Big Data para el monitoreo de precios agropecuarios

Departamento Nacional de Planeación

Julio, 2017
dnp.gov.co

1

AGENDA

Análisis exploratorio de información

Proyecto general

Objetivo: Predecir el comportamiento de los precios de los productos agrícolas comercializados en las centrales mayoristas a partir de los sus valores históricos.

Fuente de información

SIPSA

El **DANE** captura de forma **diaria y semanal** los precios mayoristas de los productos agroalimentarios que se comercializan en el país

Características de los datos

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. Se presenta asimetría en la información reportada por productos y centrales mayoristas.

3. Los precios solo se publican para días hábiles.

4. Existen errores de digitación o cálculo en la información recopilada.

Distribución de registros en las centrales mayoristas

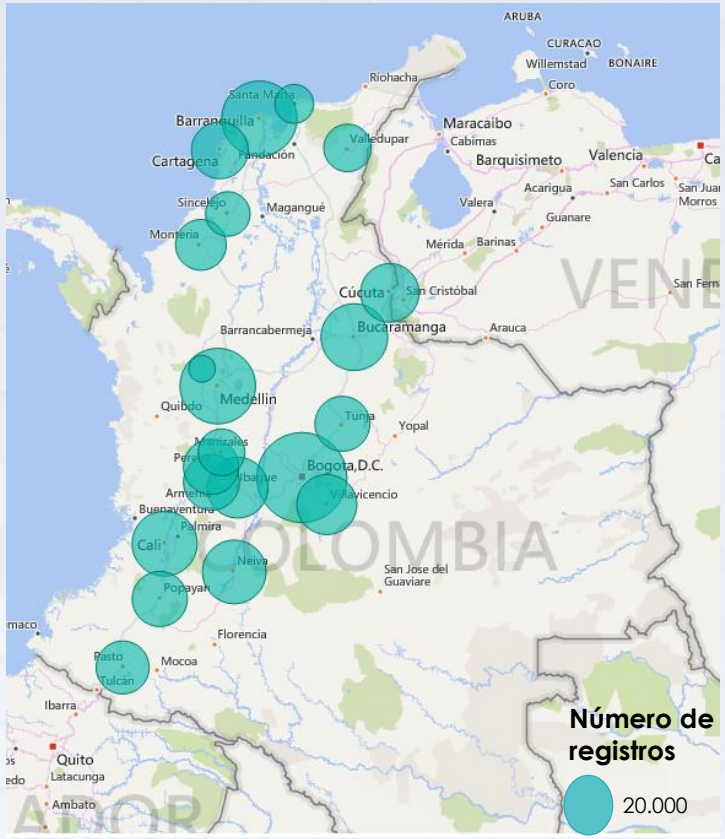
En promedio la base de datos crece diariamente **417 registros**, equivalente a ½ MB de información adicional.

Beneficios del proyecto

1. Capturar automáticamente el precio de los productos agrícolas reportados en SIPSA (diariamente).

2. Anticipar el comportamiento de los precios en las centrales mayoristas.

3. Brindar insumos a la DDRS sobre el comportamiento del mercado agrícola del país para la toma de decisiones de política pública



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
11. Villavicencio, CAV	21.123
12. Montería	21.050
13. Valledupar, Mercabastos	20.909
14. Sincelejo	19.720
15. Ibagué, Plaza La 21	16.692
16. Manizales	16.198
17. Pereira, La 41	15.754
18. Cali, Cavasa	14.228
19. Popayán	14.150
20. Pasto, El Potrerillo	12.746
21. Santa Marta	11.519
22. Cali, Santa Helena	9.108

Fuente: STEL-DDRS-DNP a partir de DANE (2014-2017)

Disponibilidad de productos y centrales

La información capturada desde el 2 de enero de 2013 hasta el 11 de septiembre de 2017* cuenta con **22 centrales mayoristas y 36 productos agrícolas**

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

Producto	Porcentaje de Centrales
Habichuela	**100%**
Lechuga Batavia	**100%**
Mora de Castilla	**100%**
Pepino Cohombro	**100%**
Pimentón	**100%**
Cebolla Cabezona Blanca	95%
Tomate de árbol	95%
Cebolla junca	91%
Granadilla	91%
Lulo	91%
Zanahoria	91%
Chócolo mazorca	86%
Guayaba	86%
Papa negra	86%
Pina	86%
Plátano hartón verde	86%
Aguacate	82%
Mandarina	82%
Maracuyá	82%
Papa criolla	82%
Papaya maradol	82%
Remolacha	82%

Producto	Porcentaje de Centrales
Tomate	82%
Arveja verde en vaina	77%
Frijol verde	77%
Mango Tommy	77%
Naranja	77%
Banano	73%
Manzana Royal Gala	73%
Limón común	68%
Limón Tahití	68%
Yuca	68%
Ahuyama	64%
Arracacha	**59%**
Coco	**23%**
Plátano guineo	**14%**

