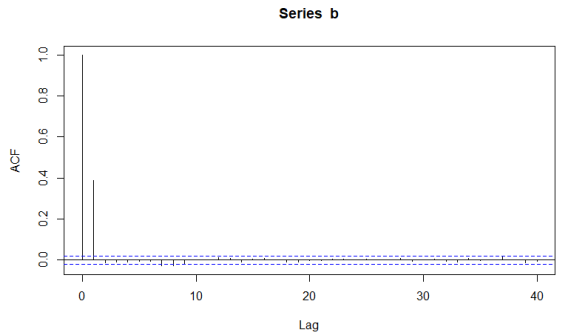
$$\gamma_h = 0 \quad \text{para todo} \quad h \geq 2. \tag{33}$$

Recapitulando toda la información retenida de las ecuaciones (26)-(33), tenemos que γ_h es finita para todo h, y además depende tan sólo de la posición temporal relativa entre las variables aleatorias del proceso $MA(1)$ que estamos considerando

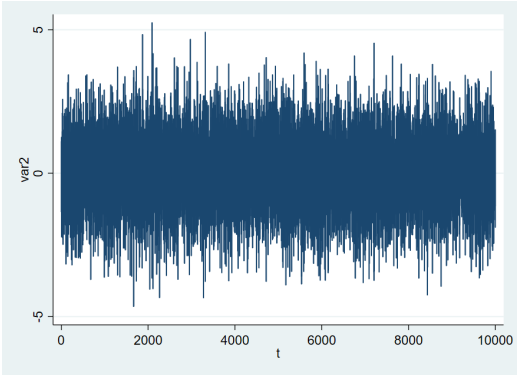
De todo esto podemos deducir que los procesos $MA(1)$ son siempre estacionarios para todo $\theta \in \mathbb{R}$. Además, podemos ver que la función de autocovarianza, y por tanto de autocorrelación, decaen a cero para $h \geq 2$. En la figura 5 podemos observar ejemplos de procesos $MA(1)$ junto a sus funciones de correlación.



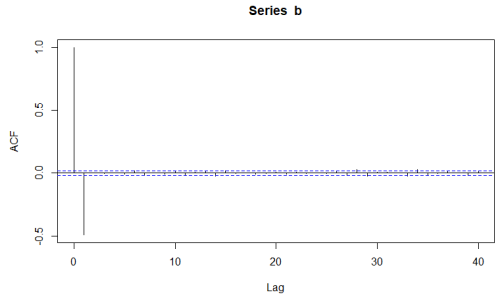
(a) Proceso $MA(1)$, $\theta = 0{,}5$.



(b) Función autocorrelación para proceso $MA(1)$, $\theta = 0{,}5$.



(c) Proceso $MA(1)$, $\theta = -0{,}8$.

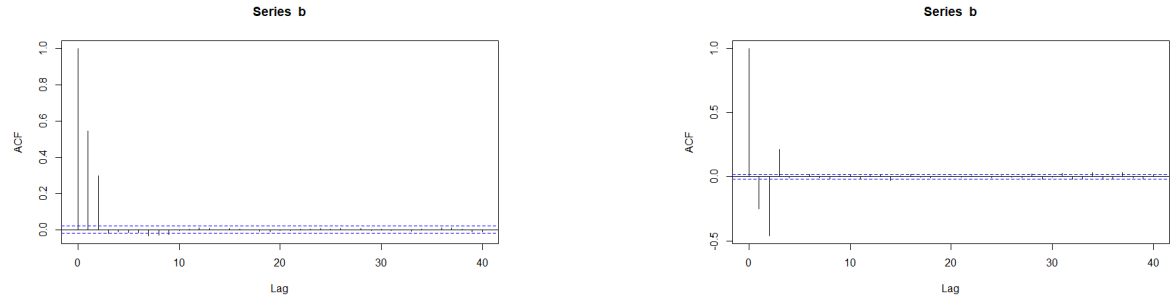


(d) Función autocorrelación para proceso $MA(1)$, $\theta = -0{,}8$.

Figura 5: Representación gráfica de diferentes procesos $MA(1)$ y su función de autocorrelación.

Al igual que hicimos con los procesos $AR(p)$, una vez desarrollados los procesos $MA(1)$ estamos en disposición de generalizar hacia los procesos $MA(q)$. En este caso, la condición

de estacionariedad se puede generalizar de los procesos $MA(1)$, siendo todo proceso $MA(q)$ estacionario para un q finito. Para comprobarlo basta ver que, para cualquier proceso $MA(q)$ su esperanza es nula al ser suma de variables de ruido blanco, la varianza será finita y constante, sólo hay que utilizar la demostración realizada para los procesos $MA(1)$ y tener en cuenta que, aunque aparezcan más términos, todos los términos de covarianza cruzada que aparezcan son cero, por lo que nos quedará un valor finito y constante. Por último, por el mismo motivo, la función $\gamma_h < \infty$ para $0 \leq h \leq q$, y $\gamma_h = 0$ para $h \geq q+1$. Por lo tanto, además tendremos que la función de autocorrelación será nula a partir del término $q+1$, igual que ocurría en el caso $MA(1)$. En la figura 6 podemos ver la función de autocorrelación para un proceso $MA(2)$ y un proceso $MA(3)$.



(a) $MA(2)$. (b) $MA(3)$.

Figura 6: Representación gráfica de la función de autocorrelación para dos procesos $MA(2)$ y $MA(3)$.

Por último, tal y como sucede con los procesos $AR(p)$, podemos escribir los procesos $MA(q)$ en función al operador de retardos, de forma que tendremos que un proceso estocástico es un proceso de medias móviles si y sólo si podemos escribirla como

$$Y_t = (1 + L + L^2 + ... + L^q)\epsilon_t, \tag{34}$$

o utilizando el polinomio de medias móviles,

$$Y_t = \Theta_q(L)\epsilon_t. \tag{35}$$

2.2.3. Procesos ARMA(p,q)

La definición de los procesos $ARMA(p,q)$ viene derivada directamente de la definición de los procesos $AR(p)$ y de los procesos $MA(q)$. Por lo tanto vamos a considerar su definición de la siguiente forma.

Definición 2.12 *Sea Y_t un proceso estocástico de media cero, sean $p,q\mathbb{N}$ y sea $\epsilon_t \sim RB(0,\sigma_\epsilon^2)$. Diremos que Y_t es un proceso $ARMA(p,q)$ si y solo si se tiene que*

$$Y_t = \sum_{i=i}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t.$$

Al igual que sucedía con los procesos $AR(p)$, hemos tomado Y_t como un proceso estocástico de media cero por simplicidad, pero no tiene por qué ser así en general. Por lo tanto, tan sólo tendremos que añadir a la definición 2.12 un término ϕ_0.

Es habitual encontrar la definición de los procesos $ARMA$ en función del operador de retardos,

$$\Phi_p(L)Y_t = \Theta_q(L)\epsilon_t. \tag{36}$$

Esta definición más compacta puede servir de gran ayuda en diferentes resultados ya que, a pesar de tratarse de un operador, $\Phi_p(L)$ y $\Theta_q(L)$ cumplen las mismas propiedades que los polinomios.

La pregunta lógica en este punto puede ser qué condición debe cumplir un proceso $ARMA(p,q)$ para ser estacionario. Dada su definición, podemos deducir de forma trivial que los procesos $ARMA(p,q)$ se basan en una parte autorregresiva y otra de medias móviles. De la misma forma, hemos visto que un proceso $MA(q)$ con $q \in \mathbb{N}$ siempre es estacionario, por lo que la condición de estacionariedad recaerá en su parte autorregresiva. Por consiguiente, como consecuencia directa del teorema 2.1 tendremos el siguiente teorema.

Teorema 2.2 *Diremos que un proceso $ARMA(p,q)$ con $p,q \in \mathbb{N}$ es estacionario si y sólo si $\Phi_p(L)$ tiene todas sus raíces fuera del círculo unidad.*

De manera análoga a lo que sucedía en los procesos $AR(p)$, cuando $\Phi_p(L)$ tiene una raíz de modulo 1, decimos que el proceso $ARMA(p,q)$ tiene una raíz unitaria, y por tanto, no es un proceso estacionario.

Como ejemplo concreto de proceso $ARMA(p,q)$ vamos a describir uno de los modelos más utilizados los procesos $ARMA(1,1)$.

• PROCESOS $ARMA(1,1)$

Empleando la definición 2.12, diremos que un proceso $ARMA(1,1)$ es un proceso que sigue la expresión

$$Y_t = \phi Y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t.$$

Aplicando los teoremas 2.1 y 2.2, sabemos que el proceso $ARMA(1,1)$ será estacionario si y sólo si $|\phi| < 1$. En cuanto a los momentos del proceso estocástico, tenemos que para un $ARMA(1,1)$ estacionario

a) En cuanto a la esperanza, por definición tenemos

$$E[Y_t] = \phi E[Y_{t-1}] + \theta E[\epsilon_{t-1}] + \epsilon_t,$$

de donde se tiene

$$(1-\phi)E[Y_t] = 0,$$

y en consecuencia

$$E[Y_t] = 0.$$

b) Ahora, con respecto a la función de autocovarianza,

$$\gamma_0 = E\left[Y_t^2\right] = E\left[(\phi Y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t)^2\right],$$

si desarrollamos el cuadrado y utilizamos la linealidad de la esperanza

$$\gamma_0 = \phi^2 E[Y_{t-1}^2] + E[\epsilon_t^2] + \theta^2 E[\epsilon_{t-1}^2] + 2\phi E[Y_{t-1}e_t] + 2\phi E[Y_{t-1}\epsilon_{t-1}] + 2\theta E[\epsilon_t\epsilon_{t-1}],$$

calculando y agrupando, teniendo en cuenta que las covarianzas cruzadas son cero, tenemos

$$\gamma_0 = \phi^2\gamma_0 + \sigma_\epsilon^2 + \theta^2\sigma_\epsilon^2 + 2\phi\theta\sigma_\epsilon^2,$$

despejando γ_0, tenemos

$$\gamma_0 = \frac{\sigma_\epsilon^2 + \theta^2\sigma_\epsilon^2 + 2\phi\theta\sigma_\epsilon^2}{1-\phi^2}.$$

c) Para calcular γ_1, vamos de nuevo a la definición

$$\gamma_1 = E[Y_t Y_{t-1}],$$

sustituimos Y_t para ponerlo en función de Y_{t-1}

$$\gamma_1 = E[(\phi Y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t) Y_{t-1}],$$

empleando el mismo procedimiento que en el caso anterior

$$\gamma_1 = \phi E[Y_{t-1}^2] + \theta E[Y_{t-1}\epsilon_{t-1}] + E[Y_{t-1}\epsilon_t],$$

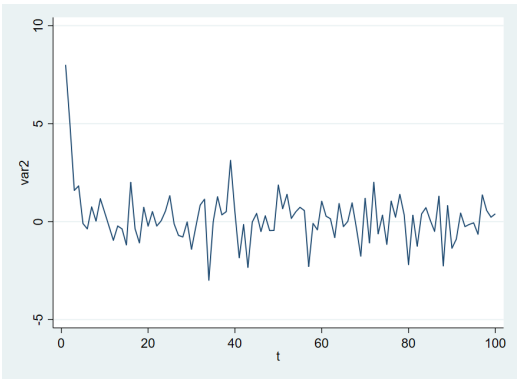
por último sustituimos cada término por su valor

$$\gamma_1 = \phi\gamma_0 + \theta\sigma_\epsilon^2.$$

Para calcular el resto, sólo hay que tener en cuenta que

$$E[Y_{t-1}\epsilon_{t-1}] = E[(\phi Y_{t-2} + \theta\epsilon_{t-2} + \epsilon_{t-1})\epsilon_{t-1}] = \phi E[Y_{t-2}\epsilon_{t-1}] + \theta E[\epsilon_{t-1}\epsilon t-2] + E[\epsilon_{t-1}^2] = \sigma_\epsilon^2.$$

En la figura 7 encontramos una representación de un proceso $ARMA(1,1)$ y su correspondiente función de autocorrelación.



(a) Proceso $ARMA(1,1)$.

(b) Función de autocorrelación de proceso $ARMA(1,1)$

Figura 7: Representación gráfica de un proceso $ARMA(1,1)$ y su función de autocorrelación.

2.2.4. Procesos ARIMA(p,d,q)

Para finalizar este primer repaso por los modelos lineales, llegamos a los procesos $ARIMA(p,d,q)$. Durante todo este apartado siempre hemos tenido una pregunta en la cabeza, ¿ es estacionario el proceso estocástico que estamos tratando? Ya dijimos, que en general es difícil modelar una serie temporal, o desde un punto de vista estocástico, es difícil encontrar el proceso estocástico que genera la serie temporal que es objeto de estudio. Pero, ¿ existe alguna forma de transformar un proceso estocástico no estacionario en uno estacionario? Pues en esta sección responderemos a esta pregunta.

Imaginemos que tenemos un proceso estocástico $ARMA(p,q)$ que no es estacionario, lo que implica según el teorema 2.2 que su polinomio de autocorrelación $\Phi_p(L)$ tiene al menos una raíz unitaria. Supongamos por simplicidad, que el proceso $ARMA(p,q)$ tiene una raíz unitaria. Como sabemos por la ecuación (36), el proceso $ARMA(p,q)$ se puede describir como

$$\Phi_p(L)Y_t = \Theta_q(L)\epsilon_t.$$

Podemos escribir el polinomio $\Phi_p(L)$ en función de sus raíces L_1,L_2,...,L_{p-1},1 de la forma siguiente

$$(1 - L_1^{-1}L)(1 - L_2^{-1}L)\cdots(1 - L_{p-1}^{-1}L)(1 - L)Y_t = \Theta_q(L)\epsilon_t. \tag{37}$$

si aunamos los $p-1$ primeros factores en un único polinomio $\Phi_{p-1}(L)$, tenemos en consecuencia que se debe cumplir la expresión

$$\Phi_{p-1}(L)(1 - L)Y_t = \Theta_q(L)\epsilon_t, \tag{38}$$

si aplicamos el operador de retardos, llegamos a la expresión

$$\Phi_{p-1}(L)(Y_t - Y_{t-1}) = \Theta_q(L)\epsilon_t. \tag{39}$$

donde el término $(Y_t - Y_{t-1}) = \Delta Y_t$ representa las primeras diferencias del proceso Y_t. Sabemos por construcción que $\Phi_{p-1}(L)$ no tiene raíces unitarias por lo que ΔY_t es un proceso $ARMA(p-1,q)$ estacionario. De este modo, hemos conseguido transformar un proceso no estacionario en uno estacionario a través de una diferenciación. Si generalizamos la demostración, podemos conseguir el mismo resultado para un proceso $ARMA(p,q)$ que tenga $d \in \mathbb{N}$ raíces unitarias. De aquí surge la definición de proceso $ARIMA(p,d,q)$.

