ESTIMACIÓN DE MODELOS PARA EL PRONÓSTICO DE PRECIOS EN EL SECTOR AGROPECUARIO EN EL DISTRITO ESPECIAL DE BUENAVENTURA, UNA APROXIMACIÓN ECONOMÉTRICA USANDO EL LENGUAJE DE PROGRAMACIÓN R

AUTORES:

JOSE LUIS GAMARRA PALACIOS Y JOHN MARIO MICOLTA GARZÓN

Trabajo de grado para optar al título:

Magíster en Ciencias de Datos

Director del Trabajo de Grado:

Ph.D. JULIO CÉSAR ALONSO CIFUENTES

TABLA DE CONTENIDO

TABLA DE ILUSTRACIONES

TABLA DE FIGURAS

LISTA DE TABLAS

RESUMEN

El pronóstico de precios de productos es un problema ampliamente documentado, en la academia y la industria. El pronosticar los precios permite una debida planeación en la cantidad de productos que se pueden sacar a la venta. Además, permite estimar las ganancias que serán obtenidas en el futuro si los productos tienen determinado precio y, si las condiciones del mercado son relativamente estables y parecidas a las condiciones del pasado.

En este trabajo abordaremos el problema de pronosticar precios para tres productos del sector agrícola en el Distrito Especial de Buenaventura: Yuca Chirosa, la Piña Gold y el Banano Criollo. Esto es importante para los campesinos pues, posibilitará la creación de planes de contingencia respecto a la siembra y cosecha de estos productos, basados en técnicas de ciencias de datos como complemento a las técnicas heurísticas, que son usadas por los campesinos basados en el conocimiento del negocio.

Las estimaciones de los pronósticos de precios de los productos antes mencionados se realizaron a través de técnicas econométricas para series de tiempo mediadas por el lenguaje de programación R, para ello se usará la metodología CRISP-DM.

Los métodos econométricos usados fueron: suavizamiento exponencial simple, exponencial lineal, Holt-Winters aditivo y multiplicativo, promedio móvil. Además de, modelos autorregresivos (AR), modelos de media móvil (MA), modelos basados en regresión, modelos ARMA y ARIMA. Por último, modelos de ensamble.

La métrica usada para validación de los pronósticos fue la raíz cuadrada del error promedio cuadrático o RMSE por sus siglas en inglés (Root Mean Squart Error). La cual especifica que el modelo con RMSE más pequeño es más idóneo para el pronóstico.

Como resultado encontramos que para la Yuca Chirosa el mejor modelo es el de suavización exponencial simple, para la Piña Gold el mejor modelo es el de mínimos cuadrados ordinarios y para la Yuca Chirosa el mejor método fue el de ensamble de Newbold y Granger.

Como complemento al análisis econométrico realizado sobre las bases de datos de los productos agrícolas mencionados. Los autores desarrollan un tablero de mando usando la plataforma Flex-Dashboard, esta interfaz gráfica permite obtener en tiempo real el precio pronosticado para los tres productos analizados y, un breve resumen sobre su comportamiento semanal, mensual y anual. Además, la desviación respecto al precio real.

1. INTRODUCCIÓN

1.1. Contexto y Antecedentes

Los métodos de estimación de pronósticos son procesos que emplea datos pasados, datos presentes, análisis de patrones y tendencias. Además, para pronosticar se debe suponer que el comportamiento futuro de la variable o las variables de interés son similares a su comportamiento pasado.

Para ejemplificar y entender los alcances, usos y bondades de los métodos de estimación de pronósticos podemos ver Acevedo (2008), donde el autor discute los diferentes modelos econométricos tales como los autorregresivos (AR), modelos de promedio móvil (MA), modelos autorregresivos de media móvil (ARMA) y los modelos integrados autorregresivos de media móvil (ARIMA).

Como ejemplo de aplicaciones podemos ver Alonso, Díaz, Estrada, Figueroa & Tamura, (2019), donde se emplean modelos jerárquicos de series de tiempo para encontrar el mejor modelo para pronosticar los galones de gasolina corriente demandados en Bogotá. A través de este modelo, la Secretaría de Hacienda Distrital de la capital de Colombia espera mejorar el recaudo del impuesto de la sobretasa a la gasolina y, hacer frente a la evasión de este tributo que se usa para el mejoramiento de la malla vial.

En el sector agrícola hay varios estudios de estimación de pronósticos. Por ejemplo, en el trabajo de Ramírez, Hernández & Zulueta, (2010), hacen análisis de series de tiempo para pronosticar la producción de la caña de azúcar. Otro estudio relacionado con el azúcar, esta vez desde el punto de vista financiero, puede consultarse en Alonso & Arcila (2019), donde los autores analizan modelos para las predicciones diarias en contratos de futuros de este commodity.

En lo que se refiere a este trabajo, abordaremos la problemática de pronosticar los precios de los siguientes productos agrícolas cultivados en el Distrito Especial, Industrial, Portuario, Biodiverso y Ecoturístico de Buenaventura: Yuca Chirosa, la Piña Gold y el Banano Criollo. Esta aproximación se hará a través de modelos estadísticos. Compararemos varios modelos y, obtendremos los mejores, vía métrica RMSE para la medición de la eficiencia de predicción respecto al dato real. A la fecha de mayo del 2021 los autores dan fe que no encontraron trabajos relacionados o parecidos al actual y, todo el estudio y análisis siguiente es de autoría intelectual propia y del director del proyecto de grado.

1.2. Justificación

En este documento se abordará el problema de estimación de pronósticos de precios del sector agropecuario en Buenaventura para tres productos típicos de la región como lo son: la Yuca Chirosa, la piña Gold y el Banano Criollo.

Las razones de la elección del sector agropecuario se deben principalmente a los indicadores de la Encuesta de Calidad de Vida (ECV) de la Población Campesina de Colombia, realizado por el Departamento Administrativo Nacional de Estadística (DANE) en julio del 2020.

Estos indicadores reflejan que la pobreza multidimensional es mayor en los hogares que se identifican como campesinos con una incidencia del 29,3%, frente al 17,5% del total nacional. Además, fue la primera vez que se incluyó población que, sin vivir necesariamente en zonas rurales, se auto reconoce como campesina. Otra de las cifras preocupantes que arroja la encuesta, es la precariedad en la cobertura de servicios públicos en los hogares campesinos. Esto se ve reflejado en diversas variables como el acceso a acueductos, donde solo el 69,3% puede acceder a agua potable, frente al 86,8% del total nacional. En cuanto al acceso a gas natural, solo el 34% de los campesinos tiene acceso a este servicio, mientras que 64% a nivel nacional cuentan con él.

Por otra parte, el 70% de los hogares campesinos tienen bajo logro educativo[1] y mientras el 37,9% de los hogares a nivel nacional se consideran pobres, este porcentaje es de 58,7% en hogares campesinos. Además de la ECV, según el estudio denominado "Campesinos, tierra y desarrollo rural: Reflexiones desde la experiencia del tercer de laboratorio de paz" realizado por la unión europea (Fuente: https://eeas.europa.eu), la población campesina de Colombia no tiene capacitación técnica ni tecnológica, están expuestos a la competencia derivada de los 16 tratados de libre comercio que tiene el país y, no tienen preparación en finanzas ni negocios. Por lo cual es imperativo que las empresas del sector privado, empresas estatales, universidades y emprendedores, aporten su grano de arena y de esta manera apoyar a los campesinos.

El estudio realizado en este documento se centra en el Distrito Especial de Buenaventura por varias razones. Buenaventura es el principal puerto de Colombia y además tiene algunos indicadores de desarrollo y pobreza más bajos en comparación con el promedio de los demás municipios del Valle del Cauca. Por ejemplo, el acceso a acueducto en Buenaventura es del 73.2%, mientras que el promedio en el Valle del Cauca es del 95.4%. Así mismo, la cobertura del alcantarillado solo alcanza el 61%, mientras que a nivel departamental el promedio es 90.9%. Por otro lado, la cobertura de servicios domiciliarios tales como el internet donde el promedio del Valle del Cauca es de 58% y el de Buenaventura tan solo de 28%. (Fuente: https://www.dane.gov.co/files/investigaciones/planes-desarrollo-territorial/100320-Info-Alcaldia-Buenaventura.pdf). Por otra parte, según el Censo Nacional de Población y Vivienda (CNPV 2018), el 88.6% de la población de Buenaventura trabaja de manera

[1] Bajo logro educativo: se considera privado el hogar donde la educación promedio de las personas de 15 años y más, es menor a 9 años de educación.

informal, mientras que el promedio en la ciudad de Cali es del 75.3%. Finalmente, el índice de Pobreza Multidimensional en Buenaventura es del 41%, frente al 11.9% de Cali. (Fuente: https://www.dane.gov.co/files/censo2018/informacion-tecnica/cnpv-2018-presentacion-3ra-entrega.pdf).

Del sector agropecuario en el Distrito Especial de Buenaventura podemos destacar que, según el Censo Nacional Agropecuario del 2014, habitaban en 11846 familias campesinas, de las cuales 9348 personas que pertenecían a alguna de estas familias eran unidades productoras; es decir, campesinos que se beneficiaban directamente de la siembra y cosecha. El estudio en cuestión que en este documento se presenta, se enfoca en estas familias campesinas.

1.3 Planteamiento del problema

Según el tercer Censo Nacional Agropecuario (CNA) realizado en Colombia en el año 2014, en Buenaventura habitaban 11846 hogares campesinos. Donde, a través de los datos de cultivos y hogares, se evidenció que estaban conformados por 44236 personas, de las cuales el 51% eran hombres y, 49% eran mujeres. Además, 87% de estos campesinos se reconocían como afrodescendiente, el 9% indígena y el 4% restante se autodenominaba palenquero.

De estas familias campesinas en Buenaventura, el 16% no sabía leer ni escribir en español. Y, el 60% no se encontraba realizando ningún tipo de estudio preescolar, escuela, colegio o universitario. En la base de datos del CNA también se encontró que el 43% de las personas tenía como logro académico más alto básica primaria y otro 20% básica secundaria. Tan solo el 15% tenía media secundaria.

A través de este documento, analizaremos cuáles son los productos más sembrados y cosechados del Distrito Especial de Buenaventura. Elegiremos tres de ellos basándonos en el criterio de disponibilidad de precios en el Sistema de Información de Precios del Sector Agropecuario (SIPSA). Luego, estudiaremos los datos y haremos análisis exploratorio, limpieza e imputación de datos perdidos. Después, realizaremos la construcción de los modelos de estimación de pronósticos de precios, para así proceder a validar los mejores modelos. Luego, se escoge el mejor modelo según la métrica de evaluación RMSE y, se efectúan los pronósticos para los productos analizados, que en este caso fueron: Piña Gold, Banano Criollo y la Yuca Chirosa. Seguidamente, se automatiza la captura de los datos oficiales del Sistema de Información de Precios del Sector Agropecuario (SIPSA) y la producción de pronósticos. Finalmente, se construye un Dashboard que permita visualizar los pronósticos.

1.3. Objetivo general

Generar una herramienta que automáticamente capture nuevos datos y presente de manera gráfica los pronósticos de los precios de tres productos agrícolas cultivados en Buenaventura para las siguientes tres semanas.

1.4. Objetivos específicos

I. Seleccionar los productos agrícolas que más se cosechan en el Distrito Especial de Buenaventura y que además sus precios estén reportados en el SIPSA.

II. Estudiar las características de las series de tiempo de los precios de los productos seleccionados.

III. Construir modelos de pronósticos de precios de los tres productos seleccionados utilizando métodos econométricos y de análisis cuantitativo.

IV. Evaluar los modelos obtenidos por medio de la métrica RMSE para seleccionar el modelo con mejor comportamiento fuera de muestra para pronosticar el precio de las siguientes tres semanas para cada producto.

V. Automatizar la actualización de los datos cada semana y la construcción de los nuevos pronósticos con los últimos datos disponibles.

VI. Crear un tablero de mando para ver un despliegue de ciertos indicadores asociados a los pronósticos realizados en cada producto.

2. MARCO TEÓRICO

Alonso (2020) estudia los objetos series de tiempo y sus componentes, los distintos métodos de suavizamiento (el exponencial simple, lineal y de Holt-Winters (aditivo y multiplicativo), el método de promedio móvil), los modelos autorregresivos (AR), los modelos de media móvil (MA), ARMA, ARIMA, métodos basados en regresión y modelos de ensamble.

A continuación, se realizará una revisión de estos modelos y de esta manera sentar las bases del análisis que haremos sobre los cultivos Yuca Chirosa, Piña Gold y Banano Criollo[2] del Distrito Especial de Buenaventura.

2.1. Muestra de estimación y evaluación

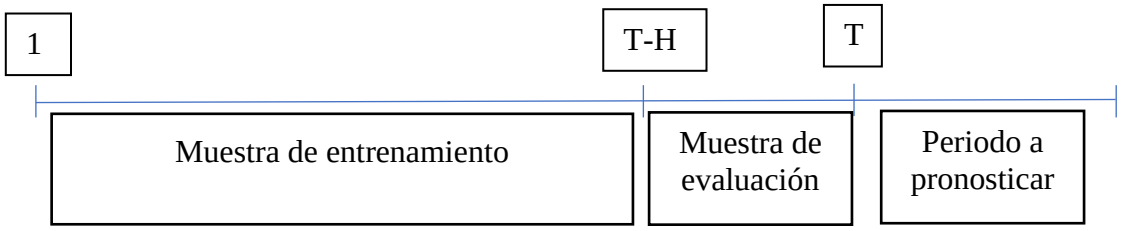
Antes de hablar de los distintos métodos de suavización para estimación de pronósticos, es pertinente hacer una introducción sobre la forma en la que los científicos de datos usan los modelos en aras de realizar extrapolaciones.

Supóngase que nuestros datos de serie de tiempo tienen un tamaño T; es decir, tienen T observaciones para a partir de ellas hacer estimaciones del comportamiento futuro. En ciencias de datos y particularmente en el Machine Learning, son muy comunes los mecanismos de separación de datos en dos bloques, uno llamado muestra de entrenamiento y otro llamado muestra de validación. Quizá uno de los métodos más usados y estudiados es el de validación cruzada. A partir de estos métodos, se encuentran los parámetros del modelo con los datos de entreno y se verifica la eficiencia con la submuestra de validación. Sin embargo, al tratarse de datos temporales y cuya importancia se da a raíz de que quizá haya una fuerte relación entre ellos en el sentido que un dato en un determinado tiempo depende del anterior, los métodos como validación cruzada no serían adecuados para tomar sub muestras de los datos y hacer análisis.

Teniendo en cuenta lo dicho anteriormente, la práctica común es dividir los datos que se tienen en formato serie de tiempo en dos grupos como lo sugiere la Ilustración 1.

[2] Mas adelante se discutirá la manera de en la que se eligieron estos productos.

Ilustración 1. Partición de observaciones en muestra de entrenamiento y muestra de evaluación.



Fuente: Elaboración propia.

Es decir, se trabaja en la estimación del modelo con la muestra de entrenamiento que contiene $T - H$ datos. La muestra restante, que contiene las últimas H observaciones, denominada muestra de evaluación, se usa para evaluar el comportamiento de los modelos en la estimación de pronósticos o extrapolación.

Es importante resaltar que con la muestra de H observaciones se evaluará el comportamiento de los modelos para pronosticar. Una vez escogido el mejor modelo, volveremos a usar toda la muestra disponible (hasta la observación T) para reestimación del modelo y para la generación de pronósticos futuros[3].

Para medir la eficiencia de los modelos en el sentido de determinar cuál es mejor en la estimación de pronósticos usaremos la métrica RMSE, definida por:

$$RMSE = \sqrt{\frac{\sum_{h=1}^{H}(y_{T+h} - \hat{y}_{T+h})^2}{H}}$$

donde y_{T+h} es el valor real de la serie para cualquier periodo $T + h$, $\hat{y}_{T+h}$es el valor pronosticado para el periodo $T + h$. Así las cosas, el error de pronóstico para un periodo especifico será $\boldsymbol{\epsilon} = y_t - \hat{y}_t$. (Alonso, 2020. p. 193).

Existen diferentes maneras para determinar si la estimación de pronósticos de un modelo en particular es eficiente; es decir, para analizar si el pronóstico es bueno en términos comparativos con los datos reales. Teniendo en cuenta que la serie de tiempo se divide en dos muestras como se explicó anteriormente, donde los $T - H$ datos dentro de la muestra sirven para construir una estimacion de los H restantes. Diferentes aproximaciones son usadas para las estimaciones de H pronósticos que sirvan de comparativo con los H que se

[3] Para más detalles ver la sección 3.1 en Alonso (2020).

tienen fuera de muestra. La aproximación utilizada en este trabajo es la ***ventana recursiva***, que consiste en actualizar los coeficientes del modelo empleando una observación más, después de hacer cada pronóstico para un periodo adelante y, dejando el inicio de la muestra siempre igual, de tal manera que la muestra crece.

