Forecasting número de nuevos fallecimientos semanales atribuidos por COVID-19 en México

(Forecasting number of new weekly deaths attributed to COVID-19 in Mexico)

Enrique Hernández-Laredo
Facultad de Medicina Universidad Autónoma del Estado de México
Toluca de Lerdo, México
ehernandezl190@alumno.uaemex.mx

I. INTRODUCCION.

El coronavirus SARS-CoV-2, causante de la enfermedad COVID-19, es capaz de generar en humanos infección, fiebre, tos, dificultad para respirar e irritación gastrointestinal, y en ciertos casos, particularmente en personas mayores o/e inmunocomprometidas, las infecciones por coronavirus pueden provocar neumonía grave y, posteriormente, la muerte del paciente(1). De este modo, la mayoría de las regiones del mundo, están utilizando nuevas formas de análisis de la población para ayudar a los estudios epidemiológicos y los estudios de predicción con el fin de controlar eficazmente la propagación viral (1,2).

El empleo de modelos estadísticos predictivos en las ciencias de la salud ha crecido en los últimos años. Estos emergen como un vínculo importante entre la estadística y la práctica médica; son de gran ayuda en la toma de decisiones y permiten la creación de diversos sistemas y herramientas útiles para reducir las incertidumbres, garantizar mejores actuaciones y establecer eficaces medidas de control para la erradicación de las enfermedades (3–5), y en los últimos meses, los investigadores han venido empleando métodos matemáticos para poder pronosticar el número de casos de COVID-19 en todo el mundo. El Método Autorregresivo Integrado de Medias Móviles (ARIMA) es el que más se ha utilizado para realizar dicho pronósticos(4).

El propósito del presente trabajo es realizar un modelo ARIMA mediante la metodología Box- Jenkins (6) que permita pronosticar el número de fallecimientos ocasionados por COVID-19 en México.

II. METODOLOGÍA.

A. Base de datos:

Se utilizó la base de datos "*owid-covid-data*" del repositorio público de *Our World in Data* (7), que contiene información del área de la salud sobre COVID-19 referente a vacunas, pruebas y positividad, pacientes en hospitales, casos confirmados, muertes confirmadas, tasa de reproducción del COVID-19 y respuestas de políticas de al menos 47 países, incluyendo México.

B. Preparación de los datos:

Se filtró la base de datos para incluir únicamente la serie de tiempo con información del número de nuevos fallecimientos atribuidos por COVID-19 en México, dicha serie contiene observaciones diarias desde el 1 de enero del 2020 hasta la fecha actual (esto dependerá de la última fecha de actualización de la base de datos "*owid-covid-data,* pero la última fecha incluida para el análisis reportado fue el 29 de mayo de 2022).

Se realizó un análisis de la calidad de los datos, encontrando que existían 87 valores nulos, los cuales se reemplazaron de la siguiente manera:

$$Y_{(t_0)} = 0$$
$$if(Y_{(t)}\ and\ Y_{(t+1)} == NAN) \leftarrow Y_{(t)} = Y_{(t-1)}$$
$$if(Y_{(t)} = NAN\ and\ Y_{(t+1)} \neq NAN) \leftarrow Y_{(t)} = \frac{Y_{(t-1)} + Y_{(t+1)}}{2}$$

Donde, $Y_{(t)}$ corresponde a la serie de tiempo con información del número de nuevos fallecimientos diarios atribuidos por COVID-19 en México, para un tiempo $t = 0,1,2,\dots n$. Donde n es el número total de muestras.

Posteriormente, se calcular los promedios semanales a partir de los datos diarios, y se redondearon al número entero más próximo, esto con el fin de cambiar la temporalidad de la serie de datos diarios a semanales, a partir de este punto la serie de tiempo $Y_{(t)}$ es considera con temporalidad semanal.

C. Análisis exploratorio de datos:

Se realizó un análisis descriptivo de la serie de tiempo $Y_{(t)}$ agrupada por mes y año, mediante los valores mínimos, máximos, media y desviación estándar (STD).

Para estudiar la distribución de los datos se ejecutó el estadístico Jarque-Bera a un nivel de significancia del 5% y se elaboraron un histograma y un grafico Q-Q.

Posteriormente, se realizó la descomposición de la serie de tiempo en sus componentes estacionales, aleatorio y tendencia.

D. Modelado de la serie de tiempo.

Se utilizó la metodología Box-Jenkins no estacional para elegir el modelo que mejor represente a la serie de tiempo $Y_{(t)}$, en la Fig. 1 se muestra las etapas correspondientes con dicha metodología.

