The Billion Prices Project:
Building Economic Indicators From Online Data

Alberto Cavallo

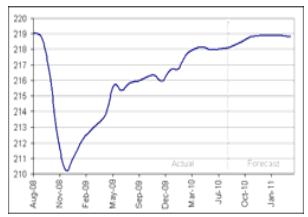
MIT Sloan

UNECE CPI Meeting, Geneva, May 31st 2012

Our approach: use online data to build real-time economic indicators around the world

1. ***Use scraping technology***
2. ***Connect to thousands of online retailers every day***
3. ***Find individual items***
4. ***Store key item information in a database***
5. ***Calculate real-time Indexes***

- *Date*
- *Item*
- *Price*
- *Description*



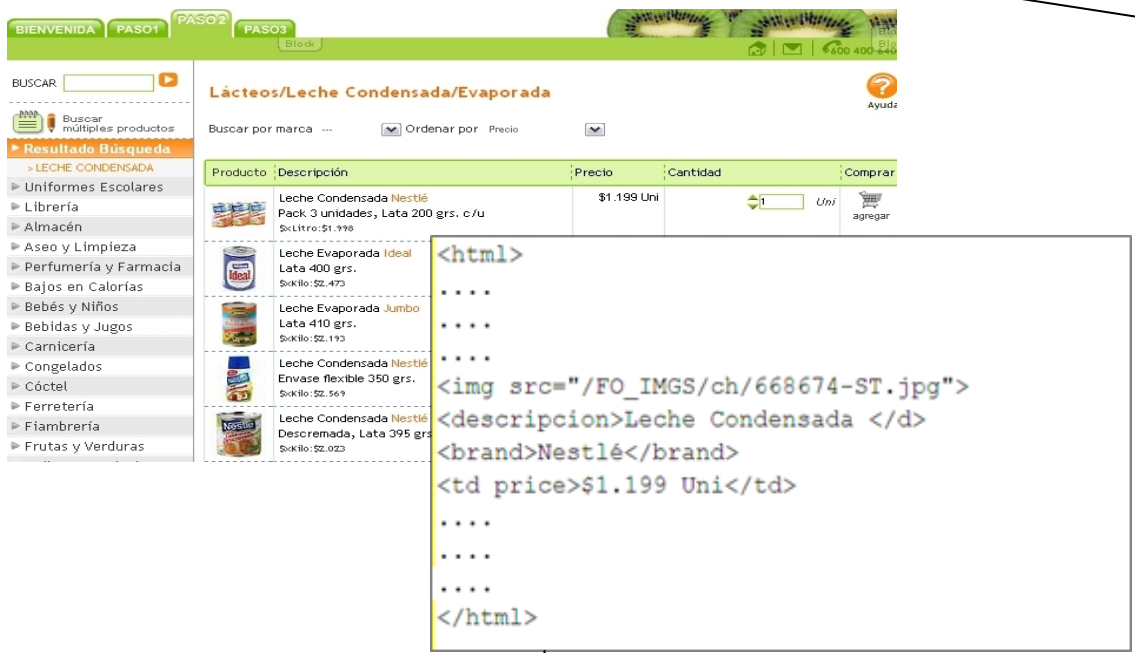
How do we collect data?

Key Scraping Guidelines

- Our prices are collected from public online sources, using a technique called “web scraping”
- A software downloads the webpage, analyses the html code, identifies price data, and stores it in a database

Scraping Example

ILLUSTRATIVE



	_ID	_ID2	_PRODUCTO	_MARCA	_TAMANO	_BULKPRICE	_PRICE
1	3429	266235-ST	Leche Condensada	Leche Sur	Lata 395 grs.	xKilo:$1.744	689
2	3422	266231-ST	Leche Condensada	Nestlé	Descremada, Lata 395 grs.	xKilo:$2.023	799
3	995	619436-ST	Leche Condensada	Nestlé	Envase flexible 350 grs.	xKilo:$2.569	899
4	3804	399781-ST	Leche Condensada	Nestlé	Lata 397 grs.	xKilo:$1.761	699
5	11676	668674-ST	Leche Condensada	Nestlé	Pack 3 unidades, Lata 200 grs. c/u	xLitro:$1.998	1.199