Definición 2.13 *Diremos que Y_t es un proceso integrado de orden $d \in \mathbb{N}$, si Y_t es un proceso no estacionario, pero su diferenciación de orden d, $\Delta^d Y_t$, es un proceso $ARMA(p-d,q)$ estacionario. Y_t es también conocido como un proceso $ARIMA(p,d,q)$.*

Tras definir los procesos $ARIMA(p,d,q)$, estamos en disposición de poder emplearlos, y para ello el siguiente paso será, dada una serie temporal, identificar que tipo de proceso es la que la produce, lo cual será presentado en la siguiente sección.

2.2.5. Identificación de los procesos ARIMA(p,d,q)

Durante toda esta sección se ha desarrollado una visión estocástica de las series temporales. Hemos visto, que una serie temporal se puede ver como una realización muestral de un proceso estocástico estacionario, pero, ¿ cómo podemos identificar que proceso estocástico genera la serie temporal sobre la que estamos trabajando? Para resolver esta pregunta para el caso de procesos $ARIMA(p,d,q)$, se desarrolló una metodología conocida como la metodología Box-Jenkins (Box and Jenkins, 1976).

La metodología Box-Jenkins se basa en un total de cinco puntos a seguir que se resumen en

(1) **Identificación.** El primer paso será identificar el proceso $ARIMA(p,d,q)$ que pueda generar nuestra serie temporal. Esto es equivalente a encontrar el valor de (p,d,q) que nos de una serie temporal lo más similar a la original.

(2) **Estimación.** Una vez identificados los valores para (p,d,q), tenemos el posible proceso $ARIMA(p,d,q)$ que ha generado nuestra serie temporal, así que el siguiente paso será el de estimar los valores $\widehat{\phi}_i$, $\widehat{\theta}_j$.

(3) **Validación.** Tras la estimación de los parámetros, tenemos que validar el modelo a través de sus residuos.

(4) **Predicción.** Para finalizar, después de estimar y validar el modelo, podemos predecir los valores futuros de la serie temporal.

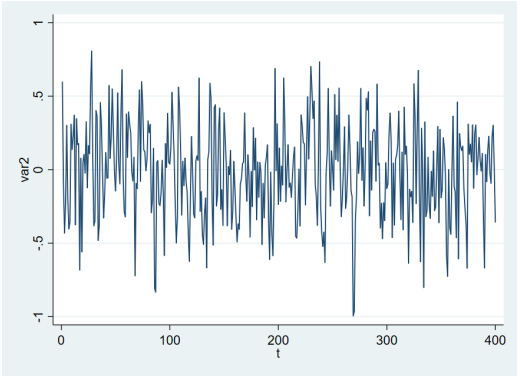
A continuación vamos a ahondar en cada uno de los pasos de la metodología Box-Jenkins.

(1) Identificación.

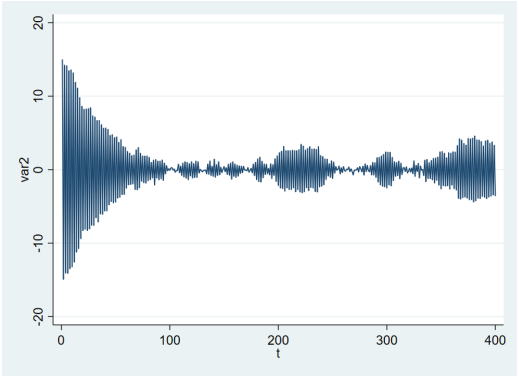
El primer paso de la metodología Box-Jenkins es identificar el proceso estocástico que genera nuestra serie temporal. Queremos que nuestro proceso estocástico sea al menos estacionario en sentido amplio, es decir, desearíamos que nuestra serie temporal tenga varianza y esperanza constante.

Para el análisis de la varianza normalmente se usa un método gráfico. Representamos la serie temporal y observamos la dispersión en torno a la media, si la dispersión se mantiene constante, la varianza de la serie temporal es constante en el tiempo. Sin embargo, si tenemos un cambio en la dispersión de la serie temporal, podemos deducir que la varianza no es constante. Otro método numérico que se puede emplear es separar la serie temporal en tramos de igual longitud y calcular la varianza muestral, si obtenemos valores similares en cada tramo, podemos asumir que la varianza de la serie es constante. En la figura 8 podemos observar el ejemplo de una serie estacionaria en varianza y no estacionaria en varianza.

En el caso de que la serie temporal no sea estacionaria en varianza, se suelen emplear ciertas transformaciones estabilizadoras que ayuden a obtener una serie temporal estacionaria en varianza. Las más habituales son las conocidas como transformaciones Box-Cox (Box and Cox, 1964), entre las que destacan las raíces cuadradas, las fuciones inversas o las transformaciones logarítmicas, siendo estas últimas el caso más habitual en las series econométricas debido a que no suelen contener valores negativos ni nulos.



(a) Ejemplo de serie temporal estacionaria en varianza.



(b) Ejemplo de serie temporal no estacionaria en varianza.

Figura 8: Representación gráfica de series temporales estacionaria y no estacionaria en varianza.

En cuanto a la estacionariedad en media consiste en comprobar si la serie temporal tiene raíces unitarias. Como se ha descrito en el apartado anterior, un proceso $ARIMA(p,d,q)$ no estacionario se puede convertir en un proceso estacionario mediante la diferenciación. Por lo tanto, bastará saber cuántas raíces unitarias tiene nuestra serie temporal para diferenciarla y obtener una serie generada por un proceso estacionario, estimando así el valor d del proceso $ARIMA(p,d,q)$ que buscamos.

Normalmente para detectar las raíces unitarias se utilizan diferentes test estadísticos entre los que destacan algunos como el test de Dickey-Fuller (Dickey and Fuller, 1979), el test de Dickey-Fuller aumentado (ADF), el test de Phillips-Perron(Phillips and Perron, 1988) o el test Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (Kwiatkowski et al., 1992), donde cada uno tiene su propia hipótesis nula para concluir si tenemos o no raíces unitarias en nuestra serie temporal. En este trabajo se utiliza la versión aumentada de Dickey-Fuller, cuya hipótesis nula es que la serie de estudio contiene una raíz unitaria y por ende no proviene de un proceso estacionario. La descripción del test no es compleja, pero se considera que está fuera de la temática del TFM, por lo que se puede encontrar en (González Casimiro, 2009; Brockwell and Davis, 2006).

Así pues, ya podemos identificar el término d del candidato a proceso $ARIMA(p,d,q)$ que ha generado la serie de estudio. Para calcular p y q debemos introducir en primer lugar el término de función de correlación parcial.

Definición 2.14 *Sean Y_m, Y_{m+h} dos variables aleatorias de un proceso estocástico separadas en h periodos. Definimos el valor de autocorrelación parcial como la correlación lineal entre ambas variables, eliminando la dependencia lineal de Y_m con respecto a todas las variables $Y_{m+1},...,Y_{m+h-1}$.*

Dada la definición 2.14, podemos considerar los coeficientes de la función de autocorrelación parcial como los coeficientes de la regresión

$$Y_{m+h} = \delta + \alpha_1 Y_{m+h-1} + \alpha_2 Y_{m+h-2} + ... + \alpha_h Y_m,$$

por lo tanto, esto implica que si consideramos un proceso $AR(p)$, los coeficientes de la función de autocorrelación parcial serán cero a partir del término $p+1$.

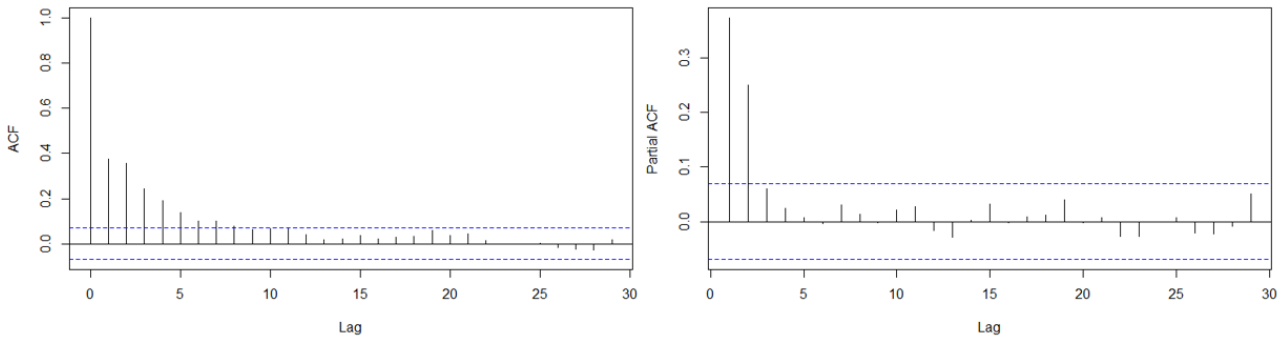
Tras esta observación ya sabemos la forma que tendrían las funciones de autocorrelación y autocorrelación parcial para el caso de los procesos $AR(p)$, $MA(q)$ y $ARMA(p,q)$, lo cual será de gran ayuda a la hora de identificar los posibles valores de p y q. En el cuadro 1 se recoge un resumen del comportamiento de las funciones de autocorrelación (FAC) y autocorrelación parcial (FACP).

PROCESO	FAC	FACP
$AR(p)$	Decrece rápidamente sin anularse	0 a partir del valor $p+1$
$MA(q)$	0 a partir del valor $q+1$	Decrece rápidamente sin anularse
$ARMA(p,q)$	Decrece rápidamente sin anularse	Decrece rápidamente sin anularse

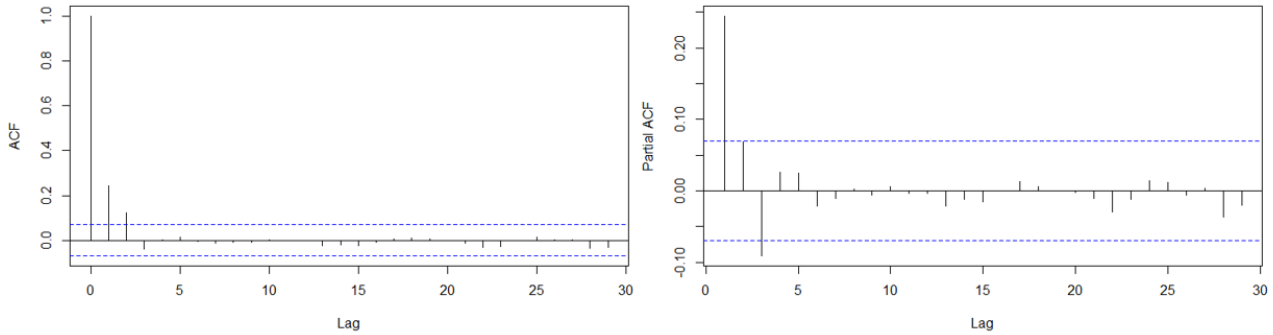
Cuadro 1: Comportamiento de las funciones de autocorrelación (FAC) y autocorrelación parcial (FACP).

En la introducción vimos como se puede estimar la función de autocorrelación a través de la serie temporal, y de la misma forma, podemos estimar la función de autocorrelación parcial para identificar el proceso $ARMA(p,q)$ que estamos buscando. La representación de ambas funciones estimadas se denomina correlograma y suele estar implementado en los diferentes programas de software. En la figura vemos representación de correlogramas para distintos procesos $ARMA(p,q)$.

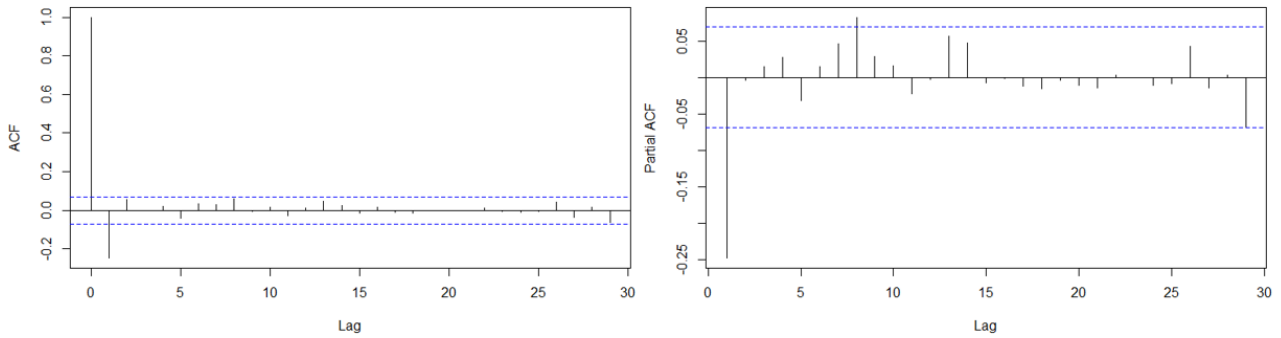
La identificación a través de correlogramas no es una ciencia exacta, teóricamente podríamos tomar como valores p y q los últimos valores significativamente no nulos de las funciones estimadas $\widehat{FACP}$ y $\widehat{FAC}$ respectivamente. Sin embargo, podemos ver como en la figura 9c tenemos el correlograma de un proceso estocástico $ARMA(1,1)$ generado por un software, sin embargo podríamos también identificarlo como un modelo $ARMA(1,2)$. Por tanto, para elegir el modelo, se suele observar el correlograma y obtener información de los posibles



(a) $AR(2)$.