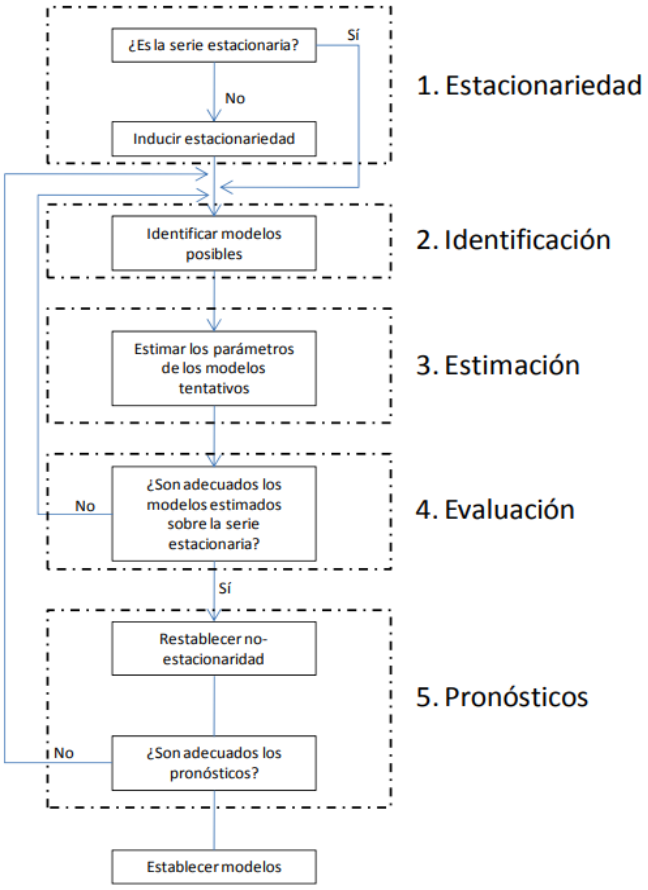

Fig. 1 Etapas de la metodología Box-Jenkins (6).

D.1) Estacionariedad

Se calculó la serie de tiempo en primera diferencia $I(1)$ a partir de la serie de tiempo a niveles $I(0)$, dada la siguiente transformación lineal:

$$I(1) = I(0) - Y_{(t-1)} \quad (1)$$

Donde, $I(0) = Y_{(t)}$, y $t = 1,2,3 \ldots n$. La estacionariedad de las series de tiempo $I(0)$ e $I(1)$, se estudió mediante la función de autocorrelación simple (ACF) y función de correlación parcial (PACF), los estadísticos Dickey-Fuller aumentada, Phillips-Perron, y KPSS, los cuales fueron ejecutadas mediante un nivel de significancia del 5%.

D.1.1) Estacionalidad

Dado que de manera empírica y visual la serie de tiempo $I(0)$ presenta tendencias en ciertos periodos de tiempo mensuales atribuidos a picos máximos locales en el número de fallecimientos por COVID-19, se estudió la estacionalidad de la serie de tiempo $I(1)$. Se ejecutó la prueba Dickey-Fuller estacional bajo un nivel de significancia del 5%, utilizando la serie $Y_{(t)}$, la serie $Y_{(t)}$ con rezago 52, y la serie $Y_{(t)}$ con diferenciada en t_{52}.

D.2) Identificación:

Posterior al cumplimiento del supuesto de estacionariedad y verificación de la no estacionalidad, se utilizaron los gráficos de ACF y PACF de la serie de tiempo $I(1)$ para comprar los mismos, con los ACF y PACF teóricos de los modelos autorregresivo integrado de promedio móvil (ARIMA), autorregresivo puro (AR) y medias móviles (AM).

El orden de autorregresivo (p) de los modelos ARIMA(p,d,q,) y AR(p), fueron seleccionados a partir de los primeros cinco retardos con correlación significativa dado el gráfico ACF de la serie de tiempo $I(1)$. Mientras que el orden q de los modelos ARIMA(p,d,q,) y MA(q) fueron seleccionados a partir de los primeros cinco retardos con correlación significativa dado el grafico PACF de la serie de tiempo $I(1)$. El orden de integración d fue definido por el número de diferencias necesarias en $I(0)$ para cumplir con el supuesto de estacionariedad. Además, se tomaron en cuenta los ordenes p,q,d calculados de manera automático por la función auto.arima() en la libreria forecast (8) de Rstudio.

D.3) Estimación:

Se estimaron todos los modelos ARIMA(p,d,q) y AR(p) posibles, dada las combinaciones de los órdenes p, d, q identificados por los gráficos ACF y PACF previamente.