*Solo en días hábiles

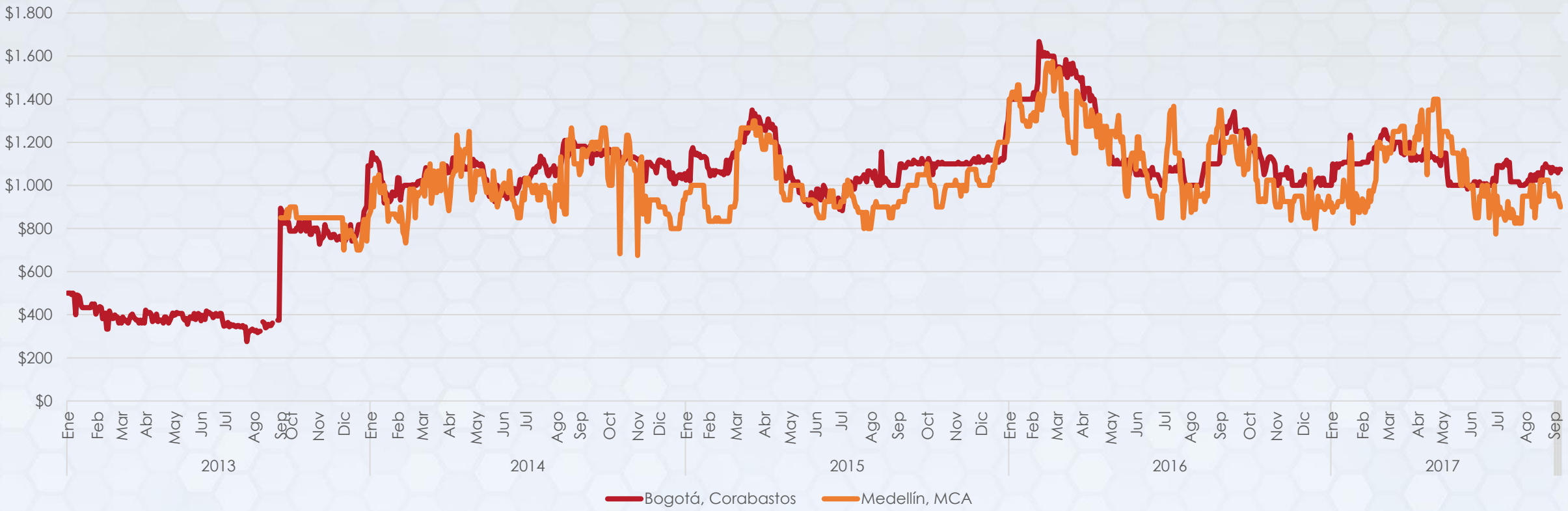
Fuente: STEL-DDRS-DNP a partir de DANE (2013-2017)

DNP Departamento Nacional de Planeación
TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

Naturaleza de la información

Luego de hacer un análisis exploratorio de los precios, se pueden evidenciar productos que tienen **datos faltantes, atípicos y cambios estructurales** en su comportamiento

Naranja. Enero 2013 – Septiembre 2017



Fuente: STEL-DDRS-DNP a partir de DANE (2013-2017)

2

AGENDA

Componentes

Componentes generales

Con el fin de brindar múltiples herramientas que aportaran información de utilidad a la DDRS se **desarrollaron 4 módulos** centrales durante la ejecución del proyecto

***Web Scraping* diario**

(Captura de información automatizada)

Identificación del mejor modelo de ajuste

(El método de pronóstico identificado es diferente para cada producto y/o central)

Pronóstico a 30 días de los precios agrícolas

***Heat map* con las similitudes entre las series de tiempo**

(Identificar similitud en comportamiento entre productos y/o centrales)

Web scraping diario

Se desarrolló un algoritmo que captura, organiza y almacena diariamente los precios publicados por el DANE en el portal SIPSA, **en promedio la base de datos crece diariamente 417 registros.**

Ciclo de sistematización para recolección de precios

Web Scraping

- Se identifica la última fecha capturada en la base de datos y se genera un vector de fechas faltantes
- Se capturan los link de cada archivo respectivo al vector de fechas faltantes

Limpieza de datos

- Se extrae cada archivo con los precios reportados y se depuran las variables a omitir
- Se genera una tabla por cada producto/central con los precios por kilogramo de cada producto

Adición a la base de datos *Master*

- Se genera un listado de registros con central, producto y precio para cada fecha
- Se adiciona el listado generado a la base de datos madre

Identificación del mejor modelo de ajuste

La validación estadística se realizó a partir de la comparación entre modelos y la selección de aquel que genere el menor promedio del valor absoluto del error para las 773 series asociadas a la combinación de productos y centrales. Para realizar la predicción se **selecciona el modelo de menor promedio del valor absoluto del error**

Ciclo de sistematización para cada central mayorista

Captura de información

PRODUCTO / CENTRAL

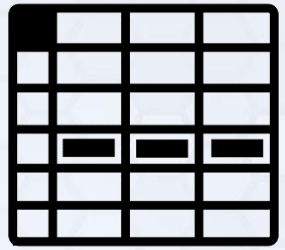
Sensibilidad de ajuste

Evaluación estadística

1. **GAM**
2. **ARIMA**
3. **TBATS**
4. **ETS**
5. **Naive**
6. **Mean Forecast**

Identificación del menor error

Registro de los errores



Gráficas con predicciones

Se realizan pruebas estadísticas bajo diferentes modelos en series de tiempo para identificar **la mejor predicción** por cada una de las 773 series asociadas a la combinación de productos y centrales

Ciclo de sistematización para cada central mayorista y/o producto

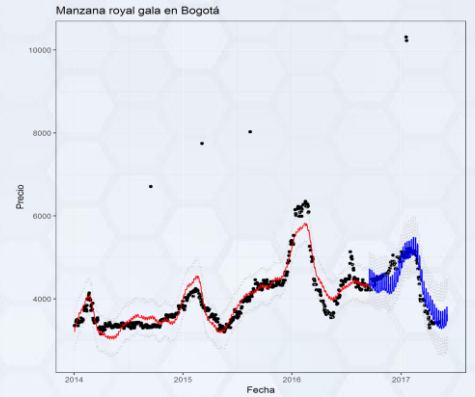
Selección

PRODUCTO

Selección de Central

Gráfica del modelo con mejor ajuste

Proyección a 30 días



Heat map con las similitudes entre series de tiempo

Se desarrolló un algoritmo que calcula las distancias entre cada serie con el fin de mostrar **la similitud entre productos y/o centrales** mediante mapas de calor.

