Ajuste de un modelo ARIMA para los precios de exportación del aguacate calibre 60

Belem A Aguilar[1], Roberto S. Acosta Abreu[2]
[1]Licenciatura en Física y Matemáticas, ESFM-IPN, México D.F., México
Teléfono (55) 3494-9649 E-mail: laguilara1700@alumno.ipn.mx
[2]Departamento de Matemáticas, ESF-IPN, México D. F., México
Teléfono (55) 5729-6000 Ext. 55011 Fax (55) 5729-55015 E-mail: racosta@esfm.ipn.mx

Resumen — **El propósito de este documento es analizar los precios de exportación a EE.UU. del aguacate calibre 60, desarrollar de manera resumida la teoría de los modelos ARIMA a la vez que se va obteniendo un modelo resultante de la metodología Box-Jenkins no estacional. El documento consiste en las justificaciones necesarias para el uso de la teoría elegida, el análisis de sus elementos; comparación de modelos AR, MA, ARMA y finalmente ARMA y ARIMA. Obteniendo un modelo ARIMA (2,1,0) cuya confiabilidad se limita a un par de pasos por delante.**

Palabras Clave – **serie de tiempo, metodología Box-Jenkins no estacional, modelo ARIMA no estacional.**

Summary — **The purpose of this document is to analyze the export prices of 60 caliber avocados to the US, to briefly develop the theory of ARIMA models while obtaining a model resulting from the non-seasonal Box-Jenkins methodology. The document consists of the necessary justifications for the use of the chosen theory, the analysis of its elements; comparison of AR, MA, ARMA and finally ARMA and ARIMA models. Obtaining an ARIMA model (2,1,0) whose reliability is limited to a couple of steps ahead**.

Keywords **– time series, non-seasonal Box-Jenkins methodology, non-seasonal ARIMA model.**

I. INTRODUCCIÓN

Para este documento se utilizaron programas como R para procesar, visualizar y analizar los precios de exportación a EE.UU. del aguacate calibre 60. Los datos fueron obtenidos de la revista virtual MODULA del Consejo Nacional de Productores de Aguacate A.C. (CONAPA) [1]. Se escogieron este conjunto de precios debido a una enorme cantidad de factores que los vuelven estocásticos entre temporadas, además de que estos varían de acuerdo al peso del aguacate. Se eligió un calibre estándar (calibre 60, aguacates que pesan entre 170-205g) con la esperanza de que este modelo guíe a un modelo para cada uno de los demás calibres.

Los datos comprendieron un periodo de tres años, comenzando en enero del año 2019 y concluyendo a finales de diciembre del año 2021.

[2] Este trabajo se realizó con apoyo del programa EDD y de la COFAA del IPN.

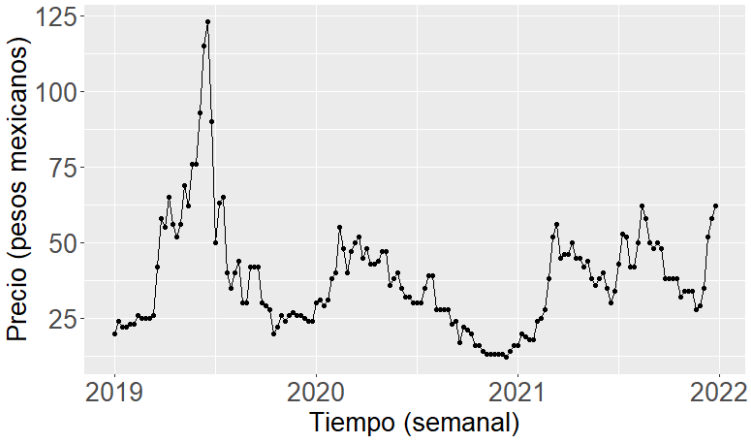

Fig. 1. Gráfica de datos finales seleccionados

II. METODOLOGÍA

A. *Visualización y limpieza de datos.*

Los datos de fecha y precio se tomaron en un principio a partir de enero 2014, hasta diciembre 2021, en intervalos de una semana, esperando un registro de 52 datos por año. Sin embargo, se encontraron con varias fechas que no se encontraban disponibles tanto en la fuente mencionada como en otros registros, dejando bastantes datos desconocidos.

Así pues, se limpiaron los datos descartando los años que tuvieran un considerable número de huecos, con lo cual se redujeron al intervalo de enero 2019 hasta diciembre 2021 [2]

La Fig.1 muestra la gráfica de los datos finales que se analizarán en este documento.

B. *Justificación del uso de series de tiempo.*

Entendemos por una *serie de tiempo discreta* a los datos recopilados, medidos o registrados de forma secuencial en el tiempo o en un intervalo fijo [3][4]. Se le denomina *análisis de series de tiempo* al enfoque sistemático mediante el cual se va a responder las cuestiones matemáticas y estadísticas planteadas por estas correlaciones de tiempo [5].

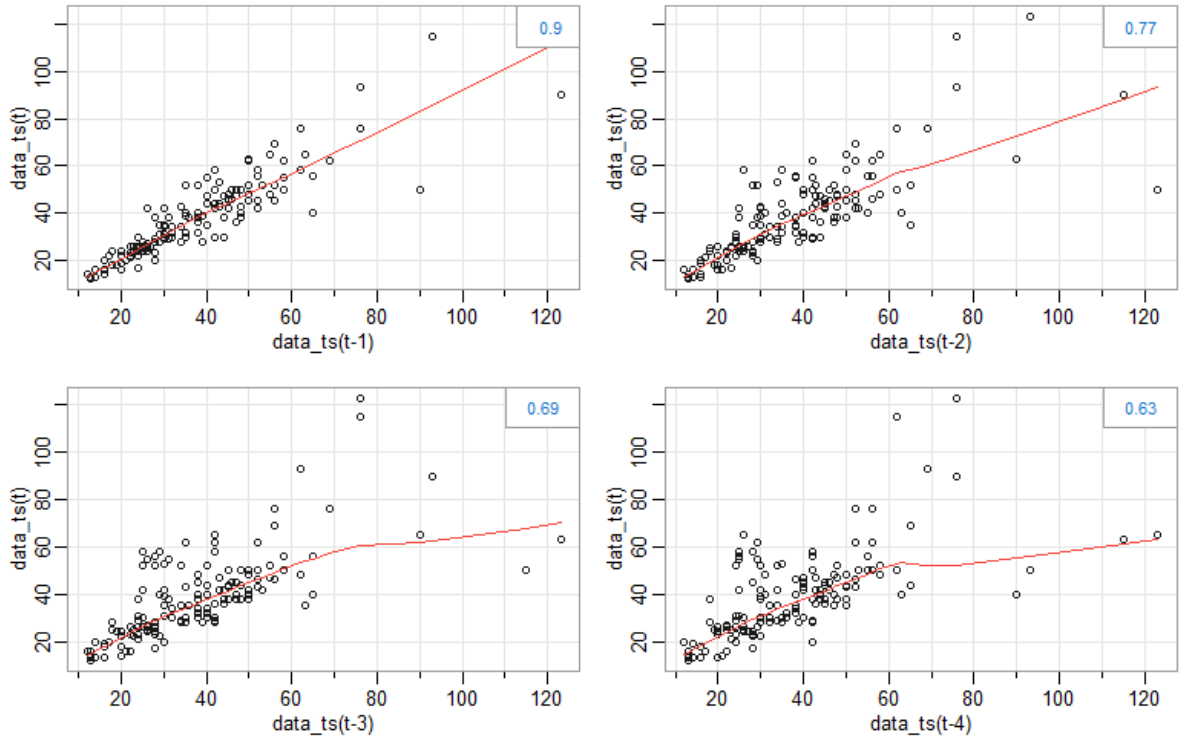

Fig. 2. Gráficos de dispersión de hasta 4 posiciones atrás

Nuestro acercamiento es de naturaleza estadística, estudiando la serie histórica de datos y ajustando un primer modelo.