Posteriormente se realizó un *benchmarking* de todos los modelos basado en error cuadrático medio (RMSE) y el criterio de información de Akaike (AIC).

Se calcularon los coeficientes del mejor modelo y su significancia.

D.4 Evaluación

Posterior a la selección del mejor modelo, se verificó su estabilidad mediante las raíces de su polinomio. Además, se calcularon sus residuos y se verificó que cumplieran con los supuestos de no autocorrelación serial, homocedasticidad de varianzas, normalidad, y estacionariedad.

D.4.1) autocorrelación serial

Se ejecutaron las pruebas Ljung-Box y Box-Pierce mediante un nivel de significancia del 5% a los residuales del mejor modelo.

D.4.2) homocedasticidad de varianzas

Se ejecuto prueba de efecto de heterocedasticidad autorregresiva condicional automática (ARCH) a nivel de significancia del 5%. Además, se calculó el cuadrado de los residuos para ejecutar la prueba Box-Ljung mediante un nivel de significancia del 5%, y se graficó el ACF del cuadrado de los residuales.

D.4.3) normalidad

Se verificó la normalidad de los residuales mediante las pruebas Jarque-Bera y Shapiro a un nivel de significancia del 5%.

D.4.3) estacionariedad

La estacionariedad de los residuos se verificó mediane los estadísticos Dickey-Fuller aumentada, Phillips-Perron, y KPSS, los cuales fueron ejecutadas mediante un nivel de significancia del 5%.

E. Pronostico:

Se graficaron los datos de la serie de tiempo $Y_{(t)}$ y los valores estimados por el mejor modelo.

Por último, se realizó un *forecasting* de las siguientes 5 semanas, y se calculó el intervalo de confianza al 95% del pronóstico.

III. RESULTADOS

A. Preparación de los datos:

En la *Fig. 2* se muestra la serie de tiempo con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México sin valores nulos ($Y_{(t)}$)

Fig. 2 Serie de tiempo a niveles I(0) con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México.

B. Análisis exploratorio de datos:

En la

Tabla *1* se muestra los estadísticos descriptivos de la serie de tiempo $Y_{(t)}$ agrupados por año, mientras que la Fig. 3 se muestra su respectivo gráfico de caja y bigotes. Por otra parte, en la

Tabla *2* se muestran los estadísticos descriptivos por mes y en la Fig. 4 su gráfico de caja y bigotes.

Se observa que los meses con mayor desviación estándar son enero y febrero, mismo que contiene el mayor registro de número de fallecimientos en una semana, con 1275 y 1144 fallecimientos respectivamente (Ver Fig. 4 y

Tabla *2*).

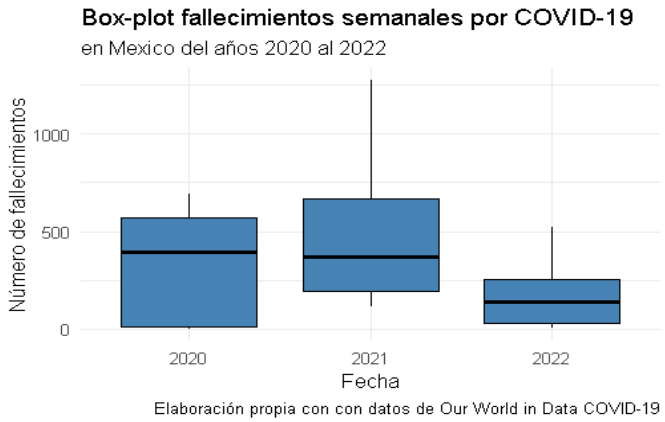

Fig. 3 Gráfico de caja y bigotes del número de fallecimientos semanales por COVID-19 en México agrupado por año.

Tabla 1 Número de fallecimientos semanales por COVID-19 en México agrupados por año.

Año	Minimo	Máximo	Media ± STD
2020	0	689	336 ± 255
2021	112	1275	462 ± 314
2022	3	178	178 ±161

Tabla 2 Número de fallecimientos semanales por COVID-19 en México agrupados por mes.