Ciclo de sistematización para cada central mayorista o producto

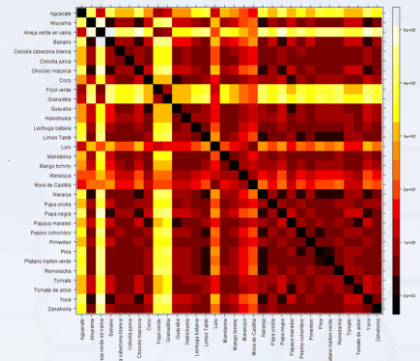
Selección de producto o central

PRODUCTO / CENTRAL

Cálculo de Distancias

1. Euclidiana
2. Manhattan
3. Minkowski
4. Infinitive Norm

Gráfica de matrices



DNP Departamento Nacional de Planeación
TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

3

AGENDA

Resultados

Web scraping diario

Se construyó una **base longitudinal** que permite hacer **seguimiento a la información histórica por producto y/o central mayorista**

Precios productos agrícolas. 8 de septiembre de 2017

Precio $/Kg	Barranquilla		Bogotá, Corabastos	
	Precio	Var %	Precio	Var %
Hortalizas y verduras				
Ahuyama	1.518	0	1.467	0
Arveja verde en vaina	5.938	-5	3.067	0
Cebolla cabezona blanca	1.508	3	1.467	-8
Cebolla junca	2.683	13	2.944	2
Chocolo mazorca	892	-3	867	-16
Pimentón	1.521	-5	1.533	2
Remolacha	1.442	7	1.556	0
Tomate*	1.267	3	1.848	0
Zanahoria	1.417	-7	1.389	9
Frutas frescas				
Aguacate *	3.500	14	3.500	0
Banano*	308	0	1.400	17
Coco	3.063	2	3.389	-1
Granadilla	n.d.	n.d.	3.333	1
Guayaba*	1.750	-1	1.333	5
Limón común	2.244	0	n.d.	n.d.
Limón Tahití	2.333	38	3.619	0
Naranja*	965	0	1.075	0
Papaya maradol	n.d.	n.d.	1.556	2
Piña *	1.045	0	693	1
Tomate de árbol	1.950	0	2.267	-3
Tubérculos y plátanos				
Arracacha*	n.d.	n.d.	1.472	0
Papa criolla	1.990	4	2.889	1
Papa negra*	628	22	767	-1
Plátano guineo	n.d.	n.d.	1.200	0
Plátano hartón verde	1.065	0	1.100	-4
Yuca*	728	7	833	-2

Fuente: DANE (2017)

Segmento base de datos madre

Fecha	Central	Producto	Precio kg
3/01/2013	Barranquilla, Granabastos	Ahuyama	n.d.
8/01/2013	Barranquilla, Granabastos	Ahuyama	686
10/01/2013	Barranquilla, Granabastos	Ahuyama	625
11/01/2013	Barranquilla, Granabastos	Ahuyama	696
14/01/2013	Barranquilla, Granabastos	Ahuyama	616
15/01/2013	Barranquilla, Granabastos	Ahuyama	589
17/01/2013	Barranquilla, Granabastos	Ahuyama	548
21/01/2013	Barranquilla, Granabastos	Ahuyama	532
22/01/2013	Barranquilla, Granabastos	Ahuyama	524
24/01/2013	Barranquilla, Granabastos	Ahuyama	539
28/01/2013	Barranquilla, Granabastos	Ahuyama	539
29/01/2013	Barranquilla, Granabastos	Ahuyama	539
31/01/2013	Barranquilla, Granabastos	Ahuyama	575
4/02/2013	Barranquilla, Granabastos	Ahuyama	575
5/02/2013	Barranquilla, Granabastos	Ahuyama	575
7/02/2013	Barranquilla, Granabastos	Ahuyama	579
12/02/2013	Barranquilla, Granabastos	Ahuyama	625
14/02/2013	Barranquilla, Granabastos	Ahuyama	662
18/02/2013	Barranquilla, Granabastos	Ahuyama	652
19/02/2013	Barranquilla, Granabastos	Ahuyama	659
21/02/2013	Barranquilla, Granabastos	Ahuyama	584
25/02/2013	Barranquilla, Granabastos	Ahuyama	589
26/02/2013	Barranquilla, Granabastos	Ahuyama	545

Fuente: STEL-DDRS-DNP a partir de DANE (2017)

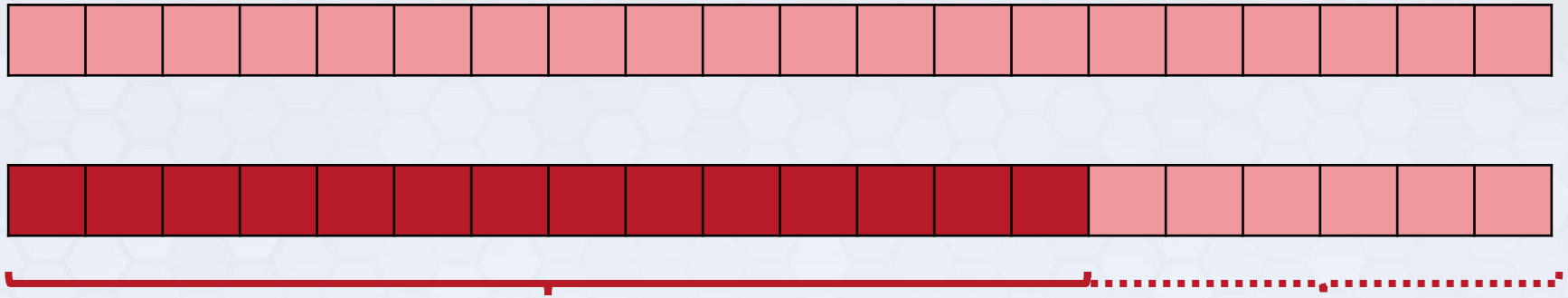
DNP Departamento Nacional de Planeación
TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

