Uso de web scraping para estadísticas de precios en IBGE

Lincoln T. da Silva, Neimar Guimarães e Vladimir Miranda – IBGE
(Lincoln.silva@ibge.gov.br, neimar.guimaraes@ibge.gov.br, vladimir.Miranda@ibge.gov.br)

Directorio de Encuestas – DPE
Coodinación de Índices de Precios – COINP/GPLACON

Seminário CEPAL:
Innovación e integración de las operaciones estadísticas

5 de abril de 2023

Estructura

i. Aplicaciones en los IPCs de IBGE

ii. Aplicaciones en el PCI en IBGE

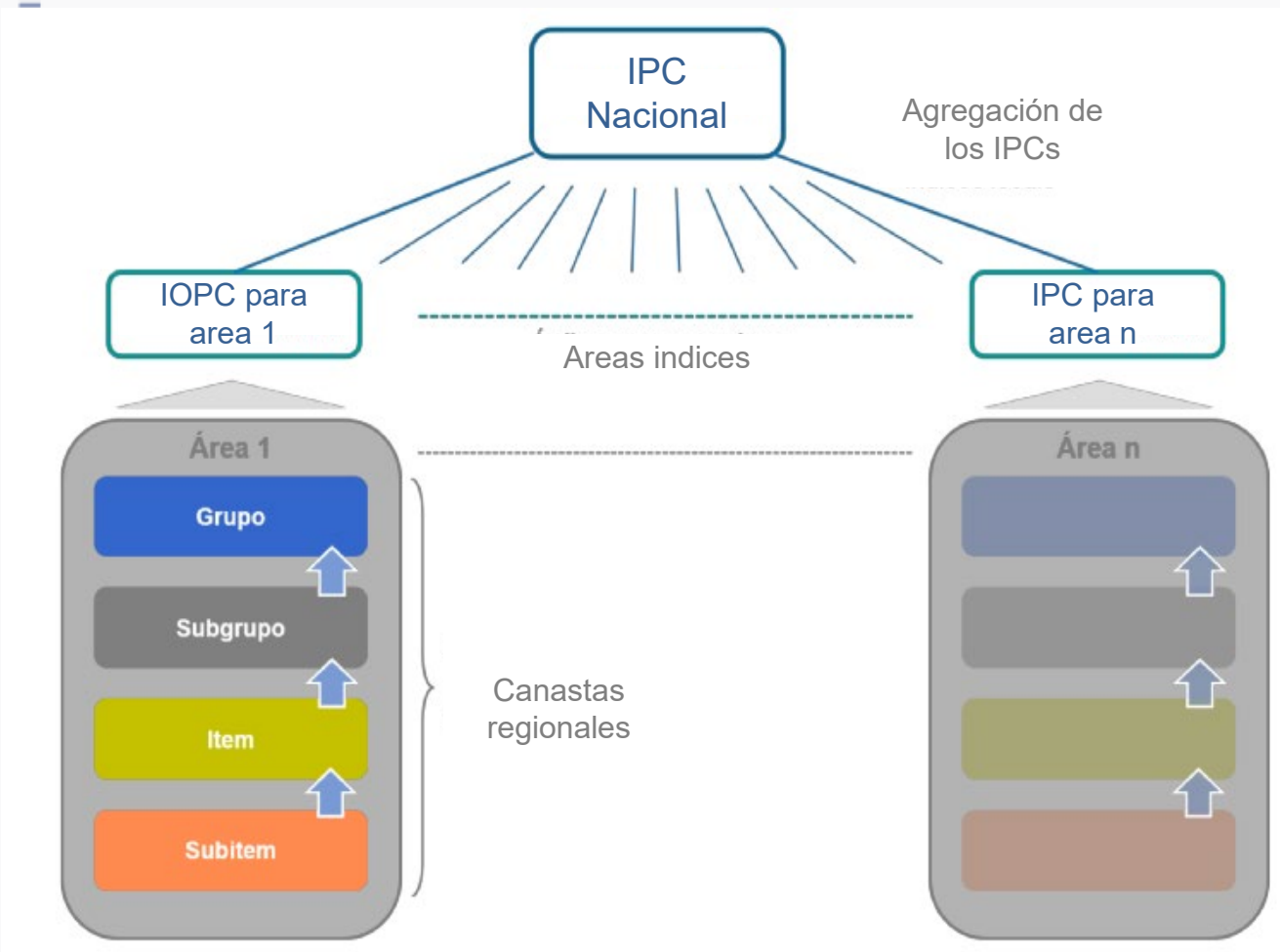
iii. Demo web scraping

Web scraping en los IPCs de IBGE

SNIPC

Indicadores para diferentes poblaciones

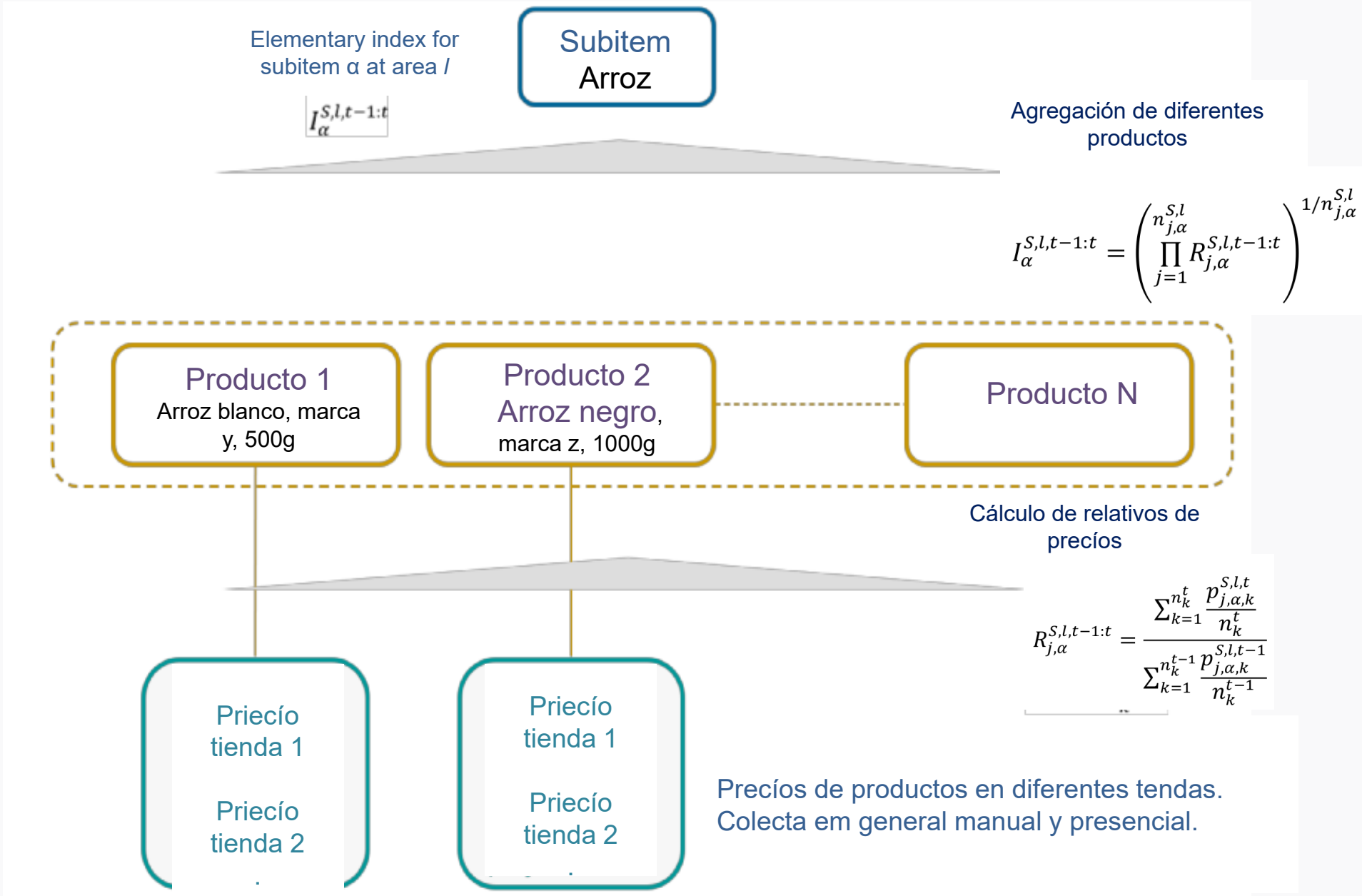
Resultado nacional = resultado agregado de las áreas



Areas de SNIPC.

Inputs principales: precios y pesos

Índices calculados en diferentes niveles. Los precios de produtos son usados en el nível más elementar. Estructura en el SNIPC:



Billetes de avión (colaboración con COMEQ/GDP)