2.2. Métodos de suavizamiento

Los métodos de suavización son mecanismos para ***suavizar*** el comportamiento de una serie de tiempo, de ahí su nombre. Una de sus bondades es necesitar pocos datos. Estos métodos son: promedio móvil (MA), suavización exponencial simple (SES), suavización exponencial lineal (Holt) y suavización exponencial lineal de Holt-Winters. (Alonso, 2020. p. 41).

2.2.1. Promedio móvil (MA)

La definición del modelo de estimación de pronósticos llamado promedio móvil para el periodo y (y_t) es:

$$\hat{y}_{T+1} = \frac{1}{m}\sum_{j=-k}^{k} y_{t-j}$$

donde, $\hat{y}_{T+1}$ corresponde al valor suavizado para el periodo T+1. La constante m=2k+1 se denomina orden del promedio móvil, donde k indica la cantidad de observaciones usadas. El método del promedio móvil es quizá el más sencillo para proyectar la serie pues ignora los componentes tendenciales, estacionales y cíclicas. (Alonso, 2020. p. 43).

2.2.2. Suavización exponencial simple (SES)

Los pronósticos del método SES[4] son promedios ponderados de observaciones pasadas. Los pesos o ponderaciones a, que se les asignan a las observaciones van decayendo exponencialmente a medida que las observaciones se alejan en el tiempo. En otras palabras, cuanto más reciente sea la observación, mayor será el peso a[5] asignado. Formalmente, se define de la siguiente manera:

Ecuación de pronóstico: $\hat{y}_{t+h|\,t} = l_t$

Ecuación de suavizamiento: $l_T = ay_t + (1-a)l_{t-1}$

[4] Para más detalles, consultar (Alonso, 2020. p. 47).

[5] El dominio de los valores de a son los reales positivos.

2.2.3. Suavización exponencial lineal (Holt)

El método SES tiene un problema al no incluir en el modelado la tendencia. Holt[6], observó este problema del método de suavizado exponencial simple. De esta manera generaliza el modelo simple incluyendo dos parámetros a y β.

Ecuación de pronóstico: $\hat{y}_{t+h|t} = l_t + hb_t$

Ecuación de nivel: $l_t = ay_t + (1-a)(l_{t-1} + b_{t-1})$

Ecuación de tendencia: $b_t = \beta * (l_t - l_{t-1}) + (1-\beta)b_{t-1}$

2.2.4. Suavización exponencial de Holt-Winters

El método SES se aplana rápido por no incluir tendencia. Holt lo extiende y define un modelo que la incluya, pero no así la estacionalidad. Holt y Winter extienden el método de suavización exponencial lineal. El método estacional de Holt-Winters[7] comprende la ecuación de pronóstico y tres ecuaciones de suavizamiento, una para cada componente en las que se caracteriza una serie de tiempo; tendencia (β_t), componente estacional (s_t) y el nivel u observación (l).

Existen dos definiciones del método de Holt-Winters dependiendo de la naturaleza de la estacionalidad: Holt-Winters aditivo y Holt-Winters multiplicativo. El método aditivo se recomienda cuando las variaciones estacionales son cuasi constantes a lo largo de la serie, mientras que el método multiplicativo se emplea cuando las variaciones estacionales cambian de manera proporcional al nivel de la serie.

Holt-Winter aditivo

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$

$$l_t = a(y_t - s_{t-m}) + (1-a)(l_{t-1} + b_{t-1})$$

$$b_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

---

[6] Para profundizar en el tema se recomienda ver (Alonso, 2020. p. 50).

[7] Para más detalles, consultar (Alonso, 2020. p. 53).

Con el método aditivo, el componente estacional se expresa en términos absolutos en la escala de la serie observada, y en la ecuación de nivel, la serie se ajusta estacionalmente restando el componente estacional.

Holt-Winter multiplicativo

$$\hat{y}_{t+h|t} = (l_t + hb_t)\, s_{t+h-m(k+1)}$$

$$l_t = a\frac{y_t}{s_{t-m}} + (1-a)(l_{t-1} + b_{t-1})$$

$$b_t = \beta * (l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$

$$l_t = \gamma\frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}$$

Con el método multiplicativo, el componente estacional se expresa en términos relativos (porcentajes), y la serie se ajusta dividiendo por el componente estacional.

Para un análisis profundo de los métodos de suavizamiento en los modelos estadísticos de pronósticos de series de tiempo para científicos de datos consultar Alonso (2020).

2.3. Modelos autorregresivos (AR)

La familia de modelos conocidos como procesos autorregresivos[8] tiene como principal característica suponer que la historia de la variable aleatoria incide en el comportamiento futuro y presente de la serie de tiempo analizada. En general, se define de la forma:

$$y_t = \delta + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + \varepsilon_t$$

donde, δ es una constante, Φ_i son los coeficientes del modelo y ε_t es un término aleatorio.

Es importante tener en cuenta que, en los modelos AR, el valor a predecir depende de un promedio ponderado de p periodos pasados. Además, asociada a la definicion de modelos AR esta el concepto de estacionariedad[9].

[8] Para profundizar en el tema, consultar (Alonso, 2020) capítulo 6.

[9] Ver (Alonso, 2020) capítulo 4 para más detalles.

2.4. Modelos de media móvil (MA)

La familia de modelos de media móvil[10] (Moving Average, por sus siglas en inglés) es parecida a la familia de modelos AR, pero esta tiene como característica suponer que la historia de la variable aleatoria incide en el comportamiento futuro y presente de la serie de tiempo analizada, través de la parte estocástica. Estas familias de modelos se definen de la siguiente manera:

$$y_t = \delta + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \cdots + \Phi_p \varepsilon_{t-p} + \varepsilon_t$$

donde, el valor a predecir depende de un promedio ponderado de q periodos pasados de un proceso denominado ruido blanco[11].

2.5. Modelos ARMA

Los procesos ARMA[12] son una mezcla formal de los métodos AR y MA. Por lo cual heredan las características ambos. La definición formal de un proceso ARMA es:

$$y_t = \delta + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

donde, la variable aleatoria y_t se puede ver como una combinación de un proceso AR y un proceso MA, es decir, es un promedio ponderado de los últimos p periodos, un promedio ponderado de los últimos q periodos pasados de un proceso ruido blanco. También, hay un término error y una constante.

2.6. Modelos ARIMA

En la sección previa se definieron los modelos ARMA, los cuales son una combinación de modelos AR y MA. Además, estos modelos satisfacen dos importantes propiedades de series de tiempo, denominadas estacionariedad y ruido blanco[11].

En la práctica no es usual hallar series de tiempo que satisfagan las propiedades antes mencionadas, pero existe un procedimiento para determinar en cuantos pasos la serie de tiempo se vuelve estacionaria. Este procedimiento se llama integración. Para mayor detalle consultar (Alonso, 2020. p. 142).

---

[10] Consultar (Alonso, 2020) capítulo 7.

[11] Para más detalles de un proceso ruido blanco, ver sección 4.3 (Alonso, 2020).

[12] Para más detalles consultar (Alonso, 2020) capítulo 8.

Para entender la integración, supongamos se tiene una serie de tiempo:

$$\{y_t | t \epsilon 1,2, \dots T\}$$

La cual no satisface la propiedad de estacionariedad, pero que cumple que la diferencia de dos términos sucesivos si es estacionaria. Es decir, la serie a continuación lo es:

$$\Delta y_t = y_t - y_{t-1}$$

Si al diferenciar (restar) términos sucesivos de la serie, obtenemos una serie estacionaria. Entonces diremos que la serie original es integrada de orden uno $I(1)$. En general, si una serie debe ser diferenciada d veces para volverla estacionaria, se denotará I(d) y significará que es integrada de orden d.

Formalmente, una serie de tiempo ARIMA[13] tiene la forma:

$$\Delta^d y_t = \delta + \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Es decir, es combinacion lineal de un modelo autorregresivo de orden $\boldsymbol{p}$, un modelo de media movil de orden $\boldsymbol{q}$ y ademas es integrada de orden $\boldsymbol{d}$, denotada por ARIMA(p, d, q)[14].

[13] Para profundizar más sobre ARIMA consultar (Alonso, 2020) capítulo 8.

[14] Tomado del capítulo 8 del texto (Alonso, 2020).

2.7. Modelos de ensamble o de combinación de pronósticos

Los ensambles, como su nombre lo indica, son modelos generados a partir de combinaciones de otros modelos definidos para la estimación de pronósticos.

El combinar o ensamblar diferentes métodos es una buena forma de mejorar los pronósticos. Para una introducción a la combinación de pronósticos o modelos de ensamble se recomienda Alonso (2020).

Definamos la notación a utilizar en adelante. $f_{i,t}$ será el pronóstico (fuera de muestra) del método i para el periodo t, f_t^c será el pronóstico combinado, H el número de períodos pronosticados y, N el número total de métodos de pronósticos que serán combinados.

En la Tabla 1 se da una breve descripción de los modelos de ensamble que serán usados. Además, las fórmulas que los definen y su funcionalidad.

Tabla 1. Métodos de ensamble de modelos.

Modelo	Fórmula	Descripción
Ensamble por el método del promedio simple	$f_t^c = \frac{1}{N}\sum_{i=1}^{N} f_{i,t}$	Combinacion lineal de $\boldsymbol{i}$ métodos de estimación de pronósticos para el periodo t.
Ensamble por el método de la mediana	$f_t^c = \begin{cases} f_{\left(\frac{N}{2}+0{,}5\right),t} \\ \frac{1}{2}\left(f_{\left(\frac{N}{2}\right),t} + f_{\left(\frac{N}{2}+1\right),t}\right) \end{cases}$	Aquí, $f_{i,t}$ corresponde al pronóstico en la posición $\boldsymbol{i}$ después de ordenar de mayor a menor todos los N pronósticos del periodo t.
Ensamble por el método de la media truncada	$f_f^c = \frac{1}{N(I-2\lambda)}\sum_{i=\lambda N+1}^{(1-\lambda)N} f_{i,t}$	Este método descarta los valores pronosticados más grandes y más bajos. Formalmente, sea λ la proporción de datos que serán eliminados.
Ensamble por el método de la media Winsor	$f_t^c = \frac{1}{N}\left[k f_{(k+1),t} + \sum_{i=K+1}^{N-K} K f_{(N-K),t}\right]$	Método que al igual que el de media truncada, también está concebido para evitar influencia de valores extremos, pero con la diferencia de que no los elimina, sino que hace una ponderación de estos valores extremos. Formalmente se define como: Donde, $K = \lambda N$.
Modelos basados en regresión	$y_t = \beta_0 + \sum_{i=1}^{N} B_i f_{i,t} + \varepsilon_t$	Enfoque introducido por (Crane & Crotty, 1967), donde la observación de la serie es la variable dependiente y las variables explicativas son cada uno de los pronósticos generados por los modelos considerados anteriormente.

Fuente: Elaboración propia, a partir del capítulo 11 de Alonso (2020).

Hasta aquí, se han descrito los diferentes métodos de estimación de pronósticos que serán usados. Es importante hacer notar que para que estos métodos sean óptimos, los datos de series de tiempo deben estar completos. Es decir, las bases de datos deben tener la totalidad de los registros. Luego, a partir de ahí se procede a realizar el análisis de las series.

Muchas veces los datos de serie de tiempo no están completos y debe procederse a llenar estos huecos de datos faltantes. El mecanismo para realizar esto, se conoce como ***imputación de datos perdidos.*** Existen diferentes técnicas o maneras de imputar datos perdidos. En la sección 2.8, abordaremos las técnicas de imputar datos para series de tiempo que están en el paquete ***ImputeTS*** (Moritz & Bartz-Beielstein, 2017).

2.8. Métodos de imputación de datos perdidos

En estadística, el proceso de reemplazar valores perdidos se llama imputación. En esta sección, daremos una introducción a este proceso para datos en formato de serie de tiempo, usando el lenguaje de programación R. Para ello nos basaremos en el trabajo desarrollado por Moritz & Bartz-Beielstein (2017).

La imputación de series de tiempo es un subcampo especial en el área de investigación de la imputación. En CRAN[15] existen varios paquetes que resuelven el problema de la imputación de datos multivariados. Los más populares y maduros son ***AMELIA*** (Honaker, 2011), ***VIM*** (Kowarik & Templ, 2016) y ***missMDA*** (Josse & Husson,2016). Sin embargo, dado que estos paquetes están diseñados para la imputación de datos multivariados, no se pueden usar para series de tiempo univariadas. El paquete ***ImputeTS*** (Moritz & Bartz-Beielstein, 2017) se dedica exclusivamente a imputación de datos en formato de series de tiempo univariadas. Sin embargo, hay algunos otros paquetes que incluyen funciones de imputación como complemento a su funcionalidad de paquete principal. Los más destacados son ***zoo*** (Zeileis y Grothendieck, 2005) y ***forecast*** (Hyndman, 2017). Ambos paquetes ofrecen también algunas funciones avanzadas de imputación de series de tiempo.

ImputeTS

El paquete ***ImputeTS*** se puede encontrar en CRAN y, es un paquete fácil de usar que ofrece varias utilidades para series de tiempo numéricas, univariadas y equidistantes. Numérica significa que las observaciones son cantidades medibles, que se pueden describir como un número. Univariadas significa que solo hay un atributo que se observa a lo largo del tiempo. Y, equidistantes significa que lleva a una secuencia de observaciones tal que, los incrementos de tiempo entre puntos de datos sucesivos son iguales.

[15] https://cran.r-project.org/

Descripción general del paquete ImputeTS

El paquete tiene como objetivo ayudar al científico de datos en el proceso de sustitución de los valores faltantes en series de tiempo a través de varios algoritmos. En la Tabla 2, se presentan los algoritmos, se dan las opciones en R y una breve descripción.

Tabla 2. Métodos de imputación en el paquete imputeTS

Función	Opción	Descripción
na.interpolation	linear	Imputación por interpolación lineal
	spline	Imputación por interpolación spline
	stine	Imputación por interpolación de Stineman
na.kalman	structTS	Imputación por modelado estructural y suavización de Kalman
	auto.arima	Imputación por representación espacial ARIMA y suavización de Kalman
na.locf	locf	Imputación por reemplazo de la última observación
	nocb	Imputación por reemplazo de la próxima observación
na.ma	simple	Imputación por reemplazo de la media móvil
	linear	Imputación por reemplazo de la media móvil lineal ponderada
	Exponential	Imputación por reemplazo de la media móvil exponencial ponderada
na.mean	mean	Imputación por reemplazo de la media
	median	Imputación por reemplazo de la mediana
	Mode	Imputación por reemplazo de la moda
na.random		Imputación por reemplazo de valor aleatorio
na.replace		Imputación por reemplazo de un valor definido a discreción
na.seadec		Imputación de valores faltantes por descomposición estacional
na.seasplit		Imputación de valores faltantes por partición estacional

Fuente: Elaboración propia a partir de (Moritz & Bartz-Beielstein, 2017).

Para más información sobre métodos de imputación y detalles del paquete ***ImputeTS*** se recomienda consultar (Moritz & Bartz-Beielstein, 2017).

3. ESTADO DEL ARTE

Las técnicas usadas en ciencias de datos para pronosticar series de tiempo son muchas; hay técnicas estadísticas, técnicas de Machine Learning (aprendizaje de maquina) y técnicas de Deep Learning (aprendizaje profundo). En el contexto de este trabajo, nuestro interés se centrará en las técnicas estadísticas, particularmente, aquellas que son útiles para las estimaciones de pronósticos de objetos denominados series de tiempo. Dado que la muestra disponible es relativamente pequeña, emplear modelos de aprendizaje de máquina o profundo no sería adecuado.

Algunos trabajos científicos que muestran las aplicaciones de las técnicas estadísticas en series de tiempo en la industria son:

- Jiménez, Miranda & Ganvita, (2008). Donde los autores estudian el sector de ganadería bovina en Colombia. Para, a través de series de tiempo estimar pronósticos del inventario de ganado.
- Hernández, Ramírez & Zuleta, (2011). En este trabajo los autores analizan a través de series de tiempo el pronóstico de producción de caña de azúcar en México.
- Quispe, (2015). El autor realiza un análisis económico de la producción agrícola y alimentaria en relación con el cambio climático, en el departamento de Puno, Perú.
- Alonso, Díaz, Estrada, Figueroa & Tamura, (2019). Emplean modelos jerárquicos para encontrar el mejor modelo para pronosticar los galones de gasolina corriente demandados en Bogotá, Colombia.
- Alonso & Arcila, (2019). Emplean un modelo de predicciones diarias para contratos de futuros del azúcar.

Para efectos de estimación de pronósticos del sector agropecuario. Particularmente, en el Distrito Especial de Buenaventura no se encontró material académico ni científico relacionado. En este trabajo se estimarán pronósticos de los precios para los productos Yuca Chirosa, Piña Gold y Banano Criollo. Así, aportar nuestro grano de arena en la construcción académica que aporte desarrollo al sector agropecuario de Colombia.

Usaremos la metodología denominada CRISP-DM, la cual es común en ciencias de datos para el desarrollo de trabajos y proyectos.

