Big Data para el monitoreo de precios agropecuarios

Unidad de Científicos de Datos

2017

Big Data para el monitoreo de precios agropecuarios

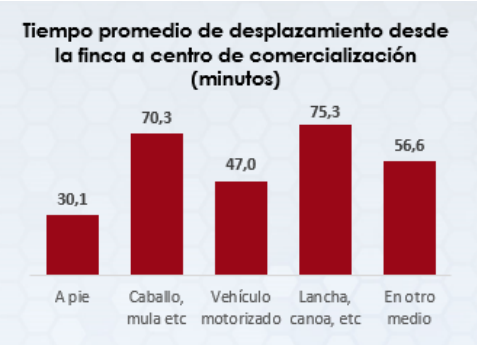
Justificación y objetivos

Justificación

De la mano de la Dirección de Desarrollo Rural Sostenible (DDRS) del DNP, se identificó que Los productores agropecuarios se enfrentan a barreras relacionadas con el desconocimiento del precio de compra del consumidor final y altos tiempos de desplazamiento a centros de comercialización, lo que genera:

- Bajo poder de negociación de los productores con los intermediarios.
- Alta vulnerabilidad a choques en dinámica del mercado.

En donde el **76,52%** de las unidades de producción con cultivos **comercian parte de su producción** según cálculos realizados con información obtenida por el Consejo Nacional Agropecuario (2014)



La DDRS-DNP y el MADR **carecen de información histórica (diaria) de fácil acceso y consumo** sobre el comportamiento de los precios de los productos agropecuarios. Razón por la cual, consultar diariamente lo publicado por el portal SIPSA del DANE y generar una base de datos histórica de la información brindaría un insumo adicional para las entidades mencionadas.

Objetivos

El objetivo general del proyecto es entonces, desarrollar una herramienta que permita monitorear el comportamiento de los precios de los productos agropecuarios comercializados en las centrales mayoristas del país, según información del DANE.

Para llevar a cabo este objetivo se establecieron 3 objetivos específicos:

- Construir una base de datos histórica, de fácil acceso y consumo que recopile la información publicada diariamente por el DANE.
- Predecir el comportamiento de los precios de los productos agrícolas comercializados en las centrales mayoristas, a partir del mejor modelo de ajuste.
- Visualizar patrones de relación entre las series precios mediante mapas de calor.

Análisis exploratorio de los datos

Características de los datos

La Fuente de información para el desarrollo de este Proyecto es el portal SIPSA del DANE, en donde se hace pública la información de los precios agropecuarios en las principales centrales mayoristas del país, así como la variación porcentual respecto al día anterior.

Las principales características de la información de este portal son:

1. El DANE **publica diariamente** un archivo Excel con información del precio por kilogramo (y su variación) de hortalizas y verduras, frutas frescas, y tubérculos y plátanos.
2. Los precios solo se publican **para días hábiles**.
3. Existen **errores de digitación** o cálculo en la información recopilada.
4. Se presenta **asimetría en la información** reportada por productos y centrales mayoristas.

En promedio la base de datos crece diariamente **417 registros**, su ubicación geográfica está distribuida de la siguiente forma:

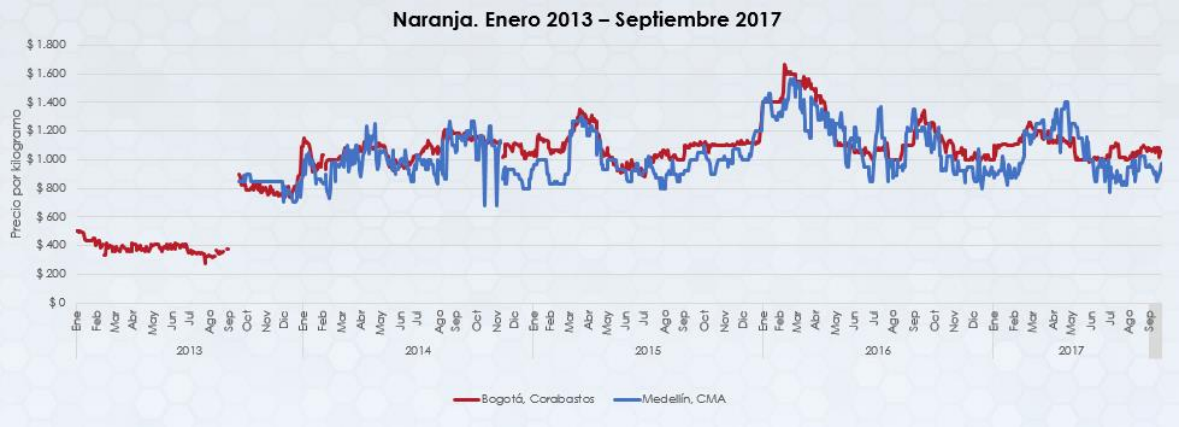
La cantidad de datos diarios al 20 de septiembre son 1.126 el top 10 de centrales con mayor número de registros son:

Ciudad y central	Número de registros
1. Bogotá, Corabastos	40.106
2. Medellín, CMA	40.052
3. Pereira, Mercasa	30.724
4. Cúcuta, Cenabastos	25.134
5. Bucaramanga, Centroabastos	25.042
6. Armenia, Mercar	23.508
7. Barranquilla, Granabastos	23.206
8. Cartagena, Bazurto	22.901
9. Tunja	21.272
10. Neiva, Surabastos	21.159

La información capturada desde el 2 de enero de 2013 hasta el 11 de septiembre de 2017 para los días hábiles solamente cuenta con **22 centrales mayoristas y 36 productos agrícolas:**

Ciudad y central	**Porcentaje de productos**		**Ciudad y central**	**Porcentaje de productos**
Medellín, CMA	97%		Pereira, La 41	81%
Armenia, Mercar	94%		Villavicencio, CAV	81%
Bogotá, Corabastos	94%		Barranquilla, Granabastos	78%
Bucaramanga, Centroabastos	94%		Cali, Santa Helena	78%
Tunja	94%		Montería	78%
Ibagué, Plaza La 21	89%		Cartagena, Bazurto	72%
Pereira, Mercasa	86%		Valledupar, Mercabastos	72%
Cúcuta, Cenabastos	83%		Neiva, Surabastos	69%
Manizales	83%		Pasto, El Potrerillo	69%
Popayán	83%		Sincelejo	58%
Cali, Cavasa	81%		Santa Marta	42%

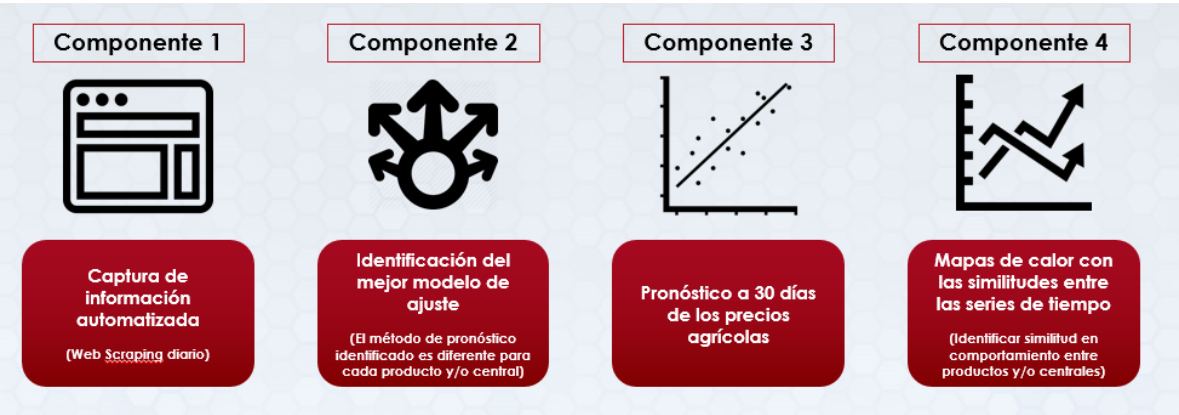
Cuando se selecciona cualquier producto y se contrastan los datos entre centrales mayoristas mediante la representación en series de tiempo se obtiene lo siguiente:



Se pueden evidenciar productos que tienen **datos faltantes, atípicos y cambios estructurales** en su comportamiento.

