Extracción, clasificación y procesamiento de datos de alta frecuencia de supermercados

Gonzalo Bueno y Marco Vega

Banco Central de Reserva del Perú

Las opiniones expresadas en este estudio corresponden a los autores y no deben ser atribuidos al Banco Central de Reserva del Perú.

Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

Episodios de inflación

- El proceso inflacionario en el Perú desde inicios de milenio se ha caracterizado por ser estable y con una tasa promedio similar al de países desarrollados.
- Desde mediados de 2021 a inicios de 2023 se ha observado casos particulares de choques de oferta.

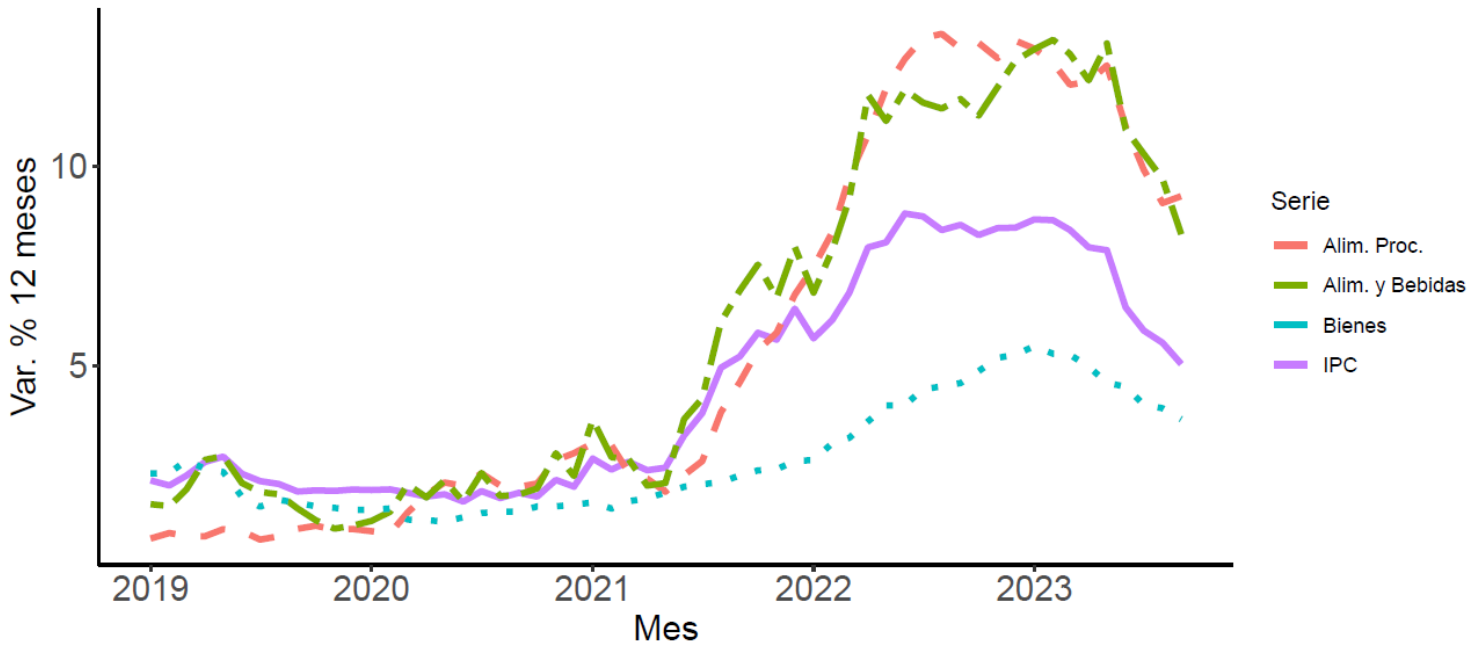


Figura: Variación a 12 meses de índices.

Fuentes de información

- El análisis de la inflación se puede nutrir a partir del estudio y comprensión del proceso de formación de precios individuales.

- Actualmente el BCRP no cuenta con información completa de alta frecuencia sobre la evolución de los precios de los alimentos procesados y otro tipo de bienes. La información recogida de otras instituciones solo muestra una selección parcial.

- En periodos de alta inflación se incrementa la frecuencia de ajustes de precios, por lo que se requiere de este tipo de datos.

Literatura relacionada

- La literatura es relativamente reciente y escasa:
  - Extracción de precios: Billions Prices Project (Cavallo & Rigobon 2016) extracción de precios en línea de tiendas de hasta 50 países.
  - Indicadores de alta frecuencia: Cavallo (2013).
  - Predicción de la inflación y nowcasting: Aparicio & Bertolotto (2020), Macias, Stelmasiak, & Szafranek (2022).
  - Duración y rigideces de precios: Gorodnichenko & Talavera (2017), Cavallo (2018), Coronado y otros (2020), Lunnemann & Wintr (2006).
- Dado el gran volumen de datos, es inevitable el uso de algoritmos de aprendizaje de máquina. Por ello, existe una literatura originada principalmente en la comunidad de científicos de datos.

Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

¿Cuál ha sido el trabajo realizado?

1. Extracción y clasificación de datos

- Se extrajeron los datos de webs de supermercados.
- Se dividieron entre alimentos, bebidas y bienes para su clasificación.
- A través de algoritmos de machine learning se asignó cada producto.

¿Cuál ha sido el trabajo realizado?

2. Procesamiento y análisis

- Las series de precios pueden tener interrupciones, o aparecer/desaparecer con el tiempo. Además, existe ruido por la presencia de descuentos y incrementos cortos en forma de saltos de precios.
    1. Se optó por un método de relleno de datos.
    2. Se filtraron las series para suavizarlas.
- Aún así, es retador buscar una forma consistente de agregación para productos con diversas variedades, presentaciones y marcas.