Online prices as a new source of data

Advantages:
- Low cost
- High-frequency (daily)
- Data can be collected remotely in ~70 countries
- Data is available in real-time, with no delays
- Product details: brand, package size, sale indicator, price control, etc.
- Information on *all* products sold in each retailer (census within retailer)
- Elementary indices can be built with hundreds of homogeneous products
- Product sampled automatically from the moment they are introduced until they disappear from the store

Disadvantages:
- Relatively few retailers covered
- Only ~60% CPI categories online (most services are not yet available online)
- No quantities (unlike in scanner data)

Our method involves three stages

Technology & Processes

- Multiple specialized applications for each step of the scraping process
- Customized scraping for each retailer
- Needs to be updated every time the “look and feel” of a website changes

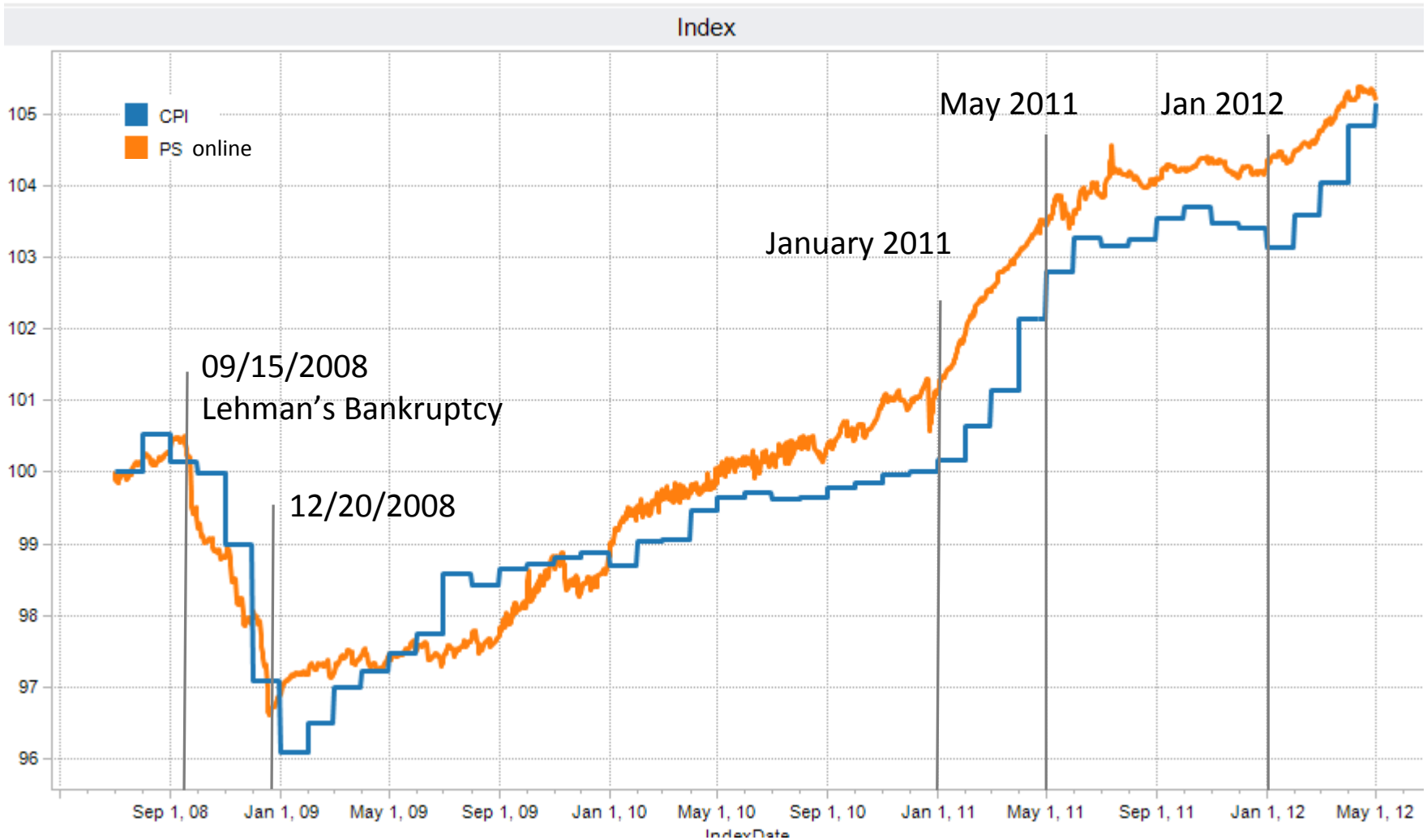
- Data is automatically cleaned and homogenized
- Need to classify and select items and retailers to be used in the indices
- Monitoring controls to check scraping, cleaning, statistics

