Universidad de la República
Facultad de Ciencias Económicas y de Administración

Trabajo final de grado para obtener el título de Licenciado en Estadística

PREDICCIÓN DE SERIES DE TIEMPO SEMANALES MEDIANTE MODELOS DE ESPACIO-ESTADO

Maximiliano Saldaña

Tutor: Ignacio Álvarez-Castro

Montevideo, Uruguay
Mayo 2024

El tribunal docente integrado por los abajo firmantes aprueba el trabajo final de grado:

Predicción de series de tiempo semanales mediante modelos de espacio-estado

Maximiliano Saldaña

Tutor: Ignacio Álvarez-Castro

Licenciatura en Estadística

Calificación: 12

Fecha: 6/5/2024

Tribunal:

Silvia Rodríguez-Collazo _____

Andrés Sosa _____

Ignacio Álvarez-Castro _____

Resumen ejecutivo

En la actualidad, la amplia disponibilidad de datos hace que sea posible contar con series de tiempo de mayor frecuencia que en el pasado, por ejemplo, en la forma de datos semanales. Las series de tiempo semanales aportan información a un grado de desagregación mayor que las series anuales o mensuales, pero presentan un conjunto de dificultades a la hora de ser modeladas. Estas residen principalmente en la modelación de la estacionalidad, debido a que su periodo estacional anual no es entero, a que cuentan con un gran número de observaciones en el periodo lo que dificulta la estimación de parámetros y además pueden presentar múltiples estacionalidades anidadas. En este trabajo se presentan y aplican un conjunto de metodologías del marco de los modelos de espacio-estado como una posibilidad para enfrentar las pecualiaridades de las series semanales, en particular en un contexto de predicción. La principal conclusión es que los modelos de espacio son una solución viable para modelizar series semanales de distintas características, destacándose la modelación de la estacionalidad mediante componentes de Fourier y el modelo TBATS, aunque cabe destacar que en ciertas series de tiempo presentan problemas para la predicción (como es el caso de series con un alto grado de variabilidad).

Palabras claves: modelos de espacio-estado, modelos estructurales, modelos dinámicos lineales, series de tiempo, predicción, ARIMA, TBATS, estacionalidad trigonométrica.

Tabla de contenidos

1 Introducción

En la actualidad la amplia disponibilidad de datos hace que sea posible contar con series de tiempo de mayor frecuencia que en el pasado, por ejemplo, en la forma de datos semanales, diarios o incluso por hora. Este tipo de series cuentan con una gran cantidad de observaciones, teniendo la ventaja de aportar información a niveles mayores de desagregación, pero contando también con un conjunto de desventajas para su modelación. El objetivo de este trabajo es presentar y aplicar técnicas para la modelación de series semanales y su predicción. El hecho de que cuenta con un alto número de observaciones por año (52 o 53) y que su periodo es no entero (un año tiene 52.18 semanas aproximadamente) dificulta la modelización de la serie, en particular debido a la modelización del componente estacional y estimación de parámetros que esto implica. Ante esto se deben recurrir a metodologías especializadas en este aspecto.

El tipo de modelos que se proponen aplicar están enmarcados en el modelado de espacio-estado, tomándose como referencia los desarrollos de Durbin & Koopman (2012) y Petris et al. (2009). Para evaluar los modelos se hace énfasis en la predicción con el menor error, siendo este último definido en base a un conjunto de métricas delineadas en Hyndman (2006). Para enfrentar las complicaciones que genera la frecuencia semanal de las series trabajadas se recurre a las especificaciones presentadas en De Livera et al. (2011) y Hyndman & Athanasopoulos (2021), que hacen uso de modelos de espacio estado donde la estacionalidad se modela con términos de Fourier.

Para la aplicación de las técnicas, se hace uso de series de precios de frutas y hortalizas registrados en la Unidad Agroalimentaria Metropolitana (UAM). En este centro de comercio mayorista se relevan datos sobre las transacciones, en el marco del Observatorio Granjero. Los datos se relevan semanalmente y las series tienen un grado alto de completitud, además de permitir contar con diversos casos de estudio para trabajar con series de dinámicas variadas, correspondiéndose cada una a un producto distinto.

La estructura del trabajo es la siguiente: en un principio se presentan los antecedentes de modelización de series de tiempo semanales y la modelización de series de precios de frutas y hortalizas. Luego, se procede a hacer una descripción del marco teórico, cubriendo los modelos de espacio estado, algunas especificaciones y metodologías de modelización de la estacionalidad y otros tópicos relevantes. Posteriormente, se realiza un análisis descriptivo de los datos, para luego presentar los resultados de la aplicación de los modelos en las series consideradas. Finalmente, se presentan las conclusiones y líneas de trabajo futuras.

2 Antecedentes

Múltiples propuestas han sido anteriormente presentadas para el modelado y predicción de series semanales, las cuales son la base del presente trabajo. A continuación, se realiza un breve recuento de estos antecedentes.

En Harvey et al. (1997) se considera el enfoque de los modelos estructurales (modelos de espacio-estado). Se formula el modelo con una frecuencia diaria, donde la estacionalidad es modelada empleando dos componentes; una función de la fecha del año y una colección de efectos asociados a feriados móviles. El patrón de estacionalidad anual se expresa de forma trigonométrica. Luego, es posible convertir el modelo a una forma semanal, modificándose las ecuaciones apropiadamente.

Por otro lado en Hyndman & Athanasopoulos (2021), se presentan distintas alternativas que se han empleado para enfrentar el problema. En particular, dentro del marco de los modelos de espacio-estado se propone la metodología de regresión dinámica armónica, propuesta en Young et al. (1999), la cual consiste en el modelado de la estacionalidad de manera determinística mediante términos de Fourier y de la dinámica a corto plazo mediante un ARMA. En Cleveland et al. (2014) se busca flexibilizar la representación de la estacionalidad empleando una regresión con pesos locales, lo cuál permite que dicho componente varíe en el tiempo. Otra opción presentada es la de los modelos TBATS[1], la cual se introduce en De Livera et al. (2011) y que comprende también el uso de términos de Fourier pero de manera estocástica, entre otras características particulares de la especificación. Esto tiene como consecuencia que este modelo permita representar una estacionalidad que evoluciona en el tiempo.

Las metodologías anteriormente descritas se aplican en conjunto en el trabajo de Godahewa et al. (2020), en conjunto con Redes Neuronales y el método Theta (una extensión del alisado exponencial), con el propósito de contar con una metodología exhaustiva de modelado de series semanales.

En lo que refiere a la predicción de precios de frutas y hortalizas, que son el tipo de series empleadas en el presente trabajo, ha sido afrontada de diversas formas. En particular, los enfoques paramétricos y los no paramétricos han resultado de interés. Los primeros constituyen un acercamiento al problema de predicción donde se plantea una estructura del modelo en la cual se asume una distribución para los datos y donde hay parámetros a estimar a partir de los mismos. Los modelos de espacio-estado sistematizados en el

[1] Sigla en inglés de estacionalidad trigonométrica con transformación Box-Cox, errores ARMA, tendencia y componentes estacionales

marco de la estadística y la econometría en la obra *Forecasting, Structural Time Series Models and the Kalman Filter* (Harvey, 1990) y el proverbial caso particular de ellos, los modelos autorregresivos integrados de medias móviles (ARIMA), sistematizados por George Box y Gwylim Jenkins en los años setenta, son ejemplos claros del enfoque paramétrico. Por otro lado las técnicas no paramétricas son relativamente más recientes y no asumen una distribución para los datos, generalmente optando por resoluciones donde los métodos computacionales juegan un papel central. Dentro de estas últimas metodologías está enmarcada la de *Artificial Neural Networks* (ANN)[2], que consiste en un sistema de nodos interconectados, los cuales intercambian información entre sí (Peng et al., 2015). Aunque las metodologías no paramétricas cuentan con el atractivo de potencialmente predecir mejor series de comportamiento volátil, los modelos de tipo paramétrico presentan la ventaja de que si los datos se ajustan a los supuestos que el modelo requiere, será más sencillo desarrollarlo, emplearlo e interpretarlo.

Yendo a trabajos previos particulares, en Dieng (2008) se predicen precios minoristas de vegetales en Senegal, haciendo uso de las metodologías ARIMA, suavizamiento exponencial (una técnica que puede ser interpretada desde la perspectiva de espacio-estado), y del análisis espectral. Analizando cuál método da mejores resultados el autor concluye que el modelo ARIMA es el que resulta mejor para el problema planteado.

Por otro lado, en Li et al. (2010) las metodologías empleadas para la predicción son ANN y ARIMA y se hace énfasis en la predicción a corto plazo de precios mayoristas del tomate en China. Los resultados que obtuvieron los autores los llevaron a concluir que el modelo no paramétrico resultó superior. También en Luo et al. (2011) se hace uso de ANN para predecir precios mayoristas de China, en este caso de hongos comestibles. El uso de esta metodología es justificada por los autores en que es particularmente apropiada cuando la información de factores del mercado y el cumplimiento de los supuestos de los modelos no resulta clara.

Otro caso de empleo de metodologías no paramétricas se expone en Peng et al. (2015). Se busca predecir precios diarios de un conjunto de hortalizas y frutas, empleando las metodologías ARIMA, mínimos cuadrados parciales (PLS, por su sigla en inglés) y ANN, siendo las dos últimas las que cuentan con el mejor desempeño. Los autores hacen uso del concepto de *online learning* en sus algoritmos predictivos, actualizando el modelo predictivo con la información nueva que se va recabando.

En Mitra & Paul (2017) también se hace uso de metodologías mixtas, al considerarse que para capturar la complejidad de los fenómenos reales es necesario considerar un elemento no lineal además del lineal, dado que la complejidad hace que estos últimos pierdan capacidad explicativa. Se predicen precios del mercado de Angra en India mediante un modelo aditivo, donde se combina el uso de un modelo ARIMA y ANN.

A nivel nacional, a pesar de que ha resultado de interés la predicción del Índice de Precios al Consumo (IPC) y de sus componentes, el sector frutas y hortalizas no ha sido un punto de foco en la literatura. Entre los trabajos previos que se enfrentan a la predicción de

[2]Redes Neuronales Artificiales

precios se encuentran Güenaga (2014), donde la autora emplea la metodología ARIMA para predecir precios de un conjunto de rubros, tanto por separado como en grupos de series agregados, bajo el planteo que esto puede resultar en una mejor predicción si las series presentan un comportamiento similar. Por otro lado en Millán & Romero (2019) se sistematizan los factores que inciden en la formación de los precios de un conjunto de frutas y hortalizas, estudiándose en particular aquello que incide en la tendencia, la estacionalidad y la variabilidad de las series.

3 Modelos de espacio-estado

Los modelos que son el foco de atención en este trabajo son los modelos de espacio-estado, los cuales son explicados en la presente sección. Esta metodología es flexible y general, lo cual permite tratar una variedad de problemas de análisis de series de tiempo (Durbin & Koopman, 2012). El enfoque tuvos sus orígenes en el campo de la ingeniería, para luego ser aplicado en la estadística y la econometría, siendo abordado en libros como Harvey (1990) y West & Harrison (1997).

Bajo este marco, se considera que el sistema o proceso que se estudia es determinado por una serie de vectores no observables $\theta_1, \dots, \theta_n$ (proceso de estado), que se asocian a una serie de vectores que sí son observados $y_1, \dots, y_n$ (proceso observado). La estructura del modelo es definida en base a estos conjuntos de vectores, matrices que definen la especificación y vectores de términos de error; todos ellos pudiendo depender de parámetros desconocidos que posteriormente se estiman. Tanto la dinámica de dependencia temporal como la estimación de los parámetros puede ser vista desde la perspectiva bayesiana, donde se definen distribuciones previas para los vectores de estado, los observados y para los parámetros (aunque los valores iniciales de estos últimos también pueden ser estimados por máxima verosimilitud). Luego, se obtienen las distribuciones posteriores de manera recursiva haciendo uso de las observaciones consideradas en su ordenamiento temporal.

Un caso particular de esta metodología son los modelos de espacio-estado lineales, donde la estructura de dependencia temporal de los procesos y entre ellos se especifica mediante funciones lineales. Este último tipo de modelos también son llamados "Modelos Dinámicos Lineales", DLM por sus siglas en inglés, y la simplificación que suponen respecto a casos más generales facilita su manejo, tanto a nivel teórico como práctico (Petris et al., 2009).

Uno de los atractivos de esta clase de modelos es su flexibilidad, que permite trabajar directamente con series de media y varianza inestable (no estacionarias) y cambios de nivel, sin la necesidad de realizar transformaciones e intervenciones para lograr la estacionariedad como las que son necesarias para trabajar con los modelos del tipo *ARIMA* tradicionales, los cuales de hecho son un caso particular de los modelos de espacio-estado lineales. Esta flexibilidad también facilita tratar el problema que suponen las observaciones atípicas en las series de tiempo, ya que la estructura de dependencia y estimación recursiva reduce el peso que tienen en la estimación final a medida que consideramos periodos de tiempo más alejados de dichas observaciones.

Asimismo, estos modelos permiten la inclusión de regresores, lo cual puede ser de utilidad para lograr una mejor predicción o dar cierta explicación al fenómeno estudiado.

3.1 Estructura general de los modelos

La estructura de un modelo de espacio-estado lineal puede describirse mediante dos procesos vinculados. El primero es el proceso de estado θ_t, sobre el cual se hace el supuesto que se comporta como una cadena de Markov. Luego, se tiene el proceso observado Y_t, que depende solamente del proceso de estado θ_t y que dado este, es condicionalmente independiente en el tiempo. Esta estructura se representa en la Figura 3.1.

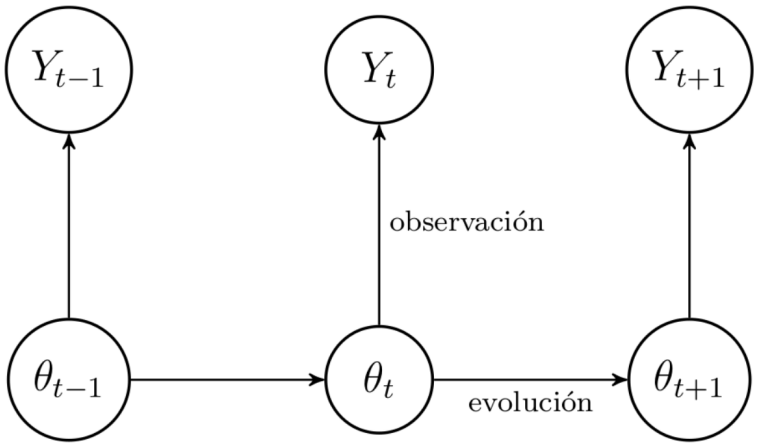


Figura 3.1: Estructura de dependencia de los DLM. Extraído de: Hernández-Banadik et al. (2021)

Un DLM se pueden especificar mediante una primer ecuación que describe el proceso observado, que es aquel sobre el cual se tienen datos, condicional a variables latentes y/o parámetros, que no son observados, y una segunda ecuación que describe la evolución de estas variables latentes (Petris et al., 2009):

$$Y_t = F_t\theta_t + v_t \quad v_t \sim N(0, V_t)$$

$$\theta_t = G_t\theta_{t-1} + w_t \quad w_t \sim N(0, W_t)$$

donde:

- Y_t es el vector m-dimensional de los valores de las m series de las variables consideradas en el momento t,

- θ_t es un vector p-dimensional de procesos no observables,
- F_t y G_t son matrices de dimensiones $m \times p$ y $p \times p$ respectivamente que permiten caracterizar la especificación del modelo,
- v_t y w_t son perturbaciones de los procesos observados y de estado respectivamente.

Sobre las perturbaciones se suele suponer que son independientes entre sí y en el tiempo, con valor esperado 0 y que se distribuyen normales con matrices de varianzas V_t y W_t. Se puede considerar que las observaciones son una medida "ruidosa" del proceso latente (Hernández-Banadik et al., 2021).

Como consecuencia de la estructura de dependencia planteada, la distribución conjunta de las variables de interés queda determinada por la previa del estado inicial $\theta_0 \sim N(m_0, C_0)$ y las densidades condicionales $p(\theta_t|\theta_{t-1})$ y $p(y_t|\theta_t)$. Usando la definición de esperanza condicional:

$$p(\theta_{0:t}, y_{1:t}) = p(\theta_0) \prod_{j=1}^{t} p(\theta_j|\theta_{j-1}) p(y_j|\theta_j)$$

donde $y_{1:t}$ es el conjunto de las observaciones $y_1, y_2, \dots, y_t$.

3.2 Inferencia

La inferencia estadística en este contexto consiste en obtener la distribución posterior de las variables latentes en cada momento del tiempo, condicional a la información observada que se tiene hasta el momento, es decir, obtener $p(\theta_t|y_{1:t})$. Si se cumplen las restricciones de normalidad y linealidad en ambas ecuaciones, esta distribución, llamada *distribución filtrada*, puede ser obtenida en forma cerrada mediante el Filtro de Kalman:

$$p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t) p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})},$$

que puede ser visto como la resolución recursiva de la estimación bayesiana en el modelo lineal dinámico gaussiano. Se supone conocida la densidad filtrada en el momento anterior $\theta_{t-1}|y_{1:t-1} \sim N(m_{t-1}, C_{t-1})$ y además, la de la ecuación de evolución del proceso latente $\theta_t|\theta_{t-1} \sim N(G_t\theta_{t-1}, W_t)$. Entonces se pueden emplear estas condiciones para calcular la densidad previa:

$$p(\theta_t|y_{1:t-1}) = \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|y_{1:t-1}) d\theta_{t-1}.$$

Se puede demostrar que $\theta_t|y_{1:t-1} \sim N(a_t, R_t)$, donde $a_t = G_t m_{t-1}$ y $R_t = G_t C_{t-1} G_t^T + W_t$. También se tiene que a partir de la ecuación de observación del modelo:

$$y_t|\theta_t \sim N(F_t\theta_t, V_t).$$

Luego, la densidad de las observaciones se obtiene mediante:

$$p(y_t|y_{1:t-1}) = \int p(y_t|\theta_t)p(\theta_t|y_{1:t-1})d\theta_t$$

Finalmente, aplicando el Filtro de Kalman se obtiene que:

$$\theta_t|y_{1:t} \sim N(m_t, C_t)$$

$$m_t = E(\theta_t|y_{1:t}) = C_t(R_t^{-1}a_t + F_t^T V_t^{-1} y_t)$$

$$C_t = (R_t^{-1} + F_t^T V_t^{-1} F_t)^{-1}$$

Para el momento inicial se define una previa θ_0, que si se distribuye normal permite obtener iterativamente la distribución filtrada para todos los estados latentes desde el momento en el tiempo 1 al T, siendo este último aquel hasta el que se cuenta con datos de y_t.

En caso de no poseerse el dato observado actual la distribución posterior se define igual a la previa:

$$p(\theta_t|y_{1:t}) = p(\theta_t|y_{1:t-1})$$

Esto resulta de utilidad para enfrentar la problemática de los datos faltantes en las aplicaciones.

