UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2021, Vol. 14, No. 1, pp. 31–52

Modelamiento del precio de la papa criolla en el departamento de Cundinamarca por medio de series de tiempo y modelos dinámicos

Modeling the price of Criolla potatoes in the department of Cundinamarca through time series and dynamic models

María Eliana Díaz Sosa[a]
mariaediaz@usantotomas.edu.co

Edwin Andrés Cruz Pérez[b]
dir.estadisticaaplicada@usantotomas.edu.co

Wilmer Dario Pineda Ríos[c]
wilmerpineda@usantotomas.edu.co

Resumen

El presente trabajo tiene como objetivo evaluar el comportamiento y pronóstico del precio de la papa criolla en el departamento de Cundinamarca, según los factores climáticos desde enero de 2012 hasta abril de 2018. Para ello, se tomaron en consideración, por un lado, análisis basados en series de tiempo (ARIMA, ARIMAX) y, por el otro, modelos lineales dinÃ¡micos (con y sin covariables). En los modelos trabajados se usaron como variables las condiciones climáticas de la zona en cuestión, a las cuales se les aplicó un método de imputación de datos debido a la ausencia de información. Luego fueron agrupados en tres factores construidos por Análisis Factorial para Series de Tiempo (TSFA). Finalmente, se procedió a comparar los indicadores de los cuatro modelos, llegando a la conclusión de que los modelos ARIMA Y ARIMAX generan las mejores predicciones respecto del precio de la papa criolla en el departamento de Cundinamarca.

***Palabras clave*:** variación de precios, variables climáticas, Modelo Autoregresivo Integrado de Media Móvil, Modelos Lineales Dinámicos, Análisis Factorial para Series de Tiempo, indicadores de modelos.

Abstract

The objective of the following research is to evaluate the behavior and forecast of the price of the papa criolla in the department of Cundinamarca, according to the climate factors from January 2012 to April 2018. Therefore, analysis based on time

[a]Estudiante de Maestría en Estadística Aplicada
[b]Docente de Maestría en Estadística Aplicada
[c]Docente de Maestría en Estadística Aplicada

series (ARIMA, ARIMAX) and, on the other hand, dynamic linear models (with and without variables) were taken into consideration. In the models worked, the climate conditions of the area in question were used as variables, to which a data imputation method was applied due to the absence of information. After that, they were grouped into three factors constructed by Time Series Factor Analysis (TSFA). Finally, the indicators of the four models were compared, concluding that the ARIMA and ARIMAX models generate the best predictions regarding the price of papa criolla in the department of Cundinamarca.

***Keywords*:** price variation, climatic variables, Autoregressive Integrated Moving Average, Dynamics Linear Model, Time Series Factor Analysis, model indicators.

1. Introducción

En Colombia, según la Superintendencia de Industria y Comercio (2011), el cultivo de papa representa, en promedio, un 32 % de la producción de los cultivos transitorios. Se trata, pues, de una actividad agropecuaria que figura como una de las bases de la economía de departamentos como Nariño, Boyacá y Cundinamarca, aunque ha ido creciendo, sobre todo, en este último.

Una muestra del crecimiento de la producción en el departamento de Cundinamarca se puede observar en los informes publicados en la revista número 43 del Fondo Nacional para el Fomento de la Papa (FEDEPAPA), donde se señala no solamente que el departamento de Cundinamarca produce actualmente el 70 % de la papa criolla en el país, sino también su posicionamiento como el epicentro a nivel nacional que registra más toneladas exportadas a países de la Comunidad Europea, Estados Unidos y Japón.

Barrientos et al. (2014) señalan que la variación del precio de la papa se ve afectada por factores tales como el clima, las plagas, las enfermedades, el manejo de cultivos, la planeación de la siembra, los precios de insumo, la estructura de consumo, entre otras. El factor climático, sin embargo, remarcan los autores, es aquel que mÃ¡s influye en la producción de la misma, debido a que si existe un cambio, ya sea, en la temperatura media, la precipitación o el incremento en la intensidad del brillo solar, se ve afectada la regularidad de la época de siembra. Desencadenado, como consecuencia, un aumento tanto en el auge de enfermedades como tambiÃ©n un incremento en el precio de la comercialización del producto. En ese sentido, si se estudia concretamente la forma en que el factor climático influye en el precio de la papa criolla se puede llegar a pronosticar la variación del precio y, de esta forma, aportar información en la toma de decisiones de los agricultores, mejorar los márgenes de utilidad y disminuir los riesgos asociados al proceso de negociación.

Nuestro interÃ©s, siguiendo esta línea de análisis, consiste en estudiar la variabilidad del precio de comercialización del kilo de papa criolla en el departamento de Cundinamarca, teniendo en cuenta, principalmente, la influencia de los factores climáticos durante su producción. Para la recolección de los datos del precio de

la papa criolla se consultó el Sistema de Información de Precios (SIPSA), entidad que desde el 2012 en Colombia es la encargada de informar sobre los precios en plazas mayoristas de los productos agroalimentarios que se comercializan en el país. Las variables climáticas que afectan la producción de la papa propuestas por Rojas (2011) son: brillo solar, precipitación máxima, valores medios mensuales de temperatura, valores mínimos mensuales de temperatura, valores máximos de temperatura y valores medios mensuales de humedad relativa, las cuales son elegidas como covariables en el presente estudio y que son recolectadas por Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM).

Para Navarro (2016) el establecer un precio por el cual los agricultores estarían dispuestos a vender sus cosechas, esta condicionado en gran parte por la fase del ciclo productivo. En la actualidad ningún productor es capaz de fijar un precio exacto para la venta, tiene que tomar decisiones basadas en predicciones, las cuales están condicionadas por el conocimiento en ese instante del mercado y de los factores que influyen los precios. Según Wheelwright et al. (1998) la mejor técnica para predecir valores futuros son las series de tiempo, las cuales se basan en inferencias estadísticas echas a los valores del pasado con el fin de describir valores presentes y pronosticar futuras cantidades. Para Tsay (2005) el análisis de series de tiempo proporciona un marco natural para estudiar la estructura dinámica de series de precios, que en su desarrollo incluye procesos de estacionariedad, función de autocorrelación, modelado y pronóstico.

Como lo expresa Coutin (2007) las series temporales presentan una característica intrínseca y es la dependencia existente entre observaciones sucesivas, es decir, la autocorrelación serial. La naturaleza de esta dependencia tiene gran interés práctico y estas correlaciones tienen la ventaja adicional que permiten detectar la presencia de estacionalidad. La modelación ARIMA utiliza la estructura de autocorrelación serial para decidir qué términos incluir en el modelo. La metodología de los modelos ARIMA (*Autoregresive Integrated Moving Average* - por sus siglas en inglés) fue formalizada por Box y Jenkins en 1976 y como describe Chatfield & Xing (2019) consiste en encontrar un modelo matemático que represente el comportamiento de una serie temporal de datos, y permita hacer previsiones únicamente introduciendo el periodo de tiempo correspondiente, por lo tanto, se plantea un modelo ARIMA para el análisis de la serie precio de la papa y además como uno de los objetivos es determinar si las variables climáticas influyen en el precio de la papa se usa también un modelo ARIMAX, el cual es descrito por Orozco et al. (2018) como una extensión del modelo ARIMA al cual se le adapta una o más variables cuando los términos del modelo no entregan un valor explicativo, es por esto, que la inclusión de covariables puede ayudar a mejorar el proceso de predicción.

