Modelización de serie de tiempo de precios mayoristas de manzana de Uruguay.

Emanuelle Marsella, Maximiliano Saldaña

Junio 2022

1. Resumen ejecutivo

2. Análisis descriptivo

La serie a ser estudiada es la de precios promedio mensuales del kilo de manzana en la Unidad Agroalimentaria Metropolitana (ex Mercado Modelo). Los precios de los distintos rubros transados en este mercado mayorista de frutas y hortalizas son relevados por el Observatorio Granjero dos veces a la semana, los lunes y los jueves, mediante encuestas a los distintos vendedores informantes. Se relevan precios por distintas variedades, calidades y calibres. Empleando los distintos precios obtenidos los técnicos del Observatorio llegan a un precio de referencia por consenso.

Se cuentan con los datos desde enero de 2013 a mayo de 2022 y se considerará el promedio mensual de los precios, por lo que se cuentan con 113 observaciones. En lugar de emplear los datos bisemanales o el promedio semanal se opta por la frecuencia mensual debido a la dificultad de emplear el herramental de los modelos SARIMA para tales tipos de series, en particular para el tratamiento de la estacionalidad.

En la figura @ref(fig:plot_precios) se presenta el gráfico de la serie a ser trabajada. La impresión inicial que da es que la serie presenta cierto patrón estacional anual, donde los precios comienzan altos para luego descender hasta el segundo trimeste de los años y luego tienden a elevarse hasta el final de año. Esto se puede observar mejor en el gráfico de los precios coloreados por año y el gráfico de la evolución de los precios año a año por mes de la figura @ref(fig:plot_precios_seas). El año 2020 presenta precios atípicamente altos y un comportamiento marcadamente distinto al de los otros años, no se observa la caída inicial de precios sino un aumento sostenido. Esto se puede deber al impacto económico que causó la pandemia de Coronavirus, que llegó a nuestro país en dicho año. Ya para 2021 y lo que va de 2022 parece haber una vuelta a patrones previos. Todo esto deberá ser tenido en cuenta a la hora de la especificación de un modelo del tipo ARIMA/SARIMA.

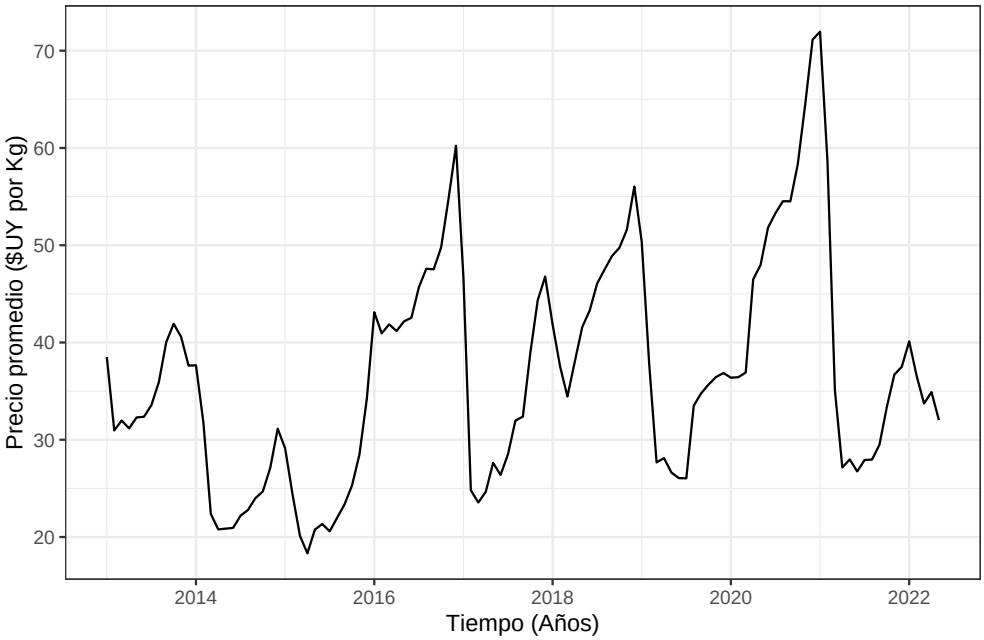


Figura 1: Serie de precios mensuales del Kg de manzana en pesos uruguayos.