3.3 Estimación de parámetros

Las fórmulas anteriormente presentadas son válidas bajo el supuesto de que los parámetros, es decir, las matrices de estructura del modelo y de varianzas-covarianzas de las perturbaciones, sean conocidos. En la práctica, este no es el caso y dichos parámetros deben ser estimados.

Las matrices de estructura suelen ser fijadas por la persona que busca ajustar el modelo, quien en este caso decide los componentes a ser incluidos y cómo son incluidos en la especificación.

Por otro lado, las varianzas y covarianzas de los residuos se suelen considerar como desconocidas y a ser estimadas. Para hacer esto, las principales alternativas son realizar una estimación máximo verosímil (MLE) o aplicar la inferencia bayesiana "pura". Esta última metodología presenta la ventaja de que permiten la inclusión de información previa sobre los parámetros, pero implica cálculos que no se pueden manejar analíticamente. La metodología MLE representa una concepción mixta de los modelos, donde se da una estimación inicial desde la perspectiva de la estadística clásica para realizar las estimaciones posteriores desde la perspectiva bayesiana. Tiene la ventaja de la simplicidad, pero no tiene en cuenta la incertidumbre sobre la estimación.

En las aplicaciones del presente trabajo se hace uso de la metodología mixta, debido a su extensiva implementación en los distintos paquetes de *software* empleados. Por este motivo, se presenta a continuación.

Considerando que tenemos el proceso estocástico $Y_1, \dots, Y_n$, la densidad conjunta de las observaciones dado un conjunto de parámetros ψ es $p(y_1, \dots, y_n|\psi)$. La función de verosimilitud $L(\psi)$ se puede definir como la densidad de los datos como función del conjunto de parámetros ψ (a menos de una constante c). Siguiendo a Petris et al. (2009), en el caso de los modelos dinámicos lineales resulta conveniente expresar la densidad conjunta de la siguiente forma:

$$p(y_1, \dots, y_n|\psi) = \prod_{t=1}^{n} p(y_t|y_{t-1}; \psi).$$

Bajo el supuesto de normalidad de los errores y aplicando la transformación logarítmica a la función de verosimilitud, la log-verosimilitud resultante es:

$$l(\psi) = -\frac{1}{2}\sum_{t=1}^{n} log|Q_t| - \frac{1}{2}\sum_{t=1}^{n}(y_t - f_t)'Q_t^{-1}(y_t - f_t)$$

donde la media f_t y varianza Q_t de la distribución dependen de ψ. La expresión anterior se maximiza numéricamente para obtener la MLE de ψ:

$$\hat{\psi} = \underset{\psi}{argmin}\ l(\psi)$$

Sea H la matriz hessiana de $-l(\psi)$, evaluada en $\psi = \hat{\psi}$, la matriz H^{-1} da una estimación de la varianza de la MLE de ψ ($\hat{Var}(\hat{\psi})$).

3.4 Diagnóstico de modelos

Una vez se ajusta un modelo y se han estimado sus parámetros, se debe verificar el cumplimiento de los supuestos realizados sobre los errores para analizar si el modelo es adecuado. En caso de que tengamos evidencia a favor del no cumplimiento, se deberá analizar qué cambios se deben tomar en la especificación planteada.

Como los errores propiamente dichos no son observables utilizamos los residuos, también llamados innovaciones, como medio de estudiar las propiedades de los primeros. Si se considera como predicción la esperanza condicional, que es la que resulta de la elección de una función de pérdida cuadrática, los residuos son:

$$e_t = Y_t - E(Y_t|y_{1:t-1})$$

donde $f_t = E(Y_t|y_{1:t-1})$ son las predicciones del momento t con los datos que se tienen hasta el momento anterior, empleando la esperanza condicional como predictor. Se requiere que cumplan los siguientes supuestos:

- $E(e_t) = 0$,
- para cualquier $s < t$, e_t y e_s están incorrelacionados,
- para cualquier $s < t$, e_t y Y_s están incorrelacionados,
- $(e_t)_{t\geq 1}$ es un proceso gaussiano (los residuos se distribuyen normales).

Para verificarlos, se emplean pruebas estadísticas, las cuales se delinean a continuación.

3.4.1 Prueba de media 0: *t-test* de Student

Para poner a prueba que los residuos tienen media 0 puede ser aplicado el *t-test* de Student de una muestra (*Student test. Encyclopedia of Mathematics*, s. f.). Considerando $e_1, \dots, e_n$ las hipótesis son:

$$H_0) \mu = \mu_0$$

$$H_1) \mu \neq \mu_0$$

El estadístico empleado es:

$$t_{n-1} = \frac{\bar{e} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

en donde:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$$

$$\hat{\sigma} = \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2}.$$

Bajo H_0 t_{n-1} tiene distribución t de Student con $n-1$ grados de libertad. Se rechaza H_0 si $t_{n-1} > t_{n-1,1-\alpha/2}$, siendo α el nivel de significación fijado.

Se realiza el supuesto de que la media de los errores ($\bar{e}$) se distribuye normal.

3.4.2 Prueba de autocorrelación nula de los residuos: *test* de Box-Ljung

Que los residuos estén incorrelacionados entre sí resulta de especial importancia, ya que en caso contrario hay una estructura de correlación subyacente que el modelo no está captando. Para poner a prueba este supuesto, que representa la bondad de ajuste del modelo, se hace uso de la prueba presentada en Ljung & Box (1978). Las hipótesis que se contrastan son:

$$H_0) \rho_1 = \rho_2 = \cdots = \rho_h = 0$$

$$H_1) \rho_k \neq 0 \text{ para algún } k = 1, ..., h$$

donde ρ_k es la función de autocorrelación a k rezagos y h es el rezago máximo a evaluar. El estadístico empleado es:

$$Q_{L-B}(h) = T(T+2)\sum_{j=1}^{h}(T-j)\hat{\rho}_j^2$$

que bajo la hipótesis nula se distribuye asintóticamente χ^2 con grados de libertad igual al número de coeficientes de autocorrelación que se evalúan (h) menos la cantidad de parámetros en el modelo. La hipótesis nula se rechaza cuando el valor del estadístico $Q_{L-B}(h)$ es mayor que el percentil $1-\alpha$ de la distribución χ^2 teórica con los grados

de libertad anteriormente especificados. El parámetro α es la significación de la prueba, usualmente fijada en 0,1 o 0,05.

Como se desarrolla en Hyndman (2014), no hay muchas referencias prácticas de cómo elegir el número h; los rezagos a evaluar en el test. En Hyndman & Athanasopoulos (2021), los autores recomiendan emplear el mínimo entre $h = 10$ y $T/5$, siendo T el número de observaciones, para datos no estacionales. En el caso de datos con estacionalidad, recomiendan el mínimo entre $2m$ y $T/5$, siendo m el periodo de la estacionalidad.

3.4.3 Prueba de normalidad: Shapiro-Wilk

Múltiples pruebas estadística de normalidad han sido desarrolladas. En el presente trabajo se emplea la prueba de Shapiro-Wilks presentada originalmente en Shapiro & Wilk (1965), debido a que es una de las que presenta mayor potencia estadística (Mohd-Razali & Yap, 2011). La implementación de la prueba empleada se describe en Royston (1982), la cual extiende el tamaño máximo de la muestra a ser empleada de $n = 50$ a $n = 2000$, y se presenta brevemente a continuación.

Las hipótesis a contrastar son:

$$H_0)\text{La muestra se distribuye normal}$$

$$H_1)\text{La muestra no se distribuye normal}$$

Sea $m' = (m_1, \dots, m_n) = (E(x_1), \dots, E(x_n))$ el vector de valores esperados de los estadísticos de orden de una normal estandar $(x_1 < x_2 < \cdots < x_n)$ y $V = (v_{ij})$ la matriz de covarianza de dimensión $n \times n$ $(i = 1, \dots, n)$.

Ahora, considerando $y' = (y_1, \dots, y_n)$, que es la muestra aleatoria en la que se quiere aplicar la prueba de normalidad, ordenada de la forma $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. El estadístico empleado, que es una extensión del test original de Shapiro y Wilks, es:

$$W = \frac{\left[\sum_{i=1}^{n} \hat{a}_i^* y_{(i)}\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde

$$\hat{a}_i^* = \begin{cases} 2m_i \ \ , i = 2, 3, \dots, n-1 \\ \left(\frac{\hat{a}_1^2}{1-2\hat{a}_1^2} \sum_{i=2}^{n-1} \hat{a}_i^{*2}\right)^{1/2} , i = 1, i = n \end{cases}$$

con

$$\hat{a}_1^2 = \hat{a}_n^2 = \begin{cases} g(n-1), n \leq 20 \\ g(n), n > 20 \end{cases}$$

donde

$$g(n) = \frac{\Gamma(\frac{1}{2}[n+1])}{\sqrt{2\Gamma\left(\frac{1}{2}n+1\right)}}$$

$g(n)$ puede ser aproximada usando la formula de Stirling:

$$g(n) = \left[\frac{6n+7}{6n+13}\right]\left(\frac{exp(1)}{n+2}\left[\frac{n+1}{n+2}\right]^{n-2}\right)^{\frac{1}{2}}.$$

Luego bajo la hipótesis nula y si $7 \leq n \leq 2000$, si se aplica la transformación normalizadora:

$$y = (1-W)^{\lambda}$$

$$z = (y - \mu_y)/\sigma_y$$

z se distribuye aproximadamente normal. El parámetro λ es estimado mediante el procedimiento desarrollado en Royston (1982, p. 118). μ_y y σ_y son la media y el desvío estándar de y, los cuales se estiman mediante sus correspondientes análogos muestrales.

Luego, se analiza si el estadístico transformado supera el percentil para el nivel de significación empleado; en caso de que lo haga se rechaza la hipótesis nula de normalidad de la muestra.

3.5 Predicción

Teniendo los datos $y_{1:t}$, se pueden obtener predicciones de valores futuros del proceso observado Y_{t+k} y del de estado θ_{t+k}. La estructura de esta clase de modelos permite calcular las predicciones a un paso naturalmente y actualizarlas secuencialmente a medida que se cuentan con nuevos datos (Petris et al., 2009, p. 66). También se puede generalizar el problema para predecir a k pasos a partir de las distribuciones de predicción, que son derivadas del filtro de Kalman. Para el proceso de estado, la recursión que define la distribución predictiva es:

$$p(\theta_{t+k}|y_{1:t}) = \int p(\theta_{t+k}|\theta_{t+k-1})p(\theta_{t+k-1}|y_{1:t})d\theta_{t+k-1}.$$

Para el proceso observado, la recursión que define la distribución de predicción a k pasos es:

$$p(y_{t+k}|y_{1:t}) = \int p(y_{t+k}|\theta_{t+k})p(\theta_{t+k}|y_{1:t})d\theta_{t+k}$$

Si se mantiene la estructura definida en las subsecciones anteriores, siendo $a_t(0) = m_t$ y $R_t(0) = C_t$. Entonces, para $k \geq 1$:

$$\theta_{t+k}|y_{1:t} \sim N(a_t(k), R_t(k))$$

donde

$$a_t(k) = E(\theta_{t+k}|y_{1:t}) = G_{t+k}a_{t,k-1}$$
$$R_t(k) = Var(\theta_{t+k}|y_{1:t}) = G_{t+k}R_{t,k-1}G'_{t+k} + W_{t+k}$$

y además

$$Y_{t+k}|y_{1:t} \sim N(f_t(k), Q_t(k))$$

donde:

$$f_t(k) = E(Y_{t+k}|y_{1:t}) = F_{t+k}a_k(k)$$

$$Q_t(k) = Var(Y_{t+k}|y_{1:t}) = F_{t+k}R_t(k)F'_{t+k} + V_{t+k}$$

Luego, a partir de dichas distribuciones se pueden realizar medidas de resumen, obtener estimaciones puntuales del proceso observado y del de estado, con sus respectivos intervalos de credibilidad.

La medida de resumen puede ser elegida en base a la teoría de decisión, donde se plantea una función de pérdida $L(\theta, a)$. Una regla de decisión bayesiana elige una acción a en un conjunto $\mathcal{A}$ que minimiza la pérdida condicional esperada $E(L(\theta, a)|y) = \int L(\theta, a)p(\theta|y)d\theta$.

La elección de la función de pérdida va a depender del problema y diferentes especificaciones darán como resultado diferentes medidas de resumen. Si se opta por emplear la función de pérdida cuadrática $L(\theta, a) = (\theta - a)^2$, entonces la medida que

minimiza la pérdida esperada posterior es $a = E(\theta|y)$, la esperanza condicional (Petris et al., 2009, p. 13). Esta medida es la que minimiza el error cuadrático medio, una medida de error usualmente empleada para cuantificar la bondad predictiva.

3.6 Representación trigonométrica de la estacionalidad (términos de Fourier)

El componente estacional de las series de tiempo puede ser representado mediante expresiones trigonométricas, llamadas términos de Fourier. Estos términos son pares de funciones en el tiempo seno y coseno. La cantidad de términos de Fourier (K) define la suavidad del patrón estacional y es elegida mediante un criterio de información, como el AIC o AICc. A menor cantidad de términos, mayor suavidad.

Las ventajas de esta metodología son su utilidad para manejar periodos estacionales largos con una cantidad reducida de parámetros, su capacidad de modelar periodos no enteros y que permite incluir distintos periodos estacionales en una misma expresión. Esto hace que sea idónea para trabajar con series de tiempo semanales, donde modelos como los ARIMA y los ETS (suavizado exponencial) presentan problemáticas incluso considerando la simplificación de que el periodo estacional anual de los datos semanales es $m = 52$. En el primer tipo de modelo, implicaría realizar una diferencia de orden 52, no permitiendo considerar un suavizado del periodo estacional para que el modelo tenga la flexibilidad necesaria al momento de representar un periodo de tiempo tan largo y a este nivel de frecuencia. En el caso de los modelos ETS, implica la estimación de 51 parámetros, lo cual es complejo bajo el método de máxima verosimilitud. Esto tiene como consecuencia que los modelos anteriores no tengan buenos resultados con series semanales estacionales (Hyndman & Athanasopoulos, 2021).

El componente estacional representado mediante términos de Fourier toma la siguiente expresión:

$$S_t = \sum_{j=1}^{K} \left[\alpha_j sen\left(\frac{2\pi}{m} jt\right) + \beta_j cos\left(\frac{2\pi}{m} jt\right) \right]$$

donde α_j y β_j son parámetros a estimar y m es el periodo estacional empleado.

En el marco de los modelos de espacio-estado, los parámetros pueden ser estimados de la manera descrita en la Sección 3.3, con sus respectivos procesos de estado. Otra alternativa consiste en considerarlos como parámetros fijos a estimar, como en la metodología de la regresión dinámica armónica presentada en Young et al. (1999). En este último trabajo, se hace uso de una estacionalidad fija en el tiempo representada mediante términos de Fourier y un término de error con la forma de un proceso ARIMA, para capturar el resto de las dinámicas de la serie.

3.7 TBATS

Los modelos de Estacionalidad Trigonométrica, transformación Box-Cox, errores ARMA, Tendencia y componentes estacionales (TBATS por su acrónimo en inglés) son una subclase de modelos de espacio-estado. Fueron presentados en De Livera et al. (2011) como una forma de predecir series de tiempo que presentan estacionalidad compleja, como lo son aquellas con múltiples estacionalidades anidadas (por ejemplo: una semanal y otra mensual) y cuando el periodo estacional no es entero, como el caso de las series de datos semanales, donde el periodo estacional es $365.25/7 \approx 52,179$.

3.7.1 Estructura del modelo

Un TBATS puede verse como una extensión de los modelos de espacio-estado de suavizamiento exponencial, donde se permite cierta flexibilidad ante no linealidades al emplear la transformación de Box-Cox. Esto presenta la limitación de restringir el modelado a las series de tiempo positivas. A pesar de esto, en el campo de la economía muchas veces las series de interés son positivas, como es el caso del presente trabajo donde se emplean series de precios. Además, otras ventajas de este tipo de modelo son que permiten un amplio espacio de parámetros al ser un modelo de espacio-estado, posibilitando mejores predicciones, y que permite tener en cuenta la autocorrelación en los residuos.

Siguiendo el desarrollo planteado en De Livera et al. (2011), se considera ahora y_t la observación de la variable de interés en el momento t y $y_t^{(\omega)}$ la transformación de Box-Cox con parámetro ω de dicha variable. El modelo se puede representar mediante el siguiente conjunto de ecuaciones:

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega}, & \omega \neq 0 \\ log(y_t), & \omega = 0 \end{cases}$$

$$y_t^{(\omega)} = \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-1}^{(i)} + d_t$$

que representan el proceso observado. Por otro lado, considerando los procesos de estado, el nivel local en el periodo t se define como:

$$\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha d_t$$

la tendencia a corto plazo en el momento t se define como:

$$b_t = (1-\phi)b + \phi b_{t-1} + \beta d_t$$

donde ϕ es un parámetro de amortiguación y b es la tendencia a largo plazo.

d_t es un proceso $ARMA(p, q)$, que representa los errores, por lo que puede definirse como:

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

donde ε_t es un proceso de ruido blanco Gaussiano de media cero y varianza σ^2. α y β son parámetros de suavizado del proceso ARMA(p,q).

En lo que refiere a la representación de los elementos que componen la estacionalidad, se expresa de la siguiente manera:

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)}$$

representa el nivel estocástico del i-ésimo componente estacional en el momento t.

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t$$

es el crecimiento estocástico en el nivel del i-ésimo componente estacional, donde a su vez:

$$s_{j,t}^{*(i)} = -s_{j,t-1} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t$$

En estas expresiones $\gamma_1^{(i)}$ y $\gamma_2^{(i)}$ son parámetros de suavizado, $\lambda_j^{(i)} = 2\pi j/m_i$, m_i es el periodo estacional en el momento del tiempo i-ésimo y k_i es el número de armónicos requerido por el i-ésimo componente estacional. Como se puede apreciar, la estacionalidad toma una forma trigonométrica que hace uso de términos de Fourier y que puede variar en el tiempo.

El modelo luego puede ser expresado en su forma matricial de espacio-estado, la cual es empleada para su implementación computacional. La notación que describe la especificación del modelo es $TBATS(\omega, \phi, \text{p}, \text{q}, \{m_1, k_1\}, \{m_2, k_2\}, ..., \{m_T, k_T\})$.

3.7.2 Estimación y selección de modelos

Para la estimación de los parámetros del modelo TBATS se puede hacer uso de una extensión de la técnica mixta para modelos de espacio-estado desarrollada en la Sección 3.3.

Además, se necesita realizar el proceso de selección de modelos, es decir, de la cantidad de parámetros y órdenes. Para esto se hace uso de criterios de información, como el de Akaike (AIC).

Para la selección de la cantidad de armónicos empleados para representar el elemento estacional se usa una aproximación de la serie desestacionalizada y se realizan pruebas de significación, para luego emplear un criterio de información como el AIC. El proceso se repite incrementando la cantidad de armónicos hasta lograr el mínimo AIC.

