The Billion Prices Project
Research and Inflation Measurement Applications

Alberto Cavallo
MIT

European Central Bank
April 2014

Micro Price Data in Macroeconomics

- Data Sources
  - Statistical Offices (CPI, PPI, IPI)
  - Scanner Data (eg. Nielsen)
  - Online Data (eg. Billion Prices Project)
- Uses
  - Research in Macroeconomics
    - Price Dynamics (Price Stickiness, Real Rigidities)
    - Market Segmentation
  - Research in International Economics
    - Pass-through and Border Effects
    - Law of One Price and PPP
    - Real Exchange Rates
  - Measuring inflation and other economic indicators

Each Data Source has Advantages and Disadvantages

CPI Data

- Purpose: measure inflation

Advantages	Disadvantages

• Representative sample
  • carefully-chosen goods
  • many retailers and locations
• Long Time Series
• Collection of ‘posted prices in stores
	• 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. market shares)

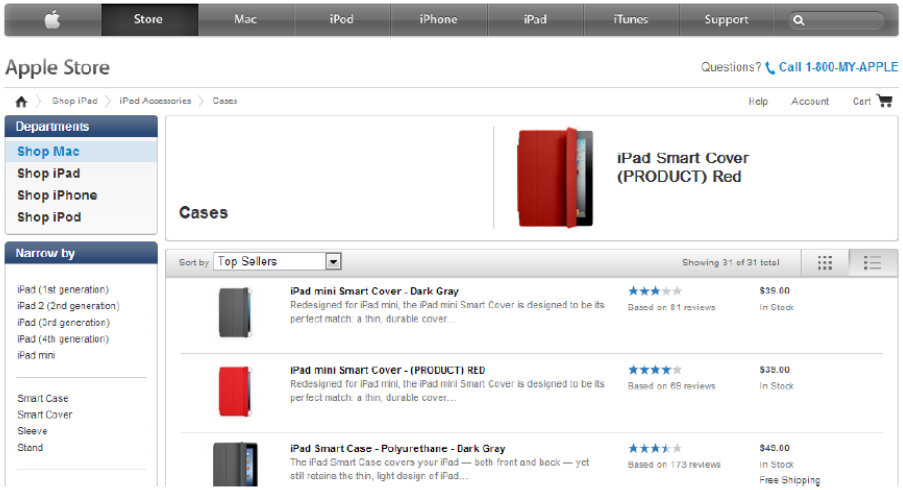
Advantages	Disadvantages

• Granularity
  • Some product details for all goods *sold*
• Transaction data
  • Contains quantities and sometimes costs
• Frequency (weekly)
	• High cost to collect/acquire
	• Limited coverage (supermarkets, department stores)
	• Data characteristics vary greatly depending on provider, location, time period, etc.
	• Extremely difficult to compare internationally
	• Unit values and time-averages (eg: prices are often calculated as sales/quantity in a week)

Each Data Source has Advantages and Disadvantages

Online Data

- Can be collected using automated web-scraping software
- Every day, a *robot* downloads a public webpage, analyses its HTML code, extract price data, and stores it in a database




```
<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 -->
.....
```

Each Data Source has Advantages and Disadvantages

Online Data

Advantages	Disadvantages

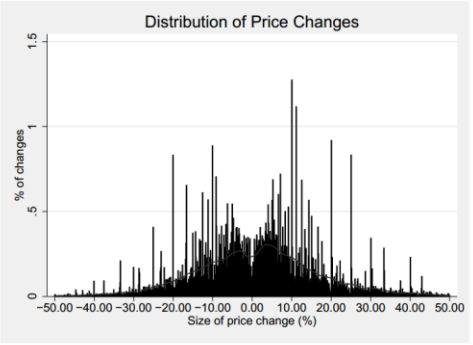
• Frequency (daily)
• Cheap to collect (but complicated)
• Granularity
  • All product details (brands, size, anything shown online)
  • All goods and varieties available for sale (census)
  • New goods automatically sampled
• Easier to compare internationally
	• Fewer retailer and locations than CPI
	• Short time series
	• Not all categories of goods and services are online (not yet)
	• Online and Offline prices may behave differently

Online Data in Macro and International Research

- Billion Prices Project at MIT
  - Daily data from 2008 to the present
  - Sample of retailers and countries has grown over time (hundreds of retailers in 70 countries)
- We have written papers on:
  - Price Stickiness
  - Border Effects
  - Law of One Price
  - Inflation Measurement

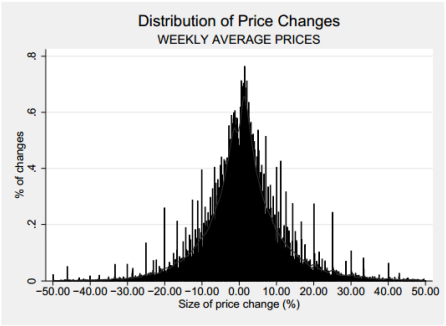
Research Examples: Price Stickiness

- We use online data to re-evaluate puzzling styled facts coming from CPI or scanner data
- For example, in contrast to the literature, we find that in most retailers and countries price-change distributions are bi-modal, with little mass close to zero percent.
  - Consistent with state-dependent models of price adjustment where small price changes are not optimal (given the existence of an adjustment or menu'cost).



(a) Chile

With unit values or weekly averages



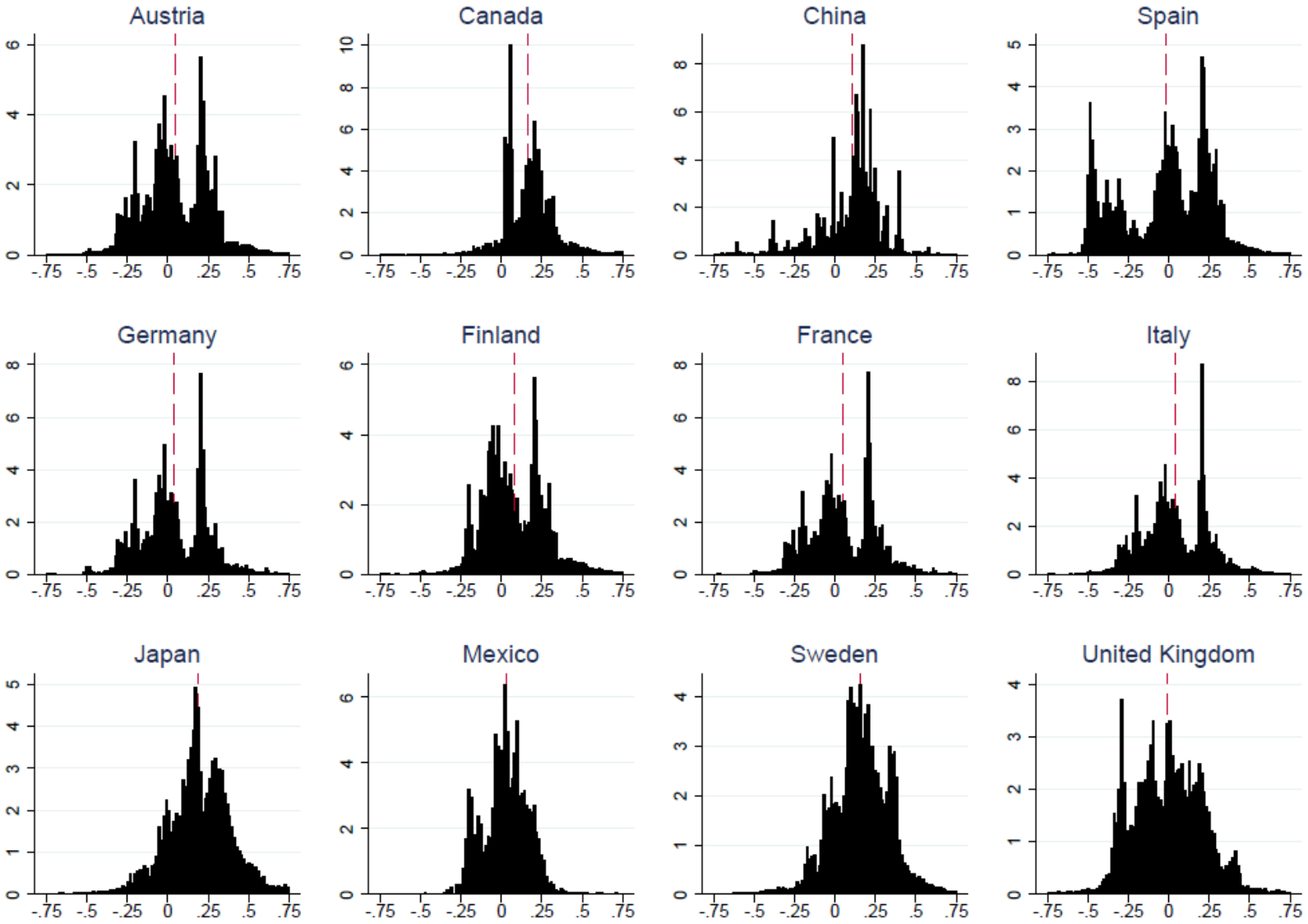
(c) Chile

See Cavallo (2012) Scraped Data and Sticky Prices", Cavallo & Rigobon (2012) "The Distribution of the Size of Price Changes"

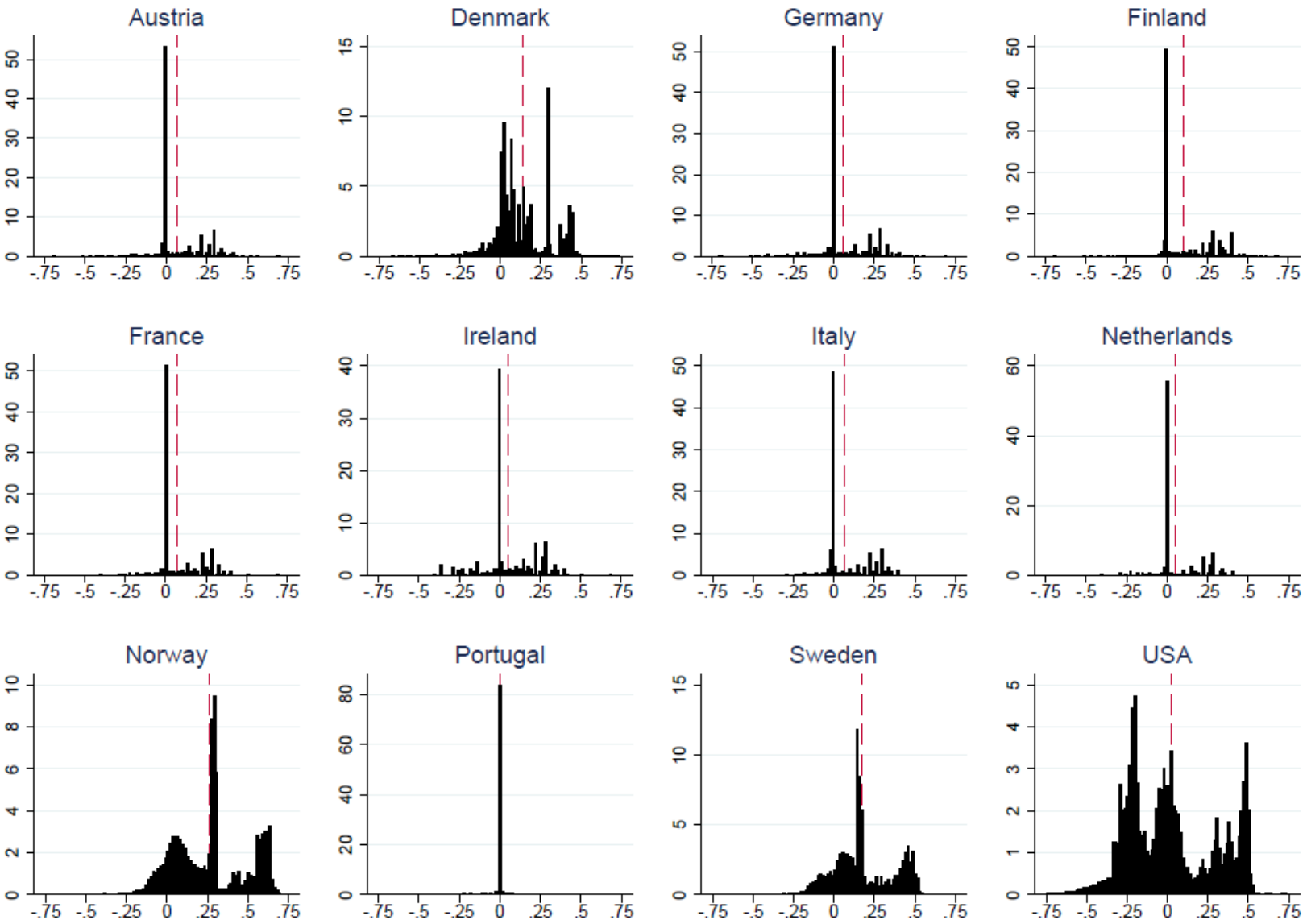
Research Examples : Law of One Price (LOP)