Por otro lado, Higuita et al. (2018) afirma que la estadística bayesiana ha tenido una alta popularidad en los últimos años para diferentes tipos de aplicaciones, entre ellas el cálculo de pronósticos en series de tiempo. Para Harrison & Stevens (1976) las técnicas de inferencia bayesiana tienen en común una distribución de probabilidad a priori, de manera que se pueda hacer un reajuste de medidas iterati-

vamente, combinándose con diferentes distribuciones de probabilidad de los datos. Una técnica bayesiana son los modelos lineales dinámicos los cuales realizan una actualización recursiva del pronóstico para cada tiempo t, y como indica Campagnoli et al. (2009) estos modelos consideran una serie de tiempo como la salida de un sistema dinámico perturbado por elementos aleatorios, permitiendo una interpretación natural de una serie de tiempo como resultado de varios componentes, como: tendencia, estacionalidad o componentes regresivos. En el presente artículo se presenta una metodología (con fundamento en la estadística bayesiana) para modelar un componente aleatorio como lo es el precio de la papa usando un Modelo Lineal DinÃ¡mico (MLD) empleando el Filtro de Kalman para determinar valores futuros, que según como lo describe West & Harrison (2006) el filtro del Kalman es aplicado a modelos dinámicos bayesianos para realizar pronósticos fundamentados en un procedimiento recursivo de actualización de error y valor puntual. Además, de forma análoga como se hace con los modelos de series de tiempo en MLD se plantean modelos usando como covariables las condiciones climáticas con el fin de realizar pronósticos del precio de la papa, para la selección del mejor modelo se usa Factor de Bayes (*Bayes Factor* - BF). Tal como lo expresa Gelman et al. (2013) el BF es una forma de comparar modelos mediante un análisis bayesiano en el que a cada modelo se le da una probabilidad previa que, cuando se multiplica por la probabilidad marginal (la probabilidad de los datos dados el modelo) produce una cantidad que es proporcional a la probabilidad posterior del modelo.

En la revisión de la literatura se encontraron trabajos destinados a estudiar el comportamiento del precio de la papa en Colombia utilizando modelos de predicción basados en series de tiempo y Red Neuronal Artificial (Rengifo (2016); Barrientos et al. (2014)). A los que se suman, ya entrando en el plano internacional, otros modelos de ecuaciones simultáneas como los esgrimidos por Chávez et al. (2004); Sabbagh-Sánchez et al. (2011); Yujra (2018); y Thiele et al. (1998), que han usado métodos de regresión para describir la variación del precio del producto; o, por otra parte, los esbozados por Rivas (2013) y García (2008), que han realizado análisis descriptivos del precio. Además, se halló que se han planteado modelos para comparar modelos frecuentistas y bayesianos (Mora Adan et al. (2020); Parra Arboleda (2015); Rojas (2018)) en diferentes campos de acción.

Después de esta introducción el artículo se presenta en cuatro capítulos: el primero corresponde a la descripción de las variables climáticas que influyen en la producción de la papa y la respectiva captura, el segundo realiza un acercamiento a los conceptos básicos de modelamiento con series de tiempo estacionarias, ARIMA y ARIMAX. El tercero presenta una breve explicación del análisis factorial para series de tiempo, luego se presenta la fundamentación de la teoría bayesiana. Posteriormente se presenta al análisis y resultados de los modelos usados para la predicción del precio de la papa criolla en Cundinamarca.

2. Generalidades de la papa criolla

La papa criolla es conocida, científicamente, como Solanum phureja. Rodríguez & Ramírez (2011) afirman que su cultivo se lleva a cabo en altitudes que varían entre los 2600 y 3500 metros sobre el nivel del mar (m.s.n.m), lo que equivale a un rango de temperatura promedio de 10 a 20 Â°C. Adicionalmente, agregan los autores, necesita un nivel de precipitaciones promedio de 900 milímetros por año, además de un tipo suelo que presente una textura franca, suelta y profunda, para así evitar la acumulación de humedad en la raíz de la papa. Una vez se ha realizado la recolección del producto, su vida útil como producto fresco es de pocos días, debido a que se brota o germina con prontitud. Su ciclo vegetativo requiere un período que oscila entre cuatro y cinco meses.

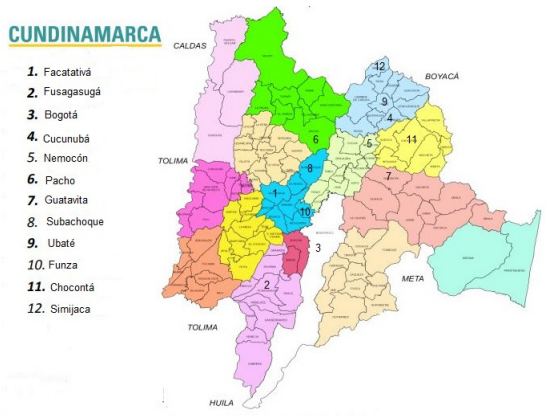

Figura 1: Estaciones hidrometereológicas para el estudio tomada de Planeación (2014)

En lo que particularmente atañe a este trabajo, se estableció el mayor número de estaciones hidrometeorológicas en Cundinamarca en la figura 1 que miden las seis variables climáticas más influyentes en la producción de la papa criolla, las cuales son: precipitación, humedad relativa, temperatura promedio, temperatura promedio máxima, temperatura promedio mínima, brillo solar. Dicha información fue recolectada de la página principal del IDEAM[1], teniendo como referente periódico -mes a mes- desde enero de 2012 hasta abril de 2018. Inicialmente se tuvieron 12 estaciones debido a que son aquellas que se encuentran ubicadas entre los 2500 y 3200 m.s.n.m., altitud ideal para el desarrollo y producción del cultivo de papa.

[1]La entidad que recopila la anterior información en el país es el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) que usa estaciones hidrometeorológicas ubicadas en los diferentes municipios de los departamentos

3. MetodologÃa Box & Jenkins

Para el presente trabajo, la metodología desarrollada por Box & Jenkins se posiciona como la más adecuada, ya que permite predecir valores futuros de una serie de tiempo basándose en valores pasados de una sola variable o entre variables entre las que existe una relación según lo que afirma Arsham (2012).

3.1. Modelo autorregresivo integrado de media móvil

El modelo ARIMA es denominado también método univariante de Box & Jenkins. Marroquín & Chalita (2011) describen la metodologÃa Box & Jenkins para la construcción de modelos de series de tiempo, cuyo procedimiento iterativo se compone de cuatro etapas: identificación, estimación, comprobación y pronóstico. En la primera etapa se utilizan datos antiguos para proponer en forma tentativa un modelo de Box & Jenkins. En la segunda etapa se utilizan datos antiguos para estimar los parámetros del modelo descrito. En la tercera etapa se emplean varios diagnósticos para comprobar si es adecuado el modelo identificado y, si es necesario, recomendar un modelo mejorado. Una vez que se obtuvo el modelo final, se usa para pronosticar valores futuros de series temporales, lo que corresponde a la cuarta etapa.