Inputs

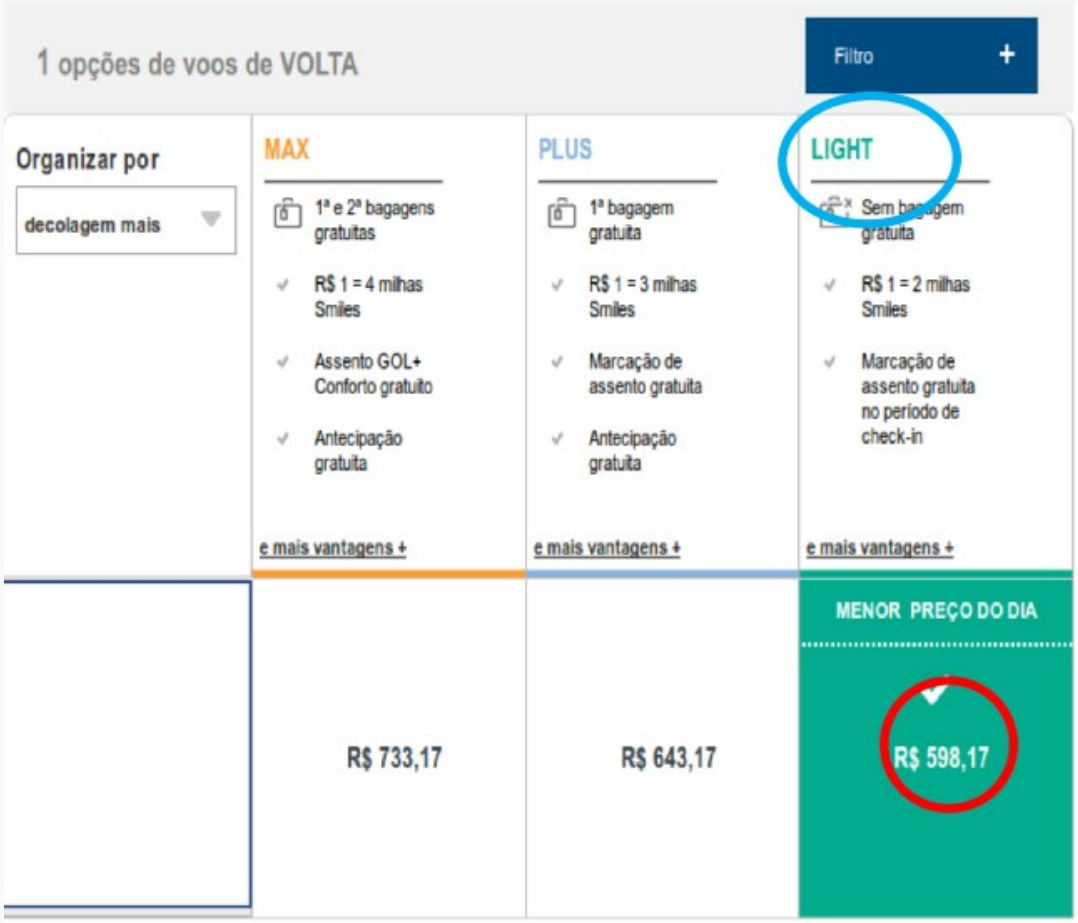
Outputs

En muchos países el comércio de billetes aéreos ocurre a través de los sítios web de las compañias.

Con eso, los precíos de este sector tambíen se extrayen de estos sítios por los recolectores en los INEs de forma manual.

Billetes de avión

Un scraper puede ser desarrollado para estas actividades.

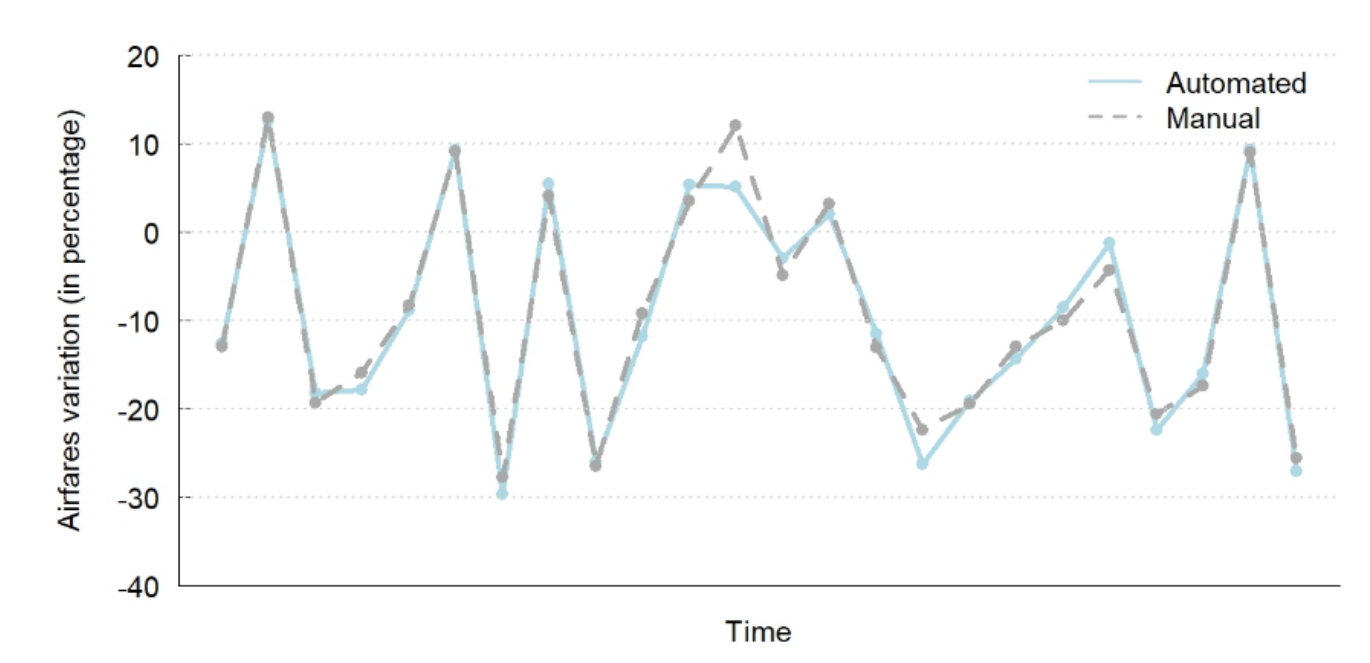



```
</td>
<td class="taxa taxaExecutiva">
  <div>
    <div class="lessPriceBox"></div>
    <div class="taxaSelected">
      <div class="checkTaxaSelected"></div>
    </div><span class="smilesAndMoneyValue"></span><label class="textIdentFareValue" for="

                    R$ 643,17</span><span>tarifa  Plus</span></label><input id="ControlGro
  </div>
</td>
<td class="taxa taxaPromocional">
  <div>
    <div class="lessPriceBox">
      <div class="lessPrice">Menor  Preço do Dia</div>
    </div>
    <div class="taxaSelected">
      <div class="checkTaxaSelected"></div>
    </div><span class="smilesAndMoneyValue"></span><label class="textIdentFareValue" for="

                    R$ 598,17</span><span>tarifa  Light</span></label><input id="ControlGr
  </div>
</td>
<td>
  <div id="market2_journey1" class="infoGrid bgGrid popupANAC"><span class="anacInformatio
              </span></div>
</td>
</tr>
```