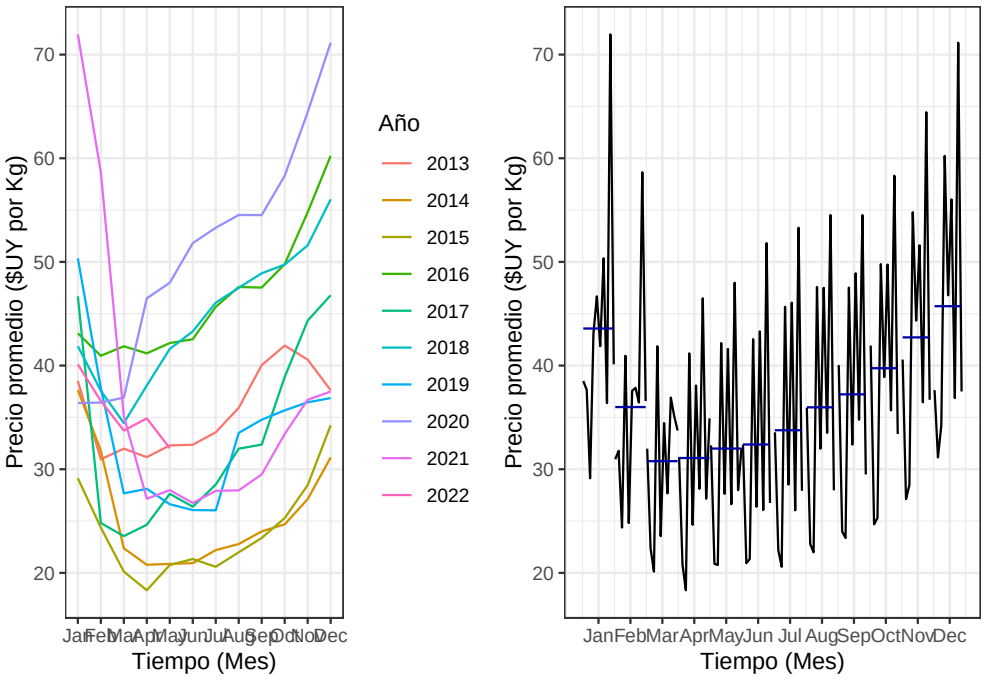
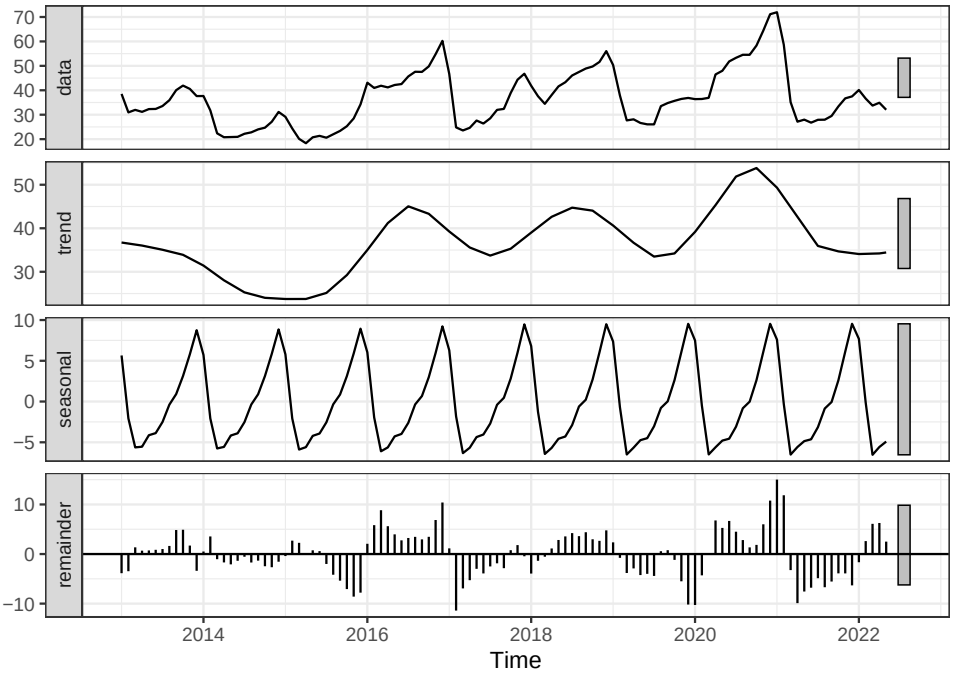


