Aplicación de modelos econométricos para el análisis de la incidencia del COVID-19 en España

Inglada-Pérez, Lucía; lucia.inglada@madrid.uned.es
Departamento de Economía Aplicada y Estadística.
UNED
Coto-Millán, Pablo; cotop@unican.es
Departamento de Economía
Universidad de Cantabria
Casares, Pedro; casaresp@unican.es
Departamento de Economía.
Universidad de Cantabria
Inglada López de Sabando, Vicente; vinglada@cee.uned.es
Departamento de Economía Aplicada y Estadística
UNED

RESUMEN

Desde que la enfermedad por coronavirus (COVID-19) apareció en China a finales de diciembre de 2019, los daños sociales, sanitarios y económicos producidos por su vertiginosa propagación en prácticamente todo el mundo, han sido devastadores. España es uno de los países donde la pandemia ha incidido con mayor virulencia, incluyendo más de 2,3 millones de casos confirmados y más de 72.900 fallecimientos hasta el 19 de marzo de 2021. Por ello, es sumamente relevante analizar, monitorizar y predecir la incidencia del COVID-19 en España con el fin de ayudar a formular políticas de salud pública que contribuyan a controlar la propagación de la epidemia de forma más eficaz. Los modelos econométricos de series temporales son importantes para predecir el impacto de la epidemia de COVID-19 y tomar las medidas necesarias para responder a esta crisis. En este estudio se aplican modelos de vectores autorregresivos (VAR) y modelos autorregresivos con retardos distribuidos (ARDL), para analizar y predecir la incidencia del COVID-19 en España, uno de los países más afectados de Europa. Los resultados del análisis pueden ayudar a comprender la evolución de la epidemia y proporcionar una base teórica para la adopción de nuevas políticas de intervención.

ABSTRACT

Since the coronavirus disease (COVID-19) emerged in China in late December 2019, the social, health, and economic damage produced by its dizzying spread virtually worldwide has been devastating. Spain is one of the countries where the pandemic has had the most virulent impact, including more than 2.3 million confirmed cases and more than 72,900 deaths as of March 19, 2021. It is therefore highly relevant to analyze, monitor and predict the incidence of COVID-19 in Spain in order to help formulate public health policies that contribute to controlling the spread of the epidemic more effectively. Econometric models are important to predict the impact of the COVID-19 epidemic and to take the necessary measures to respond to this crisis. This study applies vector autoregressive (VAR) and autoregressive distributed lag (ARDL) models to analyse and predict the COVID-19 incidence in Spain, one of the most affected countries in Europe. The results of the analysis may help to understand the evolution of the epidemic and provide a theoretical basis for the adoption of new intervention policies.

***Keywords*:** COVID-19; pandemic; VAR model; ARDL model; forecasting; time series

***Palabras claves*:** COVID-19; pandemia; modelos VAR; modelos ARDL; predicción; series temporales

***Área temática*:** A3 - Métodos cuantitativos en un entorno con incertidumbre.

1. INTRODUCCIÓN

Desde que la enfermedad por coronavirus (COVID-19) apareció en Wuhan (China) a finales de diciembre de 2019, los daños sociales, sanitarios y económicos producidos por su vertiginosa propagación en prácticamente todo el orbe, han sido devastadores. Hasta el 19 de marzo de 2021, se habían confirmado más de 12,2 millones de casos de COVID-19 en todo el mundo, incluyendo más de 2,6 millones de fallecimientos. España es uno de los países donde la pandemia ha incidido con mayor virulencia, incluyendo a más de 2,3 millones de casos confirmados y más de 72.900 fallecimientos hasta el 19 de marzo de 2021. Ante este dramático panorama, el 30 de enero de 2020, la Organización Mundial de la Salud (OMS) declaró el brote como una Emergencia de Salud Pública Internacional por causa del COVID-19, mientras que el 11 de marzo definió la situación existente como "pandemia" (Guirao, 2020). La rápida evolución de la pandemia obligó a que el 14 de marzo de 2020 el Gobierno español dictase el estado de alarma y otras medidas como el confinamiento total o la cuarentena obligatoria en todo el país para controlar la propagación del virus.

Ante este panorama desalentador, adquiere un carácter sumamente relevante analizar y predecir el impacto de la pandemia del COVID-19, con el fin de ayudar a formular políticas de salud pública que contribuyan a controlar la propagación de la epidemia de forma más eficaz. Los modelos econométricos de series temporales se presentan como una alternativa relevante para predecir el impacto del brote de COVID-19 y tomar las medidas necesarias para responder a esta crisis.

En este artículo estudiamos la incidencia del Covid-19 en España, con el objetivo de explicar y predecir la dinámica subyacente existente en la evolución de la epidemia. Con dicho objetivo, se estiman modelos de vectores autorregresivos (VAR) y modelos autorregresivos con retardos distribuidos (ARDL), para analizar y predecir la incidencia del COVID-19 en España, uno de los países más afectados de Europa. Las variables seleccionadas son el número diario de casos nuevos y el número diario de fallecidos. Las variables elegidas presentan ventajas respecto a otras alternativas como son los casos y muertes acumuladas ya que son más representativas de la gravedad de la epidemia (Wang et al., 2021). El modelo VAR es utilizado para predecir la dinámica de las dos variables mientras que el modelo ARDL es empleado para estimar las elasticidades a corto y largo

plazo de la variable correspondiente al número diario de muertos respecto al número diario de casos, así como la velocidad de ajuste al equilibrio.

Desde que la incidencia de la pandemia adquirió un cierto grado de virulencia se ha producido un aluvión de investigaciones relacionadas con la incidencia del COVID-19. Gnanvi et al. (2021) llevan a cabo una revisión sistemática de los estudios publicados entre el 1 de enero y el 30 de noviembre de 2020 sobre Covid-19 con el objetivo de resumir las tendencias en las técnicas de modelización utilizadas y evaluar la fiabilidad de las predicciones de casos y muertes por Covid-19. Como muestra del gran volumen de información existente generada en torno a la pandemia, cabe destacar que, en el periodo de tiempo citado, la investigación realizada detecta 4.311 artículos relacionados con las palabras clave definidas por los autores y en definitiva con la pandemia. Finalmente, se analizan con detalle 242 artículos.