- Cavallo, Neiman, and Rigobon (2014) QJE.“Currency Unions, Product Introductions, and the Real Exchange Rate”
- We evaluate LOP deviations using a large dataset of identical tradeable goods, sold by global retailers in three industries and dozens of countries (Apple, Ikea, Zara, H&M, and others)

Good-level RERs q_{ij} for j = United States



Good-level RERs q_{ij} for j = Spain



Research Examples : Law of One Price (LOP)

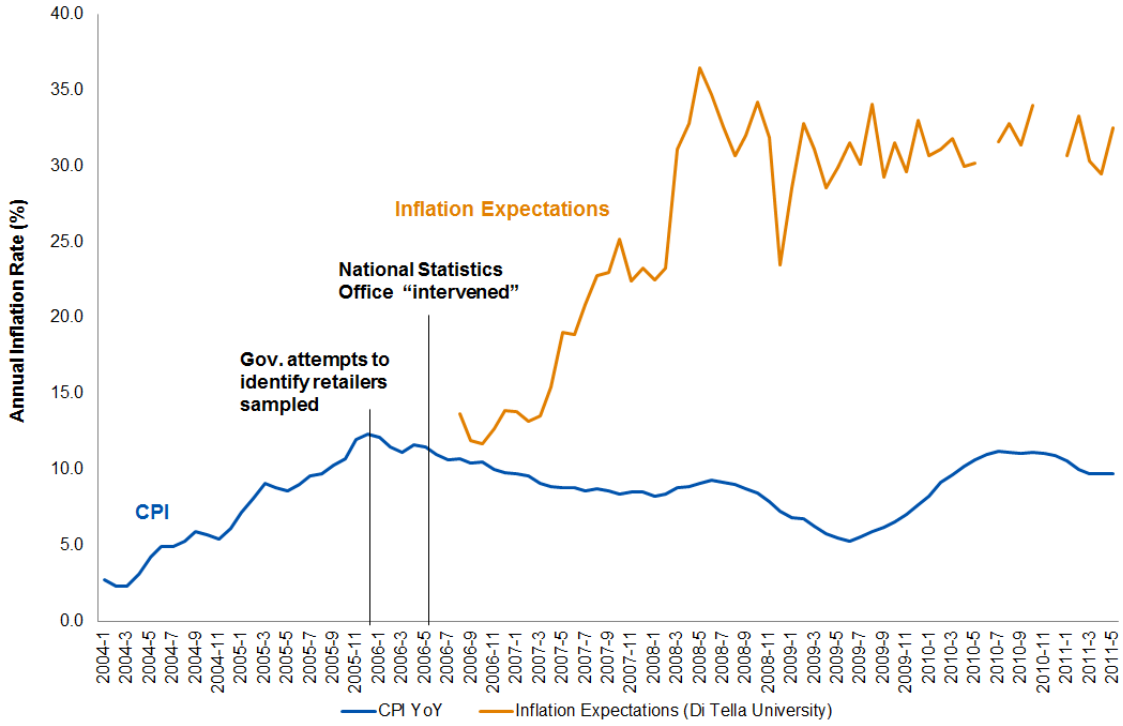
Average Absolute Value of Good-Level Log RER

		All Stores	Apple	IKEA	H&M	Zara
All Data	Currency Unions	0.062	0.005	0.117	0.021	0.087
All Data	NER Pegs	0.149	0.047	0.164	0.141	0.142
All Data	Floats	0.182	0.139	0.185	0.152	0.192

- Two main findings:
  - LOP holds within Currency Unions, fails otherwise (even in pegged regimes)
  - A new decomposition shows that the RER at the time of product introduction is most important (yet not reflected in traditional CPI-based RERs), and moves closely with the Nominal Exchange Rate.

Research Examples: Inflation Measurement

- Argentina's inflation data is widely questioned
  - Statistical Office "intervened" in 2007.
  - Since then CPI inflation has been stable around 10%
  - Inflation expectations have been consistently above 25%



Research Examples: Inflation Measurement

- Using scraped data, I showed that online price indices could closely match CPIs in four other Latin American countries

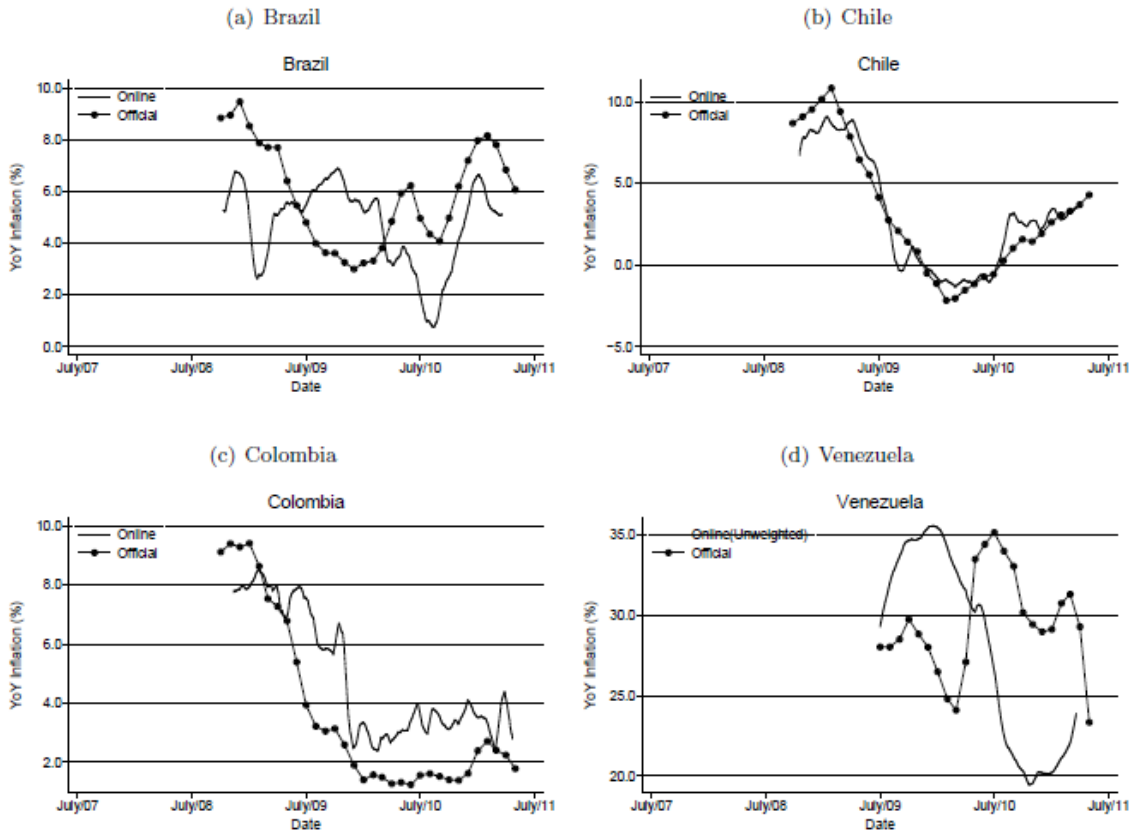


Figure 2: Online and Official Indexes - Annual Inflation Rate

Brazil, Chile, Colombia, and Venezuela

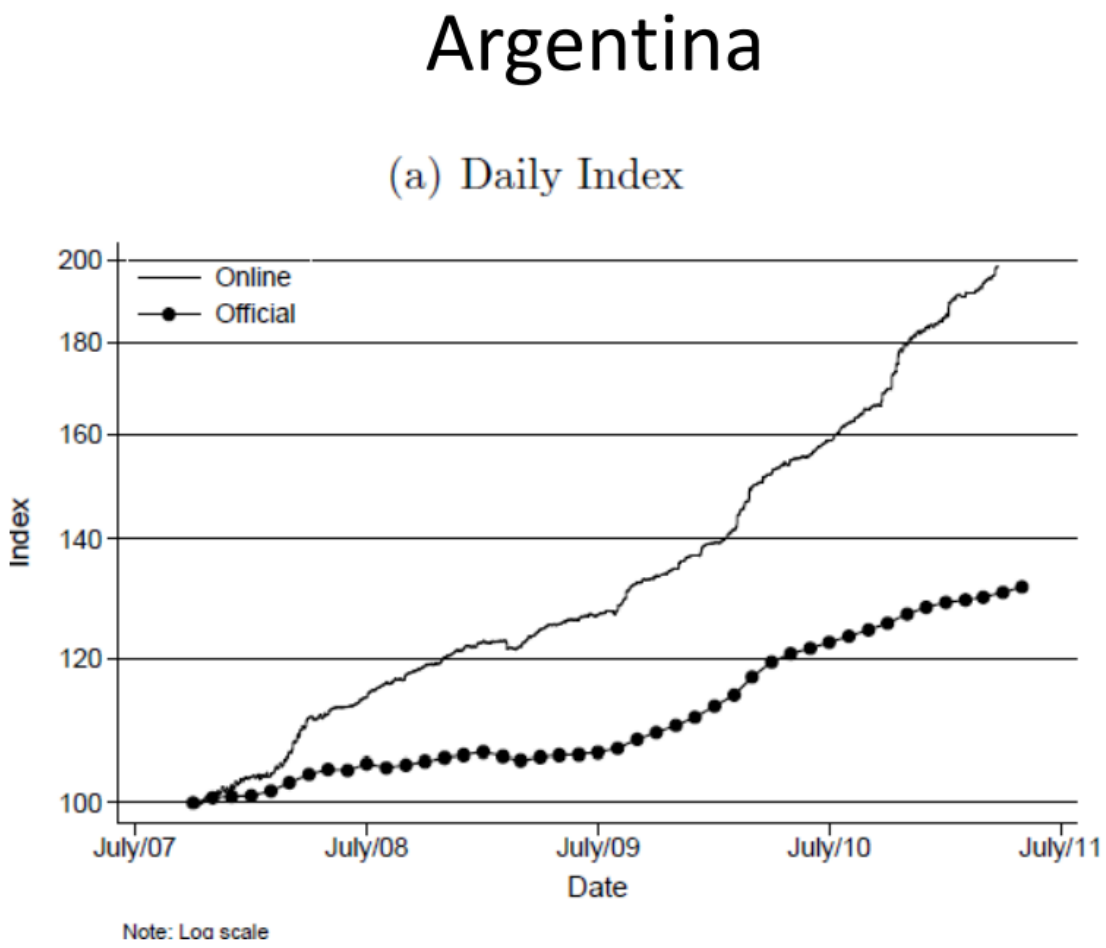
Table 3: Online vs Official Series

	Argentina	Brazil	Chile	Colombia	Venezuela
Mean Annual Inflation (%)					
Online Index		4.72	3	4.88	27.43
Official Index		5.91	3.19	3.73	29.38

- Matching is best on Chile and Colombia, where:
  - Supermarkets have larger market shares (27% and 30%, vs. only 15% in Brazil)
  - City where online data is collected accounts for most of the CPI (55% in Santiago)
- Good news for Argentina! → the supermarket I used had 28% market share, Buenos Aires is 100% of CPI data