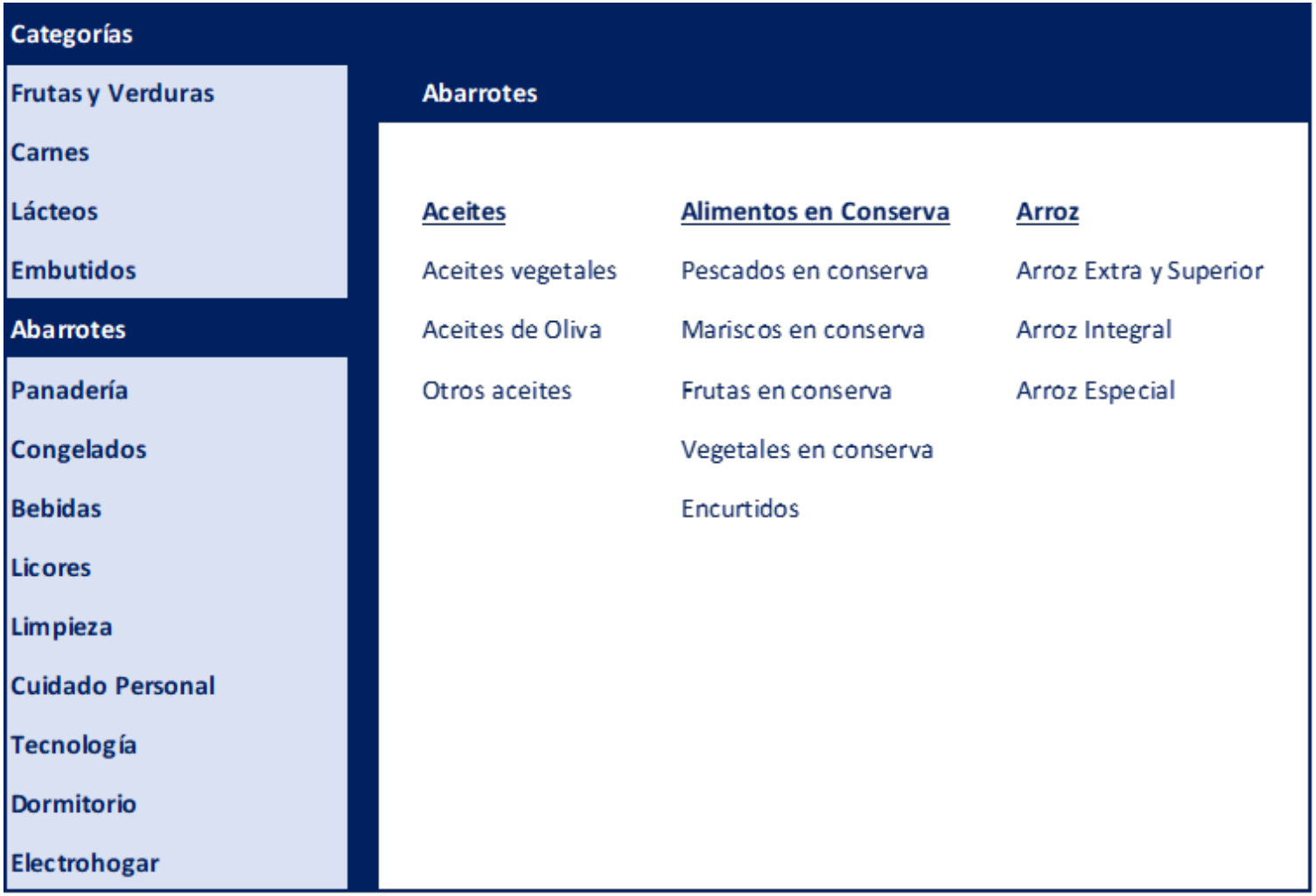
Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

Datos extraídos

- Datos desde 2020. Inicios con visitas presenciales.
- Información obtenida: (i) nombre, (ii) precio, (iii) marca y (iv) categorías.
- Se han obtenido alrededor de 134 mil registros únicos de variedades de productos que alguna vez han sido vendidos, sin importar si aún siguen disponibles.
- Existen tres precios: (i) el precio regular, (ii) el precio “en línea”, mejor precio o precio con descuento y (iii) el precio con descuento especial (tarjeta particular). El precio usado para el seguimiento de la inflación es el que tiene un descuento (o el regular en caso no esté disponible), que sería el **precio efectivo** que pagaría un consumidor.

1. Crawler


Categorías
Frutas y Verduras
Carnes
Lácteos
Embutidos
Abarrotes
Panadería
Congelados
Bebidas
Licores
Limpieza
Cuidado Personal
Tecnología
Dormitorio
Electrohogar
Abarrotes
Aceites
Aceites vegetales
Aceites de Oliva
Otros aceites
Alimentos en Conserva
Pescados en conserva
Mariscos en conserva
Frutas en conserva
Vegetales en conserva
Encurtidos
Arroz
Arroz Extra y Superior
Arroz Integral
Arroz Especial

2. Web Scraping

Aceites vegetales

Aceite Vegetal Primor Premium 900ml

Precio con Descuento S/ 11.50

~~Precio Regular~~ ~~S/ 12.10~~

Aceite de Soya Sao 900ml

Precio con Descuento S/ 8.45

~~Precio Regular~~ ~~S/ 9.15~~

Aceite Vegetal Máxima 900ml

Precio con Descuento S/ 7.90

llación de series

- Cada producto se diferencia por su nombre. Un cambio en esta variable genera una nueva serie de precios para un producto distinto.

- Por ejemplo, un vino cuyo nombre primero se registre como “Vino rosé”, con una tilde en su nombre, y luego como “Vino rose”, entrarían en la base de datos como dos productos distintos. De igual forma si se tiene “Pack galleta de chocolate” y “Galleta de chocolate pack”.

- Para evitar eso, se asignaron nombres con limpieza de caracteres raros (como tildes o guiones) y sin mayúsculas a cada producto descargado para usar como identificador.

Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

CCIF

- Las series descargadas tienen asociadas una clasificación de acuerdo a los criterios de los supermercados.

- Debido a que el objetivo es tener nuevas fuentes de información para hacerle seguimiento a la inflación, es necesario tener una agrupación de acuerdo a las categorías del IPC. Estas corresponden a la Clasificación del Consumo Individual por Finalidades (CCIF o COICOP por sus siglas en inglés).

CCIF

CLASIFICACIÓN DEL CONSUMO INDIVIDUAL POR FINALIDADES - CCIF

DIVISIONES

1. ALIMENTOS Y BEBIDAS NO ALCOHÓLICAS
2. BEBIDAS ALCOHÓLICAS, TABACO Y ESTUPEFACIENTES
3. PRENDAS DE VESTIR Y CALZADO
4. ALOJAMIENTO, AGUA, ELECTRIDAD, GAS Y OTROS COMBUSTIBLES
5. MUEBLES, ARTICULOS PARA EL HOGAR Y PARA LA CONSERVACION ORDINARIA DEL HOGAR
6. SALUD
7. TRANSPORTE
8. COMUNICACIONES
9. RECREACION Y CULTURA
10. EDUCACION
11. RESTAURANTES Y HOTELES
12. BIENES Y SERVICIOS DIVERSOS

Reconstrucción de rubros con la Enapref

- Los datos disponibles del INEI muestran la desagregación para alimentos y bebidas y bienes a niveles agregados. Se reconstruyeron las clasificaciones hasta el nivel de producto con la información disponible en la Encuesta Nacional de Presupuesto Familiar (Enapref).