(b) $MA(3)$.



(c) $ARMA(1,1)$.

Figura 9: Correlogramas de diferentes procesos $ARMA(p,q)$.

modelos que podrían haber generado nuestra serie temporal, para posteriormente utilizar algún criterio que nos permita elegir un modelo concreto. Los criterios más empleados son el criterio de información Bayesiano (BIC) (Schwarz, 1978), el criterio de Akaike (AIC) (Akaike, 1974) y el criterio de Akaike corregido (AICC), que penaliza la sobreparametrización. Matemáticamente se calculan

- $BIC = -2log(L) + klog(T)$,
- $AIC = 2k - log(L)$,
- $AICC = AIC + \dfrac{2k^2 + 2k}{T - k - 1}$,

donde T es la longitud de la serie, k es la suma de los parámetros empleados, y L el valor máximo de la función de verosimilitud para ese modelo. Como todos los modelos incluyen

el término $-log(L)$, y normalmente se busca elegir el valor de los parámetros que maximice la función de verosimilitud, elegiremos el modelo que minimice el valor del criterio que estamos empleando.

(2) Estimación.

Una vez determinado el proceso $ARIMA(p,d,q)$ elegido para recrear nuestra serie temporal, es necesario estimar los valores ϕ_i, θ_j. Existen diferentes métodos de estimación de parámetros para los procesos $ARIMA(p,d,q)$, entre los que podemos destacar el algoritmo de Burg, el algoritmo de innovación o el algoritmo de Hannan-Rissanen. Este último, consistente en una estimación a priori de los parámetros de un proceso $AR(m)$ para un $m > \text{máx}\{p,q\}$ con el que poder generar una serie de residuos $\widehat{\epsilon_t} \sim RB(0,\sigma_\epsilon^2)$, para posteriormente hacer una estimación a posteriori de los parámetros ϕ_i, θ_j del proceso $ARMA(p,q)$ que estamos estudiando mediante mínimos cuadrados es el que hemos implementado en este trabajo. Los diferentes tipos de algoritmos de estimación a los que hacemos mención se pueden encontrar en el capítulo 5 de (Brockwell and Davis, 2006).

(3) Validación.

Con nuestro modelo identificado y estimado, el último paso antes de poder predecir valores futuros es validar el modelo. Para validar el modelo, deberemos comprobar dos pasos, el primero comprobar que los coeficientes $\widehat{\phi_i}$, $\widehat{\theta_j}$ estimados son estadísticamente significativos, y el segundo que el residuo generado por el modelo sea ruido blanco.

Para comprobar si los coeficientes $\widehat{\phi_i}$, $\widehat{\theta_j}$ son estadísticamente significativos, basta con realizar sobre cada uno de ellos un contraste de hipótesis del tipo

$$\begin{cases} H_0 : \beta = 0 \\ \\ H_1 : \beta \neq 0 \end{cases} \quad \text{Para} \quad \beta \in \{\widehat{\phi_i}\}_i \cup \{\widehat{\theta_j}\}_j.$$

Para este contraste normalmente se utiliza el estadístico $t \sim N(0,1)$.

En la validación del residuo del modelo, el primer paso es estimar este residuo $\widehat{\epsilon_t}$. Para ello, se aplica emplea el modelo para hacer una estimación de los valores de la serie temporal que estamos estudiando, y se calcula

$$\widehat{\epsilon_t} = Y_t - \widehat{Y_t} \quad \text{para cada} \quad t \in [p+1, T].$$

Ahora tendremos que hacer un primer contraste de hipótesis similar al descrito en el paso anterior para saber si $E[\widehat{\epsilon_t}] = 0$.

Una vez sabemos si la esperanza del residuo es significativamente cero, tenemos que ver si es estacionario en varianza. Para conseguirlo, lo más habitual es comprobar la gráfica de representación del residuo y comprobar si la media de las desviaciones se mantiene constante sobre cero. De ser así, podemos asumir que el residuo tiene varianza constante.

Por último, tendremos que comprobar si el residuo está producido por un proceso incorrelado, es decir, si la covarianza entre dos términos cualesquiera de la serie es nula. Para hacerlo de forma sencilla, podemos utilizar la función de autocorrelación y de autocorrelación parcial, ya que el proceso estocástico será incorrelado si y sólo si estas dos funciones tienen valores estadísticamente nulos, lo que implica que $\widehat{\epsilon_t}$ será ruido blanco si sus funciones de autocorrelación y autocorrelación parcial tienen valores por debajo de $\frac{1,96}{\sqrt{T}}$, siendo T la longitud de la serie.

(4) Validación. El último paso, una vez identificado y estimado el proceso $ARIMA(p,d,q)$

que genera una serie Y_t diferenciado d veces, y tras ser validado, podremos utilizar el modelo para estimar valores futuros de la serie temporal, de forma que, si tenemos una serie con T elementos, podemos estimar

$$\widehat{Y}_{T+1} = \sum_{i=1}^{p} \widehat{\phi}_i Y_{T-i} + \sum_{j=1}^{q} \widehat{\theta}_j \widehat{\epsilon}_{T-j} + \epsilon_{T+1}.$$

2.3. Modelos ARIMA-SVR

Tras el desarrollo mostrado a lo largo de toda la sección anterior de los procesos $ARIMA(p,d,q)$, daremos el salto de los modelos lineales a los modelos no lineales. En esta sección desarrollaremos los modelos ARIMA-SVR, un modelo híbrido que combina el modelo lineal que nos aportan los procesos $ARIMA(p,d,q)$ sobre la variable objeto de observación, con una corrección del error cometido por el sistema a través de las máquinas de vector soporte de regresión (SVR).

Por lo tanto, en esta sección comenzaremos describiendo los modelos SVR, para posteriormente describir el modelo híbrido que usaremos en este trabajo. Las definiciones y resultados de esta sección se han extraido fundamentalmente de (Du and Swamy, 2013; Velásquez et al., 2010; Sujjaviriyasup and Pitiruek, 2013; Wang et al., 2018).

2.3.1. Máquinas de vector soporte de regresión (SVR)

Los modelos de máquinas de vector soporte de regresión (SVR) son un tipo de Machine Learning proveniente de una adaptación de las máquinas de vectores de soporte (SVM), algoritmo empleado normalmente para problemas de clasificación en problemas no separables mediante algún hiperplano. En nuestro trabajo, vamos a emplear los modelos SVR como un modelo de regresión que nos servirá para ajustar el error cometido por modelos lineales.

Consideremos un conjunto

$$\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R} \quad \text{con} \quad i = 1, ..., N\},$$

tal que cada x_i es un conjunto de variables predictoras e y_i es su correspondiente salida. Con los modelos SVR buscamos conseguir una estimación no linal de la forma

$$\widehat{y_i} = w' \Psi(x_i) + b, \tag{40}$$

donde $w \in \mathbb{R}^n$ es un vector de coeficientes, conocido como vector de pesos, $b \in \mathbb{R}$ es una constante, y la función Ψ una función no lineal desconocida. Por lo tanto, para definir nuestro modelo tendremos que estimar los valores de w y de b, para lo cual se minimiza la función de perdida definida como

$$\text{mín} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \|y_i - \widehat{y_i}\|_\epsilon \tag{41}$$

donde $\|x\|_\epsilon = \text{máx}\{0, \|x\| - \epsilon\}$ para un $\epsilon > 0$ y $C > 0$. Esta definición de norma es la que adapta el concepto de máquina de vector soporte al caso de la regresión. Al igual que ocurre en los problemas de clasificación donde se usan modelos SVM, y en contra de los modelos de regresión habituales, no se busca minimizar una función de pérdida en función del error, si no que se permite un margen alrededor del valor real de la variable observada para el

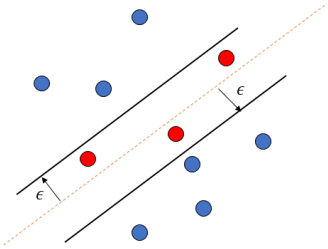


Figura 10: Tubo de ϵ de modelo SVR

cual el error se considera cero. Esto queda representado en la figura 10, donde todas las observaciones en rojo se consideran con un error cero al estar dentro del margen establecido por el valor de ϵ.

Si a la ecuación (41) le añadimos variables de holgura $\xi_i, \xi_i^* \geq 0$ y tomamos en cuenta el margen que estamos tomando para el error, podemos formular el problema de minimización como

$$\min_{w,\xi,\xi^*} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \quad \text{subjet to} \quad \begin{cases} \widehat{y_i} - y_i \leq \epsilon + \xi_i, \\ y_i - \widehat{y_i} \leq \epsilon + \xi_i^*, \end{cases} \tag{42}$$

de forma que cuando el error es menor que ϵ, las variables $\xi_i, \xi_i^* = 0$. La ecuación (42) nos deja un problema de programación cuadrática, al cual podemos aplicar la técnica de los multiplicadores de Lagranje, para finalmente llegar a la expresión a minimizar

$$\min_{\alpha,\alpha^*} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)k(x_i, x_j) + \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)y_i + \sum_{i=1}^{N}\left(\alpha_i + \alpha_i^*\right)\epsilon, \tag{43}$$

sujeto a las condiciones

$$\begin{cases} \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right) = 0, \\ 0 \leq \alpha_i, \alpha_i^* \leq C, \end{cases} \tag{44}$$

La función k que aparece en (43) se conoce como función Kernel y cumple la condición $k(x_i, x_j) = \Psi(x_i)'\Psi(x_j)$. Por otro lado, los valores α_i, α_i^* son los multiplicadores de Lagrange. Finalmente, (43)-(44) queda formulado como un problema de programación cuadrática, cuya solución se puede sustituir en (40) dando lugar a la aproximación final que buscamos

$$\widehat{y_i} = \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)k(x_j, x_i) + b, \tag{45}$$

los valores $(\alpha_j - \alpha_j^*) \neq 0$ son los llamados vectores de soporte.

2.3.2. Modelo híbrido ARIMA-SVR.

Una vez descritos los modelos SVR vamos a desarrollar el concepto del modelo híbrido entre los procesos $ARIMA(p,d,q)$ de carácter lineal, y los modelos SVR de carácter no lineal. La idea principal sobre la que gira este concepto es la de usar un factor corrector sobre el error de predicción que provocan los modelos $ARIMA(p,d,q)$.

Recordemos, que el error de un proceso $ARIMA(p,d,q)$ debe estar incorrelado, pero eso sólo implica que no habrá relaciones lineales entre el error cometido en diferentes etapas, sin embargo, esto no evita que podamos buscar otro tipo de relaciones en él que nos ayude a predecir que error debería cometer nuestro sistema, ayudándonos por tanto a corregirlo.

Así pues, si tenemos una serie temporal Y_t, diremos que el modelo ARIMA-SVR sigue la expresión

$$\widehat{Y}_{t+1} = \widehat{L}_{t+1} + \widehat{a}_{t+1} + \epsilon_{t+1}, \tag{46}$$

donde $\widehat{L}_{t+1}$ es la predicción lineal realizada por el proceso $ARIMA(p,d,q)$ sobre la serie temporal, y $\widehat{a}_{t+1}$ es la estimación realizada sobre el error por el modelo SVR.

Por lo tanto, para estimar y entrenar este algoritmo, tendremos en primer lugar que estimar el proceso $ARIMA(p,d,q)$ idóneo para nuestra serie temporal, tal y como se desarrolló en la sección de identificación de modelos, calcular el residuo del mismo, y entrenar con él la máquina de vector soporte de regresión, probando diferentes números de retardos para el error, es decir, probando con diferentes longitudes en el vector de entrada para quedarnos con la que minimice el error. Una vez está el proceso $ARIMA(p,d,q)$ estimado y el modelo SVR entrenado, podemos usar la expresión (46) para predecir valores futuros de nuestra variable observada.

3. Descripción de las bases de datos

Una vez descritos los algoritmos $ARIMA$ y $ARIMA-SVR$, pasamos a describir las bases de datos que utilizaremos para aplicar sendos modelos como ejemplo práctico.

Vamos a utilizar dos series temporales distintas, una primera serie del ámbito económico que recoge el precio diario del megavatio/hora en España, y una segunda enmarcada en el sector sanitario, la cual recoge el número de nuevos casos diarios de infectados por SARS-COV-2 en Italia.

En ambos casos emplearemos todos los datos disponibles hasta el 31 de diciembre de 2021 para construir nuestros modelos, empleando enero del año 2022 como espacio temporal en el que ejecutar las predicciones y comparar ambos modelos. Las predicciones tendrán un horizonte temporal a corto plazo (24 horas), y en las gráficas, cada punto será una predicción puntual realizada con los datos actualizados hasta el día anterior.

La elección de una serie temporal en el ámbito de la economía y otra en un sector tan separado de este como es el sanitario no es arbitrario. Normalmente, los procesos ARIMA aparecen asociado al mundo de la economía y la econometría, y con este segundo ejemplo queremos demostrar que se pueden extrapolar y emplear en otros campos, siempre que las series temporales estén tratadas de forma adecuada. Tras esto, pasamos a describir cada una de las series temporales un poco más en profundidad.

3.1. Serie temporal de precio del megavatio/hora en España

En febrero del año 2022 estalló en el norte de Europa un conflicto bélico entre Rusia y Ucrania. Como consecuencia de este hecho, toda Europa y gran parte del mundo se sumieron en una gran crisis energética que provocó una subida de producción de energía a unos niveles inéditos en la historia reciente del mundo, repercutiendo directamente en el precio del consumidor.

En España, este tema está más candente que nunca, ya que la subida en el precio del megavatio/hora ha llevado al gobierno a buscar una solución con la Unión Europea en lo que se ha conocido como excepción ibérica. Dados estos acontecimientos, se ha considerado la serie temporal del precio del megavatio/hora en España una serie de interés en el ámbito económico.

