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.

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

Publisher: American Economic Association

Persistent URL: 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

The Billion Prices Project: Using Online Prices for Measurement and Research [*] [†]

Alberto Cavallo
MIT and NBER

Roberto Rigobon
MIT and NBER

This Version: April 8, 2016

Abstract

New data-gathering techniques, often referred to as "Big Data" have the potential to improve statistics and empirical research in economics. In this paper we describe our work with online data at the Billion Prices Project at MIT and discuss key lessons for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices. We emphasize how Big Data technologies are providing macro and international economists with opportunities to stop treating the data as "given" and to get directly involved with data collection.

JEL-Codes: E31, F3, F4.

Keywords: Online Prices, Inflation Measurement, Big Data, Scraped Internet Prices.

[*] This paper was prepared for the Journal of Economic Perspectives. We thank the editors, Timothy Taylor, Ann Norman, Enrico Moretti, and Gordon Hanson for their helpful comments and suggestions. We thank Manuel Bertolotto, Maria Fazzolari, Diego Aparicio, Belen Bazano, and Augusto Ospital for outstanding research assistance at various stages of our Billion Prices Project research, and Pilar Iglesias and all the team at PriceStats for sharing their indexes and micro data. We also thank Chris Sims for detailed comments on an earlier version of this paper, as well as seminar participants at the MIT, IMF, IADB, Census Bureau, BEA, ECB, CEMLA, NYU, Federal Reserve Bank of San Francisco, UNECE/ILO CPI Experts Conference, NBER CIRW Conference, Bureau of Labor Statistics, and the State Street Annual Research Conferences. All remaining errors are ours.

[†] Financial support for this research was provided by MIT Sloan. Publicly available Billion Prices Project datasets can be downloaded at http://bpp.mit.edu.

1 Introduction

New data-gathering techniques, often referred to as “Big Data”, have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bi-monthly basis to collect prices for a pre-selected basket of goods and services. The micro data are then processed and used to construct Consumer Price Indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate adjustments for quality changes and the introduction of new goods.[1] Groves (2011) further describes other challenges faced by traditional survey-based methods of data collection, including growing levels of non-response. Shrinking resources are straining the work of national statistical offices, while recent crises have prompted policy-makers and other users of these statistics to demand faster and more accurate data.

Online prices have a natural appeal in this context. While the data are dispersed across hundreds of websites and thousands of webpages, advances in automated “scraping“ software now allow anyone to design and implement large-scale data collections on the web. Detailed information can be collected for each good, and new and disappearing products can be quickly detected and accounted for. Online data collection is cheap, fast, and accurate, making it an ideal complement to traditional methods of collecting prices, particularly in categories of goods that are well-represented online.

The first use of online data to construct inflation indexes was motivated by the manipulation of inflation statistics in Argentina from 2007 to 2015. By 2007 it had become apparent that the official level of inflation reported by the national statistical office in Argentina did not reflect the actual changes in prices. Using online data collected every

[1] See the papers in the winter 1998 issue of the Journal of Economic Perspectives for a discussion of some of these topics.

day from the websites of large retailers, Cavallo (2013) showed that while Argentina's government announced an average annual inflation rate of 8 percent from 2007-2011, the online data suggested it was actually over 20 percent, in line with the estimates of some provincial governments and local economists, and consistent with the results from surveys of household inflation expectations. The online price indexes used in that paper were automatically computed and published on a website every day from March 2008 onwards.[2] The ability to collect prices from outside the country proved particularly useful in 2011, when Argentina's government started to impose fines and to pressure local economists to stop collecting data independently. The manipulation of the official price index ended in December 2015 when a new government was elected.

Argentina's statistical debacle had a positive side effect: it showed us the potential that online prices had for inflation measurement applications. With this idea in mind, we created the Billion Prices Project at MIT in 2008 to extend our work to other countries, including the United States. The word "billion" was simply meant to express our desire to collect a massive amount of prices, though we in fact reached that number of observations in less than two years. By 2010, we were collecting 5 million prices every day from over 300 retailers in 50 countries. Half a million prices were collected every day in the United States alone (by comparison, the US Bureau of Labor Statistics collects approximately 80,000 prices on a monthly or bi-monthly basis). Although gathering this massive amount of prices was cheaper online than with traditional methods, it required funding that could not be sustained through grants. Thus, in 2011 we started a company called PriceStats that now collects the data and produces high-frequency indexes for central banks and financial-sector customers. PriceStats greatly expanded both the quantity and quality of the data. The company currently uses about 15 million products from over 900 retailers to build daily inflation indexes in 20 countries. Its micro datasets contain information from an even larger number of retailers in over 60 countries, with varying degrees of coverage. The indexes and micro data from PriceStats are available to researchers working with the Billion Prices Project, as we explain later in this paper.

Many of the other attempts to use Big Data in economics rely on social media or search data to forecast the behavior of important economic indicators. Our approach is different because we focus on measurement, not on prediction. Our objective is to experiment with

[2] See Inflacion Verdadera (http://www.inflacionverdadera.com), which was created to provide alternative prices indexes to the official ones in Argentina. The original website had two price indexes constructed with the Argentina's official National Institute of Statistics and Censuses (INDEC is the abbreviation of the Spanish translation): a "Basic Food" index and a broader "Food and Beverages" index. The website also showed the time series of prices for every good used in the index.

these new sources of information to improve the computation of traditional economic indicators, starting with the Consumer Price Index. We seek to understand whether online prices have distinct dynamics, their advantages and disadvantages, and whether they could be a reliable source of information in a “production“ setting (not just for a one-time research application).

We start this paper with a description of the methodology used to collect online prices. A first-order aspect is to realize that although the amount of data online is massive, carefully selecting the categories and retailers to sample is still crucial. The goal is to obtain data that is representative of retail transactions, so we focus our data collection efforts on large multi-channel retailers such as Walmart that sell both online and offline, instead of using online-only retailers that may have many products but a relatively small share of retail transactions. We also focus on categories of goods that are included in the official Consumer Price Index baskets, for which consumer expenditure weights are available. After describing the sources of data, we discuss the advantages and disadvantages of online data relative to other large micro-price databases (including scanner data and official price-index data), and highlight the results of a large-scale validation exercise to show how online-price levels and behaviors closely resemble those that can be obtained by physically visiting offline stores.

Next, we describe the methodology used to compute online price indexes and show how they co-move with Consumer Price Indexes in most countries. We emphasize two characteristics in greater detail. First, online indexes have the ability to approximate hedonically-adjusted price indexes in sectors with a large number of goods that come and go with overlapping lifecycles (for instance, electronics). Second, online indexes appear able to anticipate movements in the official Consumer Price Index in many countries. This anticipation extends beyond the publication lags, which suggests that online prices often adjust sooner to aggregate shocks.

We then move on to research applications and discuss two areas in macro and international economics where online price data can have a major impact. First, we show that online price data, collected daily, can significantly alter some key results in the price-stickiness literature. In particular, we document that online prices exhibit a very different distribution of price changes compared to prices collected for official Consumer Price Indexes and by scanner prices. The main reason for the difference is that online prices do not have time averages, common in scanner data, or imputed prices, common in official micro data, which create a large number of small spurious price changes. Second, tests of the "law of one price" (that there should not be large or persistent cross-country differences in the prices of identical goods when translated into a common currency) using online data give us a more nuanced picture of when and where this law works well. The existing consensus in the literature is that there are large and persistent deviations from the law of one price, with little pass-through from nominal exchange rates to relative prices, and vice-versa, causing persistent shocks to real-exchange rates that take years to dissipate. While deviations can also be large with online data, we find that the law of one price holds well across countries that use the same currency. We also show that, when goods are identically-matched goods across countries, then relative prices and nominal exchange rates co-move more closely than previously thought. This implies higher pass-through rates and less persistent real-exchange rate dynamics.

Both research examples illustrate how using data collected by others, with different purposes in mind, can distort empirical findings. They also suggest we should not treat "big data" as simply a collection of large datasets created as a byproduct of something else. Instead, we think one of the greatest opportunities of "big data" is the fact that anyone can now use new technologies such as web-scraping, mobile phones, satellite imaging, and all kinds of interconnected sensors to build customized datasets designed to fit specific measurement or research needs. We end the paper by describing how the Billion Prices Project data are publicly shared and by discussing why data collection is an important endeavor that macro and international economists should pursue more often.

2 Collecting and Processing Online Price Data

A large and growing share of retail prices all over the world are posted online on the websites of retailers. This is a massive untapped source of retail price information. Collecting these prices is not trivial because they are posted on hundreds of different websites that lack a homogeneous structure and format. And retailers do not provide historical prices, so the data has to be collected continuously and consistently over time.

To collect and process online prices we follow a "data curation" approach. It involves carefully identifying the retailers that will serve as data sources; using web-scraping software to collect the data; then cleaning, homogenizing, categorizing, and finally extracting the information so it can be used in measurement and research applications.

2.1 The Selection of Retailers and Data Source

The starting point is to select the retailers and categories of goods to sample. These decisions are driven by our need to get prices that are representative of retail transactions. We therefore focus almost exclusively on large multi-channel retailers (those retailers that sell both offline and online, such as Walmart) and tend to ignore online-only retailers (such as Amazon.com). The reason is that multi-channel retailers still are involved in the majority of all retail sales in most countries. We are also careful when we choose what categories of goods to monitor within each retailer, concentrating on those categories that are part of traditional Consumer Price Index baskets, and avoiding categories that are over-represented online such as CDs, DVDs, cosmetics, and books.

We make an effort to collect the data directly from each retailers website, rather than relying on third parties such as marketplaces, price aggregators, and price comparison websites. Data collection from individual retailers is far more challenging, but it maximizes our chances of obtaining prices linked to actual transactions and prevents third-parties from filtering or altering our samples. It also gives us full control of what we choose to collect and makes the whole process more robust, as it does not depend on a few sources of data.

Once the data are collected, we clean them, standardize them to fit a common database schema, classify individual products using Consumer Price Index categories, and start computing simple indicators to evaluate its characteristics and performance over time.

We treat each retailer as a separate sampling unit or “stratum” with potentially unique characteristics and pricing behaviors. Before including a retailer in a price index, we usually monitor its behavior for over a year to identify any special characteristics in the data, so that we can know whether it is a useful and reliable source of price information.

Most retailers that sell online have a single price for all shoppers in all locations within a country (though shipping costs and taxes may differ). Grocery retailers can sometimes show different prices for the same good depending on the zip code entered by the consumer. In such cases, we select a few zip codes corresponding to major cities and treat each case as an independent retailer.