Mes	Minimo	Máximo	Media ± STD
Enero	0	1275	414 ± 511
Febrero	0	1144	469 ± 426
Marzo	0	698	241 ± 243
Abril	11	742	180 ± 223
Mayo	3	362	184 ± 118
Junio	148	689	361 ± 232
Julio	192	642	453 ± 187
Agosto	484	716	599 ± 91
Septiembre	380	711	505 ± 112
Octubre	268	499	360 ± 75
Noviembre	183	568	184 ± 118
Diciembre	112	684	389 ±242

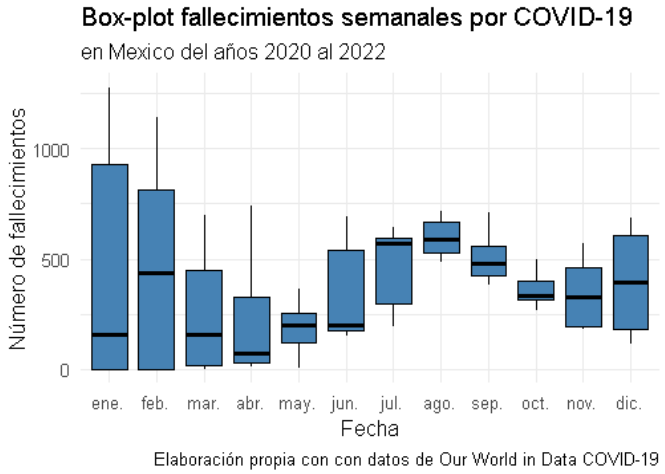

Fig. 4 Gráfico de caja y bigotes del número de fallecimientos semanales por COVID-19 en México agrupado por año.

Un p-valor=0.000 correspondiente de la prueba Jarque-Bera, indica que existe evidencia suficiente para rechazar la hipótesis nula de normalidad en los datos, este resultado se puede complementar con el histograma de la Fig. 5 donde se aprecia que los datos tienen un sesgo positivo. De igual forma los datos no se ajustan a la linea teoria del que gráfico Q-Q (ver Fig. 6), por lo que los datos de la serie de tiempo $Y_{(t)}$ no se distribuyen de manera normal.

La descomposición de la serie de tiempo $Y_{(t)}$ se muestra en la Fig. 7, en ella se observa una tendencia a largo plazo a la baja, además que el componente estacional tiene mayor peso que el aleatorio.

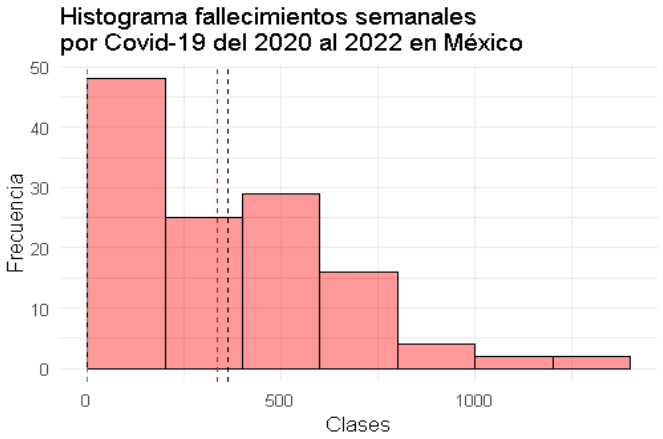

Fig. 5 Histograma serie de tiempo a niveles I(0) con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México. Linea punteada azul: moda, linea punteada café: mediana, linea punteada negra: media.

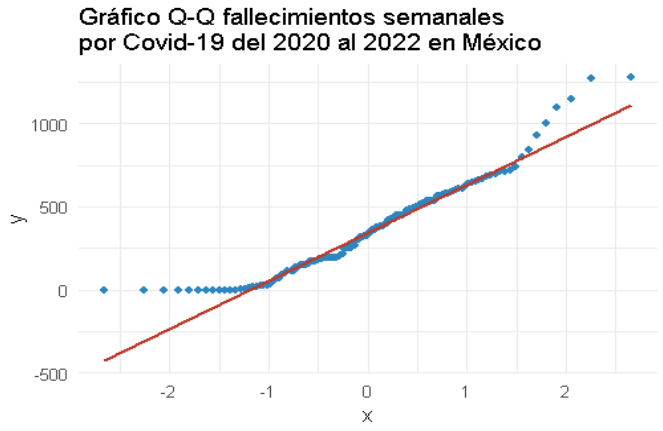

Fig. 6 Gráfico Q-Q serie de tiempo a niveles I(0) con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México

C. Modelado de la serie de tiempo.

C.1) Estacionariedad

En al Fig. 8 se muestra el resultado de calcular la primera diferencia de la serie de tiempo $Y_{(t)}$. Mientras que en la Tabla 3 se observa el resultado de aplicar las pruebas de estacionariedad a la serie de tiempo a niveles $I(0)$ y en primera diferencia $I(1)$.