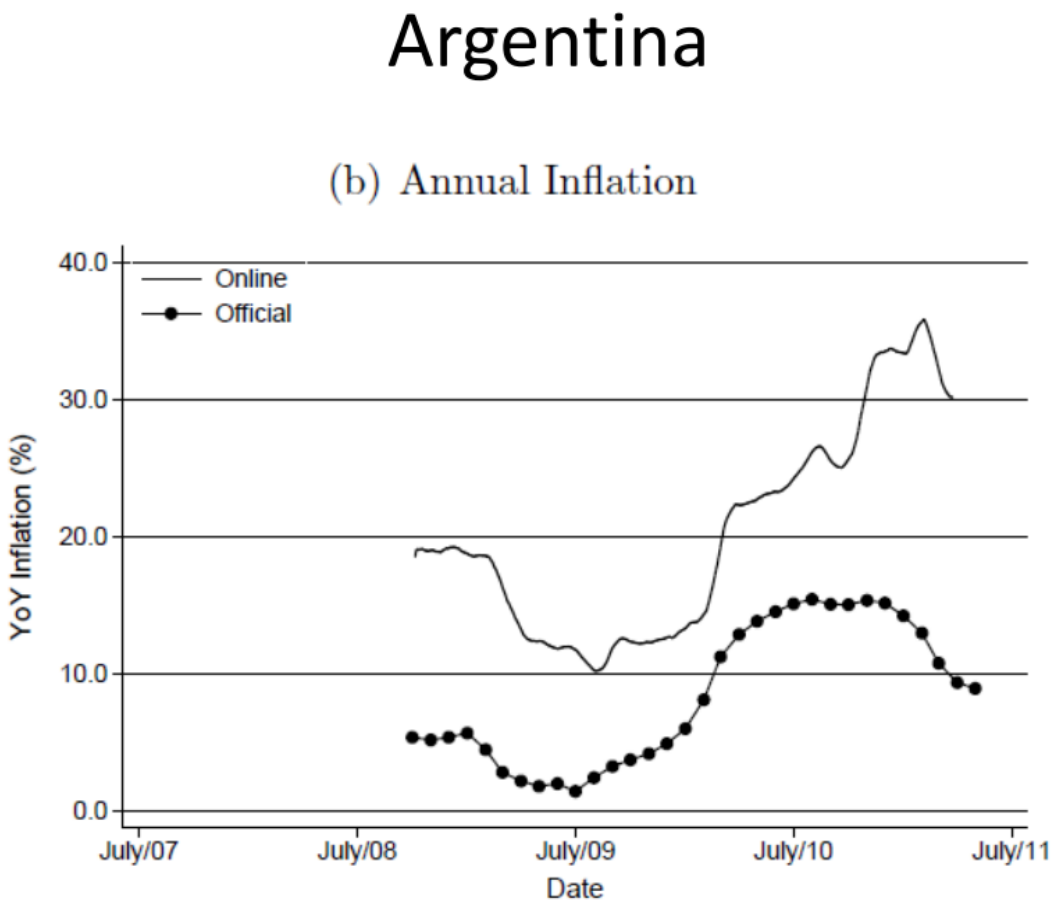
Research Examples: Inflation Measurement

- Argentina`s inflation rate



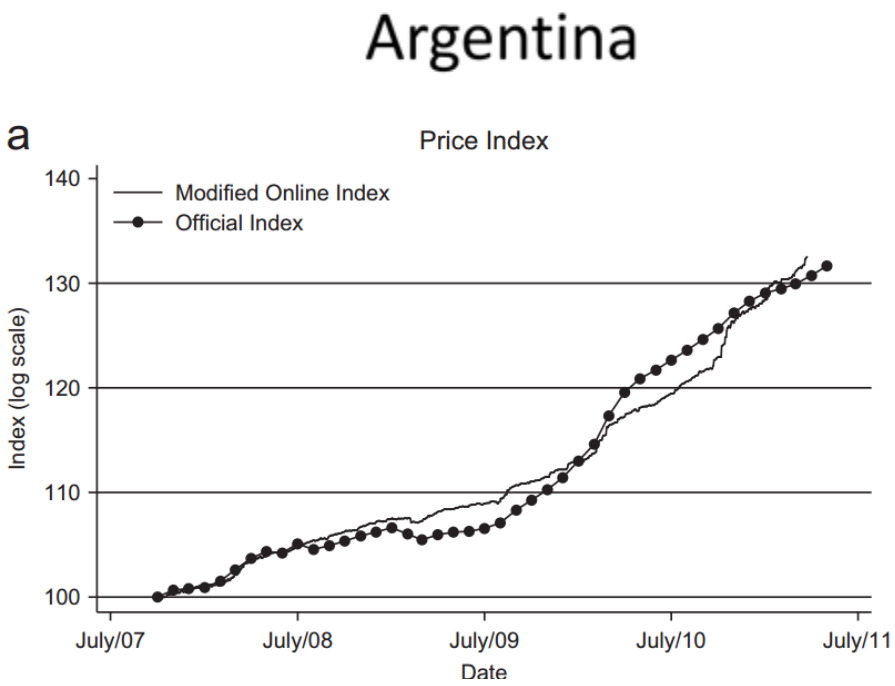
Research Examples: Inflation Measurement

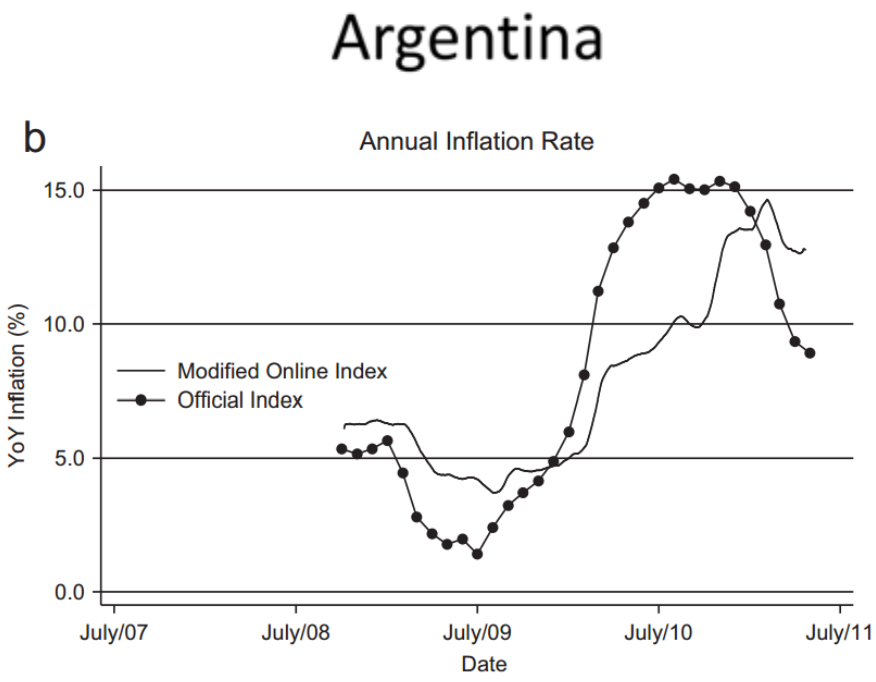
- Argentina`s inflation rate



Research Examples: Inflation Measurement

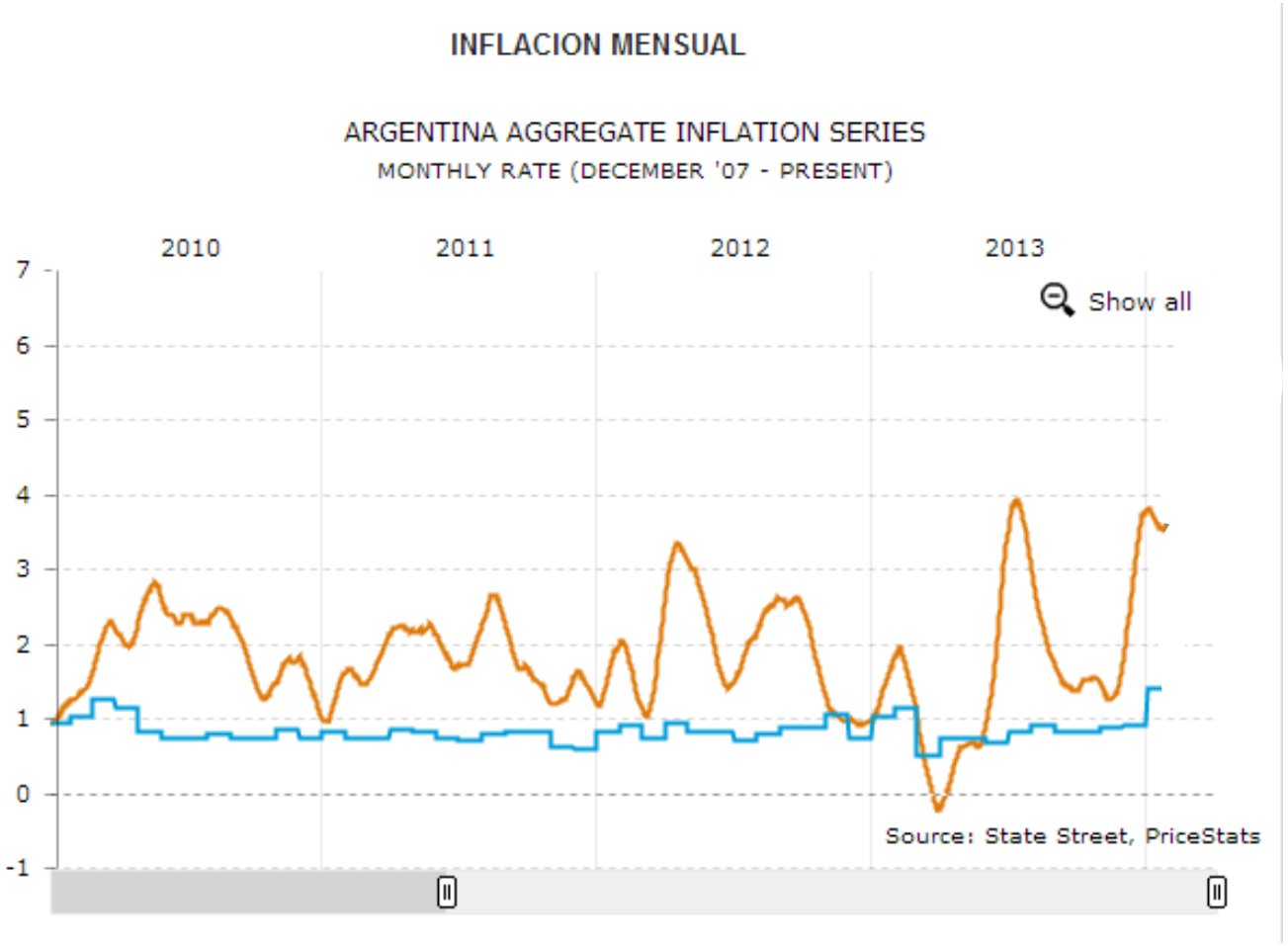
- How to best approximate the official inflation rate?
- Answer: Take the true inflation rate (online), and divide by 3.





Research Examples: Inflation Measurement

- Update: New Price Index launched in Jan 2014



Online Data and Daily Inflation Measurement

- In 2008 we started publishing a daily online index for Argentina
- In 2010, we started publishing a daily 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).