Identificación del mejor modelo de ajuste

Para cada serie (central-producto) se aplicaron diferentes modelos que permitiera identificar el de **mejor ajuste**

Serie de tiempo enero 2013 - septiembre 2017

Se **segmenta** la serie **en periodo de aprendizaje** y **periodo de prueba**



Análisis de sensibilidad: Periodo de aprendizaje varía entre el **70 %** y el **95 %**

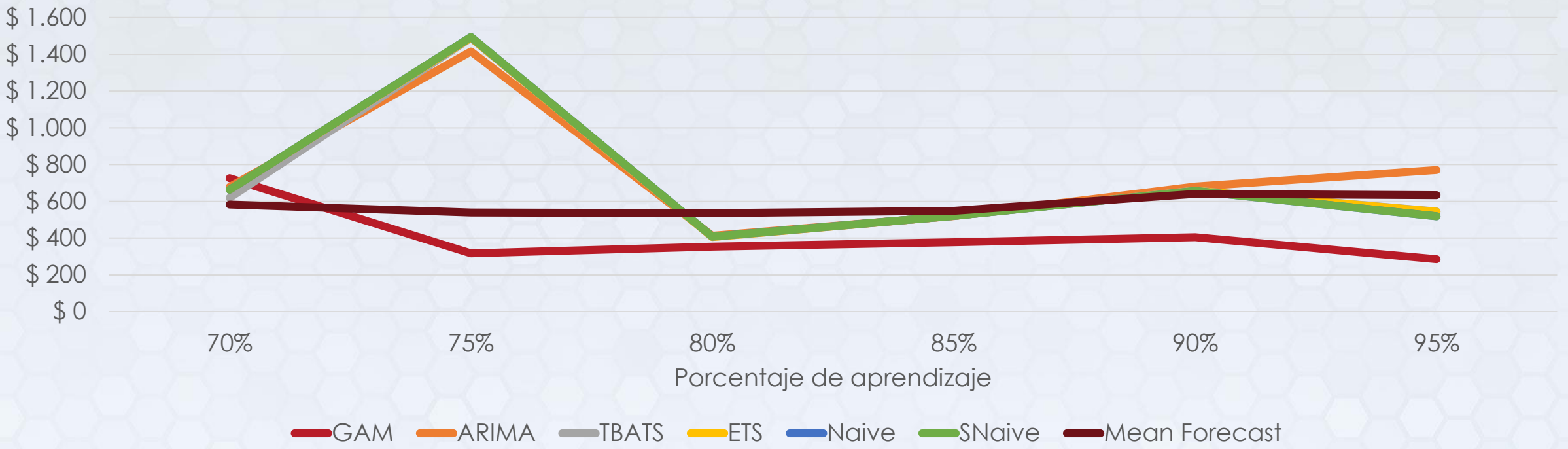
Modelos: GAM, ARIMA, TBATS, ETS, Naive, Snaive, Mean Forecast

Criterio de decisión: Promedio del valor absoluto del error de pronóstico

Identificación del mejor modelo de ajuste

La **selección del modelo de predicción** está basado en el modelo que genere el **menor error de pronóstico**

Análisis de sensibilidad del promedio del valor absoluto del error.
Producto aguacate. Central Bogotá-Corabastos



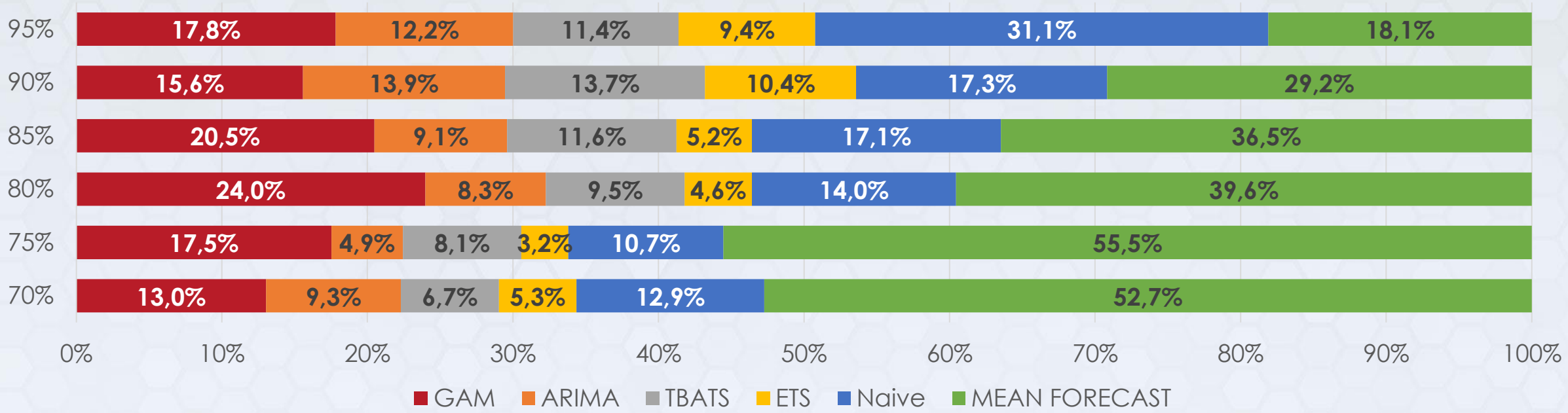
Fuente: STEL-DNP a partir de DANE (2014-2017)