The amount of data and the coverage of different categories that we can observe online vary across countries. For about 25 countries, our datasets have information on categories that cover at least 70 percent of the weights in Consumer Price Index baskets.

2.2 Data Collection Using Web Scraping Software

After selecting the sources of data, the next step is to collect the information. The technology to collect online prices on a large scale called "web scraping" is quickly improving. Just a few years ago, it required researchers to write programs in languages such as Python and PHP (for example, see the discussion in this journal by Edelman 2012). Today, there are many "point-and-click" software solutions that require almost no technical expertise. Users can simply use their mouse to teach the software what pieces of information they want to collect from a webpage. The software then creates a "robot" that is able to extract information from any other webpage with a similar structure, storing the information in a database. It identifies relevant pieces of information on a page by finding special characters of HTML code (the language that is used to create webpages) that come before and after each relevant piece of information. These characters are relatively steady as long as the page does not change its look-and-feel. The challenge in web scraping is mostly to monitor the performance of the robots over time, so that any errors in the data can be quickly detected and fixed. The robots we construct always collect a product identification number, the name, description, brand, package size, category information, and the price. When available, we also collect other variables such as sale prices and stock indicators. We provide more details of the web-scraping process in the online Appendix available with this paper at http://e-jep.org.

2.3 Advantages and Disadvantages of Online Price Data

To understand the strengths and weaknesses of this scraped online data for measurement and research applications, Table 1 offers a comparison with two other sources of micro-price data: traditional Consumer Price Index data collected offline by national statistical organizations, and scanner data recorded from consumer purchases at point-of-sale terminals by companies such as Nielsen. Detailed descriptions of these other data sources can be found in ILO (2004) and Feenstra and Shapiro (2003).

One of the most obvious advantages of online data is the low cost per observation. While not trivial, it is far cheaper to use web scraping than to hire people to visit physical stores or buy information from commercial scanner data providers such as Nielsen.

A second major advantage is the daily frequency of data collection. It is easier to detect errors in the data when it is collected at such high frequency. It also avoids time averages, which can generate spurious price changes as we discuss later on.

Third, online data includes detailed information for all products being sold by the sampled retailers. The cross-section of prices available is therefore much larger within categories than in Consumer Price Index data. Later, we discuss how this "big data" feature can be used to simplify quality adjustments and other traditional measurement problems.

Fourth, there are no censored price spells in online data. Prices are recorded from the first day a product is offered to consumers until the day it is discontinued from the store. Traditional data collection methods, in contrast, will typically start monitoring new goods only when the goods in the basket disappear from the stores. Knowing the full history of prices for individual goods can help to control for new-good biases, make both implicit and explicit quality adjustments, and study prices at the time of product introductions.

Fifth, online data can be collected remotely. This is particularly useful in situations like the one experienced by Argentina in recent years, where the government was trying to prevent independent data collection for the computation of inflation. It also allows us to centralize the data collection and homogenize its characteristics.

Sixth, and related to the previous point, online datasets can be readily comparable across countries because prices can be collected with identical methods on matching categories of goods and time periods. This is useful in research applications that use cross-country comparisons.

Finally, online data are available in real time, without any delays to access and process the information. This is particularly useful for policymakers and anyone who needs up-to-date information.

Table 1: Alternative Micro-Price Data Sources

	Online Data	Scanner Data	CPI Data
Cost per observation	Low	Medium	High
Data Frequency	Daily	Weekly	Monthly
All Products in Retailer (Census)	Yes	No	No
Uncensored Price Spells	Yes	Yes	No
Countries with Research Data*	∼60	<10	∼20
Comparable Across Countries	Yes	Limited	Limited
Real-Time availability	Yes	No	No
Product Categories Covered	Few	Few	Many
Retailers Covered	Few	Few	Many
Quantities or Expenditure Weights	No	Yes	Yes

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.

One of the main disadvantages of online prices is that they currently cover a much smaller set of retailers and product categories than a government-run survey of consumer prices do. In particular, the prices of most services are still not available on the web, and the number and type of retailers is limited compared to official Consumer Price Index data.

Another disadvantage is that online datasets lack information on quantities sold. Online prices must be combined with weights from official consumer expenditure surveys or other sources for expenditure-weighted applications. Scanner datasets, by contrast, have detailed information on quantities sold, and could potentially be a source of high-frequency expenditure weights in some categories of goods such as groceries.

2.4 Are Online Prices Different?

An important concern is whether online prices are different from offline prices; after all, most transactions still take place offline. The suspicion that online prices are different is fueled by reports that some online retailers use “dynamic pricing” strategies in which prices are varied for strategic purposes: for examples, see Mikians et al. (2012) and Valentino-DeVries et al (2012). In addition, many papers with “online prices” use data from online marketplaces such as Ebay or price-comparison websites such as Google Shopping. As Brynjolfsson and Smith (2000), Ellison and Ellison (2009), and Gorodnichenko et al (2014) have shown, these prices seem to change more frequently and in smaller sizes than in Consumer Price Index data. However, the retailers in these datasets are mostly online-only stores participating in a fiercely competitive environment, not really the type of “online data” we use.

To better understand whether online and offline prices for multi-channel retailers behave differently, in Cavallo (2015a) we simultaneously collected prices on the websites and physical stores for over 24,000 products in 56 of the largest retailers in 10 countries. This large-scale comparison was possible thanks to the combination of a smartphone app, crowdsourced workers, and web-scraping techniques. More than 370 freelance workers used their phones to scan barcodes in physical stores, manually enter prices, take photos of the price tags, and upload the information to our Billion Prices Project servers. We then used the barcodes in the offline data to collect the prices for those exact same goods in the website of the same retailer within a seven-day time window.

This direct comparison between online and offline prices revealed a high degree of similarity in price levels, as well as in both the frequency and size of price changes. On average, about 72 percent of price levels were identical in the offline and online samples. The similarity was highest in retailers that sell electronics or apparel, and lowest in drugstores and office-supply retailers that also tend to price differently across offline stores. While price changes do not have the exact same timing online and offline, they tend to have similar frequency and average sizes. This suggests that the price spells for individual goods may not be synchronized online and offline, consistent with evidence to be discussed below that online prices may anticipate later price changes. Despite the general similarity between online and offline pricing, our results also revealed a great deal of heterogeneity among pricing behaviors, suggesting some validation is needed in papers with data from a limited number of retailers.

3 Inflation Measurement

Online prices are increasingly being used in inflation measurement applications. Besides the Billion Prices Project and PriceStats, many national statistical organizations are experimenting with the use of online data, including the US Bureau of Labor Statistics, (Horrigan 2013a), the UK Office of National Statistics (Breton et al. 2015), Statistics Netherlands (Griffioen et al. 2014), Statistics New Zealand (Krsinich 2015), and Statistics

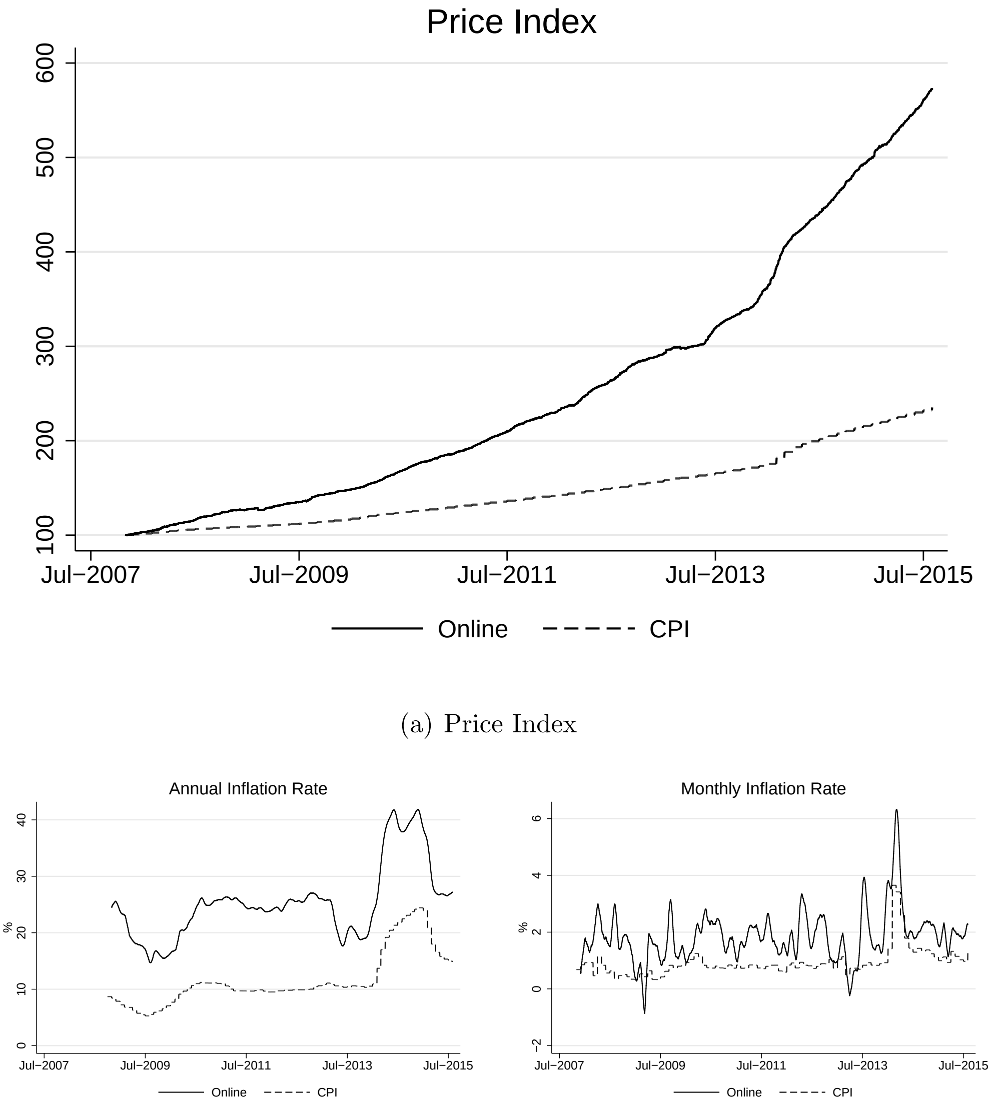
Norway (Nygaard 2015).

