Exchange Rates, Prices, and Trade

Alberto Cavallo, Associate Professor, Harvard Business School

Price Scraping and the Billion Prices Project

Sunday, August 30, 2020
Big Data in Macro and International Economics

- Macro data have many problems (Eichengreen (2015), Summers (2016))

- We rely mostly on data collected by governments → Griliches (1985) The Uneasy Alliance

“... we have shown little interest in improving it [the data], in getting involved in the grubby task of designing and collecting original data sets of our own. Most of our work is on “found” data, data that have been collected by somebody else, often for quite different purposes... “They” collect the data and are responsible for all their imperfections. “We” try to do the best with what we get, to find the grain of relevant information in all the chaff.”

- Can Big Data help? (Einav & Levin (2014))
Big Data in Macro and International Economics

- New data sources in recent years
  - Administrative data (e.g. CPI and IPI micro data, tax & property records)
  - Scanner data (e.g. Nielsen)
  - Search data (e.g. Google, Indeed)
  - Satellite data (e.g. lights, parking lots, tanker and crop heights)
  - Sensor data (smart phones, smart watches, IOT devices)
  - Crowd-sourced data (web, mobile phones)
  - Online data (e.g. Billion Prices Project)

- The 5 Vs and the origin of the BPP
Big Data

Volume
Data Size

Velocity
Speed of Change

Variety
different Forms of Data Sources

Veracity
Uncertainty of Data

Value
business value

TRUTH
DECEPTION IN ARGENTINA, 2007-2015

Government attempts to identify retailers sampled
National statistics office “intervened”

OFFICIAL CPI

PERCEIVED INFLATION
Source: Di Tella University
ONLINE PRICES PROVIDED AN ALTERNATIVE WAY TO COLLECT PRICE DATA

<html>
<descripcion> Leche Condensada </descripcion>
<br> <brand> Nestlé </brand>
<br> <td price> $1.199 Uni </td>
ONLINE INFLATION IN ARGENTINA

Online indices matched CPIs in other countries

Fig. 1. Online and official indexes in four Latin American countries: (a) Brazil; (b) Chile; (c) Colombia; and (d) Venezuela. Notes: The daily online supermarket index is constructed with an online prices and official CPI category weights. In Venezuela, the online data has no category information, and therefore the online index is built as a geometric average of all price changes observed each day. The official supermarket index is an equivalent indicator constructed by multiplying individual price changes by their own category price levels.
IS IT ONLINE DATA?

Average annual inflation (%) between 2007–2011

- Venezuela: 27.4%
- Colombia: 27.4%
- Chile: 3.8%
- Brazil: 4.9%
- Argentina: 8.5%


11.6% difference
Difference was not explained by data source, method, or type of goods

WHAT HAPPENS IF WE DIVIDE BY 3?

By underestimating inflation they overestimated growth

Statistical Uncertainty

- By underestimating inflation they overestimated growth
Los Precios Hoy

Somos un grupo de economistas que todos los días registra los precios de un listado de productos en dos grandes supermercados de Buenos Aires y los publica en esta sitio.

En esta sección mostramos un resumen de nuestras principales estadísticas, basadas en la evolución diaria de dos índices propios de inflación:

- Un Índice de Alimentos y Bebidas, equivalente al 31% del IPC.
- Un Índice de la Canasta Básica Alimentaria, como el que el INDEC usa (o usaba) para calcular el nivel de indigencia.

En la sección *Noticias* hace un análisis periódico de los datos. En las secciones de *Precios*, *Aumentos* y *Galería* se pueden ver detalles de precios individuales. Finalmente, en *Metodología* y *Preguntas* explicamos cómo los hacemos y cómo pedir ayuda.

**Aumentos de Hoy (%)**

- FIDEOS SECOS M: 0%
- POLIZA CONZELADA: 2.7%
- NAELSA: 2.6%
- TAPA DE AJIADO LOMILLERO: 3.4%
- NARANJA DE JUÍGO: 26.06%

**Aumentos Esta Semana (%)**

- ZARAHORA: 2.41%
- BATATA: 20.23%
- CEBOLA: 24.95%
- HUEVOS BLANCOS GRANDES: 23.03%

*Incluye precios que entran y salen de oferta*

**Inflación Anual**

- %: 23.7
- %: 25.4

*Canasta Básica*

**www.InflacionVerdadera.com**
DON’T LIE TO ME, ARGENTINA

“...for the vast majority of Argentines, the indices compiled by INDEC...have been grossly manipulated.”