Una revisión de la literatura nos muestra la amplia variedad de metodologías utilizadas para estudiar la incidencia del COVID-19 (Gnanvi et al., 2021; ArunKumar et al., 2020). Sin embargo, son escasos los trabajos que emplean el modelo VAR en su análisis. En este sentido, Wang et al. (2021) aplican un modelo VAR para predecir los casos diarios de la epidemia del COVID-19 en Estados Unidos. De acuerdo con la revisión de la literatura y el análisis de correlación que llevan a cabo, aparte de la variable correspondiente al número de casos nuevos, incluyen otras variables adicionales en el modelo VAR: muertes acumuladas, pacientes recuperados acumulados, temperatura y humedad. Finalmente, realizan una predicción a 30 días con resultados satisfactorios. Utilizando también un modelo VAR, Khan et al. (2020) predicen con un horizonte de diez días la incidencia de la epidemia del COVID-19 en Pakistán. En su modelo utilizan datos diarios del COVID-19, que corresponden a las variables: nuevos casos, muertes y enfermos recuperados. Obtienen que la precisión de sus predicciones es elevada.

Esta investigación contribuye a la literatura internacional al ser la primera vez que se aplica el marco metodológico adoptado en este trabajo para estudiar la incidencia del COVID-19 en el caso español. Los resultados del análisis realizado pueden ayudar a comprender la evolución de la epidemia y proporcionar una base teórica para la adopción de nuevas políticas de intervención. Asimismo, una ventaja relevante de los modelos de series temporales utilizados en este trabajo es que se pueden actualizar continuamente mediante la incorporación de nuevos datos modificando los escenarios previstos.

Específicamente, la previsión de la incidencia de la enfermedad es importante para una eficiente asignación de los recursos disponibles en el sistema sanitario (Ceylan, 2020). Debido a la gravedad de los efectos económicos y sociales generados por la pandemia, su interés se extiende a múltiples ámbitos como el diseño de la política económica, de movilidad o de empleo.

La estructura de este artículo es la siguiente. En la sección segunda se analizan los principales rasgos observados en la evolución de la pandemia en España. A continuación, se introduce el marco metodológico utilizado, que incluye a los modelos de vectores autorregresivos (VAR) y modelos autorregresivos con retardos distribuidos (ARDL). A partir de la estimación de los modelos seleccionados, en la sección cuarta se analizan los resultados obtenidos. Finalmente, en el apartado quinto se extraen una serie de conclusiones.

2. EVOLUCIÓN DE LA INCIDENCIA

En esta sección se describen los principales rasgos de la evolución de la incidencia de la pandemia del COVID-19 en España. Nos fijamos en el número diario de casos nuevos y en el número diario de fallecidos. En las figuras 1, 2, 3 y 4 se representa la evolución desde el 5 de marzo de 2020 hasta el 16 de marzo de 2021 de las series temporales correspondientes, respectivamente, al número diario de casos nuevos, número diario de casos acumulados, número diario de fallecidos y número diario de fallecidos acumulados.

Los cuatro gráficos muestran la existencia nítida de tres ciclos u oleadas de diferentes grados de intensidad y duración. El primer caso positivo en España se registra el 31 de enero de 2020 y significa el inicio de un primer ciclo caracterizado por su alta incidencia, especialmente en relación con el número de fallecimientos. La gravedad de la situación provoca que el 14 de marzo de 2020, el Gobierno aplicase el estado de alarma para limitar los contactos entre personas. Una consecuencia fue la reducción de la movilidad en un 75% en relación con los niveles existentes antes de la pandemia. El máximo número diario de muertes durante este primer ciclo se alcanza el 1 de abril cuando se registran 913 fallecimientos por COVID-19. La retirada progresiva de medidas de confinamiento y restricciones a la movilidad comienza en los primeros días de mayo y el 21 de junio todo el territorio español entra en la denominada “nueva normalidad”. En

dicha fecha el número de casos y el de muertos se había reducido a 344 y 18, respectivamente, aunque los valores acumulados en esa fecha (252.656 y 29.638) mostraban la virulencia de la incidencia de la pandemia. A partir de la segunda quincena de julio, al reducirse las restricciones impuestas, comienza la fase de expansión de un nuevo ciclo en la evolución de la incidencia de la enfermedad que se caracteriza por una tasa de crecimiento menor que el ciclo anterior. De nuevo, el 25 de octubre el Gobierno dictaba el estado de alarma para que las comunidades autónomas pudiesen imponer nuevas restricciones más severas. El 8 de noviembre se alcanza el máximo valor del número de muertos en esta segunda ola con 353 fallecimientos, sensiblemente inferior al máximo de la primera ola. Por el contrario, el máximo del número diario de casos nuevos es sensiblemente superior al de la primera ola. El mayor número de pruebas realizadas para detectar la enfermedad, que influye muy significativamente en el número de casos positivos, ha podido reflejarse en este diferente comportamiento de la segunda ola. Finalmente, el tercer ciclo se inicia a partir de la Navidades de 2020 y alcanza su máximo el 29 de enero de 2021 con 354 fallecimientos causados por la enfermedad (Guirao,2020).

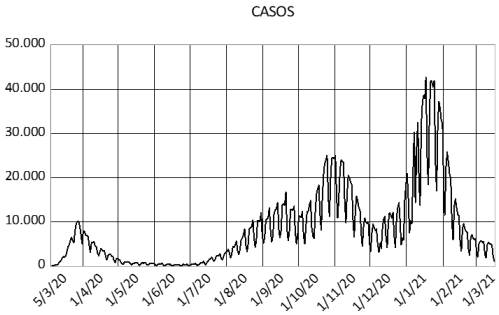

Figura 1. Evolución de nuevos casos.