En líneas generales, los modelos de pronósticos de Box & Jenkins son empleados en el análisis de series temporales estacionarias. Para Tsay (2005) una serie de tiempo (X_t) es estacionaria cuando la distribución conjunta de $X_{t_1}, ..., X_{t_k}$ es idéntica a la distribución de $X_{t_1+t}, ..., X_{t_k+t}$ para todo t, donde k es una variación arbitraria en el tiempo y $t_1, ..., t_k$ es una colección de k valores enteros positivos en el eje del tiempo. En otras palabras, la estacionariedad estricta implica invariancia de la distribución de probabilidad ante valores igualmente separados. Por lo tanto, una serie de tiempo es estacionaria, si su media, su varianza y su autocovarianza permanecen iguales sin importar el momento en el cual se midan; es decir, son invariantes respecto al tiempo. Si la serie no es estacionaria, se deben transformar los datos hasta tener un registro temporal estacionario o una varianza constante. Uno de los métodos empleados para transformar una serie temporal en estacionaria es diferenciarla (se puede diferenciar nuevamente si con la primera diferenciación no se logra), es decir, el proceso consiste en construir una nueva serie en la que cada elemento sea la diferencia de dos elementos consecutivos de la serie inicial con el fin de anular la influencia de factores ajenos a una perturbación de ciclo o tendencia.

A continuación, se explicará el modelos $ARIMA$ considerándo que: X_t es estacionario, con $\phi_1, \phi_2, ..., \phi_p$ y $\theta_1, ..., \theta_q$ como parámetros del modelo y w_t es ruido blanco (RB) con media 0 y varianza σ_w^2, es decir, $w_t \sim RB\left(0, \sigma_w^2\right)$, los parámetros p y q se conocen como los ordenes autorregresivo y de promedio móvil, respectivamente.

El modelo $ARIMA$ permite describir un valor como una función de datos anteriores y errores debidos al azar, además, puede incluir un componente cíclico o

estacional. Un modelo $ARIMA(p, d, q)$ es de la forma:

$$X_t{}^{(d)} = \phi_1 X_{t-1}{}^{(d)} + \cdots + \phi_p X_{t-p}{}^{(d)} + w_t + \theta_1 w_{t-1}{}^{(d)} + \cdots + \theta_q w_{t-q}{}^{(d)} \quad (1)$$

donde $X_t^{(d)}$ es la serie de las diferencias de orden d y $w_t^{(d)}$ es la serie de los errores que se cometen en la serie.

3.2. Modelo ARIMAX

El modelo $ARIMAX(p, d, q)$ es una variante del modelo $ARIMA$ en el cual se incluyen variables externas explicativas para generar la predicción del modelo, con p términos autorregresivos y q términos de promedios móviles, el cual se presenta de la siguiente forma:

$$Y_t^d = \phi_1 Y_{t-1}^d + \cdots + \phi_p Y_{p-1}^d + \theta_1 w_{t-1}^d + \cdots + \theta_q w_{q-1}^d + w_t + \beta_1 X_1 + \cdots + \beta_n X_n \quad (2)$$

donde $Y_t^{(d)}$ es la serie dependiente de las diferencias de orden d, X_n son las n variables explicativas y $w_t^{(d)}$ es la serie de los errores que se cometen en la serie. Además $\phi_1, \phi_2, ..., \phi_p$, $\beta_1, ..., \beta_n$ y $\theta_1, ..., \theta_q$ son parámetros del modelo y $w_t \sim RB\left(0, {\sigma_w}^2\right)$ con $\sigma_w^2 > 0$.

4. Análisis factorial para series de tiempo

Se usa un modelo factorial de series temporales (*Time Series Factor Analysis* por sus siglas en inglés TSFA) con el objetivo de reducir el número de variables, buscando expresar la variabilidad en términos de un número menor de factores. Según Suárez (2016) cuando los datos observados tienen una estructura temporal, es decir, son series temporales, el análisis factorial multivariado no es adecuado, ya que requiere la hipótesis de observaciones incorrelacionadas en cada variable. En ese sentido, los modelos TSFA resumen la mayor parte de información de un número elevado de series temporales en un número menor de factores.

En correspondencia con lo anteriormente expuesto, Gilbert & Meijer (2005) describen un modelo especificado TSFA de la siguiente forma: las variables observadas y_{it} con $i = 1, \cdots, M$, $t = 1, \cdots, T$ en cada periodo t son expresadas en términos de k procesos no observados de interés (denominados factores) donde $k < M$ para una secuencia de T periodos temporales los cuales se denotarán por ξ_{it}, con $i = 1, ..., k$, $t = 1, ..., T$. Los procesos observados (las series temporales, que se denominarán indicadores) será recogido en el vector columna $\boldsymbol{y}_t$ y los factores en un elemento similar $\boldsymbol{\xi}_t$. Con lo anterior el modelo que relaciona los indicadores con los factores está dado por:

$$\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{\xi}_t + \boldsymbol{\varepsilon}_t \quad (3)$$

con $\boldsymbol{B}$ la matriz de tamaño $M \times k$ de pesos del modelo a ser estimados y $\boldsymbol{\varepsilon}_t$ un vector aleatorio de tamaño M en el que se recogen los errores o factores poco

relevantes.[2]

Para la estimación del modelo se usa el paquete estadístico de R planteado por Paul Gilbert y Erik Meijer denominado TSFA, el cual es una extensión del análisis factorial y puede ser usado con datos provenientes de series temporales. Las funciones principales y usadas son:

* tfplot() es una herramienta gráfica para representar objetos desde TSFA.
* tframe() compila las series ha tener en cuenta para el análisis.
* FatfitStats() calcula varias estadísticas del modelo.
* estTSF.ML() ajusta un modelo utilizando un estimador del análisis factorial basado en la matriz de correlaciones. Además, utiliza el estimador de Bartlett para calcular las cargas de los factores.
* DstandardizedLoadings() extrae los pesos estandarizados.
* loadings() capta los valores de los pesos.

5. Modelos Lineales Dinámicos (MLD)

En un modelo dinámico los elementos que intervienen en el modelamiento no permanecen invariables, sino que se consideran como funciones del tiempo, describiendo trayectorias temporales. El análisis de un modelo dinámico tiene por objeto el estudio de la trayectoria temporal específica de alguno de sus elementos. A continuación, se presentan los modelos lineales dinámicos y el filtro de Kalman, siendo el segundo para Kikut-Valverde (2003) un algoritmo recursivo y óptimo de procesamiento de datos para la estimación de modelos con parámetros que cambian en el tiempo.

Bermúdez & D´Achiardi (2011) plantean, por su parte, que los MLD son una amplia clase de modelos con parámetros variables en el tiempo, útiles para el modelamiento de datos de series de tiempo. Se caracterizan por un par de ecuaciones, denominadas ecuación de observación y ecuación de evolución de parámetros.

La ecuación observacional y la ecuación del sistema, siendo el MLD determinado con una distribución a priori Normal n dimensional, son respectivamente:

$$\boldsymbol{Y}_t = \boldsymbol{F}_t^{'}\boldsymbol{\theta}_t + \boldsymbol{v}_t, \boldsymbol{v}_t \sim N(0, \boldsymbol{V}_t) \tag{4}$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \boldsymbol{w}_t \sim N(0, \boldsymbol{W}_t) \tag{5}$$

[2]Para una estimación mÃ¡s detallada de la construcción de los modelos TSFA, véase en: Gilbert, P. D., & Meijer, E. (2005). Ítime series factor analysis with an application to measuring money; University of Groningen, Research School SOM, 2005. Forni, M., Hallin, M., Lippi, M. & Reichlin, L.Ítheˆ generalized dynamic factor model: one-sided estimation and forecasting; Journal of the American Statistical Association, pp. 830-840.