Así, se desea estudiar el pasado para predecir el futuro. Para corroborar que los datos tienen relación con su histórico, se observaron sus gráficos de dispersión mostrados en la Fig.2 en donde se tiene los precios de la posición actual contra los precios de hasta 4 posiciones atrás. Notando que, dada un precio en el tiempo, este llega a tener correlación positiva mayor a 0.5 aún con el precio de un mes pasado (equivalente a 4-5 semanas).

Concluyendo de estos gráficos que los históricos se prestaron para ajustarles un modelo de series de tiempo.

C. Descomposición de la serie de precios

Las principales características de las series de tiempo son el *ciclo*, la *tendencia* y *variaciones estacionales*; definiendo la primera como una secuencia de eventos repetibles en el tiempo, donde un punto inicial de un ciclo es un mínimo local de la serie y el final es el siguiente [6], la segunda como un cambio sistemático que no aparenta ser periódico (indica la dirección general de la serie [6]) y la última como un patrón repetitivo generalmente anual [3].

Al procedimiento de descripción de estas características en una serie de tiempo se le conoce como *descomposición*. Esta permite establecer el tipo de modelo a requerir.

Los ciclos que se observaron, mostrados en la Fig. 1 son cuatro:

I. Desde inicios del 2019 hasta el tercer trimestre del mismo.
II. Desde el punto anterior hasta casi inicios del 2021
III. Desde el punto anterior hasta mediados del 2021
IV. Desde el punto anterior hasta finales del 2021

Haciendo una observación en que para inicios del 2022 la serie se encuentra en lo que parece ser un nuevo ciclo.

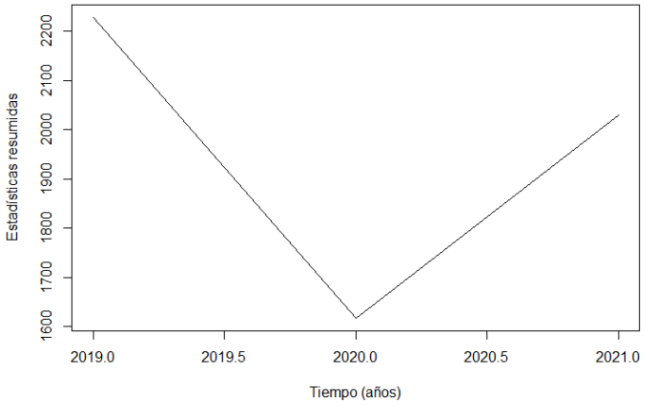

Fig. 3. Tendencia anual de los datos

Un *modelo aditivo* resulta útil cuando la variación estacional es relativamente constante en el tiempo; mientras que uno *multiplicativo* es utilizado cuando la variación estacional aumenta en el tiempo [7]. Considerando esto, para una vista más clara de la tendencia en nuestros datos, se eliminaron los efectos estacionales; obteniendo la Fig. 3, que muestra la tendencia anual de la serie de tiempo mediante la división de los datos en subconjuntos, calculando las estadísticas resumidas para cada uno graficando esto en una forma conveniente [8]. Esto indicando un comportamiento multiplicativo para posibles modelos que se especificaron.

A su vez, un resumen de los datos para cada semana, mostrados en la Fig.4: una gráfica de cajas que visualizaron de mejor manera los efectos estacionales de nuestra serie.

Las estacionalidades nos indicaron un alza de precios en las últimas semanas de Febrero (semana 8) y bajas a mediados de Mayo (semana 20), volviendo a subir a finales de Julio (semana 28) para volver a bajar a finales de Septiembre (semana 39).

Todos estos comportamientos observados en nuestra serie de tiempo resultaron útiles para la detección de posibles modelos.

D. Metodología Box-Jenkins no estacional

Existen varias metodologías que ayudan a elegir el modelo que mejor se ajusta a la serie de tiempo entre los diversos modelos que pueden existir. La metodología más usada y difundida es la que propusieron los profesores G.E.P. Box y J.M. Jenkins en la década de los 1970's, en la cual lograron varios avances en identificar, ajustar y verificar los modelos ARIMA apropiados. Comúnmente conocida como *Metodología Box-Jenkins* [9].

Antes de mencionar esta metodología, se hace énfasis en algunas consideraciones importantes sobre la naturaleza de los datos para la modelación ARIMA [10]:

- Los modelos ARIMA aplican tanto para modelos discretos como continuos, pero solo se puede aplicar a datos equidistantes en el tiempo, en intervalos discretos en el tiempo.
- Para la elaboración de un modelo ARIMA se requiere una cantidad mínima de datos. Se sugiere una cantidad mínima de 50 datos.
- Los modelos ARIMA son especialmente útiles en el tratamiento de series que presentan patrones estacionales.
- Los métodos Box-Jenkins aplican a series estacionarias y no estacionarias. Una *serie estacionaria* es aquella cuya media, varianza y función de auto correlación permanecen constantes en el tiempo.
- Se asume que las perturbaciones aleatorias presentes en la serie, son independientes entre sí. No existe correlación entre ellas, por lo tanto, ningún patrón modelable.

La metodología consta principalmente de 5 etapas [11]:

1. Estacionariedad
2. Identificación
3. Estimación
4. Evaluación
5. Pronóstico

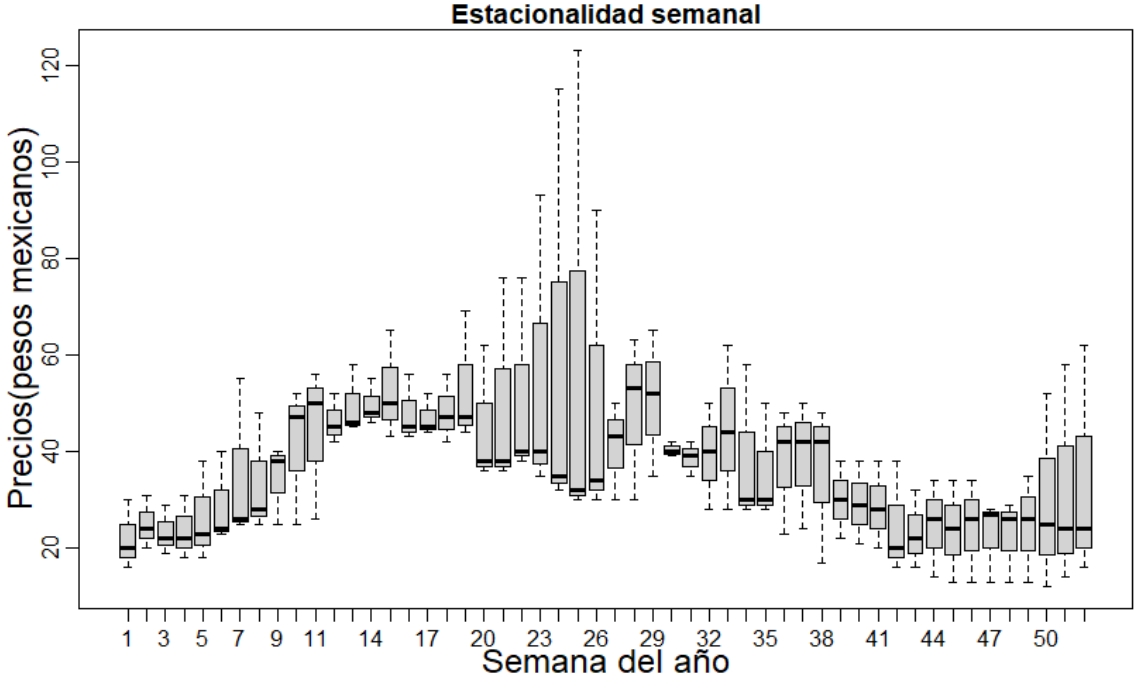

Fig. 4. Estacionalidades de cada semana

Estas se pueden apreciar en la Fig.5 y estas se desarrollarán en las siguientes secciones.