3.8 Otros tópicos relevantes al trabajo

3.8.1 STR

La descomposición estacional-tendencia usando regresión (STR por sus siglas en inglés) es un método de descomposición de series de tiempo presentado en Dokumentov & Hyndman (2022). Este tipo de metodología es útil para separar una serie de tiempo en los elementos que la componen, que en general se consideran que son la tendencia/ciclo, la estacionalidad y el remanente,también llamado componente irregular o ruido. De esta manera se puede profundizar en el entendimiento de la serie, clarificando los patrones y permitiendo identificar anomalías.

Con el método STR se supone que la serie de tiempo y_t puede descomponerse aditivamente como:

$$y_t = T_t + \sum_{i=1}^{I} S_t^{(i)} + \sum_{p=1}^{P} \Phi_{p,t} Z_{p,t} + R_t$$

en donde:

- T_t es una tendencia que puede cambiar lentamente.
- $S_t^{(i)}$ son componentes estacionales que pueden cambiar lentamente y que posiblemente tienen topología compleja.
- $z_{p,t}$ son variables auxiliares con coeficientes $\phi_{p,t}$, que pueden cambiar en el tiempo y ser estacionales.
- R_t es el componente irregular o "resto".

Bajo esta estructura el número total de coeficientes es mucho mayor a la cantidad de observaciones, lo cual representa un problema para la estimación. Para solucionar esto se impone una regularización para la estimación, siendo la innovación del método STR permitir lograr esto de una manera eficiente, utilizando operadores de diferenciación de matrices. Así, el problema se convierte a una estimación de una regresión lineal análogo a la regresión de cresta.

Este enfoque cuenta con varias ventajas respecto a métodos anteriores de descomposición, como la posibilidad de considerar múltiples tipos de estacionalidades simultáneamente (por ejemplo: anual, mensual y semanal), incluir el efecto de variables auxiliares en la estacionalidad, considerar periodos estacionales no enteros y grandes (como es el caso de la estacionalidad anual de datos semanales, donde el periodo es 52,18 en promedio) y calcular intervalos de confianza. Además, facilita trabajar con estacionalidades que cambian en el tiempo.

3.8.2 Evaluación de predicciones

Dado que en este trabajo se prioriza emplear los modelos de espacio-estado para la predicción, resulta necesario definir una metodología para evaluar su desempeño predictivo.

A partir de esto surgen dos cuestiones a tratar: en qué datos se evalúan las predicciones y las métricas que se emplean para hacerlo.

3.8.2.1 Muestra de entrenamiento y de prueba

Es de interés evaluar predicciones de datos verdaderamente futuros, es decir, si tenemos una serie de tiempo con T datos, queremos evaluar las predicciones para los datos en los momentos del tiempo $T+1$ al $T+h$, siendo h el horizonte predictivo. Esto se debe a que en general si se realizan predicciones con el modelo dentro de la muestra con la cual se estimó, su desempeño será en general mejor que si se realizan predicciones verdaderas.

Obviamente, no es posible realizar tal evaluación, dado que los verdaderos datos futuros no pueden ser conocidos. Sin embargo, podemos buscar simular esta evaluación con los datos con los que sí contamos, separando la muestra en una sección de entrenamiento y otra de prueba. En la primera se ajustará el modelo a partir del cual se realizan predicciones a un determinado horizonte predictivo. Luego, comparamos nuestras predicciones con los respectivos valores que están en la muestra de prueba, que hacen las veces de valores “futuros”.

Como se indica en Hyndman & Athanasopoulos (2021), cuando la cantidad de datos lo permite el tamaño de muestra de prueba se fija en un 20% del total de datos e idealmente tantos datos como lo requiera el horizonte predictivo deseado.

Una vez fueron separados los datos en estas muestras, se tiene que decidir qué métrica se emplea para comparar predicciones con valores reales.

3.8.2.2 Métricas de evaluación

Las métricas de evaluación del error de predicción son medidas de resumen de cuánto se equivoca un modelo en predecir determinada variable en un conjunto de datos dado. El error se define como:

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$$

donde y_{T+h} es la observación de la variable en el momento $T+h$ y $\hat{y}_{T+h|T}$ la predicción de dicha variable en ese momento.

Sea $\{y_1, y_2, \dots, y_T\}$ la muestra de entrenamiento y $\{y_{T+1}, y_{T+2}, \dots, y_{T+m}\}$ el conjunto de prueba de un largo m. En general estas métricas son un promedio de una función de los errores en la muestra de prueba. Las empleadas en este trabajo son:

- Raíz del error cuadrático medio: $RMSE = \frac{1}{m}\sqrt{\sum_{i=T+1}^{m} e_i^2}$
- Error absoluto medio: $MAE = \frac{1}{m}\sum_{i=T+1}^{m} |e_i|$

Estas métricas son dependientes de la escala y unidad de medida de la variable en cuestión, por lo que sirven para comparar predicciones que sean de una misma variable.

Ahora, sea $p_t = 100e_t/y_t$. Una medida de error porcentual es:

- Error porcentual absoluto medio: $MAPE = \frac{1}{m}\sum_{i=T+1}^{m} |p_i|$

Tiene la ventaja de ser independiente de la escala y tener la interpretación simple de ser el porcentaje promedio de error. Por otro lado, tiene desventajas, como el hecho de que penaliza más severamente los errores negativos que los positivos, no está definido cuando la variable evaluada toma el valor 0 y toma valores extremos cuando dicha variable toma valores cercanos a 0.

4 Datos

A continuación, se presentan las herramientas computacionales que se emplearon para trabajar con los datos, cómo se elijieron estos últimos para aplicar y evaluar los modelos y se realiza su análisis descriptivo.

4.1 Herramientas computacionales

Para el procesamiento, visualización y la modelización de los datos se empleó el lenguaje R (R Core Team, 2024). Para la escritura del documento se empleó R en conjunto con Quarto (Allaire et al., 2024). Se hizo uso de los siguientes paquetes de R:

- **Manipulación de datos**: *dplyr* (Wickham, François, et al., 2023), *lubridate* (Grolemund & Wickham, 2011), *tidyr* (Wickham, Vaughan, et al., 2023), *gridExtra* (Auguie, 2017), *readr* (Wickham, Hester, et al., 2023), *data.table* (Dowle & Srinivasan, 2023), *stringr* (Wickham, 2022), *tibble* (Müller & Wickham, 2023), *forcats* (Wickham, 2023)
- **Visualización de datos**: *ggplot2* (Wickham, 2016), *scales* (Wickham & Seidel, 2022), *RColorBrewer* (Neuwirth, 2022)
- **Modelización de datos temporales**: *forecast* (Hyndman & Khandakar, 2008), *dlm* (Petris et al., 2009), *statespacer* (Beijers, 2023), *Metrics* (Hamner & Frasco, 2018), *tsoutliers* (López-de-Lacalle, 2019), *stR* (Dokumentov & Hyndman, 2023)
- **Escritura de documentos**: *knitr* (Xie, 2014), *here* (Müller, 2020)

4.2 Elección de los datos

Con la finalidad de presentar una aplicación y evaluar el desempeño de los modelos presentados, en particular respecto a la predicción, se busca ajustarlos a datos que presenten una frecuencia semanal. Se optó por emplear series de precios por kilogramo correspondientes a un conjunto de frutas y hortalizas, los cuales son relevados y registrados en la Unidad Agroalimentaria Metropolitana. Dichas series tienen las características de que cuentan con una cantidad de datos suficiente para el modelado, tienen muy pocos datos faltantes y presentan distinción según la fruta u hortaliza, lo que permite emplear los modelos en diversos contextos.

Los rubros tomados en cuenta para el análisis son: manzana, papa rosada, tomate, cebolla blanca y naranja. Para su selección se tomaron en cuenta varios criterios.

En primera instancia se consideró que debían ser productos que tuvieran peso en la canasta alimenticia, de modo que tuvieran mercados de tamaño considerable donde pudiera haber dinámicas de formación de precios. En segundo lugar, se buscó un conjunto de rubros que tengan diferencias sustanciales entre sí (en su producción, estacionalidad, mercado, etc.) con la finalidad de captar una diversidad de series para evaluar la adaptabilidad de los modelos. Por último, en general se tomó en cuenta que no todas las series presenten una volatilidad excesiva a corto plazo, dada la dificultad de hacer una predicción en tal contexto. En cuanto a esto último es excepción el tomate, que se mantiene en el conjunto estudiado debido a su importancia en la canasta y a que permite ilustrar un caso de alta volatilidad.

4.3 Construcción y caracterización de los datos

Las series de precios trabajadas van desde enero de 2013, hasta junio de 2022. En su forma original, tenían una frecuencia bisemanal, dándose los relevamientos y registros los lunes y los jueves en el marco del Observatorio Granjero de la UAM. Estos días son los de mayor actividad en este mercado; aproximadamente el 40% de los ingresos semanales se acumulan en ellos según los técnicos del Observatorio. Para dicho relevamiento de datos se considera la metodología presentada en Comisión Administrativa del Mercado Modelo (2021), donde se especifica una tipificación de calidad. Además, se acota el problema a los rubros de origen nacional, dado que para los productos de origen extranjero las características de la fijación de precio presentan diferencias.

Se cuentan con series de precios para los distintos productos y sus variedades, distinguiéndose en cada caso tres calidades en orden descendiente: Extra, I y II. Para contar con una noción general del comportamiento de los precios, los técnicos del Observatorio Granjero recurren a encuestar un conjunto selecto de informantes, que consiste de vendedores mayoristas del mercado cuya información se considera fiable y "representativa". Esta información se releva cuando ya las transacciones han comenzado a disminuir y consecuentemente el clima de precios se definió. Posteriormente, los técnicos discuten la información obtenida y se llega a un consenso de cuales son los precios puntuales de mayor veracidad, tanto máximos y mínimos, observados para cada rubro. La representatividad no es en términos de muestreo probabilístico, debido a que no hay un diseño muestral subyacente en el relevamiento de la información, sino que los especialistas de la UAM consideran que el proceso permite captar con un buen grado de aproximación las señales de precios. Esto es justificado en la dificultad que representaría obtener información fidedigna de una muestra probabilística de todos los operadores; la selección de los mismos es deliberada y se toma en cuenta a aquellos considerados informantes calificados.

Esta metodología de muestreo no probabilístico se puede enmarcar dentro del llamado *muestreo de juicio o de propósito*, en el cual la selección sigue un juicio o ideas arbitrarias de los técnicos en su búsqueda por una muestra representativa (Wolf et al., 2016, p. 328). Es susceptible a sesgos debido al grado de subjetividad que implica; realizar ajustes a medida que crece la experiencia resulta vital para obtener buenos resultados.

Estas características del problema pueden ser contextualizadas dentro del modelado de espacio-estado, donde se considera que se trabaja con observaciones ruidosas de una señal subyacente, siendo tal señal en este caso el precio promedio "verdadero" de las frutas y hortalizas. La magnitud inobservable del sesgo será en última instancia otro factor que distancie el modelo del proceso real de precios.

Según los técnicos del Observatorio Granjero, la calidad Extra en general representa menos del 5% del total en los distintos rubros. Por este motivo se la excluye del análisis, considerándose de mayor relevancia capturar y modelar el comportamiento de los precios de las otras dos calidades que representan el grueso de los bienes transados en el mercado. Observando las series de precios de los rubros separando por calidad (disponibles en la sección Sección 8.2) se puede apreciar que comparten un comportamiento similar, por lo que se considera que promediar en cada rubro las categorías I y II resulta un buen resumen de su evolución en cuanto precios. Dichas series de precios promedios serán las empleadas en los análisis.

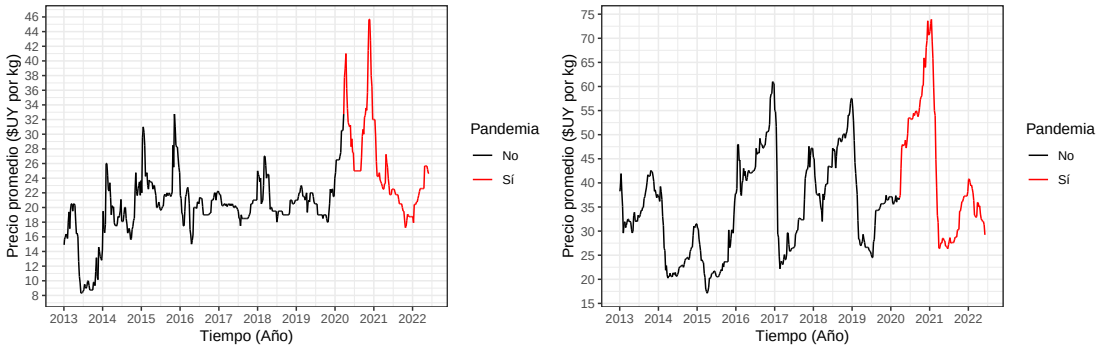
El traslado de la operativa del Mercado Modelo a la UAM, el cual a su vez fue simultáneo con la pandemia de COVID-19, puede representar un periodo de cambio en los registros, si bien la metodología de relevamiento no cambió fundamentalmente. Estos hechos llevan a prestar atención a posibles cambios estructurales.

Las series de precios son empleadas con frecuencia semanal, considerándose para cada semana el precio promedio de los dos precios relevados en ellas. A pesar de la simplificación que la serie semanal supone, permite apreciar la evolución de precios e ingresos y se gana en una mayor facilidad de manejo de las series respecto a los datos originales. El promedio se hace dentro de las semanas correspondientes al sistema de fechas ISO 8601, un sistema estándar definido sobre el calendario gregoriano de uso cotidiano. Un año tiene 365,25 días en promedio, lo que implica que tiene 52,1775 semanas de siete días en promedio. El sistema ISO opta por definir años con un número entero de semanas completas que comienzan el lunes y terminan el domingo, pudiendo ser este 52 o 53, para un total de 364 o 371 días por año. Bajo este esquema, en general los años cuentan con 52 semanas, a excepción del caso de los años donde el 1° de enero es jueves y los años bisiestos donde el 1° de enero es miércoles.

Como resultado de realizar los promedios para las series de cada rubro, se termina con 493 observaciones, con 52 observaciones en cada año, con excepción de los años 2015 y 2020, que cuentan con 53 semanas. Este número de semanas que cambia imposibilita el uso de metodologías clásicas de representar la estacionalidad de la serie (como las diferencias estacionales), que dependen de que el periodo estacional sea constante y entero. Aún simplificando la cantidad de semanas a 52, está presente el problema de que

es un periodo grande para el modelado mediante metodologías como los ARIMA, como se explicó anteriormente.

Para los distintos rubros se considera una muestra de entrenamiento que cuenta con 441 observaciones y va de la semana que comenzó el 3 de enero de 2013 a la que comenzó el 7 de julio de 2021. La muestra de prueba tiene 52 observaciones y va de la semana que comenzó el 14 de junio de 2021 hasta la que comenzó el 6 de junio de 2022. La excepción de esto es la papa, donde el comienzo de la muestra se recortó debido a tener un comportamiento notoriamente distinto al resto de los datos. Debido a la cantidad de datos con la que se cuenta y que se están removiendo observaciones del comienzo de la serie, esto no afecta sustancialmente el análisis. Para este rubro la muestra de entrenamiento va de la semana que comenzó el 30 de diciembre de 2013 hasta la que comenzó el 7 de junio de 2021 (389 observaciones). Por otro lado, la muestra de prueba va de la semana que comienza el 14 de junio de 2021 y termina en la semana que va de 6 de junio de 2022 (52 observaciones).



(a) Papa (b) Manzana

Figura 4.1: Series de precios semanales promedio por kg.

Como se puede apreciar en la Figura 4.1, entre los años 2014 a 2019 los precios de papa se mantuvieron en un intervalo estable, donde han ocurrido muy pocos eventos de fuertes alteraciones de precios. Es posible que este comportamiento se deba a que es un rubro de pocos productores y comerciantes que poseen capacidad logística y productiva, por lo que tienen capacidad de regular la oferta. Además, debe tenerse en cuenta que el elevar los precios por encima de un llamado "precio de referencia" implica que se importe el producto, lo cual también puede estar teniendo un factor disciplinante sobre el precio. Este precio de referencia es una cotización de equilibrio en la cual el productor tiene un margen de ganancia razonable y el consumidor enfrenta un precio que no es excesivo; mantenerse debajo de ese precio implica que no se de importación. Adicionalmente la papa industrializada, que en su mayoría se importa, es un sustituto fuerte de la papa fresca, lo cual también puede estar contribuyendo en el efecto disciplinante sobre el precio. Esta condición de concentración de la producción dificulta la predicción de precios, ya que puede que se estén definiendo por unos pocos actores, por encima de otros factores y la lógica de mercado.

A mediados de 2013 hubo una fuerte caída de precios y desde 2020 hay un aumento persistente hasta 2021, el cual se puede deber a las dificultades para la importación en el contexto de la pandemia de COVID-19. La atipicidad del año 2020 deberá ser evaluada a la hora de modelar porque podría llevar a malas predicciones en el periodo inmediato, ya que puede considerarse como un quiebre estructural transitorio.

En la misma figura se presenta la serie de precios de la manzana. Se puede apreciar un comportamiento estacional en el precio, en general comenzando bajo a principio del año, siendo marzo el mes que en general registra las cotizaciones más bajas, y aumentando a medida que avanza el año. La caída en los precios que se da luego del aumento a lo largo del año suele ser bastante abrupta. En general, de marzo a julio las manzanas almacenadas en cámaras comúnes mantienen buenas características de calidad. Posteriormente, hacia finales del invierno e inicios de primavera comienzan a intensificarse los problemas de conservación en parte de la oferta, lo que genera que se incrementen los valores mayoristas. Este periodo también es el inicio de la apertura de cámaras de atmósfera controlada, que permiten mejor conservación de la calidad. El producto de mayor calidad es el que tiene mayor potencial de conservación y el que se pone a disposición del público hacia el final del año, siendo su precio mayor.

Hay un aumento del rango de la serie hacia los últimos años y una tendencia a que el pico de precios se vuelva mayor, siendo el de fines de 2020 y principios de 2021 el más elevado (puede que sea un evento atípico a ser evaluado en el modelado). Esto puede indicar la necesidad de considerar que hay un elemento de tendencia.