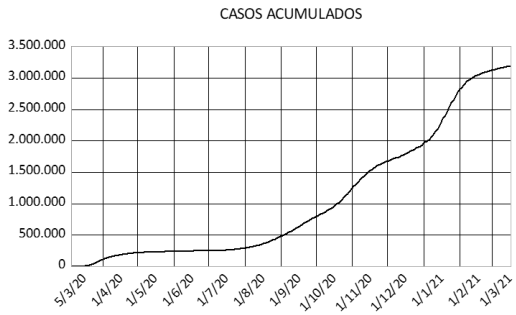

Figura 2. Evolución de casos acumulados.

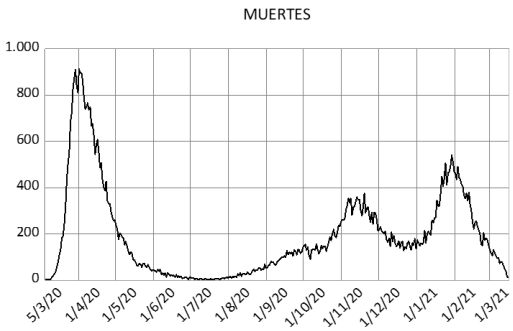

Figura 3. Evolución de nuevas muertes.

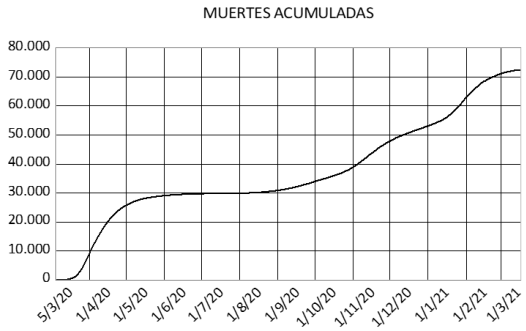

Figura 4. Evolución de muertes acumuladas.

El estudio denominado ENE-COVID (Ministerio de Sanidad, 2020) constituye una fuente de información privilegiada para conocer la evolución de la prevalencia de la enfermedad del COVID-19. Se trata de "un amplio estudio longitudinal sero-epidemiológico, de base poblacional, cuyos objetivos son estimar la prevalencia de infección por SARS-CoV-2 mediante la determinación de anticuerpos frente al virus en España y evaluar su evolución temporal". Se compone de 4 rondas (27/04-11/05, 18/05-01/06, 08/06-22/06 y finalmente del 16 al 29 de noviembre). El informe estima la prevalencia acumulada o global, es decir, el porcentaje de personas que han tenido o tienen anticuerpos IgG anti SARS-CoV-2, que reflejaría la magnitud de la pandemia. De acuerdo con los resultados de la cuarta ronda, en la que participaron 51.409 personas, la prevalencia global (porcentaje de personas en la población con anticuerpos IgG frente a SARSCoV-2 desde el inicio del estudio) se sitúa en un 9,9% (con intervalo de confianza (IC) al 95%: 9,4-10,4), siendo bastante similar en mujeres (10,1%; IC95%: 9,5-10,7) y en hombres (9,6%; IC95%: 9,0-10,2). Cabe destacar que, a pesar de la gran difusión de la epidemia, se observa una pronunciada dispersión geográfica. Mientras Coruña, Canarias y Lugo, presentan prevalencias inferiores o cercanas al 4%, el núcleo central alrededor de Madrid muestra cifras cercanas o superiores al 15%.

3. DATOS Y METODOLOGÍA

3.1. Datos

Las variables que se han utilizado para medir la incidencia de la pandemia en España son el número diario de nuevos casos y el número diario de fallecimientos. Los datos diarios han sido extraídos del sitio web de la Organización Mundial de la Salud (https://www.who.int/) y abarcan el periodo entre el 5 de marzo de 2020 y el 16 de marzo de 2021 con un total de 377 observaciones.

En la Tabla 1 se muestran los principales estadísticos descriptivos de las dos series investigadas. Se observa que la variabilidad expresada por el coeficiente de variación es próxima a la unidad y similar en las dos series. Los máximos de cada variable son 42.736 y 913 mientras que los valores mínimos son 85 y 1.

Tabla 1. Principales Estadísticos descriptivos de las variables investigadas.

	Número diario de casos nuevos	Número diario de muertes
Media	8474,45	192,11
Mediana	5950	141
Máximo	42736	913
Mínimo	85	1
Desviación Típica	8891,69	197,18
Coeficiente de variación	0,953	0,974
Número de observaciones	377	377

Fuente: Elaboración propia.

3.2. Metodología

3.2.1. Modelos VAR

Con la finalidad de analizar las interacciones globales entre las dos variables y predecir su evolución en el corto plazo, se estima un modelo VAR (Sims, 1980 y Sims et al., 1990). Los modelos VAR constituyen una ampliación de los modelos autorregresivos univariantes, que se caracterizan por el hecho de que todas las variables son consideradas como endógenas y se expresan en el modelo como una combinación lineal tanto de sus propios retardos como de los retardos de las demás variables consideradas. Así, para el caso de un modelo de dos variables como es nuestro caso, su expresión es la siguiente:

$$X_t = \propto_{10} + \alpha_{11}X_{t-1} + \alpha_{12}X_{t-2} + \cdots + \alpha_{1p}X_{t-p} + \beta_{11}Y_{t-1} + \beta_{12}Y_{t-2} + \cdots + \alpha_{1p}Y_{t-p} + u_{1t}$$

$$Y_t = \propto_{20} + \alpha_{21}X_{t-1} + \alpha_{22}X_{t-2} + \cdots + \alpha_{2p}X_{t-p} + \beta_{21}Y_{t-1} + \beta_{22}Y_{t-2} + \cdots + \alpha_{2p}Y_{t-p} + u_{2t}$$

En estas ecuaciones, los procesos X_t e Y_t se expresan como una combinación de retardos, siendo $\propto_{ij}$ y β_{ij} los coeficientes en la ecuación i del retardo j, con $i = 1, 2$ y $j = 1, \dots, p$. Además, u_{1t} y u_{2t} son dos series independientes e idénticamente distribuidas con media cero.