4. METODOLOGIA

La metodología CRISP-DM está descrita y definida por los pasos siguientes:

- Entender el modelo de negocio. En este caso, el sector agropecuario de Colombia, asociado a Buenaventura.
- Analizar y comprender las bases de datos.
- Limpieza, transformación, imputación de datos perdidos y preparación de los datos.
- Construcción de los modelos estadísticos para series de tiempo.
- Evaluación de los modelos de series de tiempo obtenidos basándonos en la métrica RMSE.
- Automatización de la actualización de la información cada semana y la construcción de los nuevos pronósticos con la última información disponible.
- Creación de un tablero de mando o dashboard.
- Conclusiones pertinentes del estudio efectuado.

4.1. Entendimiento del sector agropecuario en Buenaventura, Colombia

Características demográficas

En el año 2014 se llevó a cabo el tercer Censo Nacional Agropecuario (CNA) en Colombia. Se tenían dos censos previos del sector agropecuario llevados a cabo en los años de 1950 y 1970 respectivamente. El primero cubrió 16 departamentos, el segundo cubrió 21 de ellos. Este último cubrió la totalidad del país. Además, tuvo en cuenta toda la actividad productiva agropecuaria y no agropecuaria desarrollada en el área rural dispersa del país.

El interés principal del trabajo desarrollado en este documento, será analizar los datos de cultivos y hogares del Distrito Especial de Buenaventura, obtenidos por el DANE en el tercer CNA. (Fuente: www.dane.gov.co).

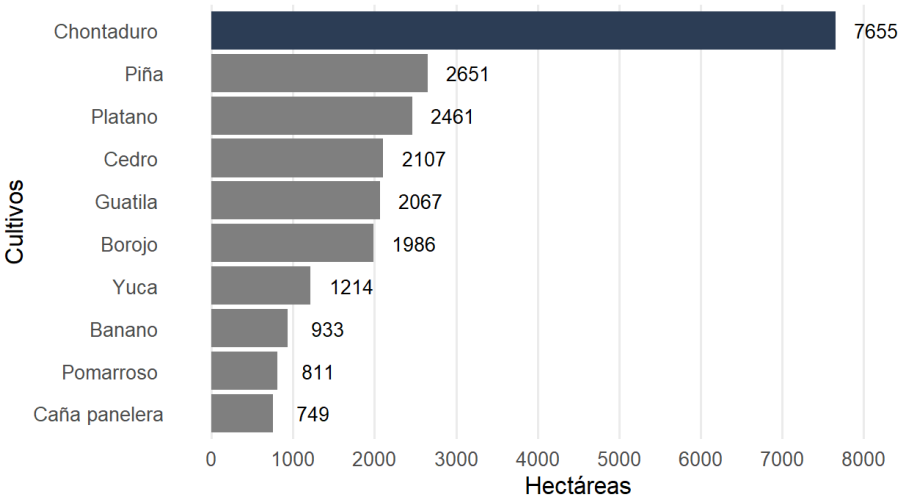
Basándonos en estos datos, se obtuvo que en Buenaventura la cantidad de hogares campesinos censados fueron 11.846, donde más de 9000 se beneficiaban directamente de la siembra y la cosecha. El 51% de los campesinos eran hombres y, el 49% eran mujeres. Además, el 90% se reconoció como afrodescendiente.

El 16% de los jefes de hogar censado era analfabeta, el 60% no se encontraba realizando ningún tipo de estudio y aproximadamente el 50% tenía como logro académico más alto básica primaria.

Características asociadas a los productos cultivados

El análisis preliminar de los datos de cultivo y hogares en Buenaventura (2014) arrojó que el Chontaduro fue el producto más sembrado en Buenaventura, con un total de 7655 hectáreas. En orden decreciente se pudo ver que la Piña Gold y el Plátano Hartón Verde eran los siguientes productos más sembrados, con un total de 2651 y 2461 hectáreas, respectivamente.

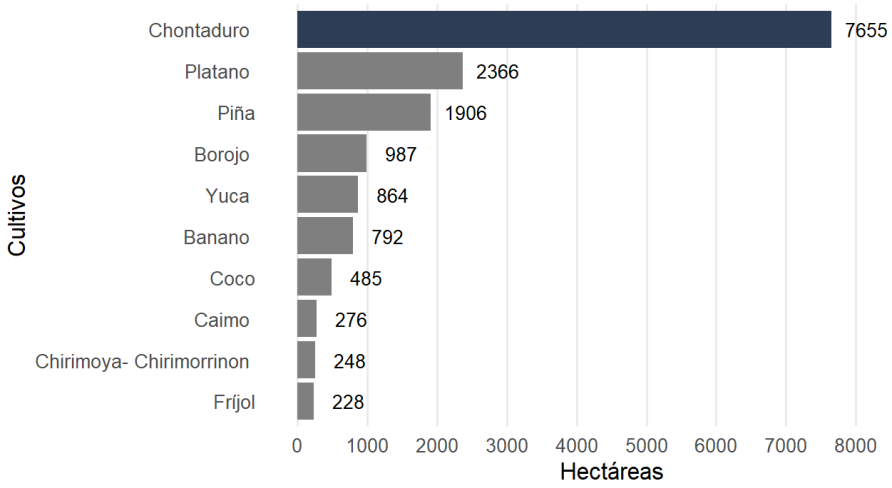
Figura 1. Cultivos más sembrados en el Distrito Especial de Buenaventura. 2014. (Hectáreas).



Fuente: Elaboración propia.

Dentro de la caracterización, el Chontaduro fue el producto más cosechado en Buenaventura con un total de 7655 hectáreas. Que, además, coincide con la cantidad de hectáreas sembradas. Siguiendo con el análisis se pudo ver que el Plátano Hartón Verde fue segundo producto más cosechado con 2366 hectáreas, seguido de la Piña Gold con 1906 hectáreas.

Figura 2. Cultivos más cosechados en el Distrito Especial de Buenaventura. 2014. (Hectáreas).



Fuente: Elaboración propia.

De las dos gráficas anteriores que representan los productos más sembrados y cosechados en Buenaventura, se observa que hay otros productos cultivados, tales como el Cedro, la Guatila, la Pomarrosa, la Caña Panelera y el Borojó.

Para efectos del trabajo efectuado en este documento se optó por trabajar con los productos cosechados, ya que finalmente son las cosechas las que realmente otorgan beneficios a las comunidades campesinas asentadas en Buenaventura.

Así las cosas, los productos candidatos fueron: Chontaduro, Piña Gold, Plátano Hartón Verde, Borojó, Yuca Chirosa y Banano Criollo. Después, se constató que en la base de datos del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA), no hay registro del Chontaduro ni Borojó. Además, del Plátano de Hartón Verde no se reportan precios desde el 31 de marzo del 2017. Por tal razón, se seleccionó la Piña Gold, el Banano Criollo y la Yuca Chirosa.

Dentro del análisis efectuado, se realizó la lista de las veredas que más hectáreas cosechan de los productos Banano Criollo, Piña Gold y Yuca Chirosa. El resultado esta resumido en la Tabla 3.

Tabla 3. Veredas con mayores áreas cosechadas de Piña Gold, Yuca Chirosa y Banano Criollo en el Distrito de Buenaventura. 2014. (Hectáreas).

	Vereda	Área
1	ZACARÍAS	883.1
2	CISNEROS	573.7
3	EL TIGRE	542.6
4	ZABALETAS	492.2
5	LLANO BAJO	411.2
6	BAZÁN	339.6
7	EL PITAL	267.2
8	SAN JOSÉ DE ANCHICAYÁ	247.2
9	PUERTO MERIZALDE	229.2
10	EXPANSIÓN URBANA	123.6
11	BOCAS DEL SAN JUAN	89.0
12	SAN ANTONIO DE YURUMANGUÍ	68.4
13	SILVA	60.0
14	EL CARMEN	42.0
15	CÓRDOBA	41.5
16	GAMBOA	39.1
17	BAJO CALIMA	35.5
18	GUADUALITO	35.3
19	SAN ISIDRO	27.9
20	PITAL	6.9
21	LA CASCADA	0.5

Fuente: Elaboración propia.

4.2. Análisis y comprensión de la base de datos

Los datos de los precios de los productos agrícolas, se obtuvieron del Sistema de Información de Precios del Sector Agropecuario (SIPSA) del DANE. El SIPSA recolecta desde el 12 de junio del 2012 la información de los precios mayoristas en las distintas centrales de abastos y plazas de mercado del país, con una periodicidad diaria, semanal y mensual. En esta base de datos tenemos disponibles los precios de venta mayorista en Buenaventura para los cultivos que son más representativos en esta región del país.

La Tabla 4 que se presenta a continuación, es descriptora de las variables y datos disponibles en el SIPSA. En ella se hace un análisis simple de la información, en busca de conocer sus características y particularidades.

Tabla 4. Descripción de las variables de los datos disponibles en el Sistema de Información de Precios del Sector Agropecuario – SIPSA.

Nombre	Descripción
NOM_ABASTO	Lugar donde se vende el cultivo
COD_ART	Código del cultivo
PROM_DIARIO	Precio promedio del cultivo
VAL_MIN	Precio mínimo del cultivo
VAL_MAX	Precio máximo del cultivo
Date	Fecha en la que se capturó el dato
NOM_ART	Nombre del cultivo
VAR_DIARIA	Precio del cultivo

Fuente: Elaboración propia.

Los boletines que proporciona el DANE para la consulta de precios mayoristas de los productos tienen una estructura, donde las filas representan el precio mínimo, máximo y promedio de un producto en una fecha determinada y, las columnas describen los datos del producto, su lugar de venta y los precios asociados.

Se encontró que el precio de la Piña se reporta para distintas variedades de ésta. Hay información disponible para la Piña Manzana y la Piña Gold. En este producto, nos enfocamos en el tipo Gold, por ser el tipo de Piña que se cultiva en el Distrito Especial de Buenaventura.

La cantidad de datos y el rango de fechas disponibles para cada cultivo se describe en la Tabla 5.

Tabla 5. Disponibilidad de datos para la Piña Gold, Yuca Chirosa y Banano Criollo en el Sistema de Información de Precios del Sector Agropecuario – SIPSA para el mercado de Buenaventura. (Semanas).

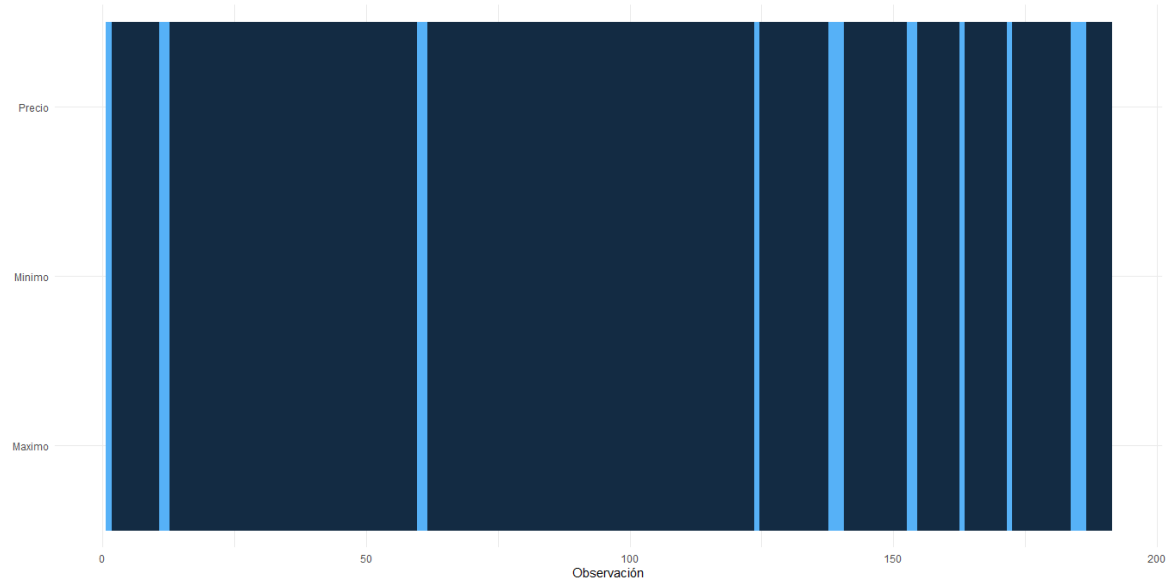
Cultivo	Cantidad de Observaciones	Desde	Hasta
Piña Gold	191	2017-07-14	2021-03-05
Yuca chirosa	455	2012-06-22	2021-03-05
Banano criollo	452	2012-07-13	2021-03-05

Fuente: Elaboración propia.

En el análisis preliminar realizado a los datos, aparentemente no había datos perdidos, pero luego de una inspección manual, se encontraron que existían espacios no registrados entre algunas fechas, lo que motivó a mapear las fechas nuevamente en el mismo intervalo de tiempo. El resultado obtenido fue el de la Tabla 5, donde se puede observar la cantidad de datos disponibles. Luego, se realizó otra inspección mediante un mapa de valores perdidos, que permite visualizar la fila exacta donde hay un valor nulo en la serie de tiempo. El eje vertical muestra las variables de nuestra base de datos y en el eje horizontal el lugar temporal donde hay un valor perdido.

En la Ilustración 2, se presenta un análisis de los valores faltantes de la Piña Gold.

Ilustración 2. Mapa de valores faltantes para la serie de la Piña Gold en el mercado de Buenaventura.

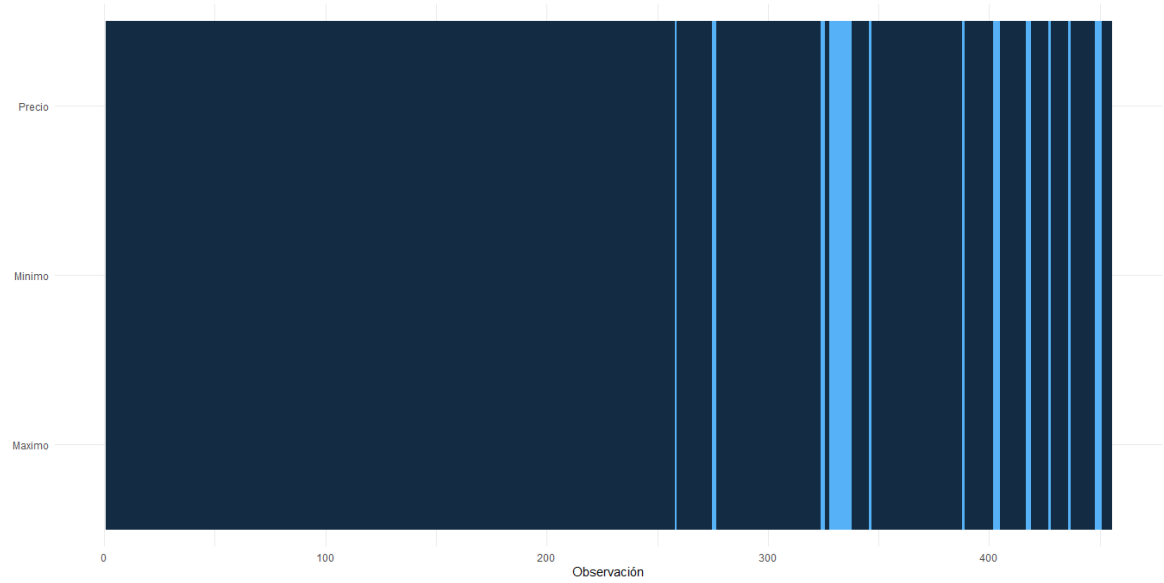


Fuente: Elaboración propia.

Para el precio de la Piña Gold, tenemos valores perdidos en las semanas del 14 de julio, 22 y 29 de septiembre del 2017 y del 31 de agosto, 7 de septiembre de 2018, 22 de noviembre de

2019, 28 de febrero, 6 y 13 de marzo, 12 y 19 de junio, 21 de agosto y 23 de octubre de 2020. Además, para las semanas que inician el 15, 22 y 29 de enero de 2021. Este cultivo no presenta patrones en los datos faltantes. Es decir, los datos de los precios faltantes son aleatorios.