El modelo está definido por la cuádrupla $\{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}$ conocidas, donde para cada t:

* $\boldsymbol{F}_t$ es una matriz conformada por las covariables de orden $n \times r$
* $\boldsymbol{\theta}_t$ es un vector de parámetros desconocidos del modelo de orden $n \times 1$
* $\boldsymbol{G}_t$ es una matriz que describe la evolución de los parámetros contenidos en $\boldsymbol{\theta}_t$ y es de orden $n \times n$
* $\boldsymbol{v}_t$ y $\boldsymbol{w}_t$ representan errores aleatorios los que se asumen tÃpicamente, normalmente distribuidos con media 0, y matrices de varianzas $\boldsymbol{V}_t$ y covarianzas $\boldsymbol{W}_t$ respectivamente
* $\boldsymbol{V}_t$ es una matriz de varianza conocida de orden $r \times r$
* $\boldsymbol{W}_t$ es una matriz de covarianza conocida de orden $n \times n$

Con distribuciones a priori:

$(\theta_0 \mid D_0) \sim N(m_0, C_0)$, siendo a priori D_0 la información en el tiempo $t = 0$ y,

$\theta \sim N\left(\mu_\theta, \frac{1}{\tau_\theta}\right)$ con hiperparámetro $\tau_\theta \sim Gamma(\alpha_0, \beta_0)$.

Los MLD, por otra parte, tal como plantea Mayoral (2013), tienen mayor flexibilidad en el tratamiento de series de tiempo no estacionarias, ya que tienen una interpretación más sencilla que los modelos ARIMA y una mayor libertad en su implementación. Por ende, consideran la serie de tiempo como una salida de un sistema dinámico que es perturbado por errores aleatorios, el cual integra componentes como tendencia, estacionalidad, entre otros. Lo que genera, a vez, que se use una estructura probabilística en la que se calcula recursivamente la distribución condicional del precio de la papa con la información disponible, usando un enfoque Bayesiano.

5.1. Filtro de Kalman

La filtración corresponde al uso de un algoritmo de procesamiento de datos óptimo recursivo, que busca estimar el vector de parámetros de las variables de interés de una manera que minimice el error cuadrático medio (ECM). En el trabajo de Munuera (2018) se define, precisamente, el filtro de Kalman como un algoritmo que estima una variable a partir de datos medidos. Lo hace siguiendo dos pasos: por un lado la predicción del sistema y, por otro, la incorporación de las observaciones recogidas una vez corregidas, obteniendo un estimador óptimo en términos del ECM, en base al comportamiento de los datos y de los errores; siendo esta una referencia para el lector que desee profundizar en el desarrollo teórico del mismo.

5.2. Factor de Bayes BF

El enfoque bayesiano provee una forma de comparación de modelos por medio del factor de Bayes BF (*Bayes Factor* - por sus siglas en inglés), siendo este definido por Acurio & Regalado (2019) de la siguiente forma:

Teniendo unos datos observados y la plausibilidad de dos modelos M_1 y M_2, parametrizados por vectores de parámetros θ_1 y θ_2 se puede medir el BF mediante

$$BF = \frac{p\left(y|M_1\right)}{p\left(y|M_2\right)} \tag{6}$$

donde $p(y|M_i)$ se denomina verosimilitud marginal o verosimilitud integrada.

Jeffrey (1992) presenta los siguientes criterios sobre el BF para saber cuando optar por M_1

Tabla 1: Criterios de decisión sobre el factor de Bayes

BF	**Fuerza de la evidencia a favor de M_1**
<1	apoya M_2
1 a 3	Muy escasa
3 a 10	Sustancial
10 a 30	Fuerte
30 a 100	Muy fuerte
>100	Decisiva

Kass (1993) presenta una gran discusión sobre las bondades y desventajas de este criterio para la comparación de modelos y Sinharay & Stern (2002) discute la sensibilidad de BF en la selección apriori sobre los parámetros de los modelos.

6. Resultados

6.1. Datos

Según lo planteado por Gil (2015) la papa es un producto básico cuyo precio depende principalmente de la oferta y la demanda. Además el sector papicultor colombiano presenta grandes incertidumbres sobre los beneficios netos que se pueden generar a corto, mediano y largo plazo, debido a la alta volatilidad de los precios del tubérculo. En la figura 2 se muestra el comportamiento de la variación del precio mensual desde enero de 2012 a noviembre de 2017, los 5 meses que faltan para completar los datos hasta abril de 2018 van a ser usados con el fin de comparar los pronósticos de los modelos planteados. Por su parte, para el periodo 2012-2017 los precios de papa criolla en Cundinamarca expuesto en la figura 2, no presentan tendencia alguna. Se destaca que en el 2016 se presentó un alza en el precio de la papa criolla, siendo posiblemente causado por el periodo denominado

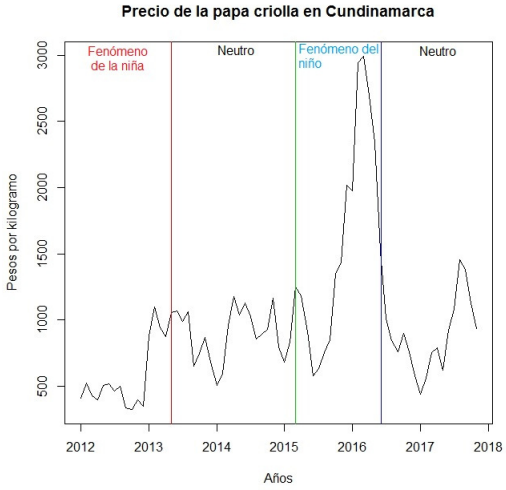

Figura 2: Precio de la papa criolla en el departamento de Cundinamarca de enero de 2012 hasta noviembre de 2017

como fenómeno del niño, que se caracteriza por una disminución de las lluvias en relación con el promedio histórico mensual y aumento de las temperaturas del aire; situaciones, las dos, que contribuyen en la baja producción del producto e incremento en el valor de venta.

En la recolección de la información de las variables climáticas se encuentran datos faltantes debido a que las estaciones que, por lo general, se encuentran alejadas del casco urbano, y si hay alguna falla en los elementos que recolectan los datos, se presenta una demora en la reparación o no es percatada por parte de las personas que realizan el seguimiento, siendo las causa más probable de la no existencia de los registros. En aquellas variables se realizó, no obstante, una imputación por medio de regresión, tal como propone Urrutia et al. (2010), donde se prueba que, para datos de tipo meteorológicos que tienen un 20 % o menos de datos faltantes, la técnica más apropiada de imputación es por medio del uso de un modelo de regresión, conclusión a la que también llegan Medina et al. (2008).

Al realizar el análisis exploratorio se observa que la estación de Chocontá presenta más de un 40 % de datos faltantes, por tal razón se decide que no se va a tener en cuenta para el análisis de la información a estudiar. Por otro lado, las variables que presentaron menos de un 20 % de valores perdidos se les aplica una imputación por regresión lineal siendo el caso de Cucunubá, Bogotá, Subachoque, Funza, Ubaté, Pacho, Simijacá, Guatavita, Fusagasugá, Facatativa y Nemocón.