- Statistics and indices are computed on a daily basis
- We use standard CPI techniques to compute elementary indices, and aggregate them using official CPI weights
- For aggregate indices, compensate for things that we cannot monitor online (eg services)

Aggregate Price Indices

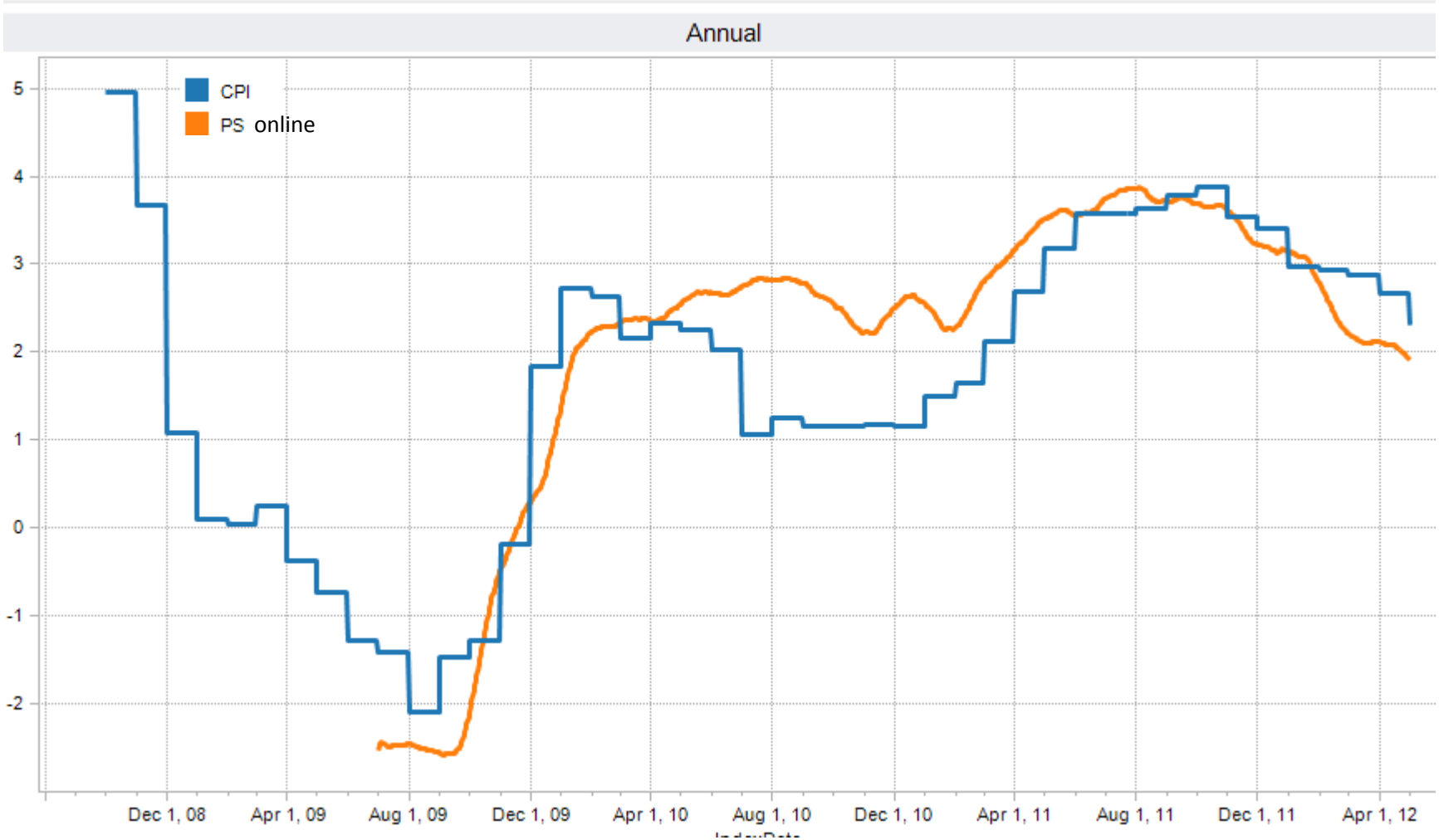
- Currently available for 18 countries: USA, Argentina, Australia, Chile, China, Colombia, France, Germany, Ireland, Italy, Japan, Netherlands, Russia, South Africa, Spain, UK, Uruguay, and Venezuela
- Daily indices, published with a 3-day lag
- US and Argentina indices are publicly available online
- Main use: detect changes in CPI trends