- Esto sirve para descartar los rubros y productos que son parte del gasto de las familias pero tienen una ponderación marginal tal que no entran en la canasta del IPC.

Clasificación con machine learning

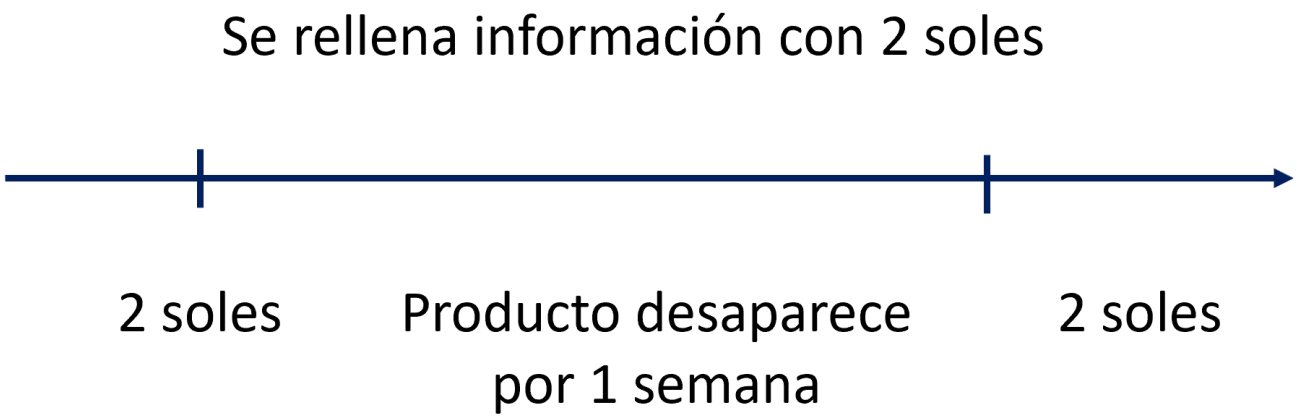
- Modelo de aprendizaje supervisado de text classification (Joulin et al. 2017).
- A cada nombre de variedad específica se le asocia al tipo de producto correspondiente. Por ejemplo, el “Aceite de maíz marca ABC 1L” se clasifica como “aceite vegetal envasado”.
- Se construyeron muestras de entrenamiento clasificando manualmente tanto alimentos, bebidas y bienes.

Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

Relleno de datos

- Para periodos anteriores a 2023, no se disponen de datos para fines de semana y para algunos productos, se tienen huecos en las series de precios.



Relleno de datos

Se mantienen *missings*.

2 soles — Producto desaparece por 32 días — 2 soles

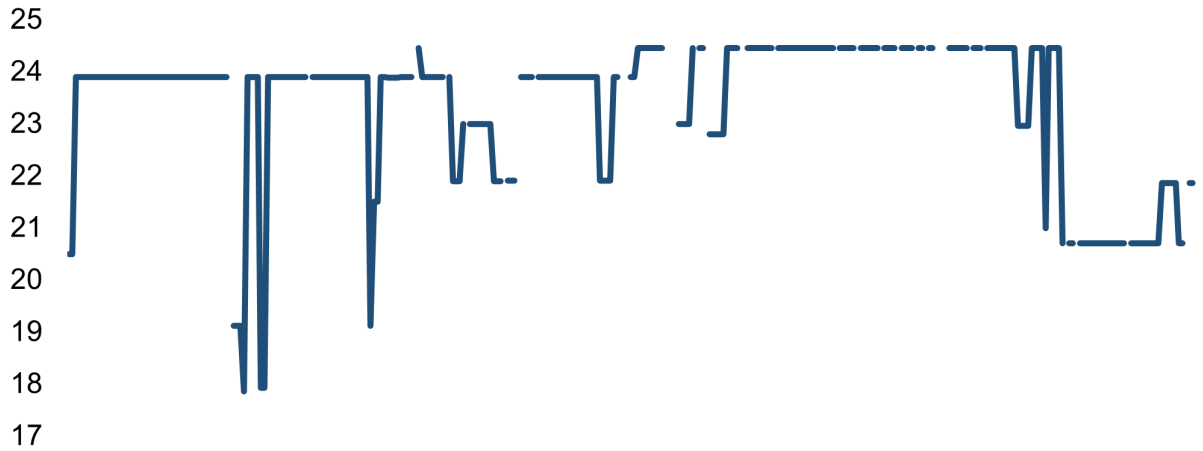
Se mantienen *missings*.

2 soles — Producto desaparece por 1 semana — 3 soles

Media móvil

- Para suavizar las series de precios, se tomó una media móvil de treinta días con al menos 7 días de datos.

- Las series de precios tienen un comportamiento volátil debido a la presencia de descuentos e incrementos de corto plazo. El filtro de media móvil a 30 días suaviza las series ante estos saltos, pero igual captura movimientos no tendenciales.

