Lecture 1: Online and Official Price Indexes: Measuring Argentina's Inflation (Cavallo, JME, 2013)

Big Data Analytics in Macroeconomics

Dooyeon Cho
Department of Quantitative Applied Economics
Sungkyunkwan University

Overview

- This paper studies whether prices collected from online retailers can match official inflation estimates in 5 Latin American countries.
  1. **Argentina**
  2. Brazil
  3. Chile
  4. Colombia
  5. Venezuela

- The availability of online prices represents a unique opportunity for **the construction of price indexes and the measurement of inflation** around the world.
  - Prices collected from online retailers can be used to construct daily price indexes that complement official statistics.

Overview (cont'd)

- The data were collected between October 2007 and March 2011 from the largest supermarket in each country
  - An unprecedented amount of micro-level price data can now be collected using special software.
  - This type of data can be collected remotely, **at much higher frequencies**.

- Among 5 Latin American countries, for Argentina, official statistics have been heavily criticized in recent years.
  - In Brazil, Chile, Colombia, and Venezuela, online price indexes approximate both the level and main dynamics of official inflation.
  - By contrast, Argentina's online inflation rate is **nearly three times higher** than the official estimate.

Introduction

- Price indexes constructed with online data can be used to obtain alternative inflation estimates in countries where official estimates have lost their credibility.
  - The Argentine government has been **manipulating the official inflation indexes** since 2007.

- A combination of i) online prices, ii) standard CPI methodologies and iii) official category weights, is used to build an "online price index" in each country.
  - Each online index is then compared to an equivalent official supermarket index, constructed as a weighted average of the official CPI components of food, beverages, and household products (the same categories available in the online data).

Introduction (cont'd)

- The matching is best in Chile, with an average annual inflation of 3.00% online and 3.19% offline, and a correlation of 0.97 in the annual inflation series.
  - The match is also close in Colombia, both for the level and dynamics of annual inflation.
  - In Brazil and Venezuela, the online index is able to match the main inflation trend of the official index, but the annual inflation series is less synchronized over time.

- For Argentina, there is **a large discrepancy between online and official price indexes** that is persistent over time.
  - For over 3.5 years, online prices had an annual inflation rate that was consistently **2 to 3 times higher** than in official statistics, with an average rate of 20.14% for the sample period compared to just 8.38% in official data.
  - Surprisingly, although the level of inflation is higher, the dynamic behavior of online inflation matches the official data quite well.

Data

- The data were collected by the Billion Prices Project at MIT using a technique called "**web scraping**" to record the price for all goods sold online, between October 2007 and March 2011, in the largest supermarket in Argentina, Brazil, Chile, Colombia, and Venezuela.

- **Web scraping**
  - Most web pages are built using a structured coding language called Hyper Text Markup Language (HTML). This code has simple "tags", such as and <bold>, that determine the style and placement of text in a page. These tags tend to remain constant over time, as they provide a distinctive "look and feel" to each page.
  - By contrast, the information within these tags, such as a product's price, changes all the time. The scraping software can be taught to use the HTML tags to locate relevant information about a product and store it in a database.
  - Repeating the process every day produces a panel database with a one record per product per day.

Data (cont'd)

- In all cases, the online data contains a combination of food, beverages, and household products. Categories range from "Eggs" to "Appliances", with about a third of them corresponding to household products (including cleaning materials, health and beauty products, furniture, appliances, and books).
  - These categories account for between 28.44% (Colombia) and 48.51% (Argentina) of CPI weights in these countries.

Table 1
Online data description.

	Argentina Retailer #1	Argentina Retailer #2[a]	Brazil	Chile	Colombia	Venezuela
Starts	10/7/2007	23/7/2007	10/10/2007	10/24/2007	11/13/2007	04/16/2008
Ends	3/24/2011	03/20/2011	03/01/2010	03/20/2011	03/24/2011	03/01/2010
Prices P/day (mean)	11,560	4790	11,000	12,000	5000	9256
Total products	26,333	10,929	21,804	35,432	9166	20,847
Price changes	204,449	136,781	25,9875	12,0112	76,979	94,808
Category indicator	Yes	No	Yes	Yes	Yes	No
CPI weights covered	48.51%	–	27.93%	31.00%	28.44%	–
Retailer market share[b]	28%	n/a	15%	27%	30%	n/a

[a] Argentina's Retailer #2 is used only in the robustness results discussed in Section 5.3.2 and Fig. 4.
[b] Market shares are based on the information posted on the corporate webpages of each supermarket.

Online price indexes

- The online price indexes use a combination of online data and official category weights.
  1. Daily data are used to construct the online price indexes.
  2. The online indexes are built using prices for all products available for purchase at each retailer.
  3. There are no forced product substitutions or adjustment for quality changes.

CBA Index Weights

- The CBA basket is constructed with a given number of grams per product. There are 45 products included in the index, detailed in Table C.2. These products and their respective grams are set by the INDEC to meet the minimum nutritional requirements for an adult male in the 30-59 age range. We calculate the daily cost of the basket using 45 items from the scraped data that were carefully chosen to match the products in the INDEC listing.