Así pues, disponemos de una serie temporal que contiene datos en el periodo comprendido entre el 01 de enero de 2021 y el 31 de agosto de 2022. Como se describe en el previo, hemos utilizado los datos recogidos entre las fechas $01-01-2021$ y $31-07-2022$ para construir nuestro modelo, utilizando el último mes como periodo para validar y comparar los resultados obtenidos. La serie temporal es pública y los datos se pueden encontrar en https://www.epdata.es.

3.2. Serie temporal de casos de SARS-COV-2

A principios del año 2020 el mundo se vio paralizado por la llegada de un nuevo virus conocido como SARS-COV-2, el cual propició una pandemia mundial cuyas consecuencias han sido las más devastadoras de la historia reciente. Por ello, desde el inicio de esta crisis sanitaria se necesitó del desarrollo de modelos predictivos sobre el número de casos que ayudaran en la planificación dentro del sistema sanitario. Dada la atención que ha captado desde su aparición, hemos elegido la serie temporal de nuevos infectados por SARS-COV-2 en Italia como serie de interés fuera del ámbito económico.

La elección de Italia para centrar el foco de nuestro estudio no es casual, ya que fue el primer país en el que se detectó la nueva enfermedad en Europa, y por ello, disponemos de

una serie temporal más larga. Por tanto, disponemos de una serie temporal con datos desde el 31 de enero de 2020 hasta el 30 de enero de 2022, disponiendo así de dos años completos de información ininterrumpida.

Como ha sido mencionado en el previo, utilizaremos los datos desde $31-01-2020$ hasta $31-12-2021$ para construir nuestro modelo, y el último mes de la serie para validar y comparar los resultados. Los datos de la serie temporal son públicos y podemos encontrarlo en la dirección https://ourworldindata.org.

4. Ejemplos prácticos

En esta sección vamos a recoger la aplicación de los modelos sobre las series temporales descritas en la sección 3. La sección se distribuirá en un primer apartado en el que explicaremos los criterios de comparación que vamos a emplear, para continuar con un epígrafe en el que construiremos paso a paso los dos modelos en cada caso, y terminaremos comparando los resultados obtenidos en base a los criterios descritos.

4.1. Criterios de comparación

Para comparar los resultados obtenidos por los procesos ARIMA y ARIMA-SVR vamos a utilizar una serie de criterios descriptivos ampliamente utilizados en el sector de los modelos predictivos y que están basados en el error cometido por el sistema.

Concretamente, vamos a utilizar el error medio absoluto (MAE), el error cuadrático medio (RMSE), el porcentaje de error absoluto simétrico (sMAPE), el error de escala absoluta de media (MASE) y el coeficiente de determinación (R^2). Para definir estos criterios matemáticamente, vamos a considerar y_t una serie temporal de longitud T y $\widehat{y_t}$ la estimación de y_t. Podemos describir los diferentes criterios descriptivos como

- $MAE = \dfrac{\sum_{t=1}^{T} |y_t - \widehat{y_t}|}{T}$.
- $RMSE = \sqrt{\dfrac{\sum_{t=1}^{T} (y_t - \widehat{y_t})^2}{T}}$.
- $sMAPE = \dfrac{1}{T} \sum_{t=1}^{T} \dfrac{|y_t - \widehat{y_t}|}{\left(|y_t| + |\widehat{t_t}|\right)/2}$.
- $MASE = \dfrac{\frac{1}{T} \sum_{t=1}^{T} |y_t - \widehat{y_t}|}{\frac{1}{T-1} \sum_{t=2}^{T} |y_t - y_{t-1}|}$.
- $R^2 = 1 - \dfrac{VAR(y_t - \widehat{y_t}}{VAR(y_t)}$.

Una vez definidos los criterios que utilizaremos, vamos a pasar a estimar los modelos.

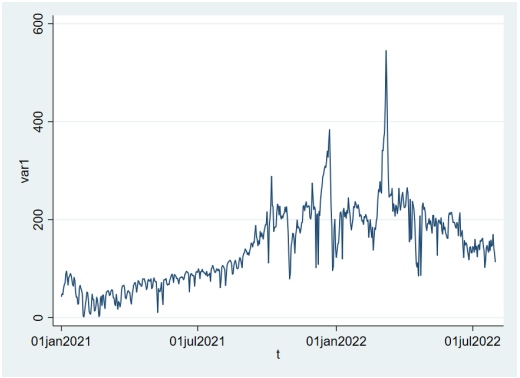
4.2. Construcción de los modelos

En este apartado vamos a construir los procesos ARIMA para cada una de las series temporales descritas siguiendo la metodología descrita en el epígrafe 2.2.5. Posteriormente ajustaremos el modelo ARIMA-SVR, calculando el número de entradas que debe tener la máquina de vector soporte.

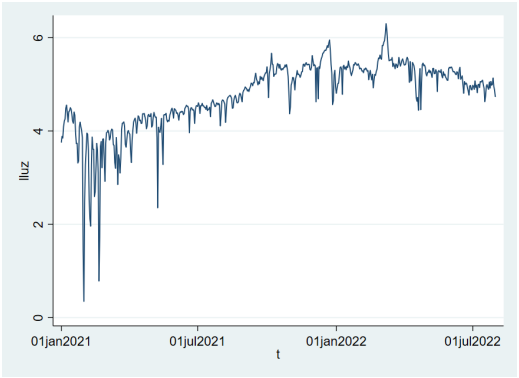
Para la identificación del proceso $ARIMA(p,d,q)$ utilizaremos el software Stata. Por otro lado, debido a su versatilidad para el trato de series temporales, tanto para construir el modelo $ARIMA(p,d,q)-SVR$, como para calcular las predicciones, utilizaremos el software R. Tanto el archivo do con los pasos para la identificación de los modelos, como las funciones implementadas en R para calcular las predicciones están añadidas en el anexo.

• Serie precio megavatio/hora en España

El primer paso para construir un proceso $ARIMA(p,d,q)$ es la identificación, es decir, tendremos que comprobar si la serie del precio del megavatio/hora en España (en adelante L_t) es estacionaria o no. En la figura 11a está representada la serie temporal L_T, y como podemos ver, la serie no tiene una media muestral constante y su varianza cambia con el paso del tiempo. Lo habitual en estos casos es tomar una de las transformaciones Box-Cox, y como es habitual en variables económicas, vamos a tomar la transformación logarítmica representada en la figura 11b. Podemos observar como se ha estabilizado la varianza de la serie, pero aún debemos de comprobar si es estacionaria o no.



(a) Serie temporal L_t.



(b) Serie temporal $log(L_t)$.

Figura 11: Serie temporal del precio megavatio/hora en España antes y después de la transformación logarítmica.

Realicemos en primer lugar una inspección visual al correlograma de la serie. En la figura 12 vemos el correlograma de la serie $log(L_t)$, el cual muestra un descenso muy lento en la función de autocorrelación con valores cercanos a la unidad. De forma teórica, esto indica la existencia de una raíz unitaria, por lo que tendremos que pasar a estudiar la serie diferenciada $\Delta log(L_t)$.

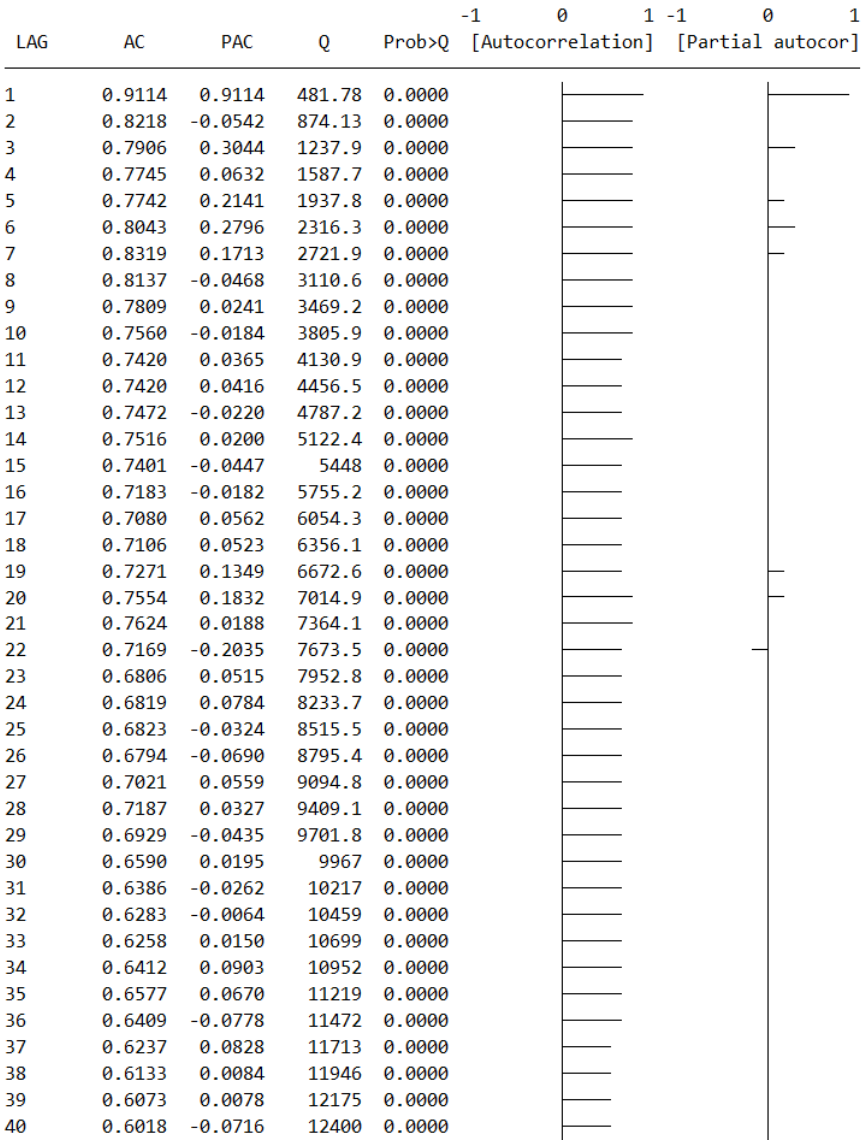
Tras la diferenciación de la serie, podemos comprobar que su comportamiento ha cambiado, visualmente en la figura 13 podemos observar como los valores se mueven en torno a su valor medio situado en el cero manteniendo las desviaciones con respecto al mismo constantes. Aunque la gráfica de la serie puede hacernos pensar que la ya es estacionaria, debemos realizar un test ADF para contrastarlo.

Stata por defecto no nos da el número de retardos óptimos que debemos añadir en el test, pero podemos probar modelando la serie temporal como distintos modelos $AR(p)$, eligiendo el retardo que menor valor nos de en el criterio AIC. En el cuadro 2 podemos ver como el resultado del test ADF nos permite rechazar la hipótesis nula de la existencia de raíz unitaria, permitiéndonos asumir que la serie $\Delta log(L_t)$ es estacionaria.

	Test Statistic	1%	5%	10%
Z(t)	-10.77	-3.43	-2.86	-2.57

Cuadro 2: Resultado test ADF en Stata para serie $\Delta log(L_t)$

En este momento hemos identificado $d = 1$ y estamos en disposición de estimar el valor de p y q para el proceso $ARIMA(p,1,q)$. Si observamos el correlograma de $\Delta log(L_t)$ que aparece en la figura 14, podemos notar que existen valores significativos hasta el valor 7 en la función FAC, y en torno al 6 de la FACP. Por lo tanto, vamos a utilizar el criterio AIC y el criterio BIC para determinarnos por el modelo que tenga el menor valor de estos criterios.



LAG	AC	PAC	Q	Prob>Q
1	0.9114	0.9114	481.78	0.0000
2	0.8218	-0.0542	874.13	0.0000
3	0.7906	0.3044	1237.9	0.0000
4	0.7745	0.0632	1587.7	0.0000
5	0.7742	0.2141	1937.8	0.0000
6	0.8043	0.2796	2316.3	0.0000
7	0.8319	0.1713	2721.9	0.0000
8	0.8137	-0.0468	3110.6	0.0000
9	0.7809	0.0241	3469.2	0.0000
10	0.7560	-0.0184	3805.9	0.0000
11	0.7420	0.0365	4130.9	0.0000
12	0.7420	0.0416	4456.5	0.0000
13	0.7472	-0.0220	4787.2	0.0000
14	0.7516	0.0200	5122.4	0.0000
15	0.7401	-0.0447	5448	0.0000
16	0.7183	-0.0182	5755.2	0.0000
17	0.7080	0.0562	6054.3	0.0000
18	0.7106	0.0523	6356.1	0.0000
19	0.7271	0.1349	6672.6	0.0000
20	0.7554	0.1832	7014.9	0.0000
21	0.7624	0.0188	7364.1	0.0000
22	0.7169	-0.2035	7673.5	0.0000
23	0.6806	0.0515	7952.8	0.0000
24	0.6819	0.0784	8233.7	0.0000
25	0.6823	-0.0324	8515.5	0.0000
26	0.6794	-0.0690	8795.4	0.0000
27	0.7021	0.0559	9094.8	0.0000
28	0.7187	0.0327	9409.1	0.0000
29	0.6929	-0.0435	9701.8	0.0000
30	0.6590	0.0195	9967	0.0000
31	0.6386	-0.0262	10217	0.0000
32	0.6283	-0.0064	10459	0.0000
33	0.6258	0.0150	10699	0.0000
34	0.6412	0.0903	10952	0.0000
35	0.6577	0.0670	11219	0.0000
36	0.6409	-0.0778	11472	0.0000
37	0.6237	0.0828	11713	0.0000
38	0.6133	0.0084	11946	0.0000
39	0.6073	0.0078	12175	0.0000
40	0.6018	-0.0716	12400	0.0000

Figura 12: Representación del correlograma de la serie $log(L_t)$.