Continuando con la exploración de las variables climáticas, se realizó un análisis de correlación para determinar información redundante. Los resultados arrojaron la existencia de correlación alta entre variables, esto debido a que las estaciones se encuentran relativamente cerca una de la otra y/o corresponden a mediciones muy

similares en la misma, al notar dicha relación se excluye una de las dos del estudio, por ejemplo entre temperatura media y temperatura máxima de cucunubá el valor es de 0.96, y de esta forma se excluyen ocho mediciones de estaciones.

6.2. TSFA

Se grafican las series climáticas para observar el comportamiento, encontrando que algunas series no muestran estacionalidad, mientras que otras sí. Como sostienen Gilbert & Meijer (2005) estos diferentes patrones en las series pueden reflejar diferencias en los factores, por lo tanto, la estacionalidad ayuda a distinguir los factores de interés y se procede a diferenciar las series.

Con el fin de plantear un modelo TSFA se hace una selección del número de factores a extraer. Para ello, se usa el gráfico de sedimentación donde se opta por tres factores, siguiendo lo citado por Lara (2008), a partir de los cuales se explica que el número de factores a conservar se encuentra en los puntos de inflexión o saltos de importancia entre factores.

Luego de determinar el número de factores a utilizar se calcula la matriz Φ que corresponde a la matriz de covarianza entre factores y donde las puntuaciones factoriales entre los mismos muestran una baja relación siendo valores menores a 0.28.

Según los pesos estandarizados, las series quedan organizadas como se muestra en la tabla 2. Estos valores se encuentran entre -1 y 1 porque una carga mayor (en tamaño absoluto) indica una relación más fuerte con el factor. Las variables se codificaron de la siguiente forma: **T+** para temperatura máxima, **TM** para temperatura mínima, **TP** para temperatura promedio, **P** para precipitación máxima, **H** para humedad relativa y **B** para Brillo.

Tabla 2: Aporte de la serie al factor con mayor carga

SERIE	FACTOR 1
NEMO.H	0,644
CUCU.H	0,538
PACHO.H	0,449
GUATA.H	0,429
UBATE.H	0,395
SUBA.P	0,304
FUSA.T+	-0,257
FUSA.TP	-0,487
BOGO.B	-0,585
FUSA.B	-0,612
FUNZA.B	-0,657
SIMI.B	-0,682
UBATE.B	-0,704
PACHO.B	-0,708
SUBA.B	-0,728
FACA.B	-0,770

SERIE	FACTOR 2
NEMO.P	0,600
SIMI.H	0,541
FUNZA.P	0,539
SUBA.H	0,523
FACA.H	0,507
BOGO.TP	0,498
NEMO.TP	0,483
UBATE.P	0,458
CUCU.P	0,436
FUSA.P	0,416
FUSA.H	0,407
GUATA.P	0,392
BOGO.P	0,336
FUNZA.H	0,325
FACA.P	0,273
SIMI.P	0,217
GUATA.B	-0,472

SERIE	FACTOR 3
SUBA.T+	0,834
NEMO.T+	0,760
FUNZA.T+	0,720
SIMI.TM	0,668
FACA.TM	0,639
GUATA.T+	0,589
UBATE.TM	0,586
CUCU.TM	0,525
PACHO.TM	0,441
SIMI.TP	0,321
GUATA.TP	0,298
FACA.TP	0,211
BOGO.T+	0,201
SIMI.T+	-0,213

Donde el primer factor es denominado Brillo y Humedad de Sabana Centro, factor 2 es Precipitación de sabana occidental y factor 3 es Temperatura.

La comunalidad es la proporción de variabilidad de cada variable que es explicada por los factores. El valor de la comunalidad para los datos del trabajo son independientes de si se usan rotaciones o no, por lo tanto no se le realiza rotación alguna.

6.3. Modelo ARIMA

Al observar el comportamiento del precio de la papa criolla en la figura **??** se determina que no es necesaria una transformación y se calculan las pruebas unitarias de Dickey Fuller y Phillips Perron con el propósito de verificar si es necesario diferenciar la serie.

Tabla 3: Resultados de las pruebas de estacionariedad

	Dickey Fuller	**Phillips-Perron**
Original	0.1909	0.2948
Diferenciada	0.04915	0.01

De los resultados encontrados en las pruebas de estacionariedad presentes en la tabla 3 se evidencia que se debe diferenciar la serie una vez y con la misma se realizan las gráficas ACF Y PACF.

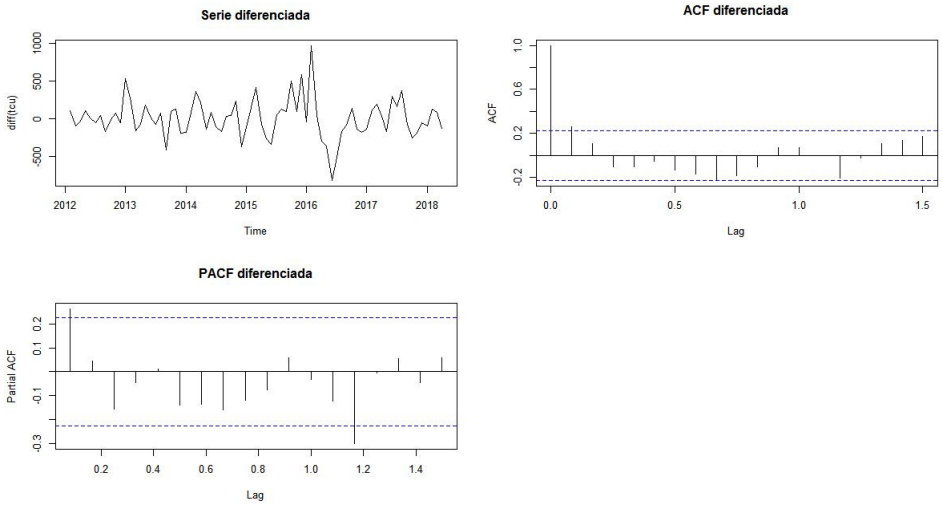

Figura 3: ACF y PACF para la serie diferenciada

A partir del gráfico 3 se realiza un análisis preliminar y estadístico que permite identificar un modelo con el mejor ajuste. Allí, se observa que la función de ACF muestra tres barras que sobresalen asentando la representatividad en esos retardos. Por otra parte, la PACF muestra dos barras sobresalientes y que decaen hacia cero lo cual puede indicar la influencia de un proceso MA. De acuerdo a lo expuesto

en la figura en cuestión se plantean dos modelos: ARIMA(1,1,0) y ARIMA(1,1,1). Para decidir cuÃ¡l de los dos se selecciona se usa el criterio AIC que, según lo planteado por Posada & Noguera (2007), tiene en cuenta los cambios en la bondad de ajuste y las diferencias en el número de parámetros entre los modelos. Los mejores modelos son aquellos que presentan el menor valor de AIC por lo tanto se calculan los valores para los dos modelos planteados y al tomar en consideración el menor se selecciona el modelo ARIMA (1,1,0) con un valor de 977.67 frente a 979.61 del otro. Así, al graficar el modelo ajustado según lo anterior y con la serie original del precio se obtiene la figura 4.