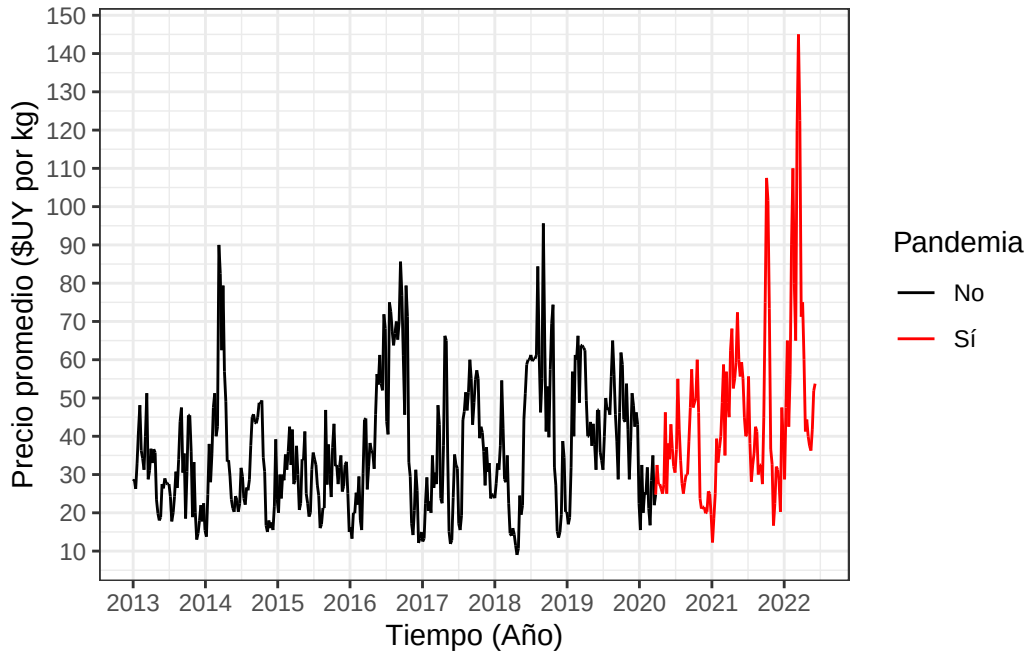
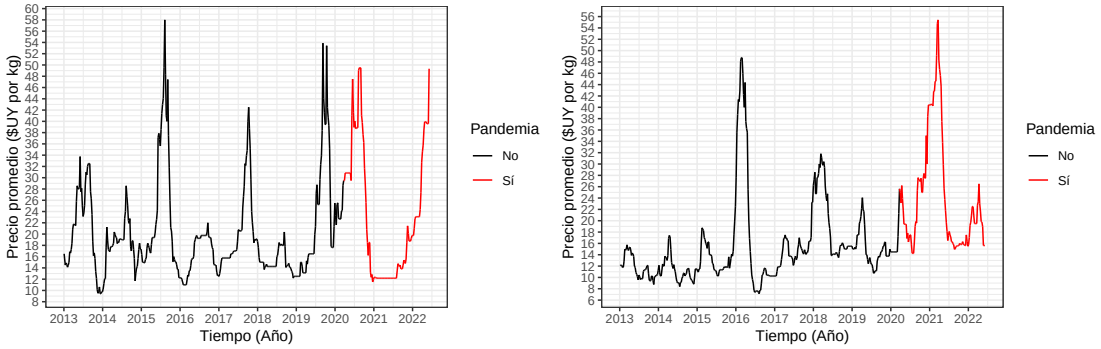


Figura 4.2: Series de precios semanales promedio por kg. Rubro: tomate.

Se presenta el gráfico de la evolución de los precios promedios del tomate en la Figura 4.2. Los precios generalmente están sujetos a comportamientos irregulares, debido a que la oferta es muy sensible a las condiciones ambientales y cualquier eventualidad climática puede ocasionar escasez o sobreoferta. Lo mismo sucede con la demanda, que es altamente variable y dependiente en gran medida de la temperatura y condiciones atmosféricas, lo cual también puede estar incidiendo en el precio. No son evidentes a primera vista un comportamiento que ocurra de forma sistemática, como la estacionalidad, por lo que esto deberá ser estudiados con el conjunto de metodologías que se trabajarán más adelante. Cabe destacar que en un principio el comienzo de la pandemia de COVID-19 no parece haber alterado sustancialmente el desarrollo de la serie.

En marzo de 2022 se registraron precios atípicamente altos, incluyendo el máximo histórico de la serie en la semana 11 ($145 por kilogramo de tomate). Dado que este valor se encuentra al final de la serie habrá que evaluar si resulta necesaria su intervención, dado que en este caso se espera que tenga un efecto sobre las predicciones mayor al que habría si tal evento hubiera ocurrido antes en el tiempo.



(a) Cebolla (b) Naranja

Figura 4.3: Series de precios semanales promedio por kg.

En la Figura 4.3 se puede observar la evolución del precio de la cebolla, resaltando en primera instancia los picos que se dan en algunos de los años, como 2015, 2017, 2019, 2020 y 2022. El año 2020 es el que cuenta con el mayor precio promedio anual, mientras que 2021 cuenta con el menor precio promedio anual, lo cual puede interpretarse como una manifestación del shock causado por la pandemia. A primera vista se puede apreciar cierto comportamiento estacional en los precios, donde hacia mitad del año hay un aumento sostenido que para mitad de año ya se ha revertido.

En términos generales, el precio de la naranja sigue un patrón donde los picos se dan en torno al comienzo del año (Figura 4.3). En los años 2016 y 2021 estos picos tomaron valores mucho mayores a los demás, lo cual puede causar alteraciones en el modelado y las predicciones, debido a que esta anomalía se encuentra a finales de la muestra.

4.3.1 Descomposición de las series de precios

En la presente sección, se realiza la descomposición de las series de precios mediante la metodología STR, para profundizar en su caracterización y visualizar patrones.

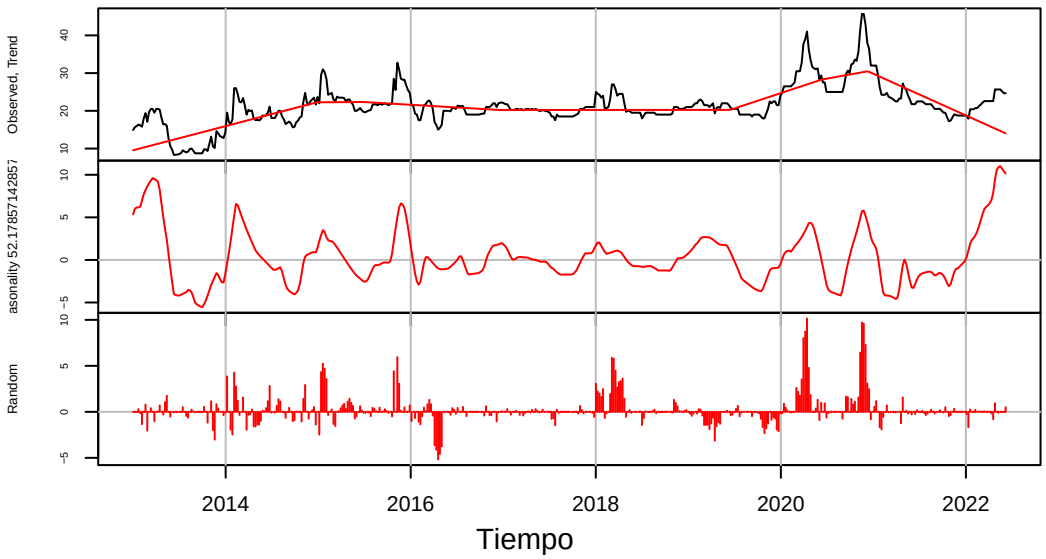


Figura 4.4: Descomposición STR de la serie de precios de papa.

Como se mencionó anteriormente, a primera vista el rubro papa presenta características que hacen difícil su predicción mediante el herramental de espacio-estado, dado que el elevado grado de concentración del rubro hace que haya cierta definición de los precios desde el lado de la oferta. Por esto, puede ser útil descomponer la serie para distinguir patrones subyacentes.

En la Figura 4.4 se muestra el gráfico de los componentes de la serie de precios promedio de papa resultado de la descomposición STR. En primera instancia, se observa que el elemento de tendencia permanece relativamente constante de 2014 a 2020 y cuando se ve alterado es por efecto de perturbaciones considerables más que por una tendencia en sí misma. Luego, el elemento estacional no parece presentar un patrón claro año a año o dentro de estos y en el elemento aleatorio o irregular es donde caen las anomalías correspondientes a los años 2020 y 2021.

En un claro contraste con la papa, ya a primera vista se aprecia un comportamiento sistemático en los precios de la manzana, en la forma de una estacionalidad aparentemente anual (Figura 4.5). El elemento estacional refleja que en los años el precio comienza siendo elevado, descendiendo hasta la mitad del año, momento a

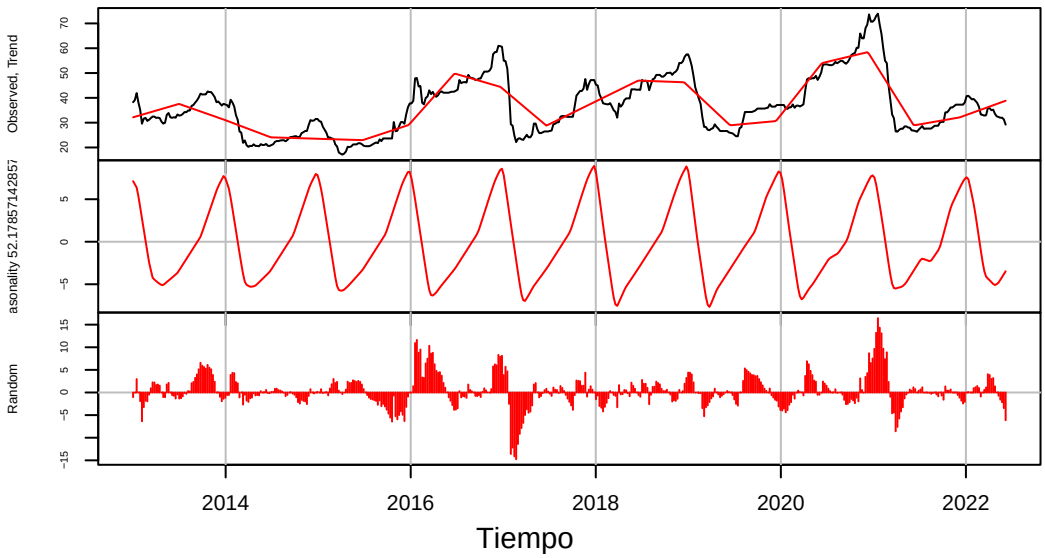


Figura 4.5: Descomposición STR de la serie de precios de manzana.

partir del cual vuelve a aumentar. Este componente presenta poca variación en el tiempo.

Por otro lado cabe destacar que a partir de 2016 es posible identificar un ciclo corto (parte superior de la Figura 4.5), donde cada dos años el precio vuelve a su nivel anterior. Esto es descripto por los especialistas del Observatorio Granjero como "añerismo", que es la hipótesis de que este comportamiento cíclico es el resultado de la observación de los precios del año pasado por parte de los productores. Cuando observan que los precios fueron altos, aumentan la siembra, lo cual tiene como resultado una baja de los precios. Al otro año, al observar esta baja, disminuyen la siembra.

Se pueden distinguir dos periodos donde el componente irregular toma más fuerza, que son en torno a los años 2016 y 2017 y luego a fines del año 2020. Si bien se podría considerar que la pandemia se manifiesta en el componente irregular como una pertubación al comportamiento sistemático de los precios, el efecto no resulta tan evidente como en el caso de la papa.

Dadas las características productivas del rubro, los precios del tomate presentan una mayor volatilidad que los rubros anteriormente tratados. Por este motivo resulta de gran utilidad la descomposición de la serie en sus elementos inobservables, para comprender en un mayor grado su comportamiento.

La visualización de la descomposición se presenta en la Figura 4.6. En primera instancia, se destaca que aparentemente la serie presenta una estacionalidad anual subyacente,

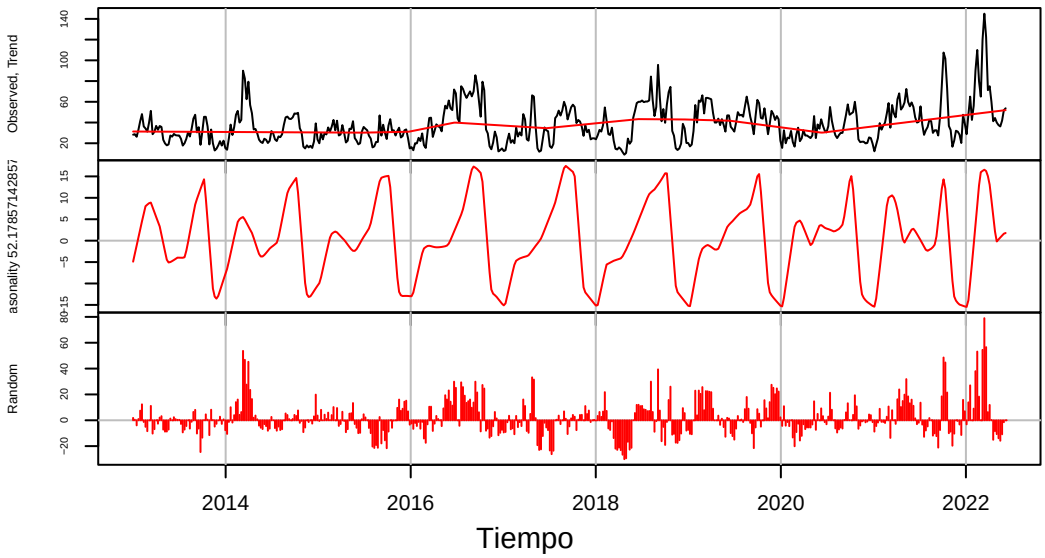


Figura 4.6: Descomposición STR de la serie de precios de tomate.

donde el pico de precios se da a fines de cada año. Esto indica que será necesario considerar un elemento estacional en la especificación del modelo. Luego, no parece haber una tendencia global al alza de precios, aunque el máximo histórico de precios está en el último año de la muestra (2022).

La descomposición STR del rubro cebolla se presenta gráficamente en la Figura 4.7. En la parte superior de la figura se puede observar el componente de tendencia-ciclo, pudiéndose apreciar cierto ciclo corto bianual, el cual puede ser una manifestación de un añerismo similar al encontrado para el rubro manzana. Por otro lado, el componente estacional presenta un patrón anual, donde a principios del año comienza bajo para ir aumentando hasta fines del año. El rango de variación de este patrón se va volviendo menor año a año. El componente irregular tiene una presencia fuerte en torno a los años 2015/2016 y el periodo de la pandemia.

Finalmente, en la figura Figura 4.8, la descomposición STR robusta de los precios de la naranja permite apreciar cierto ciclo corto de dos años y una estacionalidad con un periodo anual, donde el rango de valores de este elemento va en aumento año a año. Se destaca el comportamiento del elemento irregular en torno al año 2016 y 2021, donde hay picos de precios.

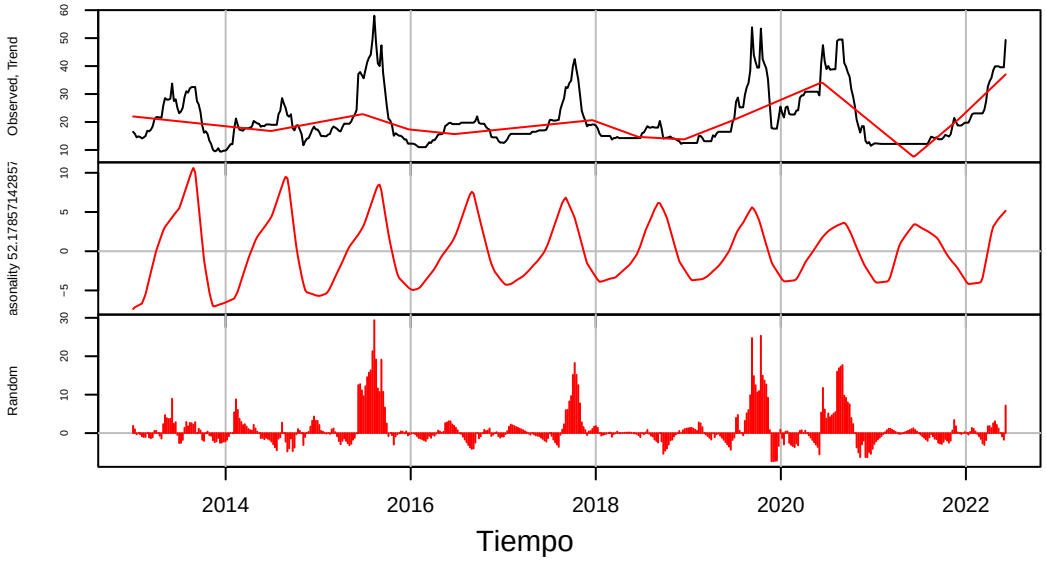


Figura 4.7: Descomposición STR de series de precios de cebolla.

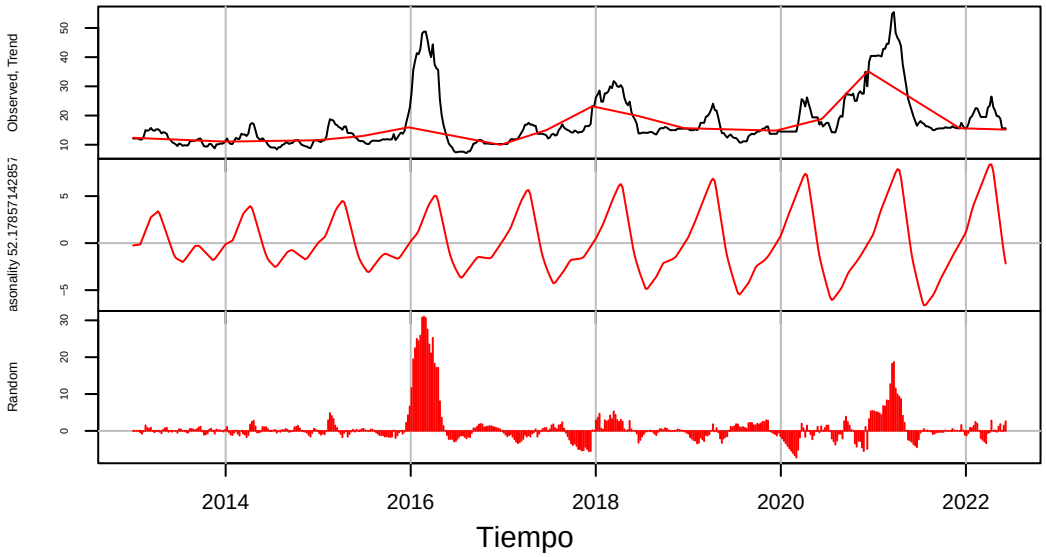


Figura 4.8: Descomposición STR robusta de series de precios de naranja.

5 Resultados

En esta sección se presenta una aplicación de los modelos de espacio estado al problema de la predicción de series semanales, haciendo uso de las metodologías presentadas en la sección Capítulo 3 y los datos descritos en la sección anterior. Se evalúan los modelos en tres horizontes predictivos:

- Corto: 1 semana, con validación cruzada en 12 semanas
- Mediano: 12 semanas
- Largo: 52 semanas

Los siguientes métodos y modelos son los aplicados: promedio simple (como referencia base), caminatas al azar (*Random walks*), ARIMA, ARIMA con estacionalidad representada con términos de Fourier y TBATS. Los modelos de corte estocástico están enmarcados en los modelos de espacio estado y desde esta perspectiva es que son estimados sus parámetros. En el horizonte de largo plazo, se consideran las predicciones mediante el promedio y los modelos estacionales.

