

The Billion Prices Project

Using Online Prices for Inflation and Research

Alberto Cavallo
MIT & NBER

MFM Conference - NYU
January 2016

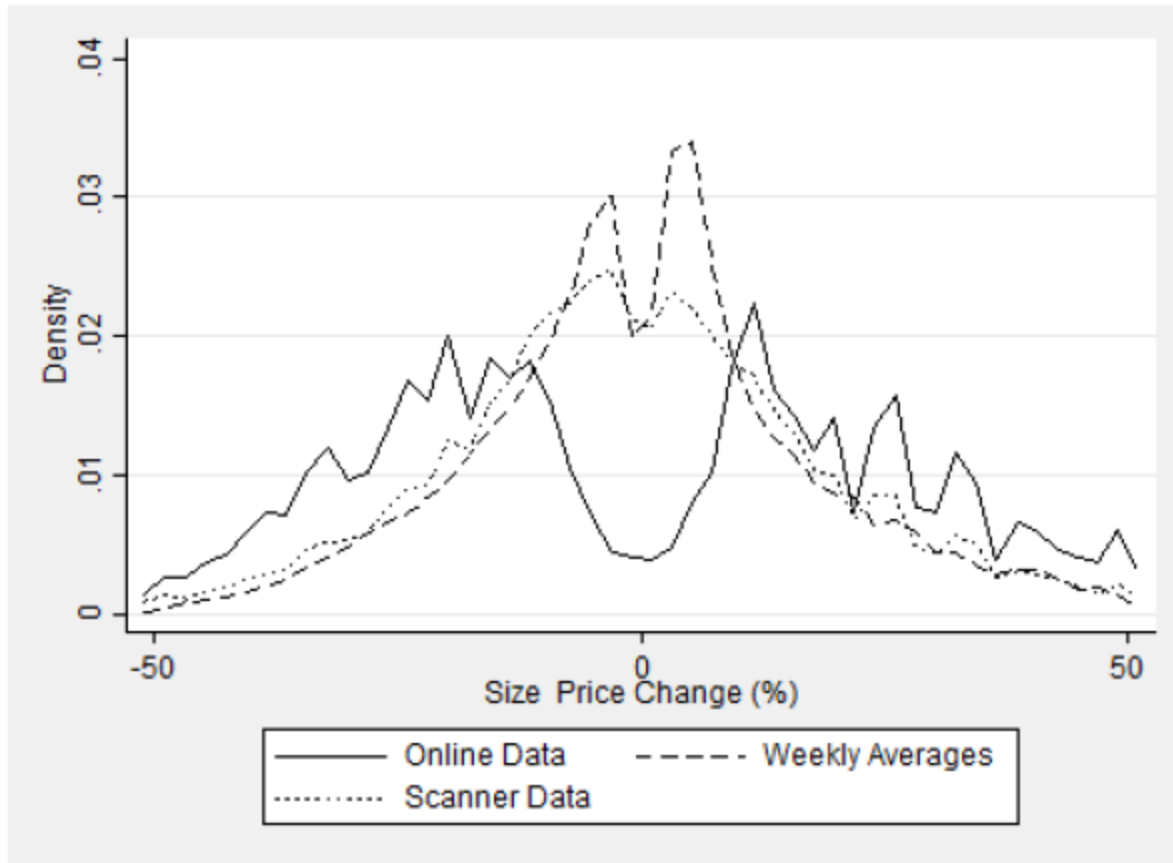
“Big Data” in Macro and International

- Quote from Griliches (AER 1985) on the “uneasy alliance” between economists and data:

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

- Big Data
 - A revolution in data collection technologies
 - Data collected to fit our research purposes

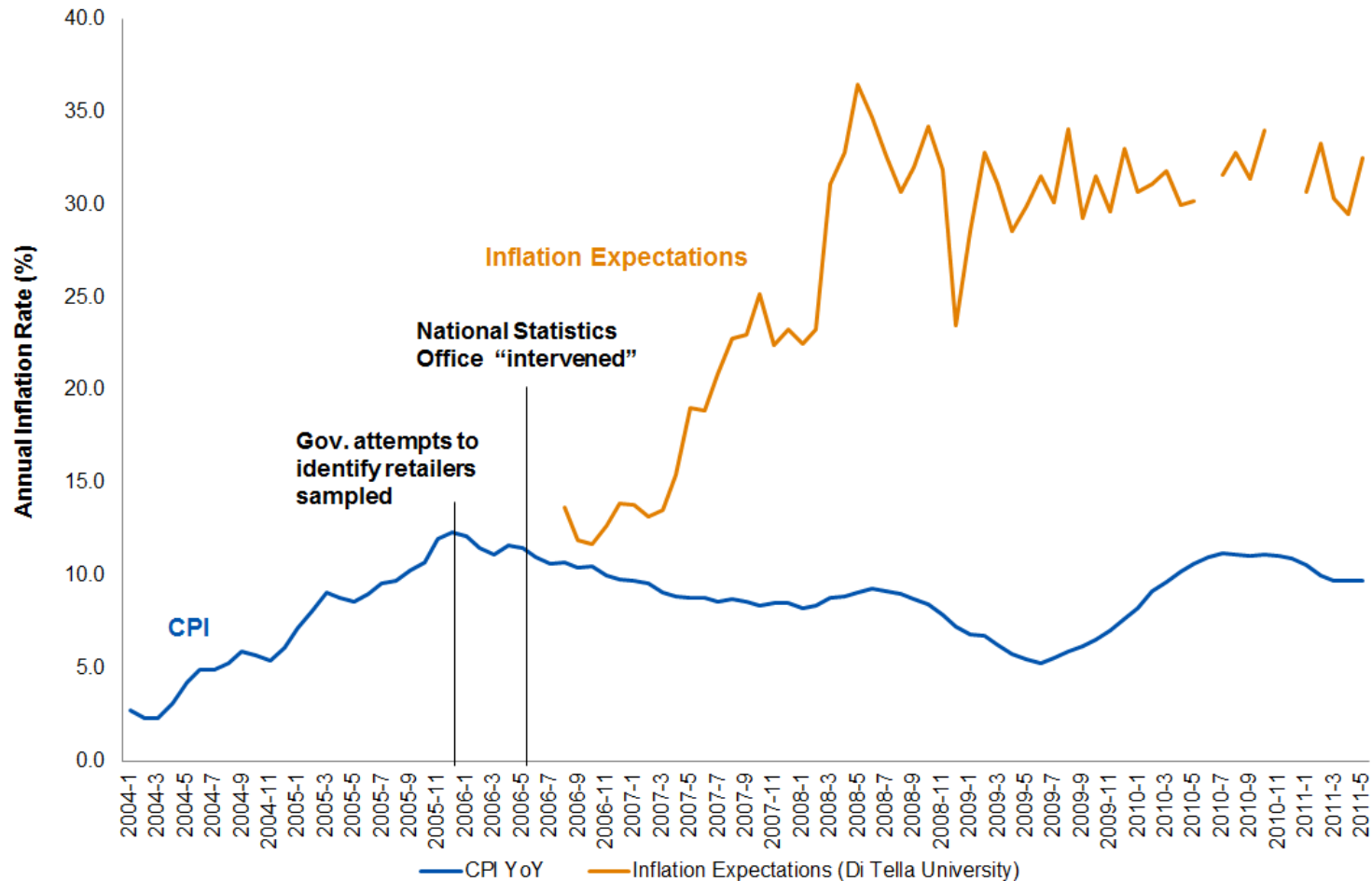
Do we need “*better*” data for research?



(a) Online vs Scanner

Do we need “*better*” data for inflation measurement?

An extreme case: Argentina since 2007

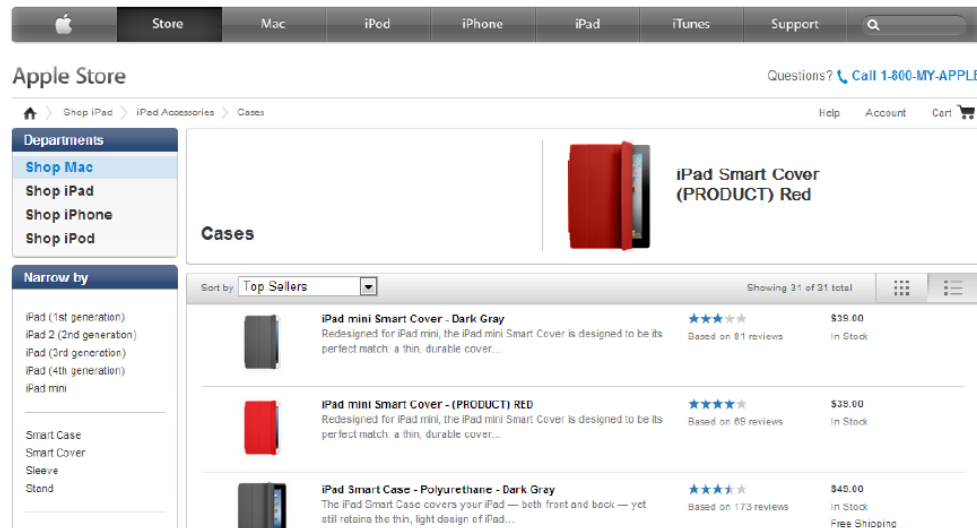


The Billion Prices Project at MIT Sloan

- Academic initiative to collect and use online price data for measurement and research applications
 - Use web scraping to collect prices from large multi-channel retailers
 - Daily data since 2008 from hundreds of retailers in 50 countries.
- Our Research is focused on
 - Inflation Measurement & Forecasting
 - Pricing Dynamics (Stickiness)
 - Real Exchange Rate and PPP

What is “Scraped Online Data”?