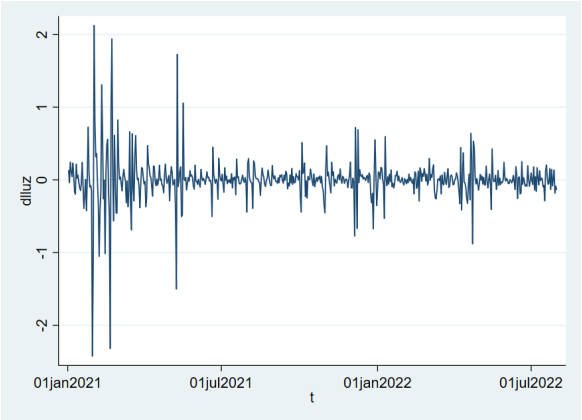
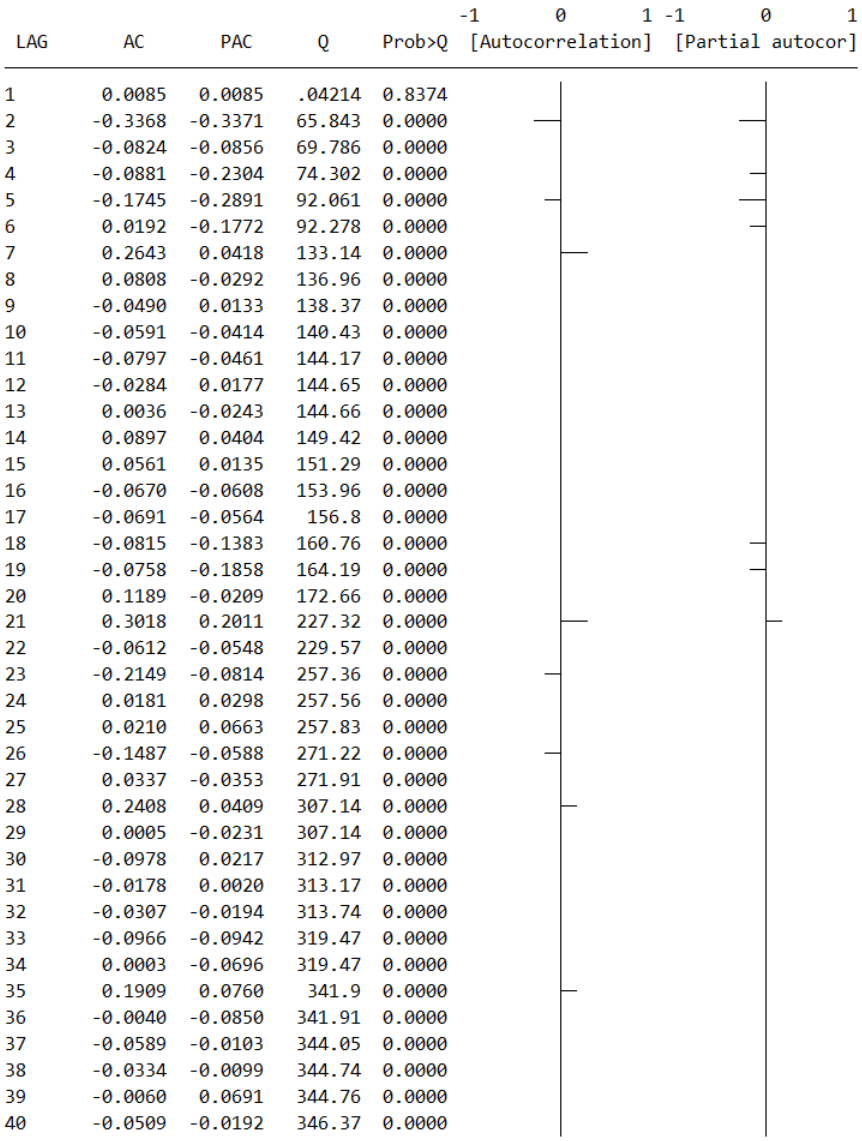


Figura 13: Serie temporal $\Delta log(L_t)$.

En el cuadro 3 se muestra el resultado de la estimación de un proceso ARIMA para distintos valores de p y q, siendo elegido finalmente el proceso estocástico $ARIMA(4,1,6)$.

Presentamos la estimación con Stata en el cuadro 4. Podemos observar como el valor L5



LAG	AC	PAC	Q	Prob>Q
1	0.0085	0.0085	.04214	0.8374
2	-0.3368	-0.3371	65.843	0.0000
3	-0.0824	-0.0856	69.786	0.0000
4	-0.0881	-0.2304	74.302	0.0000
5	-0.1745	-0.2891	92.061	0.0000
6	0.0192	-0.1772	92.278	0.0000
7	0.2643	0.0418	133.14	0.0000
8	0.0808	-0.0292	136.96	0.0000
9	-0.0490	0.0133	138.37	0.0000
10	-0.0591	-0.0414	140.43	0.0000
11	-0.0797	-0.0461	144.17	0.0000
12	-0.0284	0.0177	144.65	0.0000
13	0.0036	-0.0243	144.66	0.0000
14	0.0897	0.0404	149.42	0.0000
15	0.0561	0.0135	151.29	0.0000
16	-0.0670	-0.0608	153.96	0.0000
17	-0.0691	-0.0564	156.8	0.0000
18	-0.0815	-0.1383	160.76	0.0000
19	-0.0758	-0.1858	164.19	0.0000
20	0.1189	-0.0209	172.66	0.0000
21	0.3018	0.2011	227.32	0.0000
22	-0.0612	-0.0548	229.57	0.0000
23	-0.2149	-0.0814	257.36	0.0000
24	0.0181	0.0298	257.56	0.0000
25	0.0210	0.0663	257.83	0.0000
26	-0.1487	-0.0588	271.22	0.0000
27	0.0337	-0.0353	271.91	0.0000
28	0.2408	0.0409	307.14	0.0000
29	0.0005	-0.0231	307.14	0.0000
30	-0.0978	0.0217	312.97	0.0000
31	-0.0178	0.0020	313.17	0.0000
32	-0.0307	-0.0194	313.74	0.0000
33	-0.0966	-0.0942	319.47	0.0000
34	0.0003	-0.0696	319.47	0.0000
35	0.1909	0.0760	341.9	0.0000
36	-0.0040	-0.0850	341.91	0.0000
37	-0.0589	-0.0103	344.05	0.0000
38	-0.0334	-0.0099	344.74	0.0000
39	-0.0060	0.0691	344.76	0.0000
40	-0.0509	-0.0192	346.37	0.0000

Figura 14: Representación del correlograma de la serie $log(L_t)$.

MODELO	AIC	BIC
ARIMA(4,1,7)	92.566	144.840
ARIMA(5,1,7)	94.293	150.922
ARIMA(6,1,7)	100.677	161.662
ARIMA(4,1,6)	90.998	138.915

Cuadro 3: Valores de los criterios AIC y BIC para diferentes procesos ARIMA sobre la serie $\Delta log(L_t)$.

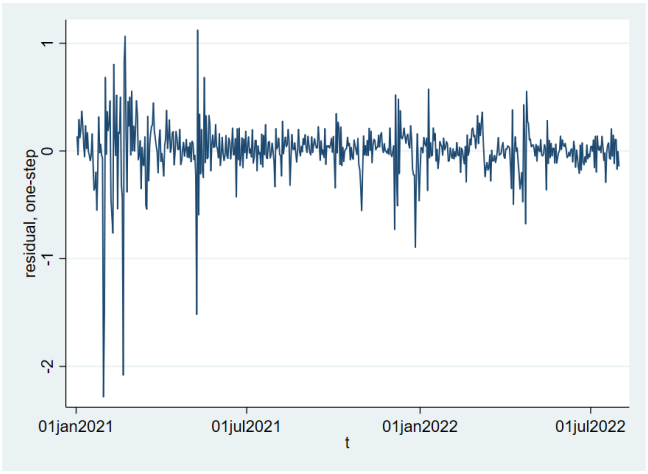
para la parte de medias móviles no es significativo, pero el resto si que son estadísticamente distintos de 0. Por ende, cuando calculemos en R el valor de la predicción, pasaremos ese valor como un valor no nulo.

El último paso que nos queda es la validación del modelo. A través del correlograma y la representación de los residuos podemos ver claramente como se trata de un ruido blanco. La representación tanto de lo la gráfica como del correlograma de los residuos queda ilustrada en la figura 15.

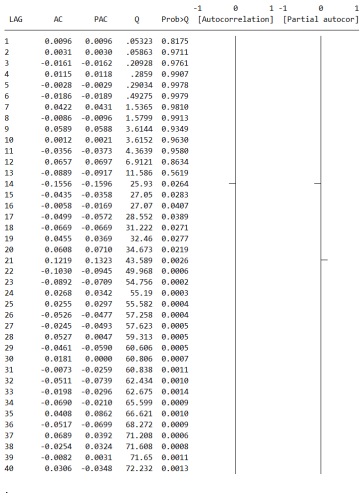
Tras la construcción del proceso $ARIMA(4,1,6)$ tenemos que estimar el modelo SVR que

D.lluz	Coefficient	std. err.	z	$P>\|z\|$	[95% conf.	interval]
ar						
L1.	0.7322399	0.0290629	25.19	0	0.6752776	0.7892022
L2.	-1.403937	0.0258169	-54.38	0	-1.454538	-1.353337
L3.	0.7624153	0.025975	29.35	0	0.7115053	0.8133253
L4.	-0.9239528	0.0276166	-33.46	0	-0.9780803	-0.8698252
ma						
L1.	-0.8908446	0.033534	-26.57	0	-0.9565702	-0.8251191
L2.	1.104977	0.0378447	29.2	0	1.030802	1.179151
L3.	-0.7809778	0.0477536	-16.35	0	-0.8745732	-0.6873824
L4.	0.4341131	0.0489993	8.86	0	0.3380763	0.5301499
L5.	0.0451572	0.037162	1.22	0.224	-0.027679	0.1179934
L6.	-0.3476224	0.0376882	-9.22	0	-0.4214898	-0.2737549
/sigma	0.2561251	0.0035463	72.22	0	0.2491745	0.2630758

Cuadro 4: Estimación del proceso $ARIMA(4,1,6)$ para $\Delta log(L_t)$ por Stata.



(a) Residuos del modelo $ARIMA(4,1,6)$.



(b) Correlograma de los residuos del modelo $ARIMA(4,1,6)$.

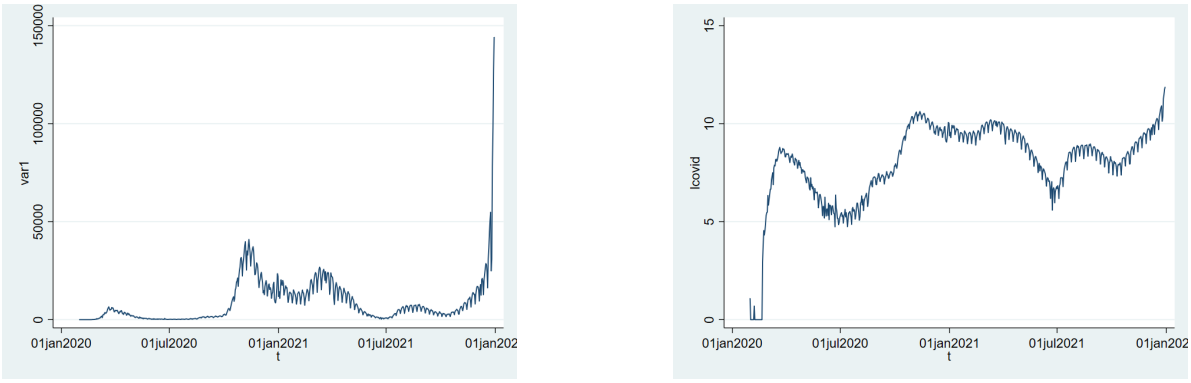
Figura 15: Representación de residuos para proceso $ARIMA(4,1,6)$ y su correspondiente correlograma.

servirá de corrector para el error del modelo. Para ello hemos separado el periodo empleado para estimar en dos partes, una primera con los datos del intervalo $01-01-2021$ hasta $30-06-2022$, y la segunda con los datos entre las fechas $01-07-2021$ hasta $31-07-2022$. Con la primera parte, calculamos el error cometido por el $ARIMA(4,1,6)$ y utilizamos la segunda para estimar el número de entradas que debemos darle a la máquina de vector soporte. Para ello, se prueba el modelo con vectores de entradas de dimensión en el intervalo $[1,25]$ en la parte SVR y elegimos la que nos devuelva menor error cuadrático medio. En este caso, se determina que los vectores de entrada para la máquina de vector soporte deben tener dimensión 7 y tendrá la forma $(\widehat{\epsilon}_{t-1},\widehat{\epsilon}_{t-2},...,\widehat{\epsilon}_{t-6})$ para cada realización en tiempo t.

• Serie temporal de casos de SARS-COV-2

Debemos repetir el proceso para la serie temporal de los casos de SARS-COV-2 en Italia, la cual denominaremos de ahora en adelante como S_t. Comenzaremos comprobando si la

serie es estacionaria. Al observar la gráfica, vemos que la serie no tiene la media ni la varianza constante, por lo que vamos a optar por transformarla mediante la transformación logaritmo. En este caso, tenemos que tener en cuenta que la serie tiene valores nulos, por lo que la transformación vendrá acompañada de una traslación, de forma que pasaremos a estudiar la serie $log(S_t+1)$. La representación tanto de la serie como de su transformada está en la figura 16.



(a) Serie temporal S_t.

(b) Serie temporal $log(S_t+1)$.

Figura 16: Serie temporal de de los casos de SARS-COV-2 en Italia antes y después de la transformación logarítmica.

Observando la figura 16b podemos ver que parece que nuevamente se ha estabilizado la varianza, pero sigue sin ser estacionario. Para confirmarlo, representamos su correlograma en la figura 17 y comprobamos como efectivamente la función de autocorrelación tiene un descenso muy bajo con valores cercanos a la unidad, por lo que tendremos la certeza de que tiene una raíz unitaria. Así pues, vamos a tomar la diferenciación de la serie, tomando $\Delta log(S_t+1)$.