El número de retardos en cada ecuación, que suele ser el mismo para cada variable, define el orden del VAR. Para determinar el orden de los retardos del modelo VAR, se suele recurrir a criterios de información tales como los de Akaike o Schwarz que serán los adoptados en esta investigación. Una vez establecido el orden del VAR, la estimación de cada una de las ecuaciones puede llevarse a cabo por mínimos cuadrados ordinarios (MCO). De este modo se obtienen estimadores eficientes.

Es comúnmente aceptado en la literatura internacional que el modelo VAR debe aplicarse en el caso de que las variables sean estacionarias. Sin embargo, en Sims (1980) y Sims et al. (1990) se argumenta que en el modelo VAR es posible utilizar variables que no son estacionarias ya que generalmente el objetivo de un análisis VAR es determinar las interrelaciones entre las variables, y no tanto determinar específicamente las estimaciones de los parámetros.

Cabe destacar que el modelo VAR contribuye a identificar las interacciones dinámicas que caracterizan el sistema de ecuaciones estimado mediante la construcción de las denominadas funciones de impulso-respuesta y la descomposición de la varianza del error de predicción.

Las funciones impulso-respuesta (FIR) recogen los impactos inducidos por los shocks en las variables del sistema. Una alteración en el comportamiento de una variable afectará directamente a la misma y se transmitirá al resto a través de la estructura dinámica que representa el modelo VAR (Díaz Fernández et al., 2014). En forma resumida se introduce una alteración en el vector aleatorio que en la práctica suele ser igual al valor de la desviación típica y se contrasta el resultado que dicha alteración ha generado sobre el sistema. Se trata de una herramienta muy empleada en la simulación de políticas. Asimismo, la descomposición de la varianza obtenida a partir de la ortogonalización del término aleatorio permite aproximar la dependencia relativa de cada variable sobre el resto. Esta herramienta nos suministra información de la potencia relativa de las innovaciones aleatorias para cada variable endógena.

3.2.2. Modelo ARDL

Para determinar si existe una relación a largo plazo entre el número diario de fallecidos y el de casos nuevos, la función que se considera tiene la forma siguiente:

$$muer_t = \beta_0 + \beta_1 cas_t + u_t$$

Donde:

$muer_t$ es el logaritmo natural del número de fallecidos;

cas_t es el logaritmo natural del número de casos;

$t = 1.....T$ es el número del día correspondiente.

u_t es el término de error.

Para estimar esta función utilizaremos el método ARDL, que fue introducido originalmente por Pesaran y Shin (1999) y posteriormente ampliado por Pesaran et al. (2001). Este método tiene numerosas ventajas en comparación con otros métodos de cointegración. La principal es que, a diferencia de otras técnicas de cointegración, el modelo ARDL no impone el supuesto restrictivo de que todas las variables consideradas deben tener el mismo orden de integración. Es decir, el enfoque ARDL puede aplicarse independientemente de si los regresores subyacentes son integrados de orden uno [I(1)] o de orden cero [I(0)]. Sin embargo, también se debe aplicar un test de raíces unitarias porque este modelo no es aplicable cuando una de las variables es integrable de orden 2.

Una representación del modelo ARDL(k,l) en el caso general con constante y tendencia es la siguiente:

$$muer_t = \alpha_c + \alpha_d t + \sum_{i=1}^{k} \alpha_{1,i}\, muer_{t-i} \sum_{i=0}^{l} \alpha_{2,i}\, casos_{t-i} + w_t$$

donde w_t es un término de error y *k y l* son las longitudes de los retardos de las variables individuales.

Esta ecuación se puede transformar en la siguiente ecuación de corrección de error:

$$\Delta muer_t = \theta_c + \theta_d \Delta t - \theta_{ect} ECT_{t-1} + \sum_{i=1}^{k} \theta_{1,i}\, \Delta muer_{t-i} \sum_{i=0}^{l} \theta_{2,i}\, \Delta casos_{t-i} + v_t$$

donde Δ es el operador de la primera diferencia, ECT_{t-1} es el término de corrección del error resultante de la relación de equilibrio a largo plazo estimada, y θ_{ect} es el coeficiente que refleja la velocidad de ajuste a largo plazo, es decir, la corrección porcentual diaria de una desviación del equilibrio a largo plazo en el día anterior**.**

Asimismo, el coeficiente de la velocidad de ajuste es $\theta_{ect} = 1 - \sum_{i=1}^{k} \alpha_{1,i}$ y el coeficiente o elasticidad a largo plazo es $\beta = \frac{\sum_{i=0}^{l} \theta_{2,i}}{\theta_{ect}} = \frac{\sum_{i=0}^{l} \theta_{2,i}}{\left(1-\sum_{i=1}^{k} \alpha_{1,i}\right)}$

La herramienta metodológica empleada en esta investigación para estimar el modelo propuesto es el análisis de cointegración. Este método proporciona un marco adecuado para contrastar las relaciones relevantes a largo plazo entre series temporales no estacionarias y nos permite contrastar la existencia de una relación en el largo plazo entre las variables de nuestro modelo.

El proceso metodológico adoptado se compone de las siguientes etapas que se describen de forma resumida a continuación: A) análisis de estacionariedad, B) análisis de cointegración, y C) análisis de la causalidad.

A) Análisis de estacionaridad

El primer paso del proceso metodológico es el análisis de las raíces unitarias o estacionaridad que nos informa sobre el orden de integración de cada variable. Como se ha mencionado anteriormente, las variables deben ser I(0) o I(1) para aplicar este modelo. En esta investigación empleamos el test de Dickey-Fuller Aumentado (ADF) para contrastar el orden de integrabilidad. Para realizar este test (Dickey and Fuller, 1979) se plantea el siguiente modelo:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

Donde α es una constante, β es el coeficiente de la tendencia temporal y p el orden del proceso autorregresivo. En general se elige el orden p del proceso autorregresivo con base en los criterios de información de Akaike o de Schwartz. La hipótesis que se contrasta es la siguiente: $H_0 \equiv \gamma = 0$ frente a la alternativa de $H_1 \equiv \gamma < 0$. Es decir, la hipótesis nula es que la variable contiene una raíz unitaria o que no es estacionaria. El estadístico para realizar el test es el siguiente: $DF_\tau = \frac{\hat{\gamma}}{SE(\gamma)}$. Este test se aplica tanto a la serie en niveles como a la serie obtenida mediante la aplicación del operador de la primera diferencia con el fin de determinar el orden de integración.