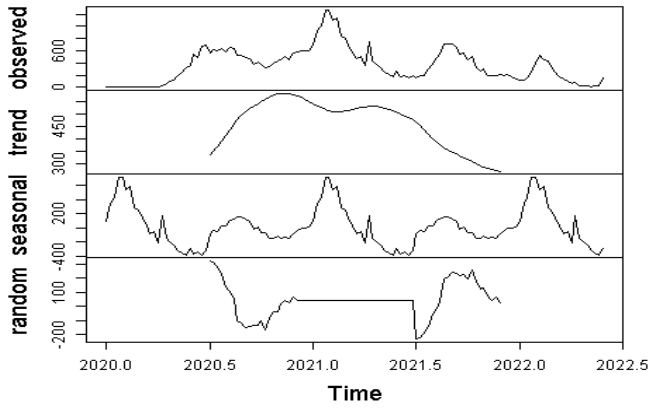

Fig. 7 Descomposición de la serie de tiempo a niveles I(0) con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México.

Tabla 3 Pruebas estadísticas de estacionariedad

Prueba	$I(0)$	$I(1)$
	p-valor	
Dickey-Fuller aumentada	0.2317	0.01
Phillips-Perron	0.5503	0.01
KPSS	0.01	0.1

En la Tabla 3 se muestra como la serie de tiempo $I(1)$ es la única que satisface el supuesto de estacionariedad, pues las pruebas basadas en medias (Dickey-Fuller aumentada y Phillips-Perron) poseen un valor menor al nivel de significancia de 0.05, por lo que existe evidencia suficiente para rechazar la hipótesis nula de no estacionariedad, por otra parte la prueba KPSS basada en varianzas tiene un p-valor 0.1 mayor al nivel de significancia del 0.05, con lo que no hay evidencia suficiente para rechazar la hipótesis nula de estacionariedad. Esto se ve complementado con los gráficos ACF y PACF, pues existe una caída exponencial en los primeros rezagos en $I(1)$, en comparación con la caída en forma de seno amortiguada en $I(0)$ (Ver Fig. 9 y Fig. 10).

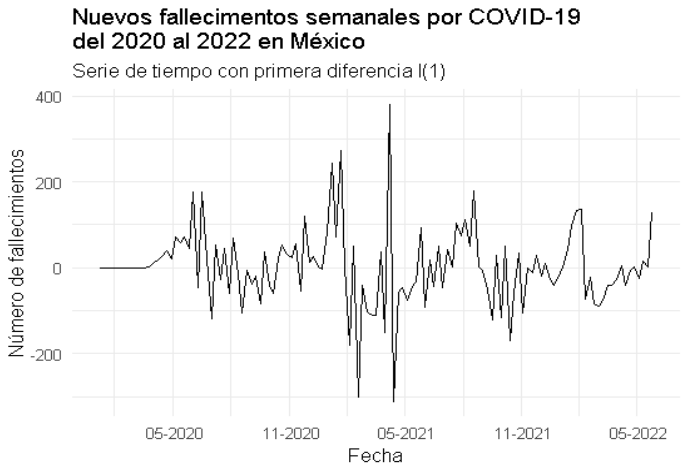

Fig. 8 Serie de tiempo en primera diferencia I(1) del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México

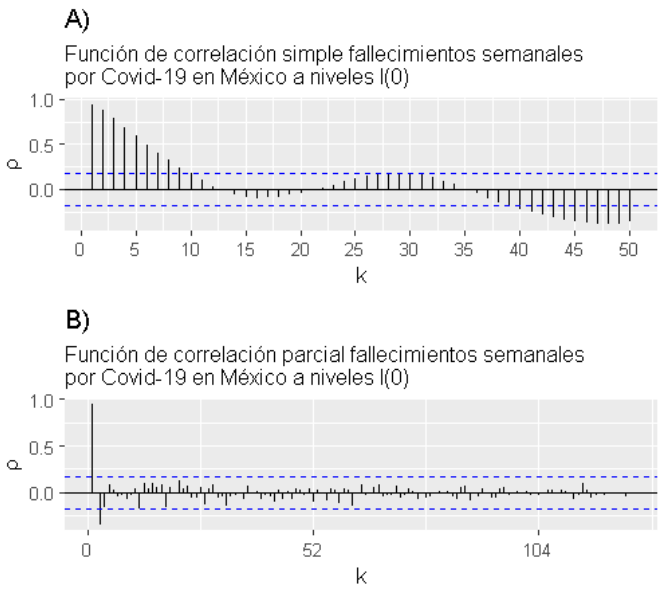

Fig. 9 Funciones de correlación de la serie de tiempo a niveles I(0) con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México, A) ACF,B) PACF