1 ***Use scraping technology***

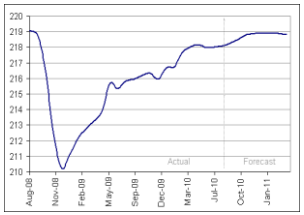
2 ***Connect to thousands of online retailers every day***

3 ***Find individual items***

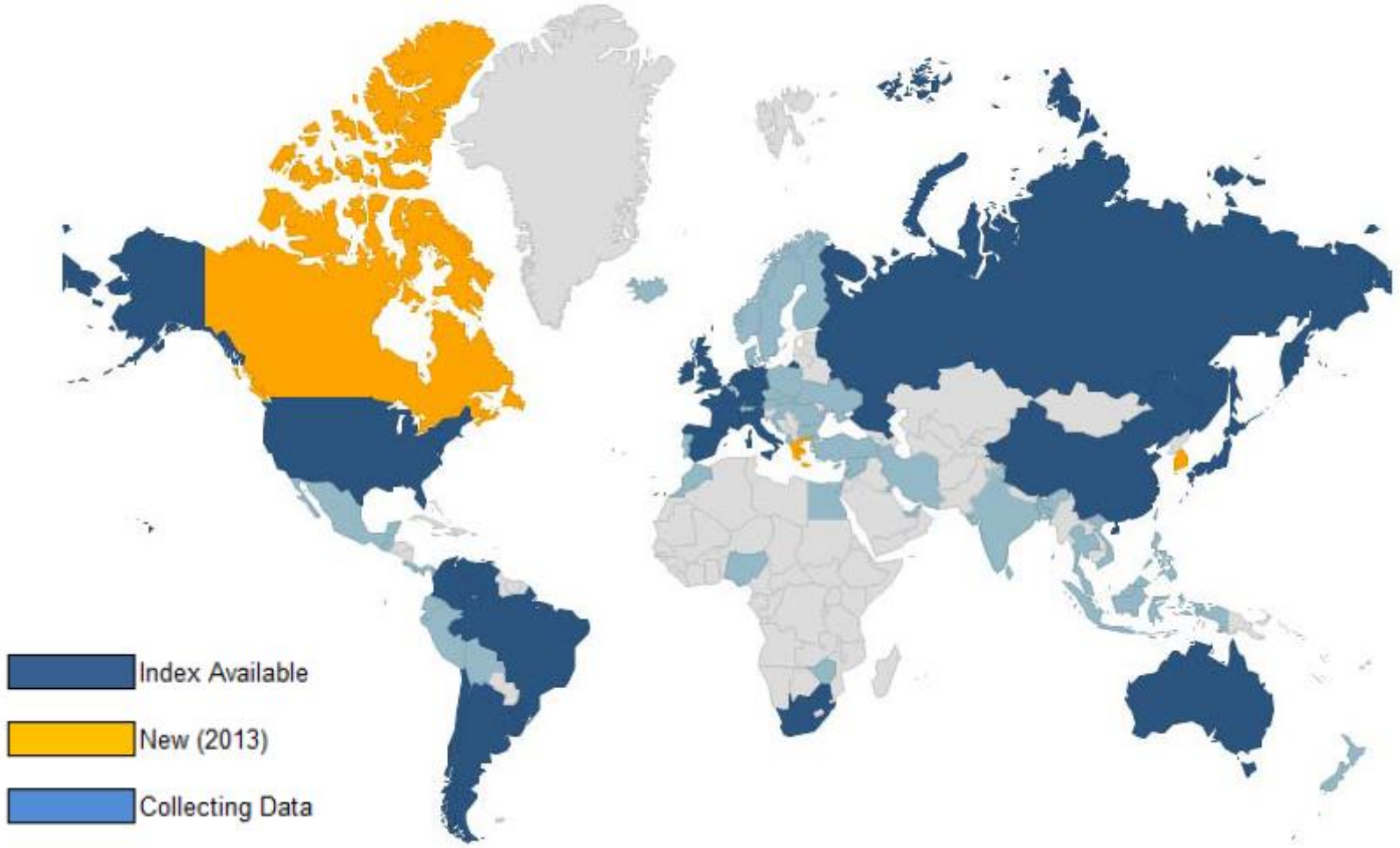
4 ***Store key item information in a database***

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

5 ***Calculate real-time Indexes***



Online Data and Daily Inflation



+ Global Series: Developed, Developing, Eurozone - All items, Fuel, Food

The Process Requires Three Stages

Technology & Processes

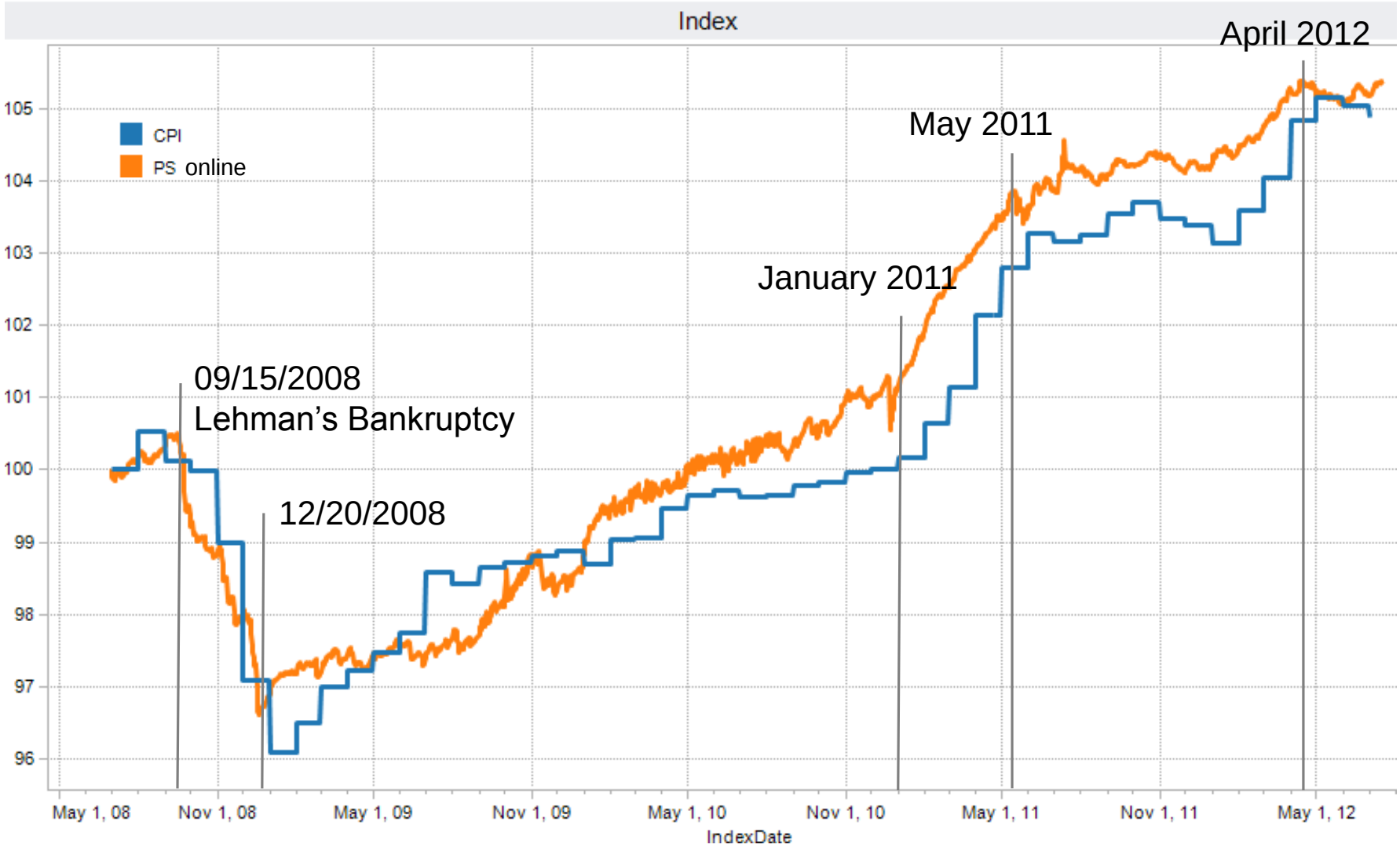
Data Capture	Data Processing	Index Computation
▪ Careful pre-selection of data sources ▪ Multiple specialized applications for each step of the data scraping process ▪ Software requires customization for each retailer ▪ Needs to be updated when there are website changes	▪ Data is automatically cleaned and homogenized ▪ Monitoring controls to check scraping, cleaning, statistics ▪ Classify and select items and sources to be used in the indices	▪ Statistics and indices are computed on a daily basis ▪ We use standard CPI techniques as much as possible ▪ For aggregate indices, compensate for things that we cannot monitor online (eg services)

Online Price Indices vs Official CPIs

- We focus on *measurement* of the *same phenomenon* (inflation) with an *alternative source* of data

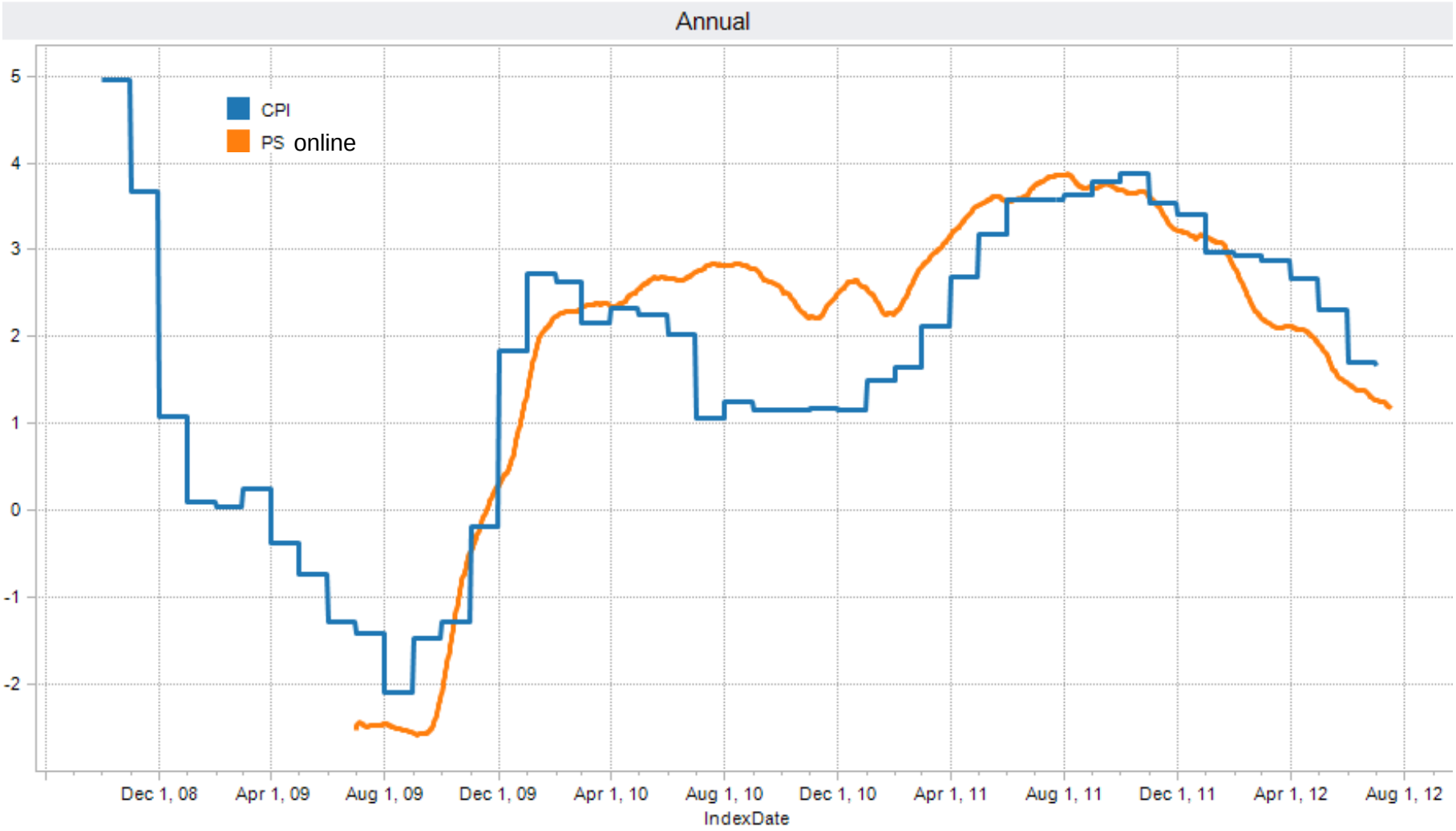
- Compared to CPIs, three main characteristics
  - Congruence
  - Differences in the short run
  - Anticipation of major changes in inflation trends