Figura 2: Serie de precios mensuales del Kg de manzana en pesos Uruguayos.



Para ahondar en el análisis descriptivo se realiza la descomposición de la serie en tendencia/ciclo, estacionalidad y componente irregular. En la figura @ref(fig:descomp) se presentan las series de los componentes resultado de una descomposición mediante *STL* (Seasonal Trend Descomposition using LOESS) por separado.

Se puede apreciar una marcada estacionalidad anual y en los últimos 6 años un ciclo corto que se repite cada dos años. La fuerza de la estacionalidad, definida como (Hyndman and Athanasopoulos 2018):

$$F_s = max\left(0, 1 - \frac{Var(R_t)}{Var(R_t + S_t)}\right)$$

toma el valor 0.51 (entre más cercano a 1 más fuerte el componente). Esto refuerza la necesidad de considerar la estacionalidad a la hora de especificar un modelo.

3. Metodología y resultados

3.1. Muestra de entrenamiento y de prueba

Resulta de interés que el modelo ajustado a la serie sea de utilidad para la predicción. Para poder evaluar la calidad de las predicciones, una manera que busca replicar el proceso de obtención de nuevos datos es dividir la serie en una muestra de entrenamiento y una de prueba. La primera se emplea para ajustar el modelo, a partir del cual se realizarán las predicciones. Se dejan las últimas 12 observaciones para la muestra de prueba, que son los precios que van desde junio de 2021 a mayo de 2022. Debe tenerse en cuenta que el periodo del final de la muestra de entrenamiento y también la muestra de prueba están enmarcados en el contexto de gran incertidumbre que presenta la pandemia, por lo que deberá tenerse especial cuidado con el tratamiento de atípicos y las conclusiones que se tomen sobre las predicciones.

3.2. Identificación

La primera fase para el modelado ARIMA de una serie de tiempo en el marco de la metodología de Box-Cox es la identificación del modelo, que consiste en determinar en un principio en detectar la estructura de autocorrelación, la cantidad de parámetros con la que contará la especificación, si la serie necesita diferenciación y si resultará necesaria alguna otra transformación.

3.2.1. Transformación logarítimica

La transformación logarítmica de una serie de tiempo puede tener como resultado una reducción del error de predicción en el caso de que estabilice la varianza (Lütkepohl and Xu 2009). Esto se cumple en particular cuando la varianza aumenta con la media de la serie, lo cual no es el caso de los precios de manzana, que si bien presentan una varianza que aumenta en el tiempo no parece haber una tendencia creciente clara. Por lo tanto, esta transformación no resultaría aconsejable de aplicar.

Para confirmar esto, se considera la transformación de Box-Cox, donde siendo y la variable transformada y x la variable a transformar:

$$y_t = \begin{cases} \frac{x_t - 1}{\lambda} & si \ \ \lambda \neq 0 \\ \ln x_t & si \ \ \lambda = 0 \end{cases}$$

Donde el parámetro λ se estima por máxima verosimilitud. En el caso de la serie planteada, dicho parámetro toma el valor -0.59, por lo que la transformación logarítmica no resulta adecuada.