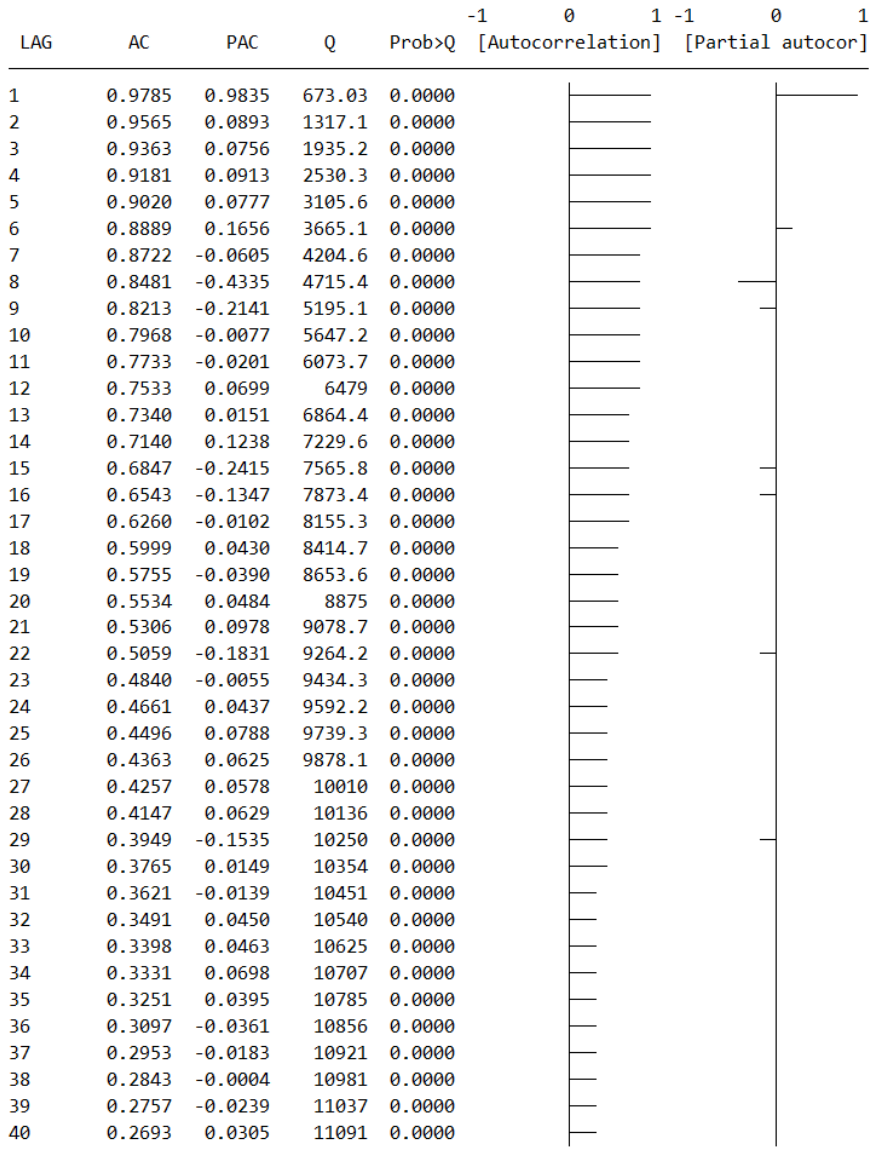
Una vez realizado la diferenciación, tenemos que comprobar si hemos eliminado la raíz unitaria. Con el correlograma y la gráfica de $\Delta log(S_t+1)$ recogidos en la figura 18, podemos intuir que ha sido eliminada, pero para tener un resultado analítico que nos lo abale, vamos a realizar el test ADF. Nuevamente, se han calculado los retardos a utilizar de manera recursiva y volvemos a contar con 8 retardos. En el cuadro 5 recogemos los resultados del test y comprobamos como efectivamente la raíz unitaria ha sido eliminada.

	Test Statistic	1%	5%	10%
Z(t)	-5.315	-3.43	-2.86	-2.57

Cuadro 5: Resultado test ADF en Stata para serie $\Delta log(S_t+1)$

Análogamente al ejemplo anterior, observando el correlograma de $\Delta log(S_t+1)$ podemos estimar los valores de p y q. Podemos observar, como los valores de la FAC son muy bajos para los primeros elementos, pero en el retardo 7 observamos un valor bastante significativo, por lo que parece claro que podemos fijar $q=7$. Ahora bien, si observamos en la figura 18b la función de autocorrelación parcial, comprobamos que los valores significativos aparecen en torno a los retardos 7 y 8, por lo que vamos a decidir empleando los criterios AIC y BIC. Dados los resultados mostrados en el cuadro 6, elegimos el proceso $ARIMA(8,1,7)$.

Por último, tan sólo nos queda validar el resultado. Con la estimación de los coeficientes en Stata presentada en el cuadro 7 podemos ver los términos que no son estadísticamente significativos, por lo que no los tendrems en cuenta al realizar la regresión. Por último, vemos en la figura 19 la representación del residuo y su correspondiente correlograma, a



LAG	AC	PAC	Q	Prob>Q
1	0.9785	0.9835	673.03	0.0000
2	0.9565	0.0893	1317.1	0.0000
3	0.9363	0.0756	1935.2	0.0000
4	0.9181	0.0913	2530.3	0.0000
5	0.9020	0.0777	3105.6	0.0000
6	0.8889	0.1656	3665.1	0.0000
7	0.8722	-0.0605	4204.6	0.0000
8	0.8481	-0.4335	4715.4	0.0000
9	0.8213	-0.2141	5195.1	0.0000
10	0.7968	-0.0077	5647.2	0.0000
11	0.7733	-0.0201	6073.7	0.0000
12	0.7533	0.0699	6479	0.0000
13	0.7340	0.0151	6864.4	0.0000
14	0.7140	0.1238	7229.6	0.0000
15	0.6847	-0.2415	7565.8	0.0000
16	0.6543	-0.1347	7873.4	0.0000
17	0.6260	-0.0102	8155.3	0.0000
18	0.5999	0.0430	8414.7	0.0000
19	0.5755	-0.0390	8653.6	0.0000
20	0.5534	0.0484	8875	0.0000
21	0.5306	0.0978	9078.7	0.0000
22	0.5059	-0.1831	9264.2	0.0000
23	0.4840	-0.0055	9434.3	0.0000
24	0.4661	0.0437	9592.2	0.0000
25	0.4496	0.0788	9739.3	0.0000
26	0.4363	0.0625	9878.1	0.0000
27	0.4257	0.0578	10010	0.0000
28	0.4147	0.0629	10136	0.0000
29	0.3949	-0.1535	10250	0.0000
30	0.3765	0.0149	10354	0.0000
31	0.3621	-0.0139	10451	0.0000
32	0.3491	0.0450	10540	0.0000
33	0.3398	0.0463	10625	0.0000
34	0.3331	0.0698	10707	0.0000
35	0.3251	0.0395	10785	0.0000
36	0.3097	-0.0361	10856	0.0000
37	0.2953	-0.0183	10921	0.0000
38	0.2843	-0.0004	10981	0.0000
39	0.2757	-0.0239	11037	0.0000
40	0.2693	0.0305	11091	0.0000

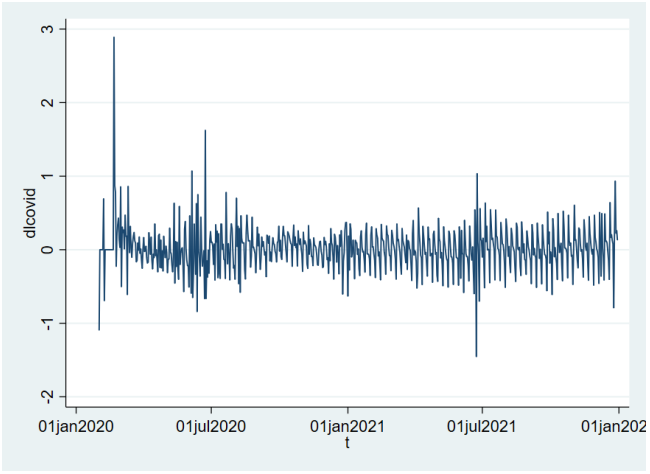
Figura 17: Representación del correlograma de la serie $log(S_t + 1)$.

MODELO	AIC	BIC
ARIMA(7,1,7)	27.935	96.180
ARIMA(8,1,7)	7.959	80.753

Cuadro 6: Valores de los criterios AIC y BIC para diferentes procesos ARIMA sobre la serie $\Delta log(S_t + 1)$.

través de los cuales podemos concluir que el residuo es ruido blanco, y en consecuencia, nuestro proceso $ARIMA(8,1,7)$ queda validado.

Una vez estimado el proceso $ARIMA(p,d,q)$, estimamos la dimensión de los vectores que servirán de entrada para la máquina vector soporte de regresión que corregirá el error. Siguiendo el mismo procedimiento, dividimos el dataset en dos partes, el primero conteniendo el intervalo de tiempo entre las fechas$31-01-2020$ y $30-11-2021$ para construir el modelo SVR, y el intervalo entre las fechas $01-12-2021$ y $31-12-2021$ para determinar la estructura definitiva de la parte no lineal del modelo. En este caso, determinamos que las entradas de la máquina de vector soporte de regresión tendrá dimensión 4.

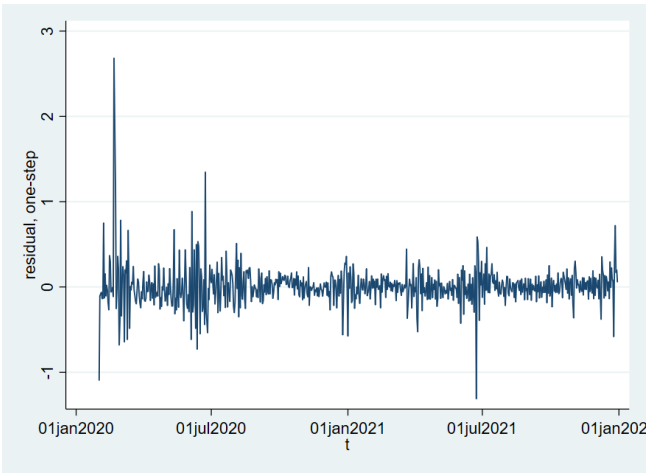


(a) Serie temporal $\Delta log(S_t + 1)$.

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial autocor]
1	-0.0939	-0.0939	6.1884	0.0129		
2	-0.0686	-0.0785	9.4977	0.0087		
3	-0.0762	-0.0925	13.581	0.0035		
4	-0.0484	-0.0759	15.231	0.0042		
5	-0.1249	-0.1616	26.253	0.0001		
6	0.1153	0.0675	35.656	0.0000		
7	0.4148	0.4381	157.48	0.0000		
8	0.0586	0.2102	159.92	0.0000		
9	-0.1001	-0.0010	167.04	0.0000		
10	-0.0398	0.0110	168.17	0.0000		
11	-0.1226	-0.0792	178.87	0.0000		
12	-0.0558	-0.0235	181.09	0.0000		
13	0.0168	-0.1315	181.29	0.0000		
14	0.4267	0.2367	311.55	0.0000		
15	0.0485	0.1219	313.24	0.0000		
16	-0.0894	-0.0069	318.97	0.0000		
17	-0.0991	-0.0604	326.03	0.0000		
18	-0.0456	0.0229	327.52	0.0000		
19	-0.1379	-0.0667	341.22	0.0000		
20	0.0284	-0.1134	341.8	0.0000		
21	0.3937	0.1758	453.81	0.0000		
22	-0.0092	-0.0006	453.87	0.0000		
23	-0.1189	-0.0476	464.12	0.0000		
24	-0.0821	-0.0816	469.01	0.0000		
25	-0.0966	-0.0645	475.79	0.0000		
26	-0.1162	-0.0590	485.63	0.0000		
27	0.0626	-0.0633	488.49	0.0000		
28	0.4116	0.1520	612.18	0.0000		
29	-0.0639	-0.0168	615.17	0.0000		
30	-0.0801	0.0119	619.87	0.0000		
31	-0.1138	-0.0470	629.37	0.0000		
32	-0.1060	-0.0481	637.63	0.0000		
33	-0.1253	-0.0714	649.17	0.0000		
34	0.0875	-0.0406	654.81	0.0000		
35	0.3351	0.0347	737.66	0.0000		
36	0.0024	0.0168	737.66	0.0000		
37	-0.1305	-0.0012	750.26	0.0000		
38	-0.0756	0.0224	754.5	0.0000		
39	-0.1400	-0.0323	769.05	0.0000		
40	-0.0999	-0.0212	776.47	0.0000		

(b) Correlograma de la serie temporal $\Delta log(S_t + 1)$.

Figura 18: Representación de la serie $\Delta log(S_t + 1)$ y su correspondiente correlograma.



(a) Residuos del modelo $ARIMA(8,1,7)$.