- Collected from public websites using web-scraping software
- A *robot* periodically downloads a public webpage, analyses its HTML code, extract price data, and stores it in a database



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

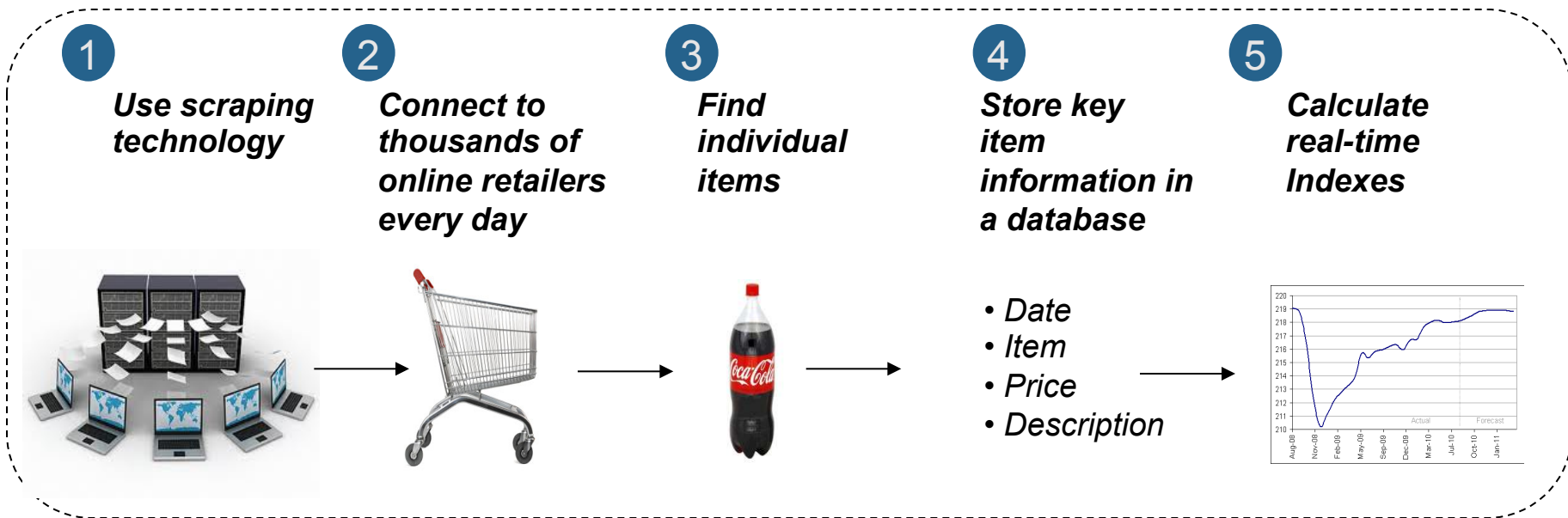
Advantages and Disadvantages

Online Data

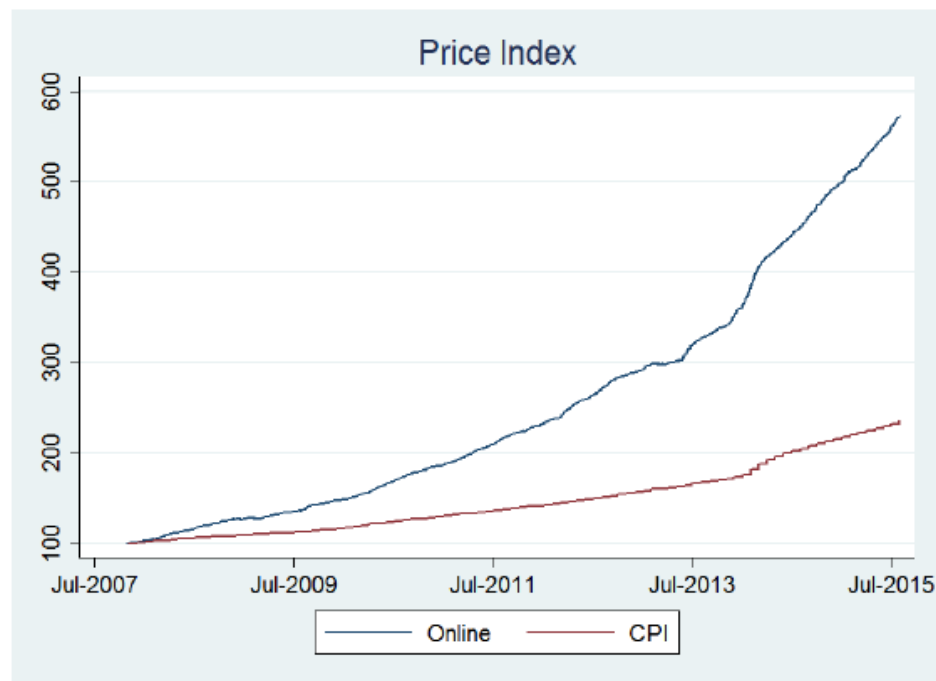
Advantages	Disadvantages
<ul style="list-style-type: none">• Cheap to collect• Frequency (daily)• Granularity<ul style="list-style-type: none">• All product details (brands, size, anything shown online)• All goods and varieties available for sale (census)• New goods automatically sampled• Easier to compare internationally	<ul style="list-style-type: none">• Not all categories of goods and services are online (not yet)• Fewer retailers and locations than CPI• Short time series relative to CPI• No quantities (present in Scanner Data)• Online and Offline prices may behave differently?

BPP and Daily Inflation Measurement

- In 2008 → daily price index for Argentina
- In 2010 → daily price index for the US on the BPP website
- Since 2011, PriceStats has been publishing daily inflation indices in 22 countries in real-time (3-day lag).