— La Nación Newspaper, March 10, 2008
DON'T LIE TO ME, ARGENTINA

“...we have decided to drop INDEC’s figures entirely...”

—The Economist, February 25, 2012
DON’T LIE TO ME, ARGENTINA

“… a declaration of censure… the Board called Argentina to… address the inaccuracy of the CPI.”

—IMF, February 2013
EVENTUALLY GOVERNMENT TELLS THE TRUTH

Annual Inflation

Source: PriceStats, INDEC, The Billion Prices Project
THE BILLION PRICES PROJECT: MEASUREMENT AND RESEARCH
The Billion Prices Project

- Academic initiative to collect and use online price data for economic measurement and research
  - Daily prices since 2008
  - From hundreds of large multi-channel retailers
  - In over 60 countries

![Workflow Diagram]

1. Use scraping technology
2. Connect to hundreds of online retailers every day
3. Find individual products
4. Store key item information in a database
5. Calculate inflation and other statistics

www.thebillionpricesproject.com
Daily Inflation Measurement

- 2008 → Daily online index for Argentina (www.inflacionverdadera.com)
- 2010 → Daily index for the US on the BPP website
- 2011 → PriceStats collects the micro data and publishes daily inflation in 22 countries in real-time.

Source: www.pricestats.com
Real-time Inflation Measurement

*Figure 2*
*United States*

A: Price index

Real-time Inflation Measurement

Figure 3
US Consumer Price Index around the Bankruptcy of Lehman Brothers

Source: Authors using online price index computed by PriceStats and the Consumer Price Index from the US Bureau of Labor Statistics.
Note: The figure highlights the events around the bankruptcy of Lehman Brothers, the fourth-largest investment bank in the United States, during September 2008.

Micro-Price Data Sources: Advantages and Disadvantages

Table 1: Alternative Micro-Price Data Sources

<table>
<thead>
<tr>
<th></th>
<th>Online Data</th>
<th>Scanner Data</th>
<th>CPI Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost per observation</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Data Frequency</td>
<td>Daily</td>
<td>Weekly</td>
<td>Monthly</td>
</tr>
<tr>
<td>All Products in Retailer (Census)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Uncensored Price Spells</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Countries with Research Data*</td>
<td>~60</td>
<td>&lt;10</td>
<td>~20</td>
</tr>
<tr>
<td>Comparable Across Countries</td>
<td>Yes</td>
<td>Limited</td>
<td>Limited</td>
</tr>
<tr>
<td>Real-Time availability</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Product Categories Covered</td>
<td>Few</td>
<td>Few</td>
<td>Many</td>
</tr>
<tr>
<td>Retailers Covered</td>
<td>Few</td>
<td>Few</td>
<td>Many</td>
</tr>
<tr>
<td>Quantities or Expenditure Weights</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: Table from Cavallo (2015b). *Approximate numbers. The Billion Prices Project (bpp.mit.edu) datasets contain information from over 60 countries with varying degrees of sector coverage. Nielsen US scanner datasets are available at the Kilts Center for Marketing of the University of Chicago. Klenow and Malin (2010) provide stickiness results with CPI data sourced from 27 papers in 23 countries. See Cavallo (2013) for more details.

## The Billion Prices Project

<table>
<thead>
<tr>
<th>Research Topic</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Macroeconomics</strong></td>
<td></td>
</tr>
<tr>
<td>Inflation Measurement</td>
<td>Cavallo (2020a)</td>
</tr>
<tr>
<td>Online Price Indices</td>
<td>Cavallo (2013), Cavallo &amp; Rigobon (2016)</td>
</tr>
<tr>
<td>Crowdsourcing with phones in Venezuela</td>
<td>Cavallo (2020b)</td>
</tr>
<tr>
<td>Online Pricing and Price Discrimination</td>
<td>Cavallo (2017), Cavallo (2018a)</td>
</tr>
<tr>
<td>Price Stickiness</td>
<td>Cavallo (2018b)</td>
</tr>
<tr>
<td><strong>International Economics</strong></td>
<td></td>
</tr>
<tr>
<td>Purchasing Power Parities</td>
<td>Cavallo, Diewert, Feenstra, Inklaar &amp; Timmer (2018)</td>
</tr>
<tr>
<td>International Wage Comparisons</td>
<td>Cavallo, Cravino, &amp; Drenik (2019)</td>
</tr>
<tr>
<td>Tariff passthrough (next session)</td>
<td>Cavallo, Neiman, Gopinath, &amp; Tang (2020)</td>
</tr>
</tbody>
</table>
Are Online Prices Relevant?