Billetes de avión

Scrapers desarrollados en casa en R para las compañías en el muestreo.

Comparación de los resultados manuales y de los scrapers.



da Silva et al, paper presented at the Ottawa Group meeting in 2019.

Scrapers implementados en producción reducen la carga manual de recolección de millares de precios todos los meses.

Proceso similar también adotado para recolección de precios de billetes de avión en el Programa de Comparación Internacional (PCI) para rutas internacionales. Sítios agregadores también són utilizados.

Servicios de viajes por apps

Resultados de la encuesta de presupuestos familiares.

Area	IPCA Taxi	IPCA Ride sharing Services	INPC Taxi	INPC Ride sharing Services
BR	**0,21**	**0,21**	**0,16**	**0,15**
AC	0,54	-	0,55	0,07
PA	0,43	-	0,32	-
MA	0,32	0,11	0,41	0,15
CE	0,18	0,15	0,15	0,16
PE	0,30	0,32	0,15	0,28
SE	0,58	0,11	0,53	0,17
BA	0,38	0,30	0,19	0,21
MG	0,24	0,19	0,17	0,16
ES	0,12	0,10	-	0,09
RJ	0,45	0,31	0,20	0,26
SP	0,16	0,20	0,11	0,12
RS	0,26	0,38	0,20	0,27
MS	0,09	0,23	-	0,28
GO	-	0,26	-	0,09
DF	-	0,25	0,11	0,16

Desafíos: ¿qué recolectar, dónde y cómo?

Componentes del precio del servício:

- Componentes “fijas”

  Tarifas base: tarifa por km,
  Tarifa de reserva.

- Componente dinámica

  Multiplicador dinámico

Servicios de viajes por apps

Se pueden desarrollar diferentes enfoques basados en los componentes del precio considerados.

Si solo se toman las componentes “fijas" para construir un viaje estándar, esto nos dá un enfoque similar al de los servicios de taxi.

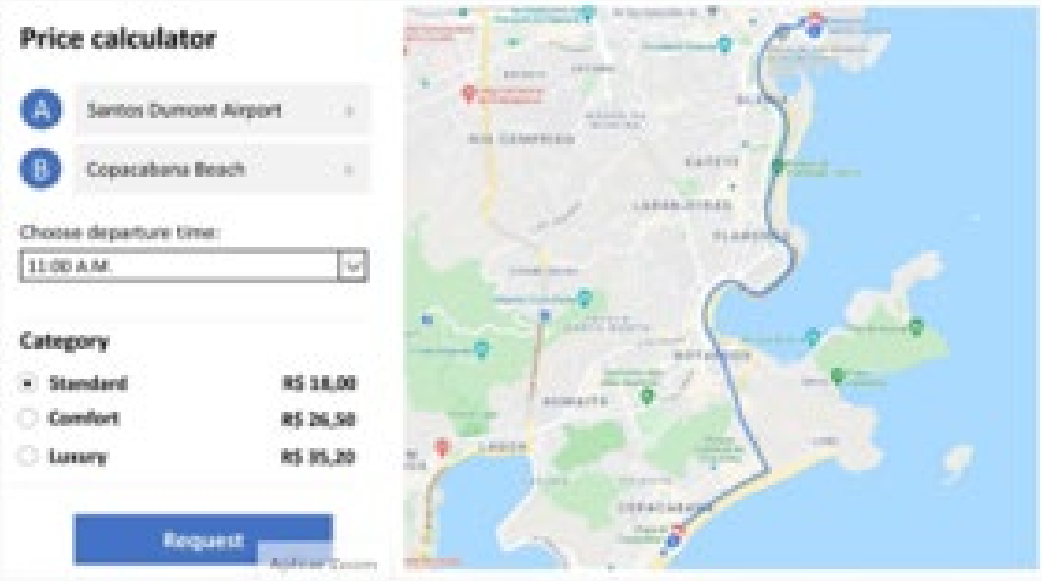
Precio = (tarifa por km) x typical distance + tarifa de reserva