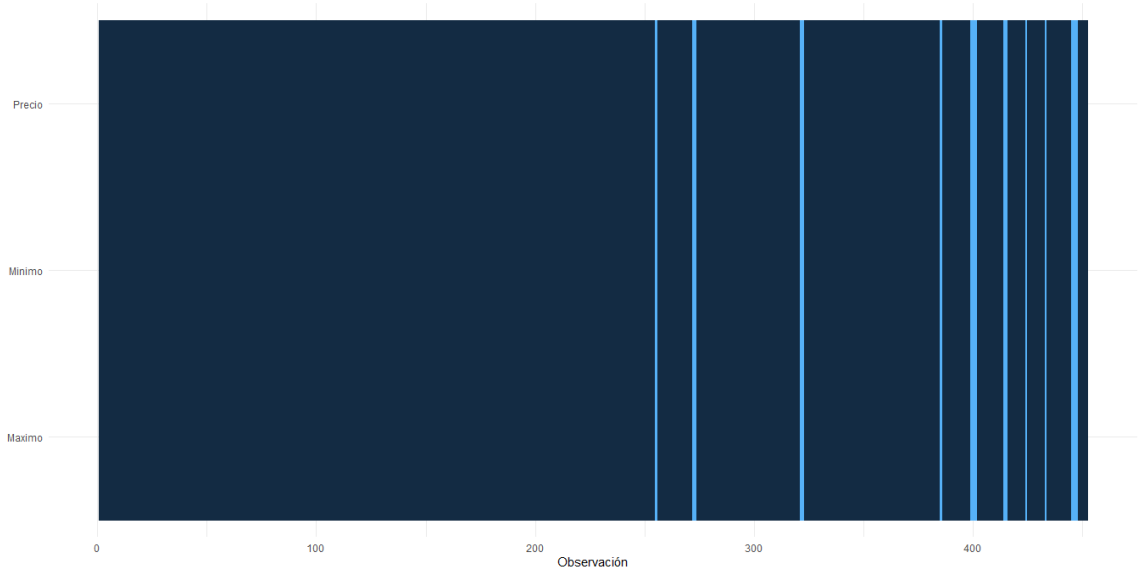
Ilustración 3. Mapa de valores faltantes para la serie de la Yuca Chirosa en el mercado de Buenaventura.



Fuente: Elaboración propia.

Para el precio de la Yuca Chirosa, los datos perdidos están en las semanas del 26 de mayo, 22 y 29 de septiembre del 2017, 31 de agosto, 7 y 28 de septiembre, 5, 12, 19 y 26 de octubre, 2, 9, 16, 23 y 30 de noviembre del 2018, 1 de febrero, 22 de noviembre de 2019, 28 de febrero, 6 y 13 de marzo, 12 y 19 de junio, 21 de agosto y 23 de octubre de 2020, 15, 22 y 29 de enero de 2021. El precio de este producto tampoco presenta patrones en los datos faltantes.

Ilustración 4. Mapa de valores faltantes para la serie del Banano Criollo en el mercado de Buenaventura.



Fuente: Elaboración propia.

Los datos faltantes correspondientes al precio del Banano Criollo, están en las semanas de: 26 de mayo, 22 y 29 de septiembre de 2017; 31 de agosto y 7 de septiembre de 2018; 22 de noviembre del 2019; 28 de febrero, 6 y 13 de marzo, 12 y 19 de junio, 21 de agosto y 23 de octubre de 2020; 15, 22 y 29 de enero de 2021. Al igual que los precios de los otros productos, este tampoco presenta patrones en los datos faltantes, es decir, los datos perdidos son aislados.

4.3. Limpieza, transformación, imputación de datos perdidos y preparación de los datos.

La metodología CRISP-DM quedó descrita al inicio de la sección 4. En este marco de trabajo para el científico de datos, la limpieza de datos y transformación de los mismos, juegan un rol importante pues, a partir de los datos y su veracidad se desprende el estudio. Para efectos del análisis que realizaremos en este documento y, después de verificar los datos de los precios de los productos, las etapas de limpieza y transformación se omitirán dado que no son necesarias. Sin embargo, faltan datos de precios de los productos, por tal razón las etapas usadas y explicadas a continuación son, imputación de datos perdidos y preparación de los datos, para posteriormente obtener los modelos de generación de pronósticos.

Imputación de datos perdidos y preparación de los datos

Como se expresó anteriormente, existen datos perdidos en las series de tiempo de los precios de los productos a analizar. Para no alterar el tamaño de las series de cada cultivo, es importante imputar los datos perdidos.

El proceso seguido por los autores para la imputación de datos perdidos en este documento consta de tres etapas a saber:

- Buscar plazas de mercados cercanas a Buenaventura y, analizar si en estas plazas se reportaron precios los días que coinciden con los datos de los precios faltantes para los productos aquí estudiados. De esta manera por proximidad de mercados usar este precio para sustituirlo en el dato faltante de las series que se tienen.
- Buscar plazas de mercados alejadas y analizar la misma situación explicada antes.
- Usar los métodos de imputación para científicos de datos descritos en la sección 2.8.

Método de imputación por proximidad de plazas de mercado a Buenaventura

Para esta etapa, se tuvieron en cuenta las series de precios de cultivos en las ciudades más cercanas al Distrito Especial de Buenaventura, donde también se registran los precios mayoristas del sector agropecuario en el sistema de información SIPSA. Es importante mencionar que, en el departamento del Valle del Cauca, las ciudades que registran precios son: Buenaventura, Cali, Cartago, Palmira y Tuluá. De esta manera, por cercanía se infiere que los precios no deben tener una variación muy alta y pueden usarse en los lugares faltantes.

La Tabla 6, muestra la distancia de las ciudades más cercanas al Distrito de Buenaventura. Así, se procedió a investigar los precios de los cultivos en los mercados de dichas ciudades, para determinar si en los días en los que no hubo registro, en el mercado de Buenaventura en estas ciudades sí se realizó el registro de los precios de los cultivos estudiados en este documento.

Tabla 6. Distancias entre el Distrito de Buenaventura y las ciudades aledañas. (Kilómetros)

Ciudad	Distancia
Cali	124
Tuluá	137
Palmira	144
Cartago	225

Fuente: Elaboración propia.

Teniendo en cuenta la cercanía entre mercados, los precios de los cultivos en Cali y Tuluá pueden ser la mejor aproximación a la hora de estimar los precios perdidos del mercado de Buenaventura. Por lo tanto, estas dos ciudades fueron las elegidas. Por otra parte, la ciudad de Cali tiene precios para los mercados de Cavasa y Santa Helena. Así, se tienen tres

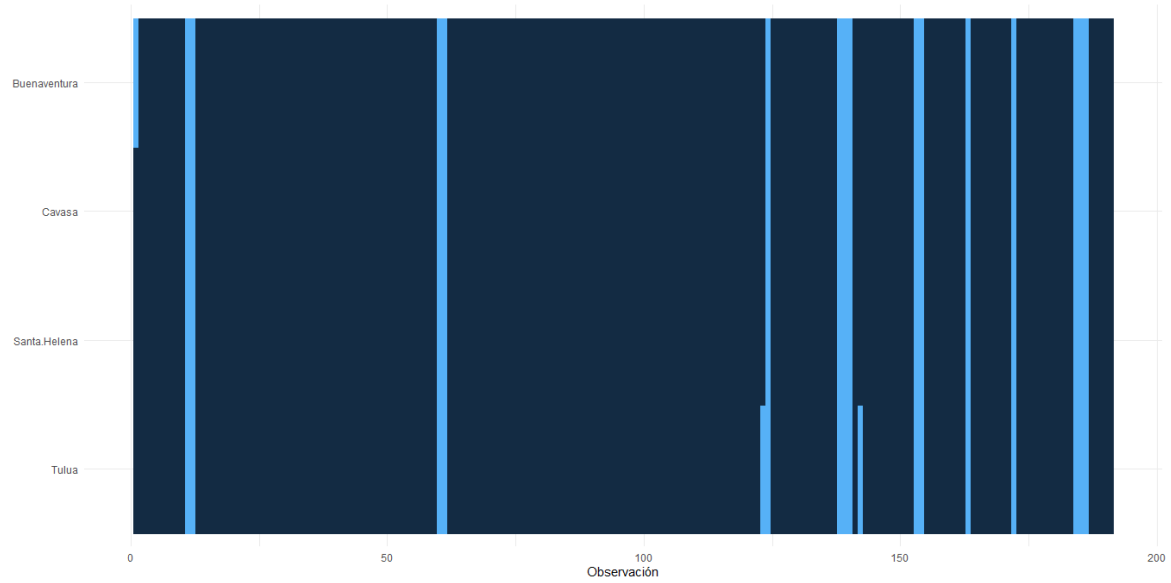
mercados para entender qué pasó con los precios de la Piña Gold, el Banano Criollo y la Yuca Chirosa, en las semanas que no hubo reporte de precios de estos productos en el mercado del Distrito de Buenaventura.

Al analizar los mercados, se encontró que Cavasa sí reporta precios para la Yuca Chirosa en las semanas del 22, 29 de junio y, 20 de julio de 2012. El mercado de Santa Helena solo reporta el 22, 29 de junio y del 10 de agosto de 2012. Desde estos mercados se reportan precios para un cultivo llamado Yuca ICA, el cual es diferente a la Yuca Chirosa. En cambio, en el mercado de la ciudad de Tuluá sí se reportan precios para la Yuca Chirosa. Se debe tener en cuenta que para realizar la comparación de precios y conocer qué pasó con el precio de los cultivos en las fechas en que no se reportó, utilizamos el mismo rango de fechas de las series que tenemos disponibles para el Distrito de Buenaventura.

Para la verificación y comparación de la calidad de datos entre mercados y, elegir el precio idóneo, se utilizó mapas de valores faltantes.

En la Ilustración 5, se presenta el análisis para la Piña Gold. Posteriormente, la Ilustración 6 e Ilustración 7 presentan un análisis similar para la Yuca Chirosa y el Banano Criollo.

Ilustración 5. Mapa de valores faltantes para la serie de la Piña Gold en el mercado de Buenaventura, Cavasa, Santa Helena y Tuluá.

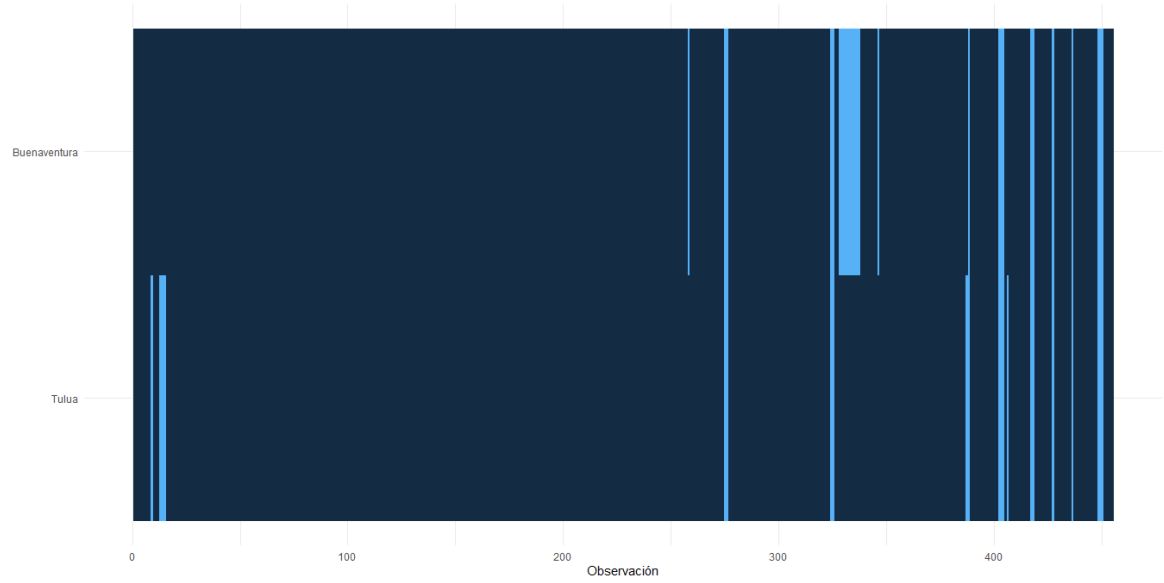


Fuente: Elaboración propia.

Se observó que todos los mercados tienen 15 valores faltantes en las mismas fechas. Las semanas sin datos para precios son las que inician el: 22 y 29 de septiembre del 2017, 31 de agosto y 7 de septiembre del 2018, 22 de noviembre de 2019, 28 de febrero, 6 y 13 de marzo,

12 y 19 de junio, 21 de agosto y 23 de octubre de 2020, el 15, 22 y 29 de enero de 2021. Además, estos valores de precios que faltan en Tuluá son aislados.

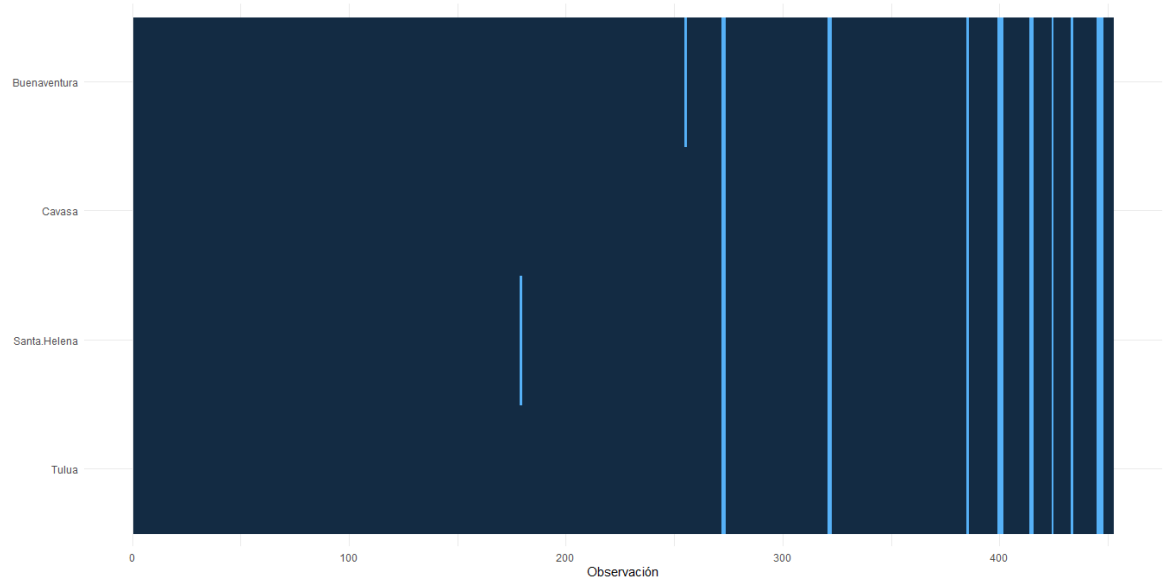
Ilustración 6. Mapa de valores faltantes para la serie de la Yuca Chirosa en el mercado de Buenaventura y Tuluá.



Fuente: Elaboración propia.

Como se mencionó anteriormente, los mercados de la ciudad de Cali no reportan precios para la Yuca Chirosa; por lo tanto, la comparación se realiza con el mercado de Tuluá. Se observó un comportamiento similar en 15 datos perdidos entre estos mercados en las semanas del 22 y 29 de septiembre del 2017, 31 de agosto y 7 de septiembre del 2018, 22 de noviembre de 2019, 28 de febrero, 6 y 13 de marzo, 12 y 19 de junio, 21 de agosto y 21 de octubre del 2020, 15, 22 y 29 de enero del 2021.

Ilustración 7. Mapa de valores faltantes para la serie del Banano Criollo en el mercado de Buenaventura, Cavasa, Santa Helena y Tuluá.



Fuente: Elaboración propia.

Se observó un comportamiento similar en la pérdida de datos para los 4 mercados. En total son 15 valores perdidos por cada mercado en las semanas del 22 y 29 de septiembre del 2017, 31 de agosto y 7 de septiembre del 2018, 22 de noviembre de 2019, 28 de febrero, 6 y 13 de marzo, 12 y 19 de junio, 21 de agosto y 23 de octubre de 2020, 15, 22 y 29 de enero de 2021.

Es importante mencionar que los mercados donde existe un patrón de pérdida simultánea de datos para la Piña Gold y Yuca Chirosa comparten las mismas fechas. No se logra identificar el evento que causo esta afectación de reportes de precios, pero parece ser de índole nacional.

Método de imputación por búsqueda de precios en plazas de mercado más alejadas

En esta etapa se buscaron los datos de los precios faltantes en las plazas de mercados que reportan precios en el SIPSA pero que geográficamente están a las afueras del departamento del Valle del Cauca o limítrofes.

Como se explicó anteriormente algunos datos perdidos para la Piña Gold, Yuca Chirosa y Banano Criollo en los mercados de Cavasa, Santa Helena y Tuluá, coinciden con la fecha de los datos perdidos de Buenaventura. Por esta razón, es necesario explorar otros mercados más lejanos, en busca de obtener datos que nos permitan sustituir los valores faltantes para los precios de los productos a analizar en las semanas que no hubo reporte en Buenaventura.

Para eso se exploraron los mercados de Palmira y Cartago, así como los departamentos de Cauca, Nariño, Tolima y Risaralda. En cuanto a Palmira y Cartago, estos mercados registran precios para la Piña Gold, pero no para la Yuca Chirosa. En el departamento del Cauca elegimos el municipio de Popayán, pero este mercado no registra precios semanales para la Piña Gold ni para la Yuca Chirosa. En Nariño, se exploró el mercado de Pasto y San Andrés de Tumaco, en cuanto a Pasto, éste no tiene registros de precios semanales de Yuca Chirosa. Por el lado de San Andrés de Tumaco, no existen precios de Yuca Chirosa ni Piña Gold. En Risaralda, exploramos los mercados de La Virginia y Pereira. En el municipio de Pereira existen dos mercados que registran precios: La 41 y Mercasa. En ninguno de éstos se registra precios para la Yuca Chirosa. En el mercado de la Virginia no se registran precios para la Yuca Chirosa ni la Piña Gold. Por último, el mercado de Ibagué, sí se registran los precios de los tres productos.