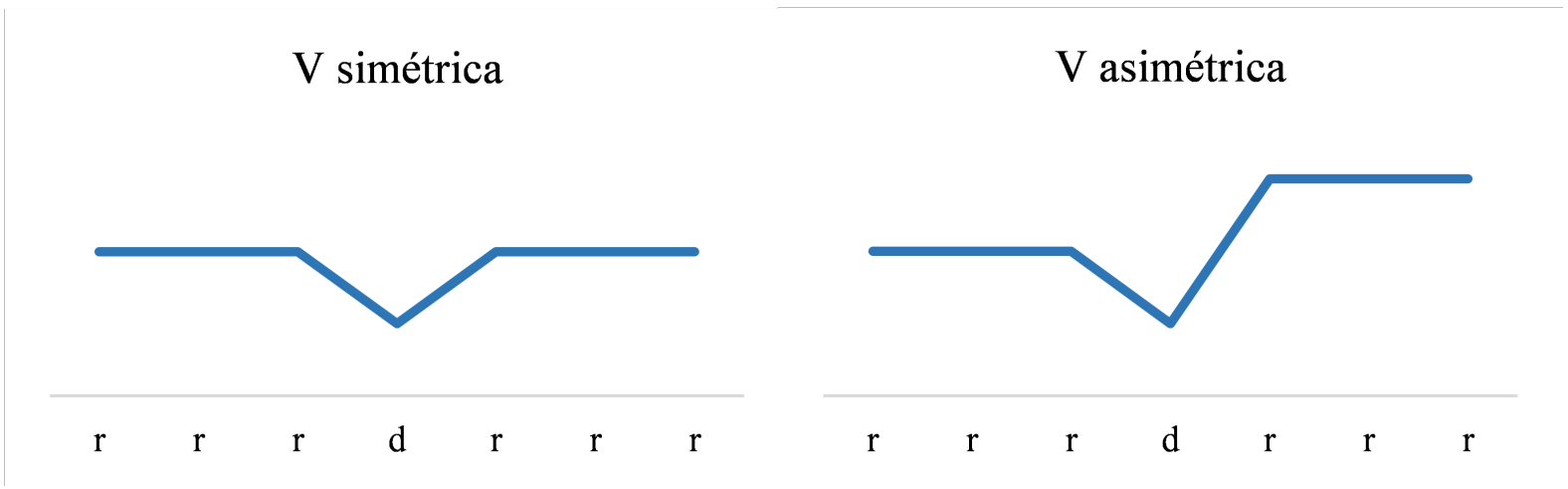


Precios regulares

Nakamura & Steinsson (2008)

- Filtro basado en Nakamura and Steinsson (2008). Este algoritmo distingue periodos de descuentos detectando patrones de precios en formas V simétricas y asimétricas.

- Por otro lado, puede darse que los descuentos persistan por una ventana temporal de una longitud tan larga que se podría asumir que el nivel de precio se ha reducido.



Precios regulares

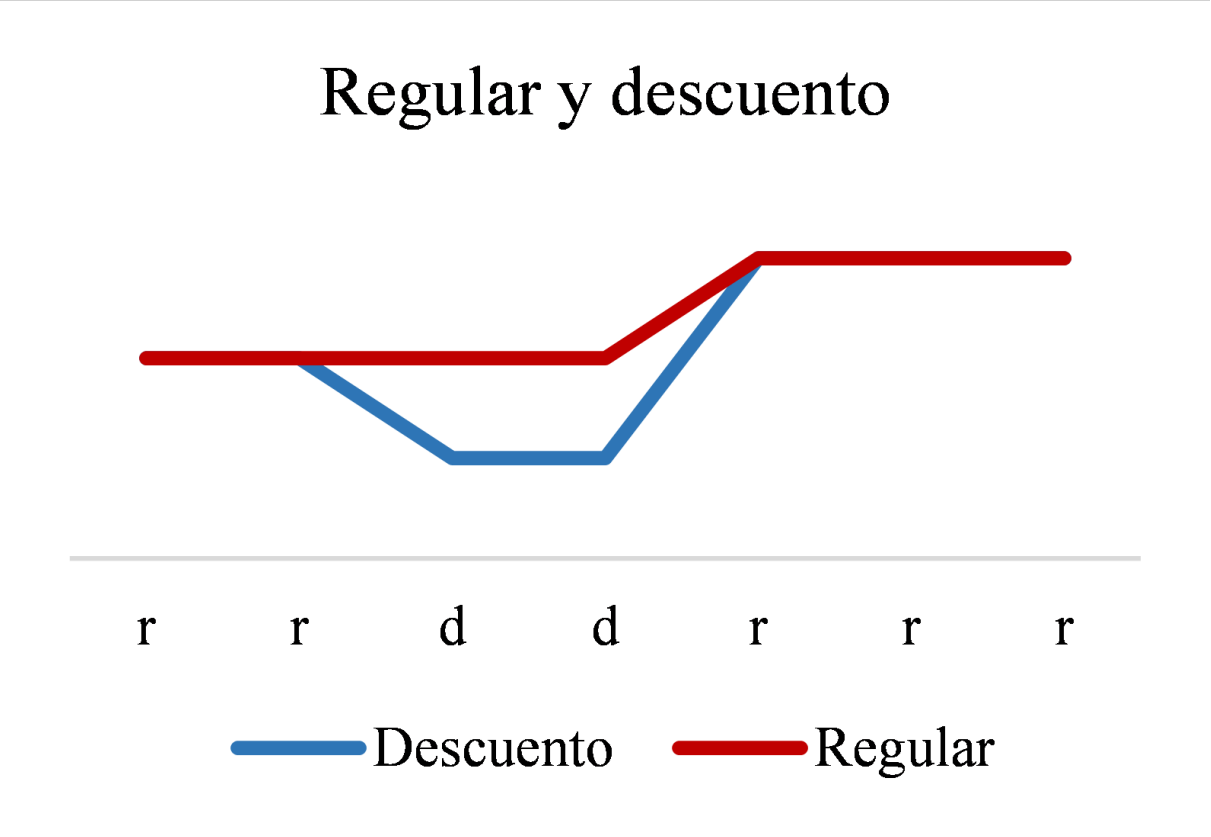
Nakamura & Steinsson (2008)

- Un problema adicional: la presencia de descuentos simultáneos a una caída en el precio regular.

Precios regulares

Nakamura & Steinsson (2008)