In this section, we show that online price indexes can closely approximate the official Consumer Price Index in a number of countries and settings. We then discuss how a large number of overlapping price series in the data can simplify quality adjustments in categories with frequent product turnover, such as electronics. Finally, we show that online price indexes can anticipate changes in the official inflation rate several months in advance.

3.1 Methodology for Comparison to Official Consumer Price Indexes

For multiple Latin American countries, Cavallo (2013) showed that online prices could be effectively used as an alternative source of price information to construct price indexes that mimic the behavior of official Consumer Price Indexes. The methodology for these daily indexes was based on a combination of online data with standard techniques used in official price indexes, including expenditure weights for each sector where online data are available. This initial work included only data from food retailers and a handful of countries. In 2010, we founded PriceStats to expand the data collection and to compute inflation measures in real-time in other sectors and countries. The company is currently publishing daily price indexes in 22 countries with only a three-day lag. In Figures 1 to 4 we plot these online indexes next to the all-item non-seasonally adjusted Consumer Price Index in each country. We first highlight the cases of Argentina and the United States, and then show some selected cases in a larger set of countries.

Figure 1 illustrates the case of Argentina from 2007 to 2015. Panel (a) compares a price index produced with online data to the official Consumer Price Index.



(a) Price Index

(b) Annual Rate

(c) Monthly Rate

Figure 1: Argentina

Notes: Online price index computed by PriceStats. Monthly Inflation Rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average a month before. Annual Inflation Rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. Consumer Price Index from the national statistical office in Argentina (INDEC). All price indexes are non-seasonally adjusted.

The fact that the two measures of inflation in 1(a) diverge so dramatically will not surprise anyone who knows the recent story of statistics in Argentina. In February 2007, the government intervened in the National Institute of Statistics and Census (INDEC) and fired the people responsible for computing the Consumer Price Index. The official index quickly stabilized, but many local economists claimed the government was manipulating the data. Household inflation expectations increased dramatically, closely tracking some alternative estimates of inflation produced by local economists. Suspicions were abundant, but before a measure of inflation based on online prices became available, there was no consistent way to confirm the magnitude of the discrepancy and track its evolution over time.

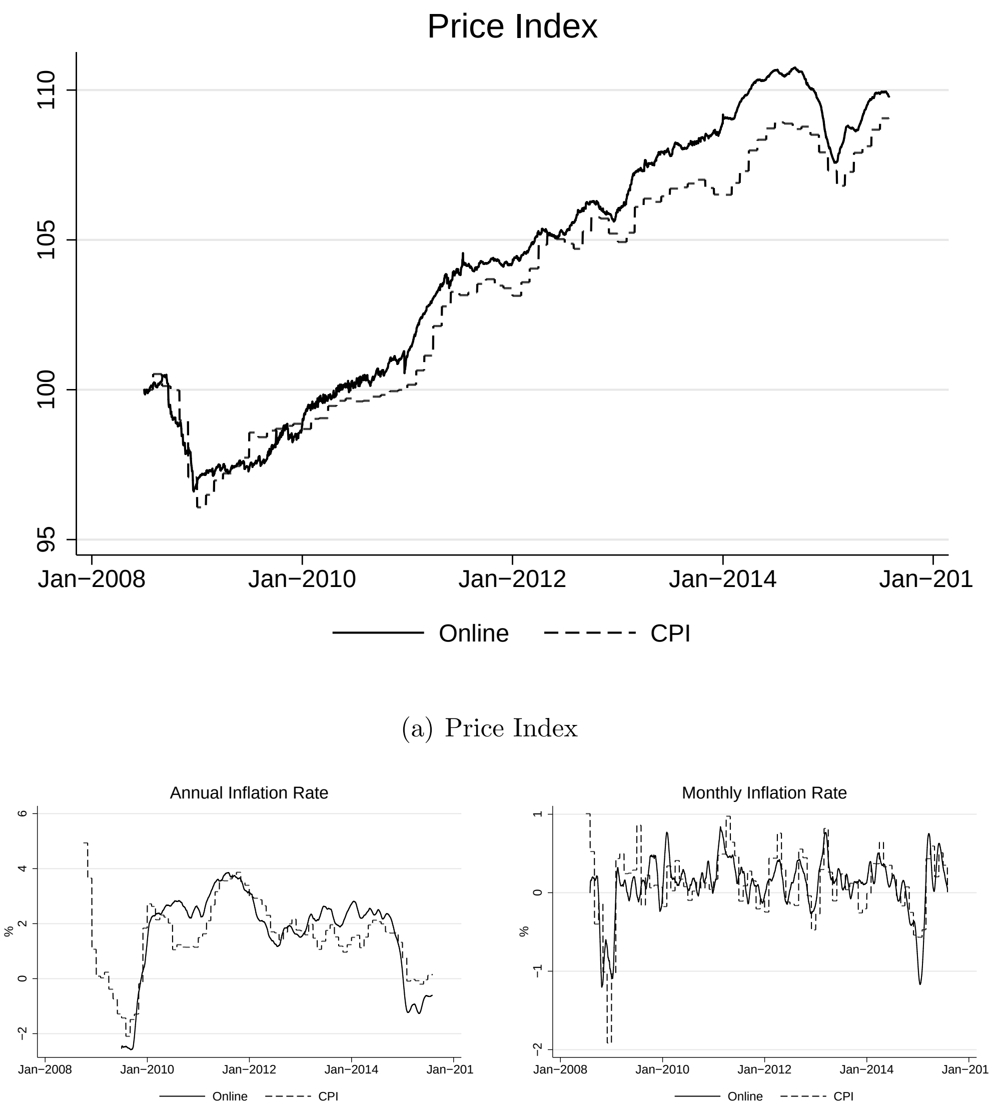
The manipulation in the official inflation data continued for almost nine years, ending in December 2015 when a new government was elected. During all this time, the monthly inflation rate shown in Figure 1(c) was almost always consistently higher than the official data reported, with the exception of a few months in 2014 when, in response to a "motion of censure" issued by the IMF in 2013 (Rastello and Katz 2013), the Argentinian government decided to launch a new Consumer Price Index. Unfortunately the change was temporary and the new official index quickly lost all credibility again.[3]

Looking only at the discrepancy in the trend of the price index or the month inflation rates, however, misses an important point. The online index tracked the dynamic behavior of the annual inflation rate over time, as shown in Figure 1(b). The difference was mostly in the level of the annual inflation rate, not its movements over time. The online index also quickly reacted to aggregate shocks, such as the massive roadblocks by farmers who protested export tax hikes in 2008. This strongly suggested that online data was capable of capturing the fundamental dynamics of inflation and prompted us to collect data in other countries.[4]

The comparison of online and offline indices in other countries is completely different. The daily US index, shown in Figure 2, is a great example.

[3] More details on the manipulation of inflation statistics in Argentina and its impact on inflation expectations can be found in Cavallo, Cruces, and Perez-Truglia (2014, 2016).

[4] It also implied that the government was not using a particularly sophisticated algorithm to change the inflation rate. In Cavallo (2013) we showed that one could closely approximate the official index by simply dividing the online inflation rate by three.



(a) Price Index

(b) Annual Rate

(c) Monthly Rate

Figure 2: USA

Notes: Online price index computed by PriceStats. Monthly Inflation Rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average a month before. Annual Inflation Rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. Consumer Price Index from the US Bureau of Labor Statistics. All price indexes are non-seasonally adjusted.

Despite the multiple reasons why we might expect inflation indexes based on online and offline prices to deviate, the US online index has co-moved closely with official Consumer Price Index for over seven years. Although there are periods where the indexes diverge, the differences are relatively small and temporary. This can also be seen in the monthly and annual inflation rates in Figures 2b and 2c.

The US online index is particularly good at anticipating major changes in inflation trends. Predicting these changes is important for participants in financial markets, policymakers, and those economists who monitor the economy closely. One remarkable example of a turning point detected with online data months before it showed up in official US Consumer Price Index data was September 16, 2008, the Tuesday after Lehman Brothers filed for bankruptcy. As Figure 3 shows, the online price index peaked that day and started falling. By October 15th, it had lost almost 1.2 percent in a single month. On October 16, the Consumer Price Index for September came out with only a 0.14 percent drop. When the official October Consumer Price Index numbers were published on November 19th, it had fallen another 1.01 percent. In other words, it took more than two months after Lehmans disaster for the official Consumer Price Index numbers to reflect the full impact on price levels. Two months later, on December 16, 2008, the online price index stopped falling and started to increase once again. The Consumer Price Index did not show this change in the trend until the estimates for January were published on February 20th, 2009. We measure the degree of anticipation in online data more formally later on.