Pero la marcada dinámica de este sector lo acerca a las estrategias de precios adoptadas por las tarifas aéreas y eso no se capta con esta solución.

No seria posible captar diferencias regionales.

¿Es posible derivar un enfoque que considera la dinámica y las diferencias regionales?

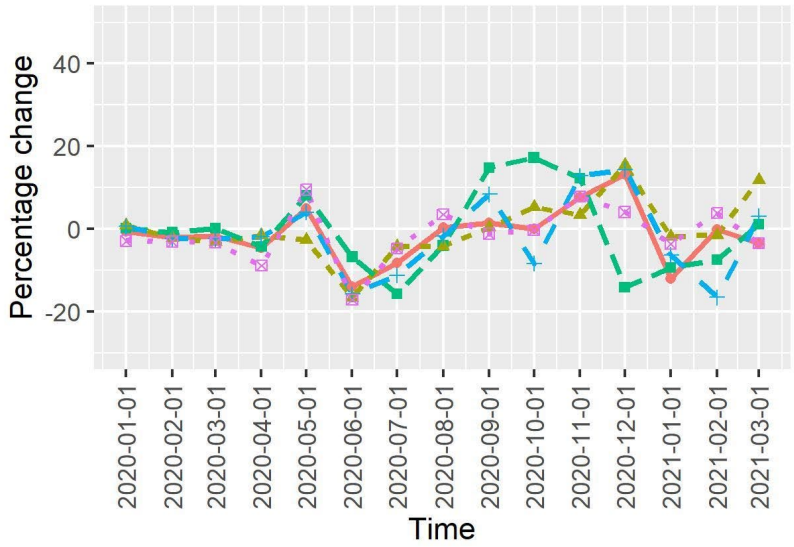
Web permite una alternativa.