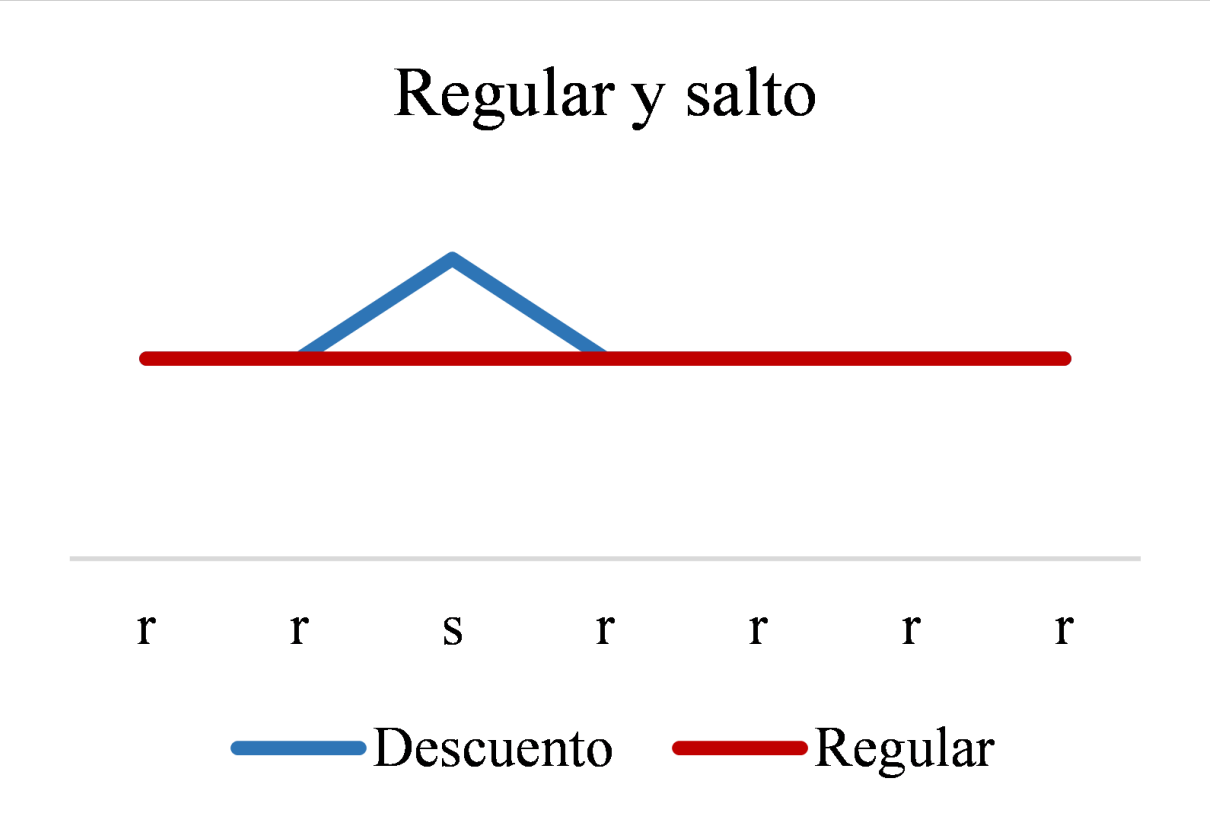
- Los parámetros del algoritmo van a depender del rubro o tipo del producto.



Precios regulares

Nakamura & Steinsson (2008)

- Después de filtrar los descuentos, se notaron series de precios que tenían saltos hacia arriba de poca duración.



Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

Mediana de variaciones interanuales

Media móvil

- Incremento en la inflación que se intensificó durante el 2022. Sin embargo, los alimentos y bebidas se diferencian del resto de bienes debido a la persistencia.

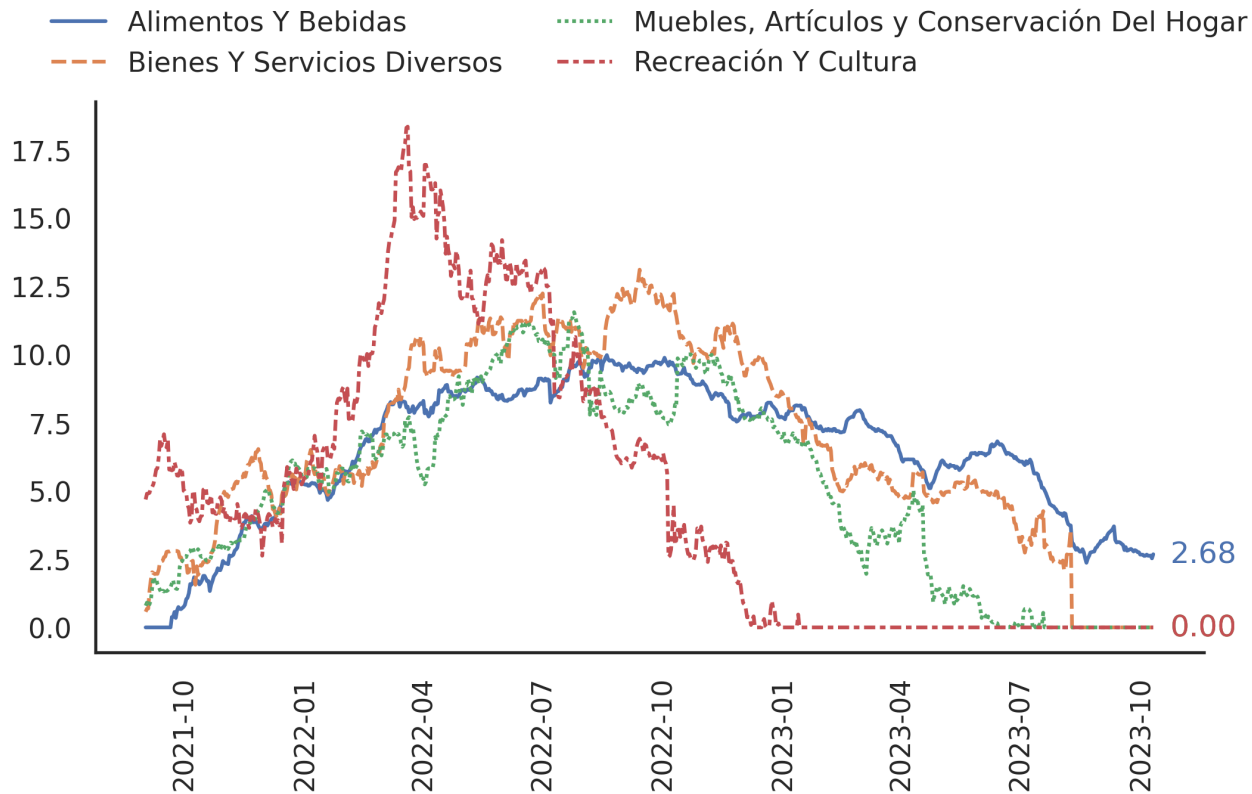


Figura: Medianas sobre la variación interanual.

Distribución de variaciones interanuales

Media móvil

- Los alimentos han tenido un comportamiento más estable. En cambio los bienes tienen una mayor proporción de productos con variación interanual cero.