B) Test de Cointegración

Los trabajos de Granger (1981) y Engle y Granger (1987) fueron los primeros en los que se formalizó el concepto de cointegración. En dichos trabajos se aportan métodos de estimación y test para contrastar la existencia de una relación a largo plazo entre un conjunto de variables en un marco dinámico. La comprobación de la existencia de cointegración es un paso necesario para establecer si un modelo presenta empíricamente relaciones significativas en el largo plazo. Si no se logra establecer la existencia de cointegración entre las variables subyacentes, resulta imperativo trabajar con las variables en diferencias. Sin embargo, con esta última forma no podremos disponer de información sobre el comportamiento en el largo plazo.

En esta investigación la relación a largo plazo entre las variables consideradas se examina mediante el denominado test "bounds" de cointegración. Antes de aplicar dicho test se determina el orden de los retardos utilizando, por ejemplo, el criterio de información de Akaike o el criterio bayesiano de Schwartz.

Para el caso de dos variables *x* e *y*, las ecuaciones utilizadas para este test son las siguientes:

$$\Delta y_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^{m} \alpha_{2,i}\, \Delta y_{t-i} + \sum_{i=0}^{n} \alpha_{3,i}\, \Delta x_{t-i} + \alpha_{4,i} y_{t-1} + \alpha_{5,i} x_{t-1} + \mu_{1t}$$

$$\Delta x_t = \beta_0 + \beta_1 t + \sum_{i=1}^{m} \beta_{2,i}\, \Delta x_{t-i} + \sum_{i=0}^{n} \beta_{3,i}\, \Delta y_{t-i} + \beta_{4,i} x_{t-1} + \beta\alpha_{5,i} y_{t-1} + \mu_{2t}$$

Donde μ_{1t} y μ_{2t} son los términos de error (ruido blanco). Los demás términos ya se describieron anteriormente.

El contraste de cointegración con esta metodología se lleva a cabo bajo las siguientes hipótesis:

$$H_0 : \alpha_1 = \alpha_2 \ldots = \alpha_n = 0$$

$$H_1 : \alpha_1 \neq \alpha_2 \ldots \neq \alpha_n \neq 0$$

Por tanto, existiría cointegración entre las variables si se rechazase la hipótesis nula consistente en que todos los coeficientes son nulos. Para contrastar la existencia de cointegración se utiliza el estadístico *F*. Los valores obtenidos del estadístico se comparan con los valores críticos definidos por Pesaran et al. (2001). En uno de los conjuntos de valores críticos se supone que todas las variables incluidas en el modelo ARDL son I(0), mientras que el otro se calcula bajo el supuesto de que las variables son I(1). Si el estadístico calculado del test supera el valor de los límites críticos superiores, se rechaza la hipótesis Ho. En el caso de que el estadístico F esté dentro de los límites, el test de cointegración no es concluyente. Finalmente, si el estadístico F es inferior al valor de los límites inferiores, entonces no se puede rechazar la hipótesis nula de que no existe cointegración. También podemos utilizar los valores críticos definidos por Narayan (2005) que son más apropiados para muestras pequeñas.

En el caso de que se confirme la existencia de cointegración en las ecuaciones anteriores, se procede a estimar los modelos a largo y a corto plazo y se obtienen las respectivas elasticidades a largo y a corto plazo.

C) Causalidad

A continuación, se aplica el método propuesto por Granger (1969) para detectar el sentido de la causalidad. De acuerdo con la definición de causalidad de Granger, una serie temporal, X_t, causa otra serie temporal, Y_t, en el sentido de Granger, si Y_t puede

predecirse con mayor precisión utilizando valores pasados de X_t que en el caso de no hacerlo. Este método de causalidad de Granger implica la comprobación de la hipótesis nula de que X_t no causa Y_t y viceversa, simplemente estimando las dos regresiones siguientes donde μ_t , ϵ_t son dos procesos ruido blanco y l denota el número de variables retardadas.

Consideramos las siguientes especificaciones para cada una de las dos variables que son similares a la descritas en el modelo VAR.

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_l y_{t-l} + \beta_1 x_{t-1} + \cdots + \beta_l x_{t-l} + \mu_t$$

$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \cdots + \alpha_l x_{t-l} + \beta_1 y_{t-1} + \cdots + \beta_l y_{t-l} + \epsilon_t$$

Los estadísticos F que consideramos son los estadísticos Wald para la hipótesis conjunta:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_l = 0$$

para cada ecuación. Es decir, la hipótesis nula es que la variable *x* no causa a la variable *y* en el sentido de Granger en la primera regresión y que la variable *y* no causa a la variable *x* en el sentido de Granger en la segunda regresión. Asimismo, la existencia de causalidad en el largo plazo vendría determinada por el estadístico *t* del coeficiente del término de corrección del error retardado.

4. RESULTADOS

4.1. Resultados del modelo VAR

Utilizando el marco metodológico descrito en la sección anterior se ha estimado un modelo VAR para las dos variables elegidas: número diario de casos nuevos y número diario de fallecidos[1]. En primer lugar, se obtiene el número de retardos que consideraremos en nuestra estimación del modelo VAR. Elegimos un número de retardos igual a 15 ya que al estimar el correspondiente modelo se obtienen los menores valores de los criterios de información de Akaike y Schwarz**.**

[1] Como se mencionó en la sección de metodología no es estrictamente necesario que las variables sean estacionarias. Independientemente de ello se ha contrastado mediante la aplicación del test de Dickey-Fuller Aumentado que las dos variables analizadas son estacionarias.

En la figura 5 se representan los valores ajustados y reales de la serie correspondiente al número diario de muertes. Cabe deducir la bondad de la estimación. En este sentido, únicamente en el máximo de la primera ola se observa una pequeña discrepancia entre los valores de ambas series.