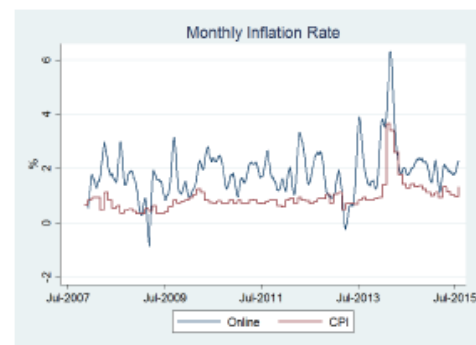
Argentina



(a) Price Index



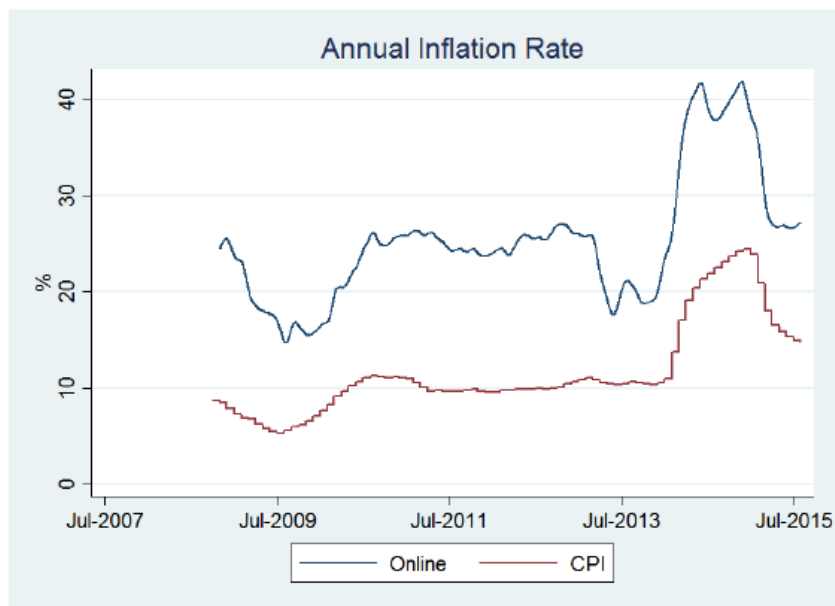
(b) Annual Rate



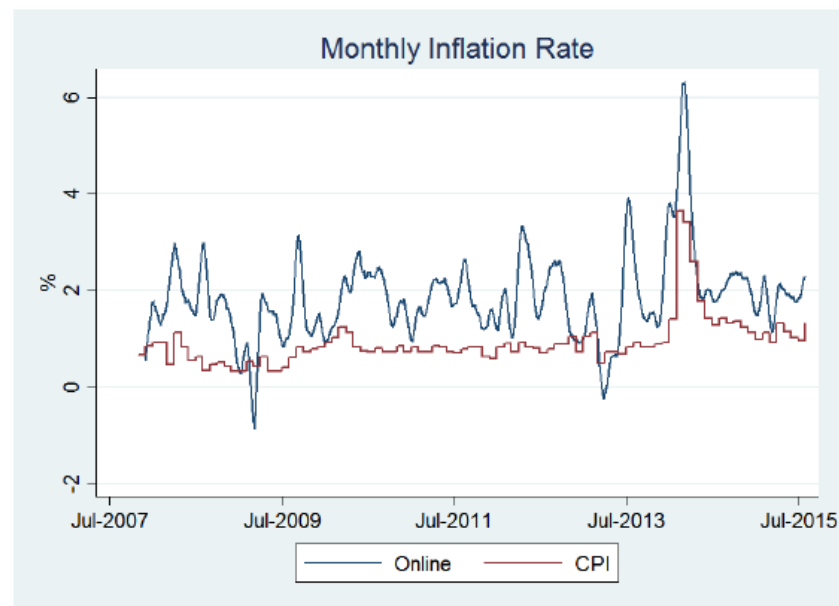
(c) Monthly Rate

Figure 2: Argentina

Argentina