E. Estacionariedad

Las series de tiempo se clasifican en *estacionarios* y *no estacionarios*. Una del tipo *estacionaria* es una con esperanza constante en el tiempo, cuya varianza es finita y su función de auto covarianza solo depende del lag de diferencia entre los tiempos [12].

Esta idea de forma intuitiva describe la estacionariedad de una serie de tiempo si sus propiedades estadísticas (media y varianza) son esencialmente constantes a través del tiempo. Una serie cuya media y/o varianza cambian a través del tiempo es una serie no estacionaria [13].

Se puede decir que una serie cuyos valores varían respecto a una media constante (sin tendencia), es una serie estacionaria.

La suposición para esto es que se trabaja con datos que tienen una media de cero [5]. El primer paso de la metodología Box-Jenkins es determinar si la serie de tiempo es estacionaria. Y en caso de no serlo es necesario aplicar una transformación para inducirla a ello.

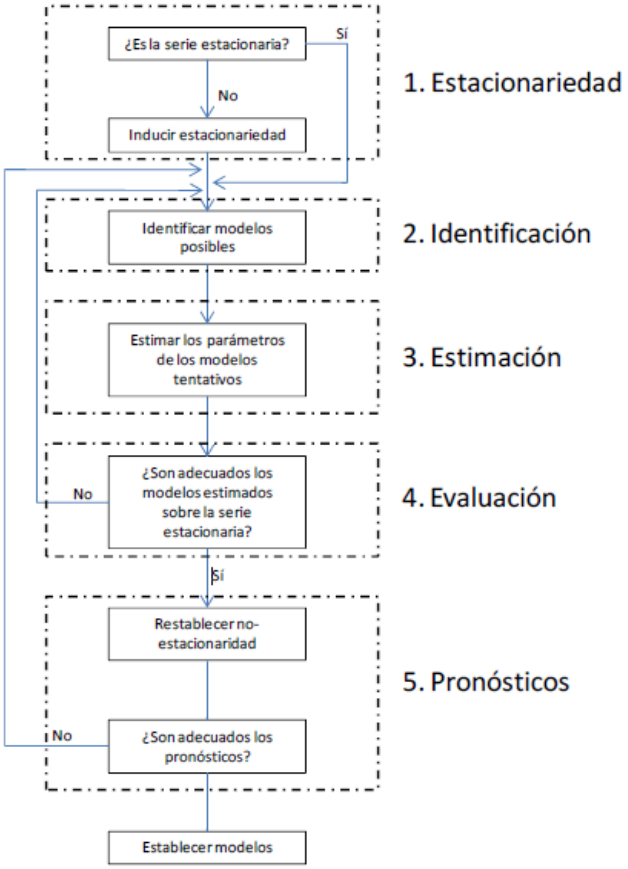

Fig. 5. Metodología Box-Jenkins

Se tienen dos maneras para determinar si una serie es estacionaria: de manera gráfica y utilizando las funciones de auto correlación.

i. Método Gráfico

Advertimos de manera visual si la serie es no estacionaria si se detectan inclinaciones en los datos conforme el tiempo avanza. En nuestro caso, observamos las Fig.1 y Fig.3 y concluimos de manera gráfica que nuestra serie no es estacionaria.

ii. Función de auto correlación

La correlación de una variable consigo misma en diferentes momentos se conoce como *auto correlación* [3]. Podemos analizar esta función usando un correlograma, que grafica el lag k contra su auto correlación. Reconocer la estabilidad de la serie se logra a partir de la variedad de comportamientos que esta función puede mostrar [13]. El caso que nos interesa es aquél donde la serie se corta o se trunca.

Se puede mostrar que si la función de auto correlación (ACF) de la serie $X_1, X_2, \ldots$ claramente se corta o se corta con rapidez, entonces se debe considerar que los valores de la serie son estacionarios; mientras que si lo hace con extrema lentitud entonces se debe considerar que los valores de la serie de tiempo no son estacionales. La Fig.6 muestra el correlograma de nuestra serie de precios. Los lags 0.2, 0.4, 0.6 corresponden a los lags $k = 10, 20$ y 31, respectivamente, ya que el periodo de nuestra serie consta de 52 semanas.

Advertimos que, nuestro correlograma decae lentamente, con lo que concluimos que nuestra serie no es estacionaria. Por lo tanto, resultó necesario aplicar transformaciones a la misma que nos permitan obtener una estacionaria.

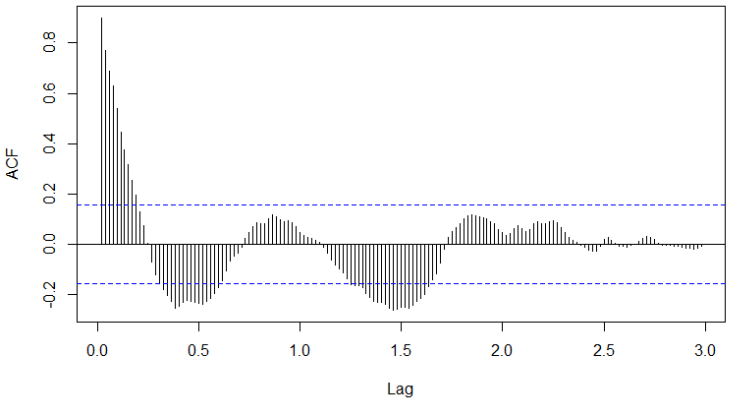

Fig. 6. ACF de la serie de precios

Se analizaron dos transformaciones usuales: primeras diferencias y logaritmo diferenciado.

i. Primeras diferencias

Diferenciar términos adyacentes de una serie puede transformarla de no estacionaria a estacionaria. Para ello, se define el operador diferencia ∇ dado por [3]:

$$\nabla x_t = x_t - x_{t-1} \qquad (1)$$

La serie resultante se observó en la Fig.7. De manera gráfica se apreció que, los precios varían alrededor de una media de cero (se puede ver esto trazando una línea horizontal desde el punto y = 0). Los datos se comprobaron estacionarios analizando su ACF dado en la Fig.8, donde se observó que la gráfica se trunca a partir del lag $k = 4$.

ii. Logaritmo diferenciado

Otras transformaciones usuales son el logaritmo y el logaritmo diferenciado. El primero se usa para disminuir una varianza que suele ser creciente, y el segundo se utiliza para precios de mercado, que puede interpretarse como aproximadamente el cambio porcentual en el precio [5]. En nuestro caso, la transformación logaritmo no resultó útil, mientras que la transformación de la serie a logaritmo diferenciada dada por:

$$\nabla \log(x_t) = \log(x_t) - \log(x_{t-1}) = \log\left(\frac{x_t}{x_{t-1}}\right) \qquad (2)$$

Derivó nuevamente en una serie estacional que se puede apreciar en las Fig. 9 y Fig.10. Debido a que el ACF de la segunda transformación resultó con un pico relevante pasado el lag $k = 30$, se eligieron las primeras diferencias como transformación a utilizar.