5.1 Ajuste de modelos, selección y diagnóstico

5.1.1 Papa

Debe evaluarse el cumplimiento de los supuestos realizados a la hora del planteo de los modelos para constatar la validez de aplicarlos. Al realizarse las pruebas de diagnóstico de autocorrelación nula y normalidad de los residuos para los modelos de corte probabilístico, para todos ellos se rechaza la hipótesis de normalidad, lo cual puede deberse a los posibles quiebres estructurales que se pudieron observar en los gráficos de la serie y/o que el proceso generador de datos no siga este supuesto distribucional. En general se puede apreciar heterocedasticidad y varios residuos atípicos, factores que en general comprometen el cumplimiento del supuesto de normalidad. Debido a que varios componentes de la inferencia se fundamentan en el cumplimiento de dicho supuesto, como por ejemplo los intervalos de confianza asociados a la predicción, resulta necesario encontrar una manera de subsanar este problema. Una manera de hacerlo es emplear técnicas de bootstrap de series de tiempo, a partir de las cuales se estiman intervalos de manera no paramétrica. No obstante, debe tenerse en cuenta que no se puede realizar un muestreo como en el bootstrap de datos no temporales, debido a la autocorrelación que presentan las series de tiempo; para subsanar esto se realiza, por

ejemplo, bootstrap en bloques (Hyndman & Athanasopoulos, 2021). Adicionalmente, se debe tener en cuenta que esta falla de la hipótesis distribucional afecta incluso la estimación de los parámetros involucrados en el modelo, que también se fundamenta en ella. Empero, como se expresa en Hamilton (1994, p. 117), las estimaciones realizadas bajo este supuesto serán razonables, aunque no se cumpla.

El cumplimiento del supuesto de autocorrelación conjunta nula de los errores resulta clave, ya que en caso de que no se dé esto implicará que la dinámica de dependencia temporal y su estructura no está siendo captada correctamente, lo cual se ve reflejado en los residuos. Para el caso de los modelos planteados y considerándose el test conjunto para los primeros 99 rezagos (cantidad elegida con el criterio de Hyndman & Athanasopoulos (2021)), para todos los modelos no se rechaza la hipótesis de autocorrelación conjunta nula.

Tabla 5.1: Tabla de errores de predicción de precios a 12 pasos. Papa.

Método	RMSE	MAE	MAPE
Promedio simple	0.8	0.6	2.8
Random walk	**0.6**	**0.44**	**2.01**
ARIMA	0.63	0.51	2.34
ARIMA + Fourier	1.11	1.03	4.65
TBATS	**0.6**	**0.44**	**2.01**

Dado que la prioridad planteada es la predicción de valores futuros, este será el criterio principal de selección de modelos. Se opta por considerar solamente el horizonte predictivo de 12 semanas para este rubro, debido a la falta de comportamiento estacional en los precios que permita tal horizonte. Las medidas del error de la predicción a 12 semanas correspondientes a cada uno de los modelos aplicados se presentan en la Tabla 5.1. Estas predicciones son realizadas empleando la esperanza condicionada a los datos de la muestra de entrenamiento, con los que son estimados los parámetros de los distintos modelos. Los errores de predicción se calculan empleando estas predicciones y comparándolas con los verdaderos valores, correspondientes a la muestra de prueba. Se puede apreciar que el modelo de caminata al azar y el TBATS son los que tienen el menor error de predicción en todas las métricas, habiendo diferencias negligibles entre ellos. Se destaca que el error porcentual absoluto medio (MAPE en adelante) es del 2,01% para ambos modelos en este horizonte.

La predicción puntual de las caminatas al azar es la predicción *naive*, donde se predice con el último valor de la muestra de entrenamiento para todo el horizonte predictivo. Por lo tanto, en este caso no se gana en precisión predictiva complejizando la modelización más allá de considerar que el proceso de generación de datos es la perturbación aleatoria de la observación previa.

Tabla 5.2: Tabla de errores de predicción a 1 paso, validación cruzada de 12 semanas. Papa

Método	RMSE	MAE	MAPE
Promedio simple	0.82	0.61	2.85
Random walk	0.37	**0.23**	**1.06**
ARIMA	**0.35**	0.31	1.43
ARIMA + Fourier	0.4	0.36	1.64
TBATS	0.37	0.24	1.1

Para simular un proceso de predicción iterativa, en los primeros 12 valores de la muestra de prueba se realiza validación cruzada con las predicciones a un paso, agregándose a cada paso el dato observado que antes había sido predicho y calculándose las métricas de error en el conjunto de iteraciones. En la Tabla 5.2 se presentan las métricas de error para los modelos anteriormente planteados, pudiéndose observar que si bien el modelo ARIMA estimado cuenta con un menor RMSE, para las otras métricas planteadas la predicción de la caminata al azar es la de menor error.

La conclusión principal de estos resultados es que los intentos de modelización estocástica más elaborados que se aplicaron no dieron resultados substancialmente mejores que simplemente emplear el último precio con el que se cuenta como predicción, es decir, la predicción *naive* que surge de una caminata al azar. Como se adelantó anteriormente, el rubro cuenta con características que pueden estar dificultando el modelado mediante las metodologías empleadas, como por ejemplo el grado elevado de concentración de la producción del rubro. Esto hace que en ciertos periodos como el evaluado haya un estancamiento de precios y/o picos súbitos, los que distorsionan el comportamiento de la serie y consecuentemente implican que no se esté ganando en desempeño predictivo respecto a interpretar el proceso de formación de precios de la papa como una caminata al azar. Además de esto, la formación de precios fue afectada por la pandemia.

5.1.2 Manzana

Se ajustan a los precios de la manzana las mismas clases de modelos anteriormente descriptos cuando se trataron los precios de papa, permitiéndose que la especificación incluya, por ejemplo, el componente cíclico y el estacional en los modelos de espacio-estado.

Al poner a prueba los supuestos hechos al momento de la especificación, se obtiene como resultado que solo los modelos de caminata al azar con componente tendencia/ciclo, ARIMA con componente estacional de Fourier determinístico y el TBATS cumplen la hipótesis de autocorrelación conjunta nula de los residuos, lo cual es evidencia a favor de que estos modelos están captando la dinámica de dependencia temporal de la serie. Ninguno de ellos cumple el supuesto de normalidad de sus residuos. En este caso se

podría pensar que el proceso generador de datos de los residuos en general no sigue una distribución normal, aunque también puede ser el resultado de residuos atípicos.

Tabla 5.3: Tabla de errores de predicción a 12 pasos. Manzana.

Método	RMSE	MAE	MAPE
Promedio simple	9.42	9.39	34.06
Random walk	1.05	0.92	3.29
ARIMA	1.85	1.15	3.18
ARIMA + Fourier	1.47	1.19	4.25
TBATS	**0.95**	**0.73**	**2.61**

Tabla 5.4: Tabla de errores de predicción a 1 paso, validación cruzada de 12 semanas. Manzana.

Método	RMSE	MAE	MAPE
Promedio simple	9.3	9.27	33.64
Random walk	0.57	**0.43**	**1.56**
ARIMA	0.64	0.44	1.58
ARIMA + Fourier	0.62	0.46	1.65
TBATS	**0.56**	0.44	1.61

Se calculan las predicciones a 12 pasos condicionando a la muestra de entrenamiento y la predicción de validación cruzada para evaluar el desempeño predictivo de los modelos estimados. Los valores de las métricas de error de predicción a 12 pasos para los distintos modelos se presentan en la Tabla 5.3. En este caso el modelo TBATS es el de mejor desempeño predictivo en todas las métricas. Por otro lado, para la validación cruzada el resultado no es tan concluyente, a muy corto plazo no hay ningún modelo que supere ampliamente a la predicción *naive*.

Tabla 5.5: Tabla de errores de predicción a 52 pasos. Manzana.

Método	RMSE	MAE	MAPE
Promedio simple	5.7	4.59	15.3
ARIMA + Fourier	**5.1**	**3.82**	11.26
TBATS	5.14	3.86	**11.15**

Dada las características de comportamiento del rubro se valora posible realizar una predicción a un año (52 pasos), mediante el uso de los modelos con componentes estacionales. En la Tabla 5.5 se presentan las métricas de error predictivo de los dos

modelos estacionales y adicionalmente el de la predicción mediante el promedio simple a modo de *benchmark*. El modelo ARIMA con términos de Fourier es el de mejor desempeño en términos de RMSE y MAE, pero no en MAPE; aunque las diferencias con el modelo TBATS no son substanciales. Ambos tienen un desempeño considerablemente mejor a la predicción mediante el promedio simple.

5.1.3 Tomate

Al igual que en el caso de los rubros anteriores, se aplica el mismo conjunto de modelos, además de la predicción mediante el promedio simple como referencia.

Entre los modelos, el que cumple el supuesto de autocorrelación conjunta es el modelo ARIMA + Fourier. Para este modelo la normalidad de los residuos no se cumple.

Tabla 5.6: Tabla de errores de predicción a 12 pasos. Tomate.

Método	RMSE	MAE	MAPE
Promedio simple	**7.47**	**5.93**	**15.04**
Random walk	18.67	17.21	49.61
ARIMA	19.55	18.13	52.13
ARIMA + Fourier	27.13	25.25	72.15
TBATS	27.46	25.61	72.96

En la Tabla 5.6 se presentan las métricas de error para las predicciones a 12 pasos de los precios del tomate. Como se puede apreciar, en este caso la mejor predicción resulta del promedio simple, con una diferencia sustancial en todas las métricas. Esta deficiencia de los modelos aplicados se puede deber a la volatilidad de la serie, ante la cual las especificaciones planteadas no pueden captar toda la dinámica y también a la muestra de prueba tomada.

Lo mismo ocurre con la predicción a 1 paso de 12 semanas con validación cruzada (Tabla 5.7), aunque la diferencia con la modelización *Random Walk* es menor.

Tabla 5.7: Tabla de errores de predicción a 1 paso, validación cruzada de 12 semanas. Tomate.

Método	RMSE	MAE	MAPE
Promedio simple	**7.47**	**5.94**	**15.08**
Random walk	8.98	7.14	18.89
ARIMA	12.83	10.11	31.14
ARIMA + Fourier	9.46	7.88	21.29
TBATS	8.91	7.17	18.84

Tabla 5.8: Tabla de errores de predicción a 52 pasos. Tomate.

Método	RMSE	MAE	MAPE
Promedio	32.92	**20.96**	**32.21**
ARIMA + Fourier	**28.32**	22.85	53.37
TBATS	30.21	22.67	46.32

Considerándose una predicción a largo plazo (52 semanas), la situación es análoga. El promedio simple resulta superior en todas las métricas. Nuevamente, esto se puede justificar en cierta medida por la incertidumbre asociada a la serie, ahora sumada al largo horizonte predictivo. Por ejemplo, considerando el modelo ARIMA + estacionalidad de Fourier (Figura 5.3) se observa la gran amplitud de los intervalos de confianza, una medida de la incertidumbre asociada a la predicción puntual.

5.1.4 Cebolla

Se realiza la estimación del conjunto de modelos anteriormente empleados pero para los precios de cebolla. En lo que refiere al cumplimiento de los supuestos de los modelos estocásticos, la hipótesis de autocorrelación conjunta nula de los residuos no se rechaza para el modelo ARIMA, el ARIMA con términos de Fourier y el TBATS. Para ninguno de los modelos se cumple el supuesto de normalidad de los errores.

Tabla 5.9: Tabla de errores de predicción a 12 pasos. Cebolla.

Método	RMSE	MAE	MAPE
Promedio simple	7.95	7.88	62.26
Random walk	**1.23**	**0.69**	**4.82**
ARIMA	4.09	3.77	29
ARIMA + Fourier	5.38	4.97	38.43
TBATS	2.33	2.09	16.21

Tabla 5.10: Tabla de errores de predicción a 12 pasos. Validación cruzada. Cebolla.

Método	RMSE	MAE	MAPE
Promedio simple	7.85	7.78	61.48
Random walk	0.54	**0.26**	**1.82**
ARIMA	**0.51**	0.47	3.62
ARIMA + Fourier	1.13	1	8.02
TBATS	0.55	0.43	3.2

Las métricas de precisión de predicción a 12 pasos y de validación cruzada se presentan en la Tabla 5.9 y la Tabla 5.10. Como se puede apreciar, el modelo de caminata al azar es el de mejor desempeño, mientras que en la validación cruzada este modelo es el mejor en dos de las tres métricas empleadas (MAE y MAPE), el ARIMA se desempeña mejor en RMSE.

Considerando la predicción a largo plazo (Tabla 5.11), los modelos estacionales no superan al promedio simple de las observaciones. En este caso no están resultando apropiados para la predicción a largo plazo.

Tabla 5.11: Tabla de errores de predicción a 52 pasos. Cebolla.

Método	RMSE	MAE	MAPE
Promedio	**9.89**	**7.54**	**35.04**
ARIMA + Fourier	16.97	13.88	56.45
TBATS	15.38	11.72	44.33

5.1.5 Naranja

Luego de ajustar los modelos anteriormente descritos y poniendose a prueba los supuestos, el supuesto de autocorrelación nula de los residuos se cumple para los modelos ARIMA, ARIMA con términos de Fourier y TBATS. Para ninguno de los modelos se da el cumplimiento del supuesto de normalidad.

Tabla 5.12: Tabla de errores de predicción a 12 pasos. Naranja.

Método	RMSE	MAE	MAPE
Promedio simple	**1.23**	**1.02**	**6.39**
Random walk	3.33	3.16	19.52
Arima	1.59	1.48	8.93
ARIMA + fourier	1.38	1.16	7.09
TBATS	2.12	1.56	9.91

Tabla 5.13: Tabla de errores de predicción a 1 paso, validación cruzada de 12 semanas. Naranja.

Método	RMSE	MAE	MAPE
Promedio simple	1.23	1.02	6.38
Random walk	0.85	0.68	4
Arima	3.99	3.26	26.5
ARIMA + Fourier	**0.8**	**0.65**	**3.86**

Método	RMSE	MAE	MAPE
TBATS	0.86	0.77	4.66

El rendimiento predictivo de los modelos a 12 pasos y en la validación cruzada, expresado en las métricas de error, se presenta en la Tabla 5.12 y la Tabla 5.13. Para el primer tipo de predicción, el promedio simple es el que presenta el menor error. Por otro lado, en la predicción a un paso de validación cruzada el modelo ARIMA con términos de Fourier es la de mejor desempeño.

Tabla 5.14: Tabla de errores de predicción a 52 pasos. Naranja.

Método	RMSE	MAE	MAPE
Promedio simple	**2.77**	**2.15**	**11.29**
ARIMA + Fourier	4.39	3.69	20.42
TBATS	6.86	5.89	32.36

Evaluando las predicciones de largo plazo, se puede observar como el error predictivo de los modelos estacionales planteados resulta mayor que realizar la predicción mediante el promedio simple de las observaciones de la muestra de entrenamiento. El modelo más cercano en desempeño a esta predicción es el modelo ARIMA con términos de Fourier, aunque su MAPE es casi el doble. Esto puede ser una señal de que el tipo de modelo y las especificaciones planteadas no están siendo capaces de captar la dinámica de largo plazo de la serie.

5.2 Predicciones

Se presentan las predicciones a mediano o largo plazo para los distintos rubros trabajados, haciendo uso de los modelos de mejor desempeño.

Para el rubro papa y considerando un horizonte predictivo de medio plazo (12 semanas), el modelo estimado de mejor desempeño es una caminata al azar, donde las varianzas de los procesos observados y de estado son 0,0013 y 2,22 respectivamente. La predicción del precio promedio es la misma para todo el horizonte: 21,75. Se da un aumento de la amplitud de los intervalos de confianza a medida que se vuelve más lejano el horizonte predictivo, lo cual refleja la incertidumbre creciente asociada. Como se mencionó anteriormente, el MAPE, que se interpreta como el porcentaje de error absoluto promedio, toma el valor 2,01%. Se presenta la visualización de la predicción a 12 pasos en la Figura 5.1.

Para el rubro manzana se presentan los resultados del modelo TBATS, debido a que en términos generales presenta un rendimiento balanceado. Presenta mejor rendimiento a

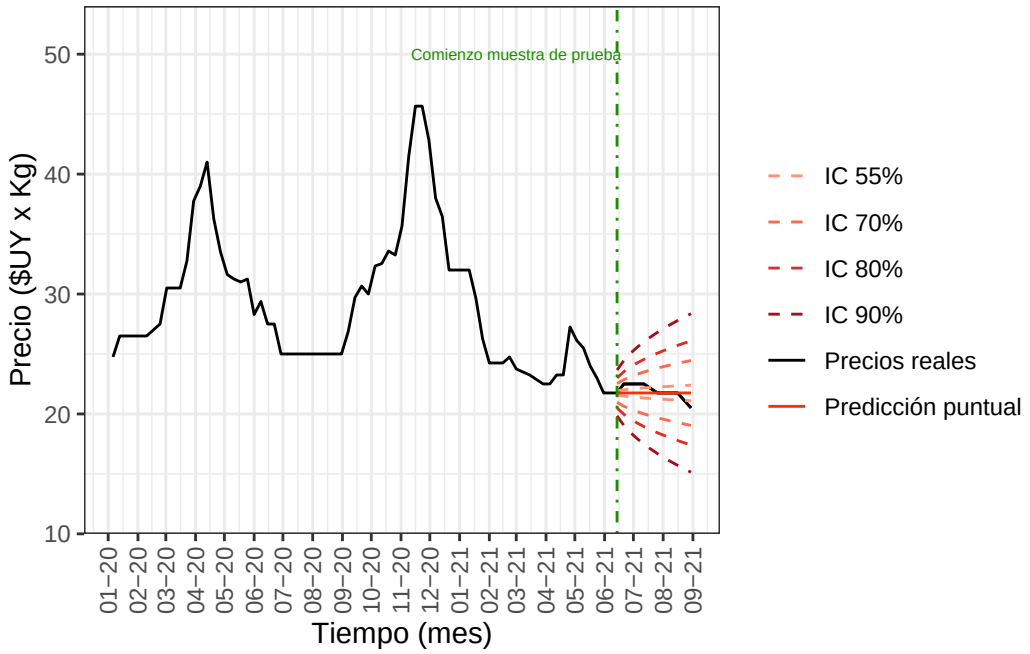


Figura 5.1: Predicción a 12 pasos. Papa. Random Walk.

12 pasos que el modelo ARIMA y para la predicción a 52 pasos la diferencia con este último es mínima.