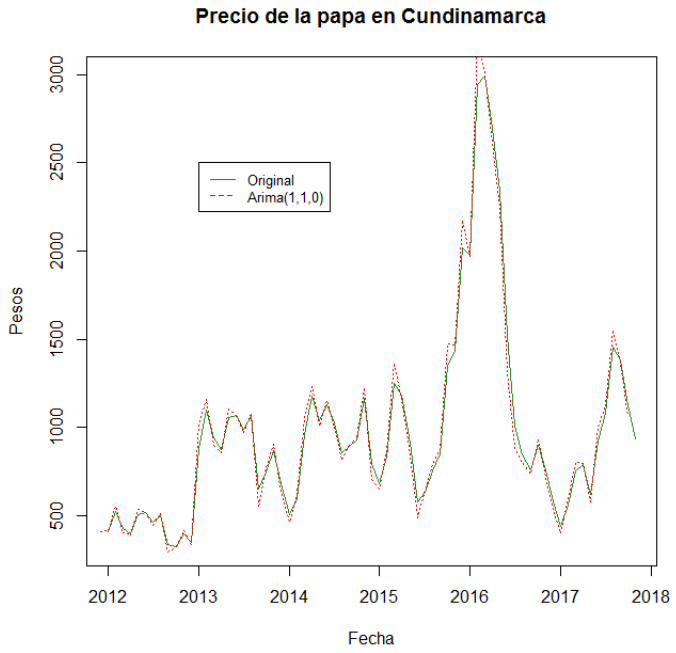

Figura 4: Estimación del precio por ARIMA

6.4. MLD

Se considera un modelo conocido como paseo aleatorio o también llamado modelo polinomial de primer orden, donde Y_t es la serie del precio de la papa criolla y el modelo se puede representar por medio del siguiente sistema de ecuaciones siguiendo lo planteado en las ecuaciones 12 y 13:

$$Y_t = \theta_t + v_t, V_t \sim N(0, V) \tag{7}$$

$$\mu_t = \theta_{t-1} + w_t, W_t \sim N(0, W) \tag{8}$$

Donde v_t y w_t son independientes, siendo un modelo con $F_t = G_t = 1$ y parámetros $V = W = 0.5$.

Se usa el filtro de Kalman con el fin de generar una estimación de la serie de precio, la cual se puede observar en la figura 5.

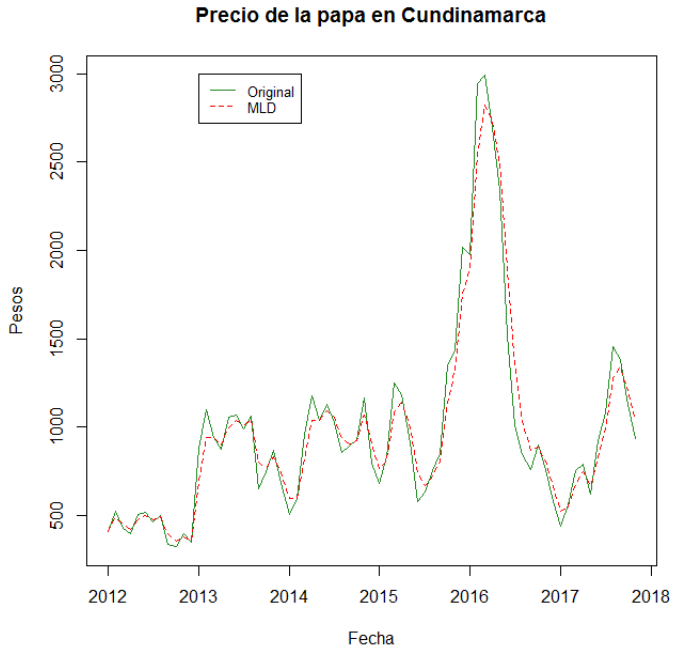

Figura 5: Estimación del precio por MLD

6.5. Modelo ARIMAX

A continuación se ajusta un modelo ARIMAX donde se considera como covariables los 3 factores hallados en la sección 3.2. Se proponen dos modelos ARIMAX, el primero (1,1,1) y (0,1,1), pero por criterio de AIC se selecciona este último, graficado en la figura **??**.

6.6. MLD con covariables

Para realizar el planteamiento del MLD se grafica los ACF del precio contando con la inclusión de los factores hallados en el TSFA, obteniendo así la figura 7.

Considerando la información presente en la gráfica 7 se plantea un modelo inicial

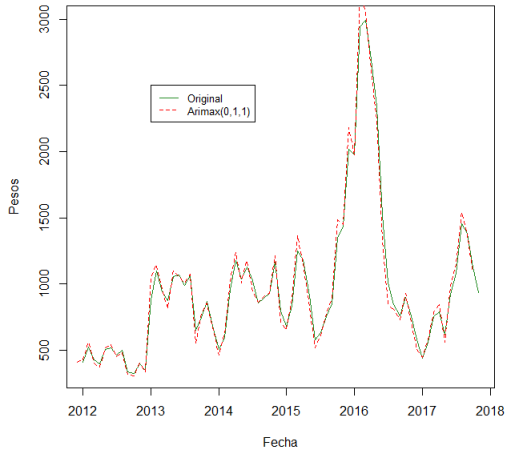

Figura 6: Estimación del precio por ARIMAX

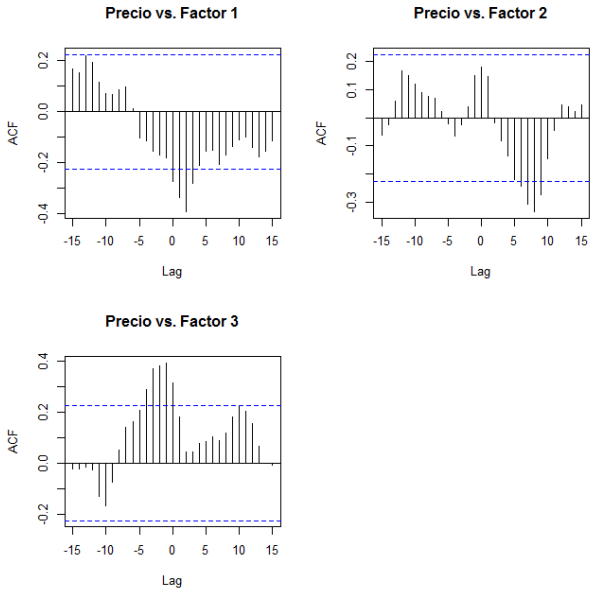

Figura 7: ACF del precio teniendo en cuenta los factores

con Y_t como variable dependiente (precio de la papa), y usando como covariables los factores de la siguiente forma: x_1 que corresponde al factor 1, x_2 factor 2 y x_3

factor 3, siendo descrito en la ecuación 9.

$$Y_t = \beta_0+\beta_1 x_{1,t}+\beta_2 x_{1,t-1}+\beta_3 x_{1,t-2}+\beta_4 x_{1,t-3}+\beta_5 x_{2,t-7}+\beta_6 x_{2,t-8}+\beta_7 x_{2,t-9}+\beta_8 x_{3,t} \tag{9}$$

En la ecuación 9 se presenta el modelo completo, y teniendo como finalidad determinar un modelo que contenga coeficientes significativamente distintos de cero y cuya contribución a la predicción de Y_t sea importante, se calculan las permutaciones generadas al incluir o no un β en el modelo final, generando de esta manera 256 modelos, que a su vez son comparados por medio del BF y DIC. En la tabla 4 se muestra, por otra parte, los 10 modelos con menor valor de DIC y BF con el fin de evidenciar otros criterios de selección que se tuvieron en cuenta.