3.2.2. Autocorrelación y autocorrelación parcial

Como primer elemento a considerar en la identificación de un modelo ARIMA resultan útiles las funciones de autocorrelación y autocorrelación parcial. Al graficarlas en ellas se puede visualizar la estructura de dependencia temporal de la serie trabajada e indicios de la estacionariedad o ausencia de ella. La función de autocorrelación (ACF) se define como:

$$\rho(t, t+j) = \frac{Cov(Y_t, Y_{t+j})}{\sigma_t \sigma_{t+j}}$$

Mientras que la función de autocorrelación parcial (PACF) es la autocorrelación entre Y_t y Y_{t+j} una vez se quita el efecto de las correlaciones intermedias que hay entre ambas variables (Collazo 2022).

Estas funciones se deben estimar a partir de los datos, obteniéndose la autocorrelación muestral. En la figura @ref(fig:acf1) del Anexo se presentan los primeros 72 valores de ambas funciones para la serie original de precios de manzana, junto con el intervalo de confianza de la prueba de hipótesis $H_0)\rho_j = 0$ vs $H_0)\rho_j \neq 0$. En la ACF se puede distinguir que las primeras 5 autocorrelaciones resultan significativas y presentan un decaimiento exponencial y que también son significativas las que están en torno al rezago 24 y entre los rezagos 36 y 48. Lo primero puede ser indicio de una estructura autorregresiva subyacente y lo segundo un indicio de estacionalidad anual o bianual.

Por otro lado, en la PACF se aprecia que los valores para los primeros dos lags son significativamente distintos a 0. Esto puede ser un indicio de que hay una estructura autorregresiva, posiblemente de segundo orden, en los datos.

3.2.3. Dominio de las frecuencias

Desde la perspectiva del dominio de las frecuencias de la serie se considera esta última en su expresión trigonométrica, mediante una suma ponderada de funciones periódicas coseno y seno. El espectro poblacional puede resultar de utilizad para observar la estructura de variabilidad de la serie, dado que el área por debajo del mismo es la variabilidad asociada a las frecuencias consideradas.

En la figura @ref(fig:espectro) se presenta la estimación no paramétrica del espectro poblacional de la serie de precios de manzana trabajada. En esta estimación se hace uso del periodograma muestral, que es la estimación del espectro poblacional a partir de las autocovarianzas muestrales y luego se realiza un promedio ponderado de sus valores mediante un *kernel* a fines de suavizar el resultado, que en general resulta difícil de interpretar inicialmente. En este caso se pondera con el *kernel* de Daniell modificado ponderando de a 3 valores del periodograma muestral 2 veces sucesivas. Se puede apreciar como las frecuencias menores, aquellas asociadas a periodos más largos (teniendo en cuenta que $p = 2\pi/w$, siendo p el periodo y w la frecuencia) son aquellas que acumulan mayor variabilidad. Esto puede considerarse como otro indicio de una dependencia temporal estacional entre las observaciones de la muestra.

3.2.4. Tests de raíces unitarias

Dado que las funciones de autocorrelación dieron indicios de que el proceso no es estacionario, resulta de interés poner a prueba si el proceso es $I(1)$, es decir, si cuenta con una raíz unitaria. En dicho caso el proceso sería no estacionario la cual implicaría que no puede ser modelado en el marco de los ARMA.

Hay múltiples pruebas de hipótesis que han sido desarrolladas con el propósito de identificar raíces unitarias. Dos de ellos son el de Dickey-Fuller aumentado y el de Phillips-Perron. En el caso del primero se especifica el proceso proceso estocástico subyacente como:

$$Y_t = \rho Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Y se contrasta $H_0)\rho = 1$ vs $H_1)\rho < 1$
. Empleamos el estadístico $(\hat{\rho} - 1)/\sigma_{\hat{\rho}}$, que si ρ es menor a 1 en valor absoluto se distribuye normal asintóticamente y si es igual a 1 debe usarse una distribución empírica tabulada por Fuller. El test además toma en cuenta la posible autocorrelación de los errores incluyendo rezagos de la variable en la regresión auxiliar para mayor robustez.