US Daily Price Index



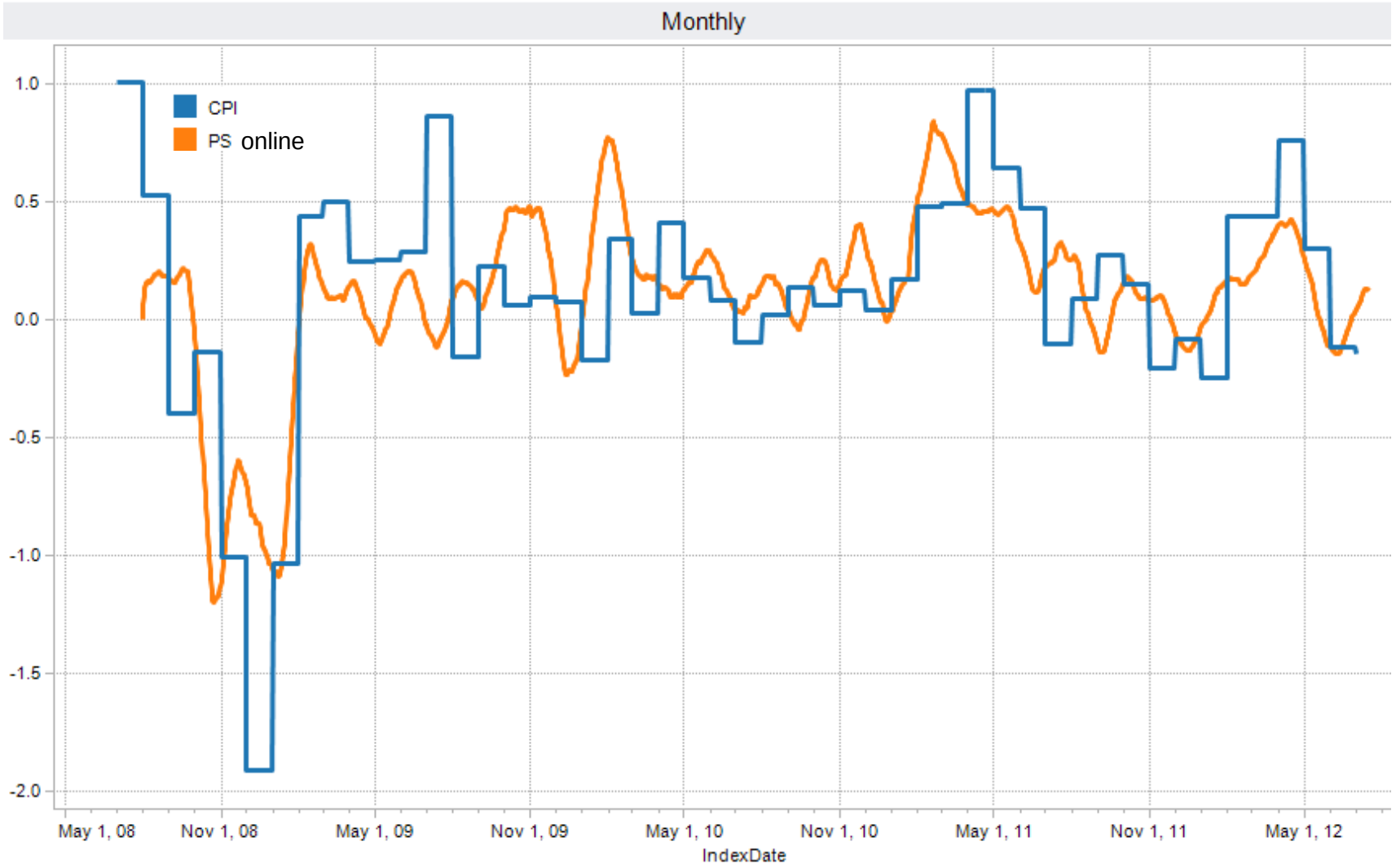
Source: BPP – PriceStats – BLS (CPI-U, US city-average, all items, NSA). Updated until 7/17/2012.

US Annual Inflation



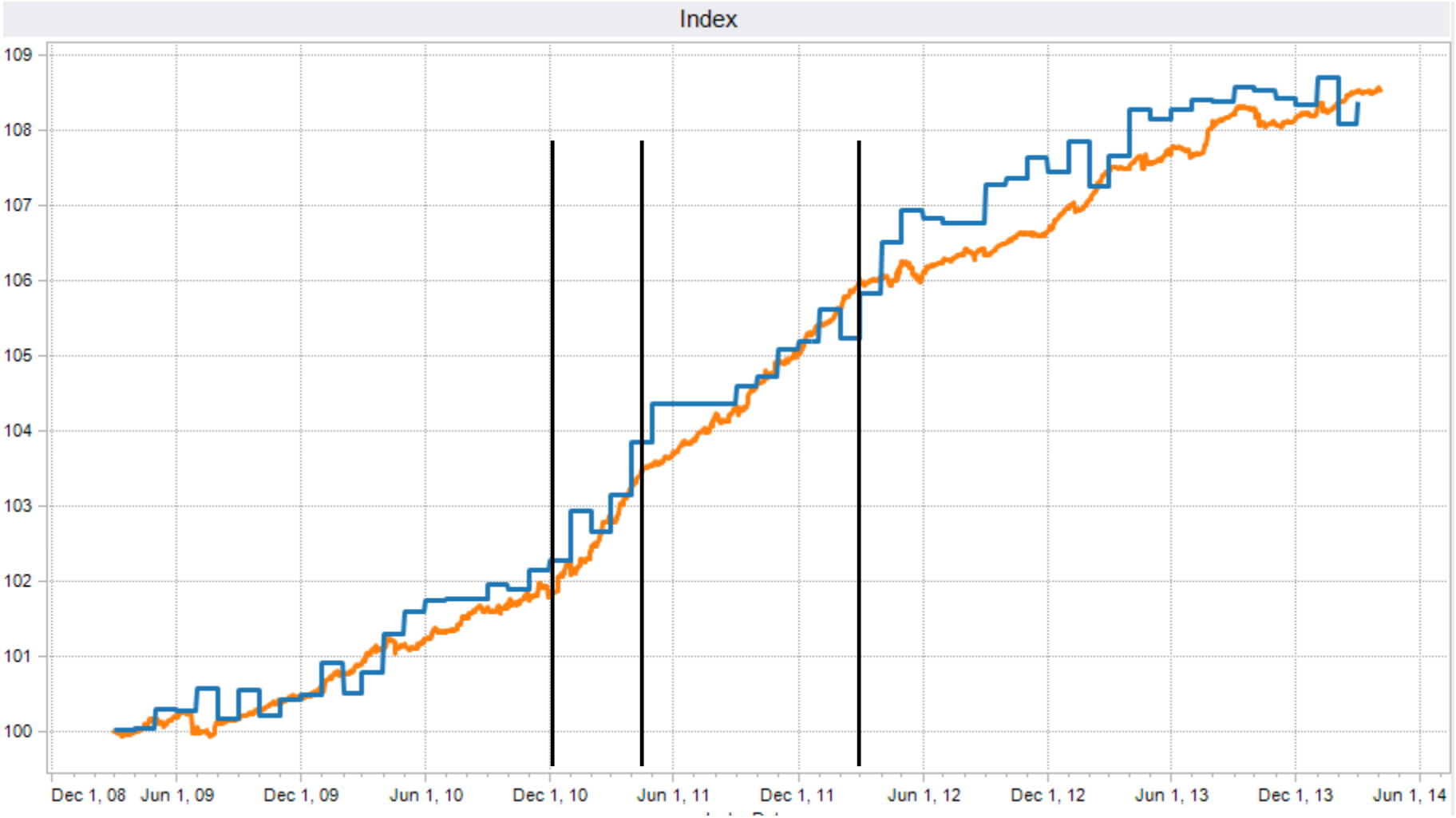
Source: BPP – PriceStats – BLS (CPI-U, US city-average, all items, NSA)

US Monthly Inflation

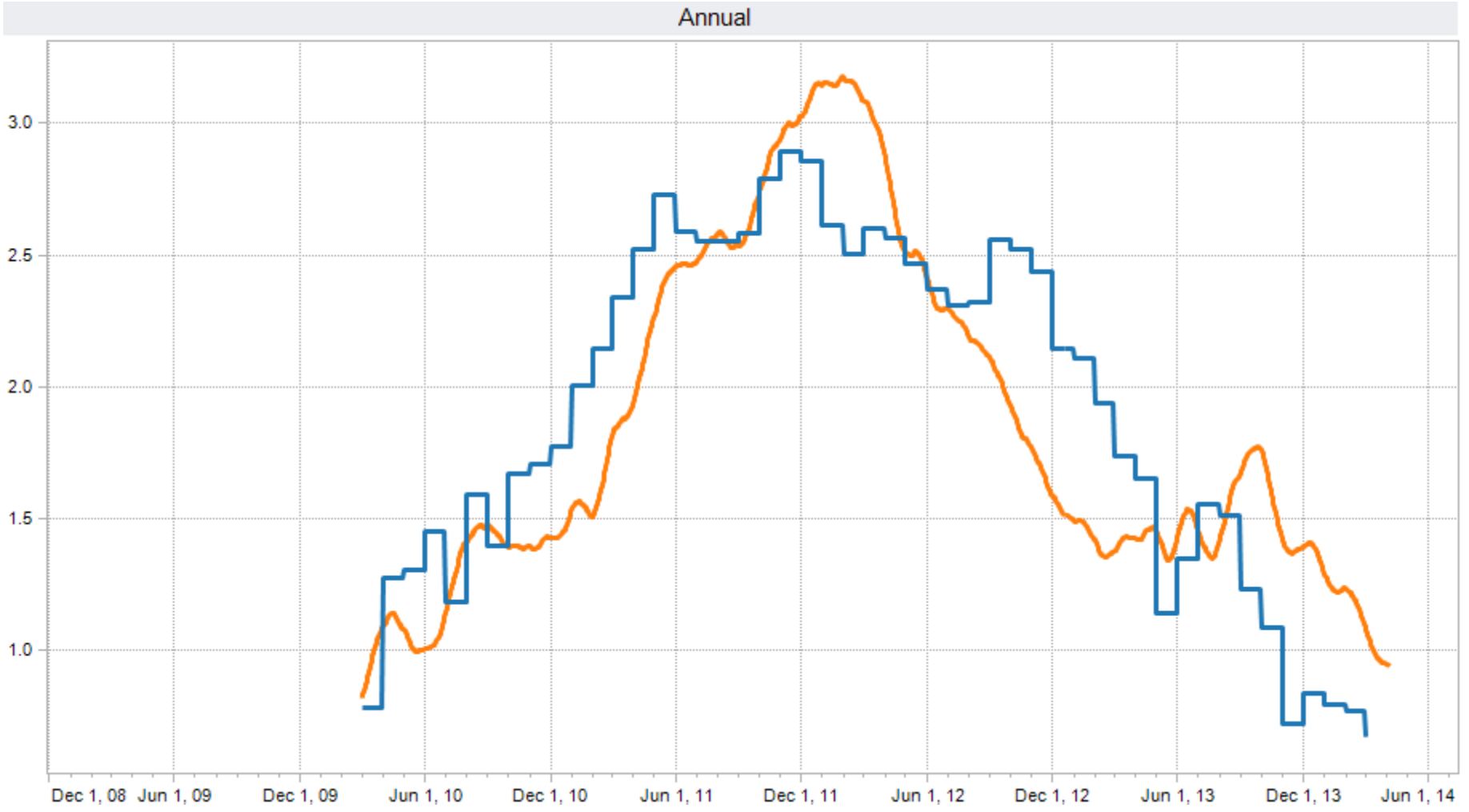


Source: BPP – PriceStats – BLS (CPI-U, US city-average, all items, NSA)