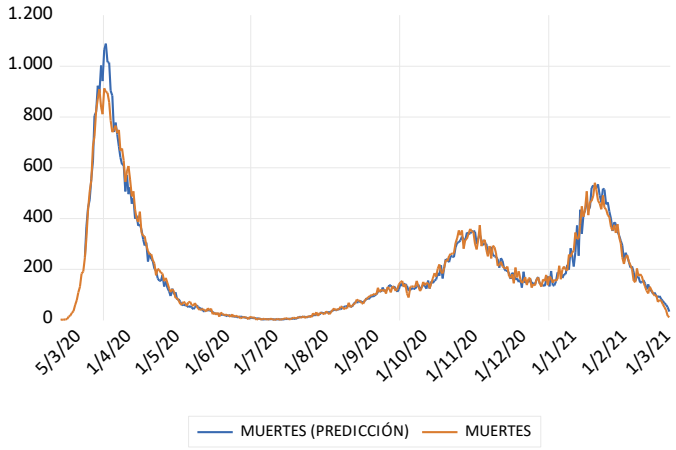

Figura 5. Comparación entre valores estimados y reales del número de muertes.

Asimismo, en las figuras 6 y 7 se muestran los valores reales y las predicciones a 10 días de las series correspondiente al número diario de casos nuevos y de fallecimientos, respectivamente. En ambos casos se observa que las predicciones tienen un comportamiento satisfactorio recogiendo en general los cambios de tendencia producidos en las series. En el caso de la serie del número de casos diarios, el valor de la predicción en el décimo día, último de la predicción, es muy similar al valor real. Asimismo, en el caso de la serie correspondiente al número de muertes, la evolución de ambas series es muy similar durante los diez días de la predicción.

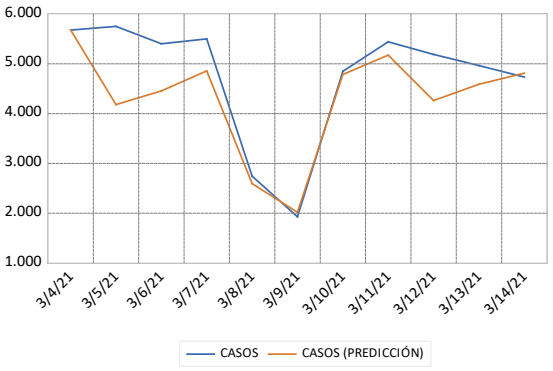

Figura 6. Predicción de casos a 10 días.

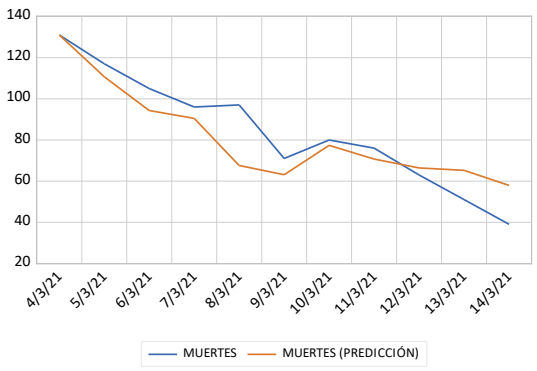

Figura 7. Predicción de muertes a 10 días.

En la Figura 8 se representa para el caso de la variable número diario de muertes, la respuesta al impulso ante variaciones equivalentes a una desviación típica de las variables. Se observa que el número diario de casos tiene un efecto creciente sobre el número de muertes, hasta alcanzar su máxima intensidad en un periodo de 15 días.

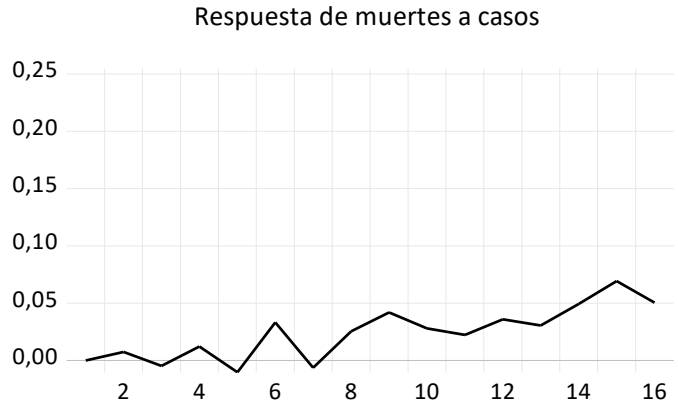

Figura 8. Respuesta del número de muertes.

5.3.3. Descomposición de la varianza

Las conclusiones anteriores se confirman con el análisis de la descomposición de la varianza que visualiza el peso de cada variable en la determinación de la desviación típica del error de predicción. En la Figura 9, que muestra la descomposición de la varianza de la variable número diario de muertes, se observa que el peso de la variable número de casos aumenta continuamente y el ritmo de crecimiento es sensiblemente mayor después de haber transcurrido 14 días.

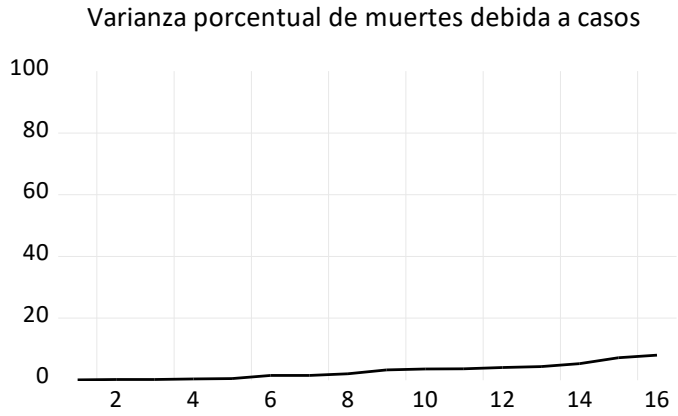

Figura 9. Descomposición de varianza del número diario de muertes.

4.2. Estimación del modelo ARDL

Utilizando el marco metodológico descrito en la sección anterior de metodología para el modelo ARDL se han estimado los efectos en el corto y largo plazo de la variable correspondiente al número diario de casos nuevos sobre la variable correspondiente al número de fallecimientos. Los resultados obtenidos se muestran a continuación.