En resumen, elegimos los mercados según la disponibilidad de precios para cada uno de los productos y listamos las fechas de los datos perdidos en el mercado del Distrito de Buenaventura en las tablas 7, 8 y 9. Estos, se compararon con los precios de los mercados descritos en el párrafo anterior. A continuación, presentamos los mercados más lejanos para la Piña Gold, Yuca Chirosa y Banano Criollo, respectivamente.

Tabla 7. Disponibilidad de datos en el mercado de Palmira, Cartago, Pasto, La 41, Mercasa e Ibagué para la Piña Gold en las fechas donde no hubo reporte para el mercado de Buenaventura. (Pesos colombianos).

Fecha	Palmira	Cartago	Pasto	La 41	Mercasa	Ibagué
2017-07-14	769	1067	900	1175	967	1100
2017-09-22	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2017-09-29	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2018-08-31	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2018-09-07	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2019-11-22	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-02-28	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-03-06	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-03-13	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-06-12	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-06-19	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-08-21	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-10-23	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2021-01-15	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2021-01-22	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2021-01-29	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*

Nota: n.a. = dato no disponible.
Fuente: Elaboración propia.

Tabla 8. Disponibilidad de datos en el mercado de Ibagué para la Yuca Chirosa en las fechas donde no hubo reporte para el mercado de Buenaventura. (Pesos colombianos).

Fecha	Ibagué
2017-05-26	868
2017-09-22	*n.a.*
2017-09-29	*n.a.*
2018-08-31	*n.a.*
2018-09-07	*n.a.*
2018-09-28	1208
2018-10-05	1117
2018-10-12	1617
2018-10-19	1667
2018-10-26	1542
2018-11-02	1458
2018-11-09	1729
2018-11-16	1563
2018-11-23	1425
2018-11-30	1458
2019-02-01	1896
2019-11-22	*n.a.*
2020-02-28	*n.a.*
2020-03-06	*n.a.*

2020-03-13	*n.a.*
2020-06-12	*n.a.*
2020-06-19	*n.a.*
2020-08-21	*n.a.*
2020-10-23	*n.a.*
2021-01-15	*n.a.*
2021-01-22	*n.a.*
2021-01-29	*n.a.*

Nota: n.a. = dato no disponible.
Fuente: Elaboración propia.

Tabla 9. Disponibilidad de datos en el mercado de Popayán, La 41, Mercasa e Ibagué para el Banano Criollo en las fechas donde no hubo reporte para el mercado de Buenaventura. (Pesos colombianos).

Fecha	Popayán	La 41	Mercasa	Ibagué
2017-05-26	1250	1250	1133	1338
2017-09-22	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2017-09-29	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2018-08-31	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2018-09-07	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2019-11-22	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-02-28	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-03-06	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-03-13	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-06-12	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-06-19	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-08-21	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2020-10-23	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2021-01-15	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2021-01-22	*n.a.*	*n.a.*	*n.a.*	*n.a.*
2021-01-29	*n.a.*	*n.a.*	*n.a.*	*n.a.*

Nota: n.a. = dato no disponible.
Fuente: Elaboración propia.

Como se describió anteriormente, el mercado de Ibagué (Tolima) es el único que reporta precios para la Yuca Chirosa.

Los datos disponibles en el mercado Plaza La 21 de Ibagué no son suficientes para cubrir la totalidad de los datos perdidos para la Yuca Chirosa en Buenaventura. Además, para la Piña Gold y Yuca Chirosa, no hay datos disponibles en ningún mercado cercano, ni en los departamentos cercanos. Así las cosas, decidimos realizar la tercera etapa. Esta consiste en usar el paquete ***ImputeTS*** y la función del mismo nombre basados en el lenguaje de programación R.

Método de imputación de datos perdidos basados en el paquete ImputeTS

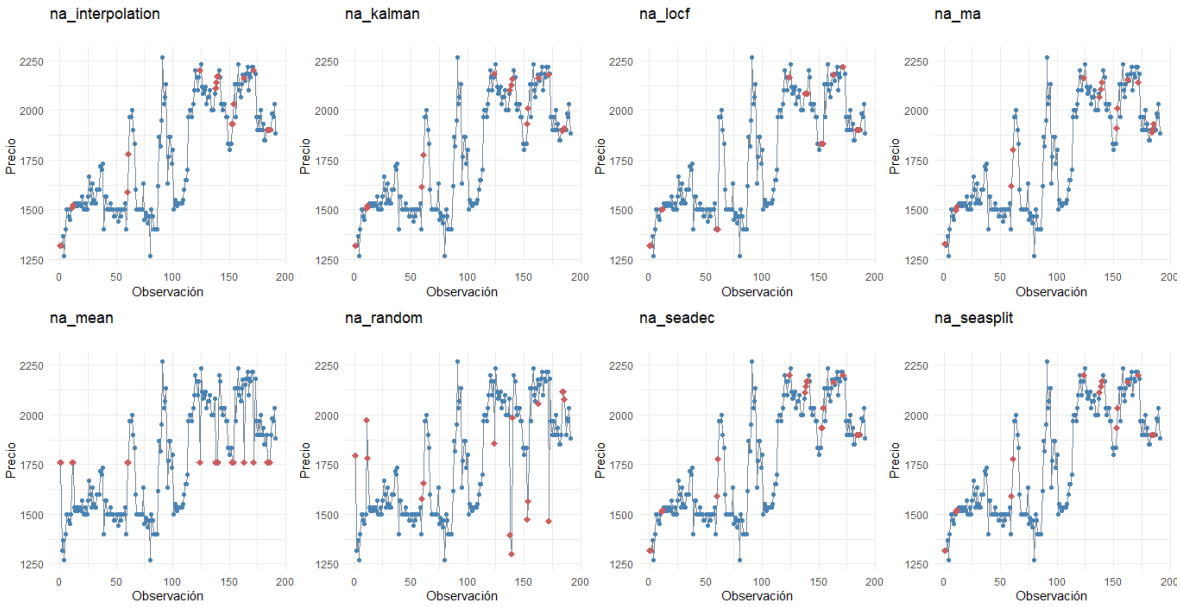
Como se mencionó anteriormente, el proceso para sustituir los valores faltantes a través de plazas cercanas y alejadas presentaba el mismo inconveniente. Es decir, en las plazas estudiadas, las fechas en que se tenían datos perdidos para los precios de los cultivos a analizar coincidía con las fechas de datos perdidos en Buenaventura. Por tal razón se procede a usar técnicas de imputación mediadas por el lenguaje de programación R.

Se hará uso del paquete ***ImputeTS*** y de la función que lleva el mismo nombre, para de esta manera sustituir los precios de los productos que no se tienen en las bases de datos obtenidas del SIPSA y de esta manera proceder a construir lo modelos de estimación de pronósticos.

El paquete ***ImputeTS*** internamente está constituido por diferentes métodos de imputación de datos faltantes y, una variedad de opciones para cada método. Una descripción breve de cómo funciona cada uno de los algoritmos de este paquete puede consultarse en la Tabla 2, de la sección 2.8.

Las ilustraciones 8, 9 y 10 describen gráficamente los resultados obtenidos después de usar el paquete ***ImputeTS*** para la imputación de valores faltantes en las bases de datos de los precios de los productos Piña Gold, Yuca Chirosa y Banano Criollo, respectivamente.

Ilustración 8. Imputación de datos en la serie de precios de la Piña Gold para el mercado de Buenaventura.



Fuente: Elaboración propia.

Se observó en la Ilustración 8 que los métodos na_mean y na_random presentan cambios bruscos en los valores imputados. En cambio, en los otros métodos, los cambios son más suaves. Las diferencias entre los otros métodos son más sutiles, por ello, se utilizó una tabla de variación porcentual para identificar estas variaciones en todos los algoritmos de imputación. De esta manera, tener un argumento complementario en la determinación del mejor método asociado a la imputación de valores faltantes para cada producto estudiado en este documento.

La Tabla 10, muestra la variación porcentual entre el valor imputado $\widehat{\boldsymbol{P}}_{\boldsymbol{t}}$, y la diferencia porcentual con respecto al precio observado inmediatamente anterior $\boldsymbol{P}_{(\boldsymbol{t-1})}\%$ y siguiente $\boldsymbol{P}_{(\boldsymbol{t+1})}\%$, para los precios de la Piña Gold.

Tabla 10. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios de la Piña Gold en el mercado de Buenaventura.

Fecha	Interpolation			Kalman			LOCF		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-07-14	1317.00	0.00	0.00	1318.50	0.00	-0.11	1317	0	0.00
2017-09-22	1511.00	0.73	0.73	1505.55	0.37	0.75	1500	0	0.00
2017-09-29	1522.00	0.73	0.72	1516.85	0.75	1.06	1500	0	2.20
2018-08-31	1589.00	13.50	11.89	1615.61	15.40	9.82	1400	0	0.00
2018-09-07	1778.00	11.89	10.63	1774.34	9.82	10.86	1400	0	40.50
2019-11-22	2200.00	1.52	1.50	2184.35	0.80	2.23	2167	0	3.05
2020-02-28	2112.25	1.40	1.38	2101.66	0.90	1.30	2083	0	0.00
2020-03-06	2141.50	1.38	1.37	2129.05	1.30	1.29	2083	0	0.00
2020-03-13	2170.75	1.37	1.35	2156.44	1.29	2.02	2083	0	5.62
2020-06-12	1933.00	5.46	5.17	1930.39	5.31	4.18	1833	0	0.00
2020-06-19	2033.00	5.17	4.92	2011.06	4.18	6.06	1833	0	16.37
2020-08-21	2164.00	-0.64	-0.65	2162.29	-0.72	-0.57	2178	0	-1.29
2020-10-23	2200.00	-0.77	-0.77	2179.14	-1.71	0.18	2217	0	-1.53
2021-01-15	1900.00	0.00	0.00	1897.85	-0.11	0.27	1900	0	0.00
2021-01-22	1900.00	0.00	0.00	1903.02	0.27	0.27	1900	0	0.00
2021-01-29	1900.00	0.00	0.00	1908.18	0.27	-0.43	1900	0	0.00

Tabla 10. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios de la Piña Gold en el mercado de Buenaventura. Cont.

Fecha	MA			Mean			Random		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-07-14	1329.20	0.00	-0.92	1758.22	0.00	-25.09	1628.35	0.00	-26.68
2017-09-22	1496.95	-0.20	1.06	1758.22	17.21	0.00	1726.57	31.60	-9.71
2017-09-29	1512.82	1.06	1.33	1758.22	0.00	-12.81	1447.75	-9.71	-13.99
2018-08-31	1619.73	15.69	11.33	1758.22	25.59	0.00	1284.91	12.67	5.09
2018-09-07	1803.18	11.33	9.08	1758.22	0.00	11.87	2057.13	5.09	18.66
2019-11-22	2162.27	-0.22	3.27	1758.22	-18.86	27.00	1627.62	-14.34	20.30
2020-02-28	2068.39	-0.70	1.87	1758.22	-15.59	0.00	2159.97	-32.99	-6.97
2020-03-06	2107.07	1.87	1.51	1758.22	0.00	0.00	1359.27	-6.97	53.02
2020-03-13	2138.89	1.51	2.86	1758.22	0.00	25.13	1292.04	53.02	10.73
2020-06-12	1910.36	4.22	5.24	1758.22	-4.08	0.00	2187.47	-19.60	6.25
2020-06-19	2010.41	5.24	6.10	1758.22	0.00	21.32	2065.22	6.25	36.23
2020-08-21	2155.20	-1.05	-0.24	1758.22	-19.27	22.28	1787.00	-5.62	4.59
2020-10-23	2140.67	-3.44	1.98	1758.22	-20.69	24.16	1449.20	-33.98	49.14
2021-01-15	1889.78	-0.54	0.85	1758.22	-7.46	0.00	1627.06	11.50	-0.07
2021-01-22	1905.93	0.85	1.29	1758.22	0.00	0.00	2119.42	-0.07	-1.97
2021-01-29	1930.50	1.29	-1.58	1758.22	0.00	8.06	1269.17	-1.97	-8.46

Tabla 10. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios de la Piña Gold en el mercado de Buenaventura. Cont.

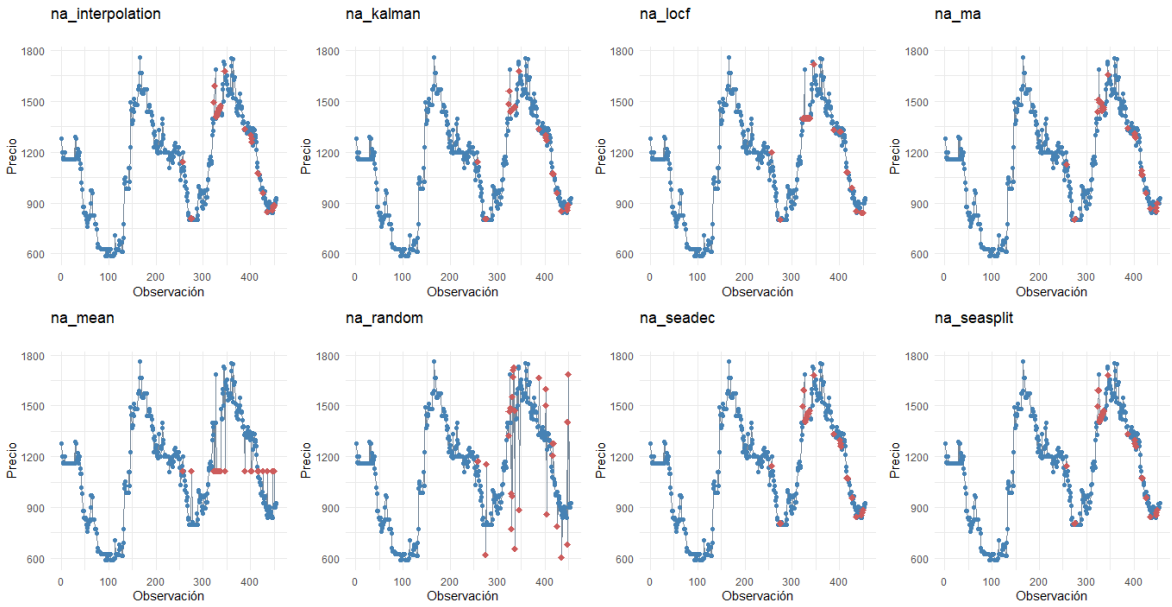
Fecha	Seadec			Seasplit		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-07-14	1317.00	0.00	0.00	1317.00	0.00	0.00
2017-09-22	1511.00	0.73	0.73	1511.00	0.73	0.73
2017-09-29	1522.00	0.73	0.72	1522.00	0.73	0.72
2018-08-31	1589.00	13.50	11.89	1589.00	13.50	11.89
2018-09-07	1778.00	11.89	10.63	1778.00	11.89	10.63
2019-11-22	2200.00	1.52	1.50	2200.00	1.52	1.50
2020-02-28	2112.25	1.40	1.38	2112.25	1.40	1.38
2020-03-06	2141.50	1.38	1.37	2141.50	1.38	1.37
2020-03-13	2170.75	1.37	1.35	2170.75	1.37	1.35
2020-06-12	1933.00	5.46	5.17	1933.00	5.46	5.17
2020-06-19	2033.00	5.17	4.92	2033.00	5.17	4.92
2020-08-21	2164.00	-0.64	-0.65	2164.00	-0.64	-0.65
2020-10-23	2200.00	-0.77	-0.77	2200.00	-0.77	-0.77
2021-01-15	1900.00	0.00	0.00	1900.00	0.00	0.00
2021-01-22	1900.00	0.00	0.00	1900.00	0.00	0.00
2021-01-29	1900.00	0.00	0.00	1900.00	0.00	0.00

Fuente: Elaboración propia.

Aunque gráficamente no es apreciable, el método LOCF es el que mejor se comporta, dado que es el único que garantiza que en los sitios donde la serie es plana, siga siendo plana. Lo anterior se constató con la tabla de variación porcentual.

El proceso descrito antes, se repite para los precios de la Yuca Chirosa.

Ilustración 9. Imputación de datos en la serie de precios de la Yuca Chirosa para el mercado de Buenaventura



Fuente: Elaboración propia.

En la Tabla 11, tenemos las variaciones porcentuales de los valores imputados para los precios de la Yuca Chirosa.

Tabla 11. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios de la Yuca Chirosa en el mercado de Buenaventura.