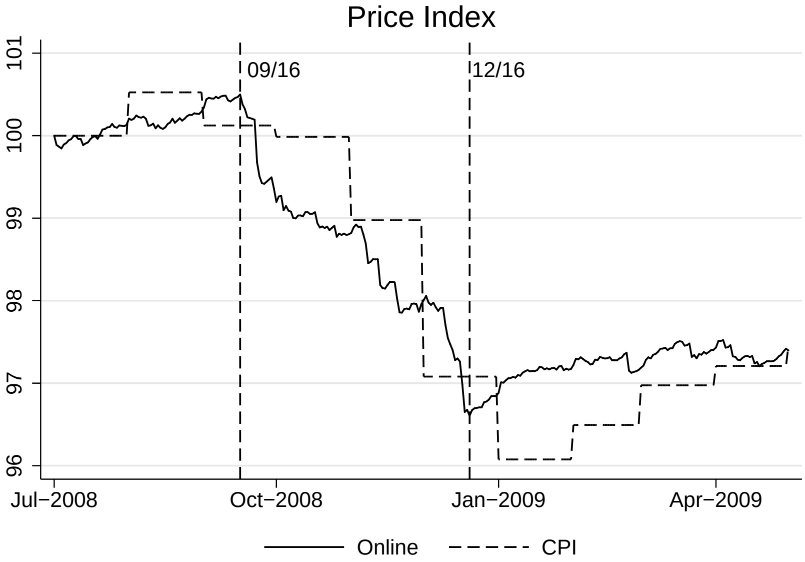
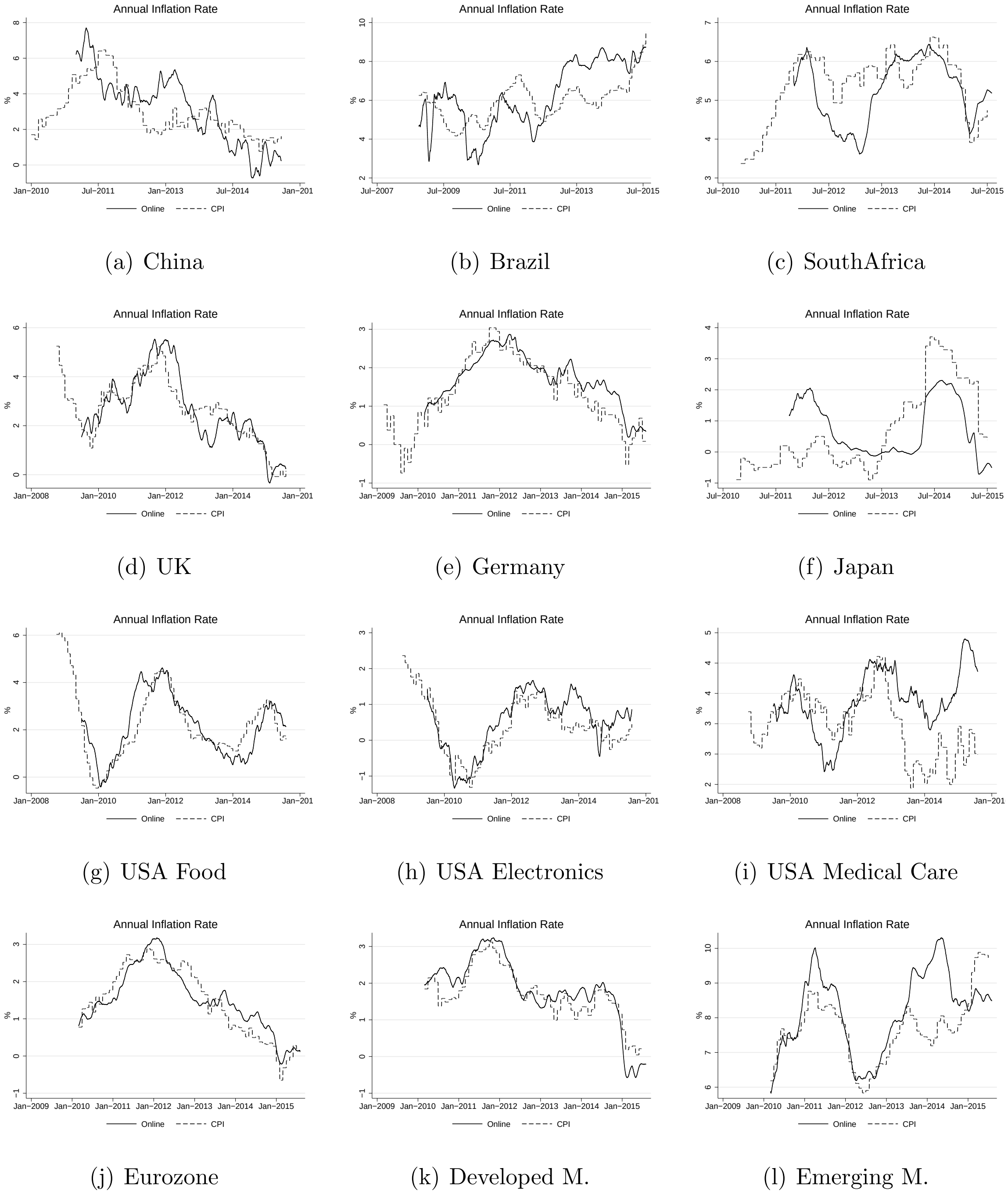


Figure 3: US Price Index around Lehman Brother's Bankruptcy

Notes: Online price index computed by PriceStats. Consumer Price Index from the US Bureau of Labor Statistics. The figure highlights the events around the bankruptcy of Lehman Brothers, the fourth-largest investment bank in the US, during September 2008.

Figure 4 compares inflation as measured by online prices and by the offline prices in the official Consumer Price Index for a selection of other countries and sectors. The main lesson of the figure is that the correspondence is reasonably close, but some more specific insights are also possible. First, we do not find evidence that China has been systematically holding its official inflation rate below the rate based on online prices, though we can only compare some sectoral indexes because official expenditures weights are not publicly disclosed. Second, the difference between the online price index and the official Consumer Price Index appears to be smaller in developed countries like the UK and Germany, and greater for countries like Brazil or South Africa, where the online sector seems to have more independent patterns. In Japan, we observed significantly more inflation after the March 2011 earthquake and an immediate impact of the sales tax changes in April 2014. While the online index in Japan does not follow its official Consumer Price Index closely, it does seem to anticipate key changes in inflation trends.



(a) China (b) Brazil (c) SouthAfrica

(d) UK (e) Germany (f) Japan

(g) USA Food (h) USA Electronics (i) USA Medical Care

(j) Eurozone (k) Developed M. (l) Emerging M.

Figure 4: Online vs CPI Annual Inflation Rates

Notes: Online price indexes computed by PriceStats. Annual Inflation Rates for daily online price indexes are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. Consumer Price Indexes are sourced from the National Statistical Office in each country. Non-seasonally adjusted series. Indexes are "all-items" with the exception of China, where an online supermarket index is shown next to the official food index. Global aggregates in the last row are computed using 2010 consumption weights in each country and CPIs from official sources.

The third row of Figure 4 shows results for a few US sectors. As one might expect, the online data matches the US Consumer Price Index better in sectors such as food and electronics, for which online information is widely available. By contrast, some official inflation patterns seen in the “Medical Care” sector are not well-captured by online prices, mostly because many services cannot be monitored online. The fourth row shows that online data can be used to provide global aggregates using country consumption weights.

It may seem surprising to some readers that indexes based on online data have the ability to mimic official Consumer Price Indexes in so many cases: for large and small countries, for developed and emerging markets, and for the aggregate and sectoral data. After all, the data differs significantly from traditional sources, and we do not apply many adjustments and methods used by national statistical organizations, such as hedonic quality adjustments. We believe there are two reasons for this closer-than-expected correspondence. First, as mentioned before, we carefully design and select the data that goes into these indexes to ensure that they are representative. Second, we learned that many sampling characteristics in our data made it simpler to deal with some traditional measurement problems. To illustrate this point, we next discuss how online data can simplify quality adjustments by providing a large number of uncensored and overlapping price spells.

3.2 Overlapping Quality Adjustments

Quality adjustment poses problem for any measure of inflation: as is widely understood, if a good rises in both quality and price, then some of the price increase is presumably due to the quality changes and should not be attributed to inflation. National statistical organizations use different methods for quality adjustments, including seeking the closest comparable substitutes when a product disappears and often relying on adjustments with hedonic regressions (in which the price change is calculated while holding constant certain attributes of a good, like the memory or hard drive capabilities of a computer). Online datasets make it easier to deal with quality adjustments because they provide uncensored price spells for a large number of models and varieties of each good. With better underlying data, online price indexes can approximate the results of more sophisticated, and often impractical, hedonic-regression methods.

To build some intuition for why this result holds true, consider a hypothetical example of a series of prices in Figure 5. It illustrates the data resulting from a traditional offline data collection process. Each line represents the price of a single good over time. Many models of electronic products, such as televisions, dishwashers, washing machines, and vacuum cleaners, tend to be introduced at relative high prices and then are discounted gradually over their life cycle, with clearance sales occurring right before the product disappears from the stores (Silver and Heravi 2005).

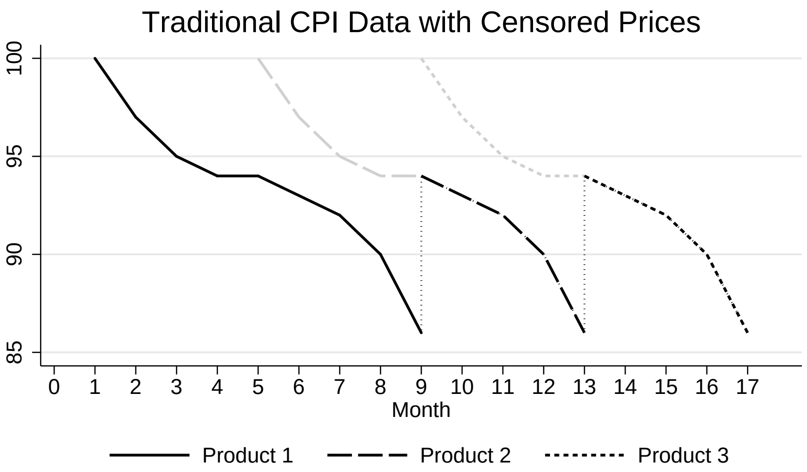


Figure 5: Overlapping Qualities

Notes: This figure shows hypothetical price series for three goods. Each downward-slopping line reflects the prices of a different good. The shaded portions of these lines illustrate the data points that would be missed with traditional data collection methods.

With traditional data collection methods, it is too expensive to collect the prices for every good available for sale at each point in time within a sampled retailer. Instead, the data collector focuses on one (or a few) of the most popular models and record its price once per month until it disappears from the store. When a particular model is no longer available, the data collector starts to sample a different model, as shown by the vertical dashed line in the figure. But at the time of the shift, the previous prices for the new model are unknown (shown where the line is shaded more lightly on the figure). The problem is to decide how much of the price gap at that point in time is attributable to quality differences. This issue is exacerbated in goods that experience extreme price movements along their life cycles, and may have steep discounts right before disappearing from the shelves.

National statistical organizations have two main ways of dealing with this problem. One preferred method is to use hedonic techniques. Again, these involve setting up a regression with the price of a good on one side and actual attributes of the good on the other side, so that future changes in the price of the good can be calculated while holding constant the attributes. While hedonic techniques have become popular in recent decades, the question of what traits should be included in the regression, how they can be measured, and what specification should be used can make hedonic techniques too data-intensive and complex to implement in practice.