4.2.1. Análisis de estacionaridad

En primer lugar, se contrasta la existencia de raíces unitarias con el fin de determinar el orden de integración de las series analizadas. En la Tabla 2 se muestran los resultados de aplicar el contraste de raíces unitarias de Dickey-Fuller Aumentado a las dos series objeto de análisis, tanto para las primeras diferencias como para los niveles de las variables. La hipótesis nula de una raíz unitaria (es decir no estacionaridad) es aceptada para todas las variables expresadas en niveles, pero es rechazada cuando las series están en primeras diferencias. Se concluye, por tanto, que todas las series son integradas de orden uno, I(1), o estacionarias en primeras diferencias y se rechaza la existencia de dos raíces unitarias para todas las series analizadas. Por todo ello se confirma la idoneidad del modelo ARDL.

Tabla 2. Contraste de raíces unitarias de Dickey-Fuller Aumentado (ADF)

Variables	Test de Dickey-Fuller Aumentado Z(t)				Conclusión sobre el orden de integración
	Test ADF en niveles		Test ADF en primeras diferencias		
	Z(t)	p-value	Z(t)	p-value	
muertos	-2,122	0,5311	-5,642***	0.0000	I(1)
casos	-2,891	0,1665	-4,2363**	0,0044	I(1)

Notas: Todas las variables en logaritmos. Los estadísticos del test se distribuyen como N(0,1) bajo la hipótesis de no ser estacionarias. Los símbolos *, ** y *** quieren decir significatividad a niveles del 10%, 5% y 1%.

4.2.2. Test de Cointegración

Una vez que se ha contrastado que no existe ninguna variable que sea I(2), se contrasta la existencia de una relación de cointegración mediante el test "F-Bounds". La Tabla 3 presenta los valores del estadístico F, calculados cuando la variable dependiente es el número diario de muertos, para contrastar la existencia de una relación en el largo plazo bajo la hipótesis nula, es decir, sin relación de cointegración entre las variables. Estos valores deben compararse con los límites críticos proporcionados en Pesaran et al. (2001) que se muestran también en la tabla. El estadístico F calculado es 5,329 que es mayor que el límite superior del valor crítico lo que sugiere la existencia de cointegración

entre el número de muertos y el número de casos, al rechazarse de forma concluyente la hipótesis nula (H_0) de que no existe una relación en el largo plazo al nivel del 1%.

Tabla 3. Resultados del test de cointegración F-Bounds.

Estadístico del test	Valor	Valores críticos		
		Signif.	I(0)	I(1)
F-statistic	5,329***	10%	3,02	3,51
K (número de variables)	1	5%	3,62	4,16
		2.5%	4,18	4,79
		1%	4,94	5,58

Notas: Ho: No existe relación de cointegración. Los símbolos ***, ** y * indican, respectivamente, que el parámetro es significativo estadísticamente con un nivel de confianza del 1%, 5% y 10%.

4.2.3. Test de causalidad

Habiendo encontrado que existe una relación de largo plazo entre las dos variables consideradas, el siguiente paso es probar la causalidad entre las variables. La causalidad en este caso se examina a través de la significatividad conjunta de las diferencias retardadas de las variables explicativas utilizando el test de Wald. Los resultados de este test de causalidad se presentan en la Tabla 4. La causalidad a corto plazo entre las variables número de casos nuevos y número de muertos, se contrasta mediante el valor del estadístico F correspondiente a cada una de las funciones. Como se observa en dicha Tabla, a partir de los valores obtenidos del estadístico F, no podemos rechazar la hipótesis de que la variable muertes no causa en el sentido de Granger a la variable casos, pero por el contrario sí rechazamos la hipótesis de que la variable casos no causa en el sentido de Granger a la variable muertes. Por lo tanto, cabe concluir que la causalidad se produce en un solo sentido, desde la variable casos a la variable muertes y no en el otro tal y como como se esperaba.

Tabla 4. Test de causalidad de Granger

Hipótesis Nula:	F-Statistico	p-valor
muertes no causa en sentido de Granger a casos	1,142	0,317
casos no causa en sentido de Granger a muertes	3,215***	0,000

Notas: Los símbolos ***, ** y * indican, respectivamente, que el parámetro es significativo estadísticamente con un nivel de confianza del 1%, 5% y 10%.

4.2.4. Estimación del modelo ARDL

De acuerdo con los resultados del test realizado, que confirma la existencia de cointegración, es decir, de una relación en el largo plazo entre las dos variables, procedemos a estimar las elasticidades a largo plazo y el correspondiente modelo de corrección de error. En dicho modelo se incorpora el término de corrección de error que permite el ajuste de las desviaciones de la variable dependiente respecto al valor correspondiente al equilibrio en el largo plazo. De esta forma estamos en condiciones de estimar las elasticidades en el corto y largo plazo para el número medio diario de muertos respecto al número diario de casos nuevos.

En la Tabla 5 se muestran los resultados obtenidos en la estimación del modelo ARDL(12,2) que ha sido seleccionado como el modelo óptimo mediante el Criterio de Información de Akaike (AIC) y el Criterio Bayesiano de Schwarz (SBC). Entre dichos resultados, cabe destacar que el coeficiente de ajuste es negativo y significativo estadísticamente al nivel de confianza del 1%, lo que está conforme con la existencia de una relación de largo plazo en nuestro modelo. Este coeficiente, que mide la velocidad del ajuste hacia el nivel de equilibrio después de un shock que desplace al modelo de su equilibrio, alcanza un valor de -0.041.

Asimismo, se observa que las elasticidades de la variable muertos respecto a la variable casos. tanto a corto como a largo plazo. son significativas estadísticamente al 1% y con el signo positivo previsto. Las magnitudes de dichas elasticidades son 0,177 y 1,159, respectivamente.

Tabla 5. Resultados de la estimación del modelo ARDL(12,2).