Independiente del nivel de aprendizaje el modelo de **mejor ajuste es el GAM** para el aguacate en Corabastos

Identificación del mejor modelo de ajuste

No existe un único método de pronóstico que logre el mejor ajuste a todas las series

Análisis de sensibilidad del porcentaje de series según método de mejor ajuste



Fuente: STEL-DDRS-DNP a partir de DANE (2014-2017)

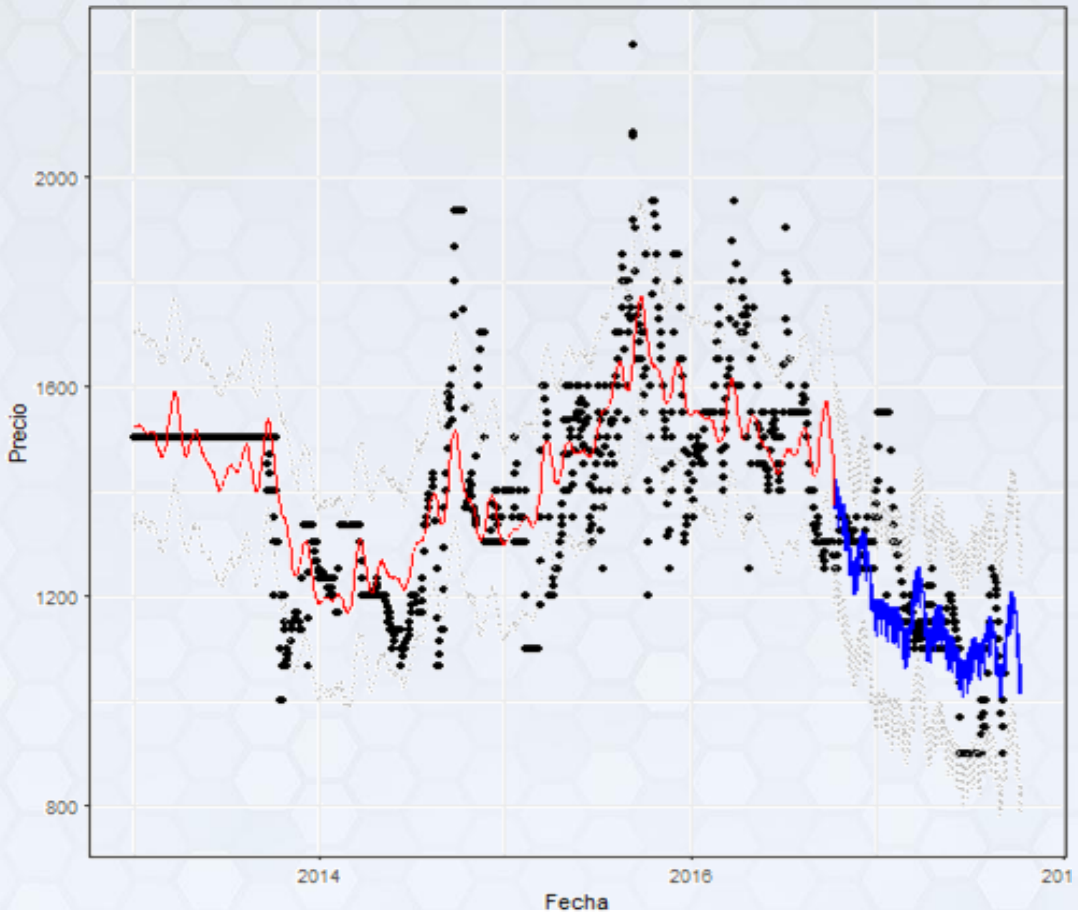
Bajo un **aprendizaje del 70 %**, el ***Mean Forecast*** es el método de mejor ajuste para la **mitad de las series**, pero si se amplía al 95% se reduce a menos de 1 de cada 5 series.

El **modelo que es menos variable** (independiente del porcentaje de aprendizaje) es el **GAM**