Por otro lado la propuesta de prueba de raíz unitaria de Phillip-Perron se basa en la de Dickey y Fuller, pero además los compatibilizan con la presencia de heteroscedasticidad y/o autocorrelación de los errores.

Para la serie planteada, al observar los p-valores, ambos tests llevan a no rechazar la hipótesis de raíz unitaria al 95 % de confianza. Esto lleva a concluir que una primera diferencia resulta necesaria para llevar la serie a la estacionariedad.

Observando los nuevos autocorrelogramas (@ref(fig:acf2) se puede notar como de entre los primeros valores de la ACF solo el primero resulta significativamente distinto a 0 (lo que puede indicar un componente de medias móviles de primer orden) y lo mismo se da para los valores en torno al lag 24 y 48. Esto último vuelve a indicar la presencia de una estructura de dependencia estacional (indicio de un componente estacional autorregresivo de orden 2). Por otro lado, en la PACF, se aprecia que solo los dos primeros valores son significativos, lo cual puede indicar cierta estructura autorregresiva, posiblemente de orden 2. A pesar de esto, hay que tener en cuenta que en las aplicaciones prácticas al tratar de identificar el orden de un proceso ARIMA mediante la FAC y PACF la distinción entre especificaciones puede volverse difusa y no resulta tan claro la cantidad de parámetros a elegir.

Teniendo en cuenta este acercamiento metodológico, inicialmente se plantea un modelo $ARIMA(2,1,1)(2,0,0)$ (en adelante " *modelo manual*").

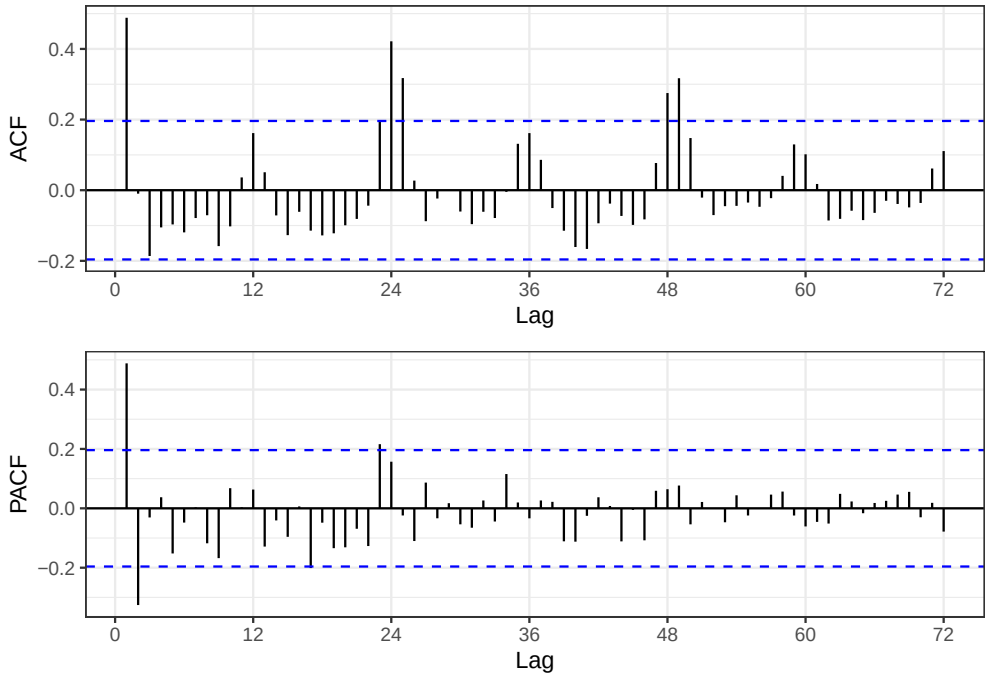


Figura 3: Autocorrelograma y autocorrelograma parcial de la serie de primeras difernecias de precios de manzana.