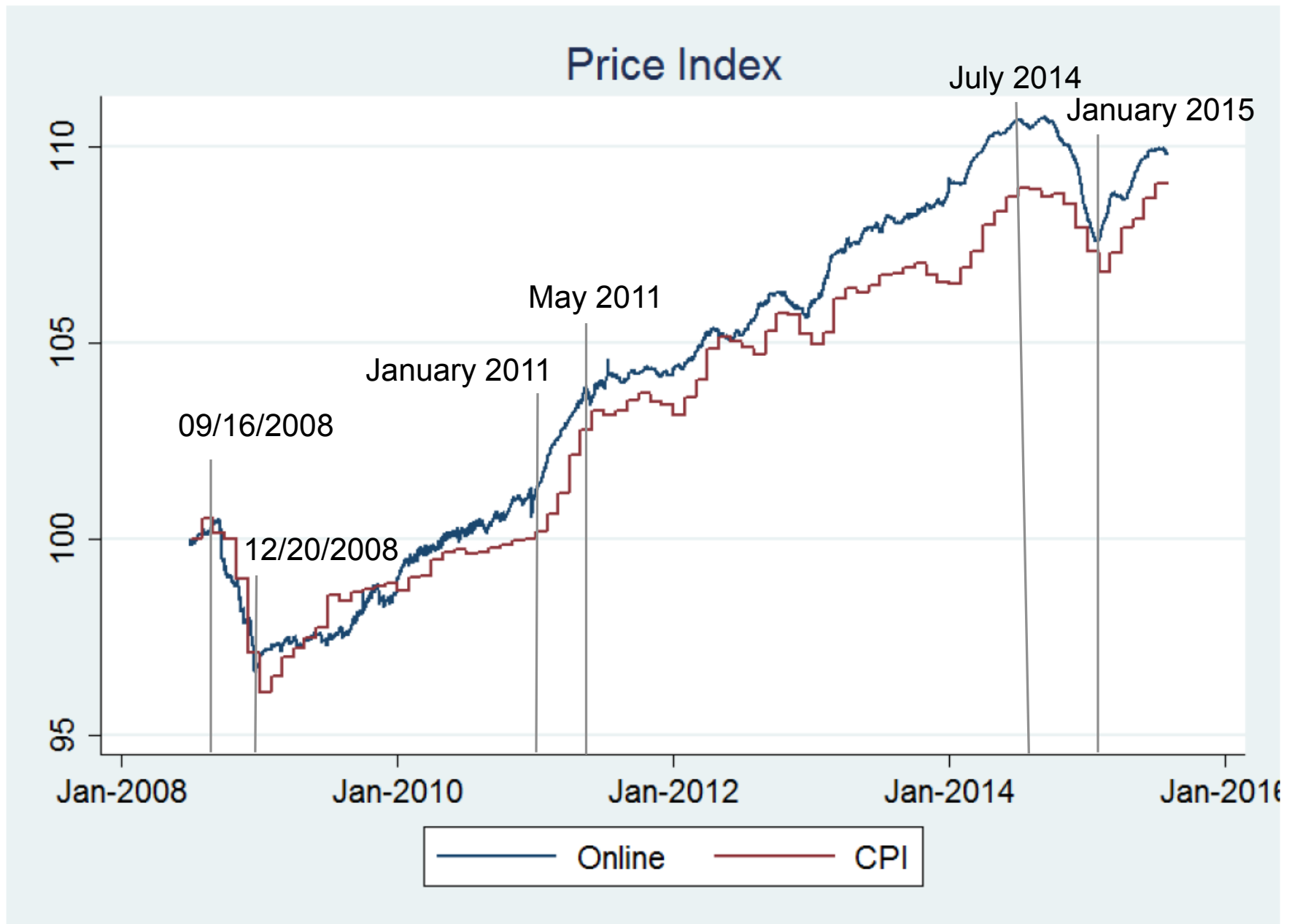
(b) Annual Rate



(c) Monthly Rate

Figure 2: Argentina

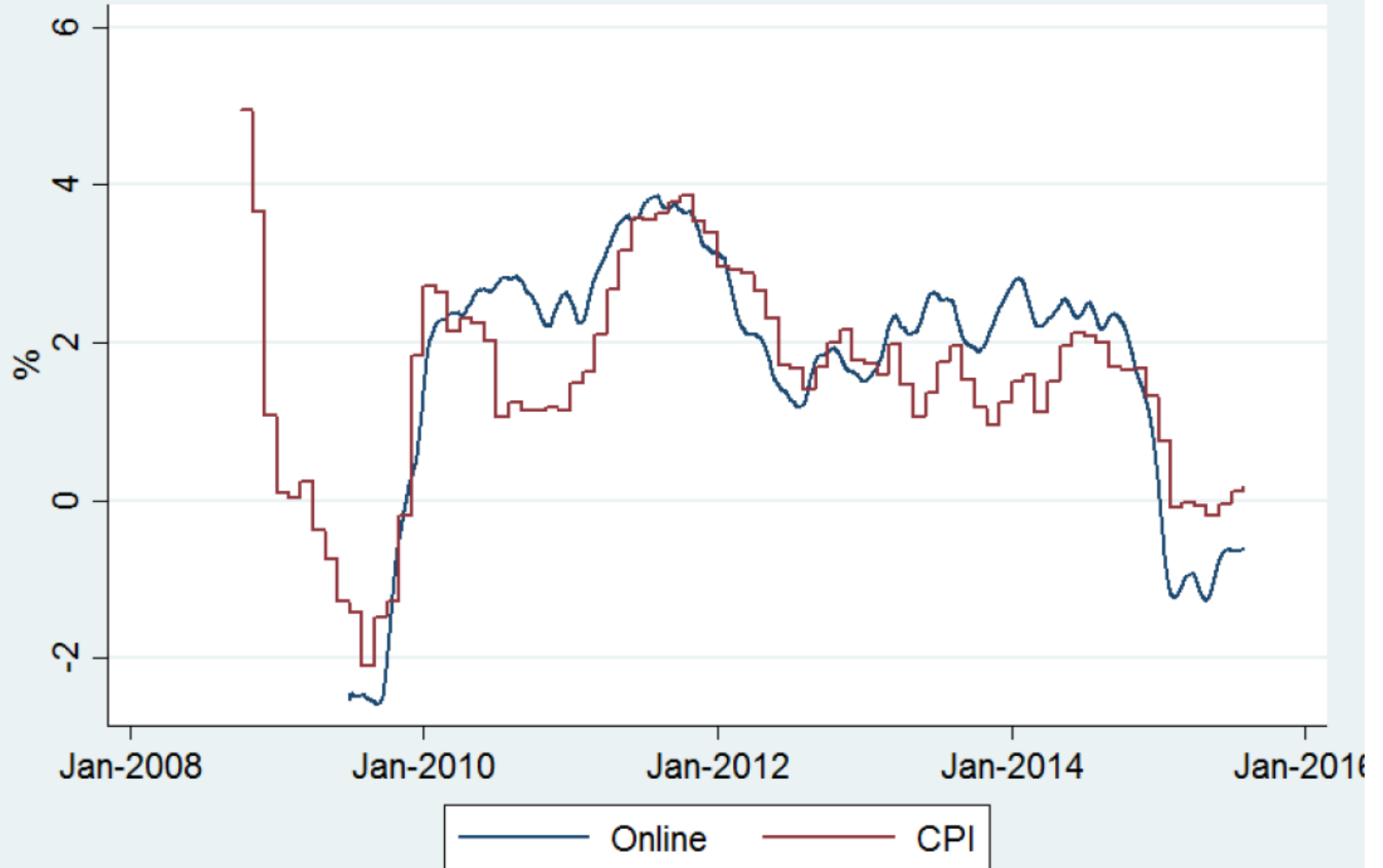
US Price Index



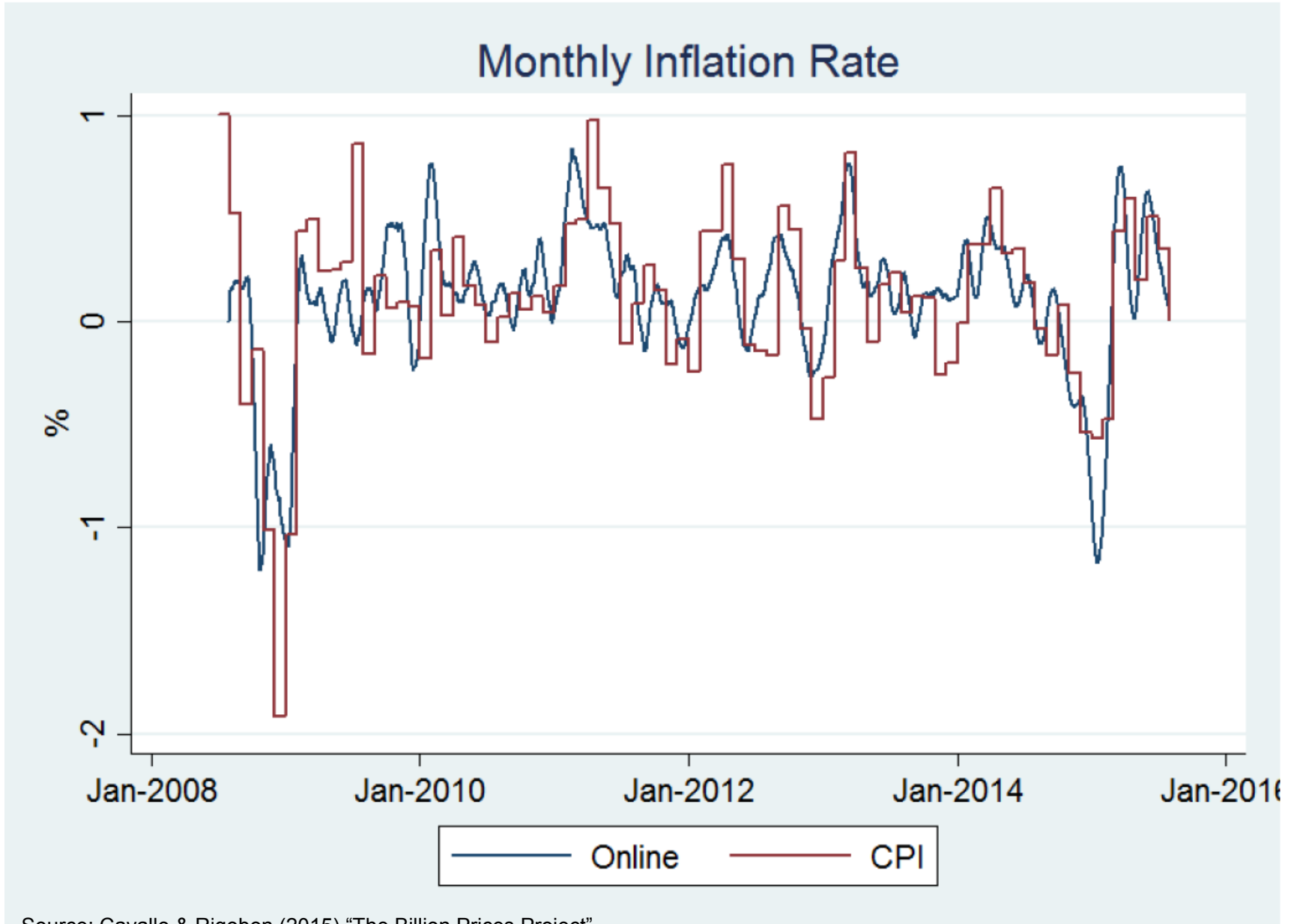
Source: Cavallo & Rigobon (2015) "The Billion Prices Project".