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial autocor]
1	0.0245	0.0245	.42176	0.5161		
2	-0.0327	-0.0333	1.1724	0.5564		
3	-0.0557	-0.0543	3.3579	0.3397		
4	0.0485	0.0505	5.0141	0.2859		
5	-0.0226	-0.0292	5.3747	0.3719		
6	0.0421	0.0445	6.6266	0.3568		
7	0.0269	0.0294	7.1406	0.4144		
8	-0.0221	-0.0269	7.4883	0.4850		
9	0.0453	0.0593	8.9425	0.4426		
10	0.0006	-0.0075	8.9427	0.5375		
11	-0.0128	-0.0118	9.0593	0.6164		
12	0.0264	0.0355	9.5557	0.6549		
13	-0.0553	-0.0707	11.739	0.5492		
14	0.0475	0.0600	13.353	0.4989		
15	0.0519	0.0468	15.286	0.4310		
16	0.0379	0.0239	16.315	0.4312		
17	-0.0220	0.0006	16.661	0.4775		
18	0.0481	0.0438	18.329	0.4342		
19	-0.0433	-0.0440	19.682	0.4140		
20	-0.0991	-0.0991	26.764	0.1420		
21	-0.0222	-0.0257	27.119	0.1669		
22	-0.0515	-0.0735	29.037	0.1438		
23	-0.0167	-0.0381	29.24	0.1725		
24	0.0184	-0.0063	29.486	0.2024		
25	0.0017	-0.0173	29.489	0.2440		
26	0.0128	0.0148	29.609	0.2841		
27	0.0015	-0.0052	29.61	0.3319		
28	0.0497	0.0652	31.416	0.2989		
29	-0.0982	-0.0583	38.473	0.1122		
30	0.0195	0.0030	38.751	0.1314		
31	-0.0093	-0.0033	38.814	0.1580		
32	-0.0397	-0.0627	39.97	0.1573		
33	0.0132	-0.0036	40.099	0.1844		
34	0.0023	-0.0255	40.103	0.2178		
35	-0.0147	-0.0186	40.261	0.2487		
36	-0.0351	-0.0311	41.174	0.2544		
37	0.0035	-0.0078	41.183	0.2925		
38	0.0264	0.0471	41.698	0.3131		
39	-0.0704	-0.0592	45.374	0.2235		
40	0.0143	0.0071	45.527	0.2529		

(b) Correlograma de los residuos del modelo $ARIMA(8,1,7)$.

Figura 19: Representación de residuos para proceso $ARIMA(8,1,7)$ y su correspondiente correlograma.

4.3. Comparación de los resultados

Después de la estimación de los modelos vamos a comparar los resultados obtenidos por ambos sistemas predictivos para comprobar si finalmente merece la pena aplicar el corrector no lineal a los procesos $ARIMA(p,d,q)$ clásicos.

Las predicciones tendrán estarán implementadas en R, pero utilizando los modelos arrojados por Stata en la sección anterior. Calcularemos 30 predicciones con un horizonte temporal de 24 horas, es decir, después de cada predicción actualizaremos al valor real para realizar la siguiente, pero sin cambiar ninguno de los valores de los parámetros calculados arriba.

Para la comparación, empleamos los criterios descriptivos descritos al comienzo de este capítulo, todos basados en el error cometido, por lo que elegiremos como mejor modelo aquel que tenga menor valor para todos los criterios salvo para el R^2, el cual recoge la varianza que es capaz de explicar cada modelo.

Los resultados obtenidos en nuestras predicciones han quedado registradas en el cuadro

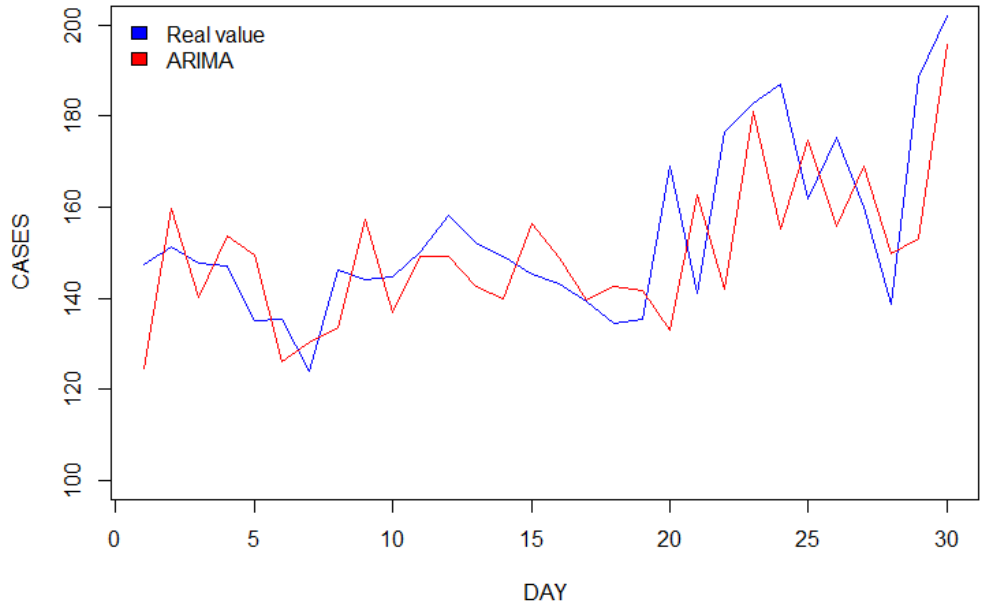
D.lluz	Coefficient	std. err.	z	$P > \|z\|$	[95% conf.	interval]
ar						
L1.	-0.3453385	0.0447634	-7.71	0	-0.4330732	-0.2576039
L2.	-0.0510202	0.0334462	-1.53	0.127	-0.1165736	0.0145332
L3.	-0.0579994	0.0331297	-1.75	0.08	-0.1229325	0.0069337
L4.	-0.0714646	0.0313745	-2.28	0.023	-0.1329573	-0.0099718
L5.	-0.0501308	0.0319287	-1.57	0.116	-0.1127099	0.0124482
L6.	-0.0434834	0.0329001	-1.32	0.186	-0.1079664	0.0209996
L7.	0.9267668	0.0312281	29.68	0	0.8655608	0.9879728
L8.	0.2852859	0.0296949	9.61	0	0.2270849	0.3434869
ma						
L1.	0.094343	.	.	.	.	.
L2.	0.0652895	0.0509511	1.28	0.2	-0.0345727	0.1651518
L3.	0.1391786	0.0817171	1.7	0.089	-0.020984	0.2993412
L4.	0.1955523	0.08658	2.26	0.024	0.0258585	0.3652461
L5.	0.0007338	0.0594805	0.01	0.99	-0.1158457	0.1173134
L6.	0.0857886	0.037695	2.28	0.023	0.0119076	0.1596695
L7.	-0.7996858	.	.	.	.	.
/sigma	0.2347382	0.006657	35.26	0.000	.2216908	.2477856

Cuadro 7: Estimación del proceso $ARIMA(8,1,7)$ para $\Delta log(S_t + 1)$ por Stata.

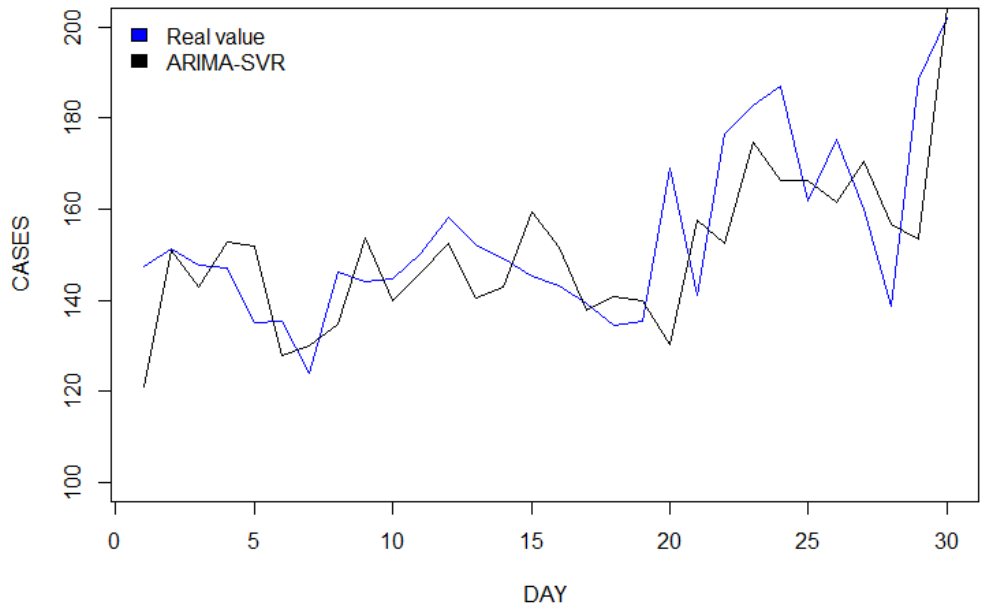
8. En el cuadro podemos comprobar como el modelo ARIMA-SVR obtiene mejores resultados en ambos casos, reduciendo de manera considerable los valores de los diferentes criterios descriptivos relacionados directamente con el error. Además, el coeficiente R^2 mejora considerablemente, sobre todo en el caso de los infectados por SARS-COV-2, y aunque es cierto que este coeficiente es sensible a la sobreparametrización, y cuantos más parámetros se empleen, mayor suele ser su valor, podemos comprobar en las gráficas cómo la predicción ARIMA-SVR se ajusta mejor a los valores reales que los procesos ARIMA. En las figuras 20, 21 se recoge la comparativa entre ambos modelos en cada una de las series.

Precio megavatio/hora en España					
MODELO	MAE	RMSE	sMAPE	MASE	R^2
ARIMA	12.994	16.296	0.084	1.105	0.222
ARIMA-SVR	11.629	14.898	0.076	0.989	0.35
Nuevos casos de SARS-COV-2 en Italia					
MODELO	MAE	RMSE	sMAPE	MASE	R^2
ARIMA	33606.22	44895.96	0.218	0.734	0.079
ARIMA-SVR	31024.47	41732.77	0.197	0.677	0.204

Cuadro 8: Resultados para predicciones de 30 días.

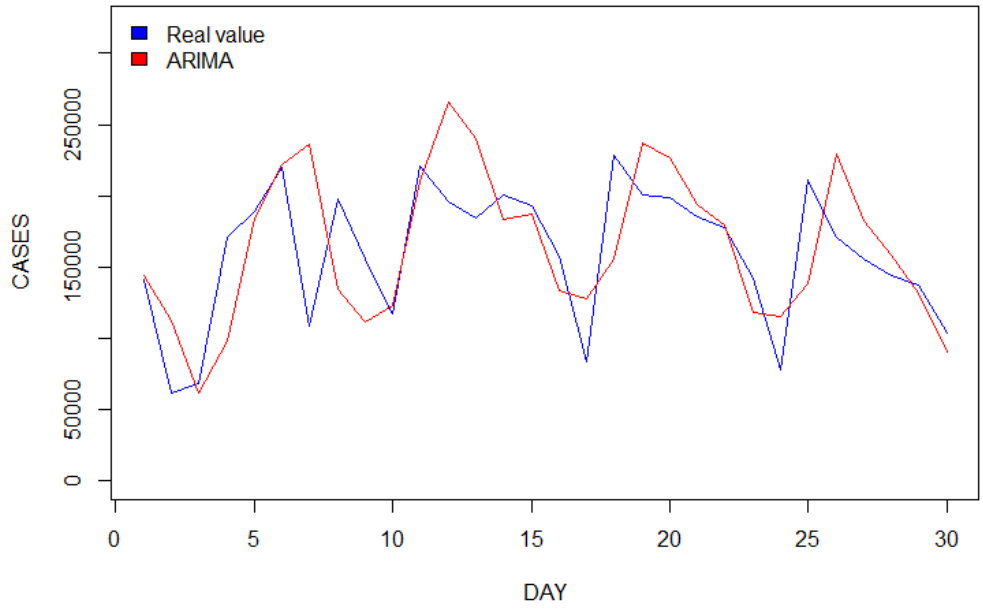


(a) Predicción por el proceso $ARIMA(4,1,6)$.

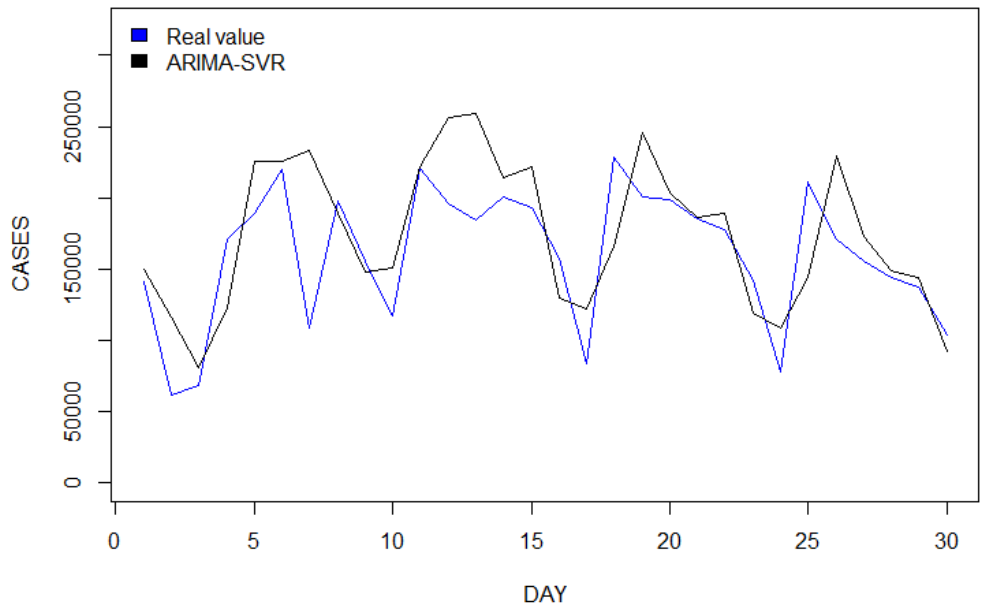


(b) Predicción por el proceso $ARIMA(4,1,6) - SVR$.

Figura 20: Comparativa de las predicciones en la serie del precio del megavatio/hora en España.



(a) Predicción por el proceso $ARIMA(8,1,7)$.



(b) Predicción por el proceso $ARIMA(8,1,7) - SVR$.

Figura 21: Comparativa de las predicciones en la serie de nuevos casos de SARS-COV-2 en Italia.

5. Conclusiones

Durante este trabajo hemos realizado un exhaustivo estudio sobre los procesos $ARIMA(p,d,q)$ y los modelos híbridos ARIMA-SVR. Los modelos $ARIMA(p,d,q)$ han sido usados en todos los ámbitos de la ciencia para modelar y predecir series temporales debido a su simpleza, ya que sólo dependen de la variable observada, y a la gran información que nos proporciona sobre cómo están relacionados los valores de una serie temporal con su pasado más cercano.