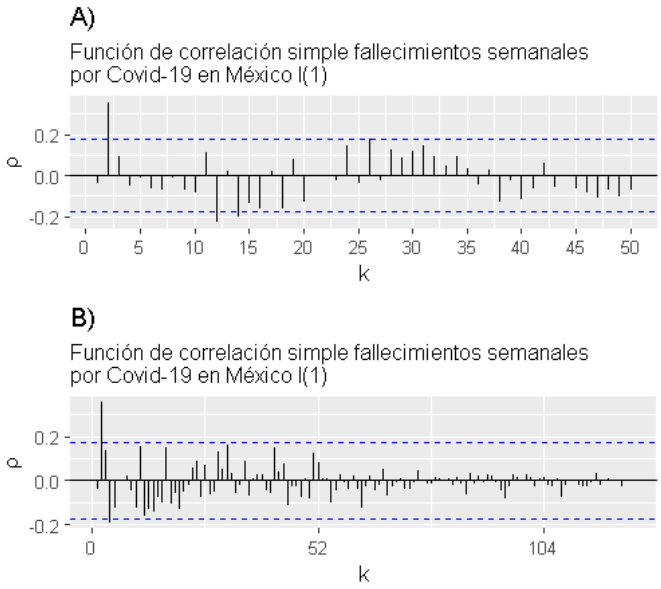

Fig. 10 Funciones de correlación de la serie de tiempo en primera diferencia I(1) con información del número de nuevos fallecimientos semanales atribuidos por COVID-19 en México, A) ACF,B) PACF

C.1.1) Estacionalidad

En la Tabla 4 se muestra el resultado de los coeficientes y sus respectivos p-valores obtenidos de aplicar el test Dickey-Fuller estacional bajo un nivel de significancia del 5% .

Tabla 4 Dickey-Fuller estacional.

Coeficiente	Estimación	p-valor
Intercepto	567.2237	0.000
$Y_{t_{(52)}}$	-1.4280	0.000

Dado que el p-valor del coeficiente $Y_{t_{(52)}}$ es asintóticamente igual a cero, y contrastado con un valor de significancia del 0.05, existe evidencia suficiente para concluir que la serie no se comporta de forma estacional.

C.2) Identificación:

De los grafico ACF y PACF de la serie de tiempo $I(1)$ se seleccionaron los órdenes p=2,4 d=1, q=1 (Ver Fig. 10). De manera que se estimaron los siguientes modelos.

Tabla 5 Estimación de modelos ARIMA

Modelo	AIC	RMSE
ARIMA(2,1,1)	1471.01	83.76
ARIMA(4,1,1)	1467.44	81.18
ARIMA(2,1,0)	1470.06	84.12

Donde ARIMA(2,1,0) fue estimado mediante la función auto.arima(). Dado los valores de AIC y RMSE el modelo ARIMA(4,1,1) fue seleccionado como el que mejor describe a la serie de tiempo $I(1)$. En la Tabla 6, se muestra los coeficientes estimado por dicho modelo.

Tabla 6 Modelo ARIMA(4,1,1)

Coeficiente	Estimación	Error estándar	p-valor
AR(1)	0.3740	0.3247	0.2494
AR(2)	0.4638	0.0934	0.000
AR(3)	-0.0313	0.1446	0.8284
AR(4)	-0.2553	0.0860	0.0029
MA(1)	-0.4439	0.3353	0.1855

C.4 Evaluación

En la Fig. 11 se muestra las raíces inversas del polinomio del modelo ARIMA(4,1,1), todas las raíces inversas del polinomio caen dentro del círculo unitario del plano Z, lo que significa que el sistema es estable.

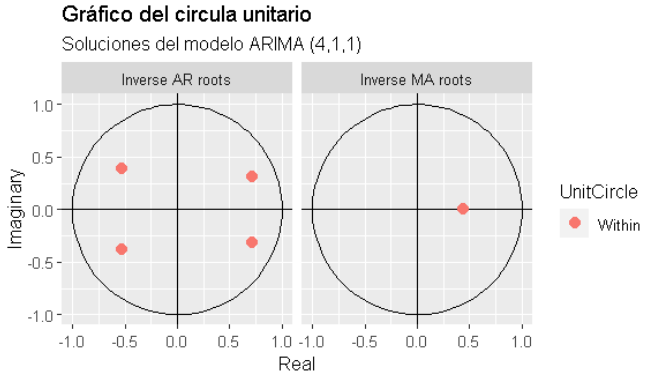

Fig. 11 Estabilidad del modelo ARIMA(4,1,1)

C.4.1) Autocorrelación serial

A pesar de que el gráfico ACF muestras los rezagos 12 y 13 fuera del intervalo de confianza (Ver Fig. 10), las pruebas estadísticas Box-Pierce y Ljung-Box, indican que los residuos presentan ausencia de autocorrelación serial, pues los p-valores calculador son 0.9734 y 0.9742 respectivamente. Por lo tanto, no hay evidencia para rechazar H0 de no autocorrelación serial.