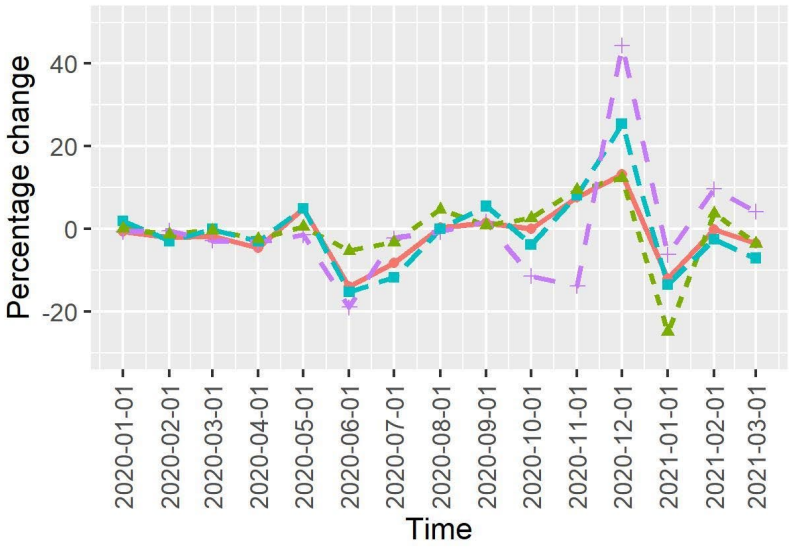


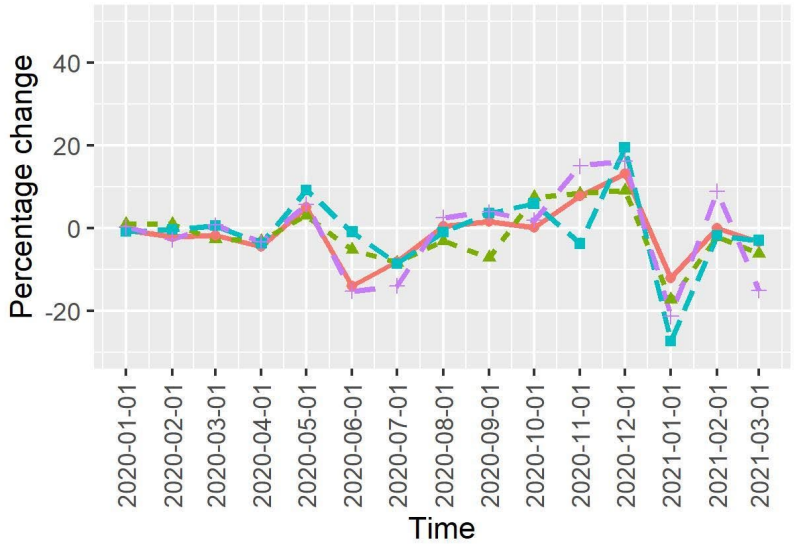
Servicios de viajes por apps

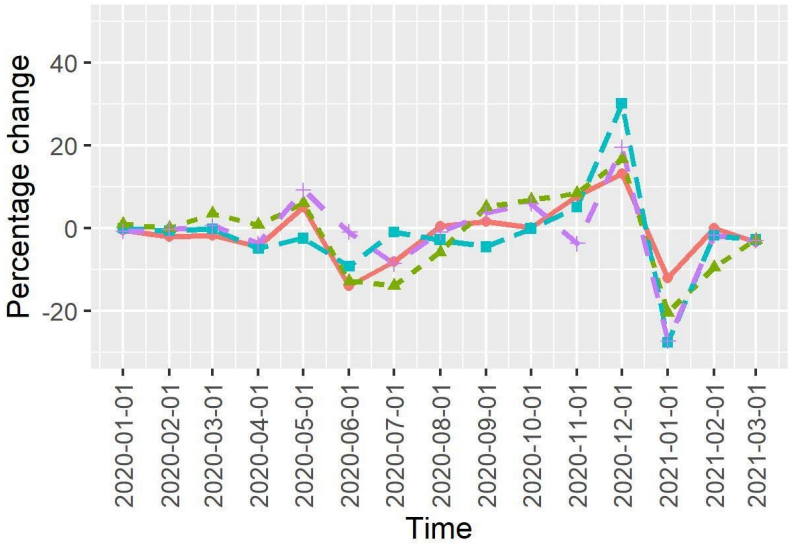
En producción desde 2020.

Resultados pueden captar la dinámica de precios y las diferencias regionales.









Otros setores en desarollo

Algunas características de los casos presentados:

1) Sectores son más monopolizados.

2) Número pequeño de sítios web con cobertura representativa del sector.

3) Reducción significativa del esfuerzo de recoleccióon manual. Millares de precios todos los meses.

4) Sin grandes demandas iniciales de cambios metodológicos para su uso.

5) Cobertura geográfica con la misma desagregación usada en nuestro IPC.

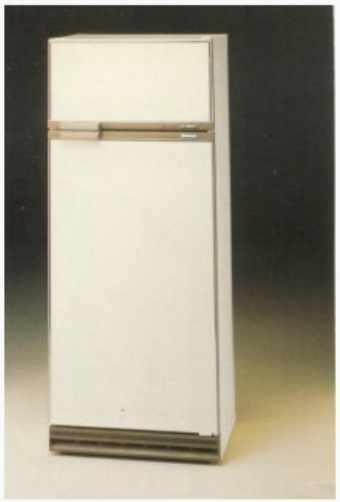
6) Prácticas de comercialización significativas en la web.

Otros sectores en desarollo con algunas características similares:

- Hoteles
- Autos usados
- Autos nuevos
- Alquiler de autos

Ajuste de calidad: Electrodomésticos y electrónicos

Evolución de productos en el tiempo.



Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n				p_n^{t+3}	p_n^{t+4}

$$R_n^{t+3,t+2} = p_n^{t+3} / p_m^{t+2}$$

Comparación directa puede llevar a sesgos.
¿Cómo medir un cambio de precio puro?

Modelos hedónicos son una herramienta propuesta para eso

$$p = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + \epsilon$$

Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n			$\hat{p}_n^{t+2}$	p_n^{t+3}	p_n^{t+4}

Comparación después del ajuste.

$$R_n^{t+3,t+2} = p_n^{t+3} / \hat{p}_n^{t+2}$$

Ajuste de calidad: Electrodomésticos y electrónicos

Los atributos y precios de los productos se pueden raspar en los sitios web con gran control y de a bajos costos.

Geladeira/Refrigerador Frost Free cor Inox 310L Electrolux (TF39S) 127V

Marca: Electrolux

★★★★½ 24 avaliações de clientes

R$2.804^{00}

Em até 10x R$ 280,40 sem juros Ver parcelas disponíveis

Total Capacity	24.52 cubic feet
Refrigerator Style	Side-by-Side
Ice Maker	Yes
Lighting Type	LED
Color Finish	Stainless steel

Ejemplo de ajuste y salida del modelo:

$$log(\mathrm{Pr}) = \beta_0 + \beta_1 \mathrm{Br} + \beta_2 \mathrm{Col} + \beta_3 \mathrm{Sty} + \beta_4 \mathrm{Defr} + \beta_5 \mathrm{Cap} + \beta_6 \mathrm{Shop}$$

