Publicly-Shared Datasets from The Billion Prices Project

Alberto Cavallo
MIT & NBER

IFM Data Project
NBER Summer Institute 2015
BPP Online Data

- Online prices collected
  - Directly from each retailer’s website
  - Largest retailers by market share in each country
  - Multi-channel retailers → sell offline and online

Every day, a software downloads a list of webpages, analyses the HTML code, extracts price data, and stores it in a database.

<table>
<thead>
<tr>
<th>Id</th>
<th>Product</th>
<th>Price</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD963LL</td>
<td>iPad Mini Smart Cover - Dark Grey</td>
<td>39</td>
<td>07/01/2012</td>
</tr>
<tr>
<td>MD953DD</td>
<td>iPad Mini Smart Cover - (PRODUCT) RED</td>
<td>39</td>
<td>07/01/2012</td>
</tr>
<tr>
<td>MD4461S</td>
<td>ipad smart case - polyurethane - Dark Grey</td>
<td>49</td>
<td>07/01/2012</td>
</tr>
</tbody>
</table>
## Online Price Data has Advantages and Disadvantages

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| • Cheaper to collect  
  • Frequency (daily)  
  • Granularity  
    • All product details (brands, size, etc)  
    • All goods and varieties available for sale (census)  
    • New goods automatically sampled  
  • No time-averages, imputations, adjustments, or any kind of third-party (data collector) interference.  
  • Easier to compare internationally | • Not all categories of goods and services are online  
  • Fewer retailers and locations than CPI  
  • Short time series  
  • Online and Offline prices may behave differently |
Publicly-Shared Datasets

1. Supermarket Data:
   Daily prices for all goods sold by some of the largest Supermarkets in Latin America and the US
   - Frequency: Daily
   - Retailers: 7
   - Countries: 6 (Argentina, Brazil, Chile, Colombia, Venezuela, and the US)
   - Dates: 2007 to 2011
   - Main variables: product id, date, price, sale indicator, category

Examples of Stylized Facts

- The Size of Price Changes

Figure 2: Effects of Weekly Averages and Scanner Data
Notes: The online and scanner data in the US was collected at the same retailer during the same time period. Scanner data collected by Nielsen and provided by the Kilts Center at Chicago Booth.

From Cavallo (2012) "Scraped Data and Sticky Prices"
Publicly-shared Datasets

2. Global Retailer Data

- Daily prices for all goods sold by APPLE, IKEA, ZARA, and H&M.
  - Frequency: Daily
  - Countries: 85 (coverage of countries varies by retailer and time)
  - Dates: 2008 to 2013
  - Main Variables: id, country, date, price, category

Example: Zara, China.
Examples of Stylized Facts

QJE 2014: Good-level RERs with the United States

Note: Log good-level RERs from Apple, Ikea, Zara, and H&M, 2008-2013
Examples of Stylized Facts

QJE 2014: Good-level RERs with Spain

Note: Log good-level RERs from Apple, Ikea, Zara, and H&M, 2008-2013
Examples of Stylized Facts

IMFER 2015: Share of Goods Satisfying LOP with Latvia

Notes: Plots show the percentage of observed good-level relative prices between Latvia and the listed country of absolute value less than 1 percent, measured each week. From Cavallo, A., Neiman, B., & Rigobon, R. (2014) “The Price Impact of Joining a Currency Union: Evidence from Latvia” • IMF Economic Review Forthcoming.
Publicly-shared Datasets

- Pre-cleaning:
  - De-identification of retailer and products names (except in global retailers data)
  - No other cleaning or adjustment → data contains errors and missing values

- Access is free:
  - Go to http://bpp.mit.edu/datasets/ for a list of available datasets
  - Data shared via Harvard’s Dataverse Repository
Steps
1. Click Download
2. Accept “Terms of Use” (only for research purposes and will not seek/reveal retailer or brand identity)
3. Sign “Guestbook”
Are Online Prices representative?

- The `online store` is effectively the *largest* store in most retailers
  - Eg: Walmart has 4759 stores in the US. The median store has 0.02% of sales. The `online store` has 8% of sales.
- Simultaneous online-offline data collection in 10 countries shows large retailers tend to have identical online and offline prices.

Preliminary results from Cavallo (2015) Are Online and Offline Prices Similar?

<table>
<thead>
<tr>
<th>Country</th>
<th>Ret.</th>
<th>Days</th>
<th>Workers</th>
<th>Prod.</th>
<th>Obs</th>
<th>(6) Ident. (%)</th>
<th>(7) High On (%)</th>
<th>(8) Low On (%)</th>
<th>(9) Differ. (%)</th>
<th>(10) On Mark. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>5</td>
<td>73</td>
<td>20</td>
<td>2271</td>
<td>3560</td>
<td>55</td>
<td>29</td>
<td>16</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Australia</td>
<td>4</td>
<td>61</td>
<td>18</td>
<td>2998</td>
<td>3603</td>
<td>72</td>
<td>21</td>
<td>7</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Brazil</td>
<td>5</td>
<td>66</td>
<td>16</td>
<td>1014</td>
<td>1487</td>
<td>25</td>
<td>39</td>
<td>36</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Canada</td>
<td>5</td>
<td>85</td>
<td>20</td>
<td>2400</td>
<td>3789</td>
<td>89</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>China</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>20</td>
<td>20</td>
<td>85</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Germany</td>
<td>3</td>
<td>44</td>
<td>6</td>
<td>756</td>
<td>1077</td>
<td>82</td>
<td>3</td>
<td>16</td>
<td>-1</td>
<td>-6</td>
</tr>
<tr>
<td>Japan</td>
<td>3</td>
<td>16</td>
<td>4</td>
<td>428</td>
<td>500</td>
<td>37</td>
<td>11</td>
<td>52</td>
<td>-9</td>
<td>-14</td>
</tr>
<tr>
<td>Southafrica</td>
<td>3</td>
<td>38</td>
<td>15</td>
<td>1139</td>
<td>1277</td>
<td>83</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>UK</td>
<td>4</td>
<td>57</td>
<td>15</td>
<td>1813</td>
<td>2420</td>
<td>87</td>
<td>4</td>
<td>9</td>
<td>-1</td>
<td>-4</td>
</tr>
<tr>
<td>USA</td>
<td>9</td>
<td>125</td>
<td>243</td>
<td>3110</td>
<td>6379</td>
<td>71</td>
<td>8</td>
<td>21</td>
<td>-2</td>
<td>-7</td>
</tr>
</tbody>
</table>

A Free **Offline** Data Collection Tool

**Download the free BPP app for android phones** (bpp.mit.edu)

**Contact us to receive a custom Project Code** → allows us to separate your data

**Steps:**

- You scan and email us the files (all within the app)
- Every day, we clean and consolidate the data, and save everything (prices, text, photos) in a shared dropbox folder.
Future Data Releases

• We plan to periodically upgrade and publish new datasets:
  • See http://bpp.mit.edu/datasets/
  • Follow us on twitter to get notified (@bppmit).