	Z(t)	T-Ratio	p-Value
Corto plazo			
Corrección de error	-0.041***	-4,01	0,000
Δcasos	0,177***	5,02	0,000
Largo Plazo			
casos	1,159***	3,81	0,000

Notas: Los símbolos ***, ** y * indican, respectivamente, que el parámetro es significativo estadísticamente con un nivel de confianza del 1%, 5% y 10%. Todas las variables consideradas están expresadas en logaritmos.

Fuente: Elaboración propia.

5. CONCLUSIONES

En este trabajo se han efectuado predicciones sobre la incidencia de la pandemia del COVID-19 en España y analizado las relaciones en el corto y largo plazo entre las variables que reflejan la evolución de la enfermedad. Las variables seleccionadas son el número medio diario de muertos y el número diario de casos nuevos. Desde la perspectiva del corto plazo, utilizando datos diarios desde el 5 de marzo de 2020 hasta el 16 de marzo de 2021, mediante la estimación del mejor modelo VAR hemos llevado a cabo la predicción de ambas variables en un horizonte de 10 días. Paralelamente, su dinámica es analizada mediante las funciones de impulso-respuesta y la descomposición de varianza. Asimismo, se ha encontrado una relación de causalidad en sentido de Granger del número diario de casos nuevos hacia el número de muertos, pero no en sentido contrario como era esperable. Finalmente, desde una perspectiva del largo plazo, el modelo ARDL estimado nos permite concluir que existe una relación de cointegración entre ambas variables. Consecuentemente, el impacto del número de casos nuevos sobre el número de muertos es estudiado mediante dicho modelo que nos facilita las correspondientes elasticidades en el corto y largo plazo.

Cabe concluir que los modelos de series temporales, tales como los que han sido utilizados en esta investigación, desempeñan un papel importante en el análisis y predicción de la pandemia del COVID-19. Los resultados de esta investigación pueden ayudar a los responsables de la política sanitaria a planificar y gestionar eficazmente los recursos disponibles, incluyendo el personal sanitario, equipamiento e instalaciones como las unidades de cuidados intensivos, que sufren el impacto de la catástrofe sanitaria producida. Debido a que la pandemia ha producido graves daños sobre todo el tejido económico y social, nuestros resultados pueden servir de ayuda para el diseño de medidas de política económica y social que contribuyan a paliar los efectos producidos.

A partir del marco metodológico de esta investigación, futuros estudios estarían dirigidos hacia la aplicación de la misma metodología para la modelización de la incidencia de la pandemia en otros países, así como la desagregación del estudio por regiones españolas. Otra línea de investigación futura consiste en la aplicación del modelo, utilizando nuevas variables como el número de personas hospitalizadas y en unidades de cuidados intensivos debido a esta enfermedad. Finalmente, se podría

extender el modelo de forma que permitiese conocer el impacto del proceso de vacunación, así como de las medidas adoptadas para contrarrestar los efectos de la pandemia.

6. REFERENCIAS BIBLIOGRÁFICAS

- ARUNKUMAR, K.E. KALAGA, D.V., KUMAR S.C.M., CHILKOOR, G., KAWAJI, M. y BRENZA, T.M. (2021). "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA)". Applied Soft Computing, 103, 107161.
- CEYLAN, Z. (2020). "Estimation of COVID-19 prevalence in Italy, Spain, and France". Science Total Environment, 729, 133817.
- DÍAZ FERNÁNDEZ, M.M., LLORENTE MARRÓN, M.M. y MÉNDEZ RODRÍGUEZ, M.P. (2014). "Un modelo vectorial autorregresivo (VAR) aplicado a la fecundidad y nupcialidad en España". Recta, 15(2), pp. 99-109.
- DICKEY, D. y FULLER, W. (1979). "Distribution of the estimators for autoregressive time series with a unit root". Journal of the American Statistical Association, 74, pp. 426-431.
- ENGLE, R.F. y GRANGER, C.W. (1987). "Co-integration and error correction: representation, estimation, and testing". Econometrica: Journal of the Econometric Society, 55, pp. 251–276.
- GNANVI, J.E., SALAKO, K.V., KOTANMI, G.B. y KAKAÏ, R.G. (2021). "On the reliability of predictions on Covid-19 dynamics: A systematic and critical review of modelling techniques". Infectious Disease Modelling, 6, pp.258-272.
- GRANGER, C.W. (1969). "Investigating causal relations by econometric models and cross-spectral methods". Econometrica, 37, pp. 424–438.
- GRANGER, C.W. (1988). "Some recent developments in a concept of causality". Journal of Econometrics, 39, pp.199–211.

- GUIRAO, A. (2020). "The Covid-19 outbreak in Spain. A simple dynamics model, some lessons, and a theoretical framework for control response". Infectious Disease Modelling, 5, pp. 652-669.
- KHAN, F., SAEED, A. y ALI, S. (2020). "Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using Vector Autoregressive model in Pakistan". Chaos, Solitons & Fractals, 140, 110189.
- MINISTERIO DE SANIDAD (2020). "Estudio ENE-COVID: Estudio nacional de sero-epidemiología de la infección por SARS-COV-2 en España". 15 de diciembre de 2020. Ministerio de Sanidad.
- NARAYAN, P.K. (2005). "The saving and investment nexus for China: evidence from cointegration tests". Applied Economics, 37(17), pp. 1979–1990.
- PESARAN, M.H. y SHIN, Y. (1999). "An autoregressive distributed-lag modelling approach to cointegration analysis". En: Strom, S. (Ed.), Econometrics and Economic Theory in the 20th Centur*y*. Cambridge University Press, Cambridge.
- PESARAN, M.H., SHIN, Y. y SMITH, R.J. (2001). "Bounds testing approaches to the analysis of level relationships". Journal of Applied Econometrics*,* 16(3), pp. 289–326.
- SIMS, C. (1980). "Macroeconomics and reality". Econometrica*,* 48, pp. 165-192.
- SIMS, C., STOCK, J. y WATSON M. (1990). "Inference in Linear Time Series Models with Some Unit Roots". Econometrica, 58(1), pp.113-144.
- WANG, Q., ZHOU, Y. y CHEN, X. (2021). "A Vector Autoregression Prediction Model for COVID-19 Outbreak". arXiv preprint arXiv:2102.04843.