3.2.5. Selección mediante criterios de información

Una forma alternativa de elegir la especificación del modelo, esto es, la cantidad de parámetros y de diferencias, es mediante los criterios de información. Con estos criterios se busca elegir entre una selección de modelos aquel con una mayor verosimilitud, penalizando por la cantidad de parámetros que tenga.

Uno de los más empleados, y en particular el que se considera en la función *auto.arima()* del software R empleada para la estimación, es el criterio de información de Akaike corregido (AICc). Su fórmula para un modelo es:

$$AICc = T \log \hat{\sigma}^2_{MV} + T\frac{1 + k/T}{1 - (k+2)/T}$$

Siendo T el número de observaciones, k el de parámetros y $\hat{\sigma}^2_{MV}$ el estimador máximo verosímil de la varianza de los errores. El modelo seleccionado es aquel con el menor valor del AICc, lo cuál a nivel algorítmico se hace empleando la selección *stepwise* a partir de un conjunto inicial de modelos (Hyndman and Athanasopoulos 2018).

El modelo seleccionado mediante esta metodología (de ahora en adelante "Modelo AICc") resulta en un $ARIMA(0,1,2)(0,0,2)$. Se puede observar que no se considera un componente autorregresivo, pero si un componente estacional de medias móviles y al igual que el anterior modelo, una diferencia (el algoritmo de *auto.arima()* realiza múltiples tests de raíz unitaria). Además, se considera un componente estacional de medias móviles de orden 2, en lugar de autorregresivo como era en el caso del modelo manual.

3.3. Estimación

Luego de haber identificado la especificación del modelo, el paso que sigue es estimar sus parámetros, dado que no es posible conocer sus verdaderos valores al ser una construcción teórica. Para esto se recurre a la estimación por máxima verosimilitud, donde las estimaciones obtenidas son las que maximizan la probabilidad

de que se haya observado la muestra con la que se cuenta. Es necesario asumir que los errores son gaussianos, un supuesto fuerte pero las estimaciones resultantes que emanen a partir de hacerlo serán razonables aunque no se cumpla (Hamilton 1994).

En particular, el método empleado es la estimación máximo verosímil condicional, donde se supone que la primera observación de la serie es determinística y se maximiza la verosimilitud condicionada a dicha observación. Esto simplifica las expresiones de las funciones, y si el tamaño de muestra es razonablemente grande, la primera observación no tendrá gran efecto sobre la verosimilitud estimada.

3.3.1. Pruebas de significación de los parámetros

Para completar la especificación se realizan pruebas de hipótesis sobre la significación de los parámetros estimados. La hipótesis para cada uno de ellos, siendo λ un parámetro cualquiera del modelo:

$$H_0)\lambda = 0$$
$$H_1)\lambda \neq 0$$

Donde el estadístico empleado es:

$$z = \hat{\lambda}/\hat{\sigma}_\lambda$$

Para el que se cumplirá la Normalidad asintótica debido a que la estimación se realizó por máxima verosimilitud (Hamilton 1994, 143).

Realizando esta prueba para los coeficientes del *modelo manual* todos resultan significativos al 5 % de confianza con la excepción del parámetro de medias móviles (MA) y el primer parámetro autorregresivo estacional (sAR). Se opta entonces por eliminar el componente MA de la especificación y restringir el primer parámetro sAR a 0.

3.3.2. La especificación del *modelo manual* final

La forma del modelo identificado de manera manual es:

$$\Phi_2(L^{12})\phi_2(L)\Delta^1 Y_t = \varepsilon_t$$

Donde:

$$\Phi_2(L^{12}) = 1 - \Phi_2 L^{24}$$

$$\phi_2(L) = 1 - \phi_1 L - \phi_2 L^2$$

$$\Delta^1 = 1 - L$$

Y se supone en un principio que los residuos son de media 0, homoscedásticos, incorrelacionados y:

$$\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Los parámetros resultantes de la estimación para el *modelo manual*, son:

- Parte autorregresiva: $\hat{\phi}_1 = 0{,}557$ y $\hat{\phi}_1 = -0{,}277$

- Parte medias móviles: $\hat{\theta}_1 = -0{,}4157$
- Parte autorregresiva estacional: $\hat{\Phi}_2 = 0{,}415$
- Varianza de las estimaciones: $\hat{\sigma}^2 = 15{,}82$

3.4. Diagnóstico

Para evaluar si el modelo es adecuado se ponen a prueba los supuestos realizados sobre los residuos, sobre los cuales se sostienen los modelos.