US Annual Inflation

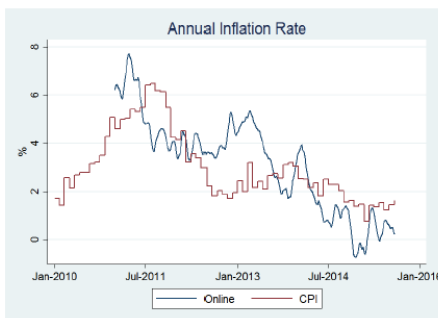
Annual Inflation Rate



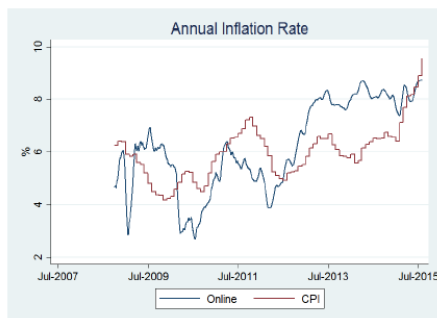
US Monthly Inflation



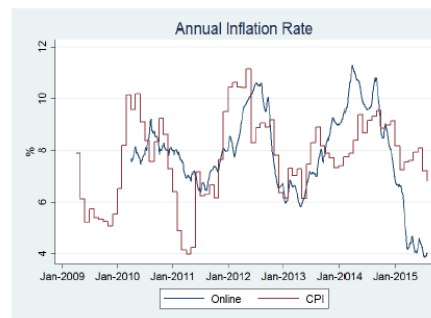
Developing vs Developed Countries



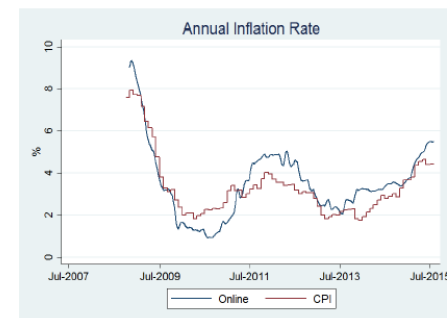
(a) China



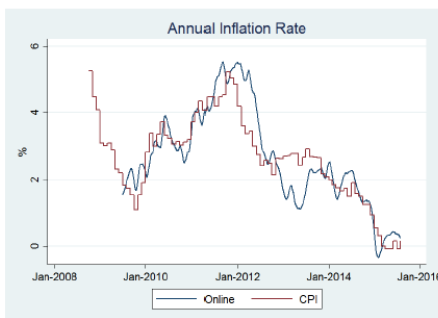
(b) Brazil



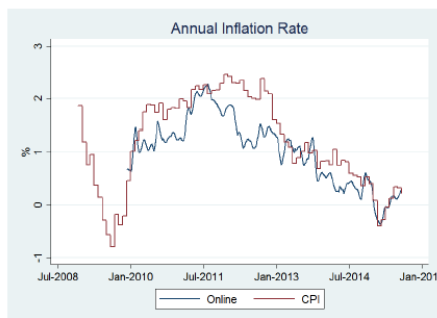
(c) Turkey



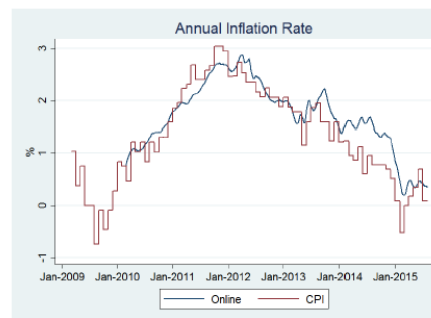
(d) Colombia



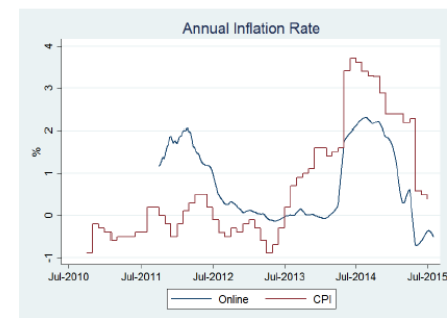
(e) UK



(f) France



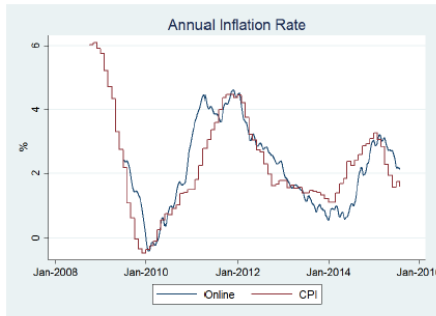
(g) Germany



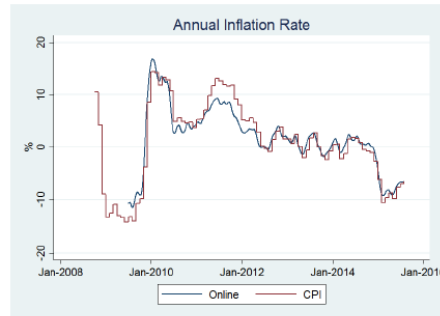
(h) Japan

Figure 5: Online vs CPI Annual Inflation Rates

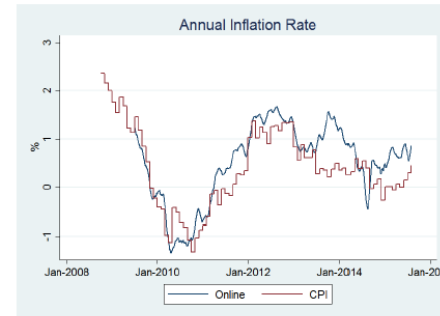
Sectors vs Global Aggregates



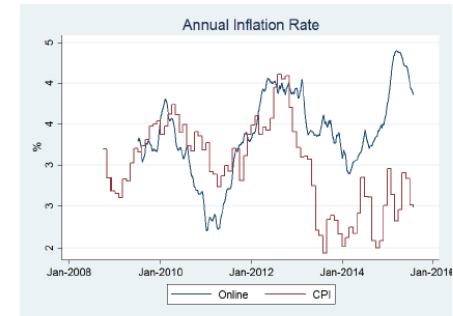
(i) USA Food



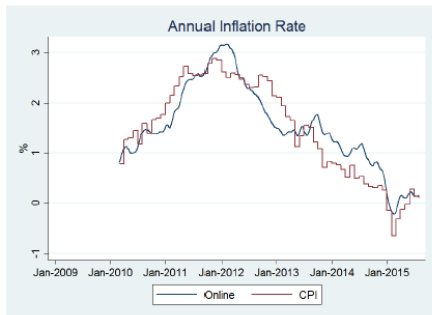
(j) USA Fuel



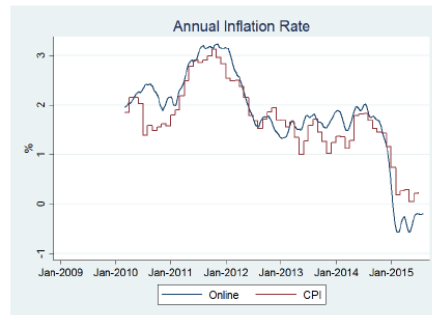
(k) USA Electronics



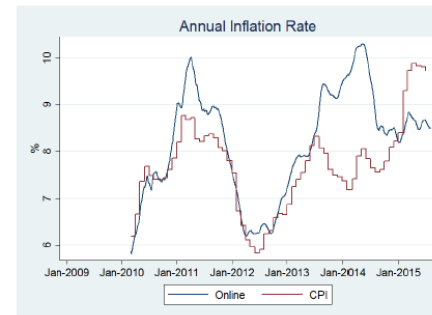
(l) USA Medical Care



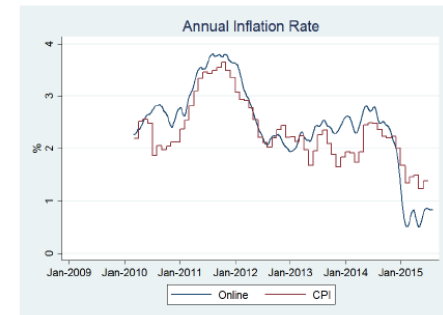
(m) Eurozone



(n) Developed M.



(o) Emerging M.

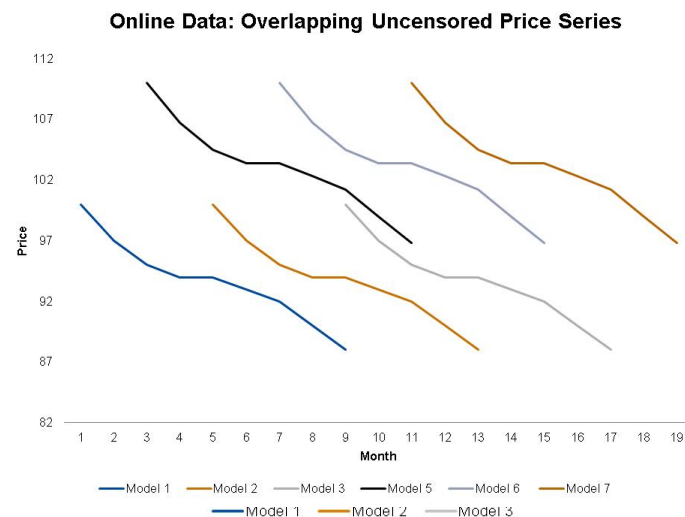
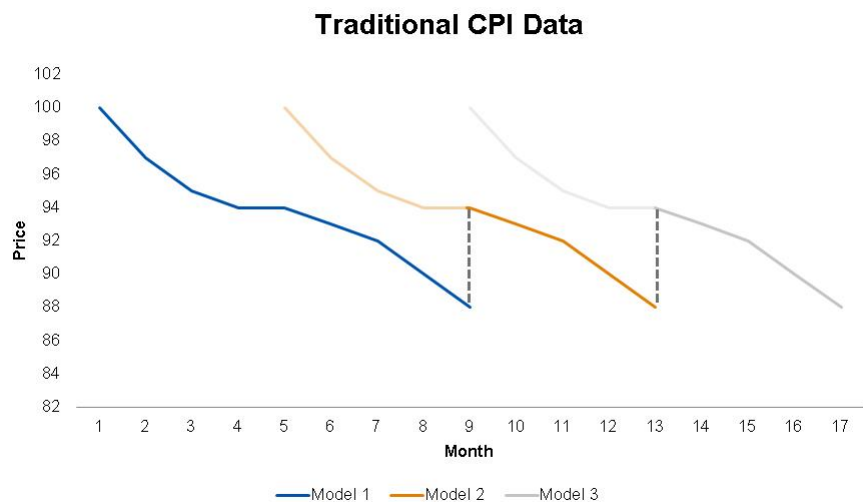


(p) World

Figure 5: Online vs CPI Annual Inflation Rates

Differences with CPI : Quality Adjustments

- Many complex techniques applied in CPI methods, such as hedonic quality adjustments, are needed because the data has inherent limitations
- Online data has “big data” advantages:
 - uncensored spells (automatically included at introduction)
 - all varieties/models on display



Differences with CPI : Quality Adjustments

- Simple indices can approximate the level and trend of CPI inflation in hedonic-adjusted categories (as suggested in Silver & Heravi (99), Aizcorbe, Corrado & Doms (2003))

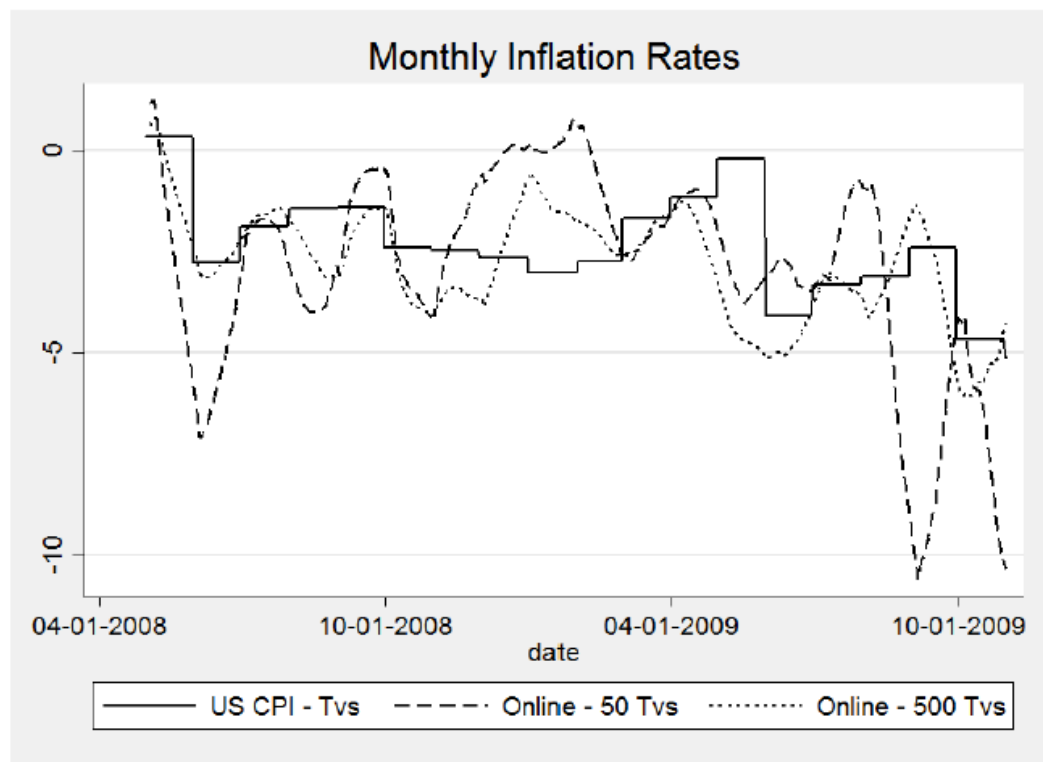


Figure 7: US CPI with Hedonics vs Online Jevons Index

Are Online and Offline Prices Similar?