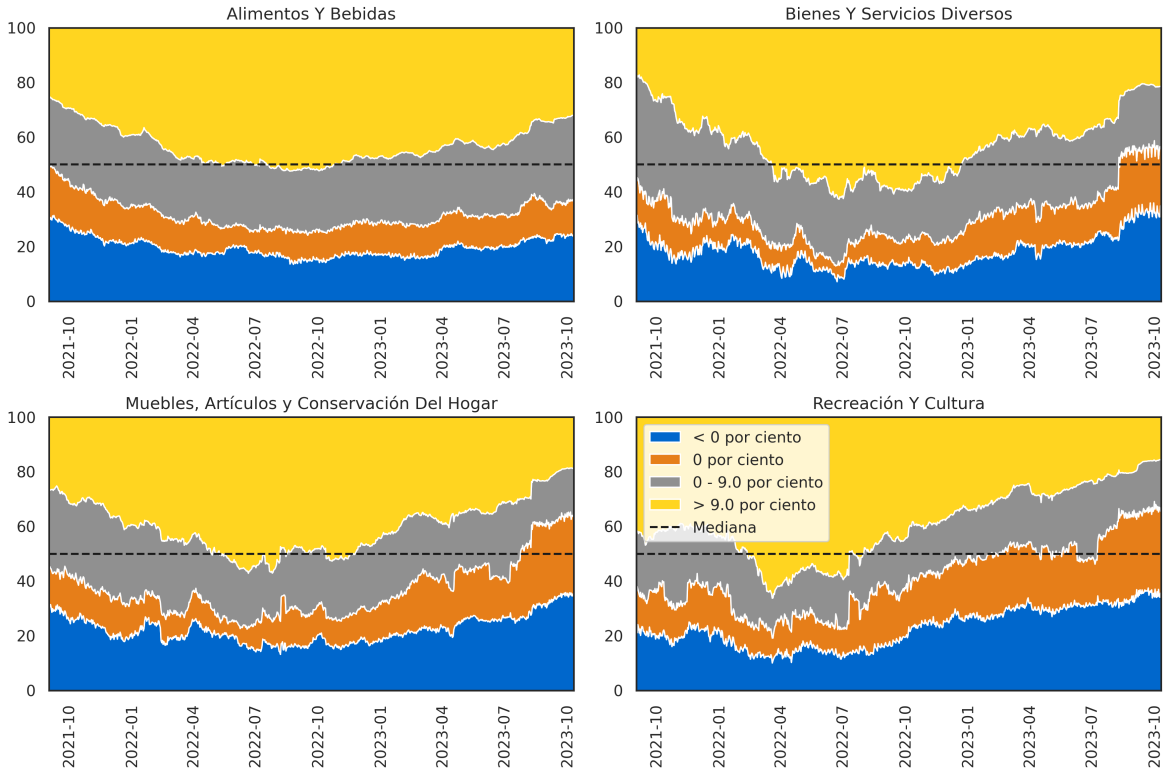


Figura: Distribución por rangos de la variación interanual de precios con filtro de media móvil.

Mediana de variaciones interanuales

Media móvil

- Los alimentos perecibles han tenido un incremento de precios más persistente a comparación de los procesados. Se evidencia el efecto de choques y formación de precios particulares.

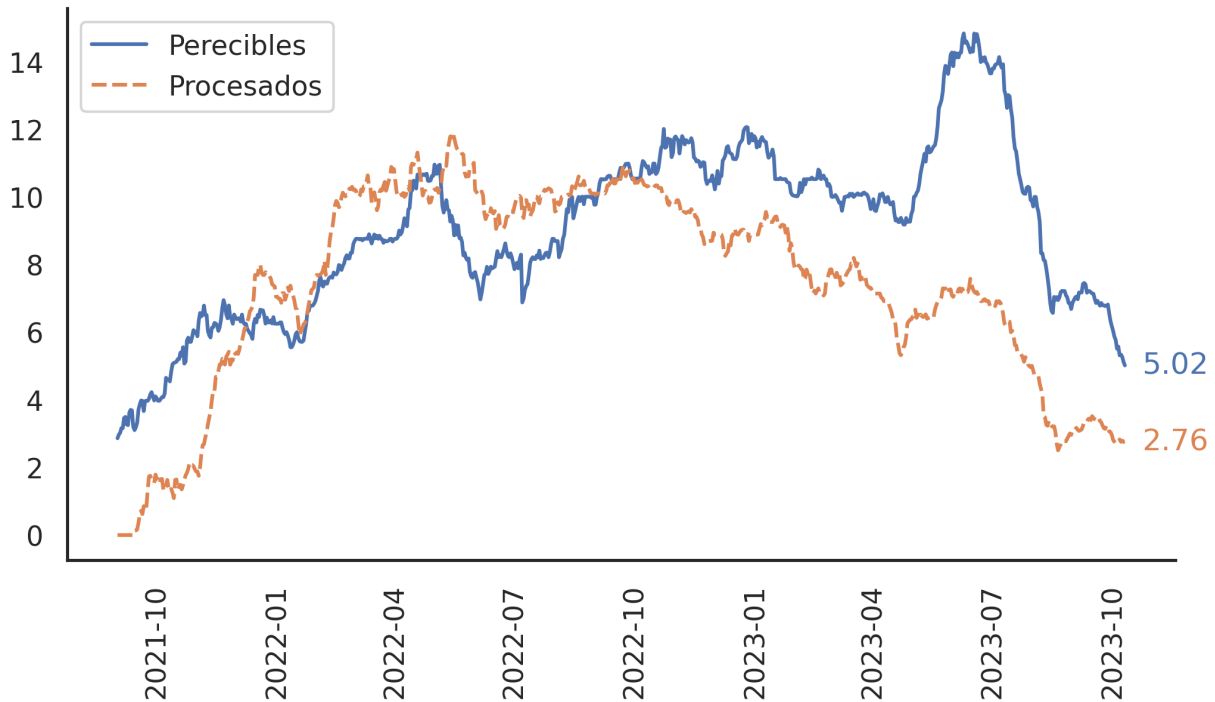


Figura: Medianas sobre la variación interanual.

Mediana de variaciones interanuales

Precios regulares

- Las tendencias, sobre todo el calentamiento en 2022 y la moderación posterior se mantienen.