Tabla 4: 10 Modelos con menor valor de DIC y BF

NÂ°	**MODELO**
1	$\beta_0 + \beta_1 x_{2,t-7} + \beta_2 x_{2,t-9}$
2	$\beta_0 + \beta_1 x_{2,t-8} + \beta_2 x_{2,t-9}$
3	$\beta_0 + \beta_1 x_{2,t-7} + \beta_2 x_{2,t-8}$
4	$\beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{1,t-2} + \beta_3 x_{1,t-3}$
5	$\beta_0 + \beta_1 x_{1,t} + \beta_2 x_{1,t-1} + \beta_3 x_{1,t-2}$
6	$\beta_0 + \beta_1 x_{1,t} + \beta_2 x_{1,t-3} + \beta_3 x_{3,t}$
7	$\beta_0 + \beta_1 x_{1,t} + \beta_2 x_{1,t-2} + \beta_3 x_{1,t-3}$
8	$\beta_0 + \beta_1 x_{2,t-9} + \beta_2 x_{3,t}$
9	$\beta_0 + \beta_1 x_{2,t-7} + \beta_2 x_{3,t}$
10	$\beta_0 + \beta_1 x_{2,t-8} + \beta_2 x_{3,t}$

Para seleccionar uno de los modelos presentes en la tabla 4 se tuvieron en cuenta las siguientes condiciones, siendo éstas definidas por el proceso de producción de la papa criolla:

* El proceso de siembra, cosecha y recolección de la papa criolla no supera los cuatro meses.
* La papa puede ser almacenada en bodega por un lapso máximo de dos meses.
* Las temperaturas extremas no afectan el precio de la papa criolla debido a que en los periodos donde se presentan dichos cambios los agricultores no tienen papa sembrada que se vea afectada.

Teniendo en cuenta las caracterÃsticas descritas anteriormente se descartan los modelos que contienen el factor 3 (x_3), además los que incluyen la temperatura y también aquellos donde se encuentran periodos mayores a 6. Se selecciona, por consiguiente, el modelo 4 que corresponde a $\beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{1,t-2} + \beta_3 x_{1,t-3}$ el cual se ajusta y se presenta en la figura 8 donde a simple vista se traslapan el precio y el ajuste por MLD.

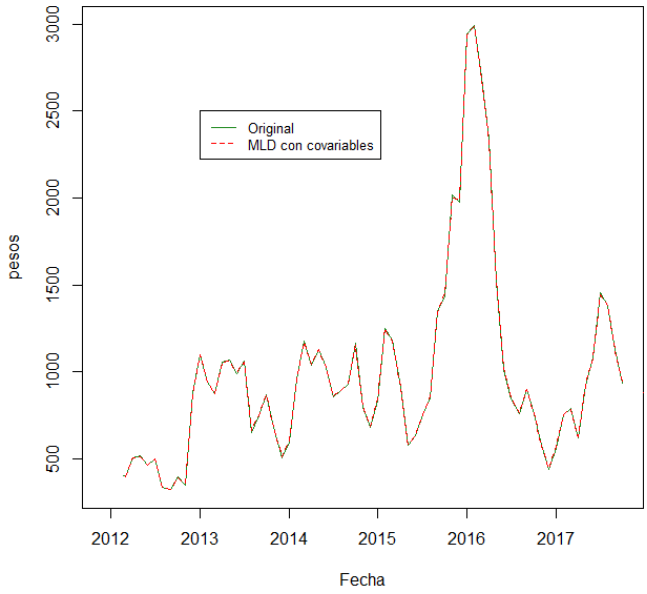

Figura 8: Estimación del precio por MLD con covariables

6.7. Comparación de modelos

Como parte de la comparación de los modelos se calcula en cada uno de ellos el valor MAPE, MAD y RMSE, indicadores usados para medir la dispersión de los pronósticos frente a los valores reales; teniendo en cuenta que entre más pequeño sea el valor, la diferencia entre lo predicho y lo real será tambien menor.

En la tabla 5 se observa que el modelo ARIMA fue el más asertivo para pronosticar el precio. Allí se logró calcular los pronósticos con un valor MAPE de 6 % y siendo en el MAD y RMSE también los de menor valor en comparación con los otros tres modelos. El modelo ARIMAX, por su parte, mostró ser una alternativa al permitir, a través de la técnica usada, incorporar las covariables climáticas. Se trata de un modelo, que seguido al ARIMA, tiene un menor valor en cada uno de los criterios tenidos en cuenta para la confrontación de los resultados.

7. Conclusiones

A lo largo del presente trabajo, se analizaron cuatro modelos con el objetivo de generar predicciones del precio de la papa criolla en el departamento de Cundinamarca. Los modelos usados fueron ARIMA, ARIMAX y MLD con o sin covariables. El

Tabla 5: Comparación de modelos

MES	VERDADERO	ARIMA	ARIMAX	MLD	MLD CON COVARIABLES
Diciembre	880	883,03	838,78	1156,67	979,5
Enero	790,4	869,57	826,88	1115,08	977,1
Febrero	917	866,01	832,34	1081,79	989,9
Marzo	1000	865,07	825,18	1055,14	963,6
Abril	873	864,83	804,16	1033,82	995,3
	MAPE	0,06	0,09	0,23	0,12
	MAD	55,25	81,21	196,42	103,56
	RMSE	73,69	95,4	218,18	115,2

primero de éstos, generó un adecuado grado de predicción, específicamente dentro del periodo establecido entre diciembre de 2017 y los primeros cuatro meses de 2018 (periodo que se estableciÃ^3*comopruebadedichosmodelos*).*Mientrasquelosmodelosqueseplantearonconcovariablesparasuesti*

Los datos obtenidos no desestiman, sin embargo, a continuar en trabajos futuros con la búsqueda de variables adicionales que expliquen el precio de la papa criolla como, por ejemplo, el área sembrada, que se consultó pero que no fue posible encontrarla discriminada mensualmente, sino anual y aproximada. De igual manera, resultaría oportuno empezar a realizar un procedimiento similar para los departamentos de Boyacá o Nariño y, así, comparar los resultados de predicción entre los modelos trabajados.

Aunque la metodología de Box-Jenkins y los modelos lineales dinámicos tuvieron un desempeño de pronóstico similar en nuestros datos, este último es más flexible para tratar con diferentes escenarios, como conjuntos de datos de áreas pequeñas o cuando se establecen nuevos sistemas de vigilancia. Las comparaciones cuantitativas informadas entre ARIMA y MLD dependen en gran medida de las series de tiempo y las medidas de error elegidas. Se necesita una evaluación adicional de ambos métodos, incluido el uso de conjuntos de datos simulados y la comparación con otras metodologías de series de tiempo (por ejemplo, redes neuronales o series de tiempo bayesianas). Además, si uno tiene series de tiempo que, razonablemente, se pueden suponer que son observaciones de valor continuo, una opción natural sería aplicar el enfoque de Box-Jenkins. Esto es así porque el uso de estos modelos en muchos campos induce al uso de un software que simplifique y acelere el proceso de modelado en este caso el software R.

Para la construcción de los modelos se emplearon las siguientes librerías del software R: *mice, VIM, rSHAPE, dplyr, rlang, tseries, forecast, car, dlm, corrplot, tsfa, TSA, dynlm, knitr, rjags, R2jags, coda, gtools, bayestestR, BayesFactor, logspline*, las cuales permitieron el planteamiento de los cuatro modelos descritos anteriormente y realizados en un ordenador de procesador "Intel (R) Core (TM) i5-1035G1 de 1.19 Ghz", el cual tiene memoria de 4Gb, y donde los tiempos empleados para las secuencias no sobrepasaban el minuto.