- Simultaneous collection of online and offline prices in 50 large retailers in 10 countries

amazonmechanical turk
Artificial Artificial Intelligence

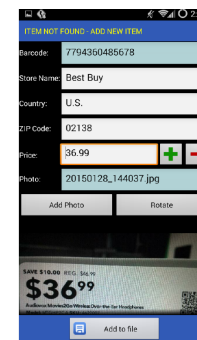
Elance upwork™
formerly oDesk

+



Android App

<https://play.google.com/store/apps/details?id=com.mit.bpp>



- Results:
 - 70% of price levels are identical ([see details](#))
 - Lowest in food/drugstores and highest in electronics/apparel.
 - Similar frequency and size of price changes

International Relative Prices

- Classic questions with lots of “puzzles”
 - Law of One Price and market segmentation
 - PPP puzzle → slow mean reversion of RERs
 - Low exchange rate pass-through
- Huge data limitations:
 - CPIs → no levels, different baskets, products, and methods
 - “Big Mac” indices → only 1 good!
 - World Bank ICP → low frequency (5 years)

International Relative Prices

- We use online prices to improve (i) coverage of countries, (ii) quality of matches, (iv) quantity of goods, and (iii) frequency of observations
- Main challenge: matching product ids across countries
- QJE 2014 : RERs for goods sold by Apple, IKEA, H&M, Zara, and other global retailers → $RER = 1$ in currency unions but not in pegs or floats.

Table 3: Absolute Value of Good-Level Log RER

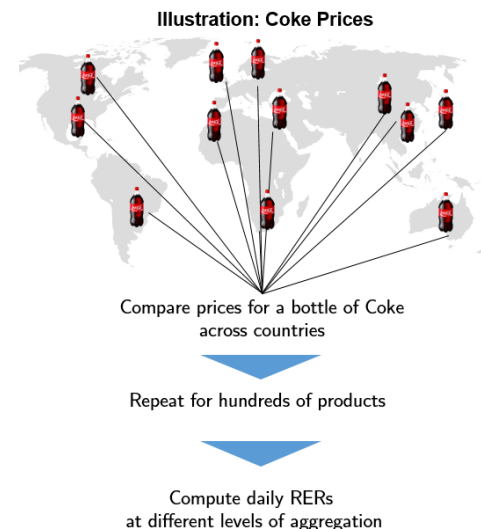
		All Stores	Apple	IKEA	H&M	Zara
(i)	Currency Unions	0.076	0.023	0.129	0.020	0.102
(ii)	NER Pegs	0.116	0.085	0.145	0.119	0.115
(iii)	Floats	0.187	0.143	0.216	0.145	0.207

Notes: Extracted from Cavallo et al. (2014a). Unconditional means of the average (across weeks in the data) of the absolute value of each good's log RER, separated by the currency regime. We exclude the small number of observations where $|q_{ij}| > 0.75$. Currency regime definitions closely follow Ilzetzi et al. (2008) and are described in Cavallo et al. (2014a).

International Relative Prices

- Now→ Much broader set of closely-matched goods sold by largest retailers in each country
- A “*Big Mac index*” with hundreds of products (e.g. “coffee, decaf, ground”) and thousands of varieties (brands, sizes, and retailers)

- 50 thousand individually matched items
- 300 narrowly defined products
- Food, fuel, and electronics
- Sector and Country-Level Indices
- Daily frequency
- 7 countries



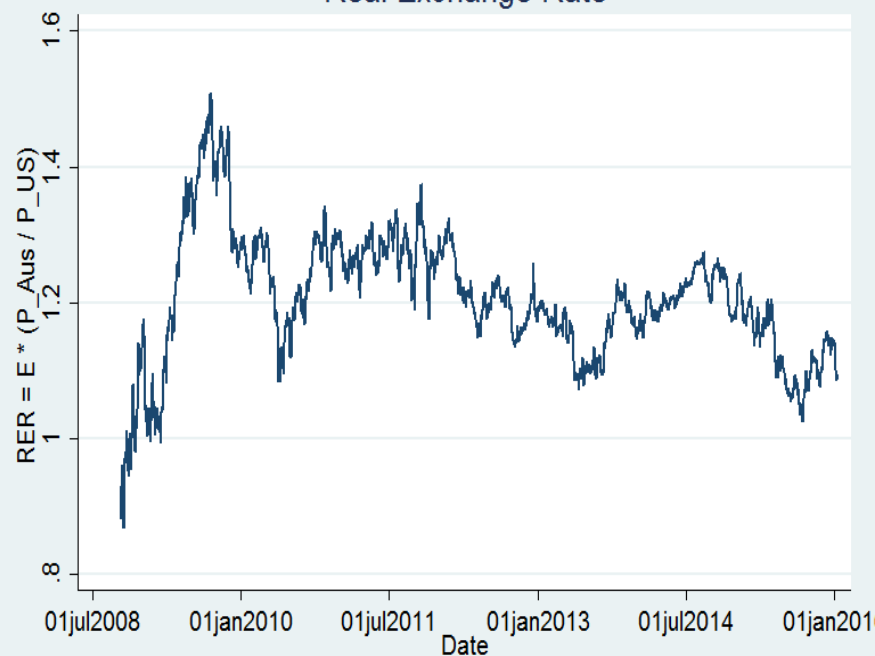
- We use machine learning to classify ids. The models train on language specific, hand-categorized items.