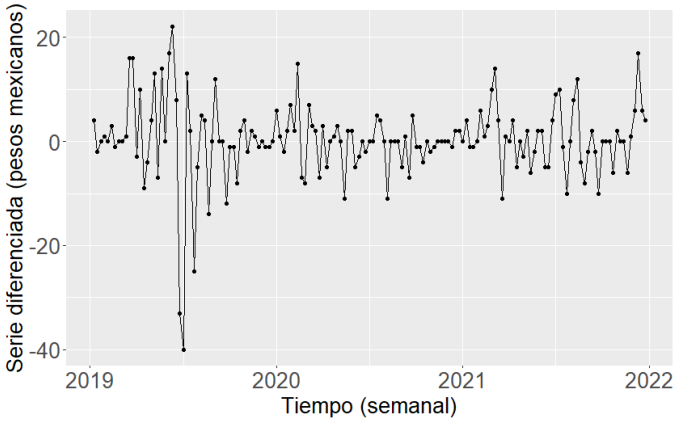

Fig. 7. Serie de precios diferenciada

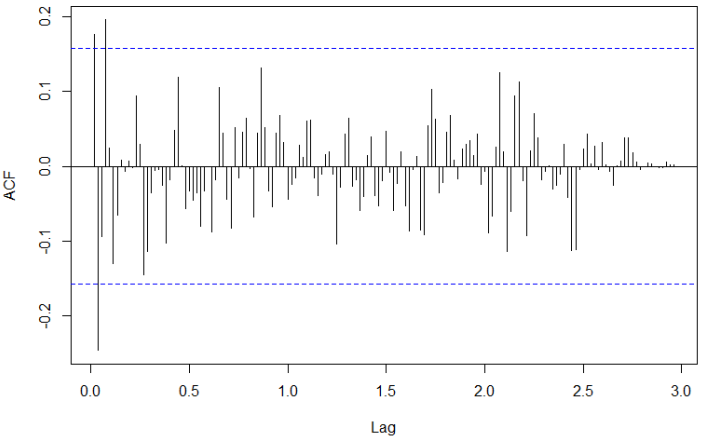

Fig. 8. ACF de la serie de precios diferenciada

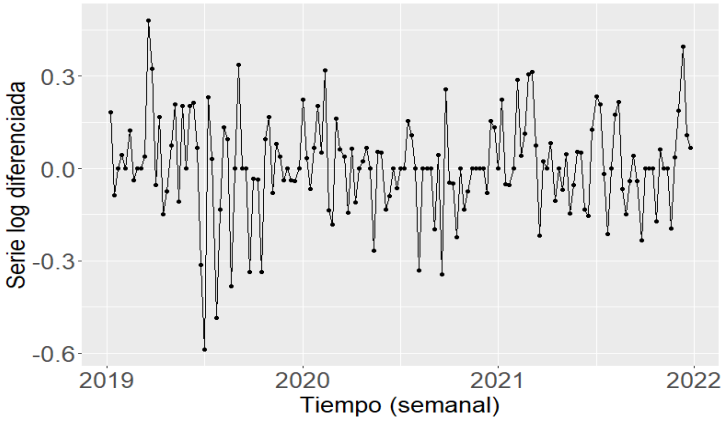

Fig. 9. Serie de precios logaritmo diferenciado

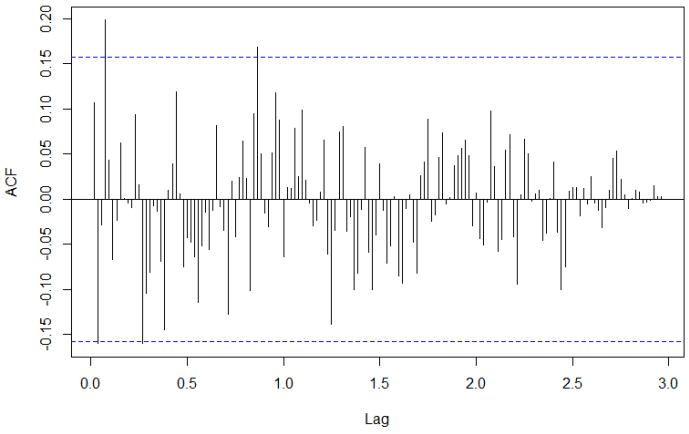

Fig. 10. ACF de la serie de precios logaritmo diferenciada

TABLA I.
CARACTERÍSTICAS DE ACF Y PACF TEÓRICAS

Modelo	ACF	PACF
AR(p)	Decae a cero	Se trunca o corta después del retraso p
MA(q)	Se trunca o corta después del retraso p	Decae a cero
ARMA(p,q)	Decae a cero	Decae a cero

F. Identificación para serie diferenciada

El siguiente paso en la metodología Box-Jenkins consiste en la identificación del posible modelo que rige el proceso de la serie de precios.

Las primeras ideas básicas son las siguientes:

- La serie de tiempo estudiada cuenta con sus respectivas funciones de auto correlación (ACF) y correlación parcial (PACF).
- Cada configuración ARMA posee su ACF y PACF teóricas asociadas al modelo
- S i la ACF y PACF obtenidas de la serie a estudiar se asemeja a una o varias ACF y PACF teóricas, entonces podemos decir que el modelo ARMA teórico es un modelo tentativo para la serie.

Así, la identificación del posible modelo se realiza por medio de la comparación de las funciones de auto correlación calculadas contra las teóricas. Las principales características que se observan de las funciones de auto correlación teóricas se encuentran en la Tabla I [5].

Modelo ARIMA(p,1,0): La Fig. 11 muestra la PACF de la serie de precios diferenciada. El último pico significativo se encuentra en el lag $k = 4$, luego se ajustó un modelo AR(4) [6] a la serie diferenciada de la siguiente manera:

$$y_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3} + \alpha_4 x_{t-4} + w_t \qquad (3)$$

con $y_t = x_t - x_{t-1}$ la serie diferenciada, x_t la serie de precios y w_t ruido blanco con la suposición de que w_t son independientes, con distribución normal de media cero y desviación estándar común σ_w^2. Lo que se convertiría en un modelo ARMA (4,1,0) para la serie de precios.

Debido a que la esperanza de nuestros datos diferenciados estacionarios es relativamente pequeña ($\mu = -0.03846$) no se trabajará con un modelo con constante [14] [12].

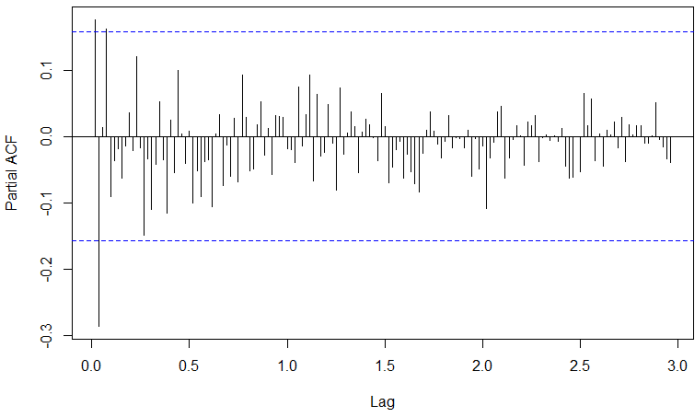

Fig. 11. PACF de la serie de precios diferenciada

Modelo ARIMA(0,1,q): En el ACF de la serie diferenciada, mostrada en la Fig. 8 tiene su último pico significativo en el lag $k = 4$, después del cual los valores decaen a cero [13]; así, se ajusta un segundo posible modelo MA(4) a la serie diferenciada dada por [3]:

$$y_t = w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2} + \beta_3 w_{t-3} + \beta_4 w_{t-4} \qquad (4)$$

Sustituyendo y_t obtenemos el modelo ARIMA(0,1,4) para la serie de precios.

Modelo ARIMA(p,1,q): Para un proceso ARMA(p,q), se analizaron las posibles configuraciones con los límites $p = 4$ y $q = 4$ y se escogió el modelo con el menor AIC, resultando ser el ARMA(4,2) para la serie diferenciada, lo que resulta en el modelo:

$$y_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3} + \alpha_4 x_{t-4} + w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2} \qquad (5)$$

Lo que nos deja para la serie de precios con el modelo ARIMA (4,1,2).

G. Estimación para la serie diferenciada

A continuación, se estimaron los coeficientes de los modelos escogidos anteriormente. Esto se realizó utilizando paquetes del programa R. En esta etapa, se compararon los modelos escogidos y se seleccionó el más adecuado para nuestra serie.