Componentes desarrollados en el proyecto

Dada la naturaleza de la información que se está tratando y el fin general que es crear una herramienta para monitorear dichos precios, se determinó realizar el proyecto en 4 módulos principales:



Componente 1

Para este módulo se desarrolló un algoritmo que estuviese en la capacidad de determinar cuáles fechas son faltantes en la base de datos histórica y fuese directamente a la página del DANE y bajase los archivos correspondientes a ese vector de fechas faltantes.

Una vez se encuentran descargados los archivos, el algoritmo genera un ciclo donde va buscando por cada archivo el nombre de la central mayorista, el producto y su correspondiente precio en kg para luego adicionar a la base de datos principal.

Como resultado final se obtiene una base de datos long-term que guarda la información de los precios diarios para 740 posibles combinaciones de productos y centrales.

Componente 2

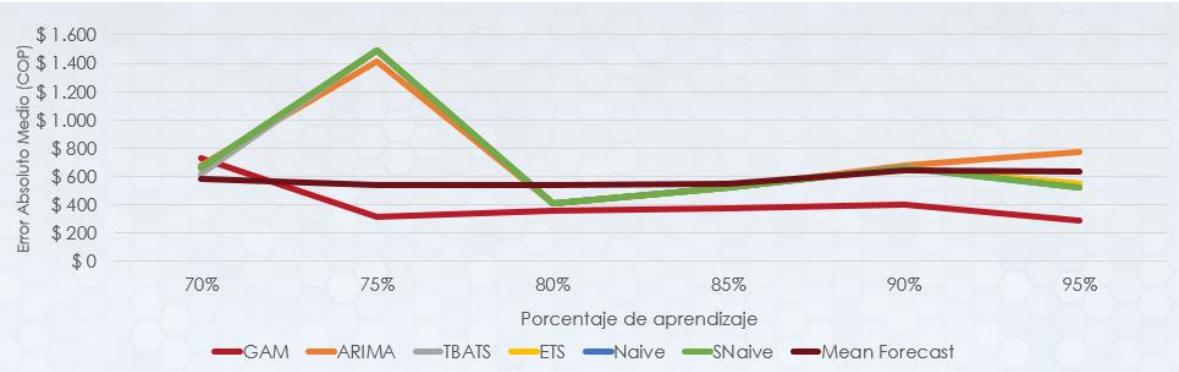
Para este módulo se desarrolló un algoritmo que evaluara 6 diferentes tipos de regresiones para series de tiempo que según el paper *Forecasting at Scale cuyos autores son: Sean J. Taylor – Benjamin Letham los cuales fueron:*

- TBATS model
- ARIMA
- Error, trend, seasonality
- Naive
- Generalized Additive Model
- Mean Forecast