International Relative Prices

- We find faster mean reversion in RERs than observed with CPI-based RERs

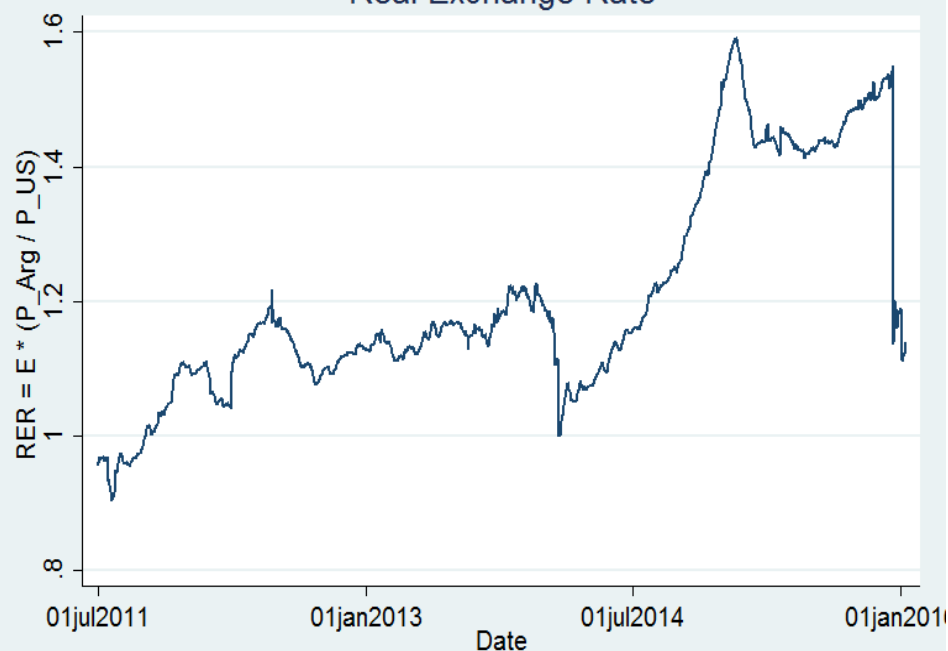
Australia vs US

Real Exchange Rate



Argentina vs US

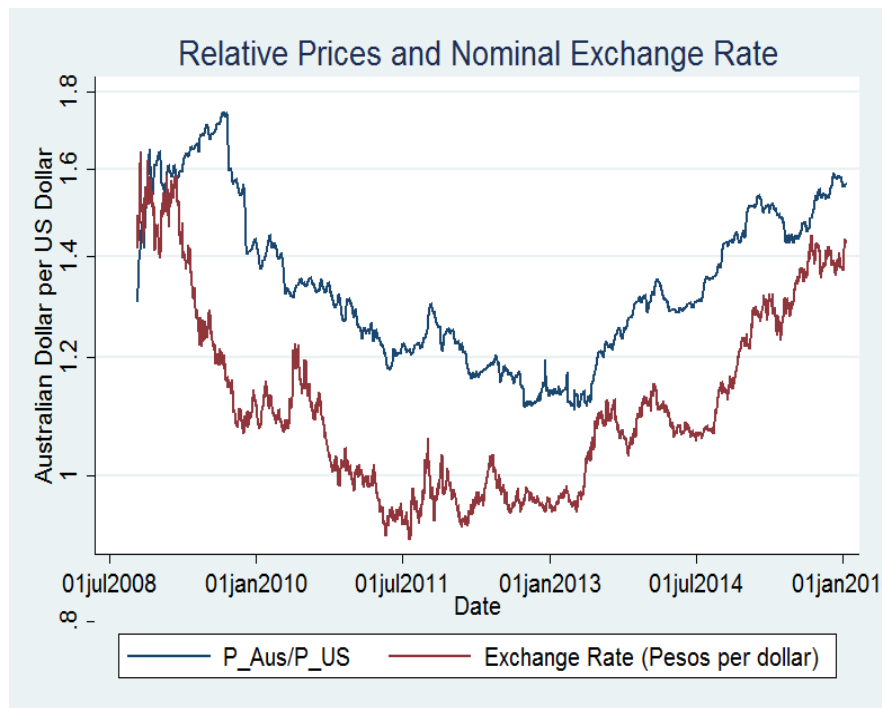
Real Exchange Rate



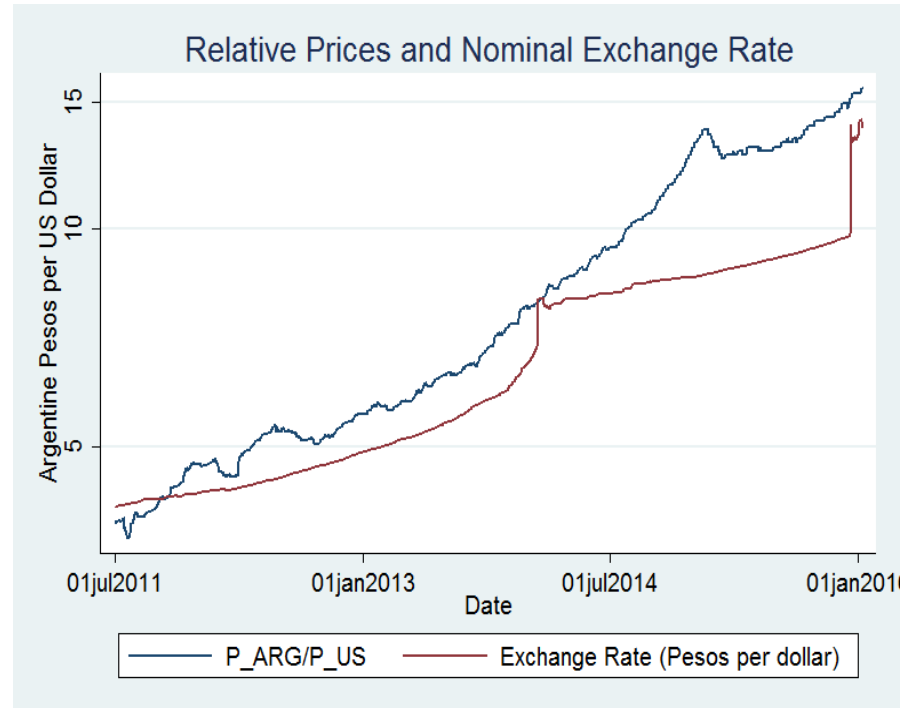
International Relative Prices

- Driven by **faster co-movement of relative prices and nominal exchange rates**

Australia vs US



Argentina vs US



Conclusions

- Online data can be a reliable source of information for inflation measurement
 - Cheaper, faster, provides anticipation
 - *Big Data* characteristics can simplify measurement
- Online prices increase the frequency and quality of micro price data available for research
 - Re-evaluate old empirical puzzles or address questions that could not be answered before
- *Big Data* → opportunity to get involved in the “grubby task” of data collection, creating datasets that fit our specific research needs.



Extra Slides

Each Data Source has Advantages and Disadvantages

CPI Data

- Purpose: measure inflation

Advantages	Disadvantages
<ul style="list-style-type: none">• Representative sample<ul style="list-style-type: none">• carefully-chosen goods• many retailers and locations• Long Time Series• Collection of 'posted prices in stores	<ul style="list-style-type: none">• Very costly to collect and access• Low frequency (monthly)• Limited number of goods and varieties• Some unit values and imputed prices• Difficult international comparisons

Each Data Source has Advantages and Disadvantages

Scanner Data

- Purpose: marketing analytics (eg. brand market shares)

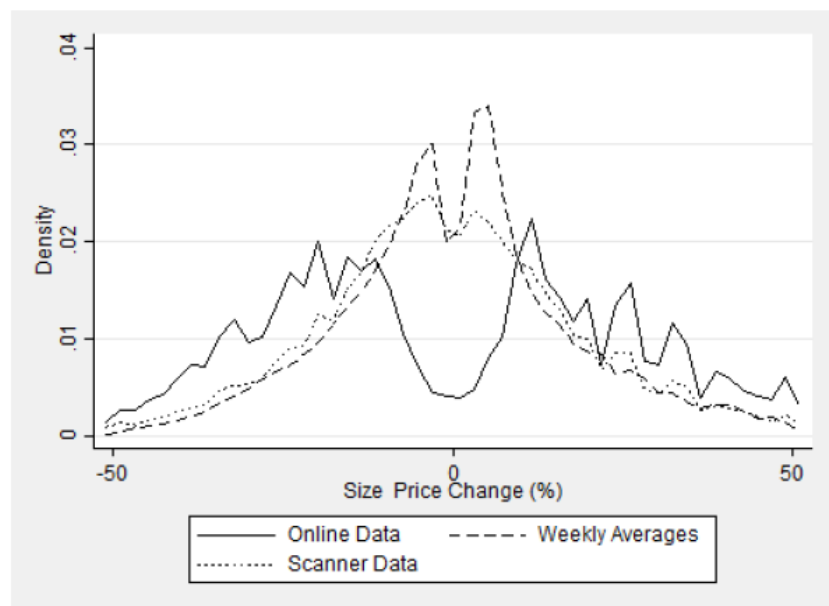
Advantages	Disadvantages
<ul style="list-style-type: none">• Transaction data<ul style="list-style-type: none">• Contains quantities and sometimes costs• Granularity<ul style="list-style-type: none">• Some product details for all goods <i>sold</i>• Frequency (weekly)	<ul style="list-style-type: none">• Limited coverage (supermarkets)• High cost to collect/acquire• Data characteristics vary greatly depending on provider, location, time period, etc.• Hard to compare internationally• Unit values and time-averages (eg: prices are often calculated as sales/quantity in a week)

Price Stickiness

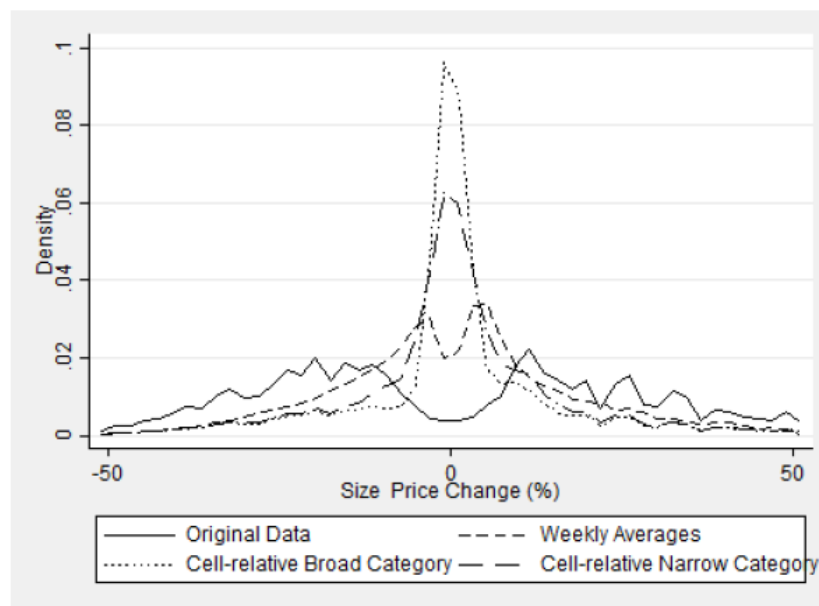
- I use online data to show that many of the previous findings in the literature were driven by measurement biases in CPI and scanner data:
 - Weekly average prices in scanner data
 - Cell-relative imputation for temporary missing prices in CPI data
- Effects:
 - Reduce the duration (stickiness) of price changes
 - Create spurious small changes → altering the shape of distribution
 - Cause downward sloping hazard functions

Distribution of the Size of Price Changes

- The distribution is bimodal, with little mass near 0% → more consistent with adjustment/menu cost models