Sin embargo, queda patente por los resultados obtenidos que los modelos $ARIMA(p,d,q)$ tienen algunas debilidades a la hora de modelar el comportamiento de series temporales. La limitación más evidente es que, al tratarse de modelos lineales, cuando la serie temporal tiene mucha variabilidad entre valores consecutivos, los proceso $ARIMA(p,d,q)$ no son capaces de captar bien los cambios de tendencia.

Por ello, hemos tratado de solucionar estos problemas añadiendo al algoritmo una parte no lineal que incide directamente en el error cometido por el proceso $ARIMA(p,d,q)$. En la realización de los dos ejemplos recogidos en este trabajo, hemos comprobado que el modelo ARIMA-SVR mejora las predicciones conseguidas por los procesos $ARIMA(p,d,q)$, proporcionando un mejor ajuste de los valores predichos a la serie real y disminuyendo el error cometido por el sistema primigenio.

La utilización de este modelo híbrido nos proporciona por tanto un modelo que sigue siendo intuitivo y con el cuál podemos entender el funcionamiento de la serie temporal que se quiere modelar y a la vez mejora las predicciones sin que exista una pérdida en la intuición acerca de lo que está ocurriendo.

Sin embargo, aunque hayamos conseguido una mejora, el modelo ARIMA-SVR aún tiene limitaciones que habría que solventar, para lo que se pueden abrir dos vías de futura investigación que podrían ser interesantes. Por un lado,si queremos seguir trabajando con procesos ARIMA, existe toda una línea de investigación basada en modelos ARIMA a los que se le aplican modificaciones o se aplican filtros, como por ejemplo los modelos que incluyen el Filtro de Kalman, un algoritmo muy empleado en el sector de la economía sobre el que podemos encontrar trabajos recientes (Lagos-Álvarez et al., 2019; Borrero and Mariscal, 2022). También podemos quitar la limitación de linealidad en el modelo utilizando modelos autorregresivos no lineales como son las redes neuronales autorregresivas no lineales. Este sector está siendo muy prolifero debido al auge de la investigación en redes neuronales y muestra de ello es la cantidad de trabajos que podemos encontrar sobre el tema (Sun et al., 2020; Alsumaiei and Alrashidi, 2020; Liu et al., 2020).

Por todo ello, se puede considerar los procesos ARIMA como una fuente de investigación sobre la que aún se pueden obtener buenos resultados, tanto en su versión clásica como formando parte de otros modelos más complejos.

Anexo

Archivos do de Stata

```
/* Vamos a hacer el estudio ARIMA para la base de datos del precio
 del MWH en España */

/*Creamos una variable temporal t que tendrá comportamiento diario*/

gen t=td(01jan2021)+_n-1

/* Le damos estructura de serie temporal y comprobamos que la fecha
coincide con las del dataset original*/
tsset t, daily

/* Vamos a pintar la serie temporal para hacernos una idea de lo que
sucede*/

tsline var1

/* Vemos que claramente tiene un comportamiento no estacional, por
lo que vamos a tomar logaritmos.*/

gen lluz=log( var1)

/*Volvemos a pintarla*/


tsline lluz

/*Se ha suavizado la gráfica, pero aún así se ve claramente como
al menos no es estacionaria en varianza. Vamos a estudiar
el correlograma*/

corrgram  lluz
/*El correlograma da indicios claros de una raíz unitaria. Vamos
a diferenciar la serie*/


gen dlluz=d.lluz

/* Si la pintamos, vemos como el comportamiento ha cambiado, ya
sus valores se encuentran en torno a la media 0.*/

tsline dlluz


/*Vemos el test de dfgls para la diferenciación*/

dfgls dlluz
```

```
/*Comprobamos que todos los retardos están en la zona de rechazo,
por lo que hemos eliminado la raíz unitaria. Podemos asumir la
serie como estacionaria. Vamos a estudiar el correlograma para
estimar el orden.*/

corrgram dlluz



arima lluz , arima(4,1,7)   noconstant

estat ic
/*Tenemos AIC = 92.56645  - BIC =  144.8397*/

arima lluz , arima(5,1,7)   noconstant

estat ic

/*Tenemos AIC =  94.29291    - BIC =  150.9223  */

arima lluz , arima(6,1,7)   noconstant

estat ic

/*Tenemos AIC = 92.56645  - BIC =  144.8397*/

arima lluz , arima(4,1,6)   noconstant

estat ic
/*Tenemos AIC =   90.99811 - BIC = 138.9153  */

/*Elegimos el modelo ARIMA(4,1,7). Sólo nos queda validar
el modelo con los residuos.*/


predict residuos, resi

corrgram residuos

tsline residuos

/*Comprobamos con el correlograma y la gráfica de los
residuos que se trata de ruido blanco, por lo que queda
 validado el modelo.*/
```

Funciones ARIMA y ARIMA-SVR de R

```
dividir_SVR<-function(serie_temporal,p){
```

```
  n_serie<-length(serie_temporal)
  p1<-p+1
  n_patrones<-n_serie-p
  M_patrones<-matrix(0,n_patrones,p1)
  for(i in 1:n_patrones){
    for(j in 1:p1){
      M_patrones[i,j]<-serie_temporal[i+p-j+1]

    }

  }
  M_patrones<-data.frame(M_patrones)
  M_patrones

}
#############################################################



Arima_SVR<-function(base_datos,orden,n_svr,n_dias,
        coeficientes_ar,coeficientes_ma,fecha){

  base_datos_original<-base_datos
  base_datos<-subset(base_datos,FECHA<=fecha)
  p=orden[1]
  d=orden[2]
  q=orden[3]
  m=max(p,q)+1
  serie_temporal<-ts(base_datos$VARIABLE)
  serie_temporal<-diff(serie_temporal,differences = d)
  n_serie<-length(serie_temporal)
  error_arima<-matrix(0,n_serie,1)
  y_estimada<-0
  resultados<-c(1:n_dias)

  for(i in m:n_serie){

    for(j in 1:p){
      y_estimada<-coeficientes_ar[j]*serie_temporal[i-j]+y_estimada

    }

    for(k in 1:q){
      y_estimada<-coeficientes_ma[k]*error_arima[i-k]+y_estimada

    }

    error_arima[i]<-serie_temporal[i]-y_estimada
    y_estimada<-0
```

```
}
error_modelo<-error_arima
error_arima<-ts(error_arima)
M_error<-dividir_SVR(error_arima,n_svr)
modelo_error<-svm(X1~.,M_error)
y_estimada<-0
fecha<-fecha+1

for (i in 1:n_dias) {

  base_datos<-subset(base_datos_original,FECHA<=fecha)
  serie_temporal<-ts(base_datos$VARIABLE)
  serie_temporal_original<-serie_temporal
  n_original<-length(serie_temporal_original)-1
  serie_temporal<-diff(serie_temporal,differences = d)
  n_serie<-length(serie_temporal)

  for(j in 1:p){
    y_estimada<-coeficientes_ar[j]*serie_temporal[n_serie-j]+y_estimada

  }

  for(k in 1:q){
    y_estimada<-coeficientes_ma[k]*error_modelo[n_serie-k]+y_estimada

  }

  M_test<-matrix(0,1,n_svr+1)

  for(l in 2:n_svr){
    M_test[1,l]<-error_modelo[n_serie-l+1]

  }
  M_test<-data.frame(M_test)
  M_test<-select(M_test,-X1)
  prediccion_error<-predict(modelo_error,M_test)
  prediccion_error<-as.numeric(prediccion_error)
  prediccion<-y_estimada+prediccion_error
  prediccion_1<-y_estimada+prediccion_error
  error_modelo<-rbind(error_modelo,serie_temporal[n_serie]-y_estimada)
  y_estimada<-0

  if(d>=2){
    res<-c(1:d)
    serie<-serie_temporal

    for (h in 1:d) {
      n_serie<-length(serie)
```

```
        res[h]<-serie[n_serie]
        serie<-diff(serie)

      }


      for (h in d:1) {

        prediccion_1<-prediccion_1+res[h]

      }

      resultados[i]<-prediccion_1


    }

    if(d==0){resultados[i]<-prediccion}

    if(d==1){
      resultados[i]<-prediccion+serie_temporal_original[n_original]
    }


    fecha<-fecha+1



  }


  resultados



}

#################################################################

Arima_coeficientes<-function(base_datos,orden,n_dias,
        coeficientes_ar,coeficientes_ma,fecha){

  base_datos_original<-base_datos
  base_datos<-subset(base_datos,FECHA<=fecha)
  p=orden[1]
  d=orden[2]
  q=orden[3]
  m=max(p,q)+1
  serie_temporal<-ts(base_datos$VARIABLE)
  serie_temporal<-diff(serie_temporal,differences = d)
```

```
n_serie<-length(serie_temporal)
error_arima<-matrix(0,n_serie,1)
y_estimada<-0
resultados<-c(1:n_dias)

for(i in m:n_serie){

  for(j in 1:p){
    y_estimada<-coeficientes_ar[j]*serie_temporal[i-j]+y_estimada

  }

  for(k in 1:q){
    y_estimada<-coeficientes_ma[k]*error_arima[i-k]+y_estimada

  }

  error_arima[i]<-serie_temporal[i]-y_estimada
  y_estimada<-0


}
error_modelo<-error_arima
error_arima<-ts(error_arima)
y_estimada<-0
fecha<-fecha+1


for (i in 1:n_dias) {

  base_datos<-subset(base_datos_original,FECHA<=fecha)
  serie_temporal<-ts(base_datos$VARIABLE)
  serie_temporal_original<-serie_temporal
  n_original<-length(serie_temporal_original)-1
  serie_temporal<-diff(serie_temporal,differences = d)
  n_serie<-length(serie_temporal)

  for(j in 1:p){
    y_estimada<-coeficientes_ar[j]*serie_temporal[n_serie-j]+y_estimada

  }

  for(k in 1:q){
    y_estimada<-coeficientes_ma[k]*error_modelo[n_serie-k]+y_estimada

  }

  prediccion<-y_estimada
  prediccion_1<-y_estimada
  error_modelo<-rbind(error_modelo,serie_temporal[n_serie]-y_estimada)
```

```
    y_estimada<-0

    if(d>=2){
      res<-c(1:d)
      serie<-serie_temporal

      for (h in 1:d) {
        n_serie<-length(serie)
        res[h]<-serie[n_serie]
        serie<-diff(serie)

      }


      for (h in d:1) {

        prediccion_1<-prediccion_1+res[h]

      }

      resultados[i]<-prediccion_1


    }

    if(d==0){resultados[i]<-prediccion}

    if(d==1){
      resultados[i]<-prediccion+serie_temporal_original[n_original]
    }


    fecha<-fecha+1



  }


  resultados




}
```

Referencias

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alsumaiei, A. A. and Alrashidi, M. S. (2020). Hydrometeorological drought forecasting in hyper-arid climates using nonlinear autoregressive neural networks. *Water*, 12(9).

Borrero, J. D. and Mariscal, J. (2022). Predicting time seriesusing an automatic new algorithm of the kalman filter. *Mathematics*, 10(16).

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control (Revised Edition)*. Holden-Day, revised edition.

Brockwell, P. J. and Davis, R. A. (2006). *Introduction to time series and forecasting*. Springer texts in statistics. Springer, 2 edition.

Dickey, D. A. and 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.

Du, K.-L. and Swamy, M. (2013). *Neural Networks and Statistical Learning*.

González Casimiro, M. P. (2009). *Análisis de series temporales: Modelos ARIMA*. SARRIKO-ON.

Hernandez-Matamoros, A., Fujita, H., Hayashi, T., and Perez-Meana, H. (2020). Forecasting of covid19 per regions using arima models and polynomial functions. *Applied Soft Computing*, 96:106610.

Jamil, R. (2020). Hydroelectricity consumption forecast for pakistan using arima modeling and supply-demand analysis for the year 2030. *Renewable Energy*, 154:1–10.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178.

Lagos-Álvarez, B., Padilla, L., Mateu, J., and Ferreira, G. (2019). A kalman filter method for estimation and prediction of space-time data with an autoregressive structure. *Journal of Statistical Planning and Inference*, 203:117–130.

Liu, M., Li, G., Li, J., Zhu, X., and Yinhong, Y. (2020). Forecasting the price of bitcoin using deep learning. *Finance Research Letters*, 40:101755.

Mehmood, Q., Sial, M., Riaz, M., and Shaheen, N. (2019). Forecasting the production of sugarcane crop of pakistan for the year 2018-2030, using box-jenkin's methodology. *Journal of Animal and Plant Sciences*, 5.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

Sujjaviriyasup, T. and Pitiruek, K. (2013). Hybrid arima-support vector machine model for agricultural production planning. *Applied Mathematical Sciences*, 7:2833–2840.

Sun, Z., Li, K., and Li, Z. (2020). Prediction of horizontal displacement of foundation pit based on nar dynamic neural network. *IOP Conference Series: Materials Science and Engineering*, 782:042032.

Uriel Jímenez, E. (1991). *Analisis de series temporales - modelos ARIMA*. Paraninfo.

Velásquez, J., Olaya, Y., and Franco, C. (2010). Time series prediction using support vector machines. *Ingeniare*, 18:64–75.

Wang, B., Liu, P., Chao, Z., Junmei, W., Chen, W., Cao, N., O'Hare, G., and Wen, F. (2018). Research on hybrid model of garlic short-term price forecasting based on big data. *Computers, Materials & Continua*, 57:283–296.

Xu, S., Chan, H., and Zhang, T. (2018). Forecasting the demand of the aviation industry using hybrid time series sarima-svr approach. *Transportation Research Part E Logistics and Transportation Review*, 122:169–180.

Yang, H. and O'Connell, J. (2020). Short-term carbon emissions forecast for aviation industry in shanghai. *Journal of Cleaner Production*, 275:122734.