A simpler alternative method is to use "overlapping qualities." As Armknecht and Weyback (1989) point out, if two goods coexist for some time, their overlapping prices can be used to obtain an estimate of quality change. In practice, this approach tends to assume that the price gap at the time of introduction of the new variety mostly reflects a quality difference. The problem with traditional data, however, is that the price of the new good is not observed at the time of introduction, but much later, when the old good disappears from the stores. This is noted in the Consumer Price Index Manual of the International Labor Organization (ILO 2004, p. 27-28): *"When there is overlap, simple linking ... may provide an acceptable solution ... In practice, however, this method is not used very extensively because the requisite data are seldom available ...The information needed for this...will never be available if price collectors are instructed only to introduce a new quality when an old one is dropped."*

Online prices offer a simple solution to this data problem by providing a large number of uncensored price spells for all models on sale at any point in time. With this type of data, a simple index using overlapping qualities can closely approximate official indexes that use complex hedonic quality-adjustment methods. Similar results were documented earlier in the price-index literature using scanner data. For example, Aizcorbe et al. (2000) and Aizcorbe et al. (2003) used scanner prices to demonstrate that, with high-frequency data, matched-model price indexes could yield results that are numerically close to those obtained using hedonic techniques in samples where product characteristics did not change much over time. More generally, the extent to which a simple matched-model price index can capture quality change will depend on several factors (Silver and Heravi 2005). First, both varieties of the product need to have a substantial degree of overlap in their prices. Second, there needs to be a large number of models so that continuing varieties can capture aggregate effects without being overly affected by idiosyncratic price movements of goods that enter and disappear from the sample.

As evidence of this effect, consider the data in Figure 6. It contains three price indexes for televisions in the United States from 2008 and 2009. The solid line shows the official Consumer Price Index for televisions as computed by the US Bureau of Labor Statistics using hedonic methods. The line with long dashes shows an online price index based on 50 distinct models of televisions from a large US retailer. The line with short dashes shows an online price index with 500 models from the same source. As we increase the number of models included in the index, we more closely approximate the results of the hedonic price index constructed by the US Bureau of Labor Statistics during this time period. Intuitively, the more overlapping price series being used, the less important the extreme price movements of goods being sold at clearance prices or newly introduced will be for our price index.

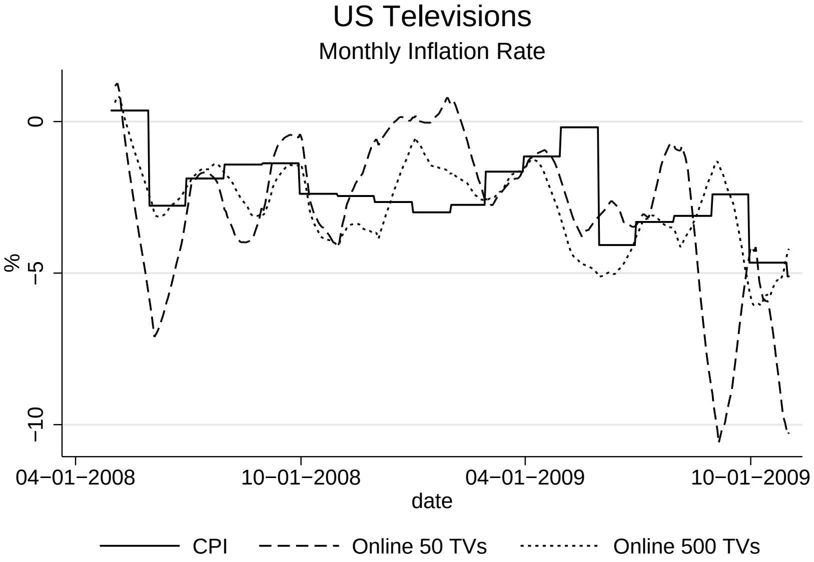


Figure 6: Hedonics CPI vs Online Index for US Televisions

Notes: The solid line shows the official Consumer Price Index for televisions as computed by the US Bureau of Labor Statistics using hedonic methods. The line with long dashes shows an online price index based on 50 distinct models of televisions from a large US retailer. The line with short dashes shows an online price index with 500 models from the same source.

This example illustrates one of the “size” advantages of online datasets. We may not need or want to use every single data point available in these data, but being able to extract and use uncensored spells for a large number of models can greatly simplify measurements. Even if the goal is to run a hedonic regression, online data can supply the detailed information needed to make it practical. And with more data, simpler methods can be applied. For example, Krsinich (2015) showed that online data can be used to construct a time-product dummy index that is equivalent to a fully-interacted time dummy hedonic index based on all product characteristics.

3.3 Anticipation of Future Changes in the Consumer Price Index

As mentioned before, online price indexes can sometimes anticipate changes in official inflation. In this section, we document this pattern more formally and conjecture about some possible explanations. To document the degree of anticipation, we estimate a simple auto-regression equation with the Consumer Price Index as the dependent variable and our online price index as the exogenous variable, and compute an impulse response to see how shocks to the online index impact the official price index over time. The regression is expressed in monthly changes: specifically, we use monthly log changes in the Consumer Price Index and monthly log changes of the online index on the last day of each month. We include six lags of each variable, plus the contemporaneous value of the online price index to account for the early availability of the online price information.[5]

Figure 7 shows the cumulative impulse response of the CPI to a shock in the online index over time, together with the 95 percent confidence intervals.

[5] For each month t, the specification is as follows:

$$\Delta ln(CPI_t) = \alpha + \beta \Delta ln(Online_t) + \sum_{i \in [1,6]} \alpha_i \Delta ln(CPI_{t-i}) + \sum_{i \in [1,6]} \beta_i \Delta ln(Online_{t-i}) + \epsilon_t \qquad (1)$$

The autoregressive distributed lag (ADL) specification is equivalent to a vector auto regression (VAR) with an exclusion restriction. The confidence bands are computed by bootstrapping in blocks. This specification gives the online price index the highest chance to explain the observed variation. There is, however, no unambiguous way of identifying the system given that under the null hypothesis both indexes are valid measures of the underlying inflation. We chose this specification because it matches the actual availability of data at the end of each month: the online index is immediately available, while the CPI has a publication lag of 15 days in most countries. The results are robust to the elimination of the contemporaneous effect of the online price index from the equation.

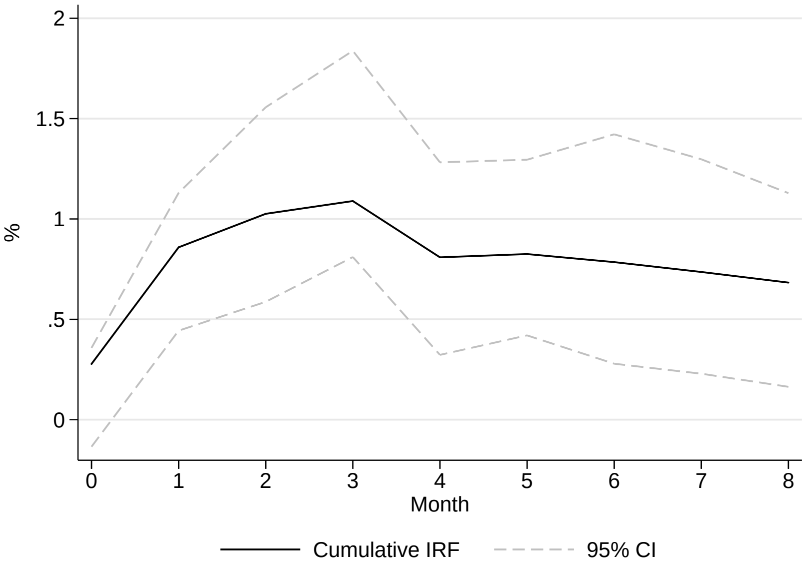


Figure 7: Impulse Response of the US CPI to a Online Price Index shock

Notes: Cumulative impulse response to a 1% shock in the online index. The online price index was computed by PriceStats. The CPI is a US city average, all items, non-seasonally adjusted from the Bureau of Labor Statistics. Data from July 2008 to January 2015.

In the US, it takes several months for the Consumer Price Index to fully incorporate the shock to the online price inflation. At the sector level, the impact is quickest in fuel (transportation) and slowest in food and electronics (see Appendix for details). The result is robust to the elimination of the contemporaneous effect of the official price index from the vector autoregression. In most cases, the anticipation significantly exceeds the typical publication delays in official statistics. Moreover, we find similar degrees of anticipation in other countries.

Possible reasons for why online prices can anticipate shocks in the Consumer Price Index include delays embedded in the methodology used for the official data, differences in mixture of stores sampled, and faster adjustment of online prices in some sectors or retailers. Understanding what drives the anticipation is sill an open question for future research, but the patterns in Figure 7 suggest that online data can be a useful addition to inflation forecasting models. This is explored by Aparicio and Bertolotto (2016), who show that out-of-sample inflation forecasts using online data can outperform a large number of alternative forecasting models in the US and UK economies.

4 Lessons for Macro and International Research

In this section, we illustrate how online data can change empirical results in macro and international research by focusing on price-stickiness and real exchange rate behaviors. Our main objective is to show how online datasets constructed to fit specific research needs can help mitigate biases and other empirical challenges that are so frequent in traditional datasets collected for other purposes.

4.1 Price Stickiness and the Distribution of Price Changes

Sticky prices are a fundamental element of many macroeconomic models. In the past decade, a large empirical literature has tried to measure price stickiness and understand its micro-foundations (for an example in this journal, see Dhyne et al. 2006; for a survey of the literature, see Nakamura and Steinsson 2013, and the references cited there). This research has been possible due to an unprecedented access to micro-level Consumer Price Index data and scanner datasets in several countries. Over time, the literature has settled on a set of stylized facts, summarized by Klenow and Malin (2010). In Cavallo and Rigobon (2011) and Cavallo (2015b), we use online data to argue that the sampling characteristics of official Consumer Price Index and scanner data can introduce measurement biases that affect the stylized facts in the literature on patterns of price changes.

As one prominent example, a pattern that has received a lot of attention in the literature is the shape of the distribution of the size of price changes. Most papers using scanner or Consumer Price Index data found bell-shaped (unimodal) distributions centered around zero percent with a significant share of small price changes, which seemed inconsistent with standard menu-cost models that predict periods of unchanging prices followed by relatively large changes (a bi-modal distribution centered around zero). This finding motivated a surge in papers trying to adapt sticky-price models to account for this fact (for example, Woodford 2009; Midrigan 2011).

However, the shape of the distribution of price changes is greatly affected by the sampling characteristics of the data. This can be seen in Figure 8, where we show a distribution of prices changes for both online and scanner data obtained from exactly the same US retailer, zip code, and time period. While the raw prices that generate these distributions are in principle the same, the results are strikingly different. The online data distribution is strongly bimodal, with very few price changes close to zero percent. There is a simple explanation for the difference. Scanner data are reported as weekly averages.

As noted by Campbell and Eden (2014), this can create a large number of spurious small changes. For example, in a three-week period with a single price change in the middle of the second week, taking weekly averages would yield two small price changes: one from the first week to the second, and another from the second week to the third. These spurious changes can be seen explicitly in Figure 8, where we approximate the shape of the scanner data distribution by simply taking weekly averages of the raw online data.