El criterio principal y el que se utilizó para el cálculo de los parámetros es el llamado *Estimación de máxima verisimilitud* o *Maximum Likelihood* (ML). Se eligió este método debido a que tiene un mejor efecto para series de longitud moderada y también para modelos estocásticos estacionales [14], el cual, como se analizó en la sección C, es nuestro caso.

Además de que se ha demostrado que los parámetros estimados utilizando este método reflejan con gran exactitud las características presentes en los datos de la serie de tiempo [10]. Su ventaja es que usa toda la información de la serie a comparación de otros criterios [14]. Así, se define para cualquier conjunto de observaciones $Y_1, Y_2, \dots, Y_n$ la *función de verisimilitud L* como la densidad de probabilidad conjunta obtenida de los datos observados [5]:

$$L(\mu, \phi_i, \sigma_w^2) = f(x_1, \dots, x_n | \mu, \phi_i, \sigma_w^2) \qquad (6)$$

Con ϕ_i los parámetros a estimar. Sin embargo, es considerada como función de los parámetros desconocidos del modelo con los datos obtenidos fijos. Los *estimadores de máxima verosimilitud* se definen entonces como aquellos valores de los parámetros para los cuales los datos realmente observados son más probables, es decir, los valores que maximizan la función de verosimilitud [14]. Como regla general, es más conveniente trabajar con el logaritmo de la función de verisimilitud y maximizar esta función. Esto se realizó de manera numérica en R.

A continuación, se utiliza el *Criterio de Información de Akaike* (AIC) para seleccionar el modelo a escoger. Este es un estimador del promedio de la divergencia Kullback-Leibler del modelo estimado al modelo real. Este indica que se debe seleccionar el modelo que minimice [5]:

$$AIC = -2\log\big(L(\mu, \phi_i, \sigma_w^2)\big) + 2k \qquad (7)$$

Donde $k = p + q + 1$ para un modelo con un término de intersección o constante y $k = p + q$, otro caso. Como desventaja se tiene que el AIC es un estimador sesgado y esto puede notarse para relaciones grandes entre parámetros. En la Tabla II se muestran los parámetros estimados para cada modelo propuesto para la serie de precios diferenciada con sus respectivos valores de máxima verisimilitud y AIC.

Una vez obtenidos los parámetros estimados de los modelos propuestos, estos deben cubrir los siguientes requerimientos [10]:

- El modelo elegido debe tener *parsimonia de parámetros*: el modelo ajusta la información disponible sin usar coeficientes innecesarios. En otras palabras, siempre se opta por el modelo con el menor número de coeficientes. El objetivo de la modelación ARIMA es encontrar el modelo que mejor se aproxime al verdadero proceso y que exprese el comportamiento de la variable estudiada de forma apropiada y práctica, más que encontrar el modelo exacto que represente al proceso generador de las observaciones.
- El modelo debe ser *estacionario* e *invertible*.
  Una definición que nos será útil para reescribir los modelos es el *operador de retraso* (backward shift) dado por [3]:

TABLA II
PARÁMETROS ESTIMADOS Y CARACTERÍSTICAS

Modelo	Estimaciones	Error estándar	Valor $\|t\|$	Valor p
ARIMA(4,1,0)	$\alpha_1 = 0.2298$	0.0794	2.8942	0.0038
	$\alpha_2 = -0.2385$	0.0811	2.9408	0.0032
AIC = 1057.3	$\alpha_3 = -0.0219$	0.0820	0.2670	0.7892
RMSE = 7.0656	$\alpha_4 = 0.1630$	0.080	2.0375	0.0415
ARIMA(0,1,4)	$\beta_1 = 0.2294$	0.0807	2.8426	0.0044
	$\beta_2 = -0.1576$	0.0807	0.1885	0.0593
AIC = 1058.7	$\beta_3 = -0.1535$	0.1001	1.5334	0.1250
RMSE = 7.0995	$\beta_4 = 0.1241$	0.0831	1.4933	0.1354
ARIMA(2,1,2)	$\alpha_1 = -0.1099$	0.2618	0.4197	0.6746
	$\alpha_2 = -0.6088$	0.1248	5.5128	0.0000
AIC = 1057	$\beta_1 = -0.3592$	0.3026	1.1870	0.2351
RMSE = 7.0609	$\beta_2 = 0.4406$	0.1618	2.7231	0.0064

$$\mathbf{B}x_t = x_{t-1} \qquad (8)$$

Y a la reescritura de las ecuaciones para los modelos AR(p) y MA(q) igualada a cero se les conoce como *ecuación característica*.

La *estacionariedad* también la determina las raíces d la ecuación característica generada por la parte autorregresiva del modelo, estas deben exceder la unidad en valor absoluto [3].
La *invertibilidad* se refiere a que cualquier modelo ARIMA puede representar a la serie en función de las observaciones pasadas; mientras que esto es claro para un modelo autorregresivo, no lo es tanto para uno de medias móviles, por lo que esto se determina cuando las raíces de la ecuación característica generada por la parte de medias móviles exceden la unidad en valor absoluto [3].

- Los coeficientes del modelo deben ser *estadísticamente significativos*.
  Esto se determina usando los valores del error estándar y el valor *t* asociado a cada coeficiente:

$$t = \frac{coeficiente\ calculado}{error\ estándar\ asociado\ al\ coef\ calculado} \qquad (9)$$

Como regla práctica, se dice que es razonable incluir en el modelo un parámetro cuya estadística t absoluta es mayor a 2.

- El modelo debe proporcionar un ajuste adecuado. Puede ocurrir que, para una misma serie de tiempo, existan distintos modelos que cumplen con las características de un modelo adecuado y brindan

resultados semejantes, sin embargo, debe elegirse aquel que se ajuste mejor a la serie.

Una medida útil que ayuda a saber el grado de ajuste del modelo a la serie es la *RMSE* (root-mean-squared error). Esta medida da a conocer la desviación estándar de los residuales del modelo. Está dada por [6]:

$$RMSE = \sqrt{\frac{1}{n}\sum {residuo_t}^2} \qquad (10)$$

Con n el número de residuales del modelo.

Usualmente se utiliza este valor para compararlo con el de otros modelos estimados para la misma serie. Se escogerá aquel modelo cuyo RSME tenga un menor valor.

- *Redundancia de parámetros.* Cuando los polinomios generados comparten un factor en común, el modelo puede ser simplificado [3]. Esto solo se analiza en un modelo ARMA(p,q) que cuente con parte autorregresiva y de medias móviles.

El primer punto a revisar sería ver si los coeficientes son estadísticamente significativos, esto con el fin de averiguar si algún modelo necesita ser reducido a otro. Y, en efecto, esto ocurriría en todos los modelos teóricos propuestos. Los modelos reducidos finales se encuentran en la Tabla III. A continuación, se explicarán estas reducciones. De la columna de los valores t asociados del modelo ARIMA (4, 1, 0) observamos que solo 3 de los 4 coeficientes estimados son estadísticamente significativos, por lo que se disminuye en un grado el modelo. Al momento de ajustar un ARIMA (3, 1, 0) resulta que solo dos coeficientes son significativos. Lo que resulta finalmente en un ARIMA (2, 1, 0) con todos sus coeficientes significativos, la eliminación de los coeficientes α_3 y α_4 lo indicaban sus valores t y p, respectivamente; ya que, si bien el coeficiente α_3 tendría un valor t mayor a 2, su valor p se encontraba cerca al valor 0.05. En el modelo ARIMA (0, 1, 4), resulta que tres de los cuatro coeficientes no son significativos, resultando en un modelo ARIMA (0, 1, 1).

Para el modelo ARIMA (2, 1, 2) se tiene un coeficiente autorregresivo y uno de medias móviles no son significativos, reduciendo el modelo a un ARIMA (1, 1, 1); pero este a su vez indicaría que el coeficiente autorregresivo no es significativo, por lo que resta el modelo ARIMA (0, 1, 1), mismo del caso anterior.