D.4.2) Homocedasticidad de varianzas

En la Fig. 12 se muestran los resultados de las pruebas Portmanteau-Q y Lagrange-Multiplier, contenidos en la prueba ARCH, en donde se muestra a y través del test Portmanteau-Q que el modelo posee heterocedasticidad en la varianza de los residuales, por otra parte , la prueba Lagrange-Multiplier indica que la heterocedasticidad solo se presenta en los primeros 8 rezagos.Por otra parte, un p-valor=0.00163 de la prueba Box-Ljung y el grafico ACF (Ver Fig. 13) refuerza la existencia de heterocedasticidad en la varianza de los residuales.

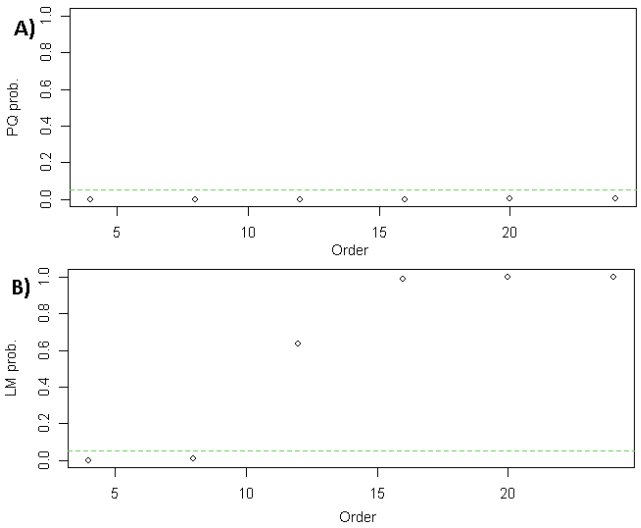

Fig. 12 prueba ARCH, A) Portmanteau-Q, B) Lagrange-Multiplier

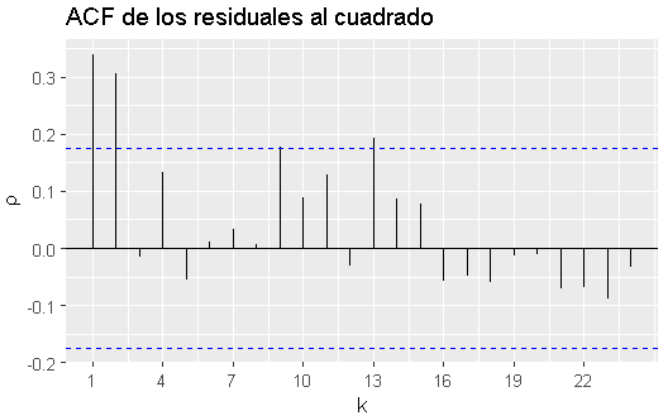

Fig. 13 Función de correlación simple en los residuos al cuadrado

C.4.3) Normalidad

Con p-valores de $2.764x10^{-14}$ y $1.168x10^{-05}$ en las pruebas Jarque-Bera y Shapiro respectivamente, se asume que los residuales no se distribuyen de manera normal.

C.4.3) estacionariedad

En la Tabla 3 se observa el resultado de aplicar las pruebas de estacionariedad a los residuos del modelo ARIMA(4,1,1). De este modo, los p-valores permiten asumir que existe estacionariedad en los residuos.

Tabla 7 Estacionariedad en residuos

Prueba	p-valor
Dickey-Fuller aumentada	0.01
Phillips-Perron	0.01
KPSS	0.1

D.5) Pronostico:

En la Fig. 14 se muestran los datos de la serie de tiempo $Y_{(t)}$ y los valores estimados por el modelo ARIMA(4,1,1). Además, en la Fig. 15 se grafica el *forecasting* correspondiente a los fallecimientos pronosticados por COVID-19 en México de las siguientes 5 semanas, así como su intervalo de confianza al 95%.