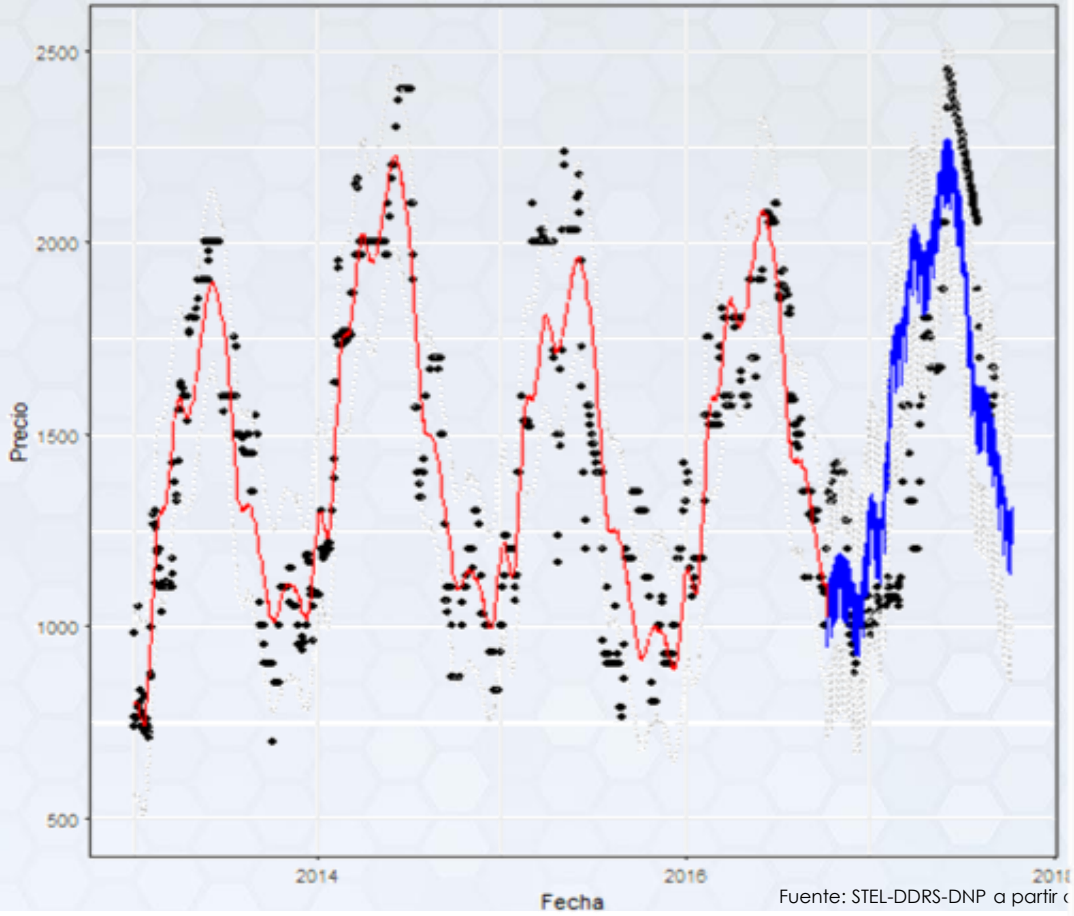
Pronóstico a 30 días de los precios

El modelo **GAM** es **resistente a los efectos de los valores atípicos**, y soporta datos recopilados en una escala de tiempo irregular (sin presencia de datos faltantes) y sin necesidad de interpolación.

Tomate de árbol. Central mayorista de Medellín. 2013-2017



Piña. Central mayorista de Pasto. 2013-2017

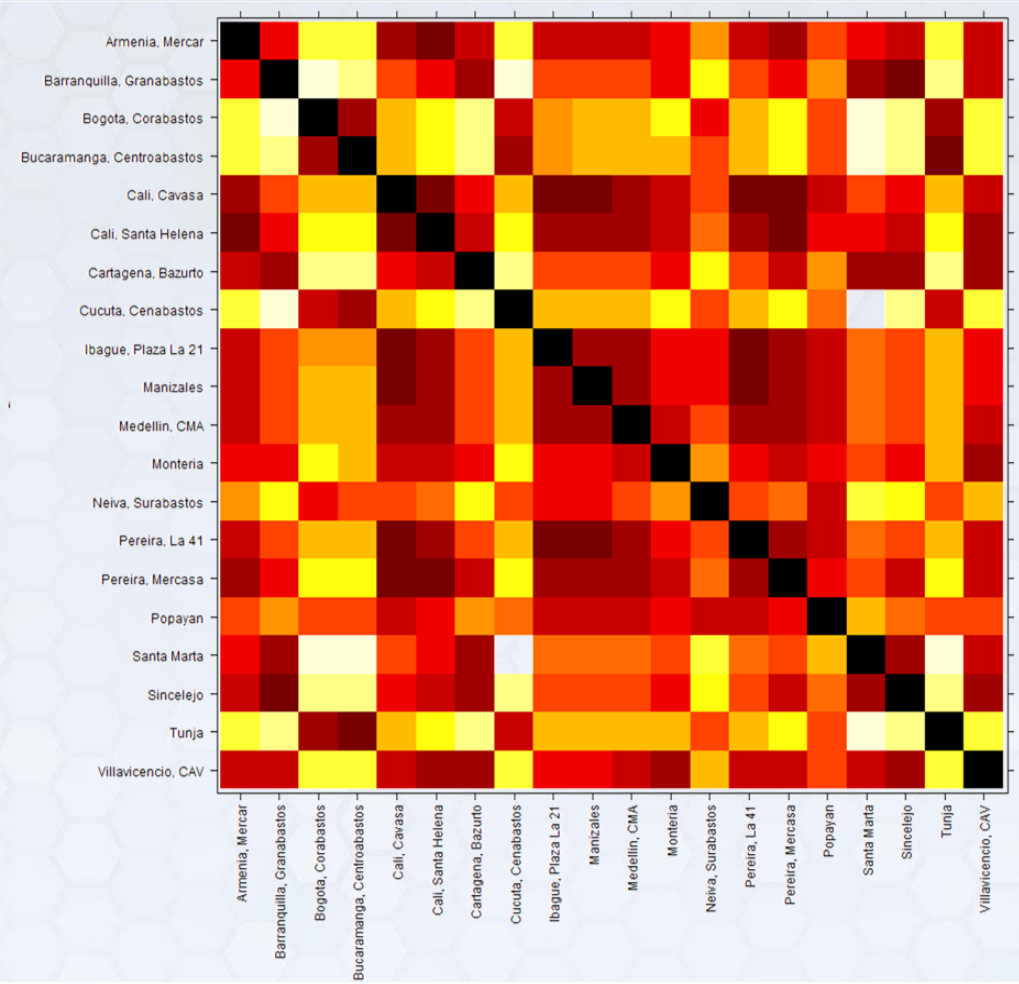


Fuente: STEL-DDRS-DNP a partir de DANE (2017)