- Cavallo (2017) Are Online and Offline Prices Similar: Evidence from Large Multi-Channel retailers, American Economic Review

- Large-scale comparison of online and offline prices collected simultaneously in ~50 retailers in 10 countries.

- Crowdsourced workers scan random barcodes, enter prices, send emails with data files.

- We then scraped the online price for the same good-retailer (within 7 days).
Prices are identical ~70% of the time

### Table 3: Country - Level Differences

<table>
<thead>
<tr>
<th>Country</th>
<th>Ret.</th>
<th>Obs</th>
<th>Identical (%)</th>
<th>High On (%)</th>
<th>Low On (%)</th>
<th>Markup (%)</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>5</td>
<td>3699</td>
<td>60</td>
<td>27</td>
<td>13</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Australia</td>
<td>4</td>
<td>3797</td>
<td>74</td>
<td>20</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Brazil</td>
<td>5</td>
<td>1915</td>
<td>42</td>
<td>18</td>
<td>40</td>
<td>-7</td>
<td>-4</td>
</tr>
<tr>
<td>Canada</td>
<td>5</td>
<td>4031</td>
<td>91</td>
<td>3</td>
<td>5</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>China</td>
<td>2</td>
<td>513</td>
<td>67</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Germany</td>
<td>5</td>
<td>1604</td>
<td>74</td>
<td>4</td>
<td>23</td>
<td>-8</td>
<td>-2</td>
</tr>
<tr>
<td>Japan</td>
<td>4</td>
<td>2186</td>
<td>40</td>
<td>7</td>
<td>45</td>
<td>-13</td>
<td>-7</td>
</tr>
<tr>
<td>South Africa</td>
<td>5</td>
<td>3212</td>
<td>85</td>
<td>6</td>
<td>9</td>
<td>-3</td>
<td>-1</td>
</tr>
<tr>
<td>UK</td>
<td>4</td>
<td>2094</td>
<td>91</td>
<td>2</td>
<td>7</td>
<td>-8</td>
<td>-1</td>
</tr>
<tr>
<td>USA</td>
<td>17</td>
<td>15332</td>
<td>72</td>
<td>8</td>
<td>22</td>
<td>-5</td>
<td>-1</td>
</tr>
</tbody>
</table>

### Table 4: Sector - Price Level Differences

<table>
<thead>
<tr>
<th>Sector</th>
<th>Ret.</th>
<th>Obs</th>
<th>Identical (%)</th>
<th>High On (%)</th>
<th>Low On (%)</th>
<th>Markup (%)</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>10</td>
<td>5953</td>
<td>52</td>
<td>32</td>
<td>15</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Clothing</td>
<td>7</td>
<td>2534</td>
<td>92</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Household</td>
<td>9</td>
<td>7875</td>
<td>79</td>
<td>5</td>
<td>16</td>
<td>-8</td>
<td>-2</td>
</tr>
<tr>
<td>Drugstore</td>
<td>4</td>
<td>3053</td>
<td>38</td>
<td>11</td>
<td>52</td>
<td>-5</td>
<td>-3</td>
</tr>
<tr>
<td>Electronics</td>
<td>5</td>
<td>3712</td>
<td>83</td>
<td>4</td>
<td>13</td>
<td>-9</td>
<td>-1</td>
</tr>
<tr>
<td>Office</td>
<td>2</td>
<td>1089</td>
<td>25</td>
<td>37</td>
<td>38</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Multiple/Mix</td>
<td>18</td>
<td>14149</td>
<td>80</td>
<td>5</td>
<td>15</td>
<td>-9</td>
<td>-2</td>
</tr>
</tbody>
</table>

Note: Results updated 5 Apr 2016. Markup excludes identical prices. Difference includes identical prices.

Practice some web scraping

- It is not as hard as you think:
  - All webpages use an HTML code with tags that provide a stable structure to identify the data → you can teach a software to recognize them
  - Many alternative tools: R, python, specialized software, scraping services

- Similar Steps:
  1) Create a template for parsing the HTML code
  2) Create a list of URLs with the relevant data
  3) Run scraper, analyze data, debug

- Not just for retail prices!
  - Wage postings (Cavallo, Cravino & Drenik 2019), stockouts (Cavallo & Kryvtsov 2020), real estate listings, customs data, wholesale prices, or any structured data that can be found online


Eichengreen, B., 2015. An Economics to Fit the Facts, Project Syndicate.