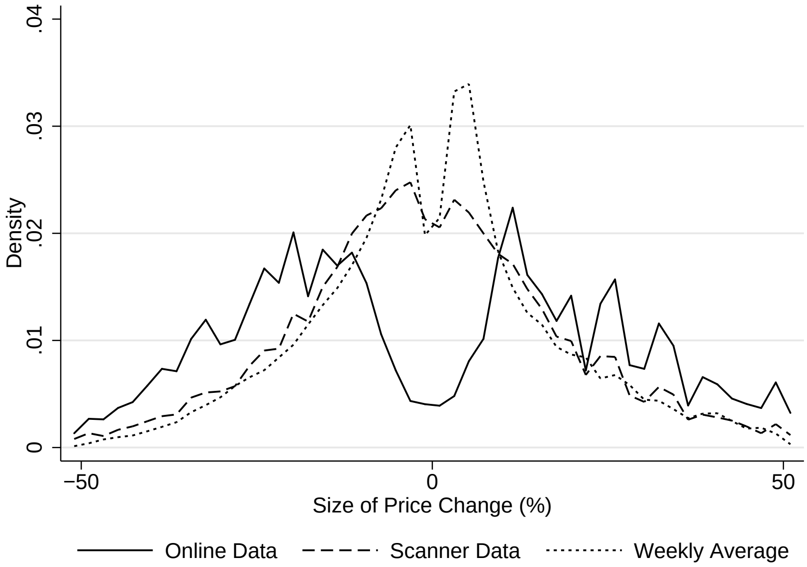


Figure 8: The Distribution of the Size of Price Changes in the US

Notes: Online and Scanner data collected from the same retailer, zip code, and time period. Weekly averages computed using the daily online data. Nielsen Scanner Data provided by the Kilts Marketing Center at Chicago Booth. The figure is from Cavallo (2015b).

Something similar happens with Consumer Price Index data, although the source of measurement bias is different in nature, as discussed in Cavallo (2015b). In particular, micro data from the Consumer Price Index will often contain imputed prices for temporarily missing items, which is a sensible thing to do when measuring inflation. This imputation is often done with the average price change of related goods, resulting in an artificial pattern of many small changes. If these imputations are not identified or removed when generating the distributions, the result is a unimodal-shaped distribution similar to those found in the literature. Furthermore, other forms of measurement biases can have a similar impact. For example, Eichenbaum et al. (2014) use Consumer Price Index and scanner data from multiple stores to show how “unit-value prices,” which are reported as the ratio of sales revenue to the quantity sold, also create a large number of spurious small changes.

Controlling for measurement bias is important, but to better understand price stickiness and its determinants, the literature also needs data with similar characteristics from multiple countries and economic settings. This is very hard with traditional data sources. For example, to obtain frequency estimates in 24 countries, Klenow and Malin (2010) had to source them from 27 different papers, each with its own particular data and methodologies. Appearing in this journal, Dhyne et al. (2006) is one of the few papers with data from multiple countries, thanks to the coordination provided by the European Inflation Persistence Network. But even in this case, each European national statistical organization was unwilling to share its micro data with Eurostat, so the frequency analysis had to be conducted independently in each country by a different team, each facing a dataset with different characteristics.

Instead, online prices have the potential to provide datasets with identical sampling characteristics in a large number of countries. At the Billion Prices Project we are currently working to standardize stickiness statistics in all our data and to be able to produce them on a ongoing basis. The goal is not only to share with other researchers a wide range of indicators that can be used to study stickiness, but also provide policy-makers with more up-to-date information about its behavior over time.

4.2 International Prices and the Law of One Price

The global nature of online data also makes it appealing for research in international economics. In particular, the relation between relative prices and exchange rates is a classic question in international economics. A basic hypothesis is the "law of one price," which implies that there should not be large or persistent cross-country differences in the prices of identical goods when translated into a common currency. When considering a group of many traded goods, the law of one price implies that exchange rates and relative prices will adjust to maintain stable purchasing power parities ("PPP"). Modest deviations from PPP are not surprising in a world with transport costs and other barriers to arbitrage. However, a huge literature documents a large failure of the law of one price for many traded goods at retail prices, resulting in significant volatility in the relative cost of consumption across countries. This failure occurs not only in price levels ("absolute PPP"), but also in changes over time ("relative PPP"). Furthermore, nominal exchange rate shocks tend to have persistent effects on the real-exchange rate, leading to what Kenneth Rogoff called the "PPP puzzle." At the core of this puzzle is the fact that relative prices do not seem to adjust quickly to nominal exchange rate shocks. Many papers have documented the slow response of prices by measuring very low exchange "pass through" rates.[6]

The literature concerning the law of one price and PPP is hampered by the formidable difficulties in obtaining prices in a large number of identical goods sold simultaneously in a large number of countries, as discussed by Taylor (2001). In practice, researchers are forced to settle on having prices for identical goods from two countries (typically the US and Canada), or use price indexes from a large number of countries (constructed with different methods and baskets, and precluding any price level comparisons). Some micro sources of data, such as the index published by The Economist magazine based on the prices of McDonalds Big Mac sandwiches, provide information on many countries but are limited to a single good. The World Banks International Comparison Program makes a worldwide effort to collect price data and to estimate PPP-adjusted GDPs in dozens of countries. But carrying out this task with traditional methods of collecting prices and adjusting for quality is so daunting that it can only be done every five years or so, severely limiting its use for research on real-exchange rate levels and dynamics.

In principle, online prices can be obtained in high frequency, for large number of goods, in dozens of countries. The main challenge is not in the raw data collection, but rather in matching identical products across countries, as product identification codes in the data tend to be specific to the good, country, and retailer where the product is sold.

In Cavallo, Neiman, and Rigobon (2014a) we addressed the matching problem by using prices collected from global retailers such as Apple, Ikea, Zara, and H&M, who sell identical goods with the same identifying information in several dozen countries. This allowed us to directly study conditions under which the law of one price holds. Much to our surprise, we found that the law of one price only holds well in countries that share the same currency: for example, countries within the euro area, or countries that use the US dollar such as El Salvador and Ecuador. What really seems to matter for these global retailers is simply whether prices have to be shown to customers in the same currency, not whether countries are physically close, in a trade union, or even strongly pegging their currencies. In Cavallo, Neiman, and Rigobon (2014b) we used the introduction of the euro in Latvia in January 2014 to show that the adjustment towards the law of one price can take place within a matter of days after a country joins a currency union. This type

[6] See Rogoff (96) for a description of the "PPP puzzle" and Taylor and Taylor (2004) for a review of the PPP literature. Burstein and Gopinath (2013) provide a review of the empirical literature on relative prices and exchange rates, and a discussion of some theoretical advances, including accounting for non-tradeables or tradeables that are only locally consumed, variable markups, and pricing-to-market.

of price convergence was, after all, one of the objectives of the euro.

The main implication of this line of work is that choice of currency units is far more important for defining the boundaries between markets for goods than has previously been suspected. Conversely, factors that were traditionally thought to be important - such as physical distance, political and tax territories, language, and culture- do not seem to matter as much. Furthermore, these patterns also point to the importance of customer psychology, organizational structure, and the internet for price-setting behavior. For example, firms may fear antagonizing customers who see prices posted on the web in the same currency across borders. Such considerations do not yet feature prominently in most macroeconomic models.

Ideally though, for some applications we need to have both global and local retailers. So since 2014, PriceStats has been expanding the product matching to include local retailers as well, classifying over 30,000 individual goods into 300 product categories. The challenge is to classify a large set of heterogeneous individual products (with varying package sizes, flavors, and retailers where they are sold) into narrowly defined product categories such as “Basmati White Rice, 1kg” or “LG Basic Blu-Ray Player, 1 unit.” This is achieved by using supervised machine learning (specifically a “Naive Bayes” classifier) that trains on language-specific, hand-categorized items. The process is described in detail in Bertolotto (2016). The output resembles a collection of hundreds of “Big Mac”-type indexes for different kinds of goods.

These matched indexes can be used to study real-exchange rate levels and dynamics, as in Cavallo and Neiman (2016). To illustrate this, Figure 9 shows PPP metrics constructed by PriceStats for an average of more than 250 goods in food, electronics, and fuel, in Argentina and Australia relative to the United States (examples for other countries are provided in the online Appendix available with this paper at http://e-jep.org). The top panel shows the relative prices (in local currencies) and the nominal exchange rate (defined as local currency per US dollar). For the case of Argentina, we also plot the black-market exchange rate. The bottom panel shows the real exchange rate constructed from the other two variables (as the ratio between the relative prices and the nominal exchange rate). This is simply the relative cost of the basket when expressed in the same currency.

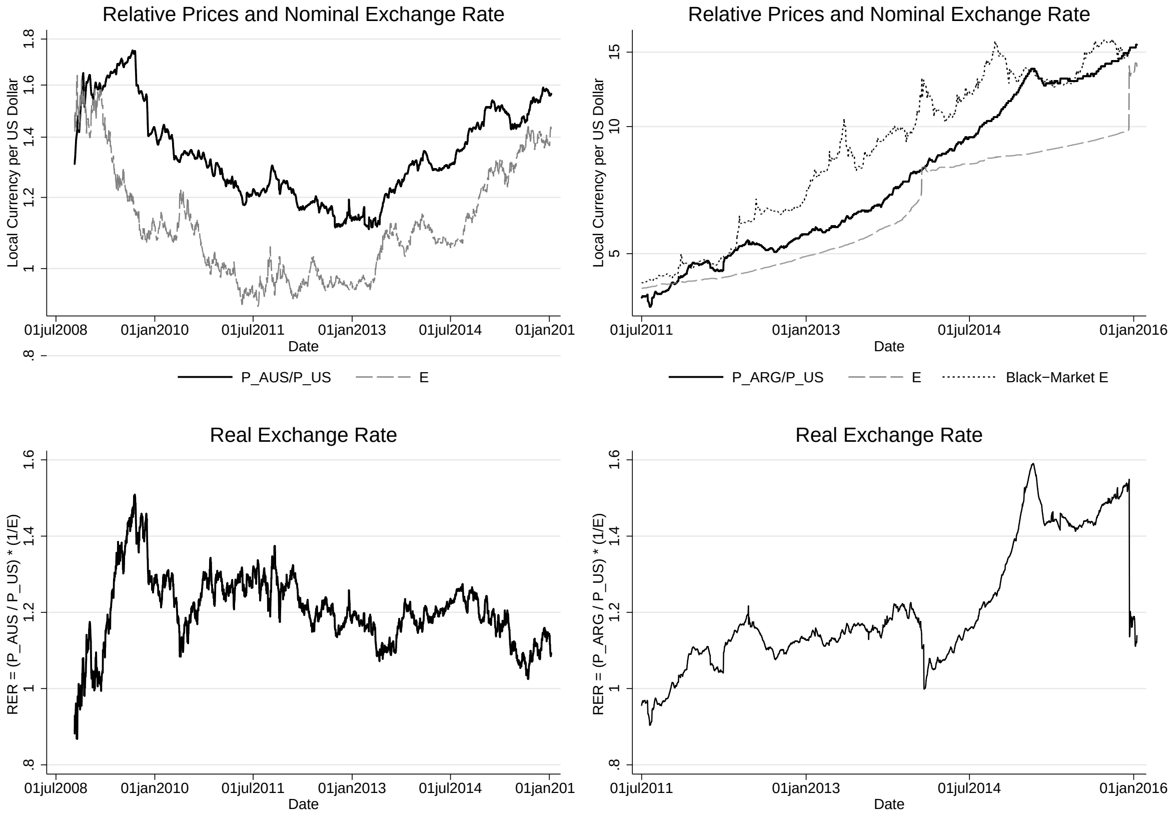


Figure 9: Relative Prices and Exchange Rates

Notes: The top 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 bottom 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 expenditure weights for food, fuel, and electronics.