Como observación en estas reducciones está el hecho de que varios de los coeficientes no significativos de los modelos propuestos se encontraron en términos intermedios o iniciales del modelo, caso para el cual no se contaban con herramientas para estimar el modelo reducido resultante de la eliminación específica de estos elementos intermedios.

TABLA III
COEFICIENTES ESTADÍSTICAMENTE SIGNIFICATIVOS: ARIMA (2,1,0)

Modelo	Estimaciones	Error estándar	Valor $\|t\|$	Valor p
ARIMA(2,1,0)				
AIC = 1057.4	$\alpha_1 = 0.2274$	0.0769	2.9570	0.0031
RMSE = 7.1625	$\alpha_2 = -0.2844$	0.0767	3.7079	0.0002
ARIMA(0,1,1)				
AIC = 1064.3	$\beta_1 = 0.2971$	0.0839	3.5411	0.0003
RMSE = 7.373377				

Sin embargo, siguiendo el principio de parsimonia, se decidieron por continuar con los modelos reducidos. Examinando ahora los modelos ARIMA (2, 1, 0) y ARIMA (0, 1, 1) observaron que, a pesar de que el modelo ARIMA (0, 1, 1) tiene menos parámetros, este tiene valores AIC y RMSE más altos. Para despejar este dilema, analizamos la estacionariedad e invertibilidad de los modelos:

Modelo ARIMA (2,1,0)

i. Estacionariedad e invertibilidad

Reescribiendo el modelo en términos del operador de retraso:

$$(1 - 1.2274\mathbf{B} + 0.2844\mathbf{B^2})y_t = w_t \qquad (11)$$

El cual tiene como ecuación característica:

$$1 - 1.2274\mathbf{B} + 0.2844\mathbf{B^2} = 0 \qquad (12)$$

La cual tiene raices, iguales sus respectivos valores absolutos por ser reales:

$$\boldsymbol{B_1} \approx 1.09 > 1$$

$$\boldsymbol{B_2} \approx 3.22 > 1 \qquad (13)$$

Por lo tanto, el proceso es estacionario e invertible por definición.

Modelo ARIMA (0,1,1)

i. Estacionariedad e invertibilidad

Reescribiendo el modelo en términos del operador de retraso:

$$y_t = (1 + 0.2971\mathbf{B})w_t \quad (14)$$

El cual, tiene como ecuación característica:

$$1 + 0.2971\mathbf{B} = 0 \qquad (15)$$

Con raíz:

$$\boldsymbol{B} = 3.36 \qquad (16)$$

Ambos modelos son estacionarios e invertibles. Así que se eligió el ARIMA (2, 1, 0) por sus valores AIC y RMSE, los cuales son menores y más cercanos a los modelos teóricos. Además de que es de los métodos estadísticos prioritarios.

H. Evaluación del modelo

Una vez estimados los coeficientes del modelo propuesto, la siguiente etapa es la evaluación o diagnóstico del mismo. En este punto se verifica la eficiencia del modelo y se decide si es estadísticamente adecuado. Esta investigación incluye el análisis de *residuos*, definidos por:

$$e_t = x_t - \hat{x}_t \qquad (17)$$

Donde $\hat{x}_t$ son los valores estimados de la serie resultantes del modelo.

También se analizarán los *residuos estandarizados* dados por [5]:

$$\bar{e}_t = \frac{e_t}{\sqrt{P_t}} \qquad (18)$$

Con P_t la varianza estimada del error del valor estimado al tiempo t. Estandarizar nos ayuda a visualizar los residuos de tamaño inusual más fácilmente [14].

Si el modelo está correctamente especificado y las estimaciones de los parámetros son razonablemente cercanos a los valores reales de la serie, entonces el modelo ha recolectado toda la correlación serial de los datos, por lo que los residuos deberían estar no correlacionados seriamente entre sí [14] [3]; es decir, los residuos y los residuos estandarizados deberían tener un comportamiento aproximado de variables independientes, idénticamente distribuidos con distribución (se supone normal, aunque no en todos los casos) de esperanza cero y desviación estándar común [14] [5]. Para ello, primero se graficó la serie de residuos estandarizados, mostrados en la Fig.12. Si el modelo resulta adecuado, se espera que la gráfica sugiera una dispersión rectangular alrededor de un nivel horizontal cero sin algún tipo de tendencia [14]. Sin embargo, es visible el aumento de la variación al inicio de la serie, lo que se aleja un poco a lo esperado.

Como segundo análisis, para la verificación de la independencia de los residuos, consideramos la función de auto correlación muestral de los mismos, denotado $\hat{r}_k$ [14] y analizamos su ACF, mostrada en la Fig. 13. De donde se observó que no se tiene algún pico con valor alto de importancia o patrones. Más aún, todos los lags se encuentran por debajo de las líneas de dos varianzas, esto indica que no existen patrones que el modelo no haya capturado ya [6]. Concluimos que el gráfico no muestra evidencia suficiente de auto correlación no cero.

A menos que la serie de tiempo sea Gaussiana, no es suficiente con que los residuos no estén correlacionados. Así, como segundo diagnóstico, investigación acerca de la normalidad marginal se puede realizar visualmente mediante la visualización de un histograma de los residuos. Agregando a esto, una gráfica de probabilidad Q-Q (quantil-quantil) puede ayudar a identificar desviaciones de la normalidad. Por último, se puede realizar un test general que toma en consideración las magnitudes de la auto correlación muestral en conjunto con el método estadístico llamado *Q-estadística Ljung-Box-Pierce* dada por [5]:

$$Q = n(n+2)\sum_{h=1}^{H}\frac{\hat{r}_k}{n-h} \qquad (19)$$

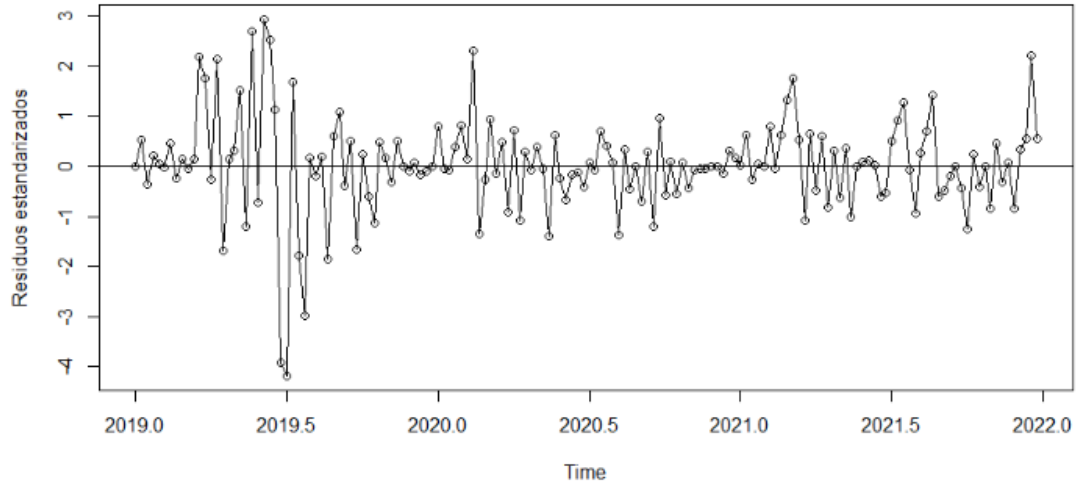

Fig. 12. Gráfica de residuos estandarizados del modelo ARIMA (2,1,0)