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.592e+00  2.905e-02 226.935  < 2e-16 ***
BrConsul       -1.619e-01  1.486e-02 -10.896  < 2e-16 ***
BrElectrolux   -4.476e-02  1.106e-02  -4.046 5.78e-05 ***
ColInox         1.003e-01  1.126e-02   8.909  < 2e-16 ***
StyDuplex       1.166e-01  1.717e-02   6.791 2.35e-11 ***
StyInverse      2.210e-01  2.212e-02   9.991  < 2e-16 ***
DefrFrost Free  1.615e-01  1.045e-02  15.445  < 2e-16 ***
Cap             2.684e-03  6.284e-05  42.707  < 2e-16 ***
ShopOnline     -1.094e-01  8.593e-03 -12.736  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1001 on 713 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8832
F-statistic: 682.5 on 8 and 713 DF,  p-value: < 2.2e-16
```

Aplicación en el PCI en IBGE

Estructura

i. Contexto del PCI 2021

ii. Cobertura y aplicación

iii. Resultados y oportunidades de mejorías

PCI 2021: El contexto

El PCI hace uso de precios de un catálogo de productos enviados a los países para construir los indicadores PPPs.

11.05.61.1.01.05: *Detergente de lavavajillas, MC*

Tipo de marca: Conocida
Minimum quantity: 500
Cantidad máxima: 1200
Unidad de medida Mililitros
Tipo: Detergente para lavavajillas concentrado
Forma Líquido
Embalaje: Botella plástica
Excluir: Productos especiales para pieles sensibles.
Especificar: Marca
Cantidad de referencia: 500
Unidad de medida: Mililitros

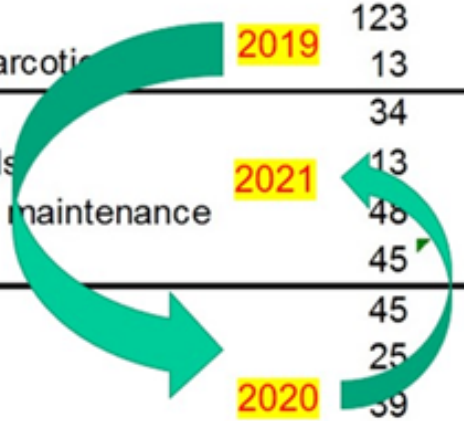
- A menudo, las especificaciones del catálogo son muy cerradas y restrictivas con la finalidad de garantizar la comparabilidad de los precios por medio de una canasta común de bienes y servicios entre los países.

- Webscraping: cumplir los requerimientos establecidos por el catálogo del PCI.

PCI 2021: El contexto

- Compartir los mismos recursos de recolección (profesionales, sistemas de informática) entre nuestro IPC y el PCI;
- Esfuerzo adicional: durante la ronda del PCI 2017, 66% de los productos eran de recolección exclusiva para el PCI;

	2021 List of Products	Number of Products	
01	Food and non-alcoholic beverages	123	
02	Alcoholic beverages, tobacco, and narcoti	13	136
03	Clothing and footwear	34	
04	Water, electricity, gas, and other fuels	13	
05	Furnishings, household equipment & maintenance	46	
06	Health	45	140
07	Transport	45	
08	Communication	25	
09	Recreation and culture	39	
11	Restaurants and hotels	16	
12	Miscellaneous goods and services	22	147
	Total	423	



- Antes de la pandemia: Acordada la recolección de 1/3 de artículos entre los años de 2019 a 2021.
- Pandemia: incertidumbre con el fin de la rotación para la recolección de los artículos del PCI 2021. De hecho, tuvimos la recolección de 2/3 del panel de artículos. Oportunidad: ampliar el uso del Webscraping.

PCI 2021: Cobertura y aplicación

- Programas internos desarrollados para web scraping en páginas web de minoristas (33 productos), concesionarios de autos (2 carros) y compañías aéreas (5 tipos de vuelos). Recolección del webscraping hecha entre abril y diciembre de 2021

Encabezado Básico del PCI	Artículo
Bienes domésticos no durables	Detergente en polvo, Limpiador doméstico, Insecticida, Vela
	Papel de cocina, Servilleta de papel, Detergente de lavavajillas
Aparatos domésticos mayores	Microondas, Cocina a gas con 4 quemadores
Aparatos y equipos terapéuticos	Monitor automático de presión arterial
Equipos audiovisuales, fotográficos	Cámara CANON, Laptop LENOVO, Software MS Office 2019
	Televisión, LED, 43", LG/SAMSUNG, Computador portátil, APPLE/ HP Pavilion
Medios grabadores	USB memoria flash, 32 GB; Disco duro externo, 2 TB
Otros items y equipos recreativos	Pelotas de tenis WILSON; Pelota de fútbol, MC.
	Muñeca Barbie, MATTEL; Paquete de 52 carta de naipe
Joyas, relojes y relojes pulseras	Reloj de pulsera para niños/hombre
Otros efectos personales	Carrito maleta SAMSONITE, Paraguas de bastón/plegable
Mantención y reparación de la vivienda	Azulejos, cerámicas
Equipo de teléfono y fax	Teléfono inteligente, XIAOMI Redmi 9
Aparatos, artículos y productos para el cuidado personal	Máquina de afeitar eléctrica, recargable, PHILIPS