3.4.1. Media 0 de los residuos

Para poner a prueba este supuesto se realiza una prueba con el estadístico t, donde se testea $H_0)\mu_\varepsilon = 0$ contra $H_1)\mu_\varepsilon = 0$. En el caso de la muestra no se rechaza la hipótesis nula al 99 % de confianza.

3.4.2. Incorrelación de los erores

Un supuesto cuyo cumplimiento es clave es la incorrelación de los errores, debido a que toda estructura de dependencia que no se esté captando en el modelo irá a parar a los residuos, el proxy con el que contamos para conocer el comportamiento de los errores. Si hay correlaciones entre los errores, la especificación del modelo todavía no capta el comportamiento de los datos.

Para evaluar el cumplimiento de este supuesto un contraste usual es el de autocorrelación conjunta de Ljung-Box, donde la hipótesis nula es que los residuos son incorrelacionados contra la alternativa de que no lo son. El estadístico para realizar el contraste con los primeros h rezagos es:

$$Q_{L-B}(h) = (T(T+2))\frac{\sum_1^h(\hat{\rho}_j)^2}{T-j}$$

El cual se distribuye asintóticamente χ^2_{h-m} bajo la hipótesis nula, donde m es el número de parámetros del modelo.

Para el **modelo manual**, $m = 3$. En la figura @ref(fig:ljungbox) se presentan los p-valores de la prueba conjunta, aumentando sucesivamente el número de rezagos considerados. Se puede observar que considerando hasta los primeros 36 rezagos, no se rechaza la hipótesis nula de incorrelación de los residuos, por lo que el modelo resulta apropiado en este aspecto.

3.4.3. Homoscedasticidad

Otro supuesto que se realizó fue sobre la varianza de los errores, la cual se desea constante. Para ponerlo a prueba se realiza el test desarrollado en McLeod and Li (1983). Es básicamente una prueba de Box-Ljung sobre los residuos al cuadrado del modelo. La hipótesis nula de esta prueba es la homoscedasticidad entre los k rezagos considerados.

En la figura @ref(fig:mcleodli) se presentan los valores de los p-valores de la prueba para los primeros 19 rezagos. Se puede apreciar que considerando en conjunto hasta el décimo rezago hay presencia de heteroscedasticidad. Esto es una indicación que los modelos del tipo SARIMA puede no ser los más adecuados y se tendría que recurrir a los del tipo ARCH/GARCH donde se busca modelizar la varianza de los errores.