Eurozone Index

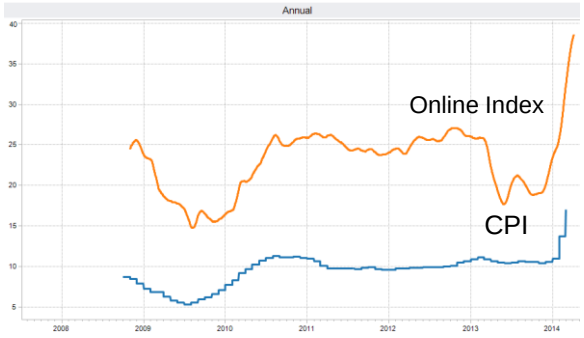


Eurozone Annual Inflation

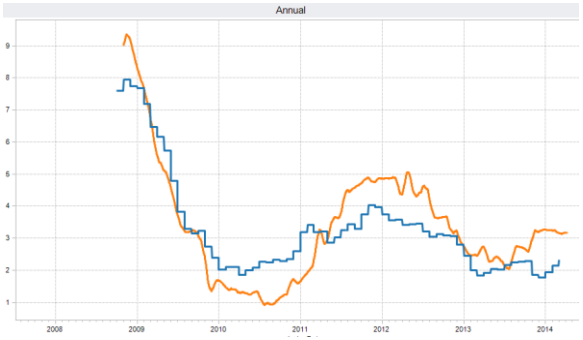


Annual Inflation Rates in Other Countries

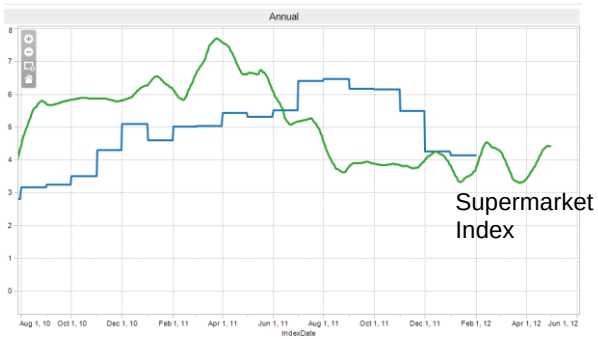
Argentina



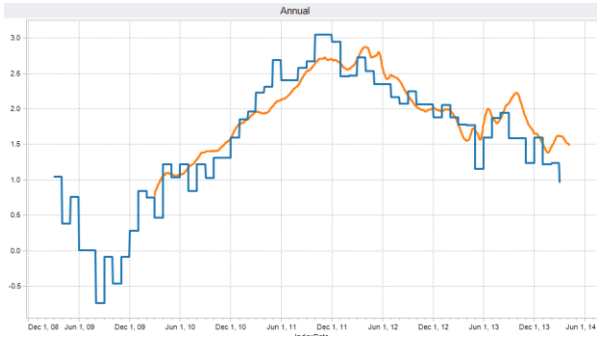
Colombia



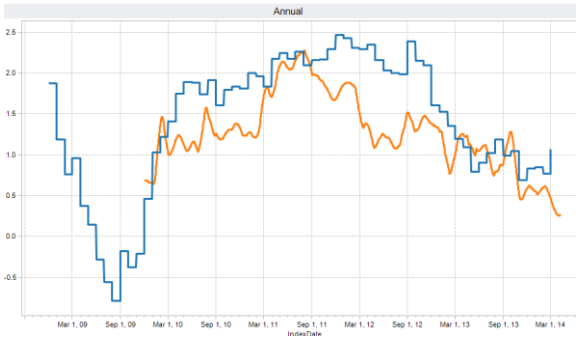
China



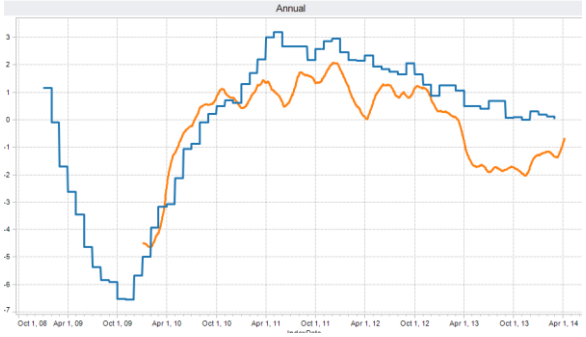
Germany



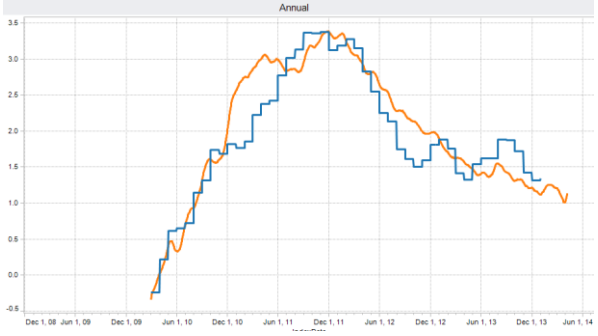
France



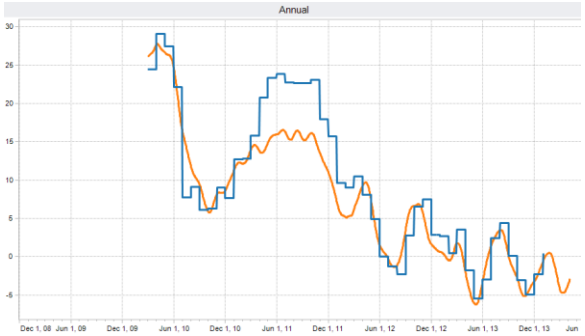
Ireland



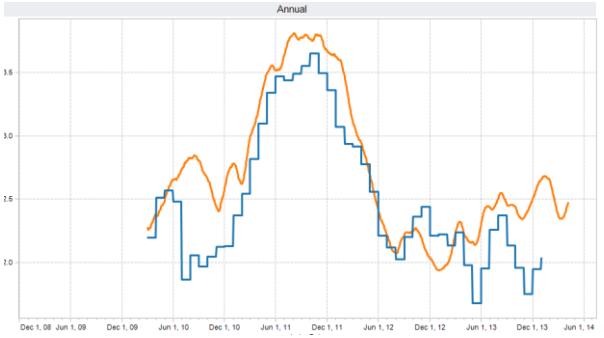
Developed Food



Developed Fuel



Global - Aggregate



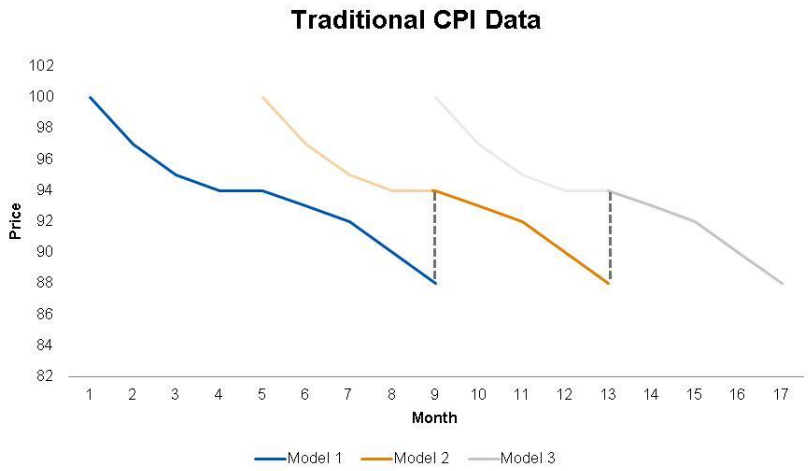
Source: PriceStats - StateStreet

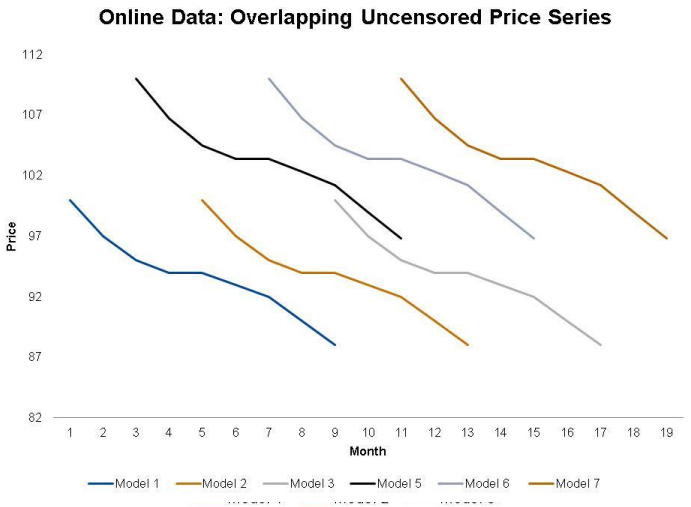
Differences: Online vs Offline Markets

- Are online prices representative of the economy?
  - Online sales are still only about 10% of retail sales in developed countries
- However:…
  - Online and Offline market are tightly integrated in many countries (people search online even when they end up buying offline)
  - Studies with simultaneous sampling show retailers tend to have either identical online and offline prices (eg: Apple, Ikea, Zara, H&M, Cavallo et al 2014) or stable online 'markups' (see Cavallo (2012))
  - 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
- Still, as always with macro, the right answer is "it depends" (on the country, on the particular conditions, etc).

Differences: 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



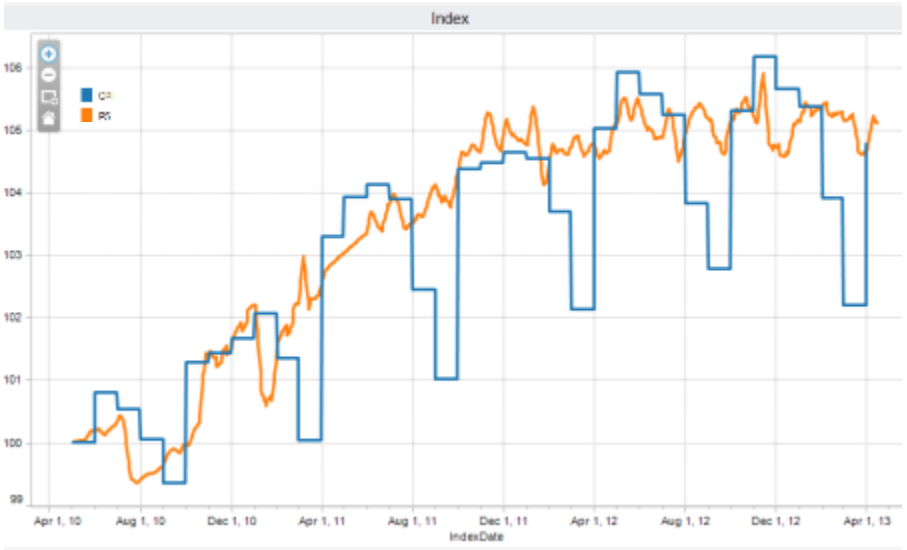


Differences: Quality Adjustments

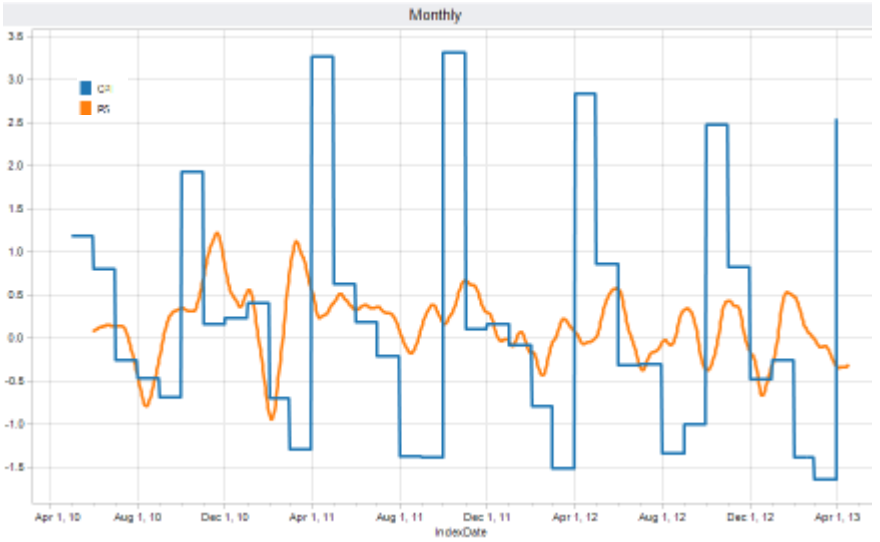
- Simple indices can approximate the level and trend of CPI inflation in hedonic-adjusted categories (as suggested in Silver & Heravi (99), Aizcorbe, Corrado & Doms (2003))