Fecha	Interpolation			Kalman			LOCF		
	$\hat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\hat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\hat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-05-26	1145.00	-4.58	-4.80	1142.26	-4.81	-4.58	1200	0	-9.17
2017-09-22	804.33	0.54	0.54	803.73	0.47	0.43	800	0	0.00
2017-09-29	808.67	0.54	0.54	807.18	0.43	0.72	800	0	1.62
2018-08-31	1495.67	6.83	6.40	1482.04	5.86	5.32	1400	0	0.00
2018-09-07	1591.33	6.40	6.01	1560.92	5.32	8.08	1400	0	20.50
2018-09-28	1407.27	0.52	0.52	1438.81	2.77	0.26	1400	0	0.00
2018-10-05	1414.55	0.52	0.51	1442.61	0.26	0.26	1400	0	0.00
2018-10-12	1421.82	0.51	0.51	1446.31	0.26	0.25	1400	0	0.00
2018-10-19	1429.09	0.51	0.51	1449.92	0.25	0.24	1400	0	0.00
2018-10-26	1436.36	0.51	0.51	1453.42	0.24	0.23	1400	0	0.00
2018-11-02	1443.64	0.51	0.50	1456.83	0.23	0.23	1400	0	0.00
2018-11-09	1450.91	0.50	0.50	1460.14	0.23	0.22	1400	0	0.00
2018-11-16	1458.18	0.50	0.50	1463.34	0.22	0.21	1400	0	0.00
2018-11-23	1465.45	0.50	0.50	1466.45	0.21	0.21	1400	0	0.00
2018-11-30	1472.73	0.50	0.49	1469.46	0.21	0.72	1400	0	5.71

2019-02-01	1680.00	-2.33	-2.38	1677.94	-2.45	-2.26	1720	0	-4.65
2019-11-22	1335.00	0.38	0.37	1337.35	0.55	0.20	1330	0	0.75
2020-02-28	1300.00	-1.52	-1.54	1301.83	-1.38	-1.21	1320	0	0.00
2020-03-06	1280.00	-1.54	-1.56	1286.09	-1.21	-1.23	1320	0	0.00
2020-03-13	1260.00	-1.56	-1.59	1270.30	-1.23	-2.39	1320	0	-6.06
2020-06-12	1075.67	-0.40	-0.40	1077.58	-0.22	-0.58	1080	0	0.00
2020-06-19	1071.33	-0.40	-0.40	1071.30	-0.58	-0.40	1080	0	-1.20
2020-08-21	959.00	-3.23	-3.34	959.54	-3.17	-3.39	991	0	-6.46
2020-10-23	846.50	-0.76	-0.77	850.95	-0.24	-1.29	853	0	-1.52
2021-01-15	855.00	1.79	1.75	856.63	1.98	1.69	840	0	0.00
2021-01-22	870.00	1.75	1.72	871.10	1.69	1.67	840	0	0.00
2021-01-29	885.00	1.72	1.69	885.60	1.67	1.63	840	0	7.14

Tabla 11. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios de la Yuca Chirosa en el mercado de Buenaventura. Cont.

Fecha	MA			MEAN			Random		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-05-26	1126.03	-6.16	-3.20	1113.96	-7.17	-2.15	1173.14	-2.24	-7.09
2017-09-22	802.36	0.30	0.29	1113.96	39.25	0.00	621.03	-22.37	85.84
2017-09-29	804.73	0.29	1.03	1113.96	0.00	-27.02	1154.14	85.84	-29.56
2018-08-31	1438.19	2.73	5.11	1113.96	-20.43	0.00	1322.56	-5.53	10.74
2018-09-07	1511.74	5.11	11.59	1113.96	0.00	51.44	1464.55	10.74	15.19
2018-09-28	1495.67	6.83	0.00	1113.96	-20.43	0.00	1488.81	6.34	-48.17
2018-10-05	1495.67	0.00	0.00	1113.96	0.00	0.00	771.64	-48.17	26.96
2018-10-12	1495.67	0.00	0.00	1113.96	0.00	0.00	979.71	26.96	58.72
2018-10-19	1495.67	0.00	-0.26	1113.96	0.00	0.00	1554.99	58.72	-37.87
2018-10-26	1491.75	-0.26	-3.13	1113.96	0.00	0.00	966.17	-37.87	77.28
2018-11-02	1445.00	-3.13	1.04	1113.96	0.00	0.00	1712.86	77.28	-2.51
2018-11-09	1460.00	1.04	0.00	1113.96	0.00	0.00	1669.88	-2.51	3.40
2018-11-16	1460.00	0.00	-0.32	1113.96	0.00	0.00	1726.74	3.40	-62.06
2018-11-23	1455.29	-0.32	0.54	1113.96	0.00	0.00	655.08	-62.06	124.98
2018-11-30	1463.13	0.54	1.15	1113.96	0.00	32.86	1473.77	124.98	0.42
2019-02-01	1659.50	-3.52	-1.18	1113.96	-35.23	47.22	883.66	-48.62	85.59
2019-11-22	1343.37	1.01	-0.25	1113.96	-16.24	20.29	1665.44	25.22	-19.54
2020-02-28	1311.50	-0.64	-1.09	1113.96	-15.61	0.00	1597.20	21.00	-5.90
2020-03-06	1297.14	-1.09	-0.75	1113.96	0.00	0.00	1502.94	-5.90	-42.88
2020-03-13	1287.39	-0.75	-3.68	1113.96	0.00	11.31	858.51	-42.88	44.44
2020-06-12	1090.00	0.93	-2.19	1113.96	3.14	0.00	1205.23	11.60	6.18
2020-06-19	1066.18	-2.19	0.08	1113.96	0.00	-4.22	1279.67	6.18	-16.62
2020-08-21	960.93	-3.03	-3.53	1113.96	12.41	-16.78	789.82	-20.30	17.37
2020-10-23	865.73	1.49	-2.97	1113.96	30.59	-24.59	604.95	-29.08	38.85
2021-01-15	852.61	1.50	2.43	1113.96	32.61	0.00	1402.75	66.99	-51.32
2021-01-22	873.36	2.43	2.54	1113.96	0.00	0.00	682.86	-51.32	146.64
2021-01-29	895.56	2.54	0.50	1113.96	0.00	-19.21	1684.24	146.64	-46.56

Tabla 11. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios de la Yuca Chirosa en el mercado de Buenaventura. Cont.

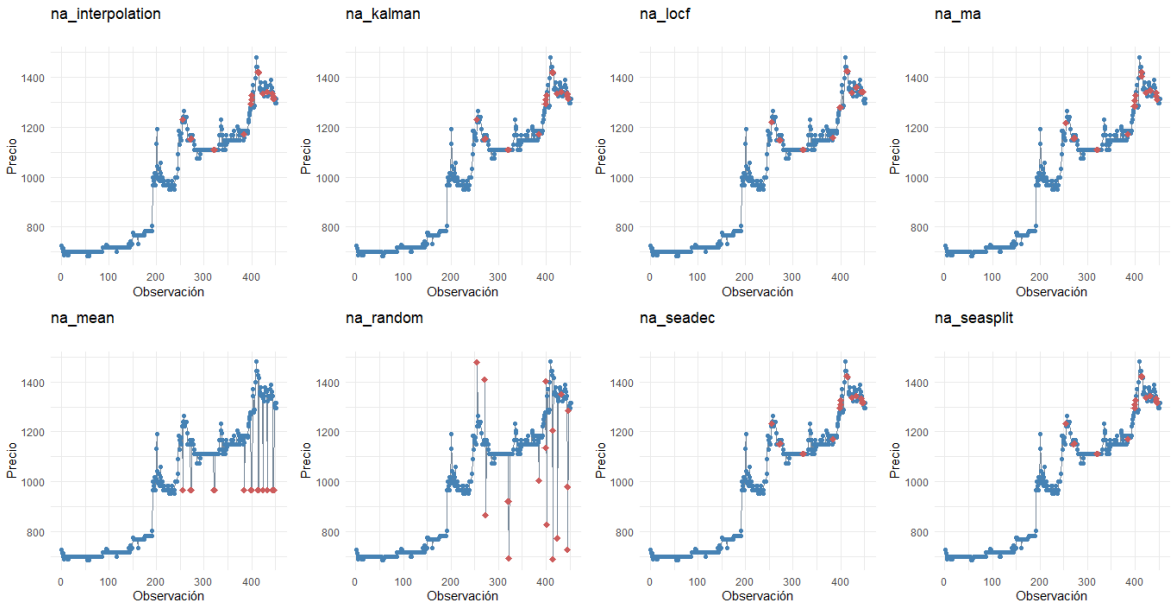
Fecha	Seadec			Seasplit		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-05-26	1145.00	-4.58	-4.80	1145.00	-4.58	-4.80
2017-09-22	804.33	0.54	0.54	804.33	0.54	0.54
2017-09-29	808.67	0.54	0.54	808.67	0.54	0.54
2018-08-31	1495.67	6.83	6.40	1495.67	6.83	6.40
2018-09-07	1591.33	6.40	6.01	1591.33	6.40	6.01
2018-09-28	1407.27	0.52	0.52	1407.27	0.52	0.52
2018-10-05	1414.55	0.52	0.51	1414.55	0.52	0.51
2018-10-12	1421.82	0.51	0.51	1421.82	0.51	0.51
2018-10-19	1429.09	0.51	0.51	1429.09	0.51	0.51
2018-10-26	1436.36	0.51	0.51	1436.36	0.51	0.51
2018-11-02	1443.64	0.51	0.50	1443.64	0.51	0.50
2018-11-09	1450.91	0.50	0.50	1450.91	0.50	0.50
2018-11-16	1458.18	0.50	0.50	1458.18	0.50	0.50
2018-11-23	1465.45	0.50	0.50	1465.45	0.50	0.50
2018-11-30	1472.73	0.50	0.49	1472.73	0.50	0.49
2019-02-01	1680.00	-2.33	-2.38	1680.00	-2.33	-2.38
2019-11-22	1335.00	0.38	0.37	1335.00	0.38	0.37
2020-02-28	1300.00	-1.52	-1.54	1300.00	-1.52	-1.54
2020-03-06	1280.00	-1.54	-1.56	1280.00	-1.54	-1.56
2020-03-13	1260.00	-1.56	-1.59	1260.00	-1.56	-1.59
2020-06-12	1075.67	-0.40	-0.40	1075.67	-0.40	-0.40
2020-06-19	1071.33	-0.40	-0.40	1071.33	-0.40	-0.40
2020-08-21	959.00	-3.23	-3.34	959.00	-3.23	-3.34
2020-10-23	846.50	-0.76	-0.77	846.50	-0.76	-0.77
2021-01-15	855.00	1.79	1.75	855.00	1.79	1.75
2021-01-22	870.00	1.75	1.72	870.00	1.75	1.72
2021-01-29	885.00	1.72	1.69	885.00	1.72	1.69

Fuente: Elaboración propia.

Al igual que con las imputaciones de los precios de la Piña Gold, el método LOCF sigue siendo el que genera la menor variación porcentual para el cultivo de la Yuca Chirosa.

El proceso se repite finalmente con el Banano Criollo.

Ilustración 10. Imputación de datos en la serie de precios del Banano Criollo para el mercado de Buenaventura.



Fuente: Elaboración propia.

Tabla 12. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios del Banano Criollo en el mercado de Buenaventura.

Fecha	Interpolation			Kalman			LOCF		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-05-26	1231.50	0.78	0.77	1229.71	0.63	0.92	1222	0	1.55
2017-09-22	1151.00	0.26	0.26	1152.10	0.36	0.23	1148	0	0.00
2017-09-29	1154.00	0.26	0.26	1154.75	0.23	0.19	1148	0	0.78
2018-08-31	1111.00	0.00	0.00	1111.00	0.00	0.00	1111	0	0.00
2018-09-07	1111.00	0.00	0.00	1111.00	0.00	0.00	1111	0	0.00
2019-11-22	1171.00	1.21	1.20	1171.31	1.24	1.17	1157	0	2.42
2020-02-28	1294.25	1.27	1.26	1294.72	1.31	1.25	1278	0	0.00
2020-03-06	1310.50	1.26	1.24	1310.90	1.25	1.23	1278	0	0.00
2020-03-13	1326.75	1.24	1.22	1327.08	1.23	1.20	1278	0	5.09
2020-06-12	1423.00	-0.21	-0.21	1422.22	-0.26	-0.33	1426	0	0.00
2020-06-19	1420.00	-0.21	-0.21	1417.53	-0.33	-0.04	1426	0	-0.63
2020-08-21	1336.50	-0.26	-0.26	1336.64	-0.25	-0.27	1340	0	-0.52
2020-10-23	1342.50	-1.36	-1.38	1343.17	-1.31	-1.43	1361	0	-2.72
2021-01-15	1333.75	-0.69	-0.69	1333.31	-0.72	-0.68	1343	0	0.00
2021-01-22	1324.50	-0.69	-0.70	1324.23	-0.68	-0.69	1343	0	0.00
2021-01-29	1315.25	-0.70	-0.70	1315.15	-0.69	-0.70	1343	0	-2.76

Tabla 12. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios del Banano Criollo en el mercado de Buenaventura. Cont.

Fecha	MA			Mean			Random		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-05-26	1217.70	-0.35	1.91	964.17	-21.10	28.71	1478.36	20.98	-16.06
2017-09-22	1154.82	0.59	0.14	964.17	-16.01	0.00	1408.45	22.69	-38.62
2017-09-29	1156.45	0.14	0.05	964.17	0.00	20.00	864.44	-38.62	33.84
2018-08-31	1111.00	0.00	0.00	964.17	-13.22	0.00	920.36	-17.16	-24.95
2018-09-07	1111.00	0.00	0.00	964.17	0.00	15.23	690.76	-24.95	60.84
2019-11-22	1172.07	1.30	1.10	964.17	-16.67	22.90	1003.24	-13.29	18.12
2020-02-28	1284.67	0.52	1.79	964.17	-24.56	0.00	1136.50	-11.07	23.33
2020-03-06	1307.71	1.79	1.54	964.17	0.00	0.00	1401.62	23.33	-41.12
2020-03-13	1327.83	1.54	1.14	964.17	0.00	39.29	825.31	-41.12	62.73
2020-06-12	1420.41	-0.39	-1.17	964.17	-32.39	0.00	686.87	-51.83	75.19
2020-06-19	1403.73	-1.17	0.95	964.17	0.00	46.97	1203.34	75.19	17.76
2020-08-21	1339.73	-0.02	-0.50	964.17	-28.05	38.25	773.05	-42.31	72.43
2020-10-23	1347.23	-1.01	-1.72	964.17	-29.16	37.32	1350.59	-0.76	-1.97
2021-01-15	1337.72	-0.39	-0.95	964.17	-28.21	0.00	977.79	-27.19	-25.82
2021-01-22	1325.00	-0.95	-1.10	964.17	0.00	0.00	725.36	-25.82	77.22
2021-01-29	1310.39	-1.10	-0.33	964.17	0.00	35.45	1285.48	77.22	1.60

Tabla 12. Variación porcentual entre los datos imputados y el dato inmediatamente anterior y siguiente en la serie de precios del Banano Criollo en el mercado de Buenaventura. Cont.

Fecha	Seadec			Seasplit		
	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$	$\widehat{P}_t$	$P_{(t-1)}\%$	$P_{(t+1)}\%$
2017-05-26	1231.50	0.78	0.77	1231.50	0.78	0.77
2017-09-22	1151.00	0.26	0.26	1151.00	0.26	0.26
2017-09-29	1154.00	0.26	0.26	1154.00	0.26	0.26
2018-08-31	1111.00	0.00	0.00	1111.00	0.00	0.00
2018-09-07	1111.00	0.00	0.00	1111.00	0.00	0.00
2019-11-22	1171.00	1.21	1.20	1171.00	1.21	1.20
2020-02-28	1294.25	1.27	1.26	1294.25	1.27	1.26
2020-03-06	1310.50	1.26	1.24	1310.50	1.26	1.24
2020-03-13	1326.75	1.24	1.22	1326.75	1.24	1.22
2020-06-12	1423.00	-0.21	-0.21	1423.00	-0.21	-0.21
2020-06-19	1420.00	-0.21	-0.21	1420.00	-0.21	-0.21
2020-08-21	1336.50	-0.26	-0.26	1336.50	-0.26	-0.26
2020-10-23	1342.50	-1.36	-1.38	1342.50	-1.36	-1.38
2021-01-15	1333.75	-0.69	-0.69	1333.75	-0.69	-0.69
2021-01-22	1324.50	-0.69	-0.70	1324.50	-0.69	-0.70
2021-01-29	1315.25	-0.70	-0.70	1315.25	-0.70	-0.70

Fuente: Elaboración propia.

Nuevamente, para este cultivo, el método LOCF es el que presenta menor variación porcentual. Luego, es el mejor método de imputación para los precios del Banano Criollo según las series dadas.

En las tablas 13, 14 y 15, se muestran las relaciones de fecha, producto y tipo de imputación usada para completar los datos de los precios de la Piña Gold, Yuca Chirosa y Banano Criollo respectivamente. De esta manera, se procede posteriormente a la construcción de los modelos.