El modelo estimado es un TBATS donde el parámetro de la transformación Box-Cox es $\omega = 0,015$, los parámetros del modelo ARMA(p,q) de los errores son ambos 0 (es un ruido blanco), su desvío estandar estimado es $\sigma = 0,0504$, el parámetro de amortiguación del componente de tendencia es $\phi = 0,8$ y son empleados 3 armónicos para representar el componente estacional anual $m = 52,18$. Los parámetros de suavizado de los errores ARMA para el componente de nivel local, la tendencia de corto plazo y de los componentes estacionales son $\alpha = 1,0093$, $\beta = 0,113$, $\gamma_1 = 0,00045$ y $\gamma_2 = -0,00042$.

En la Figura 5.2 se presentan las predicciones a 52 semanas para los precios de la manzana mediante el modelo TBATS. Estas van desde la semana que comienza el 14/05/2021 a la que comienza el 06/06/2022 y adicionalmente se presentan los verdaderos valores de los precios observados. Se destaca que el comportamiento de los precios tiene un buen nivel de precisión en este horizonte, dado su MAPE del 11,15%. A pesar de esto, desde marzo de 2022 el modelo comienza a flaquear, prediciendo un crecimiento de los precios mientras que los reales descienden. No obstante, se debe tener en cuenta que esto ocurre al final del horizonte predictivo, donde la incertidumbre es mayor para el modelo, lo que se ve reflejado en la amplitud de los intervalos de confianza.

Si bien para el rubro tomate a corto y mediano plazo la predicción considerando el promedio simple es la de mejor desempeño, a largo plazo el modelo ARIMA con términos

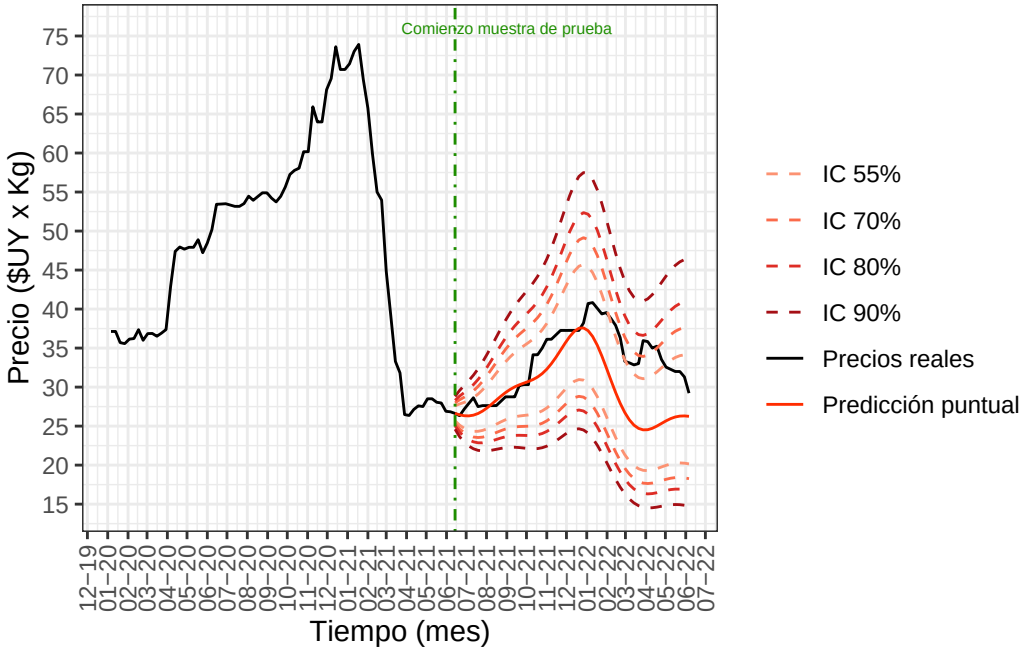


Figura 5.2: Predicción a 52 pasos. Manzana. TBATS

de Fourier capta el comportamiento estacional del rubro y permite el cálculo de intervalos de predicción. Por este motivo, se presentan sus resultados además de la predicción mediante el promedio simple.

El modelo ARIMA en cuestión es un ARIMA(0, 1, 4), donde los parámetros de medias móviles son: $\theta_1 = -0,2153$, $\theta_2 = -0,3019$, $\theta_3 = -0,1145$ y $\theta_4 = 0,0827$, la estacionalidad se representa con 3 términos de Fourier y el desvío de los términos de error es $\sigma = 9,8$.

Para comparar, se grafica la predicción empleando el promedio simple, que es 36,6 para el periodo de prueba.

En comparación con el resto de los rubros, el error en términos porcentuales promedio es mayor, siendo el MAPE para el promedio simple 15,04% y 72,15% para el modelo ARIMA + Fourier. Esto refleja la dificultad para predecir los precios del tomate con los modelos aplicados.

Como se describió anteriormente en el Capítulo 4, el tomate es un rubro cuyo volumen de producción se ve afectado rápidamente por un conjunto de factores rápidamente cambiantes (como lo son las condiciones climáticas), lo cual dificulta su predicción. Incluso contándose con las variables que cuantifican estos factores, resulta necesaria su predicción para ser incorporada al modelo, lo cual es una tarea que excede el alcance de este trabajo.

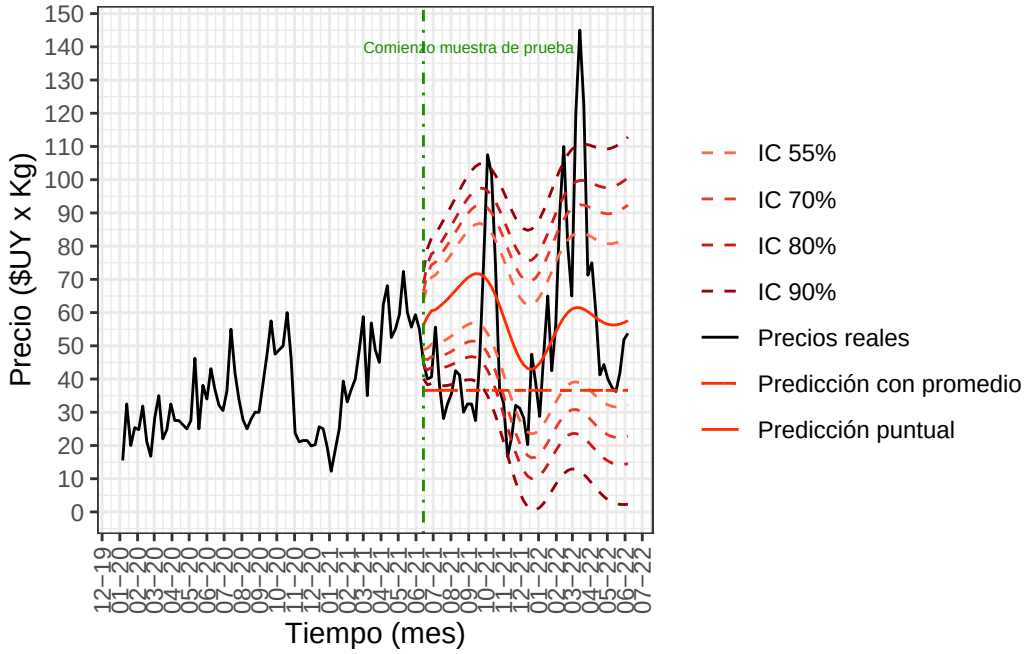


Figura 5.3: Predicción a 52 pasos, tomate.

Se presentan las predicciones a mediano plazo del modelo de mejor desempeño para el rubro cebolla, una caminata al azar. Como se puede apreciar, el comportamiento atípico de los precios a lo largo de 2021, donde se mantuvieron fijos luego de picos súbitos en 2020, puede ser una causa del pobre desempeño de los otros modelos.

El modelo estimado es una caminata al azar donde la varianza del proceso observado es 0,001 y del proceso de estado 6,66. La predicción puntual de la caminata al azar es 12,17 para todo el horizonte predictivo y el MAPE resultante es del 4,82% en este caso.

Si bien la predicción en el mediano plazo tiene un buen grado de precisión dada la estabilidad de los precios en el periodo de prueba, quiebres estructurales como el que se dio unos meses antes de que comenzara son difíciles de predecir para cualquier tipo de modelo.

En el caso de la naranja, se presenta la predicción del promedio simple, que es la de mejor desempeño, junto a la del modelo ARIMA con estacionalidad representada por términos de Fourier, que es el que le seguía en desempeño a largo plazo y tiene como punto a favor que la predicción es más informativa de las variaciones resultantes del componente estacional de la serie. Este último es un modelo $ARIMA(4,1,0)$, donde los parámetros de la parte autorregresiva son $\phi_1 = 0,2082$, $\phi_2 = 0,1123$, $\phi_3 = 0,1321$ y $\phi_1 = 0,1084$, mientras que el desvío estandar es $\sigma = 1,58$. La estacionalidad se considera en este contexto como determinística y se modela mediante dos componentes de Fourier.

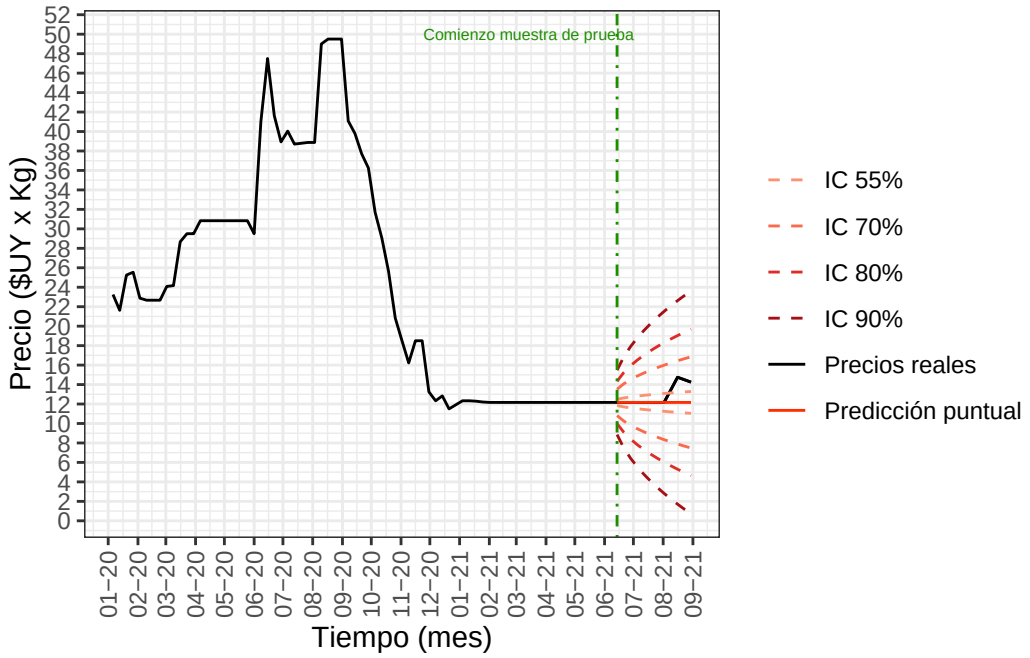


Figura 5.4: Predicción a 12 pasos. Cebolla. Random Walk.

La predicción mediante el promedio es de 17, 26 para todo el periodo de evaluación, la cual se grafica para comparar. En la Figura 5.5 se puede ver como el modelo ARIMA + Fourier capta el comportamiento estacional, a pesar de su peor desempeño en las métricas de error. El error porcentual absoluto promedio para la predicción mediante el promedio es del 6,39% y para el modelo 7,09%.

A modo de resumen, los métodos de predicción de mejor desempeño para cada rubro y horizonte temporal se presentan en la Tabla 5.15. Como se puede observar, no hay un modelo que predomine sobre el resto para todos los rubros y horizontes predictivos.

Tabla 5.15: Modelo o método de mejor desempeño para cada rubro y horizonte predictivo.

	Corto plazo	Mediano Plazo	Largo plazo
Papa	Random Walk	Random Walk	-
Manzana	Random walk	TBATS	ARIMA+Fourier
Tomate	Promedio simple	Promedio simple	Promedio simple
Cebolla	Random Walk	Random Walk	Promedio simple
Naranja	ARIMA+Fourier	Promedio simple	Promedio simple

En la Figura 5.6 se presentan las métricas de bondad predictiva a 12 pasos, para cada

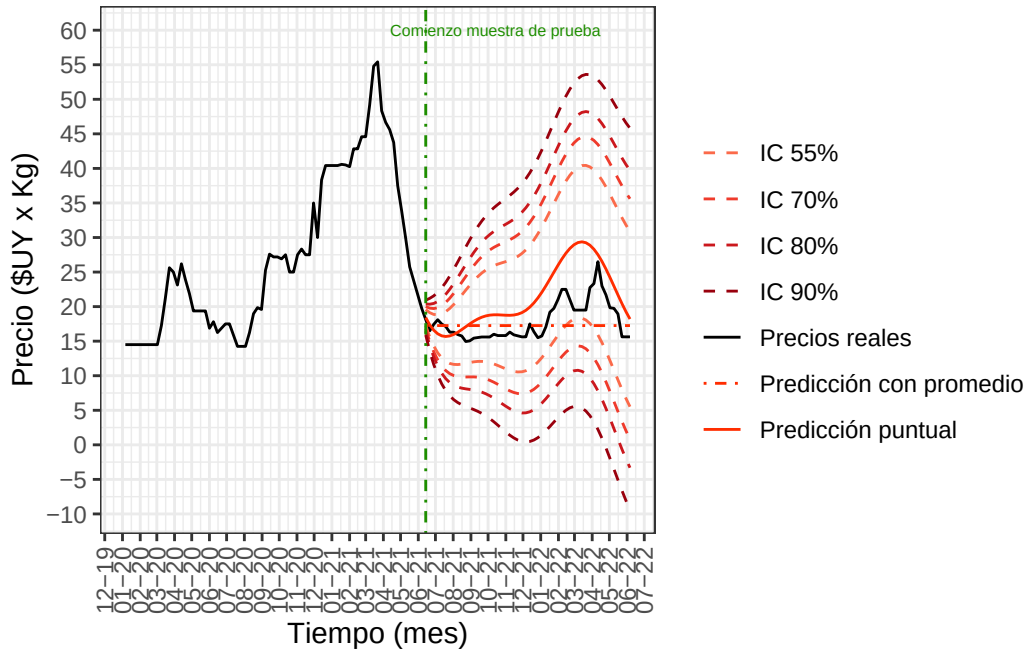


Figura 5.5: Predicción a 52 pasos. Naranja. ARIMA + Fourier.

modelo y rubro[1]. A través de esta visualización se puede apreciar que en general las tres métricas están en consonacia respecto a los modelos de mejor desempeño y cómo varían los márgenes de diferencia entre los distintos modelos; en el caso de la cebolla es considerablemente mayor al resto y para la papa es menor.

[1]Se excluye al tomate, dado que presenta valores altos en la métricas de error que no permite apreciar las del resto de los rubros

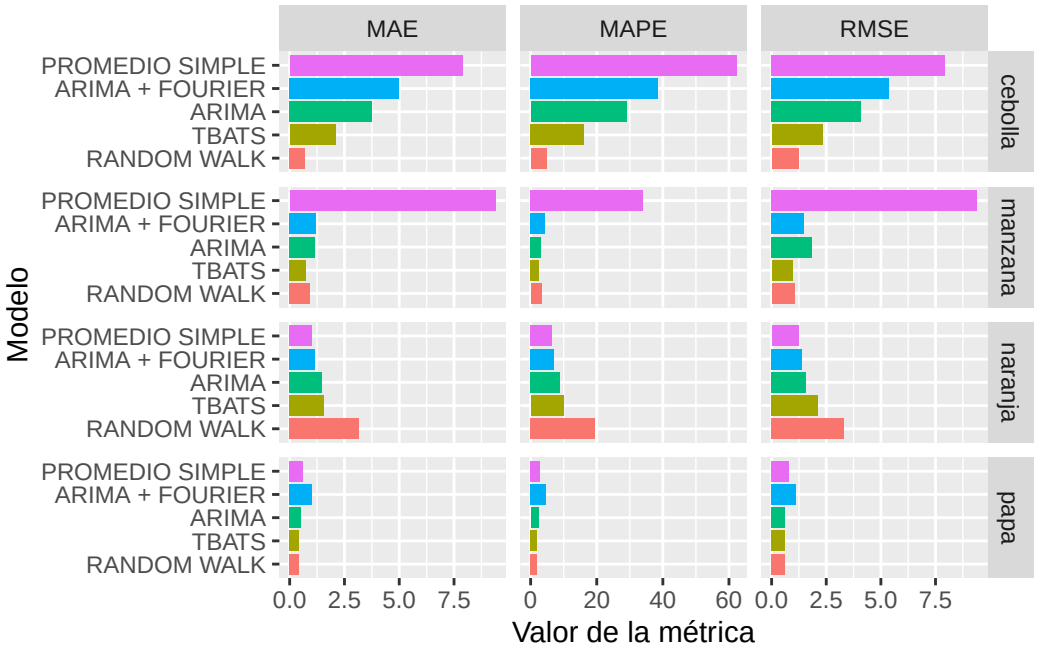


Figura 5.6: Métricas de bondad predictiva según rubro y modelo, predicción a 12 pasos.

6 Conclusiones y pasos futuros

Los modelos de espacio estado representan una opción paramétrica viable y relativamente sencilla de aplicar para enfrentarse a la problemática del modelado de series semanales. Hay planteos previos de especificaciones presentadas por varios autores, implementaciones computacionales y además los modelos permiten la flexibilidad suficiente para ser adaptados a distintas aplicaciones.

En lo que refiere a los ejemplos de aplicación que se trataron en este trabajo, no se puede concluir que hay un tipo de modelo particular que tenga desempeño superior para todos los horizontes predictivos y todas las series consideradas, sino que para cada caso hay un método de predicción particular que se destaca por sobre los otros. Esto en parte responde a la diversidad de las series consideradas y a que en algunas casos las dinámicas de estas son dificiles de modelar.

Para la manzana es donde los modelos de espacio estado estacionales planteados tienen en general el mejor desempeño a través de los distintos horizontes. Los precios de rubros como la papa y el tomate resultan difíciles de predecir a cualquier horizonte predictivo. El primero debido a la concentración de la oferta, que hace que los precios dependan de las decisiones de pocos actores. En el caso del segundo se debe al elevado grado de variabilidad que presentan sus precios. La cebolla y la naranja comparten que cuentan con cierto componente estacional pero afectado por periodos con picos elevados de precios. Esto último es un factor que dificulta la predicción de sus precios.

Como posibles líneas de investigación futura se consideran la inclusión de variables auxiliares regresoras, la aplicación de otras especificaciones de modelos de espacio-estado y a otras series temporales. Adicionalmente, también queda abierta la posibilidad de emplearse otro tipo de modelos, como la aplicación de la técnica STR en su forma predictiva, redes neuronales y otro tipo de modelos de *deep learning*, que en algunos casos permiten lidiar con series de alta frecuencia.