Mapas de calor con similitudes entre centrales

El precio del banano presenta una **similitud alta entre centrales mayoristas** respecto a la habichuela. Se destacan las relaciones entre Pereira, Armenia, Cali, Ibagué y Manizales

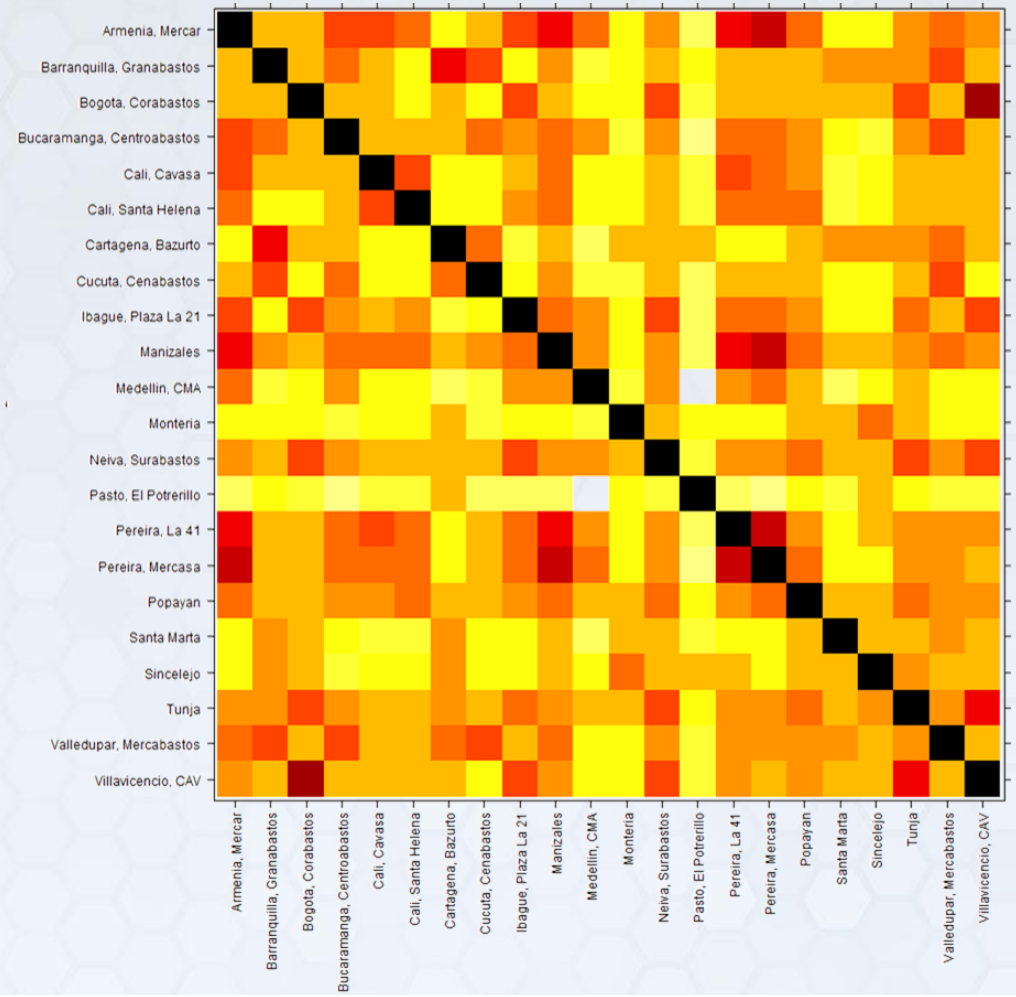
Banano – Distancia Euclidiana



Baja relación

Alta relación

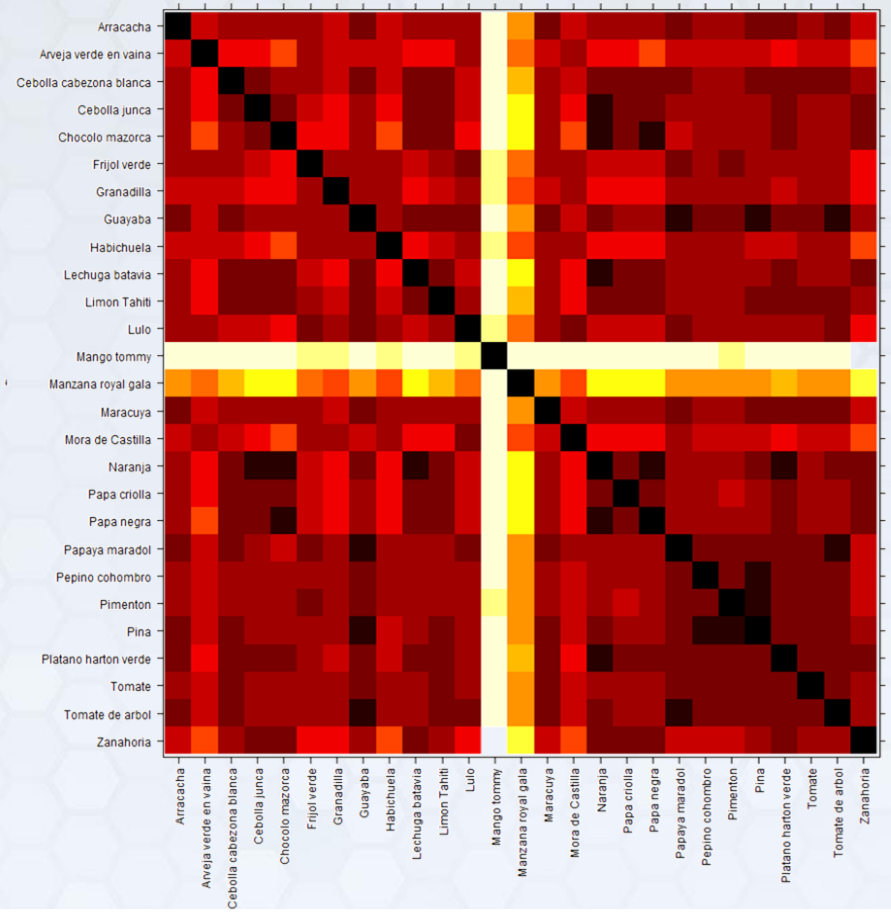
Habichuela – Distancia Euclidiana



Mapas de calor con similitudes entre productos

El comportamiento de los precios, en las centrales, es heterogéneo. Se requiere un **análisis diferenciado territorialmente** para el diseño de política pública

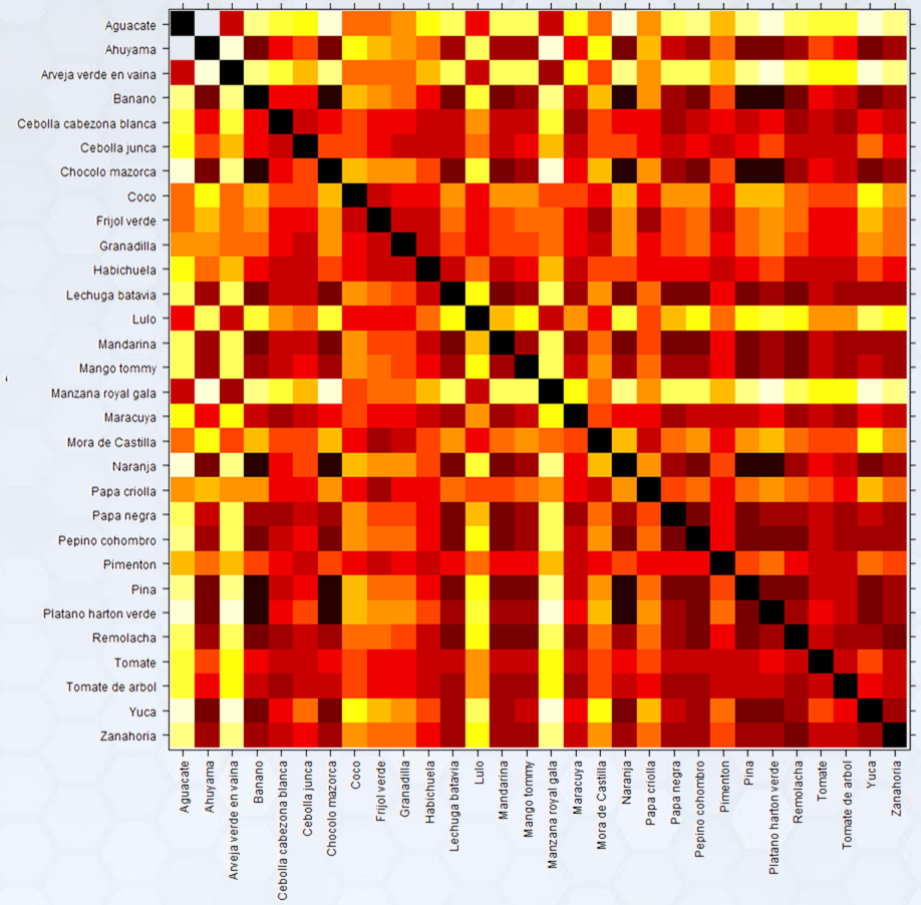
Pasto, El Potrerillo – Distancia Euclidiana