Tabla 13. Método de imputación para los precios de la Piña Gold en las fechas donde no hubo reporte para el mercado de Buenaventura.

Fecha	Tipo de imputación
2017-07-14	ImputeTS - locf
2017-09-22	ImputeTS – locf
2017-09-29	ImputeTS – locf
2018-08-31	ImputeTS – locf
2018-09-07	ImputeTS – locf
2019-11-22	ImputeTS – locf
2020-02-28	ImputeTS – locf
2020-03-06	ImputeTS – locf
2020-03-13	ImputeTS – locf
2020-06-12	ImputeTS – locf
2020-06-19	ImputeTS – locf
2020-08-21	ImputeTS – locf
2020-10-23	ImputeTS – locf
2021-01-15	ImputeTS – locf
2021-01-22	ImputeTS – locf
2021-01-29	ImputeTS - locf

Nota: Tipos de imputación: Método de plaza cerca, método de plaza o lejana algoritmos imputeTs.
Fuente: Elaboración propia.

Tabla 14. Método de imputación para los precios de la Yuca Chirosa en las fechas donde no hubo reporte para el mercado de Buenaventura.

Fecha	Tipo de imputación
2017-05-26	ImputeTS - locf
2017-09-22	ImputeTS – locf
2017-09-29	ImputeTS – locf
2018-08-31	ImputeTS – locf
2018-09-07	ImputeTS – locf
2018-09-28	ImputeTS – locf
2018-10-05	ImputeTS – locf
2018-10-12	ImputeTS – locf
2018-10-19	ImputeTS – locf
2018-10-26	ImputeTS – locf
2018-11-02	ImputeTS – locf
2018-11-09	ImputeTS - locf
2018-11-16	ImputeTS – locf
2018-11-23	ImputeTS – locf
2018-11-30	ImputeTS – locf
2019-02-01	ImputeTS – locf
2019-11-22	ImputeTS – locf
2020-02-28	ImputeTS – locf
2020-03-06	ImputeTS – locf
2020-03-13	ImputeTS – locf
2020-06-12	ImputeTS - locf
2020-06-19	ImputeTS – locf
2020-08-21	ImputeTS – locf
2020-10-23	ImputeTS – locf
2021-01-15	ImputeTS – locf

2021-01-22	ImputeTS – locf
2021-01-29	ImputeTS – locf

Nota: Tipos de imputación: Método de plaza cerca, método de plaza o lejana algoritmos imputeTs.
Fuente: Elaboración propia.

Tabla 15. Método de imputación para los precios del Banano Criollo en las fechas donde no hubo reporte para el mercado de Buenaventura.

Fecha	Tipo de imputación
2017-05-26	ImputeTS – locf
2017-09-22	ImputeTS – locf
2017-09-29	ImputeTS – locf
2018-08-31	ImputeTS – locf
2018-09-07	ImputeTS – locf
2019-11-22	ImputeTS – locf
2020-02-28	ImputeTS – locf
2020-03-06	ImputeTS – locf
2020-03-13	ImputeTS - locf
2020-06-12	ImputeTS – locf
2020-06-19	ImputeTS – locf
2020-08-21	ImputeTS – locf
2020-10-23	ImputeTS – locf
2021-01-15	ImputeTS – locf
2021-01-22	ImputeTS – locf
2021-01-29	ImputeTS – locf

Nota: Tipos de imputación: Método de plaza cerca, método de plaza o lejana algoritmos imputeTs.
Fuente: Elaboración propia.

4.4. Construcción de los modelos estadísticos para proyecciones

Una vez se consiguió tener las series de tiempo de los precios de los cultivos completas. Se inicia el proceso de creación de modelos. Para ello, se utilizaron las familias de métodos de promedio móvil, suavizamiento exponencial simple, lineal, modelos ARIMA y ensambles.

La métrica utilizada para evaluar el rendimiento de los modelos creados fue el RMSE, pero la siguiente tabla muestra además las métricas ME, MAE, MPE, MAPE, en caso tal que la métrica RMSE no sea concluyente. También, se utilizó una ventana recursiva con pronósticos de un paso adelante.

La muestra utilizada para evaluar el rendimiento de cada uno de los modelos para cada uno de los cultivos, corresponde a los últimos 20 registros de cada serie de tiempo (h = 20). Los detalles de los rangos de la muestra de estimación y evaluación se muestran en la Tabla 16.

Tabla 16. Rangos de fechas de los periodos de entrenamiento y evaluación para la estimación de los modelos para las series de la Piña Gold, Yuca Chirosa y el Banano Criollo

Cultivo	Muestras	
	Estimación	Evaluación
Piña Gold	14-7-2017 al 16-10-2020	23-10-2020 al 5-3-2021
Yuca Chirosa	22-6-2012 al 16-10-2020	23-10-2020 al 5-3-2021
Banano Criollo	13-7-2012 al 16-10-2020	23-10-2020 al 5-3-2021

Fuente: Elaboración propia.

Tabla 17. Rendimiento de los modelos según las métricas de evaluación ME, RMSE, MAE, MPE, MAPE con la serie de tiempo de la Piña Gold disponible para el mercado de Buenaventura.

Modelos	Métricas				
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Media móvil 3	-64.6247	132.1579	96.3516	-3.3741	5.0043
Media móvil 4	-68.2098	131.0488	96.6670	-3.5724	5.0299
Media móvil 5	-70.7045	130.5737	95.5489	-3.7051	4.9858
Media móvil 6	-82.9511	130.9943	96.1014	-4.3124	4.9990
Media móvil 7	-77.0512	122.4399	87.0721	-4.0269	4.5464
Media móvil 8	-127.9349	208.3892	143.4974	-6.7196	7.4299
Media móvil 9	-108.7350	178.6046	126.1890	-5.7101	6.5108
Suavización exponencial simple	-18.6689	78.2884	55.2519	-0.9992	2.8483
Suavización exponencial lineal	-23.4496	79.5950	56.2664	-1.2429	2.9011
Arima (0,1,1)	-18.7153	78.3165	55.6427	-1.0003	2.8679
Promedio simple (comb_SA)	2.3672	90.6124	57.3724	-0.0437	3.2835
Mediana (comb_MED)	2.4486	101.9214	61.1602	-0.0426	3.4728
Media Truncada (comb_TA)	2.3672	90.6124	57.3724	-0.0437	3.2835
Media Winzor (comb_WA)	2.3672	90.6124	57.3724	-0.0437	3.2835
Método de Bates y Granger (comb_BG)	2.3551	88.4967	56.7483	-0.0423	3.2545
Método de Newbold y Granger (comb_NG)	-6.8919	79.7148	56.0162	-0.5750	3.2614
Ranking inverso (comb_ invw)	2.1225	82.8098	55.1263	-0.0516	3.1808
Mínimos cuadrados ordinarios (comb_OLS)	-1.8230	76.2081	53.2182	-0.2914	3.0827
Mínima desviación absoluta (comb_LAD)	-1.9136	76.9093	52.9656	-0.2484	3.0714
Estándar (comb_EIG1)	2.4083	93.5243	58.2828	-0.0432	3.3278
Con corrección de sesgo (comb_EIG2)	-0.7833	93.6687	58.2170	-0.2256	3.3299
Truncada (comb_EIG3)	2.3124	80.7825	56.3654	-0.0390	3.2644
Truncada y con corrección de sesgo (comb_EIG4)	-1.2599	80.9027	56.2427	-0.2463	3.2609

Fuente: Elaboración propia.

Para el precio de la Piña Gold y en este marco temporal, el mejor modelo para pronosticar es el de ensamble por el método de mínimos cuadrados ordinarios.

Tabla 18. Rendimiento de los modelos según las métricas de evaluación ME, RMSE, MAE, MPE, MAPE con la serie de tiempo de la Yuca Chirosa disponible para el mercado de Buenaventura.

Modelos	Métricas				
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Media móvil 3	9.9872	53.9146	46.3917	0.9572	5.2698
Media móvil 4	3.9299	65.5784	56.2401	0.2382	6.4105
Media móvil 5	3.4914	64.3177	54.5657	0.1855	6.2179
Media móvil 6	-3.2328	68.0744	58.7230	-0.5994	6.7015
Media móvil 7	-4.1949	65.4351	56.1370	-0.7042	6.4105
Media móvil 8	-11.8104	64.0009	53.9835	-1.5689	6.1808
Media móvil 9	-12.9242	60.8062	51.4154	-1.6869	5.8927
Suavización exponencial simple	3.9317	21.5259	16.5226	0.4070	1.8632
Suavización exponencial lineal	8.2614	23.5134	17.6103	0.9073	1.9810
Promedio simple (comb_SA)	0.2527	44.7926	25.9093	-0.1096	2.2540
Mediana (comb_MED)	0.2916	51.6998	29.2387	-0.1135	2.5457
Media truncada (comb_TA)	0.2527	44.7926	25.9093	-0.1096	2.2540
Media Winzor (comb_WA)	0.2527	44.7926	25.9093	-0.1096	2.2540
Método de Bates y Granger (comb_BG)	0.2261	35.8737	21.1184	-0.0848	1.8272
Método de Newbold y Granger (comb_NG)	-1.5591	25.4321	15.3798	-0.1663	1.3385
Ranking inverso (comb_ invw)	0.2744	38.0815	22.3421	-0.0961	1.9377
Mínimos cuadrados ordinarios (comb_OLS)	0.0431	25.3291	15.1270	-0.0421	1.3213
Mínima desviación absoluta (comb_LAD)	0.1017	25.4733	14.8322	-0.0317	1.3013
Estándar (comb_EIG1)	0.3968	29.6048	17.3872	0.0186	1.5592
Con corrección de sesgo (comb_EIG2)	0.0647	29.6266	17.3521	-0.0253	1.5528
Truncada (comb_EIG3)	0.2276	27.0782	16.0987	-0.0462	1.3895
Truncada y con corrección de sesgo (comb_EIG4)	0.0120	27.0969	16.1756	-0.0676	1.3979

Fuente: Elaboración propia.

Para el precio de la Yuca Chirosa y en este marco temporal, se obtiene que el mejor modelo para realizar pronósticos es el de suavización exponencial simple.

Tabla 19. Rendimiento de los modelos según las métricas de evaluación ME, RMSE, MAE, MPE, MAPE con la serie de tiempo del Banano Criollo disponible para el mercado de Buenaventura.

Modelos	Métricas				
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Media móvil 3	-23.2351	32.0889	29.4894	-1.7604	2.2136
Media móvil 4	-24.0126	41.3052	35.1111	-1.8347	2.6419
Media móvil 5	-23.6509	40.3400	35.0332	-1.8068	2.6354
Media móvil 6	-29.4620	49.1305	41.1381	-2.2504	3.0978
Media móvil 7	-27.5465	47.4742	38.6083	-2.1064	2.9095
Media móvil 8	-31.7688	53.9046	43.6973	-2.4268	3.2927
Media móvil 9	-30.5088	51.4784	41.6543	-2.3303	3.1392
Suavización exponencial simple	-2.4673	22.4045	16.2268	-0.1991	1.2164
Suavización exponencial lineal	-3.9659	22.5887	16.7321	-0.3111	1.2548
Promedio simple (comb_SA)	0.8328	18.6167	9.4650	0.0609	0.8809
Mediana (comb_MED)	1.1689	21.3129	10.1181	0.0933	0.9331
Media truncada (comb_TA)	0.8328	18.6167	9.4650	0.0609	0.8809
Media Winzor (comb_WA)	0.8328	18.6167	9.4650	0.0609	0.8809
Método de Bates y Granger (comb_BG)	0.4665	14.5823	7.7224	0.0283	0.7202
Método de Newbold y Granger (comb_NG)	0.1472	10.3371	5.6797	0.0086	0.5373
Ranking inverso (comb_ invw)	0.6226	15.6428	8.3235	0.0428	0.7781
Mínimos cuadrados ordinarios (comb_OLS)	-0.0502	10.3388	5.6367	-0.0144	0.5315
Mínima desviación absoluta (comb_LAD)	-0.5574	10.5384	5.2857	-0.0670	0.4924
Estándar (comb_EIG1)	-0.2860	12.2002	6.2189	-0.0323	0.5895
Con corrección de sesgo (comb_EIG2)	-0.0841	12.2282	6.2912	-0.0126	0.5971
Truncada (comb_EIG3)	-0.0389	10.9391	5.8132	-0.0170	0.5421
Truncada y con corrección de sesgo (comb_EIG4)	-0.0174	10.9466	5.8283	-0.0146	0.5438

Fuente: Elaboración propia.

Para el precio del Banano Criollo y en este marco temporal, se obtuvo que el mejor modelo de generación de pronósticos es el de ensamble por el método de Newbold y Granger.

4.5. Automatización de la captura de los datos del Sistema de Información de Precios del Sector Agropecuario

Dado que los datos del Sistema de Información de Precios del Sector Agropecuario se actualizan con una frecuencia diaria, semanal y mensual, se automatizó la captura de éstos para tener un producto que requiriera la menor intervención humana para su funcionamiento.

Los datos se capturan automáticamente mediante una petición GET a la API *selectAllInfoProduct* proveída por el DANE para actualizar los gráficos interactivos que dispone en su portal web. Para tal fin, se utilizó el paquete httr[16] que provee las funciones HTTP para realizar peticiones a servicios API REST.

Como cuerpo de la petición, se envían los datos de la Tabla 20.

Tabla 20. Descripción de los parámetros utilizados para consumir los datos expuestos en la API de SIPSA

Parámetro	Descripción	Valor
depcod	Código del departamento	76
codmun	Código del municipio	76109
codart	Código del artículo	Piña Gold: 144 Yuca Chirosa: 191 Banano Criollo: 93
archivoCsv	Concatenación entre depcod, codmun y codart	76109+codart
fechaIni	Fecha de inicio de los datos a consultar	20120101
fechaFin	Fecha final de los datos a consultar	20301231
tipoReporte	Frecuencia de los datos (day, week o year)	week

Fuente: Elaboración propia.

Los valores asignados a los parámetros como cuerpo de la petición GET, garantizan la consulta de todos los datos semanales de cada uno de los cultivos estudiados en este documento.

[16] Para más detalles, consultar: Tools for Working with URLs and HTTP. https://cran.r-project.org/web/packages/httr/index.html

4.6. Construcción del Dashboard

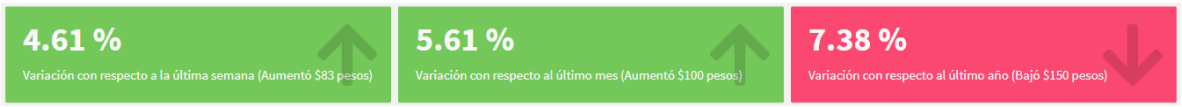
Dada la importancia de contar con una herramienta visual que facilite la interpretación de las proyecciones, se desarrolló una interfaz gráfica usando el paquete Flex-Dashboard[17]. La interfaz gráfica permitió disponer los pronósticos para los tres productos analizados de manera organizada. Y a la vez, mostrar un resumen sobre el comportamiento de los precios para cada cultivo con respecto a la semana, mes y año anterior. Por último, se describe el comportamiento de cada uno de los tres modelos elegidos con respecto al dato real de la semana inmediatamente anterior. Estos elementos gráficos se utilizaron para permitirle al productor agrícola la fácil interpretación y análisis de los resultados.

Los elementos gráficos utilizados para representar la información son los siguientes:

Tarjetas

La Ilustración 11 es una tarjeta y, en su interior contiene la variación porcentual del precio de cada uno de los cultivos estudiados en este documento, con respecto a la semana, mes y año anterior. Para este elemento gráfico, se consideró importante utilizar colores que representaran situaciones positivas (verde) o negativas (rojo) según el comportamiento del precio para las unidades productivas del Distrito Especial de Buenaventura.

Ilustración 11. Tarjetas que facilitan comprender el comportamiento del precio de los cultivos con respecto a la semana, mes y año anterior.

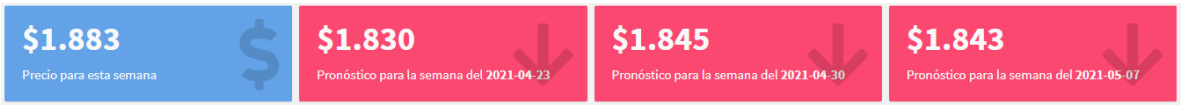


Fuente: Elaboración propia.

Al igual que en la Ilustración 11, en la Ilustración 12 se utilizaron colores que representaran situaciones positivas y negativas según los pronósticos de las próximas semanas.