A common finding, present also in other countries, is that relative prices co-move closely with the nominal exchange rate movements. For example, as the Australian dollar appreciated from 2008 to 2011, relative prices in Australia fell to compensate, and when the Australian dollar started to depreciate again in 2013, relative prices rose. In Argentina, the steady increase in relative prices was matched by the overall trend of depreciation in the currency, which is gradual in the black-market and lumpy on the official exchange rate. There are long periods where prices kept rising and the official exchange rate was held fixed by the government, causing "deviations" in the real-exchange rate, but there were sudden adjustments in the two occasions when the country devalued its currency, in January 2014 and December 2015.

This co-movement between relative prices and exchange rates implies high rates of pass-through, which can go in both directions. In Australia there is evidence that nominal exchange rate shocks affect retail prices (as the literature tries to capture in traditional "pass-through" estimates). In Argentina, retail price movements tend to precede nominal exchange rate adjustments.

Another unique feature of online data is that they provide information on relative price *levels*, which are not available when using Consumer Price Indexes. For example, the real exchange rates in Figure 9 shows that the basket tends to be 20 percent more expensive in Australia relative to the US. In Argentina, the cost is about 10% higher than in the US when the currency is allowed to float. Recognizing these patterns is useful for estimating the degree of currency misalignment at different points in time, particularly in countries with managed exchange rates.

For example, in December 2015 the new government of Argentina wanted to remove all foreign-exchange market restrictions. It was unclear what the free-market exchange rate would be, and what effect it would have on tradable prices. The nominal exchange rate implied by purchasing power parity was 14.3 pesos per dollar, suggesting that the official rate of 9.6 pesos per dollar was greatly overvalued while the black market rate of 15 pesos per dollar was slightly undervalued. When the market was freed, the new exchange rate quickly settled around 14 pesos per dollar, closely matching the implied PPP exchange rate (the ratio of relative prices). This can be seen in the jump of the official exchange rate in the top right panel of Figure 9.

While we do not expect these metrics to help predict exchange rates so closely in every country and situation, they can provide better measures of the amount of deviation of real-exchange rates from "normal" levels at a given point in time.

So far, our micro data has only been matched for seven countries and the time series is still too short to make strong inferences, but it is clear that some key puzzles in international economic and macroeconomics that emerged from studies using official price indexes appear quite different when viewed through the perspective of online data.

5 Access to the Billion Prices Project Data

As an academic project, we share as much data and results as possible on our webpage (bpp.mit.edu). Most of the micro data and indexes used in our papers are currently available to download on that page, together with detailed scripts that allow others to replicate and extend our results. The micro data are posted with little pre-treatment, so other researchers can apply their own methods. We will upgrade the shared data periodically, both increasing the number of databases and retailers and also expanding the time series.

The US and Argentina inflation indexes used in this paper are published with a 30-day lag on the Billion Prices Project website, while the PPP exchange rate information discussed in the previous section are currently published with a one-year lag on the PriceStats website. The raw micro data collected by PriceStats are not publicly available but can be shared with academic researchers who collaborate with the Billion Prices Project and sign a data-access agreement.

6 Final Remarks

The need for economists to get involved in data collection was eloquently pointed out many years ago (1985) by Zvi Griliches (also see his Presidential Address at the American Economic Association in 1994). In his words,

> *We have shown little interest in improving [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.*

Big Data technologies are finally providing macro and international economists with opportunities to stop treating the data as "given" and get personally involved with data collection. We can now build datasets customized to fit specific measurement and research needs. This will help mitigate issues in empirical research such as sample selection, endogeneity, omitted variables, and error-in-variables, which are so frequent in traditional datasets.

The Billion Prices Project is just one example of the use of "Big Data" in economics.[7] Although online price data are the focus of this paper, we hope to have convinced other economists and perhaps a few policymakers of the benefits of experimenting with alternative data sources. Other examples include various types of "scraped" data, such as labor and real estate information available on the web, along with data from mobile phones,

[7] Einav and Levin (2014) provide a more general discussion of this topic, including new granular data sources, computational techniques such as machine learning, and the role of theory in analyzing large, unstructured datasets. In this journal, Varian (2014) describes in detail some new "Big Data" techniques that are useful to analyze large datasets.

satellite images, GPS signals, and many other sensors that are increasingly part of our daily lives.

While many governments have been active in searching for alternative data sources, hoping to increase the quality of statistics and to reduce cost, their use will require not only the will of policymakers and statisticians working on the field, but also the involvement of more economists and academics who can help identify the best ways to collect, treat, and use these new sources of information.

References

Aizcorbe, A. M., C. A. Corrado, and M. E. Doms (2000). Constructing price and quantity indexes for high technology goods. *Industrial Output Section, Division of Research and Statistics, Board of Governors of the Federal Reserve System, July 19.*

Aizcorbe, A. M., C. A. Corrado, and M. E. Doms (2003). When Do Matched-Model and Hedonic Techniques Yield Similar Measures? *FRB of San Francisco Working Paper* (2003-14).

Aparicio, D. and M. Bertolotto (2016). Forecasting Inflation with Online Prices. *Working Paper - MIT*.

Armknecht, P. A. and D. Weyback (1989). Adjustments for quality change in the US Consumer Price Index. *Journal of Official Statistics 5*(2), 107.

Bertolotto, M. (2016). Real Exchange Rates using Online Data. *Universidad de San Andres Working Paper.*

Breton, R., G. Clews, L. Metcalfe, N. Milliken, C. Payne, J. Winton, and A. Woods (2015). Research indices using web scraped data.

Brynjolfsson, E. and M. D. Smith (2000, April). Frictionless Commerce? A Comparison of Internet and Conventional Retailers. *Management Science 46*(4), 563–585.

Burstein, A. and G. Gopinath (2013). International Prices and Exchange Rates. *Handbook of International Economics (Forthcoming).*

Campbell, J. R. and B. Eden (2014). Rigid prices: evidence from U.S. scanner data . *International Economic Review 55*(2), 423–442.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, 152–165.

Cavallo, A. (2015). Scraped Data and Sticky Prices. *NBER Working Paper 21490.*

Cavallo, A. (2016). Are Online and Offline Prices Similar? Evidence from Multi-Channel Retailers. *NBER Working Paper 22142.*

Cavallo, A., E. Cavallo, and R. Rigobon (2014). Prices and Supply Disruptions during Natural Disasters. *Review of Income and Wealth 60.*

Cavallo, A., G. Cruces, and R. Perez-Truglia (2014). Inflation Expectations, Learning and Supermarket Prices: Evidence from Field Experiments. *NBER Working Paper 20576.*

Cavallo, A., G. Cruces, and R. Perez-Truglia (2016). Learning from Potentially-Biased Statistics: Household Inflation Perceptions and Expectations in Argentina. *NBER Working Paper 22103.*

Cavallo, A. and B. Neiman (2016). Real Exchange Rate Behavior: Evidence from Online Retailers in Nine Countries. *Working Paper - MIT*.

Cavallo, A., B. Neiman, and R. Rigobon (2014a). Currency Unions, Product Introductions, and the Real Exchange Rate. *Quarterly Journal of Economics 129*(2).

Cavallo, A., B. Neiman, and R. Rigobon (2014b). The price impact of joining a currency union: evidence from Latvia. *IMF Economic Review Forthcoming.*

Cavallo, A. and R. Rigobon (2011). The Distribution of the Size of Price Changes. *NBER Working Paper w16760*, 1–40.

Dhyne, E., L. J. lvarez, H. L. Bihan, G. Veronese, D. Dias, J. Hoffmann, N. Jonker, P. Lnnemann, F. Rumler, and J. Vilmunen (2006). Price Changes in the Euro Area and the United States: Some Facts from Individual Consumer Price Data. *Journal of Economic Perspectives 20*(2), 171–192.

Edelman, B. (2012). Using Internet Data for Economic Research. *The Journal of Economic Perspectives 26*(2), 189–206.

Eichenbaum, M., N. Jaimovich, S. Rebelo, and J. Smith (2014). How frequent are small price changes? *American Economic Journal: Macroeconomics 6*(2), 137–155.

Einav, L. and J. Levin (2014, November). Economics in the age of big data. *Science 346*(6210), 1243089.

Ellison, G. and S. F. Ellison (2009, March). Search, Obfuscation, and Price Elasticities on the Internet. *Econometrica 77*(2), 427–452.

Feenstra, R. C. and M. D. Shapiro (2003). Scanner Data and Price Indexes. *NBER.*

Gorodnichenko, Y., V. Sheremirov, and O. Talavera (2014). Price Setting in Online Markets: Does IT Click? *NBER Working Paper Series* (20819).

Griffioen, R., J. de Haan, and L. Willenborg (2014). Collecting clothing data from the Internet. In *meeting of the group of experts on consumer price indices*, pp. 26–28.

Griliches, Z. (1985). Data and econometriciansthe Uneasy Alliance. *The American Economic Review*, 196–200.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly 75*(5), 861–871.

