The Billion Prices Project: Using Online Prices for Measurement and Research

Citation: Cavallo, Alberto, and Roberto Rigobon. “The Billion Prices Project: Using Online Prices for Measurement and Research.” Journal of Economic Perspectives 30.2 (2016): 151–178.

Published Version: http://dx.doi.org/10.1257/jep.30.2.151

Publisher: American Economic Association

Permanent Link: http://hdl.handle.net/1721.1/105176

Version: Author's final manuscript: final author’s manuscript post peer review, without publisher’s formatting or copy editing

Appendix

The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo
MIT & NBER

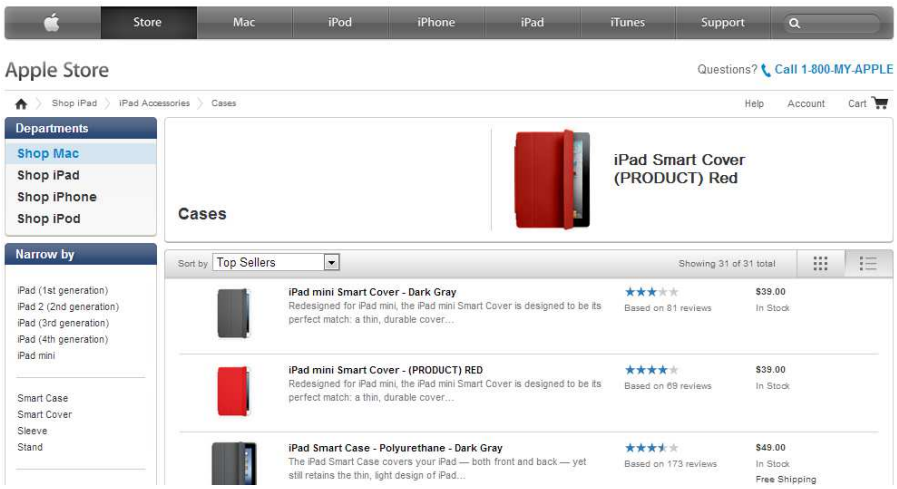
Roberto Rigobon
MIT & NBER

This Version: March 16, 2016

A1 Web Scraping

Our scraping process is based on a combination of programming languages and software optimized for scanning the code of publicly available websites, identifying relevant pieces of information, and storing them in a database. It follows three basic steps.

First, every day the software downloads a selected list of public web-pages where product and price information are shown. These pages are individually retrieved using their web-address (URL). For example, the software may visit the Apple Store and look at the page where iPad covers are listed, as shown in Figure A1.




```
<html>
<!-- START product -->
<ahref="productId=MD963LL"></a>
<p class="productname">Ipad Mini Smart Cover – Dark Grey</p>
<td class="Price">$39.00</td>
<!-- END product -->
.....
```

Figure A1: Example of HTML code used for web scraping

Note: This is an example of how the html code in a webpage can be used to identify different variables to be scraped. The scraping "robot" can be instructed to use a set of characters in the code to know when to start and stop collecting information for each variable. In this hypothetical example, each product is contained between the $<!--START\ product-->$ tag and the $<!--END\ product-->$ tag, and the price is shown between the $"Price" >$ and $</td>$ characters.

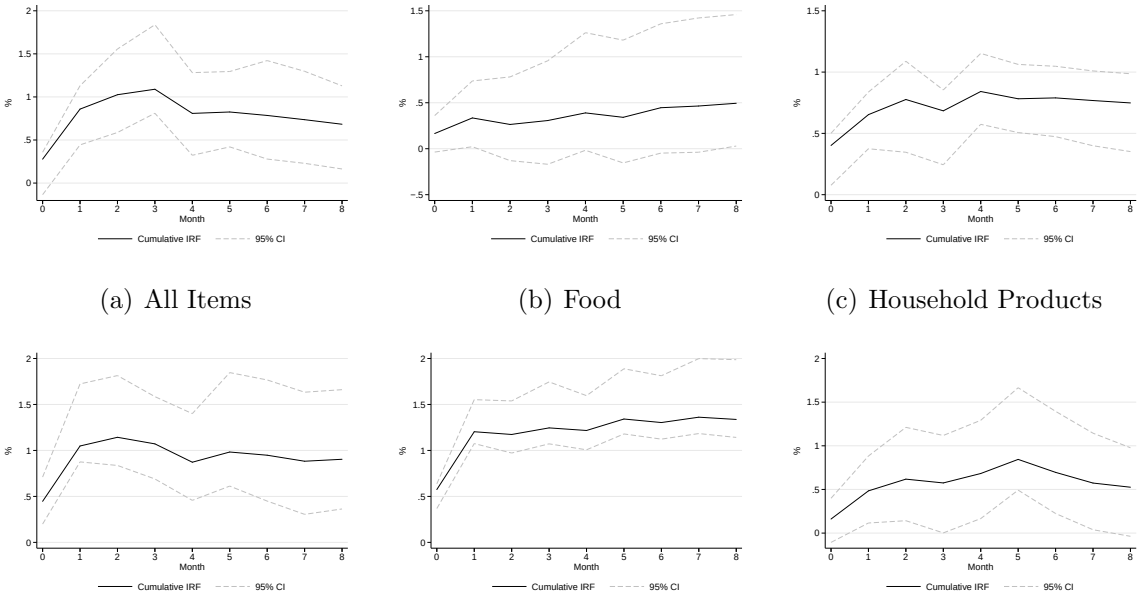
The set of URLs that the robot visits are carefully chosen by the scraping team based on the categories of goods that we want to sample. We also follow the robots.txt exclusion protocols of the servers where the data is located.