7 Bibliografía

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2024). *Quarto* (Versión 1.4). https://doi.org/10.5281/zenodo.5960048

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra

Beijers, D. (2023). *statespacer: State Space Modelling in R*. https://dylanb95.github.io/statespacer/

Casares, I. (2022). *Encuesta de papa: Primavera 2021*. Ministerio de Ganadería, Agricultura y Pesca, DIEA.

Cleveland, W. P., Evans, T., & Scott, S. (2014). *Weekly Seasonal Adjustment - A Locally-weighted Regression Approach* (N.º 473). U.S. Bureau of Labor Statistics.

Comisión Administrativa del Mercado Modelo. (2021). *Manual de procedimientos y referencias técnicas para la tipificación de la calidad de frutas y hortalizas frescas*. Unidad Agroalimetaria Metropolitana.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, *106*, 1513-1527.

Dieng, A. (2008). Alternative Forecasting Techniques for Vegetable Prices in Senegal. En *Revue sénégalaise des recherches agricoles et agroalimentaries* (N.º 3; Vol. 1).

Dokumentov, A., & Hyndman, R. J. (2022). STR: Seasonal-Trend Decomposition Using Regression. *INFORMS Journal on Data Science*, *1*(1), 50-62. https://doi.org/10.1287/ijds.2021.0004

Dokumentov, A., & Hyndman, R. J. (2023). *stR: STR Decomposition*. https://CRAN.R-project.org/package=stR

Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

Durbin, J., & Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Godahewa, R., Bergmeir, C., Webb, G. I., & Montero Manso, P. (2020). *A Strong Baseline for Weekly Time Series Forecasting*. https://arxiv.org/abs/2010.08158#:~:text=Many%20businesses%20and%20industries%20require,approaches%20dedicated%20to%20this%20task.

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3), 1-25. https://www.jstatsoft.org/v40/i03/

Güenaga, M. (2014). *Componentes variables del IPC: Frutas y Verduras*. Trabajo final de grado. Universidad de la República. Facultad de Ciencias Económicas y de Administración. https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/

22574
Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.
Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation Metrics for Machine Learning*. https://CRAN.R-project.org/package=Metrics
Harvey, A. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.
Harvey, A., Koopman, S. J., & Riani, M. (1997). The Modeling and Seasonal Adjustment of Weekly Observations. *Journal of Business & Economic Statistics*, *15*, 354-368.
Hernández-Banadik, M., Álvarez-Castro, I., Da Silva, N., & de Mello, S. (2021). *Modelos Bayesianos para series diarias: Modelado de temperaturas extremas en Uruguay*. (Serie Documentos de Trabajo; 4/21). Universidad de la República (Uruguay). Facultad de Ciencias Económicas y de Administración, IESTA. https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/34862
Hyndman, R. J. (2014). *Thoughts on the Ljung-Box test*. https://robjhyndman.com/hyndsight/ljung-box-test/.
Hyndman, R. J., & Anne B. Koehler. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679-688.
Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice*. OTexts.
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1-22. https://doi.org/10.18637/jss.v027.i03
Instituto Nacional de Estadística. (2022a). *Cotizaciones al publico de las principales monedas*. Recuperado 21 de diciembre de 2022 de https://www.gub.uy/instituto-nacional-estadistica/datos-y-estadisticas/estadisticas/cotizacion-monedas.
Instituto Nacional de Estadística. (2022b). *Estadísticas Económicas*. Recuperado 21 de diciembre de 2022 de https://www.ine.gub.uy/.
Instituto Nacional de Investigación Agropecuaria. (2024). *Banco de Datos Agroclimáticos*. Portal INIA. http://www.inia.uy/gras/Clima/Banco-datos-agroclimatico.
Li, G., Xu, S., & Li, Z. (2010). Short-Term Price Forecasting For Agro-products Using Artificial Neural Networks. *Agriculture and Agricultural Science Procedia 1*, 278-287.
Ljung, G. M., & Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, *65*(2), 297-303.
López-de-Lacalle, J. (2019). *tsoutliers: Detection of Outliers in Time Series*. https://CRAN.R-project.org/package=tsoutliers
Luo, C., Wei, Q., Zhou, L., Zhang, J., & Sun, S. (2011). Prediction of Vegetable Price Based on Neural Network and Genetic Algorithm. *Computer and Computing Technologies in Agriculture IV*, 672-681.
Millán, J., & Romero, D. (2019). *Aportes para la construcción de un modelo de predicción de precios mayoristas de frutas y hortalizas en el Uruguay*. Tesis de grado. Universidad de la República. Facultad de Agronomía. https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/29570
Mitra, D., & Paul, R. K. (2017). Hybrid time-series models for forecasting agricultural

commodity prices. *Model Assisted Statistics and Applications*, *12*, 255-264.
Mohd-Razali, N., & Yap, B. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Analytics*, *2*.
Müller, K. (2020). *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here
Müller, K., & Wickham, H. (2023). *tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble
Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer
Peng, Y.-H., Hsu, C.-S., & Huang, P.-C. (2015). *Developing Crop Price Forecasting Service Using Open Data from Taiwan Markets* (pp. 172-175). IEEE.
Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic Linear Models With R*. Springer Science + Business Media.
R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Royston, P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(2), 115-124. http://www.jstor.org/stable/2347973
Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality (complete samples)†. *Biometrika*, *52*(3-4), 591-611. https://doi.org/10.1093/biomet/52.3-4.591
Student test. Encyclopedia of Mathematics. (s. f.). https://encyclopediaofmath.org/index.php?title=Student_test.
West, M., & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer New York.
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org
Wickham, H. (2022). *stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr
Wickham, H. (2023). *forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats
Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr
Wickham, H., Hester, J., & Bryan, J. (2023). *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr
Wickham, H., & Seidel, D. (2022). *scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales
Wickham, H., Vaughan, D., & Girlich, M. (2023). *tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr
Wolf, C., Joye, D., Smith, T. W., & Fu, Y. (2016). *Forecasting, Structural Time Series Models and the Kalman Filter*. SAGE Publications Ltd.
Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in R. En V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Computational Research*. Chapman; Hall/CRC.

Young, P. C., Pedregal, D. J., & Tych, W. (1999). Dynamic Harmonic Regression. *Journal of Forecasting*, *18*, 679-688.

8 Anexo

8.1 Series de variables auxiliares

Inicialmente, se consideró la construcción de variables auxiliares para incluir en el modelado de las series de precios, con la finalidad de incrementar la capacidad predictiva de los modelos. En última instancia, esta especificación no fue profundizada, debido a que para los modelos donde fue aplicado no implicó un aumento en la capacidad predictiva, o en otros casos, como en el TBATS, su implementación supone una dificultad considerable. Por estos motivos, se optó por dejar este elemento de la especificación para trabajos futuros, pero se incluye lo que se avanzó sobre las variables durante la escritura del trabajo a modo informativo.

Las series auxiliares disponibles para el periodo considerado son las de ingresos de los distintos rubros al Mercado Modelo y la UAM, es decir, el volumen de mercadería que ingresa a plaza. Se cuenta con estimaciones semanales, dado que no todos los transportistas declaran el volumen de productos con el que cuentan a su ingreso. Esta información no se encuentra desglosada por calidades y también es susceptible a errores de estimación, aunque con el traslado de las actividades a la UAM los registros presentan una mayor fiabilidad al haber una única entrada para los transportes (en el antiguo Mercado Modelo habían 13 entradas, dificultando el registro). Se pueden visualizar las series en la Sección 8.1.1.

También se seleccionó un conjunto de variables auxiliares consideradas de importancia para la explicación de fenómenos que afectan los precios de los rubros, las cuales se esperaría que mejoren la capacidad de predicción al incluirse en una lógica de regresión. Tal conjunto fue sopesado en varias reuniones con técnicos de la UAM, considerándose inicialmente un grupo deseable amplio, más allá de consideraciones de disponibilidad de los datos.

Posteriormente, se procedió a la búsqueda de las distintas series especificadas. En algunos casos fue posible encontrar series completas para el periodo considerado, recurriendo por un lado a informes de organismos nacionales como la Dirección de Estadísticas Agropecuarias (DIEA) del Ministerio de Ganadería, Agricultura y Pesca (MGAP), las mediciones meteorológicas del Instituto Nacional de Investigación Agropecuaria (INIA), el Banco Central del Uruguay (BCU) y por otro a los de organismos internacionales como la *Centrais de Abastecimiento do Rio Grande do Sul*, entre otros. No obstante, no fue posible construir muchas de las variables que en un principio se habían planteado debido a la ausencia total o parcial de datos.

Una de las mayores problemáticas que se encontró al incluir variables auxiliares en las especificaciones de los modelos fue que, en la mayoría de los casos, no es claro cuánto es el periodo de rezago entre que se da un fenómeno medido por las variables (por ejemplo: la lluvia medida en milímetros) y su impacto en los precios. Tal ambigüedad dificulta la implementación de los modelos y suma otra fuente de variabilidad, por lo que se vio acotado el conjunto de variables efectivamente empleado. El listado de las variables sistematizadas y sus respectivas fuentes se presentan a continuación:

- Papa:
  - Area Sembrada anual (Casares, 2022)
  - Rendimiento anual (Casares, 2022)
  - Producción anual (Casares, 2022)
  - Cotización en la región (datos de Conab: https://www.conab.gov.br/)
  - Precipitaciones (Instituto Nacional de Investigación Agropecuaria, 2024)
  - Cotización del dólar (Instituto Nacional de Estadística, 2022a)
- Manzana:
  - Producción anual estimada (Estimaciones de DIGEGRA-MGAP)
  - Horas de frío acumuladas (Instituto Nacional de Investigación Agropecuaria, 2024)
  - Precipitaciones (Instituto Nacional de Investigación Agropecuaria, 2024)
  - Meses desde la cosecha (elaboración propia)
  - Índices de costo de fertilizante, combustible, energía, mano de obra (Instituto Nacional de Estadística, 2022b)
- Tomate:
  - Producción anual estimada (Estimaciones de DIGEGRA-MGAP)
  - Estacionalidad de la demanda (elaboración propia)
- Cebolla:
  - Producción anual estimada (Estimaciones de DIGEGRA-MGAP)
  - Área sembrada anual estimada (Estimaciones de DIGEGRA-MGAP)
  - Precipitaciones (bajas y exceso hídrico) (Instituto Nacional de Investigación Agropecuaria, 2024)
- Naranja:
  - Producción anual estimada (Estimaciones de DIGEGRA-MGAP)
  - Demanda de verano (elaboración propia)

8.1.1 Gráficos de ingresos de mercadería

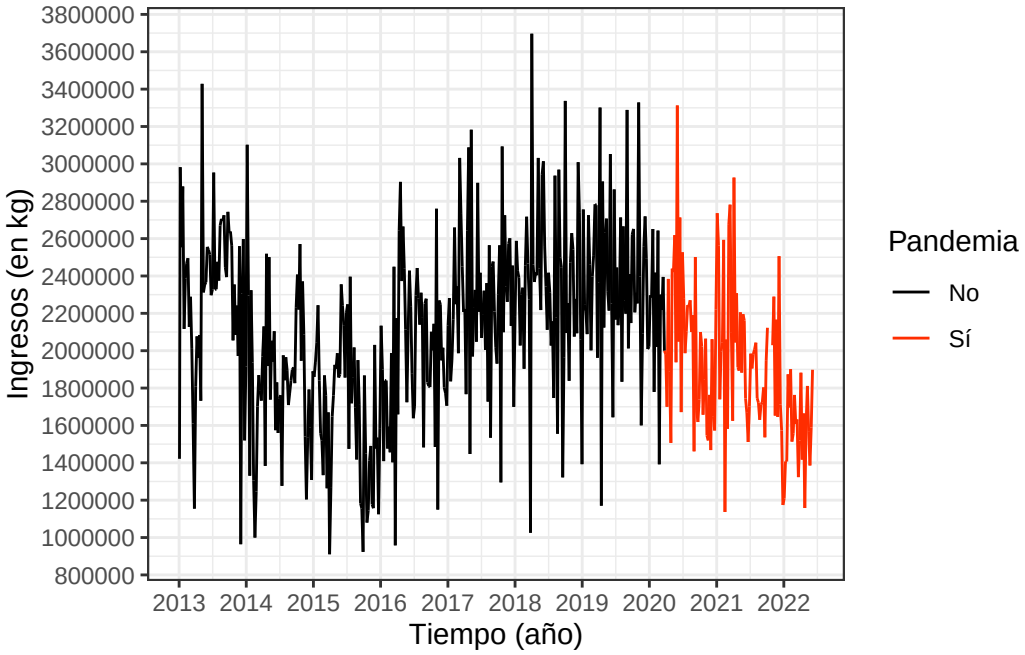


Figura 8.1: Serie de ingresos de mercadería semanales. Rubro: papa.

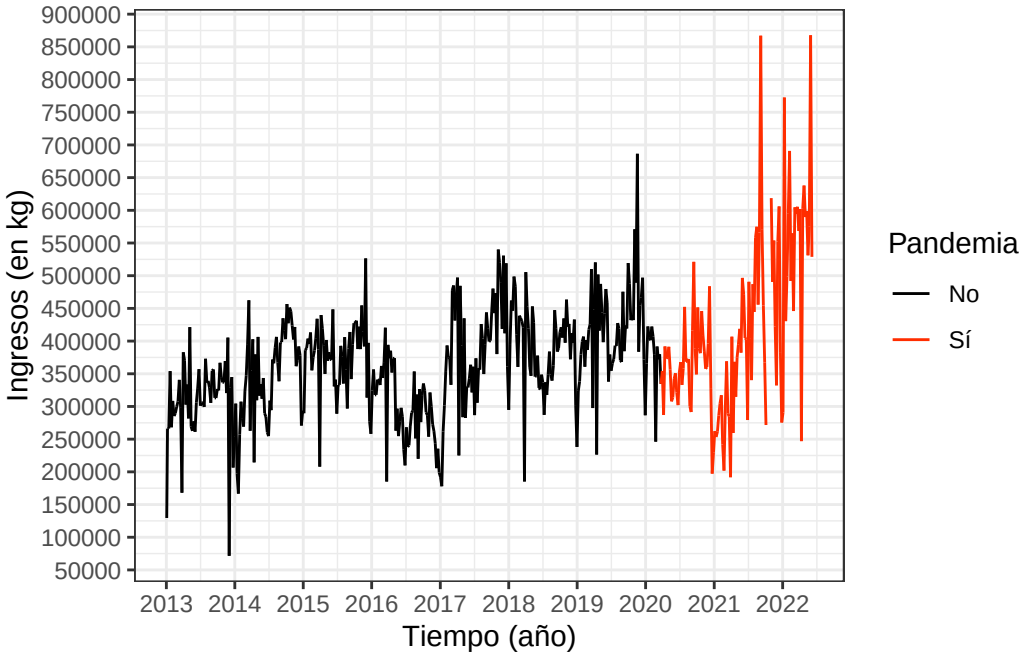


Figura 8.2: Serie de ingresos de mercadería semanales. Rubro: manzana.

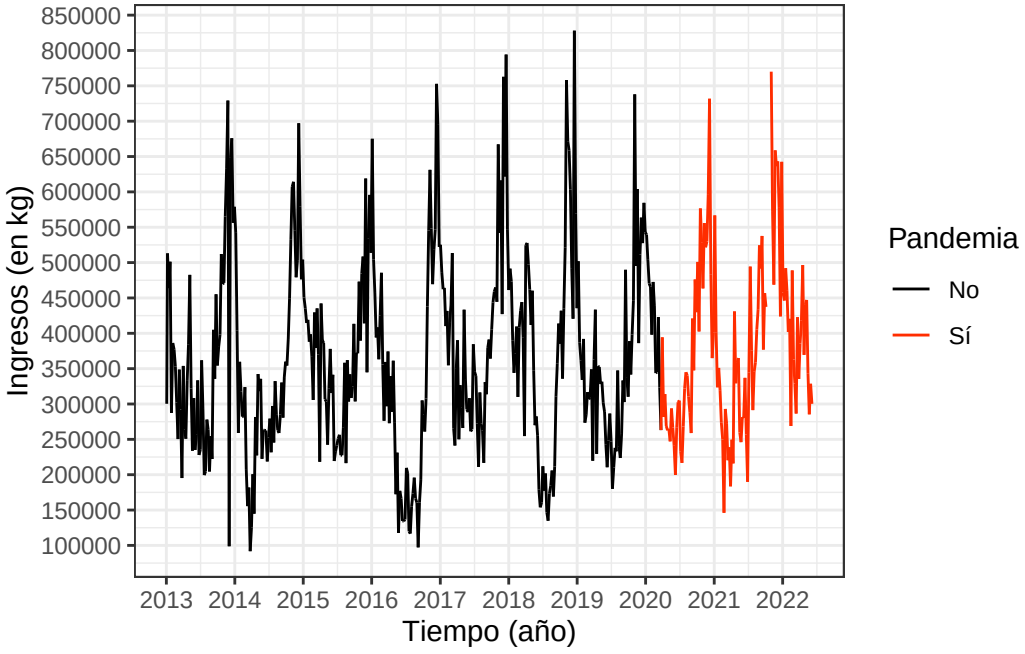


Figura 8.3: Serie de ingresos de mercadería semanales. Rubro: tomate.

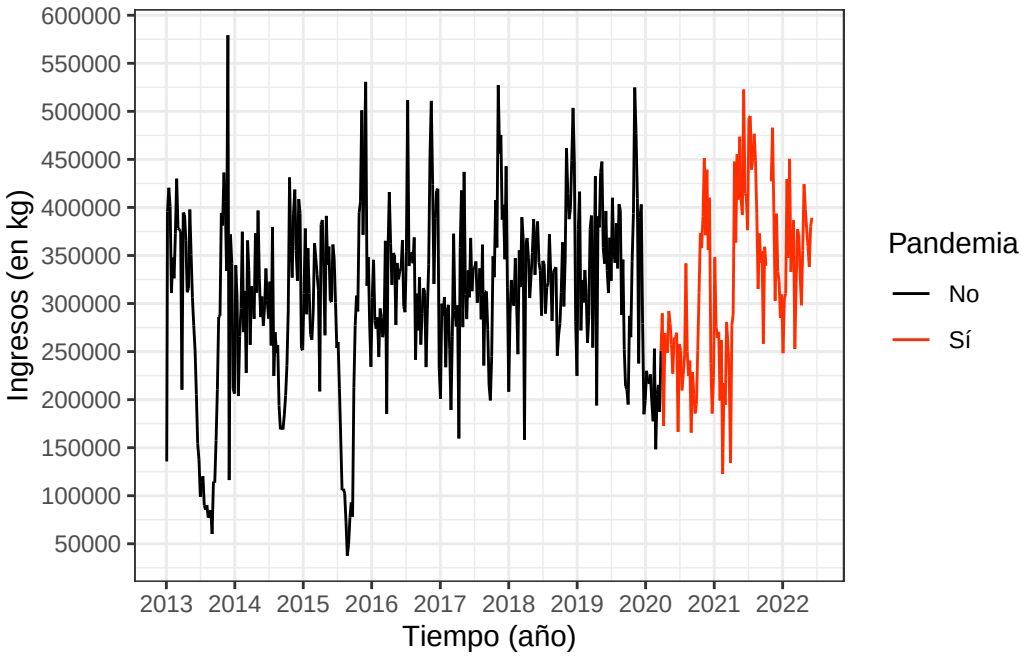


Figura 8.4: Serie de ingresos de mercadería semanales. Rubro: cebolla.