- Implication: online series tend to be smoother than CPIs

Greece Price Index



Greece Monthly Inflation

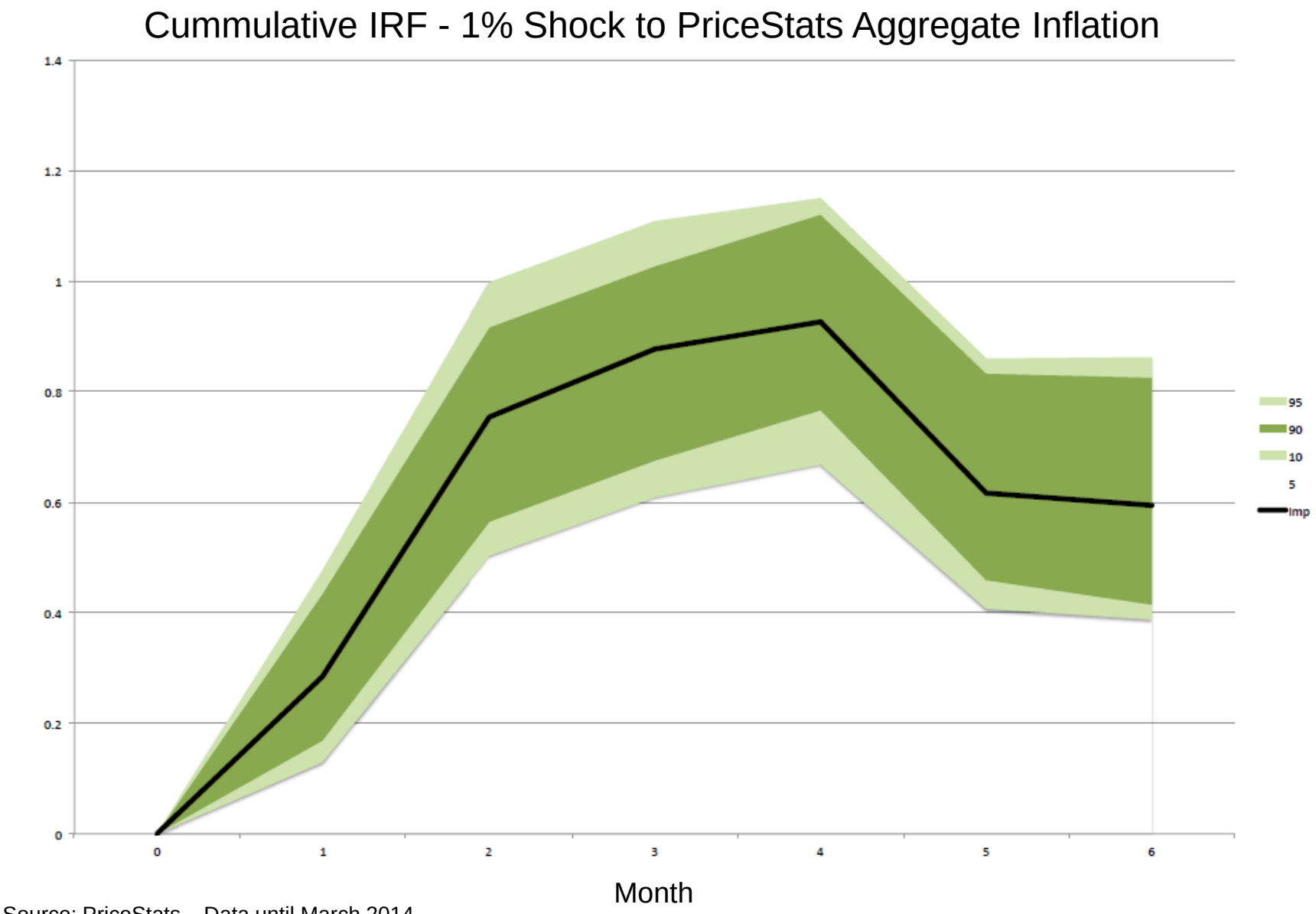


Anticipation

- We consistently find that online prices tend to anticipate changes in inflation trends.
  - This goes beyond the ability to collect and publish the data faster
  - 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

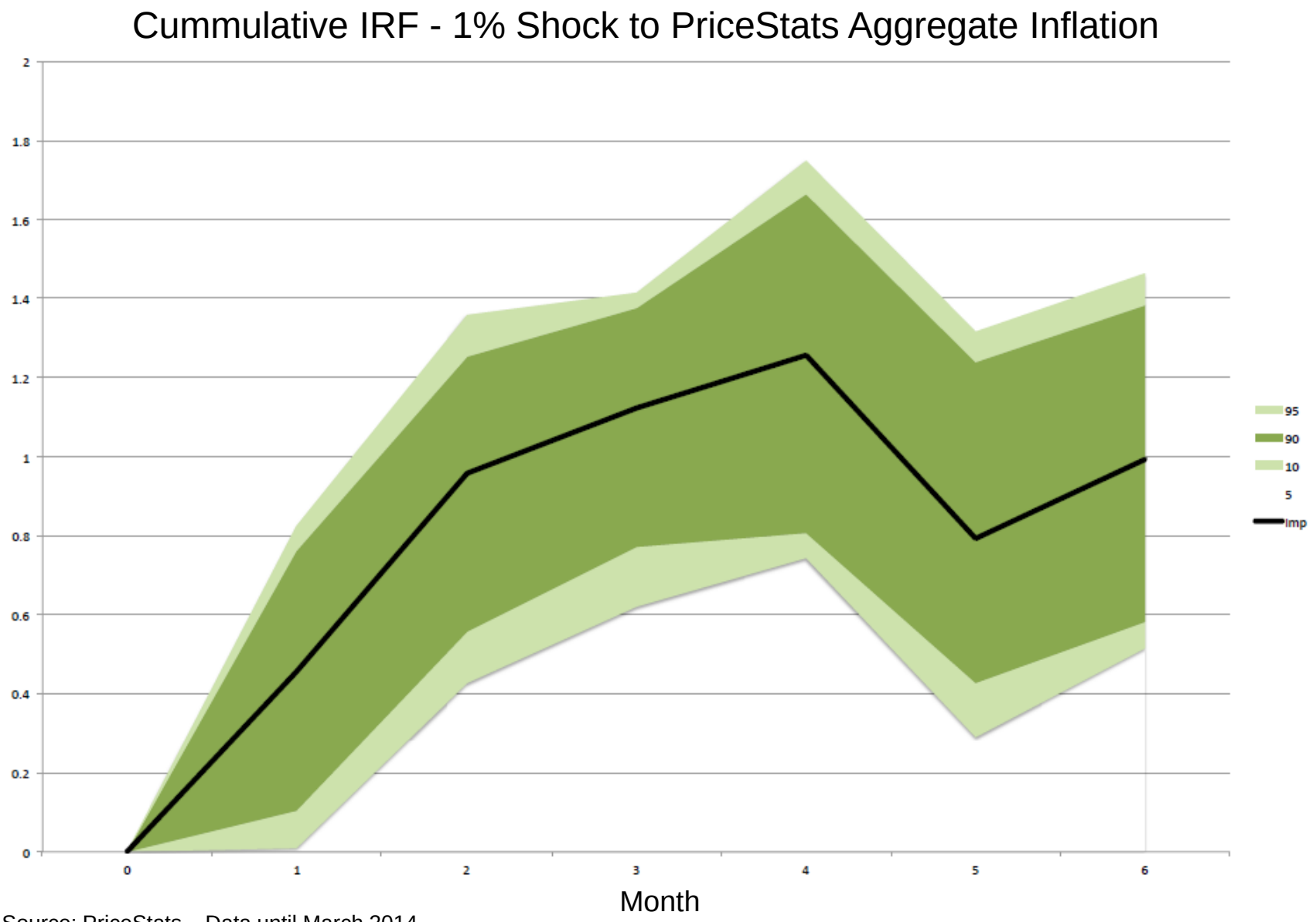
- We can study the link between online data and CPIs using simple VARs.
  - VAR regressions with Δ CPI on the LHS and lags of Δ CPI and ΔPS on the RHS (monthly data)
  - Impulse responses show the impact of a 1% shock in PriceStats on future CPI (reflecting additional information not contained in lagged CPI)

Impulse Response USA



Source: PriceStats – Data until March 2014

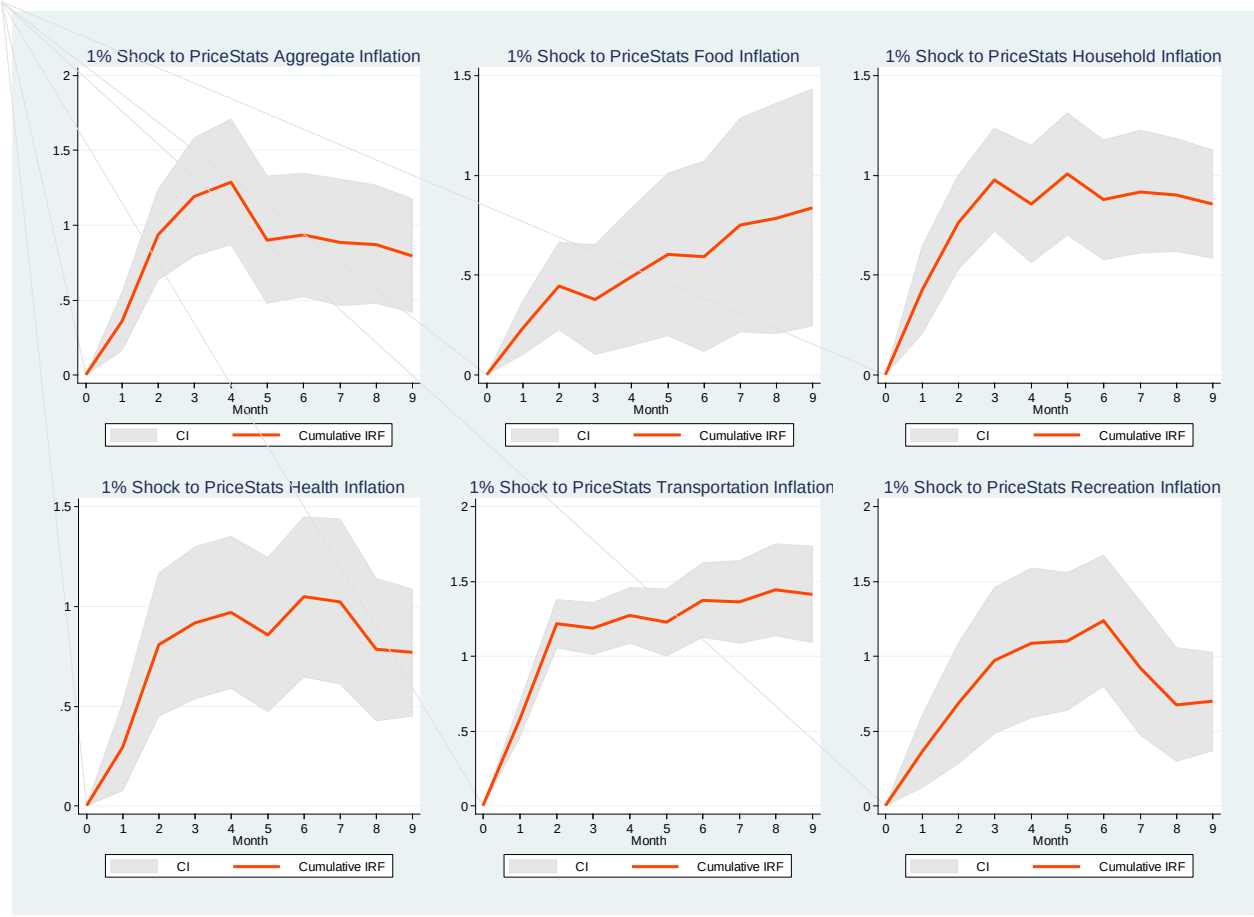
Impulse Response Eurozone



Source: PriceStats – Data until March 2014

Impulse Responses for US Sectors

- There are differences in anticipation across US sectors
  - Shortest in Fuel (1 month)
  - Longest in Recreation (5 months)



Source: PriceStats – Data until June 2013

Other variables?

- Do online prices add more information than Gas Prices?
- We can test it with a simple extension:
  1. Run a regression of ΔCPI on lags of ΔCPI.
  2. Regress the residuals of 1) on Δ Gas Prices.
  3. Regress the residuals of 2) on Δ PS Aggregate and Δ PS Transportation (US)

Results	R2
1)	38.915%
2)	60.596%
3)	22.325%

- PriceStats series explain 22% of the variability of CPI inflation that is not explained by Gas Prices and lagged CPI
- This is all within-sample

Forecasting Example

- We forecast the US CPI inflation rate for 24 months, from January 2012 to January 2014.
- Each month t we first estimate the following model:

$$\Delta\log\mathrm{CPI}_{t-1,t-2} = c_0 + \alpha_0\Delta\log\mathrm{HFD}_{t-1,t-2} + \sum_{i=1}^{l}\alpha_i\Delta\log\mathrm{HFD}_{t-1-l,t-2-l} + \sum_{i=1}^{l}\beta_i\Delta\log\mathrm{CPI}_{t-1-l,t-2-l}$$

  - Each variable is first seasonally adjusted by regressing it on monthly dummies and using the residuals.
  - The high frequency data (HFD) is either Gas Prices (values from last week of month) or any of the PriceStats (PS) series (values from the last day of month).
  - We try models with different number of lags $l \in \{1, 2, 3\}$, and using different estimation windows $\mathrm{W} \in \{24, 36, 48, 60\ months\}$
- We compute the fitted value for $\Delta\log\widehat{\mathrm{CPI}}_{t,t-1}$ using the coefficients from the previous regression and the high frequency data available on month t. This forecast is repeated for 24 months.