Second, the HTML code underlying the webpage is analyzed to locate each piece of relevant information. This is done by using special characters in the code that identify the start and end of each variable, and have been placed by the page programmers to give the website a particular look and feel. In the example in Figure A1, prices are shown with a dollar sign in front of them and enclosed within the $< tdclass = "Price" >$ and $</tags>$ tags.

Third, the software stores the scraped information in a database that contains one record per product per day. Our datasets typically include a product identification number, the price, the date, some category information, and an indicator for whether the item was on sale or not. In other cases we are able to also record product details, such as the description, model, unit, size, brand, an indicator for whether the product is out-of-stock, and country of origin.

A2 Impulse Responses at the Sector Level

The figure below shows the sector-level monthly cumulative impulse responses of the US Consumer Price Index to a 1% shock in the online price index. The fastest impact takes place in transportation, which includes fuel. The slowest is food, where the impact is gradual and incomplete.



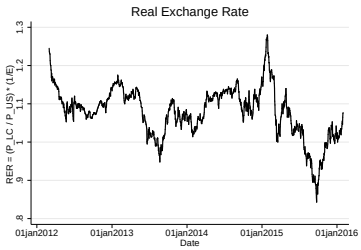
(a) All Items (b) Food (c) Household Products

(d) Health (e) Transportation & Fuel (f) Electronics

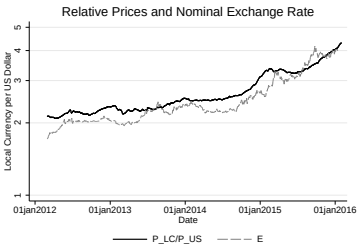
Figure A2: Sector Impulse Responses of the US CPI to an Online Price Index Shock

Notes: Sector-level cumulative impulse responses to a 1% shock in the online index. The online price index was computed by PriceStats. The CPIs are US city averages, non-seasonally adjusted, from the Bureau of Labor Statistics. Data from July 2008 to January 2015.

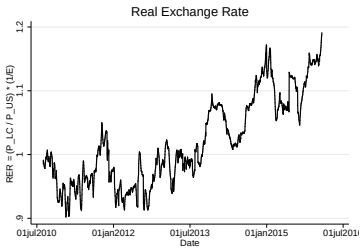
A3 PPP in other Countries



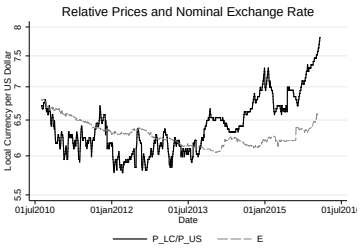
(a) Brazil RER



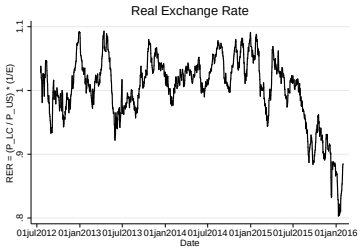
(b) Brazil RP and E



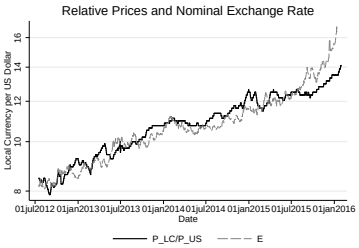
(c) China RER



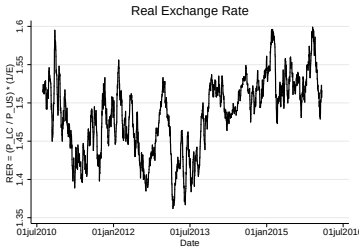
(d) China RP and E



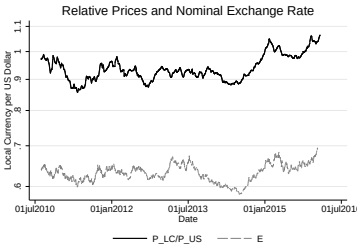
(e) South Africa RER



(f) South Africa RP and E



(g) UK RER



(h) UK RP and E

Figure A3: Relative Prices and Exchange Rate - Additional Countries

Notes: The right panel shows the ratio of relative prices (in local currencies, P/P_US) and the nominal exchange rate (E, defined as local currency per US dollar). The left panel shows the real exchange rate computed as $(P/P\_US) * (1/E)$. It is the relative cost of the basket in each country relative to the US, when expressed in the same currency. Real exchange rates and relative price series are computed by PriceStats at the product level and aggregated using a Fisher index with official CPI expenditure weights for food, fuel, and electronics.