US Daily Price Index



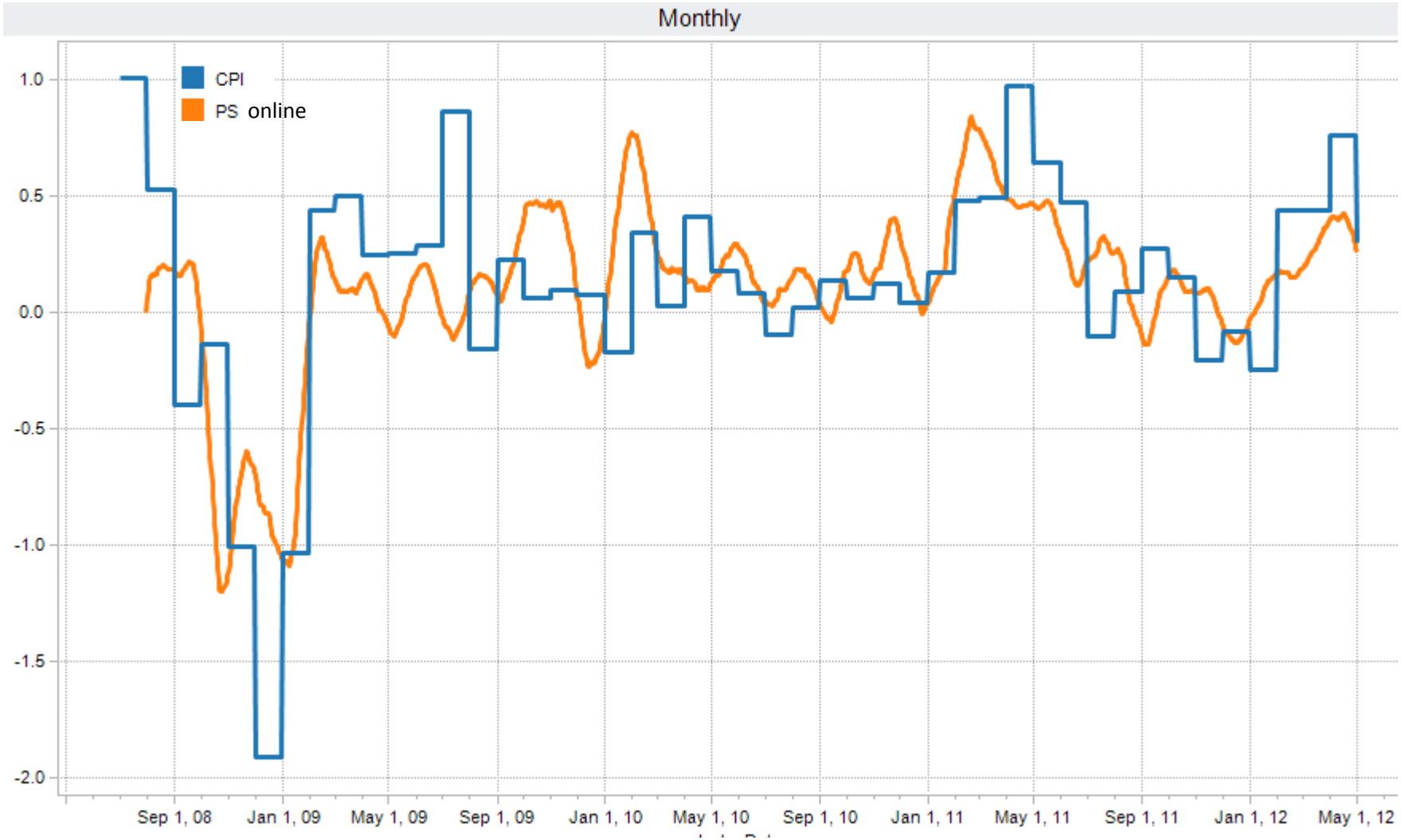
Source: BPP – PriceStats – BLS (CPI-U, US city-average, all items, NSA)