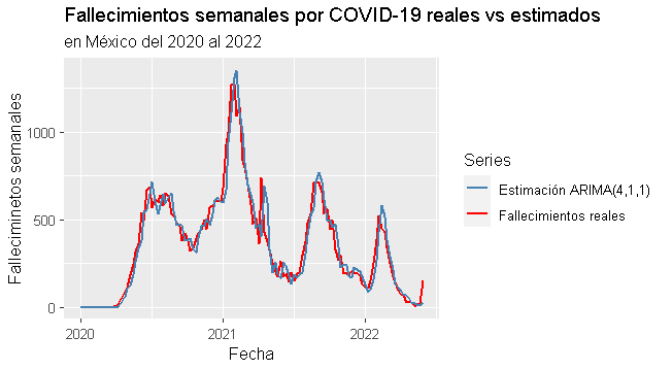

Fig. 14 Ajuste de los valores estimados vs valores reales del número de fallecimientos por COVID-19 en México.

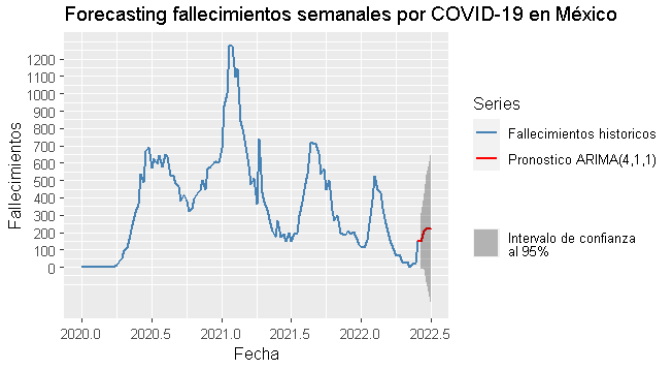

Fig. 15 Forecasting fallecimientos semanales por COVID-19 en México del mes de junio y la primera semana de julio del año 2022.

IV. CONCLUSIONES

Los pronósticos obtenidos mediantec el modelo ARIMA(4,1,1), comparados con los datos históricos reales, muestran un ajuste adecuado. Sin embargo, la principal limitante del modelo ARIMA(4,1,1) es la presencia de heterocedasticidad en los residuos, por lo que se sugiere utilizar un modelo de tipo ARCH. A pesar de ello se sugiere que el modelo ARIMA(4,1,1) presentado en este trabajo puede considerarse una herramienta simple e inmediata para aproximar el número de fallecimientos por COVID-19 en México.

V. ANEXO

El código para realizar el pronóstico del número de nuevos fallecimientos por COVID-19 en México basado en un modelo ARIMA(4,1,1), se encuentra disponible en: Github-enriquehdez98

VI. REFERENCIAS

1. Sharma A, Ahmad Farouk I, Lal SK. COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission, Detection, Control and Prevention. Viruses [Internet]. 2021 Feb 1 [cited 2022 May 29];13(2). Available from: /pmc/articles/PMC7911532/

2. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases [Internet]. 2020 May 1 [cited 2022 May 29];20(5):553. Available from: /pmc/articles/PMC7158569/

3. Felipe J, Mendieta M, Cortés Cortés ME, Cortés Iglesias M, del Carmen Pérez Fernández A, Cabrera MM. Estudio sobre modelos predictivos para la COVID-19 en Cuba Study on predictive models for COVID-19 in Cuba. [cited 2022 May 29]; Available from: http://medisur.sld.cu/index.php/medisur/article/view/4703

4. Sotomayor DAC, Carlos FBSM, Sotomayor DAC, Carlos FBSM. Aplicación del método autorregresivo integrado de medias móviles para el análisis de series de casos de covid-19 en Perú. Revista de la Facultad de Medicina Humana [Internet]. 2021 Jan 12 [cited 2022 May 29];21(1):65–74. Available from: http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2308-05312021000100065&lng=es&nrm=iso&tlng=es

5. León-Álvarez AL, Betancur-Gómez JI, Jaimes-Barragán F, Grisales-Romero H, León-Álvarez AL, Betancur-Gómez JI, et al. Ronda clínica y epidemiológica. Series de tiempo. Iatreia [Internet]. 2016 Jul 1 [cited 2022 May 29];29(3):373–81. Available from: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-07932016000300373&lng=en&nrm=iso&tlng=es

6. Modelación ARIMA. [cited 2022 May 29]; Available from: http://www.ptolomeo.unam.mx:8080/jspui/bitstream/132.248.52.100/363/7/A7.pdf

7. Appel Cameron, Beltekian Diana. Data on COVID-19 (coronavirus) by Our World in Data [Internet]. 2022 [cited 2022 May 20]. Available from: https://github.com/owid/covid-19-data/tree/master/public/data

8. auto.arima function - RDocumentation [Internet]. [cited 2022 May 29]. Available from: https://www.rdocumentation.org/packages/forecast/versions/8.16/topics/auto.arima