Forecasting Example

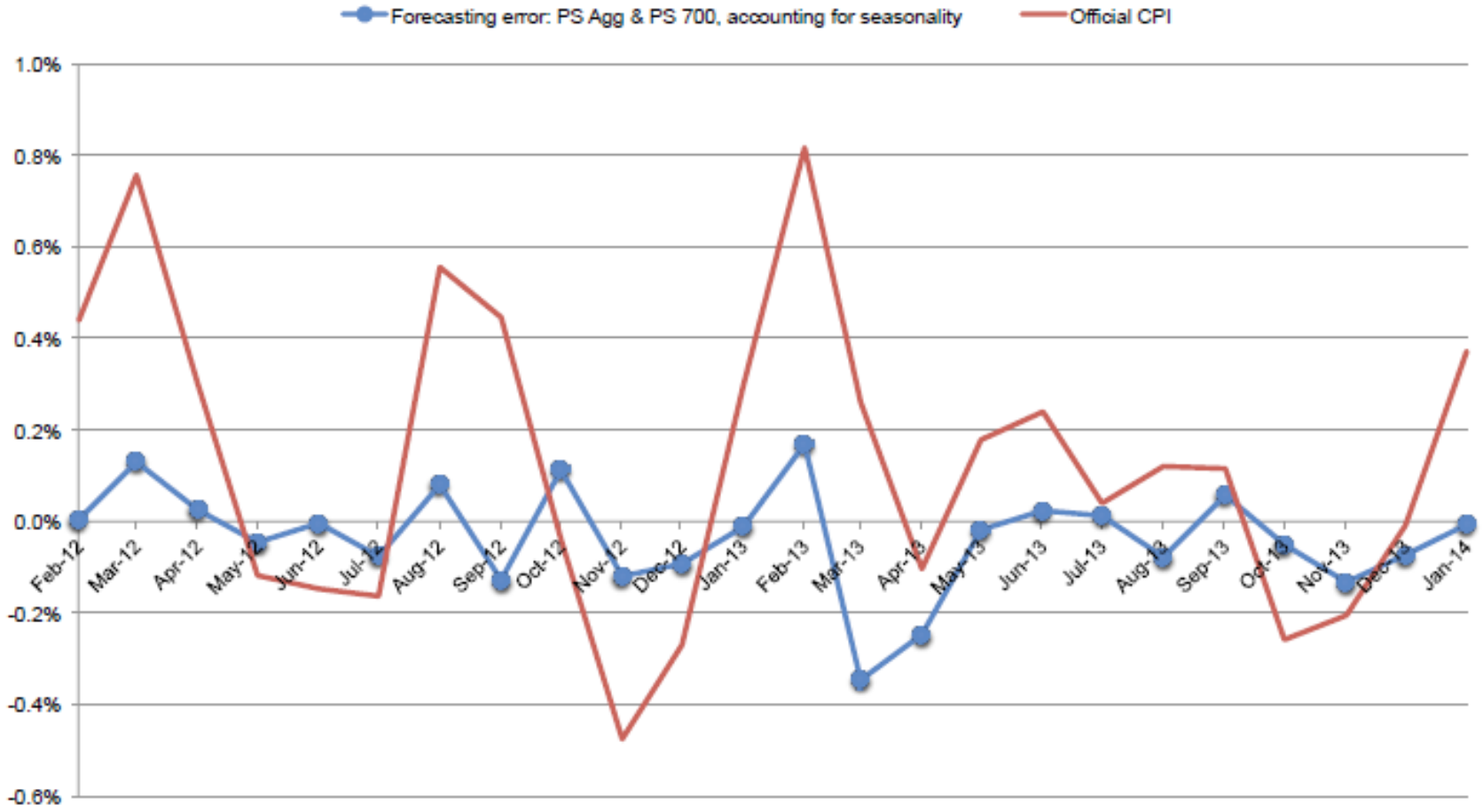
- We tried alternative models using
  - Only CPI
  - CPI and Gas Prices
  - CPI and PriceStats Series

Regressors	MAE %	RMSE %
CPI	0.164	0.221
CPI + Gas Prices	0.101	0.149
CPI + PS Aggregate	0.143	0.184
CPI + PS Food + PS Transportation	0.110	0.134
CPI + PS Aggregate + PS Transportation	0.085	0.117

Note: Averages of models with 1, 2, and 3 lags

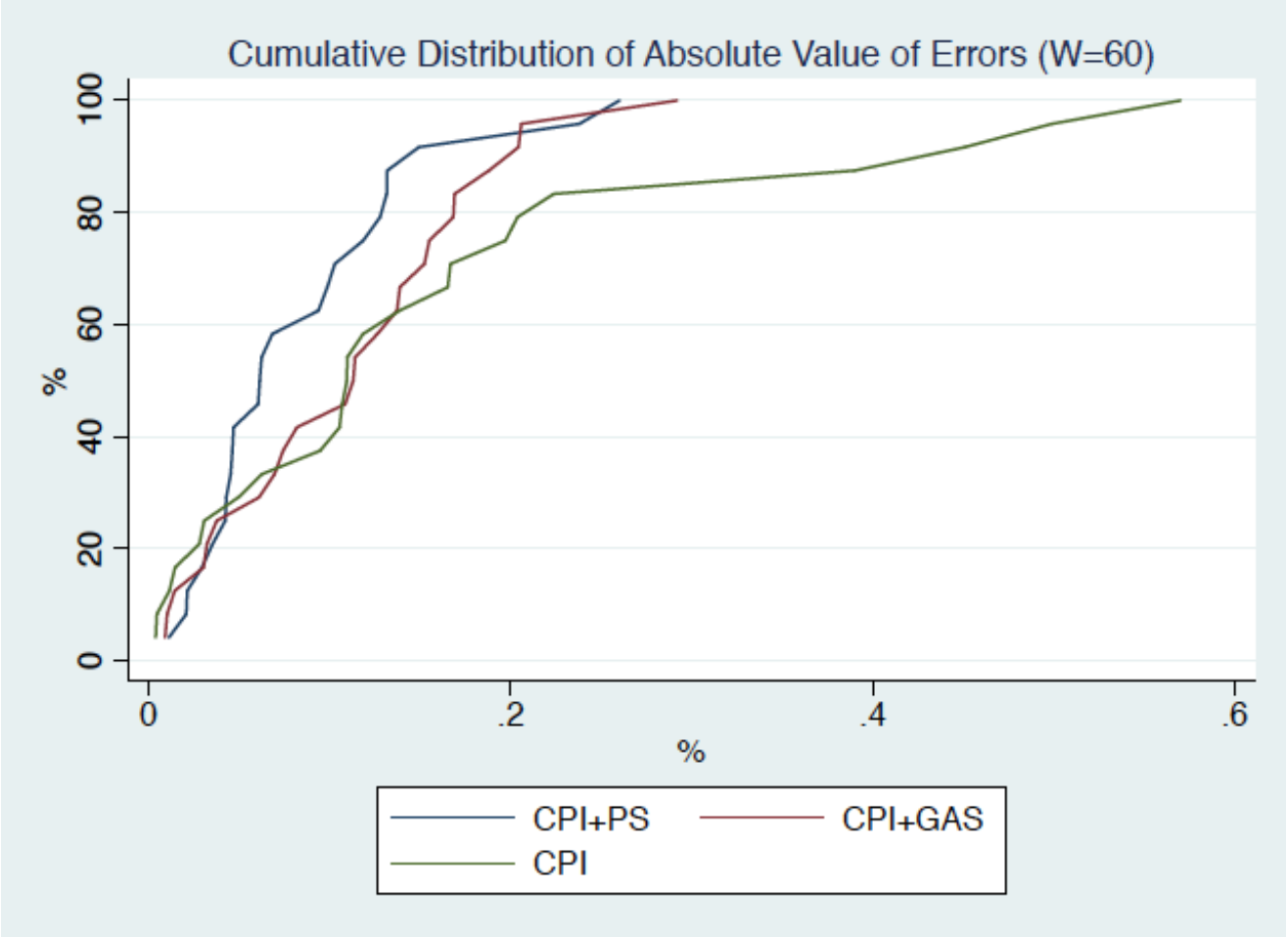
Forecasting Example

- Errors for model with W=24, seasonally adjusted, average of lags



Forecasting Example

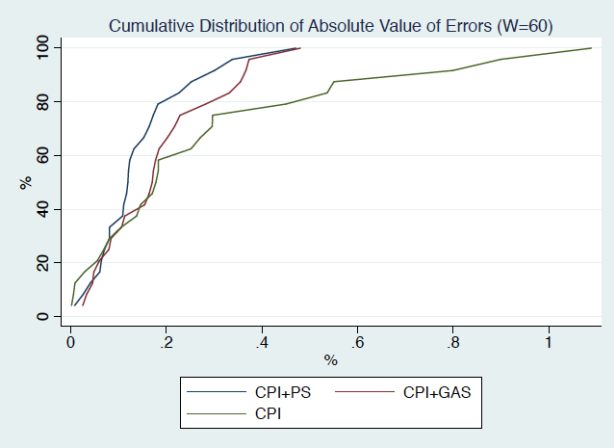
- Cumulative Distribution of Abs Value of Errors



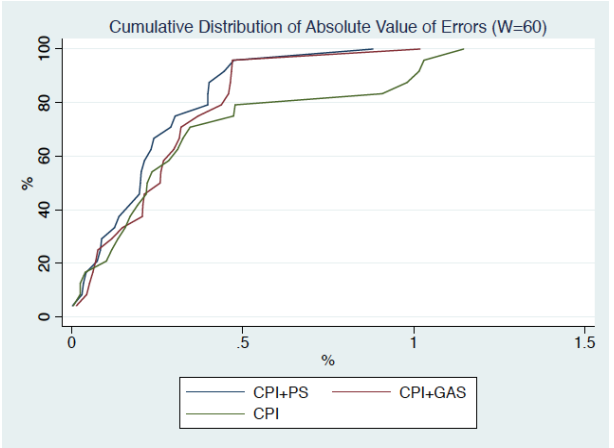
Forecasting Example

- We repeated the same exercise, but this time forecasting the inflation from t to t+i, where i ∈ $\{2, 3, 4, 5\ months\}$

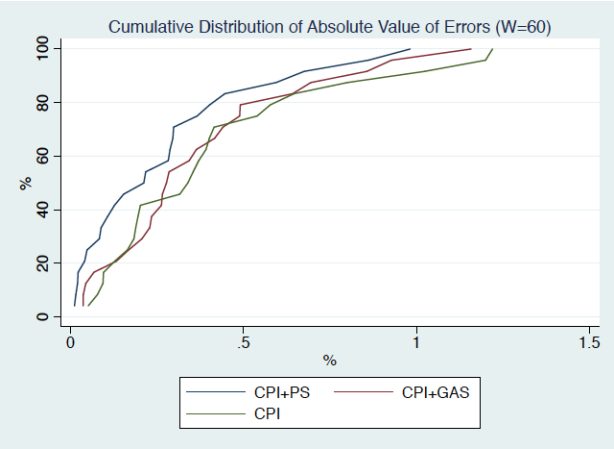
i = 2



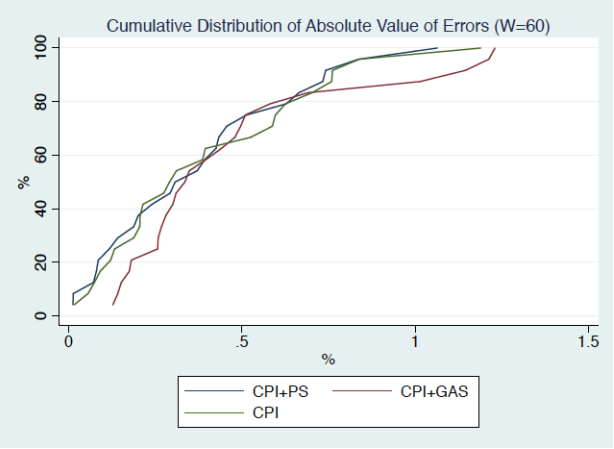
i = 3



i = 4



i = 5



Conclusions

- Online prices have the potential to dramatically increase the amount and quality of micro price data available for academic research.
  - Data characteristics help re-evaluate old empirical puzzles and answer questions that could not be tackled before
- Online data can also be a reliable source of information for inflation measurement
  - Congruence, differences, and heterogeneity
  - ``Big data`` characteristics can greatly *simplify* measurement
  - Best when used as an alternative source of data, not as a separate sector that needs special treatment
- Online prices tend to anticipate changes in inflation trends
  - Typical anticipation is 2-3 months
  - Provides unique information in real-time, different from other sources such as gas prices,
  - Particularly useful to forecast times of inflation ``surprises``
- Access to online data will become widespread in the following years.