Equipo de teléfono y fax	Teléfono inteligente, XIAOMI Redmi 9
Household textiles	Toalla de baño, algodón
Transporte aéreo de pasajeros	Vuelos internacionales
Automóviles	Toyota Yaris/Corolla

- Precios recolectados (recolección tradicional-manual + webscraping): 189.872 precios validados para la ronda 2021;

- Registros de Webscraping válidos: 25.599 (13,4% de todos los precios validados);

- Cobertura de 15 Encabezados (17,2% en la encuesta de consumo de hogares).

PCI 2021: Cobertura y aplicación

Ajustes contínuos en los programas para cumplir con los requerimientos:

11.05.61.1.01.05: *Detergente de lavavajillas, MC*

Tipo de marca: Conocida
Minimum quantity: 500
Cantidad máxima: 1200
Unidad de medida Mililitros
Tipo: Detergente para lavavajillas concentrado
Forma Líquido
Embalaje: Botella plástica
Excluir: Productos especiales para pieles sensibles.
Especificar: Marca
Cantidad de referencia: 500
Unidad de medida: Mililitros

- Es posible filtrar palabras claves para identificar los productos elegibles. En el ejemplo del detergente, son excluidos los productos con los términos “Kit” (paquetes con precios para varios productos) y “glicerina” (sustancias usadas en productos de pieles sensibles) en sus descripciones.

Producto en revisión	Precio	Aprobado
Kit Com 06 Detergentes Limpol Cristal 500Ml	18,35	NO
Detergente Limpol Neutro Com Glicerina 500Ml	8,15	NO
Detergente 500 Ml Limpol Neutro	2,16	Sí
Detergente Líquido Neutro Limpol 500ml	1,99	Sí

PCI 2021: Resultados y oportunidades de mejorías

Artículo	Coeficiente de variación (% antes)	Coeficiente de variación (% después)	Precio promedio (antes)	Precio promedio (después)	Número de precios (antes)	Número de precios (después)
Detergente de lavavajillas	98,3	10,3	6,99	2,34	418	130
Pelotas de tenis, multipack	155,8	16,0	95,35	66,37	224	205
Paquete de 52 cartas	96,4	17,1	37,98	15,10	29	19
Paraguas de bastón	170,9	32,5	311,35	45,36	215	15
Servilleta de papel	140,5	-	225,85	-	112	0
.						
.						
Total					10.894	4.626
PCI: Julio 2021						

1) Gran número de registros recolectados mensualmente. Sin embargo, ni todos fueron efectivamente aprovechados (ejemplo encima, 42% en julio de 2021). Muchos registros no cumplen con los requerimientos del catálogo;

•2) Puntos positivos: alivio para el equipo de recolección, tiempo adicional para desarrollar otros proyectos (revisión de las encuestas especiales del PCI, PPPs subnacionales);

•3) Puntos de atención: Se necesita tener profesionales envueltos en la revisión continua de los datos (crítica estadística); Ampliar el uso y cobertura; Con cambios en las páginas web de referencia, es necesario el mantenimiento continuo de los programas (robots en R, Python).

Demo web scraping

Web scraping en la práctica

Sitio web del que queremos extraer información

Enfoque dirigido vs bulk

Sitio web dinámico (javascript) vs sitio web estático

Ejemplo

https://www.kayak.com/

Ejemplo

Enfoque dirigido

Busca de vuelos con origen, destino, fichajes de salida e vuelta y otros.

Ejemplo:

Vuelo: EZE (Buenos Aires) – GIG (Rio de Janeiro)
Fechas: 2023/04/13 - 2023/04/21

Ejemplo

Sitio dinámico

Configuración ->
Privacidad y seguridad ->
Configuración de sitios ->
Javascript

Si no permitimos que el sitio use Javascript, no se cargan los datos.

Hay que emular el comportamiento humano.

Ejemplo

El enlace del sitio web facilita el trabajo.

la consulta se puede cambiar directamente a través del enlace del sitio web

Mostrar los precios más bajos primero

Ejemplo

Como extraer los datos del sitio web dinámico? RSelenium

Identificar los elementos que queremos extraer. **Inspeccionar**.

Podemos usar **css selector** o **xpath**.

¿Cómo saber si el sitio web se ha cargado completamente?

```
<div id=”B5T6-advice” class=”value “ aria-busy=”false”>
        Buy now
</div>
```

Xpath = ‘//div[@aria-busy]’ > element attribute

Ejemplo

Precio

```
<div class=”f8F1-price-text”>
        $522
<\div>
```

css selector = ‘.f8F1-price-text’ > element text

Taller de práctica en web scraping

Introducción al web scraping aplicado a índices de precios

Gracias por la atención!