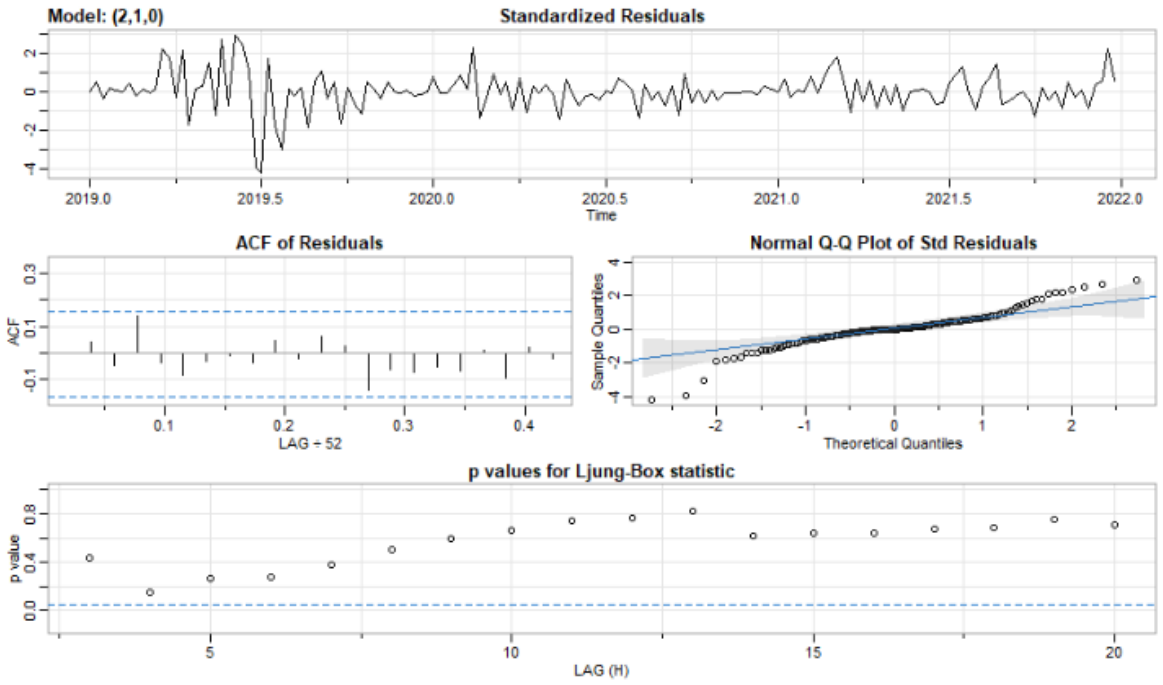

Fig. 13. Resumen de residuos del modelo ARIMA (2,1,0)

Donde el valor H es elegido de alguna forma arbitrario, usualmente tomando el valor $H = 20$. Bajo la Hipótesis Nula H_0 de que el nivel de correlación entre la serie y sus lags es igual a cero, luego las observaciones estudiadas son independientes [6]; asintóticamente $(n \to \infty)$ $Q \sim \chi^2_{H-p-q}$. Luego, se rechazará la hipótesis a un nivel α si el valor de Q sobrepasa el $(1-\alpha)$-cuantil de una distribución χ^2_{H-p-q} [5].

Para nuestro análisis, utilizamos $\alpha = 0.05$, lo que nos resulta en un Q-test de $Q = 0.00036$ y un valor p de $p = 0.9849$, lo cual es mucho más grande que 0.05, por lo que se falla en rechazar la hipótesis nula del test y se concluyó que los valores de los residuos son independientes.

Todos estos métodos se pueden observar en la Fig. 14 y de ellos advertimos que, aunque los residuos estandarizados posean una varianza incrementada al inicio de la serie y la gráfica Q-Q muestre valores alejados de una distribución normal en las orillas de la serie, tanto la gráfica ACF de los residuos como los valores-p de los mismos muestran que el modelo es adecuado.

Finalmente, se realizó un análisis sobre ajustando un modelo ARIMA (3,1,0) a la serie de precios. El modelo original ARIMA (2,1,0) sería confirmado si [14]:

- El estimado del parámetro adicional, en esta caso, de α_3 es no diferente significativamente de cero, y
- Los estimados para los parámetros en común de ambos modelos, es decir, de α_1 y α_2 no cambian significativamente de las estimaciones originales.

Dicho esto, en la Tabla IV se muestran los coeficientes estimados para un modelo ARIMA(3,1,0) para la serie de precios, con sus respectivos errores estándar y valores $|t|$. Primero notando que, el estimado para α_3 es no diferente significativamente de cero, además de que su valor $|t|$ es menor a 2, lo que nos sugiere fuertemente el descartarlo. En segundo, nótese que los estimados para α_1 y α_2 son bastantes cercanos (en especial si consideramos el cambio de magnitud de los errores estándar asociados) y su valor $|t|$ disminuye. Aunado a esto, el modelo tiene un valor AIC más alto, con un valor de $AIC = 1059.39$.

Las desventajas por ajustar un modelo más complejo ARIMA (3,1,0) son suficientes para escoger el modelo ARIMA (2,1,0).

TABLA IV
SOBREAJUSTE: COEFICIENTES ESTIMADOS ARIMA (3,1,0)

Coeficiente	Errores estándar asociados	Valor $\|t\|$
$\alpha_1 = 0.2312$	$s.e. = 0.0806$	2.8684
$\alpha_2 = -0.2873$	$s.e. = 0.0787$	-3.6505
$\alpha_3 = 0.0130$	$s.e. = 0.0813$	1.599

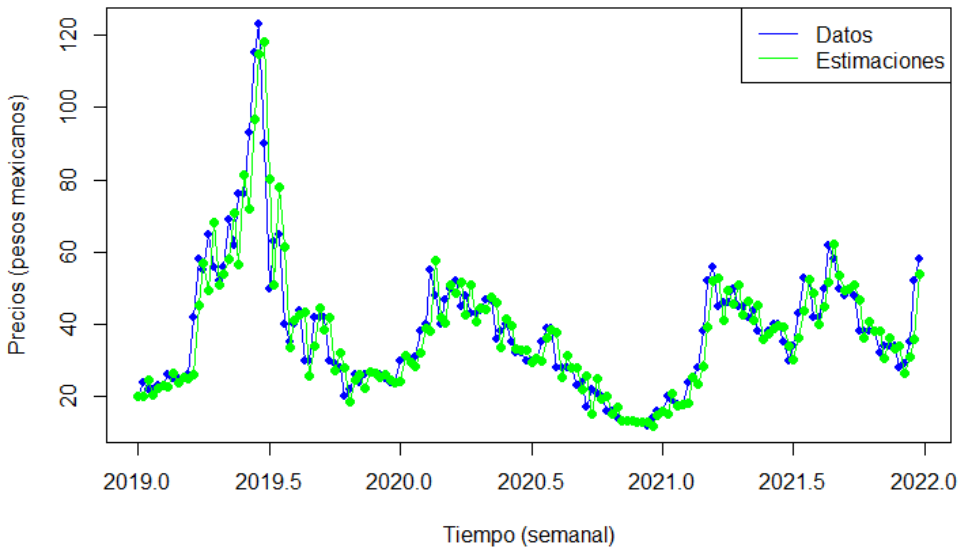

Fig. 14. Serie de precios vs modelo ARIMA (2,1,0) ajustado

Por todo el diagnóstico realizado en esta sección, se concluyó que el modelo ARIMA (2,1,0) es adecuado para la serie de precios.

Pronóstico del modelo

Nuestro objetivo principal en la construcción de un modelo ARIMA para la serie de precios es poder predecir valores de la serie para valores futuros. Esto se realiza en la fase final de la metodología Box-Jenkins. En este paso se realizan *predicciones puntuales*, se estiman límites de probabilidad alrededor de una estimación puntual, conocidos como *intervalos de confianza*, se conocerá y probará la *suficiencia de los pronósticos*. Así, dada el historial disponible de la serie $x_1, \dots, x_t$, interesaron predecir el valor Y_{t+l} que ocurrirá l tiempos en el futuro. Llamamos al tiempo t el *origen de la predicción* y a l el *tiempo dirigido*.