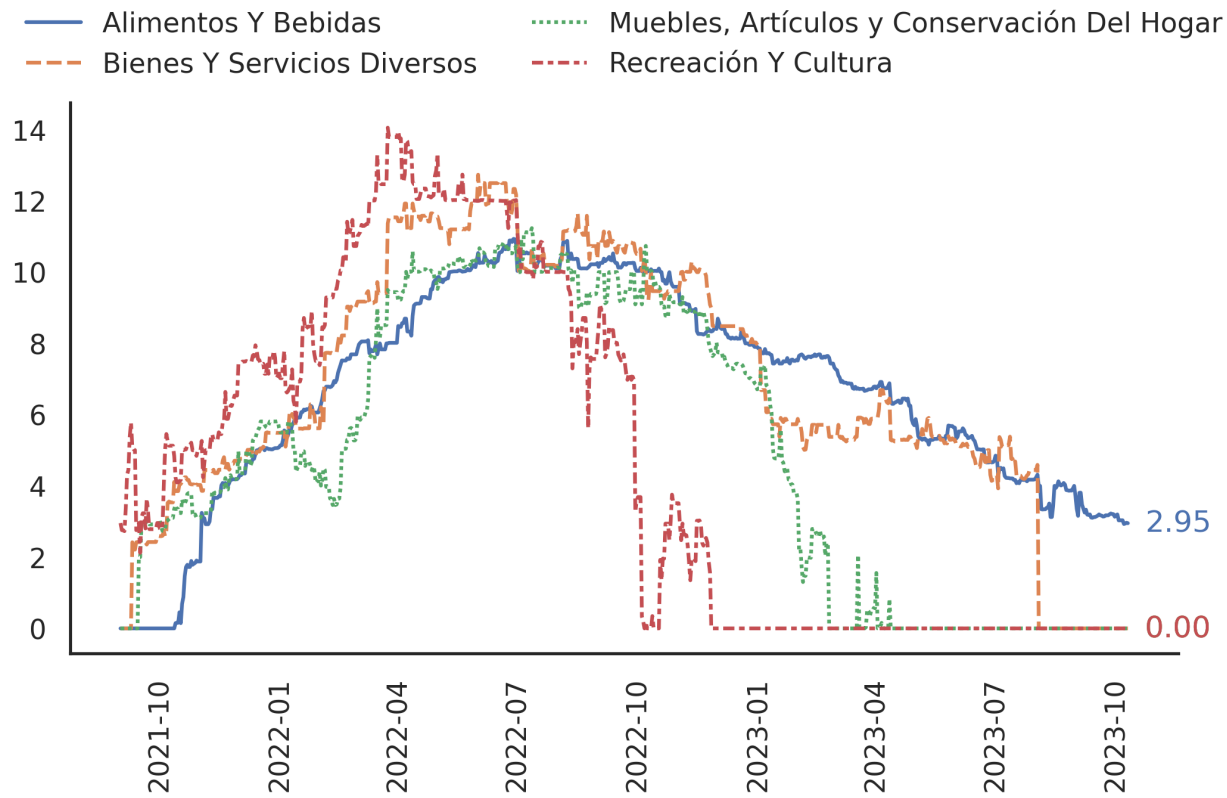


Figura: Medianas sobre la variación interanual.

Distribución de variaciones interanuales

Precios regulares

- El filtro captaría cambios más bruscos a diferencia del suavizamiento de la media móvil.

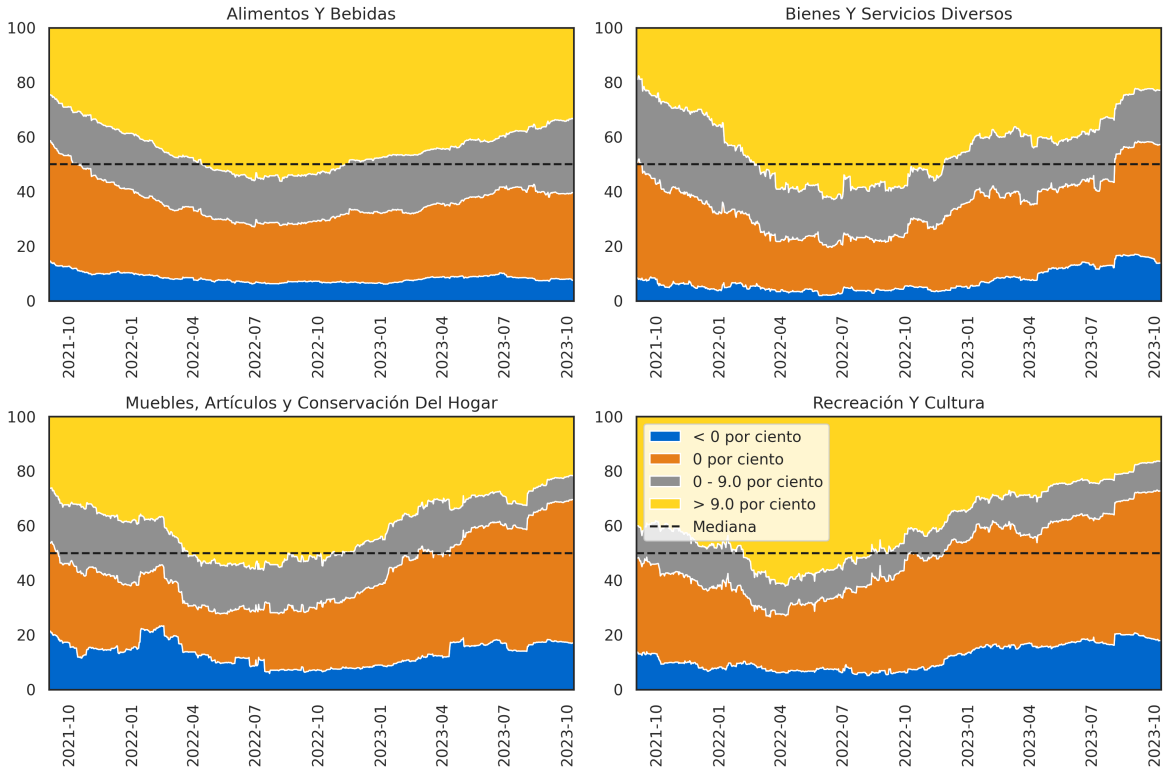


Figura: Distribución por rangos de la variación interanual de precios con filtro de media móvil.

Mediana de variaciones interanuales

Precios regulares

- Se nota con mayor claridad cómo los alimentos perecibles y procesados tienen una similitud en el comportamiento tendencial que se rompe en la segunda mitad de 2022.