US Annual Inflation



Source: BPP – PriceStats – BLS (CPI-U, US city-average, all items, NSA)

US Monthly Inflation



Source: BPP – PriceStats – BLS (CPI-U, US city-average, all items, NSA)

Annual Inflation Rates

Argentina

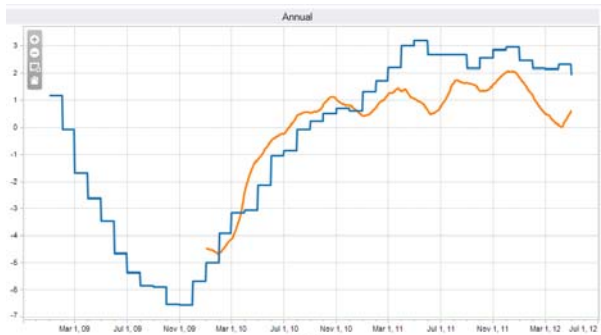
Australia

China

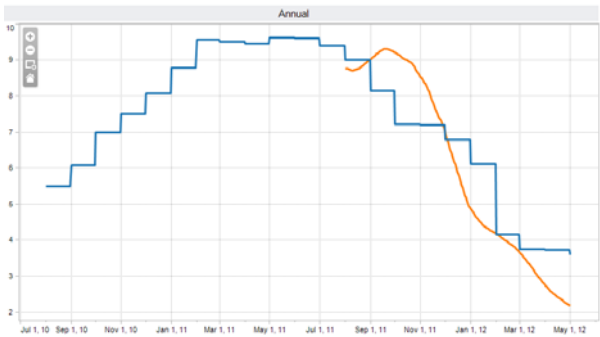
Colombia

Germany

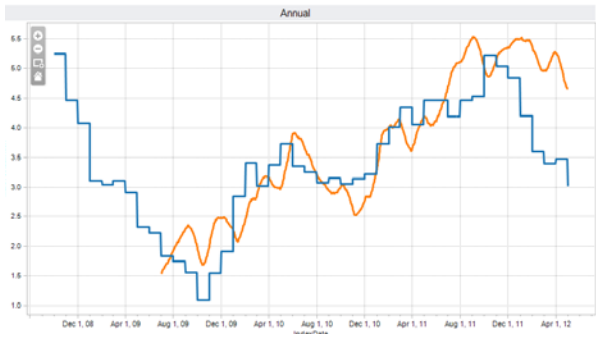
Ireland



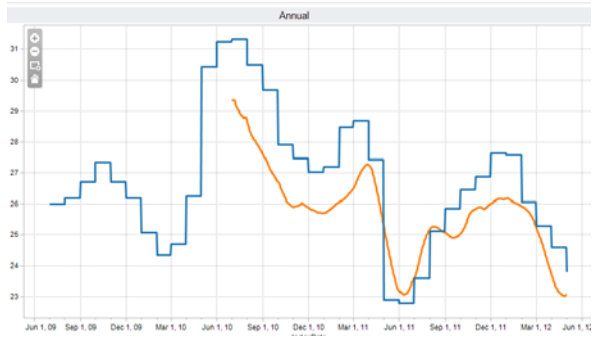
Russia



UK



Venezuela



Source: PriceStats - StateStreet

Things we have learned

- Most retailers are ok with data scraping for statistical purposes as long as you:
  1) Follow their rules and avoid putting strain on their servers
  2) Explain to them that you are trying to compute aggregate statistics, not to disclose their pricing practices to competitors (which some companies are trying to do)
- Online prices behave like offline prices even if the online market is small
  - In some countries (eg US) online and offline markets are closely integrated → similar price levels & pricing strategies
  - Often online prices have a markup over offline data, but it tends to be constant over time→ implies similar inflation rates
  - In countries where online markets are not well developed, online and offline prices can be identical because firms do not have a differentiated pricing strategy online (they simply show their offline database in their website)
- Data needs to be "pulled" from the retailers, not "pushed" by them (i.e. collected, not self-reported)
  - Ensures reliability
  - Prevents self-reporting biases