El pronóstico del error cuadrático medio mínimo viene dado por [14]:

$$\bar{X}_t(l) = E(X_{t+l}|X_1, \dots, X_t) \qquad (20)$$

Resolviendo para el caso AR (2), se tiene que esta predicción de un paso adelante para la serie diferenciada está dado por [5]:

$$y_{t+1} = \alpha_1 y_t + \alpha_2 y_{t-1} \qquad (21)$$

Para obtener la ecuación de predicción puntual ARIMA (2,1,0) para la serie de precios, sustituyendo $y_t = x_t - x_{t-1}$, y valores resulta en:

$$x_{t+1} = 1.2261x_t - 0.5239x_{t-1} + 0.2978\,x_{t-2} \qquad (22)$$

La cual es la ecuación utilizada para obtener los ajustes mostrados en las Tablas IV y V. Además, para una mejor visualización, en la Fig. 14 se aprecia un gráfico con los valores reales de la serie (en azul) y los pronosticados por el modelo (en verde). Ahora bien, interesa evaluar la precisión de las predicciones; así, cada pronóstico posee un intervalo de confianza asociado a la probabilidad de que el valor observado futuro se encuentre dentro de dicho intervalo. Usualmente se establece que la probabilidad sea del 95%. A esta combinación de probabilidad y de intervalo se le conoce como *intervalo de confianza del 95%*. La *varianza del error de pronóstico* está dada por [11]:

$$Var(e_t(l)) = \sigma_e^2 \sum_{j=0}^{l-1} \psi_j^2 \qquad (23)$$

Donde, en la práctica, σ_e^2 será desconocida y debe ser estimada de la serie de tiempo estudiada.

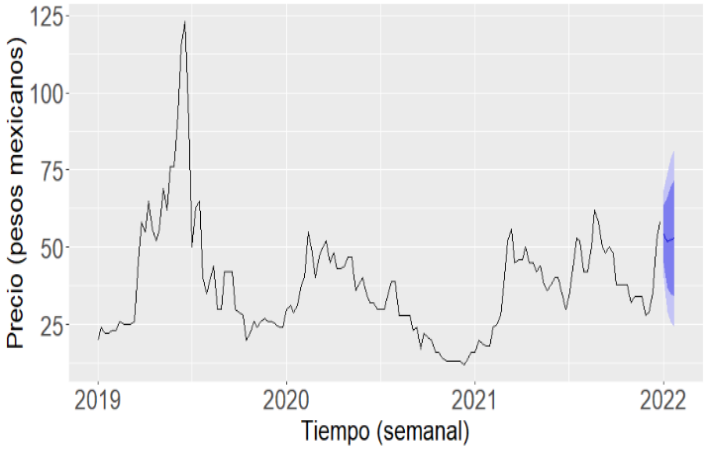

Fig. 15. Gráfica de predicciones con sus intervalos de confianza

TABLA V
PRONÓSTICOS, INTERVALOS DE CONFIANZA Y ERRORES

Predicciones	Valor mínimo	Valor máximo	Valor Enero	Error porcentual
$x_{157} = 61.2030$	47.0277	75.3782	60	2%
$x_{158} = 59.8840$	37.4419	82.32619	62	3.4%
$x_{159} = 59.8108$	33.3070	86.3146	47	27.2%
$x_{160} = 60.1693$	30.8943	89.4443	58	3.7%

Los pesos ψ también son desconocidos debido a que son ciertas funciones dependientes de los coeficientes del modelo (los cuales igual se estiman). Así, la *desviación estándar del error de pronóstico* queda de la forma:

$$S.E.\left(e_t(l)\right) = \sqrt{\sigma_e^2 \sum_{j=0}^{l-1} \psi_j^2} \qquad (24)$$

Con la suposición de errores distribuidos normalmente, un intervalo de confianza del 95% para x_{t+l} el futuro valor de la serie en el tiempo $t + l$ es:

$$x_{t-l} \pm 1.96\left[S.E.\left(e_t(l)\right)\right] \qquad (25)$$

Tanto estas desviaciones estándar como los intervalos de confianza del 95% se calculan de manera numérica. La Tabla VII muestra una regresión de 4 semanas de la serie de precios (aprox 1 mes) con sus respectivos intervalos de confianza y la Fig. 15 la gráfica correspondiente.

III. RESULTADOS

La generación de pronósticos es satisfactoria si los valores futuros se encuentran dentro del intervalo de confianza [6]. La Tabla V muestra los valores reales que se dieron en lo correspondiente a los tiempos $157, 158, 159$ y 160, que equivaldrían a los precios del mes de Enero 2022. Así como el error porcentual de la estimación.

IV. DISCUSIÓN

Pese a que aún están presentes algunas variaciones en la varianza, como se observó en la gráfica de los residuos estandarizados, se alcanzó a obtener la información necesaria de la serie para ajustar un modelo práctico, adecuado, que modela el comportamiento de la serie de manera aceptable.
Lo que corresponde a el análisis numérico de residuos nos verifica esta aceptación.

Volviendo al punto de la varianza; en el primer paso de la metodología, se eligió la serie con la transformación de primeras diferencias debido a lo conveniente que resultó ser su gráfica ACF para modelar; sin embargo, se descubrió que la media de la serie bajo la transformación logaritmo diferenciado era la más cercana a cero (-0.002145611 vs -0.03846154 de la transformación primeras diferencias), lo que en un principio podría disminuir el detalle de varianza resultante en el modelo final.
Ahora, si consideramos el precio x_{156} = 62, y describimos el comportamiento de la serie en enero como: "baja, sube, baja, sube", el modelo falla en la segunda predicción. Para los intervalos de confianza (especialmente los valores máximos), el modelo brinda intervalos razonables.
De las predicciones puntuales en la Fig podemos rescatar que estos valores son bastante cercanos.

V. CONCLUSIONES

En general, el modelo obtenido simula de manera suficiente el comportamiento de la serie de precios para un par de pasos adelante, con intervalos de confianza bastante confiables.
Aun así, se recomienda seguir ajustando conforme surjan nuevos datos debido a que la información estudiada contiene eventos externos únicos, como una temporada de escases (mostrada a mediados del año 2019) y los efectos tardíos de una pandemia (mediados del 2020).

AGRADECIMIENTOS

Este trabajo fue realizado con apoyo de la COFAA del IPN y del programa EDD del IPN.

1. REFERENCIAS

[1] https://www.productoresdeaguacate.com/MODULArevista/modulos/web/www/precioseua.php
[2] Nielsen, A. (2019). *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media.
[3] Copertwait, P. and Metcalfe, A. (2009). *Introductory Time Series with R*. NY, USA: Springer.
[4] Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R* (2nd 2008 ed.). Springer.
[5] Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples* (4th 2017 ed.). Springer.
[6] Krispin, R. (2019). *Hands-On Time Series Analysis with R*. Packt Publishing.
[7] https://online.stat.psu.edu/stat510/lesson/5/5.1
[8] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate
[9] Hanke, J. (2006). *Pronósticos En Los Negocios* (8va Ed). Pearson Educación.
[10] Pankratz, A. (1983). *Forecasting with Univariate Box - Jenkins Models*. Wiley.
[11] Jaime, A. A. (1994). *Introducción Al Tratamiento De Series Temporales. Aplicación A Las Ciencias De Salud*. Ediciones Díaz De Santos.
[12] Shumway, R., & Stoffer, D. (2019). *Time Series: A Data Analysis Approach Using R* (1.a ed.). CRC Press.
[13] Bowerman, B. L., & O'Connell, R. T. (2007). *Pronósticos, series de tiempo y regresión: Un enfoque aplicado*. International Thomson Editores.
[14] Cryer, J. D., & Chan, K. (2008). *Time Series Analysis: With Applications in R* (2nd ed.). Springer.