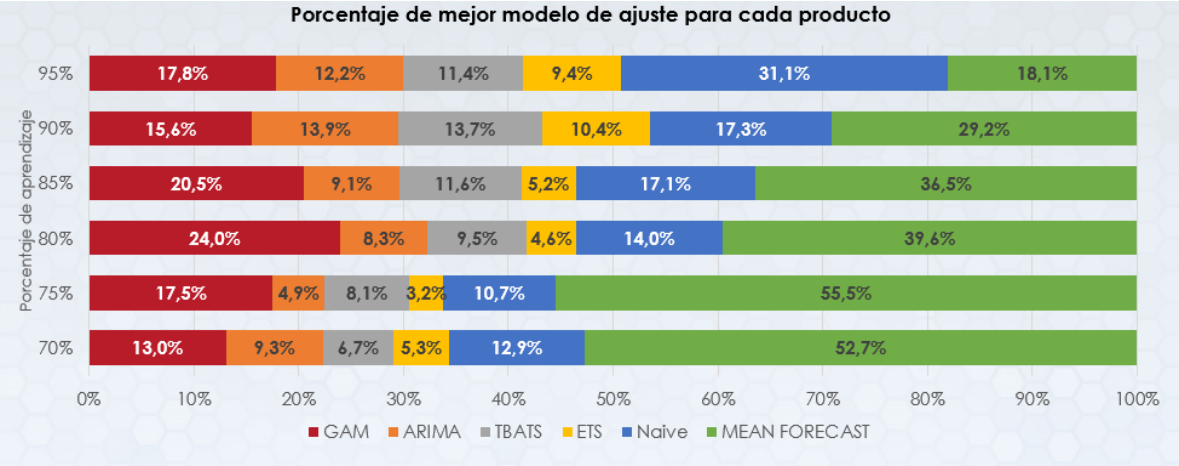
El motivo de evaluación con diferentes modelos es básicamente determinar para cada serie de tiempo disponible (producto X central) cual es el mejor ajuste basado en la comparación de los errores absolutos medios de cada modelo y variando el índice de aprendizaje del módulo de 5% en 5%.

El índice de aprendizaje es un porcentaje de sensibilidad que permite al modelo ver únicamente una parte de toda la serie en cuestión, de esta forma se obtendrá un set de datos complementarios para realizar el pronóstico y contrastarlo con los datos observados y de esa manera ver que tan bien son las predicciones realizadas.

El resultado para un producto aleatorio al graficar la diferencia de errores ante diferentes porcentajes de sensibilidad es:



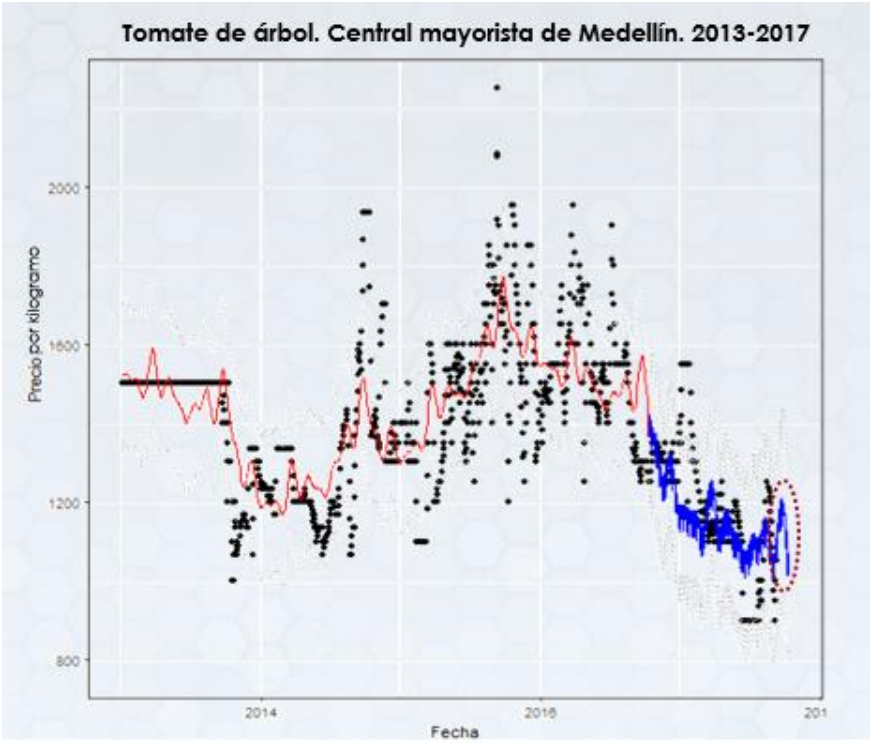
Luego, si este procedimiento se realiza para cada posible combinación de productos y se determina la distribución de cada modelo según el número de productos que determinar tener el mejor ajuste en este tenemos:

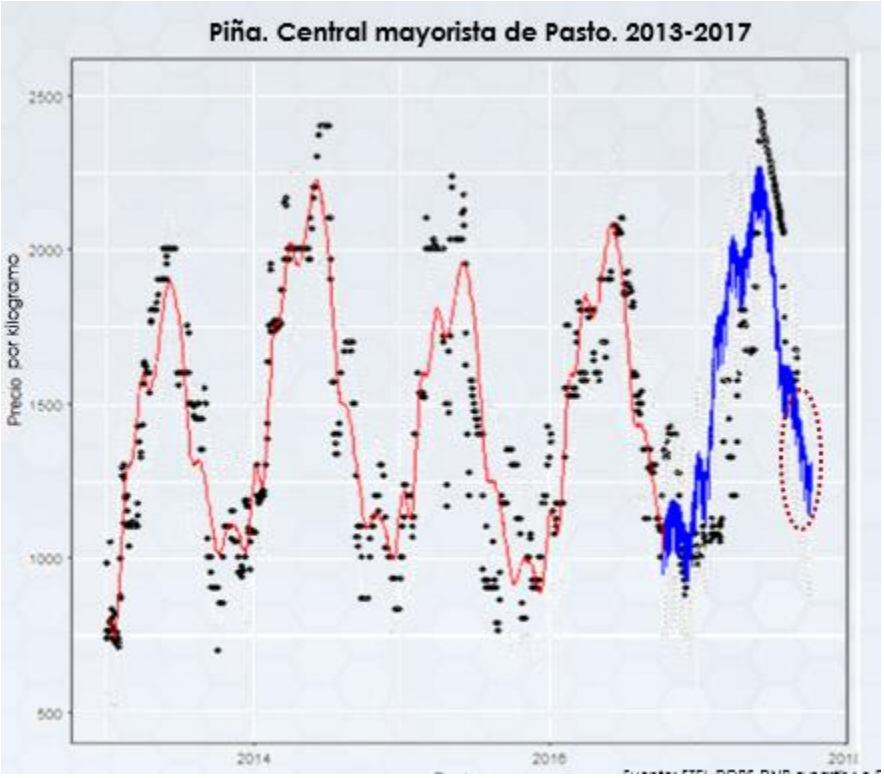


Finalmente, el algoritmo está en la capacidad de generar una tabla de "Algoritmo Ganador" para cada producto y que posteriormente sea utilizado por el componente de predicción (componente 3)

Componente 3

Luego de poder determinar cuál es el mejor modelo de ajuste se procede a realizar la predicción para los próximos 30 días contando desde la última fecha que se tiene registro en la base de datos. Para ello el algoritmo verifica que modelo aplicar y luego de generar las estimaciones necesarias procede a exportar una imagen formato png con el resultado de la predicción delineado con un tono azul, mientras que la parte de aprendizaje o la parte visible por el modelo con el cual se utilizó para realizar dicha predicción tiene una tonalidad roja. A continuación, algunos ejemplos:





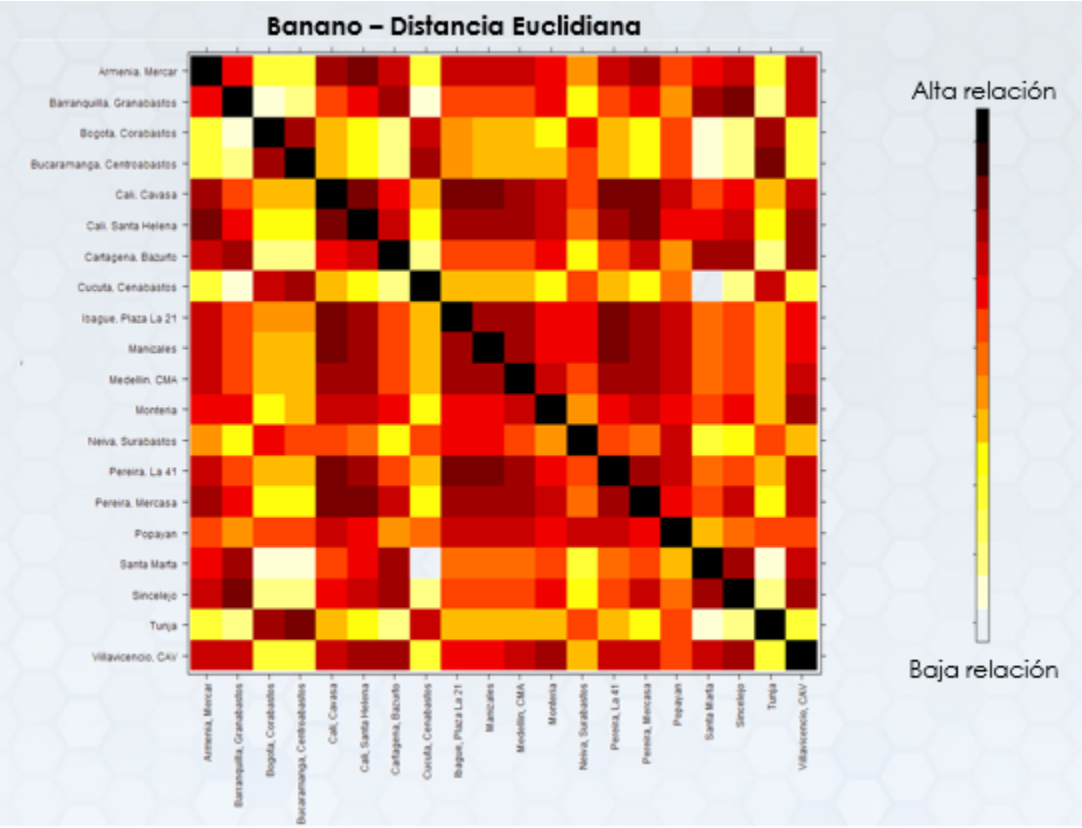
Componente 4

Para demostrar que la información capturada en el componente 1 puede ser de gran utilidad aparte de predecir el comportamiento de los precios en el mercado, se adicionó este componente buscando comparar el comportamiento de un producto en diferentes centrales mayoristas, logrando identificar un comportamiento geográfico crucial para la determinación de los precios en las zonas representadas por las centrales mayoristas.

Para este análisis se utilizaron 4 pruebas vectoriales distintas:

- Euclidian Distance
- Manhattan Distance
- Minkowski Distance
- Infinitive Norm Distance

Los resultados son imágenes png con mapas de calor determinando el nivel de cercanía que existe entre dos centrales para un mismo producto.



Los principales resultados de este módulo fueron encaminados a que visualmente fuese mucho más sencillo consumir las distancias calculadas, por tal razón el algoritmo está diseñado para correrse una vez y de volverse a ejecutar se esperaría encontrar cambios si han ocurrido choques externos que hayan alterado el comportamiento de los precios hacia el pasado.

Acceso al código de desarrollado

La herramienta se desarrolló en el lenguaje de programación R versión 3.3.3 y el código fuente de todo el proyecto se encuentra en el siguiente link:

- https://github.com/rojasdaniel/Monitoreo-de-precios-AGRO/tree/codigo_final

También se encuentra corriendo en el servidor VanaliticaDev del DNP el código para capturar información diaria desde el 1 de septiembre de 2017 actualizándose a las 7:00 p.m. y enviando la información a una carpeta compartida en OneDrive con todos los involucrados en el proyecto.