Recibido:

Aceptado:

Referencias

Acurio, G. B. & Regalado, Z. R. (2019), 'Crisis financieras y contagio en mercados latinoamericanos: una aplicación empírica usando un modelo de cambio de régimen con distribución normal sesgada'.

Arsham, H. (2012), 'Toma de decisiones con periodos de tiempo crítico en economía y finanzas'.

Barrientos, J. C., Rondón, C. & Melo, S. E. (2014), 'Comportamiento de precios de las variedades de papa parda pastusa y diacol capiro en colombia (1995-2011)', *Revista Colombiana de Ciencias Hortícolas* **8**(2), 272–286.

Bermúdez, D. & D´Achiardi, E. (2011), 'Estudio del caudal a través de modelos lineales generalizados dinámicos', *Matemática* **9**(1), 7–15.

Campagnoli, P., Petris, G. & Petrone, S. (2009), *Dynamic Linear Models with R*, Springer.

Chatfield, C. & Xing, H. (2019), *The analysis of time series: an introduction with R*, CRC press.

Chávez, M. C., Mata, R. G., Díaz, S. L., Flores, J. S. M. & Salazar, J. A. G. (2004), 'Efecto del precio internacional sobre el mercado de la papa en méxico, 1990-2000', *Revista Fitotecnia Mexicana* **27**(4), 377–384.

Coutin, M. G. (2007), 'Utilización de modelos arima para la vigilancia de enfermedades transmisibles', *Revista Cubana de Salud Pública* **33**(2).

García, B. M. (2008), 'Análisis del comportamiento de precios de cinco productos hortícolas en costa rica de 1999 al 2003', *Revista Tecnología en Marcha* **21**(2), ág–45.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, CRC press.

Gil, V. V. D. (2015), 'Modelo de simulación de estrategias de inversión para papicultores colombianos', *Lámpsakos* (13), 81–87.

Gilbert, P. D. & Meijer, E. (2005), 'Time series factor analysis with an application to measuring money', *University of Groningen, Research School SOM Research Report 05F10* .

Harrison, P. J. & Stevens, C. F. (1976), 'Bayesian forecasting', *Journal of the Royal Statistical Society: Series B (Methodological)* **38**(3), 205–228.

Higuita, A. D., Valencia, C. M. & Correa, M. J. C. (2018), 'Combination forecasting method using bayesian models and a metaheuristic, case study', *Dyna* **85**(207), 337–345.

Jeffrey, R. (1992), *Probability and the Art of Judgment*, Cambridge University Press.

Kass, R. E. (1993), 'Bayes factors in practice', *Journal of the Royal Statistical Society: Series D (The Statistician)* **42**(5), 551–560.

Kikut-Valverde, A. C. (2003), Uso del filtro de kalman para estimar tendencia de una serie, Technical report, Banco Central de Costa Rica.

Lara, C. P. (2008), 'Técnicas estadísticas multivariantes para la generación de variables latentes', *Revista Escuela de Administración de Negocios* (64), 89–99.

Marroquín, M. G. & Chalita, T. L. E. (2011), 'Aplicación de la metodología box-jenkins para pronóstico de precios en jitomate', *Revista mexicana de ciencias agrícolas* **2**(4), 573–577.

Mayoral, T. A. (2013), Modelos Dinámicos Lineales Aplicados a Series de Tiempo, PhD thesis, Universidad Nacional Autónoma de México.

Medina, R., Montoya, E. & Jaramillo, A. (2008), 'Estimación estadística de valores faltantes en series históricas de lluvia'.

Mora Adan, P. A. et al. (2020), 'Comparación de modelos clásicos en series de tiempo y modelos bayesianos para pronosticar tres acciones colombianas en el último año'.

Munuera, R. M. C. (2018), 'Filtro de kalman y sus aplicaciones'.

Navarro, S. A. J. (2016), 'Predicción de precios de venta de hortalizas'.

Orozco, M. A. M., Aguilar, D. S. G., Ramírez, F. O. P. & Rodríguez, N. J. M. (2018), 'Modelo cuantitativo arimax-egarch para la predicción de la tasa de cambio colombiana (cop/usd)', *Revista Espacios* **39**(7), 16.

Parra Arboleda, L. F. (2015), 'Modelamiento conjunto del número de siniestros y pagos por reclamación en seguros mediante una cópula mixta desde la perspectiva frecuentista y bayesiana', *Departamento de Estadística* .

Posada, S. L. & Noguera, R. R. (2007), 'Comparación de modelos matemáticos: una aplicación en la evaluación de alimentos para animales', *Revista Colombiana de Ciencias Pecuarias* **20**(2), 141–148.

Rengifo, C. L. (2016), 'Comportamiento del precio de la papa industrial diacol capiro, caso colombia a partir del mejor pronóstico entre un sarima y una red neuronal artificial'.

Rivas, S. T. (2013), 'Estructura de mercado y determinantes del precio de la papa para consumo fresco'.

Rodríguez, D. R. & Ramírez, L. N. (2011), 'La agroindustria de la papa criolla en colombia. situación actual y retos para su desarrollo', *Gestión y Sociedad* **4**(2), 17–30.

Rodriguez, Y., Pineda, W. & Olariaga, O. D. (2020), 'Air traffic forecast in post-liberalization context: a dynamic linear models approach', *Aviation* **24**(1), 10–19.

Rojas, A. J. S. (2018), Pronóstico del precio en bolsa de la energía eléctrica en colombia, utilizando inferencia bayesiana, B.S. thesis, Uniandes.

Rojas, B. E. O. (2011), Evaluación del desarrollo del cultivo de papa bajo escenarios de variabilidad climática interanual y cambio climático, en el sur oeste de la Sabana de Bogotá, PhD thesis, Universidad Nacional de Colombia.

Sabbagh-Sánchez, A., García-Salazar, J. A., Matus-Gardea, J. A., Jiménez-Sánchez, L. & Hernández Juárez, M. (2011), 'Comportamiento del consumo de papa (solanum tuberosum l.) fresca en méxico', *Revista mexicana de ciencias agrícolas* **2**(4), 559–572.

Sinharay, S. & Stern, H. S. (2002), 'On the sensitivity of bayes factors to the prior distributions', *The American Statistician* **56**(3), 196–201.

Suárez, G. S. L. (2016), 'Técnicas estadísticas multivariantes de series temporales para la validación de un sistema reconstructor basado en redes neuronales.'.

Thiele, G., Bustamante, J., Mansilla, J. & Scott, G. (1998), *Los precios de papa, arroz y trigo en Bolivia: Un analisis del periodo 1980-96.*, International Potato Center.

Tsay, R. S. (2005), *Analysis of financial time series*, Vol. 543, John wiley & sons.

Urrutia, J. A., Palomino, R. & Salazar, H. D. (2010), 'Metodología para la imputación de datos faltantes en metereología', *Scientia et technica* **3**(46), 44–49.

West, M. & Harrison, J. (2006), *Bayesian forecasting and dynamic models*, Springer Science & Business Media.

Wheelwright, S., Makridakis, S. & Hyndman, R. J. (1998), *Forecasting: methods and applications*, John Wiley & Sons.

Yujra, I. D. P. (2018), 'Evaluación de precios de la papa (solanum tuberosum), en bolivia', *Revista Estudiantil Agro-Vet* **2**(1), 79–93.