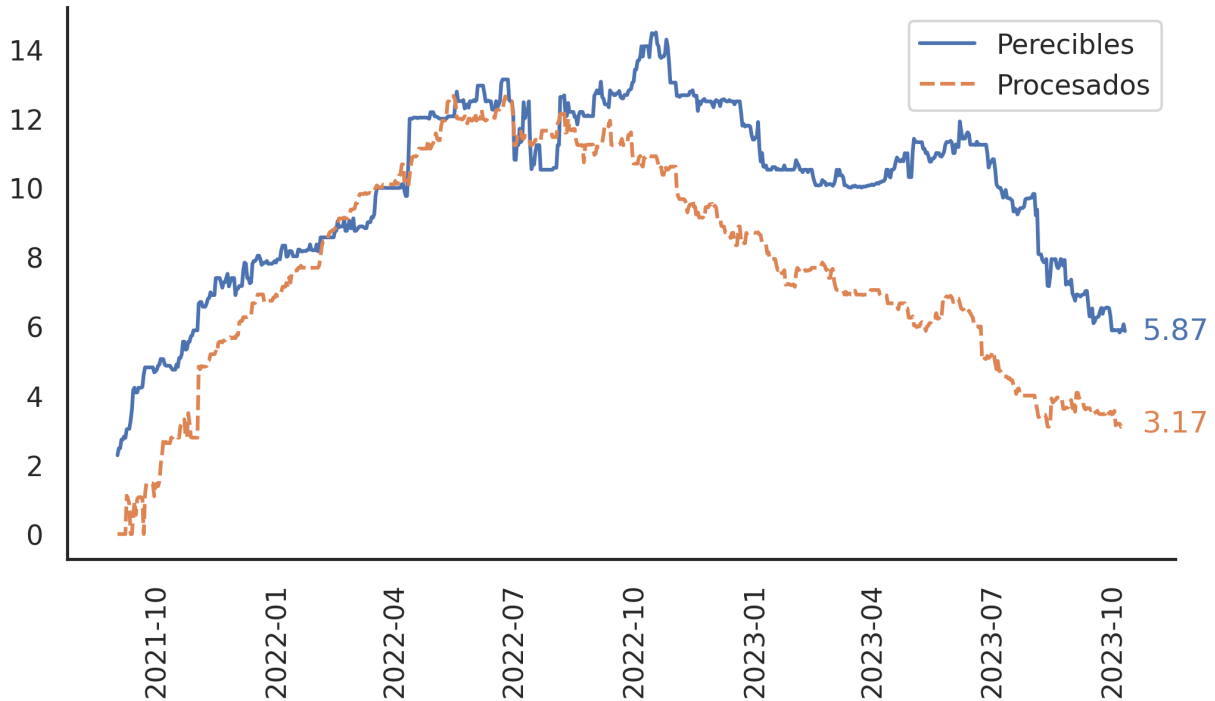


Figura: Medianas sobre la variación interanual.

Frecuencia de actualización

Precios regulares

- La frecuencia de actualización se mide como el número de veces que cambia el precio regular.
- Se incrementó del tercer trimestre de 2020 al cuarto trimestre de 2022.

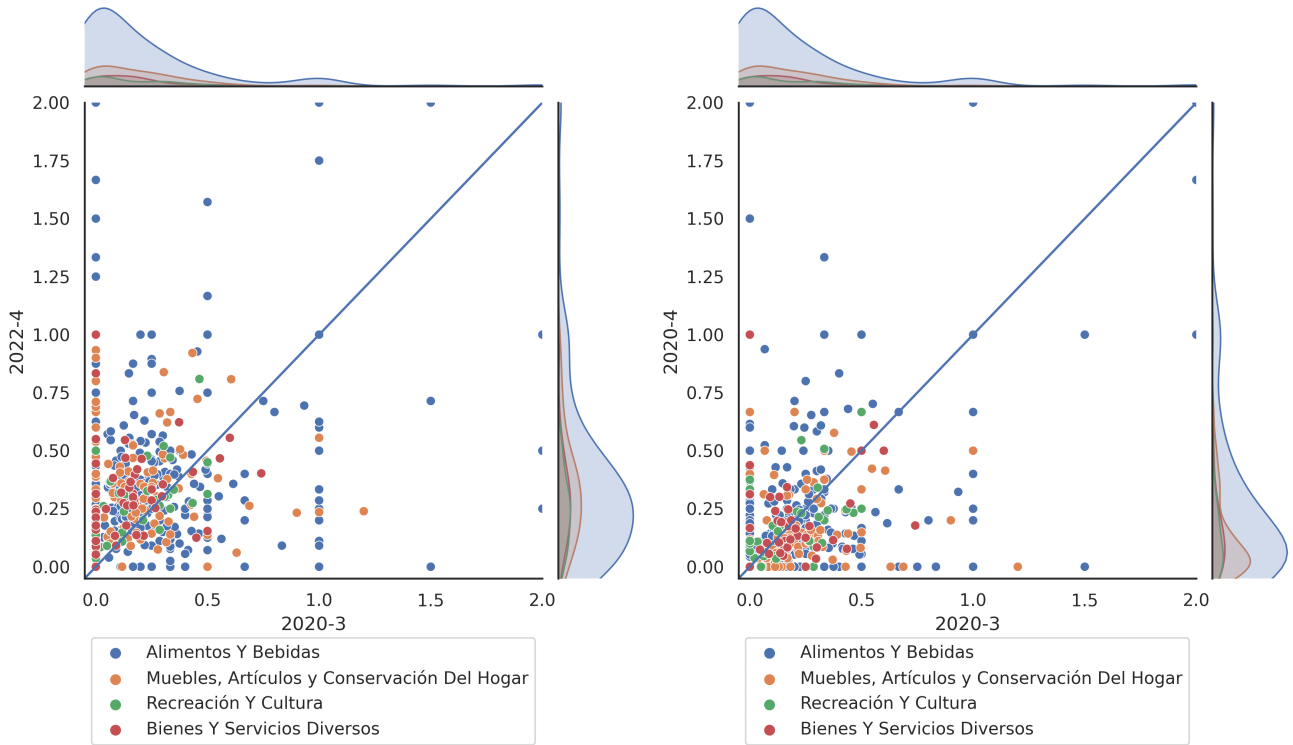


Figura: Frecuencia de actualización por tipo de producto

Contenidos

1. Episodios de inflación y fuentes de información
2. ¿Cuál ha sido el trabajo realizado?
3. Extracción de datos
4. Clasificación
5. Filtros de series
6. Resultados
7. Conclusiones

Conclusiones

1. Base de datos de alta frecuencia de precios que servirá como fuente de información para el monitoreo de la inflación, como también para comprender mejor el proceso de formación de precios.

2. Como inspección inicial se revela el incremento de precios seguido de una moderación tanto para alimentos, bebidas y bienes en 2023. Además, la frecuencia de actualización parece haberse incrementado, lo cual se condice con la noción de diferente comportamiento para distintos regímenes inflacionarios.

Conclusiones

3. Agenda pendiente:

   – Verificación y mejora de clasificación de alimentos y bienes de la muestra de supermercados. La información de bienes sería la primera en su tipo y cobertura en ser explotada para las funciones de seguimiento de los precios.

   – Probar distintas parametrizaciones para el filtro actual de precios como también implementar variaciones.

   – Estudiar a mayor profundidad la duración y actualización de precios, la probabilidad de ocurrencia de estos y cómo reaccionan frente a choques y distintos regímenes de inflación.

Extracción, clasificación y procesamiento de datos de alta frecuencia de supermercados

Gonzalo Bueno y Marco Vega

Banco Central de Reserva del Perú

Las opiniones expresadas en este estudio corresponden a los autores y no deben ser atribuidos al Banco Central de Reserva del Perú.