The Billion Prices Project
Using Online Data for Measurement and Research

Alberto Cavallo and Roberto Rigobon
MIT & NBER

Journal of Economic Perspectives - Spring 2016

UNECE CPI Meeting - Geneva 2016
The Billion Prices Project at MIT

- Academic initiative to collect and use online price data for economic measurement and research
- Daily data since 2008 from large multi-channel retailers in over 60 countries

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**

- Date
- Item
- Price
- Description

The Billion Prices Project at MIT

- Uses and applications of online data
  - Measure inflation and other price indicators
  - Research in Macroeconomics
    - Price Dynamics (Price Stickiness, Real Rigidities)
    - Market Segmentation
  - Research in International Economics
    - Pass-through and Border Effects
    - Law of One Price and PPP
    - Real Exchange Rates

Web-Scraping Online Data

- Every day, our software downloads public webpages, scans the HTML code, extracts price data, and stores it in a database.
## Micro-Price Data Sources For Research Applications

### Table 1: Alternative Micro-Price Data Sources

<table>
<thead>
<tr>
<th>Feature</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.

Daily Inflation Measurement

- 2008: daily online index for Argentina (www.inflacionverdadera.com, now also published in The Economist)

- 2010: daily index for the US on the BPP website (bpp.mit.edu)

- Since 2011 PriceStats has been publishing daily inflation indices in 22 countries in real-time. (www.pricestats.com)

Source: www.pricestats.com

US Daily Price Index

US Annual Inflation

Developing vs Developed Countries

(a) China  
(b) Brazil  
(c) Turkey  
(d) Colombia  
(e) UK  
(f) France  
(g) Germany  
(h) Japan  

Figure 5: Online vs CPI Annual Inflation Rates

Sectors vs Global Aggregates

Figure 5: Online vs CPI Annual Inflation Rates

Overlapping Quality Adjustments

- Online data has big data advantages:
  - uncensored spells (automatically included at introduction)
  - all varieties/models on display

Overlapping Quality Adjustments

- Simple indices can approximate the level and trend of CPI inflation in hedonic-adjusted categories

Figure 7: US CPI with Hedonics vs Online Jevons Index
Online vs Offline Prices

- Are online prices representative?

- Online sales are still less than 10% of retail sales in developed countries

- The `online store` is effectively the *largest* store for most retailers. E.g.: Walmart has 4759 stores in the US. The median store has 0.02% of sales. The `online store` has 8% of sales

- The online price behaves similarly to the offline price
  - Close matching in price indices
  - Simultaneous data collection

Which Online Prices?

- We study retailers that sell both online and offline (multi-channel).
Simultaneous Offline and Online Data Collection

- Crowdsourced workers scan barcodes, enter prices, and email data files

![Amazon Mechanical Turk and Elance](image)

- We automatically process and consolidate the offline data

- We use the barcodes to check prices online using customized web scrapers
  - Up to 7 days later

Price Level Comparison by Country

Table 3: Country - Level Differences

<table>
<thead>
<tr>
<th>Country</th>
<th>(1) Ret.</th>
<th>(2) Obs</th>
<th>(3) Identical (%)</th>
<th>(4) High On (%)</th>
<th>(5) Low On (%)</th>
<th>(6) Markup (%)</th>
<th>(7) 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>87</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>48</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>69</td>
<td>8</td>
<td>22</td>
<td>-5</td>
<td>-1</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>56</td>
<td>38383</td>
<td>72</td>
<td>11</td>
<td>18</td>
<td>-4</td>
<td>-1</td>
</tr>
</tbody>
</table>

Note: Results updated 17 Mar 2016. Column 3 shows the percentage of observations that have identical online and offline prices. Column 4 has the percent of observations where prices are higher online and column 5 the percentage of price that are lower online. Column 6, is the online markup, defined as the average price difference excluding cases that are identical. Column 7 is the average price difference including identical prices.

PPP series with Online Data

- Main Characteristics
  - 350 narrow product categories
  - 30 thousand individually matched items
  - 7 countries
  - Food, fuel, and electronics
  - Sector and Country-Level Indices
  - Daily frequency

Illustration: Coke Prices

Compare prices for a bottle of Coke across countries

Repeat for hundreds of products

Compute daily RERs at different levels of aggregation
A country-level RER for tradable goods

Real Exchange Rate

- The relative cost of a large basket of identical goods when expressed in the same currency

\[ RER = \frac{P_{lc} \cdot E}{P_{us}} \]

Components: Prices and E

- Is there a link between Relative Prices and Exchange Rates?
Aggregate Results

Argentina

Real Exchange Rate (RER)

Components of RER

Brazil

Real Exchange Rate (RER)

Components of RER

South Africa

Real Exchange Rate (RER)

Components of RER

Sources: BPP, PriceStats, State Street,

Aggregate Results

United Kingdom

Australia

China

Sources: BPP, PriceStats, State Street,

PPP Implied Exchange Rates

Final Remarks

- Big Data in economics is a *measurement* opportunity
  - New data collection tools (web, sensors, phones, gps, satellites)
  - Build customized datasets that fit specific measurement and research needs

- Online Price data
  - Advantages for inflation and PPP measurement
  - Increase the quantity and quality of micro-price data available for research, reevaluating old empirical puzzles and questions that could not be answered before