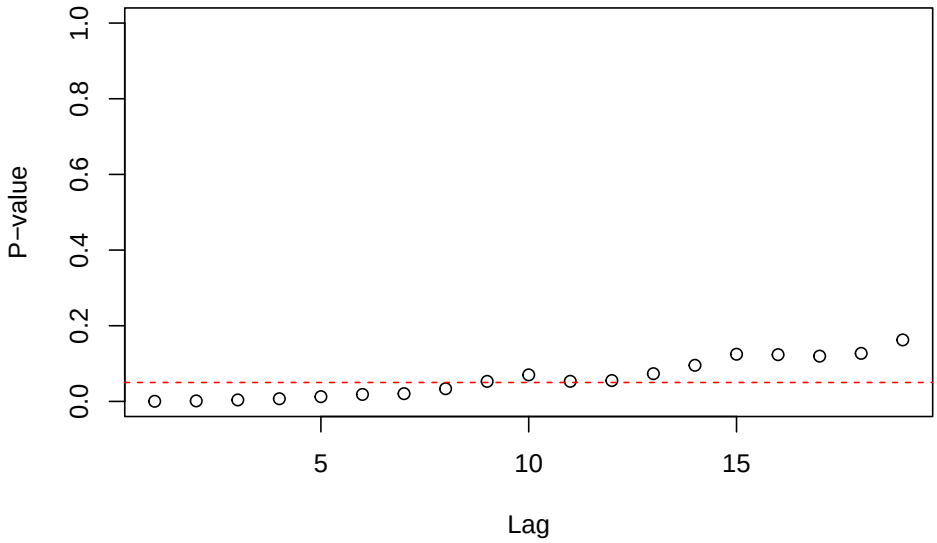


Figura 4: P-valores del test de McLeod-Li para los primeros 19 rezagos

3.4.4. Normalidad

4. Predicción

5. Conclusiones

6. Anexo

Collazo, Silvia Rodríguez. 2022. *Series Cronológicas: Notas de Curso*.

Hamilton, James Douglas. 1994. *Time Series Analysis*. Princeton University Press.

Hyndman, Robin John, and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. 2nd ed. Australia: OTexts.

Lütkepohl, Helmut, and Fang Xu. 2009. "The Role of the Log Transformation in Forecasting Economic Variables." Working paper No. 2591. Munich, Alemania: CESifo.

McLeod, A. I., and W. K. Li. 1983. "Diagnostic Checking Arma Time Series Models Using Squared-Residual Autocorrelations." *Journal of Time Series Analysis* 4 (4): 269–73. https://doi.org/10.1111/j.1467-9892.1983.tb00373.x.

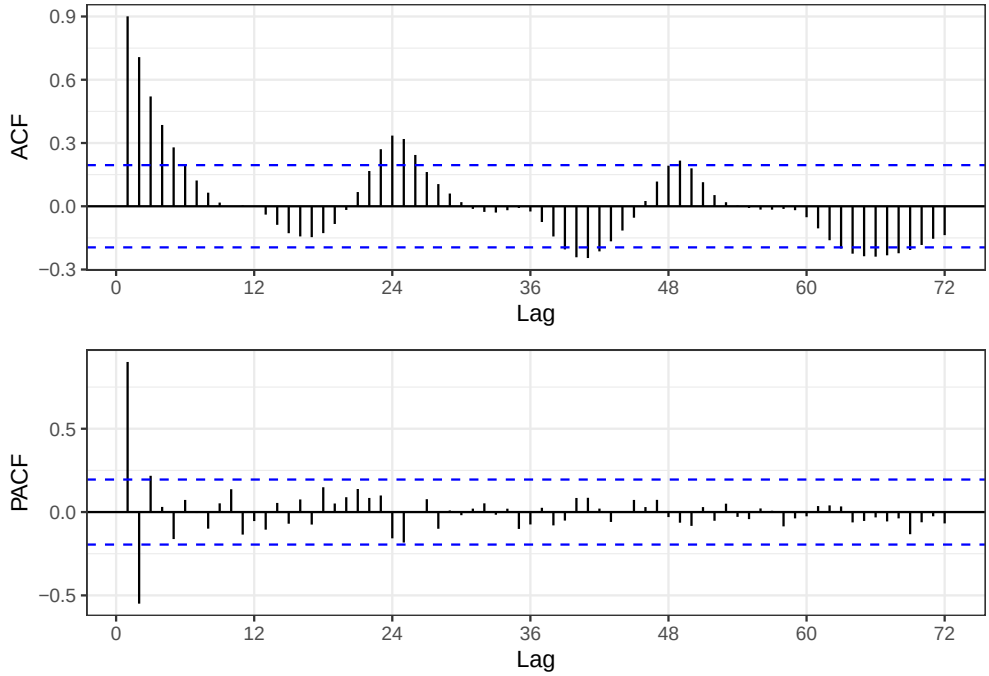


Figura 5: Funciones de autocorrelación y autocorrelación parcial muestrales de la serie de precios de manzana.