Horrigan, M. (2013). Big Data: A Perspective from the BLS | Amstat News.

ILO, IMF, W. Bank, OECD, Eurostat, and U. Nations (2004). *Consumer Price Index Manual: Theory and Practice.*

Klenow, P. J. and B. A. Malin (2010). Microeconomic evidence on price-setting. In *Handbook of Monetary Economics*, Volume 3. Elsevier.

Krsinich, F. (2015). Price Indexes from online data using the fixed-effects window-splice (FEWS) index. *Statistics New Zealand.*

Midrigan, V. (2011). Menu Costs, Multiproduct Firms, and Aggregate Fluctuations. *Econometrica 79*(4), 1139–1180.

Mikians, J., L. Gyarmati, V. Erramilli, and N. Laoutaris (2012). Detecting Price and Search Discrimination on the Internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, HotNets-XI, New York, NY, USA, pp. 79–84. ACM.

Nakamura, E. and J. Steinsson (2013). Price Rigidity: Microeconomic Evidence and Macroeconomic Implications. *Annual Review of Economics 5*(1), 133–163.

Nygaard, R. (2015). The use of online prices in the Norwegian Consumer Price Index. *Statistics Norway*.

Rastello, S. and I. Katz (2013). Argentina Is First Nation Censured by IMF for Economic Data.

Rogoff, K. (1996). The Purchasing Power Parity Puzzle. *Journal of Economic Literature 34*(June), 647–668.

Silver, M. and S. Heravi (2005). A Failure in the Measurement of Inflation. *Journal of Business & Economic Statistics 23*(3).

Taylor, A. M. (2001). Potential Pitfalls for the Purchasing-Power-Parity Puzzle? Sampling and Specification Biases in Mean-Reversion Tests of the Law of One Price. *Econometrica 69*(2), 473–498.

Taylor, A. M. and M. P. Taylor (2004). The purchasing power parity debate. *Journal of Economic Perspectives 18*(4), 135–158.

Valentino-DeVries, J., J. Singer-Vine, and A. Soltani (2012). Websites Vary Prices, Deals Based on Users' Information. *Wall Street Journal*.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives 28*, 25.

Woodford, M. (2009). Information-constrained State-dependent Pricing. *Journal of Monetary Economics 56*, S100– S124.

Appendix

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

Alberto Cavallo
MIT & NBER

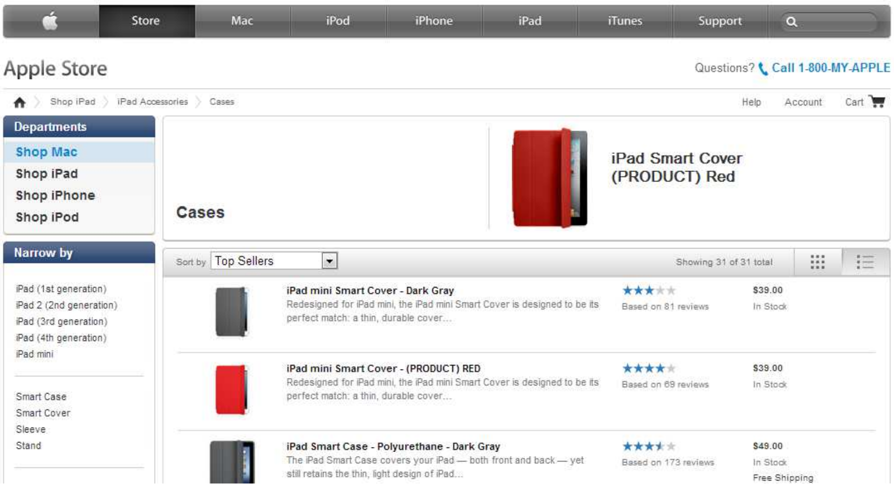
Roberto Rigobon
MIT & NBER

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.



iPad mini Smart Cover - Dark Gray
Redesigned for iPad mini, the iPad mini Smart Cover is designed to be its perfect match: a thin, durable cover...
★★★☆☆ Based on 81 reviews
$39.00
In Stock

iPad mini Smart Cover - (PRODUCT) RED
Redesigned for iPad mini, the iPad mini Smart Cover is designed to be its perfect match: a thin, durable cover...
★★★★☆ Based on 69 reviews
$39.00
In Stock

iPad Smart Case - Polyurethane - Dark Gray
The iPad Smart Case covers your iPad — both front and back — yet still retains the thin, light design of iPad...
★★★½☆ Based on 173 reviews
$49.00
In Stock
Free Shipping

```
<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.

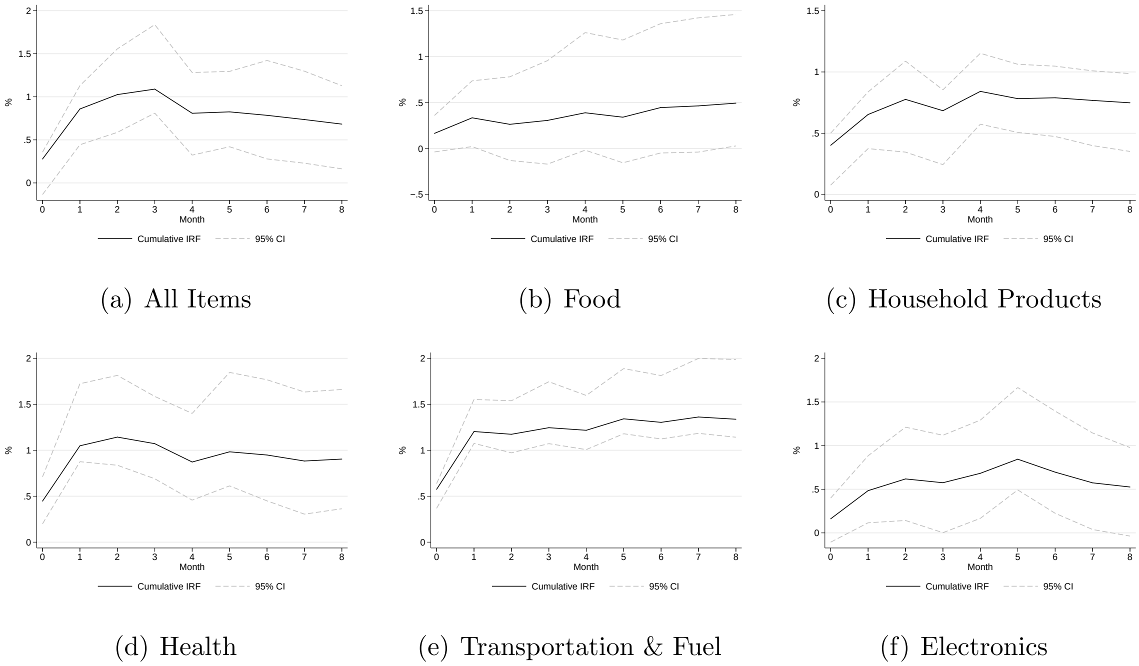
The set of URLs that the robot visits are carefully chosen by the scraping team based on the categories of goods that we want to sample. We also follow the robots.txt exclusion protocols of the servers where the data is located.

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

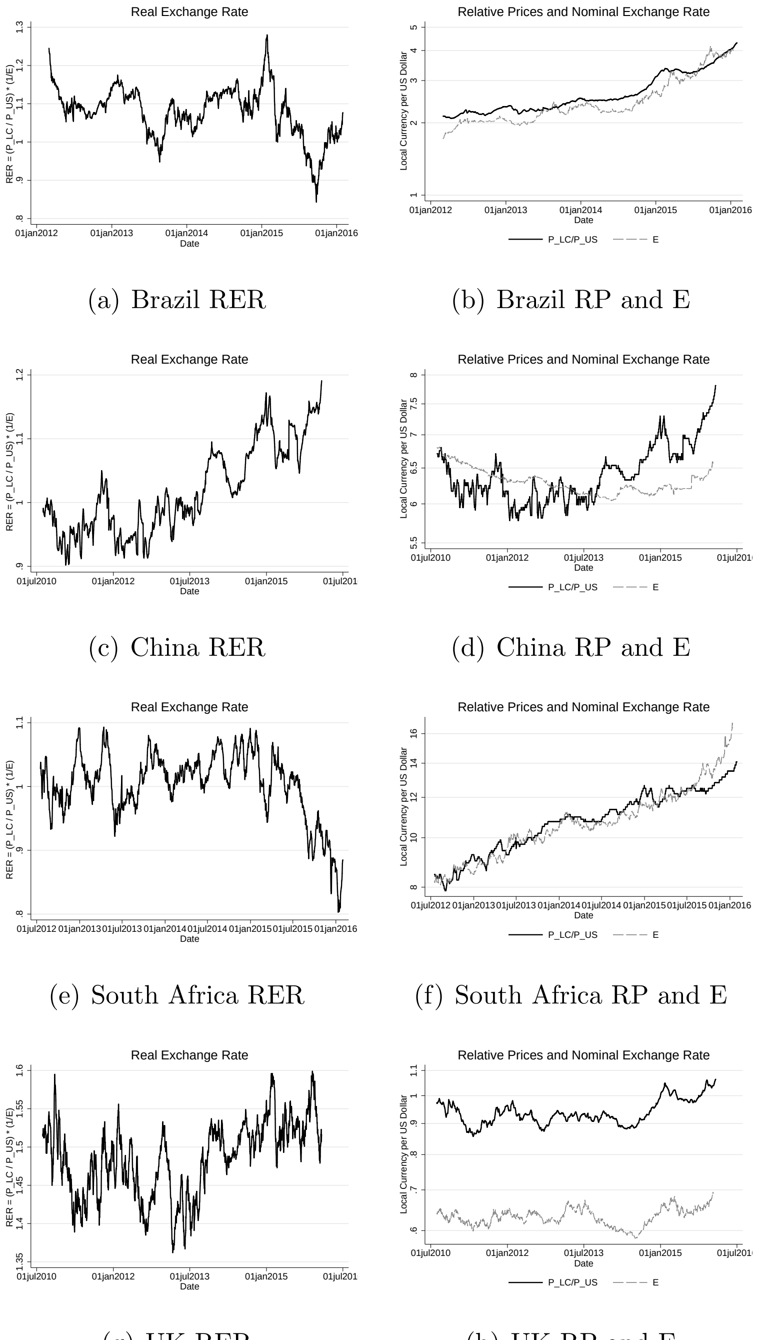
(e) Transportation & Fuel

(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.