(a) Online vs Scanner



(b) Online vs CPI-simulation

Figure 9: The Distribution of the Size of Price Changes in the US

Advantages of Online Data to Measure Stickiness

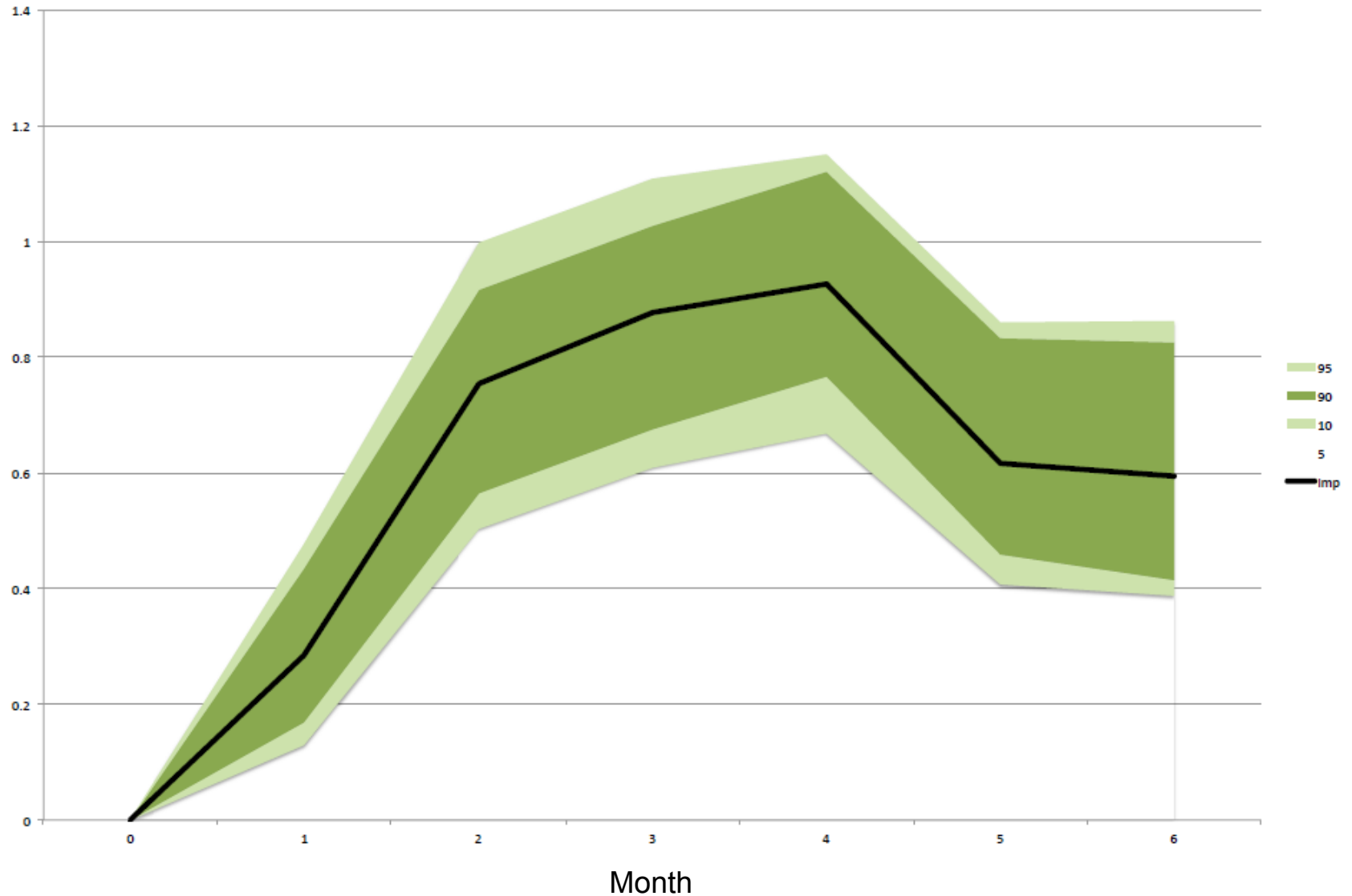
- Free from “measurement bias”
 - High frequency
 - Posted prices
 - All products
 - Uncensored price spells
- Available in multiple countries → no differences in methods, time periods or type of data.
- Real-time data, with no delays → policy applications

Anticipation in Online Prices

- Online prices tend to anticipate changes in CPI inflation trends.
 - More than just quicker access to data
 - Online prices tend to *react faster* to shocks.
 - Why?
 - Lower adjustment (or menu) costs
 - Online shoppers may be less sensitive to price changes
 - More intense and transparent competition
- We can study the link between online data and CPIs using simple VARs.
 - VAR regressions with Δ CPI on the LHS and lags of Δ CPI and Δ OPI on the RHS (monthly data)
 - Impulse responses show the impact of a 1% shock in online series (OPI) on future CPI (reflecting additional information not contained in lagged CPI)

Impulse Response USA

Cummulative IRF - 1% Shock to Online Aggregate Inflation

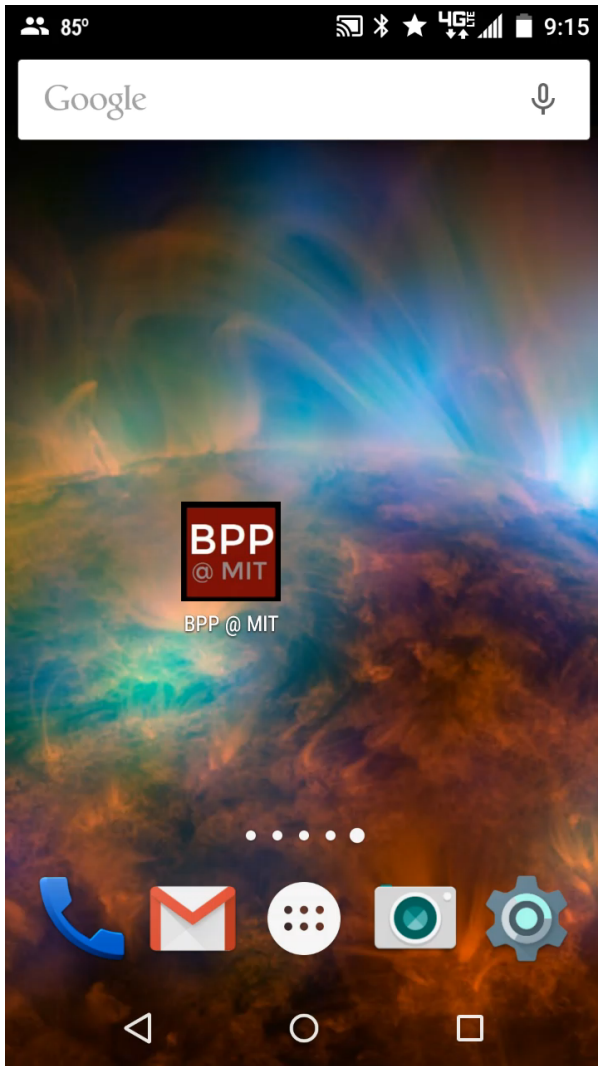


Source: PriceStats – Data until March 2014

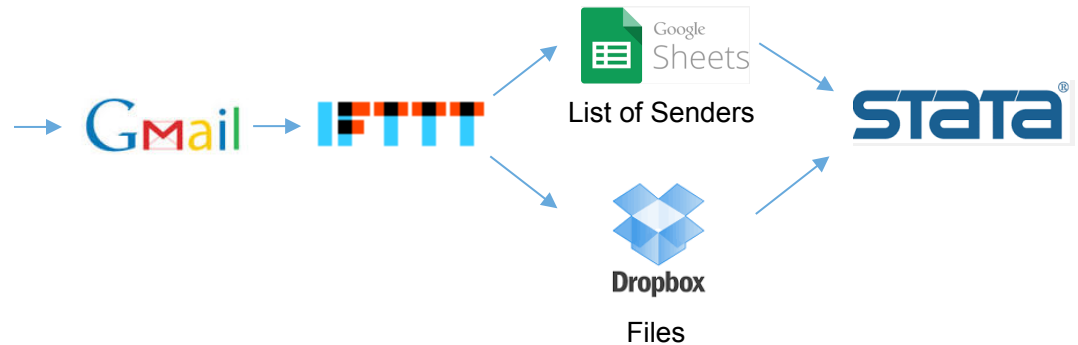
Online vs Offline Prices

- Are online prices “representative”?
 - “No”
 - Online sales are still only about 10% of retail sales in developed countries
 - “Yes” ...
 - The `online store` is effectively the *largest* store for 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
 - Even if online transactions are rare, the online price can be a good proxy for the offline price\
 - Close matching in price indices
 - Cavallo (2015) : simultaneous data collection for 50 large retailers in 10 countries shows they tend to have either identical online and offline price levels.



The BPP App



- Every day we process and consolidate the offline data



	A	B	C	D	E	F	H	I	J
1	DEVICE ID	DATE	TIME	BARCODE	STORE NAME	STORE LOCATION	PRICE	PHOTO	COMMENTS
2	45c653f06cc750a8	10/20/2014	12:11	98071000050369	home depot	andrew	19.98	20141020_121214.jpg	
3	45c653f06cc750a8	10/20/2014	12:12	9807390862	home depot	andrew	7.88	20141020_121310.jpg	
4	45c653f06cc750a8	10/20/2014	12:13	9807203925	home depot	andrew	179	20141020_121357.jpg	
5	45c653f06cc750a8	10/20/2014	12:15	01178841	home depot	andrew	34.97	20141020_121558.jpg	

- We then use the barcode id to check prices online

Are Online and Offline Prices Similar?

PRELIMINARY

Table 3: Country - Level Differences

Country	(1) Ret.	(2) Obs	(3) Identical (%)	(4) High On (%)	(5) Low On (%)	(6) Markup (%)	(7) Difference (%)
Argentina	5	3935	52	34	13	5	3
Australia	4	4093	72	21	7	5	1
Brazil	5	2241	37	21	41	-7	-4
Canada	5	4255	88	6	6	2	0
China	1	490	83	15	2	13	2
Germany	3	1243	81	4	16	-5	-1
Japan	4	1760	45	9	46	-14	-8
South Africa	5	2762	84	8	8	1	0
UK	4	2413	87	4	9	-3	0
USA	16	11144	67	10	24	-7	-2
ALL	52	34336	69	13	18	-3	-1

Note: Results updated 27 Jan 2016. Column 3 shows the percentage of observations that have identical online and offline prices. Column 4 has the percent of observation where prices are higher online and column 5 the percentage of price that are lower online. Column 6, is the online markup, defined as the average price difference excluding cases that are identical. Column 7 is the average price difference including identical prices.