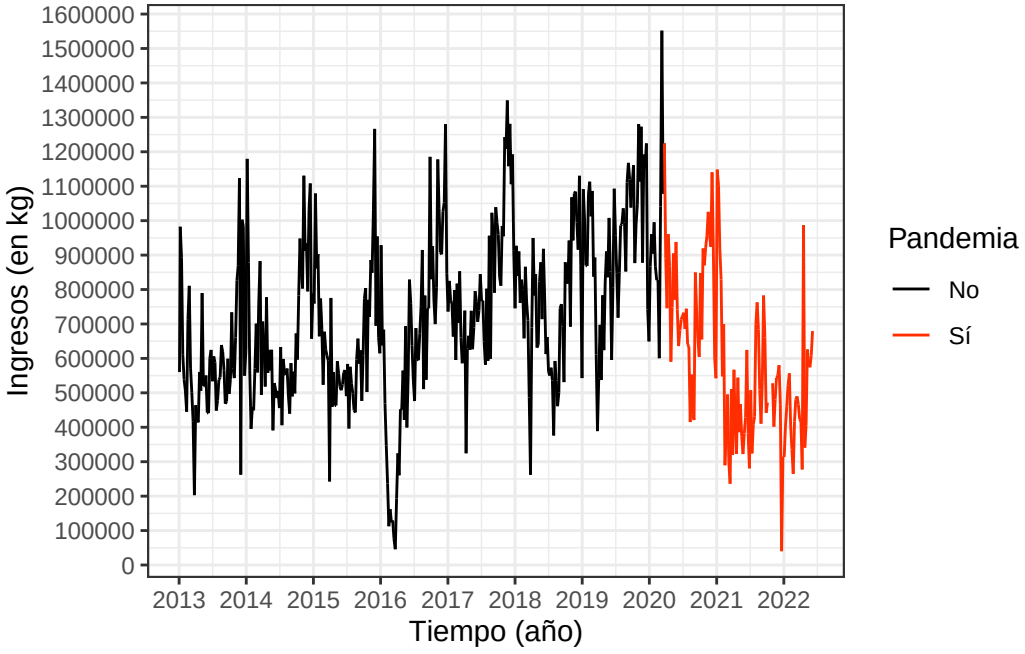


Figura 8.5: Serie de ingresos de mercadería semanales. Rubro: naranja.

8.2 Precios desglosados por calidades

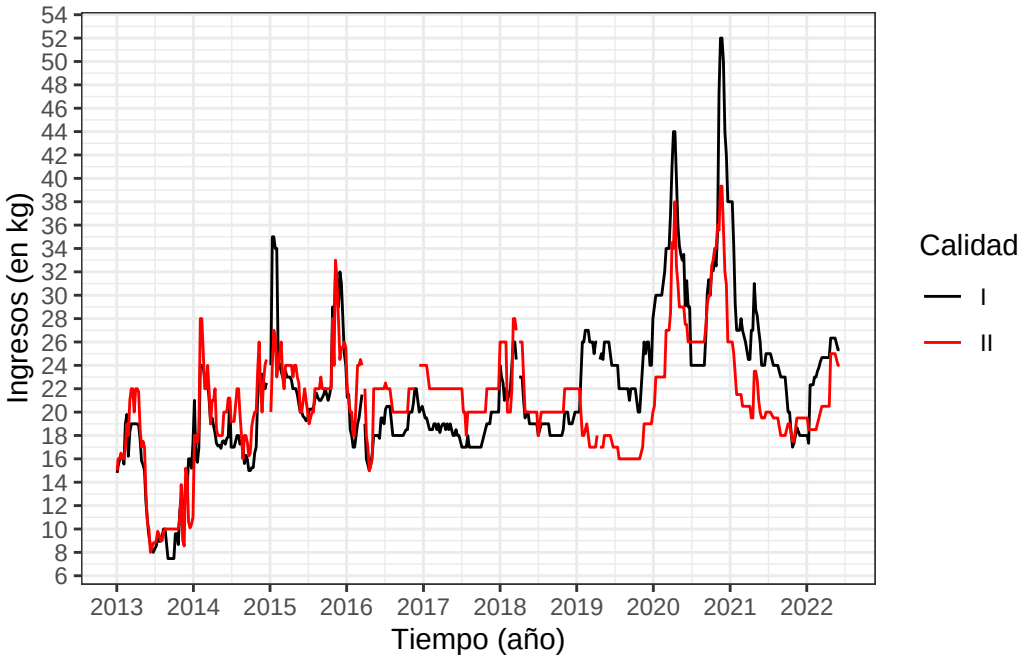


Figura 8.6: Series de precios semanales, distinguiendo calidades. Rubro: papa.

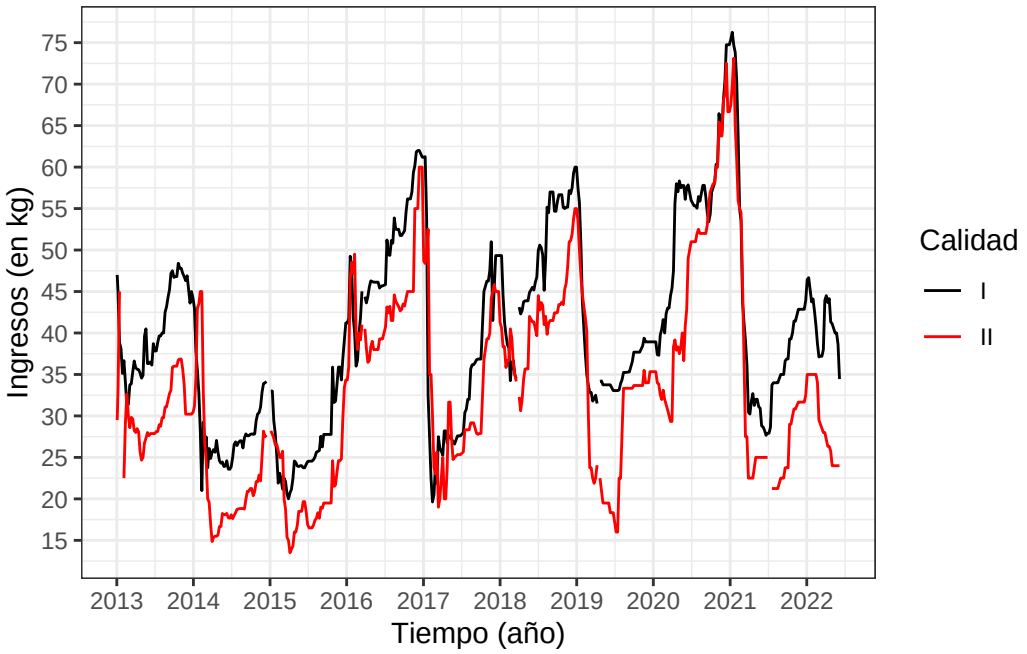


Figura 8.7: Series de precios semanales, distinguiendo calidades. Rubro: manzana.

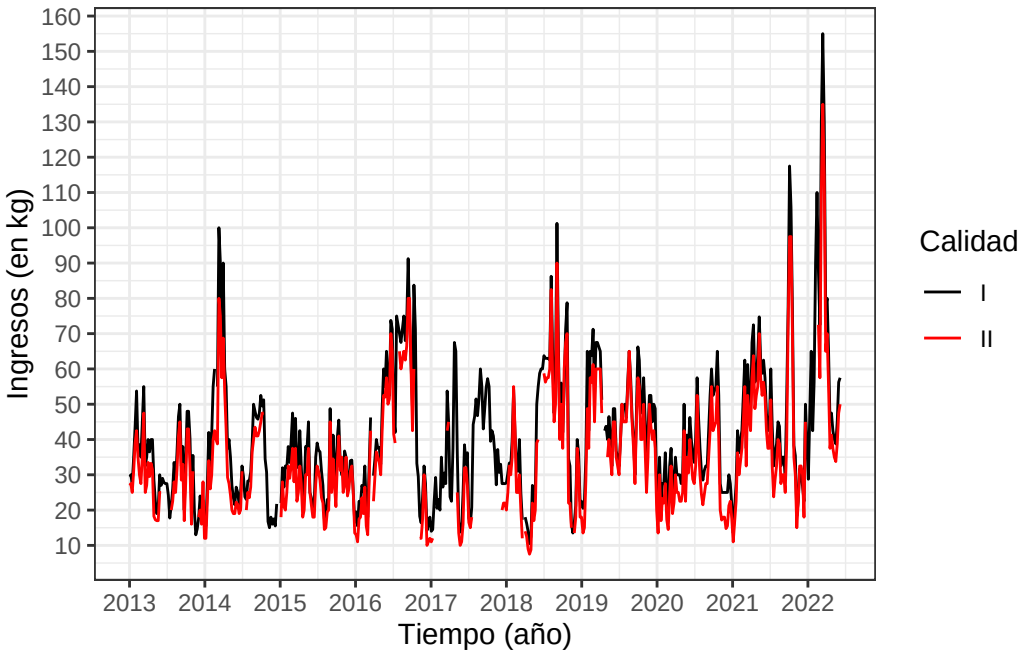


Figura 8.8: Series de precios semanales, distinguiendo calidades. Rubro: tomate.

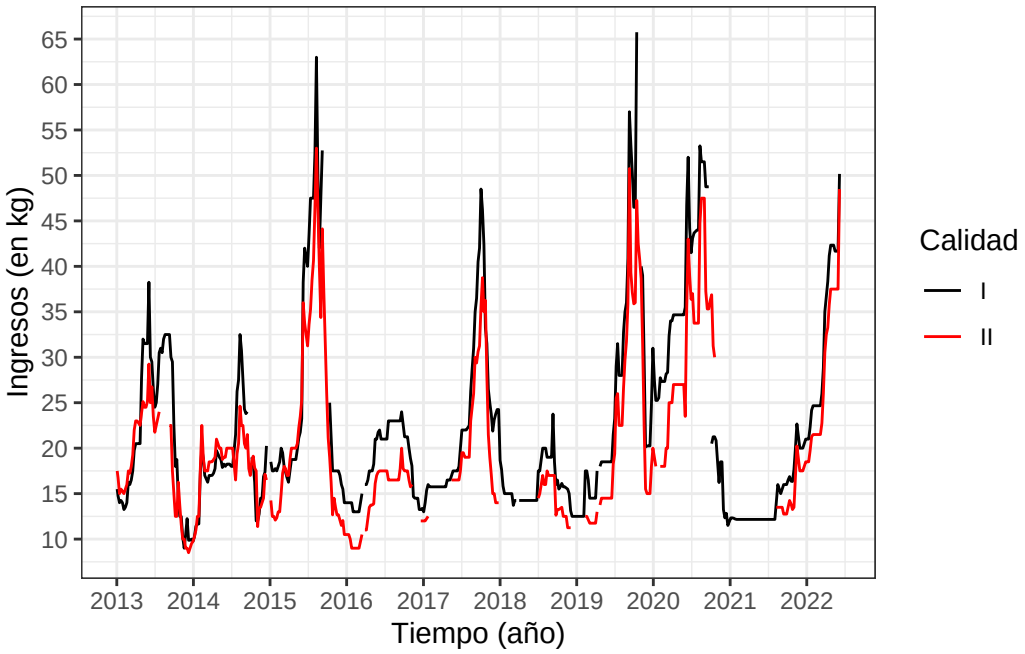


Figura 8.9: Series de precios semanales, distinguiendo calidades. Rubro: cebolla.

8.3 Tablas de predicciones

Tabla 8.1: Predicciones a 52 pasos de precios promedio semanales de manzana. Modelo TBATS.

Fecha inicio sem.	Precio promedio	Predicción	IC. inf. 90%	IC. sup. 90%
2021-06-14	26.62	26.62	24.60	28.80
2021-06-21	26.33	26.44	23.64	29.57
2021-06-28	27.18	26.33	22.95	30.20
2021-07-05	27.91	26.31	22.44	30.83
2021-07-12	28.64	26.40	22.10	31.53
2021-07-19	27.50	26.61	21.90	32.32
2021-07-26	27.62	26.93	21.82	33.22
2021-08-02	27.62	27.34	21.83	34.22
2021-08-09	27.62	27.82	21.91	35.30
2021-08-16	27.62	28.33	22.03	36.41
2021-08-23	28.19	28.83	22.14	37.50
2021-08-30	28.75	29.29	22.24	38.55
2021-09-06	28.75	29.69	22.29	39.52
2021-09-13	28.75	30.02	22.29	40.38
2021-09-20	30.19	30.29	22.25	41.16

Fecha inicio sem.	Precio promedio	Predicción	IC. inf. 90%	IC. sup. 90%
2021-09-27	30.31	30.51	22.19	41.88
2021-10-04	30.31	30.72	22.13	42.58
2021-10-11	34.14	30.97	22.10	43.33
2021-10-18	34.14	31.29	22.12	44.19
2021-10-25	35.00	31.73	22.23	45.20
2021-11-01	36.13	32.30	22.43	46.41
2021-11-08	36.13	33.00	22.72	47.82
2021-11-15	36.70	33.82	23.10	49.42
2021-11-22	37.26	34.73	23.52	51.14
2021-11-29	37.26	35.64	23.95	52.90
2021-12-06	37.26	36.48	24.33	54.56
2021-12-13	37.26	37.15	24.58	55.98
2021-12-20	37.26	37.54	24.66	57.00
2021-12-27	38.21	37.59	24.51	57.49
2022-01-03	40.71	37.24	24.10	57.38
2022-01-10	40.83	36.48	23.44	56.63
2022-01-17	40.10	35.38	22.56	55.31
2022-01-24	39.38	34.00	21.52	53.53
2022-01-31	39.55	32.45	20.40	51.46
2022-02-07	38.78	30.86	19.26	49.28
2022-02-14	37.92	29.32	18.17	47.15
2022-02-21	36.34	27.93	17.18	45.23
2022-02-28	33.31	26.74	16.34	43.60
2022-03-07	33.11	25.80	15.66	42.35
2022-03-14	32.84	25.13	15.15	41.51
2022-03-21	33.01	24.71	14.80	41.07
2022-03-28	35.96	24.52	14.60	41.02
2022-04-04	35.83	24.54	14.52	41.30
2022-04-11	35.00	24.72	14.54	41.85
2022-04-18	35.24	25.00	14.62	42.58
2022-04-25	33.54	25.34	14.73	43.40
2022-05-02	32.56	25.67	14.84	44.23
2022-05-09	32.28	25.96	14.92	44.98
2022-05-16	32.00	26.17	14.95	45.60
2022-05-23	32.00	26.28	14.93	46.05
2022-05-30	31.31	26.30	14.86	46.34
2022-06-06	29.22	26.26	14.75	46.51

Tabla 8.2: Predicciones a 52 pasos de precios promedio semanales de tomate. Modelo ARIMA+Fourier.

Fecha inicio sem.	Precio promedio	Predicción ARIMA+F.	Predicción promedio	IC. inf. 90%	IC. sup. 90%
2021-06-14	45.00	56.23	36.6	40.11	72.35
2021-06-21	40.00	58.74	36.6	38.25	79.23
2021-06-28	40.62	60.55	36.6	38.63	82.47
2021-07-05	55.62	60.85	36.6	38.15	83.56
2021-07-12	38.12	61.82	36.6	37.98	85.66
2021-07-19	28.12	62.88	36.6	37.96	87.81
2021-07-26	32.50	64.05	36.6	38.08	90.01
2021-08-02	35.62	65.31	36.6	38.35	92.28
2021-08-09	42.50	66.65	36.6	38.72	94.57
2021-08-16	41.25	68.00	36.6	39.15	96.86
2021-08-23	30.00	69.30	36.6	39.54	99.06
2021-08-30	32.50	70.43	36.6	39.79	101.06
2021-09-06	32.50	71.27	36.6	39.79	102.75
2021-09-13	27.50	71.71	36.6	39.40	104.02
2021-09-20	45.00	71.64	36.6	38.52	104.76
2021-09-27	75.00	70.96	36.6	37.05	104.87
2021-10-04	107.50	69.63	36.6	34.95	104.31
2021-10-11	101.67	67.66	36.6	32.23	103.09
2021-10-18	73.33	65.10	36.6	28.93	101.27
2021-10-25	36.67	62.06	36.6	25.16	98.95
2021-11-01	32.71	58.69	36.6	21.09	96.29
2021-11-08	16.67	55.20	36.6	16.90	93.50
2021-11-15	22.50	51.80	36.6	12.82	90.79
2021-11-22	32.08	48.72	36.6	9.07	88.38
2021-11-29	31.25	46.15	36.6	5.83	86.47
2021-12-06	28.33	44.26	36.6	3.29	85.23
2021-12-13	20.25	43.16	36.6	1.56	84.77
2021-12-20	47.50	42.91	36.6	0.67	85.15
2021-12-27	38.75	43.48	36.6	0.62	86.34
2022-01-03	28.75	44.79	36.6	1.32	88.26
2022-01-10	47.50	46.70	36.6	2.63	90.77
2022-01-17	65.00	49.03	36.6	4.36	93.70
2022-01-24	42.50	51.58	36.6	6.32	96.84
2022-01-31	56.25	54.13	36.6	8.29	99.97
2022-02-07	90.00	56.49	36.6	10.08	102.90
2022-02-14	110.00	58.49	36.6	11.51	105.47
2022-02-21	80.00	60.02	36.6	12.49	107.56
2022-02-28	65.00	61.03	36.6	12.94	109.12

Fecha inicio sem.	Precio promedio	Predicción ARIMA+F.	Predicción promedio	IC. inf. 90%	IC. sup. 90%
2022-03-07	120.00	61.49	36.6	12.86	110.13
2022-03-14	145.00	61.46	36.6	12.29	110.64
2022-03-21	122.50	61.02	36.6	11.32	110.73
2022-03-28	71.25	60.29	36.6	10.06	110.53
2022-04-04	75.00	59.40	36.6	8.64	110.16
2022-04-11	60.00	58.47	36.6	7.19	109.75
2022-04-18	41.25	57.62	36.6	5.83	109.41
2022-04-25	44.38	56.93	36.6	4.63	109.23
2022-05-02	40.00	56.47	36.6	3.67	109.27
2022-05-09	37.50	56.26	36.6	2.96	109.56
2022-05-16	36.25	56.29	36.6	2.50	110.08
2022-05-23	41.88	56.55	36.6	2.26	110.83
2022-05-30	51.88	56.99	36.6	2.22	111.75
2022-06-06	53.75	57.56	36.6	2.32	112.81