Things we have learned

- Online prices can react faster to shocks, providing anticipation in inflation trends
  - Lower menu costs or consumer anger
- Online data makes it easier to deal with product introductions and overlapping quality adjustments
  - Every model or version of a product is automatically included in the sample as soon as they are available to consumers
- Online data work best as an alternative source of data, not as a separate sector or location that needs special treatment
- There are lots of issues that still need to be addressed:
  - Deal with sectors whose prices are not yet online
  - Determine mechanisms to select retailers sampled
  - Explore potential complementarities with scanner data to update weights

Other Indicators

GLOBAL PULSE

Harnessing innovation to protect the vulnerable

HOME

BLOG

RESEARCH

TECHNOLOGY

LABS

MULTIMEDIA

ABOUT

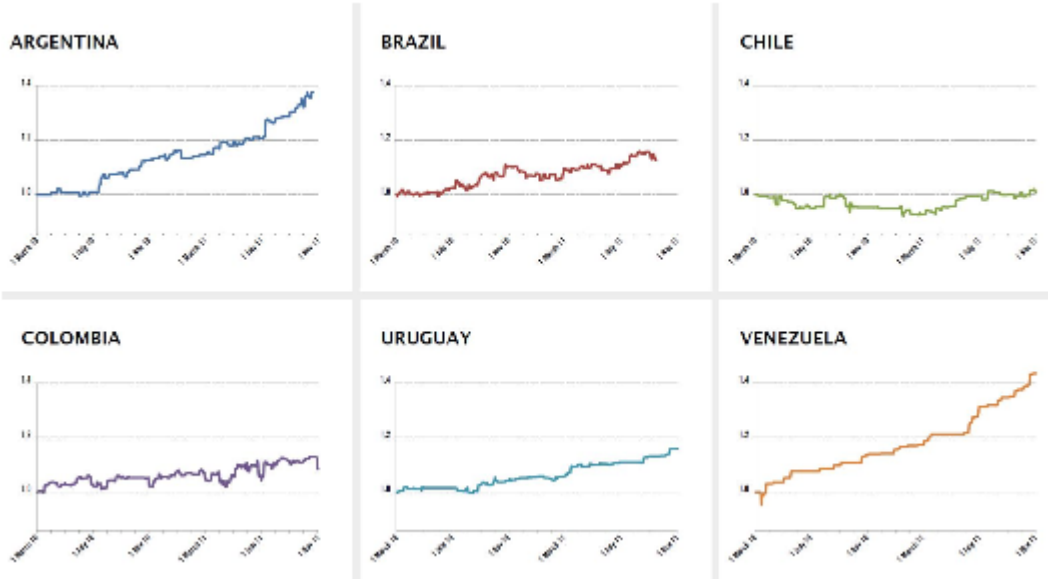
CONTACT

RESEARCH

DAILY TRACKING OF COMMODITY PRICES: THE E-BREAD INDEX

Description:

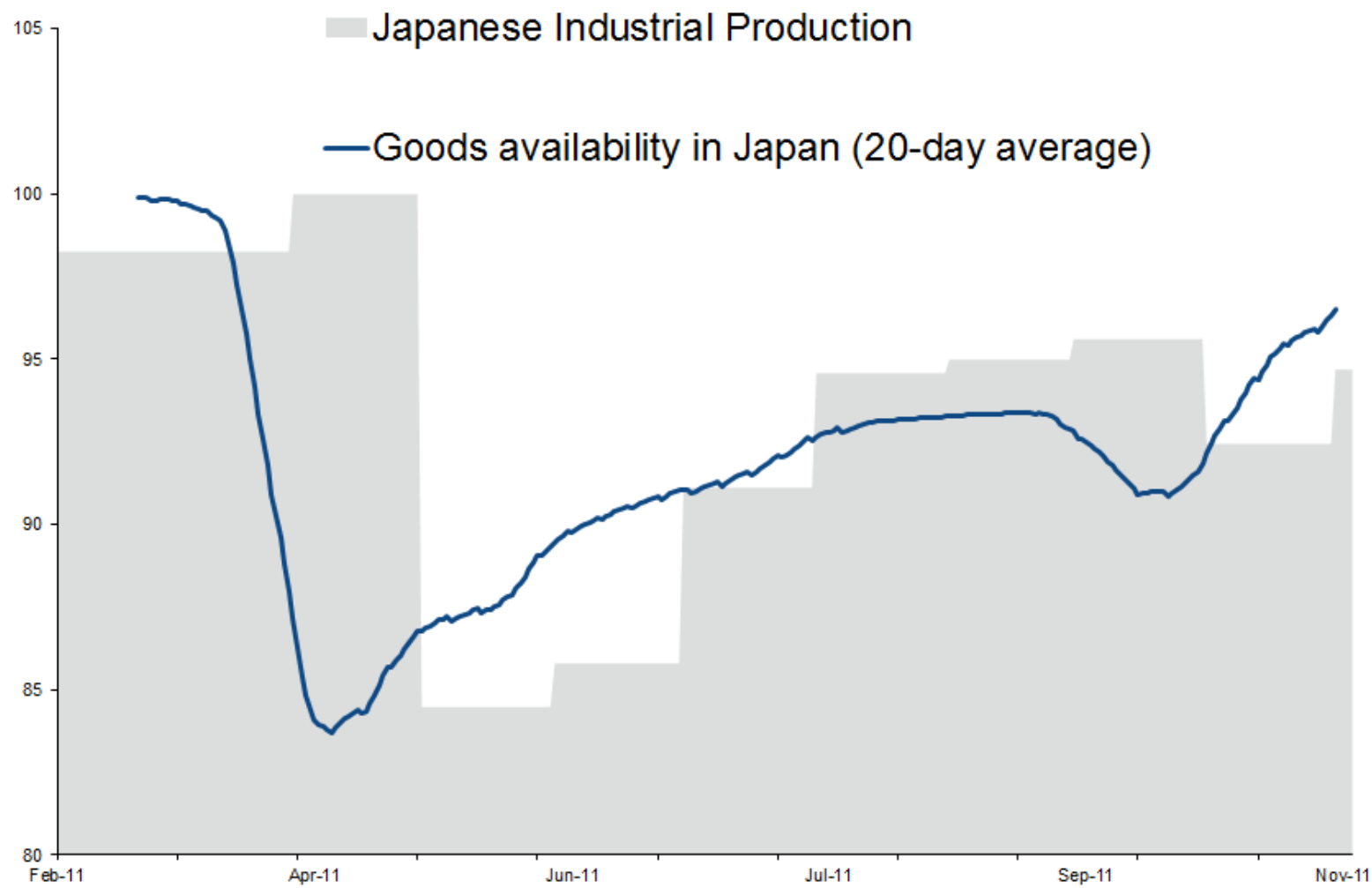
This project investigates and shows how scraping online prices could provide real-time insights on price dynamics, focusing on the case of bread in 6 Latin American countries.



Partners: Billion Prices Project at MIT

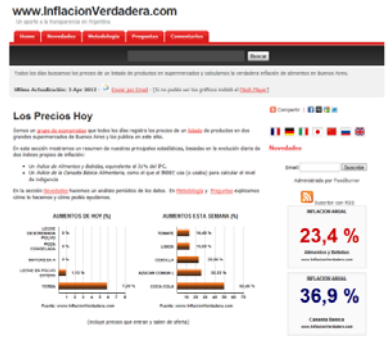
Economic Activity and Natural Disasters

The Impact of Japan's 2011 Earthquake on Product Availability



Related Initiatives

- Argentina's "Real Inflation"
  - www.inflacionverdadera.com
  - Started in 2007 to provide an alternative inflation estimates when official data became unreliable



- Billion Prices Project at MIT
  - bpp.mit.edu
  - Started in 2008
  - Conduct academic research on inflation and pricing behaviors
  - Joint work with Prof. Roberto Rigobon (MIT Sloan)

- PriceStats LLC
  - www.pricestats.com
  - Founded in 2011 to collect data and produce daily price indexes
  - Partners with State Street Bank to distribute indices to the financial sector
  - PriceStats' Argentina index is published in The Economist every week

Conclusions

- Scraped online data can provide a reliable complementary source of information for consumer price indices
  - Immediate application in sectors that are available online: e.g. electronics, apparel, household products, and supermarkets
- This is a long-term research agenda

Massachusetts
Institute of
Technology