[17] Para más detalles consultar: R Markdown Format for Flexible Dashboards. https://cran.r-project.org/web/packages/flexdashboard/index.html

Ilustración 12. Tarjetas que facilitan la interpretación del precio actual y los pronósticos para las próximas tres semanas de cada uno de los cultivos. (Pesos colombianos)

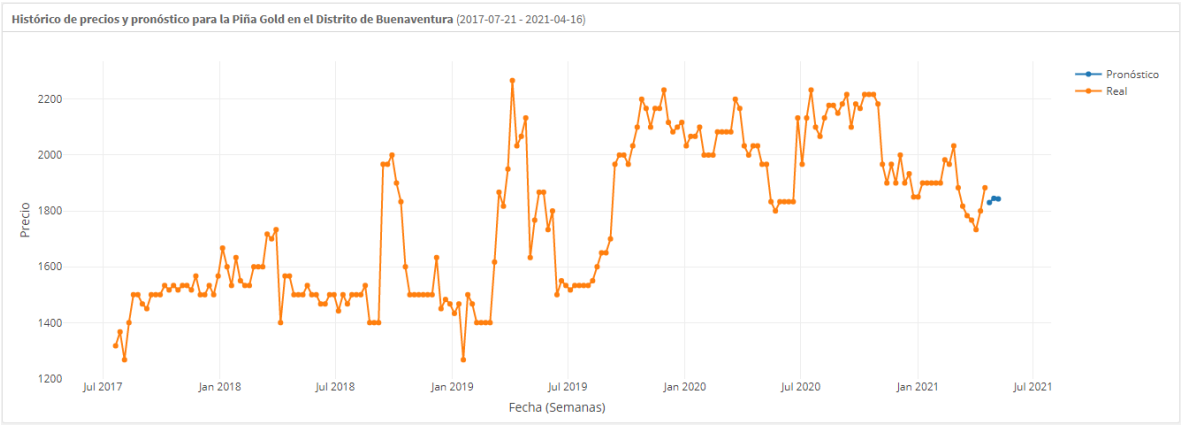


Fuente: Elaboración propia.

Gráficos

El gráfico de la Ilustración 13 se creó utilizando el paquete Plotly[18], que facilita la creación de visualizaciones de datos atractivas e interactivas que permiten la exploración de los mismos. Se eligió un gráfico de líneas que discriminara la serie histórica y el pronóstico para las próximas tres semanas. La serie histórica se representa con el color naranja y el pronóstico con el color azul.

Ilustración 13. Gráfico de líneas para representar los precios históricos de las series de tiempo de los cultivos y su pronóstico.



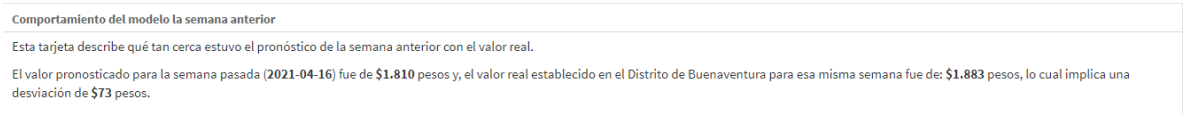
Fuente: Elaboración propia.

[18] Para más información, consultar: Create Interactive Web Graphics via 'plotly.js'. https://cran.r-project.org/web/packages/plotly/index.html

Cuadros de texto

Se incluyó un espacio para consultar el rendimiento del modelo con respecto al pronóstico realizado para la semana actual. Esto se logró almacenando los pronósticos de forma persistente y, junto a la captura automática de los datos (sección 4.5), se comparan y se calcula la diferencia entre el valor real y el pronosticado.

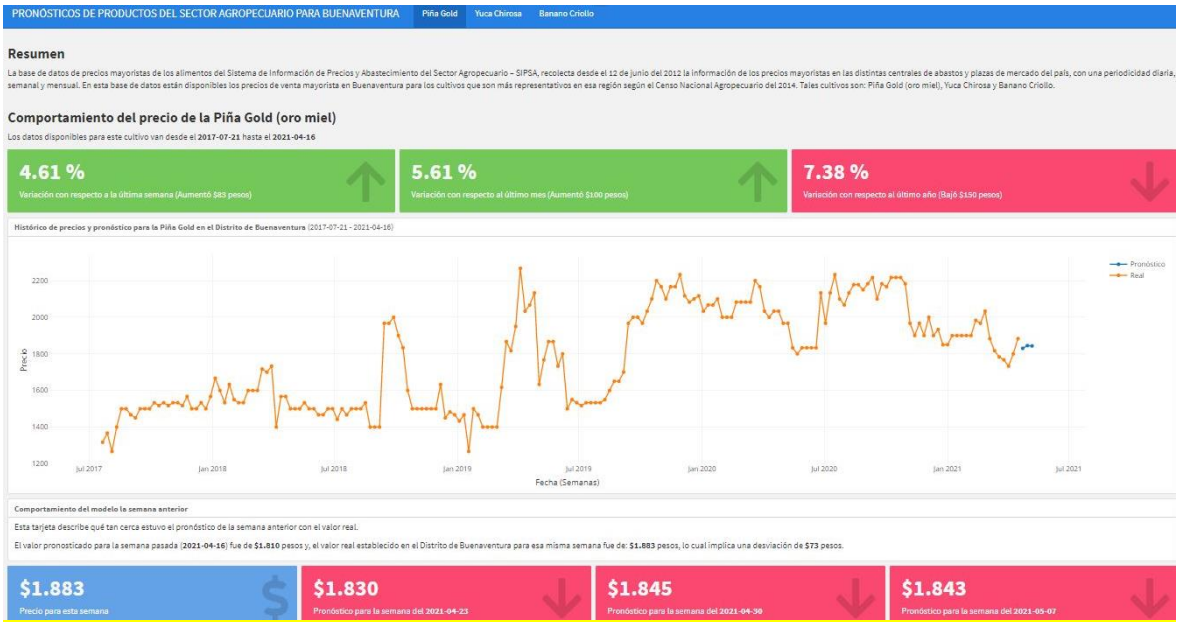
Ilustración 14. Cuadro de texto que describe el rendimiento del modelo



Fuente: Elaboración propia.

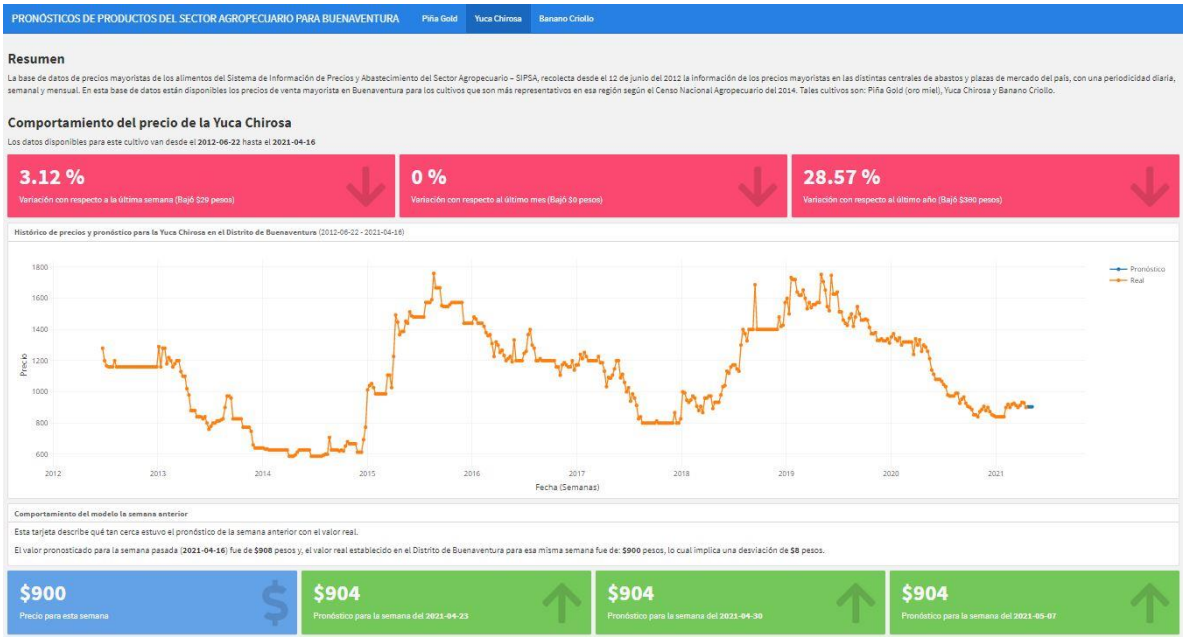
En las ilustraciones 15, 16 y 17, se presenta el Dashboard completo para la Piña Gold, Yuca Chirosa y Banano Criollo respectivamente. Se puede apreciar que, en la parte superior, hay un menú de color azul que permite navegar por los diferentes productos.

Ilustración 15. Dashboard completo para la serie de precios de la Piña Gold.



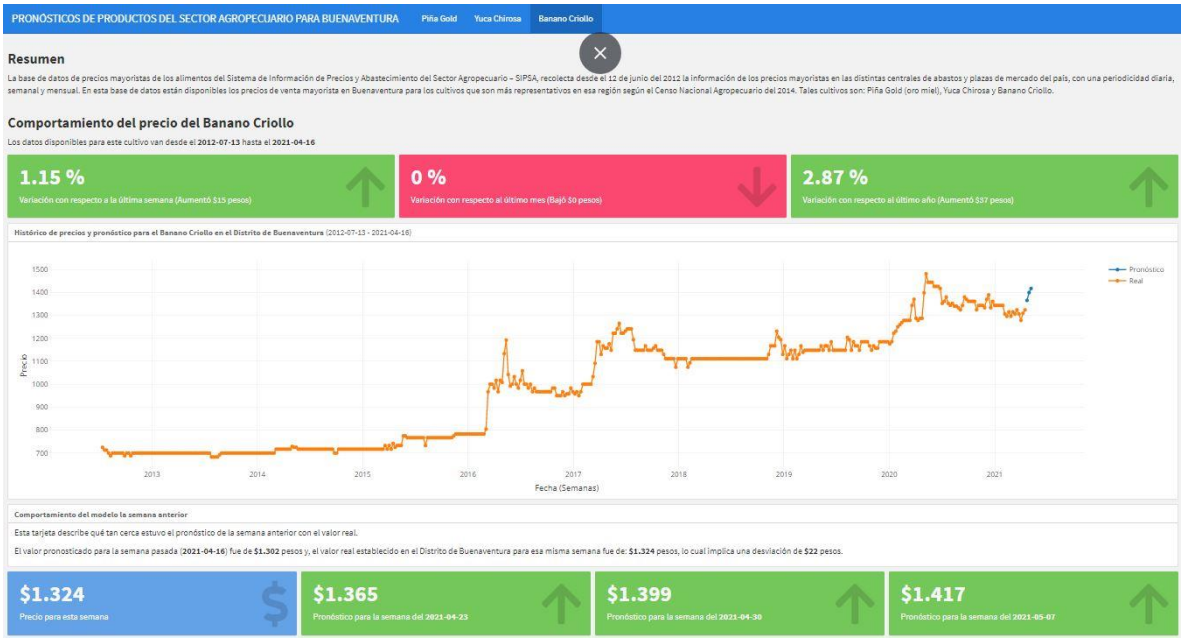
Fuente: Elaboración propia.

Ilustración 16. Dashboard completo para la serie de precios de la Yuca Chirosa.



Fuente: Elaboración Propia.

Ilustración 17. Dashboard completo para la serie de precios del Banano Criollo.



Fuente: Elaboración propia.

5. CONCLUSIONES

Basados en los datos del tercer Censo Nacional Agropecuario (CNA) llevado a cabo por el DANE en el año 2014, se recogió información del campo colombiano. De la base de datos de cultivos y hogares para el Distrito Especial de Buenaventura, podemos destacar que de la totalidad de los hogares campesinos censados (11846), el 59% (6949) se benefició directamente de la siembra y la cosecha. Por lo tanto, este trabajo, potencialmente beneficiará a 6949 familias campesinas del distrito de Buenaventura.

También, se obtuvo información de los cultivos más sembrados y cosechados. En algunos casos críticos como el Cedro, se encontró que se cosechaba muy poco en comparación con l sembrado. Pero, en otros casos como el chontaduro, se encontró que se cosechó todo lo que se sembró, pero el Sistema de Información de Precios y Abastecimiento del Sector Agropecuario - SIPSA, ente adscrito al DANE para la captura de precios en los diferentes mercados del país, no lleva registro de la actividad comercial asociada a este cultivo.

Teniendo en cuenta que uno de los objetivos específicos de este trabajo de grado era seleccionar los productos agrícolas que más se cosechan Pacífico Colombiano, específicamente en el Distrito de Buenaventura, se tuvo en cuenta dos criterios para la selección de los cultivos. El primer criterio fue que la cosecha fuera sustancialmente importante en comparación con la siembra y, además que el precio estuviera reportado por el SIPSA, pues en concordancia con otro objetivo específico, se había propuesto encontrar modelos estadísticos para proyecciones de series de tiempos de los cultivos. De esta manera crear modelos basados en técnicas de ciencias de datos, econometría y análisis cuantitativo para realizar estimación de pronósticos de precios.

Considerando lo antes descrito, los precios de los cultivos elegidos fueron: Piña Gold, Yuca Chirosa y el Banano Criollo. De donde obtuvimos que el mejor modelo de proyecciones de precios para la Piña Gold es el método de ensamble de Mínimos Cuadrados Ordinarios, para la Yuca Chirosa es Suavización Exponencial Simple y para el Banano Criollo es el ensamble por el Método de Newbold y Granger. La métrica que determinó cuál fue el mejor modelo fue el RMSE.

Finalmente, se creó un Dashboard para de esta manera tener una interfaz gráfica amigable que brinde información importante de las estimaciones como lo son: los pronósticos, el mejor modelo, el periodo de tiempo estudiado y la desviación del modelo en términos del precio real y, de automatizó la captura de los datos del Sistema de Información de Precios del Sector Agropecuario, de tal manera que se tiene un producto totalmente autónomo en la actualización y pronóstico de los precios de los cultivos estudiados en este documento.

Para trabajos futuros se puede aplicar este mecanismo desarrollado en otros productos y/o otras ciudades. Además, es expandible a otro tipo de perspectiva como estimación de cantidad de ventas por dar un ejemplo.

6. REFERENCIAS BIBLIOGRÁFICAS.

Acevedo, Raziel. (2008): Los modelos jerárquicos lineales: fundamentos básicos para su uso y aplicación. Serie cuadernos metodológicos del instituto de investigaciones psicológicas de la universidad de Costa Rica.

Alonso, J. (2020): Introducción a los pronósticos con modelos estadísticos de series de tiempo para científicos de datos (en R). Series libros de texto.

Alonso, J. C., Díaz, J. G., Estrada, D., Figueroa, C. A., & Tamura, G. (2019): Empleando modelos jerárquicos para encontrar el mejor modelo para pronosticar los galones de gasolina corriente demandados en Bogotá (Colombia). Revista De Métodos Cuantitativos Para La Economía y la Empresa, (28)113-123.

Alonso, J., & Arcila, A. (2019): Un modelo de predicciones diarias para contratos de futuros del azúcar. Economía & Región, 6(2), 33-51.

Correa, J. C. & Salazar, J. C. (2016): Introducción a modelos mixtos. Escuela estadística facultad de ciencias de la universidad nacional de Colombia, sede Medellín, Centro editorial.

Holt, C. (1957): Forecasting seasonals and trends by exponentially weighted moving averages, ONR Memorandum (Vol. 52), Pittsburgh, PA: Carnegie Institute of Technology. Available from the Engineering Library, University of Texas at Austin.

Honaker, J., King, G., & Blackwell, M. (2011): Amelia II: A program for missing data. Journal of Statistical Software, 45(7):1–47. URL https://doi.org/10.18637/jss.v045.i07.

Hyndman, R. J. (2017): forecast: Forecasting Functions for Time Series and Linear Models. URL http: //github.com/robjhyndman/forecast.

Jiménez, N., Miranda, F. & Gantiva, O. (2008): El sector de ganadería bovina en Colombia. Aplicación de modelos de series de tiempo al inventario ganadero. Revista Facultad de Ciencias Económicas, Universidad Militar Nueva Granada, XVI,(1).

Josse, J. & Husson, F. (2016): missMDA: A package for handling missing values in multivariate data analysis. Journal of Statistical Software, 70(1):1–3. URL https://doi.org/10.18637/jss.v070.i01.

Kowarik A. & Templ, M. (2016): Imputation with the R package VIM. Journal of Statistical Software, 74(7): 1–16. URL https://doi.org/10.18637/jss.v074.i07.

Moritz, S. & Bartz-Beielstein, T. (2017): ImputeTS, Time Series Missing Value Imputation in R, The R Journal, Vol. 9/1

Quispe, J. (2015): Producción agrícola alimentaria y cambio climático: un análisis económico en el departamento de Puno, Perú. Revista IDESIA (Chile). Vol. 33-2.

Ruiz, J., Hernández, G. y Zulueta, R. (2010): Análisis de series de tiempo en el pronóstico de la producción de caña de azúcar. Terra latinoamericana. (29): 103-109.

Winters, P. R. (1960): Forecasting sales by exponentially weighted moving averages. Management science, 6(3), 324-342.

Zeileis, A. & Grothendieck, G. (2005): zoo, S3 infrastructure for regular and irregular time series. Journal of Statistical Software, 14(6):1–27. URL https://doi.org/10.18637/jss.v014.i06.