Baja relación

Alta relación

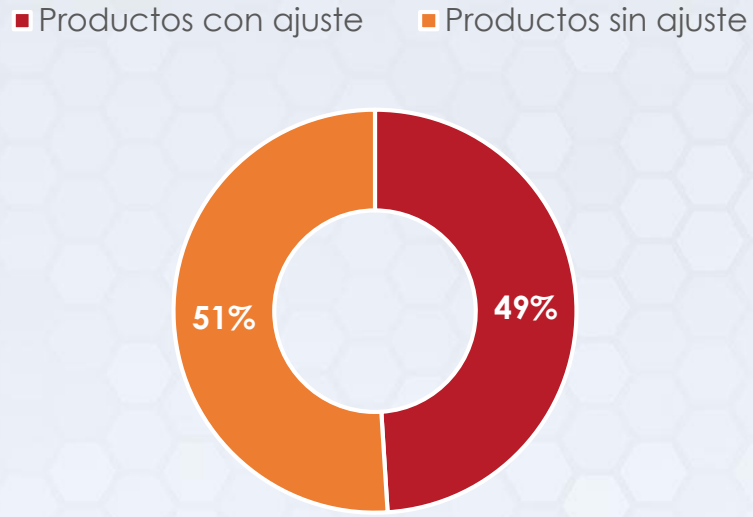
Montería – Distancia Euclidiana



Aprovechamiento del insumo generado por la STEL

Con datos adicionales (precipitaciones , variabilidad climática y abastecimiento) es posible mejorar el ajuste de los modelos.

Porcentaje de productos pronosticados correctamente



Número de productos: 617

La Dirección de Desarrollo Rural Sostenible y el Ministerio de Agricultura y Desarrollo Rural cuentan ahora con una base de datos de actualización diaria y de fácil acceso al consumo, para futuras investigaciones o análisis sobre los precios agrícolas.

Departamento Nacional de Planeación

www.dnp.gov.co