Table C.2: CBA Index Weights

INDEC		**Scraped Data**		
Product (1)	Grams (2)	Product (3)	Package size in grams (4)	Index Weight (2)/(4)
Pan	6,060	Pan	500	12.12
Galletitas Saladas	420	Galletas Saladas	1000	0.42
Galletitas Dulces	720	Galletas Dulces	160	4.50
Arroz	630	Arroz Largo Fino	500	1.26
Harina De Trigo	1,020	Harina De Trigo	1000	1.02
Otras Harinas (Maz)	210	Harina Maiz	500	0.42
Fideos	1,290	Fideos	1000	1.29
Papa	7,050	Papa Negra	1000	7.05
Batata	690	Batata	1000	0.69
Azcar	1,440	Azucar	1000	1.44
Dulce De Leche	80	Dulce De Leche	250	0.32
Dulce De Batata	80	Dulce De Batata	1000	0.08
Mermeladas	80	Mermelada De Frutilla	454	0.18
Legumbres Secas				
Lentejas	80	Lentejas	500	0.16
Porotos	80	Porotos	500	0.16

CBA Index Weights (cont'd)

Arvejas	80	Arvejas Verdes	350	0.23
Cebolla	655	Cebolla	1000	0.66
Lechuga	655	Lechuga Francesa	1000	0.66
Tomate	655	Tomate	1000	0.66
Zanahoria	655	Zanahoria	1000	0.66
Zapallo	655	Zapallo	1000	0.66
Tomate En Lata	655	Tomate Perita	400	1.64
Manzana	2,010	Manzana	1000	2.01
Naranja	2,010	Naranja De Jugo	1000	2.01
Asado	896	Asado Centro Novillito	1000	0.90
Carne Picada	896	Carne Picada	1000	0.90
Carnaza	896	Carnaza Comun	1000	0.90
Cuadril	896	Colita De Cuadril	1000	0.90
Nalga	896	Nalga	1000	0.90
Paleta	896	Paleta De Cerdo	1000	0.90
Pollo	896	Pollo Sin Piel	1000	0.90
Huevos	630	Huevo	300	2.10
Leche	7,950	Leche Entera	1000	7.95
Fresco	90	Queso Crema	200	0.45
De Rallar	90	Queso Rallado	210	0.43
Crema	90	Queso Cremoso	1000	0.09
Aceite	1,200	Aceite Mezcla	1000	1.20
Bebidas Edulcoradas	4,050	Coca Cola	2000	2.03
Bebidas Gaseosas Sin Edulcorar	3,450	Soda	2000	1.73
Sal Fina	150	Sal Fina	500	0.30
Sal Gruesa	90	Sal Gruesa	500	0.18
Vinagre	90	Vinagre Alcohol	1000	0.09
Caf	60	Cafe Molido	500	0.12
T	60	Te	50	1.20
Yerba	600	Yerba	500	1.20

Index computation

- Price changes are calculated at the product level, then averaged inside categories using unweighted geometric means, and then aggregated across categories with a weighted arithmetic mean.

- The first step is to obtain the unweighted geometric average of price changes in category j for each day t:

$$R^j_{t,t-1} = \prod_i \left(\frac{p^i_t}{p^i_{t-1}} \right)^{\frac{1}{n_{j,t}}}$$

  - where p^i_t is the price of good i at time t, and $n_{j,t}$ is the number of products in category j that are present in the sample that day.

Index computation (cont'd)

- The second step is to compute the category-level index at t:

$$I_t^j = R_{1,0}^j \cdot R_{2,1}^j \ldots \cdot R_{t,t-1}^j$$

- Finally, the Supermarket Index is the weighted arithmetic average of all category indexes:

$$S_t = \sum_j \frac{w^j}{W} I_t^j$$

  - where w^j is the official CPI weight for that category and W is the sum of all the weights included in the sample.

- Compared to CPI statistics, these online indexes have an advantage in terms of **frequency and the number of items** sampled within each category.

Monthly and Annual Inflation using Daily Data

- Formally, the monthly inflation rate π_t^m at time t is defined as the percentage change in the average of the index from t to $t-29$ and the average from $t-30$ to $t-59$.

$$\pi_t^m = 100 \times \left[\frac{\frac{1}{30} \sum_{x=0}^{29} S_{t-x}}{\frac{1}{30} \sum_{x=30}^{59} S_{t-x}} - 1 \right]$$

  - On the last day of each calendar month, π_t^m is comparable to the monthly inflation reported in official statistics.

Monthly and Annual Inflation using Daily Data (cont'd)

- Similarly, the annual inflation rate is computed as the percentage change in the index in the past 30 days over the same period 365 days ago,

$$\pi_t^y = 100 \times \left[\frac{\frac{1}{30} \sum_{x=0}^{29} S_{t-x}}{\frac{1}{30} \sum_{x=365}^{394} S_{t-x}} - 1 \right]$$

  - On the last day of each calendar year, π_t^y is comparable to the annual inflation rate reported in official statistics.

Results: Four Latin American countries of Brazil, Chile, Colombia, and Venezuela

- All indexes are weighted using official CPI weights, with the exception of Venezuela, where the online data could not be classified into the standard official categories.
  - In that case, an unweighted index is constructed with a simple geometric average of all price changes observed each day.

- Figure 1 shows a remarkable ability of online indexes to **track the main inflation trends** over long periods of time.

- Table 2 shows that the average annual inflation rates for the period are nearly identical for all countries.

- The differences in the online-official matching across countries appears to be linked to **the representativeness of the retailer** used in online data.

Results: 4 Countries (cont'd)

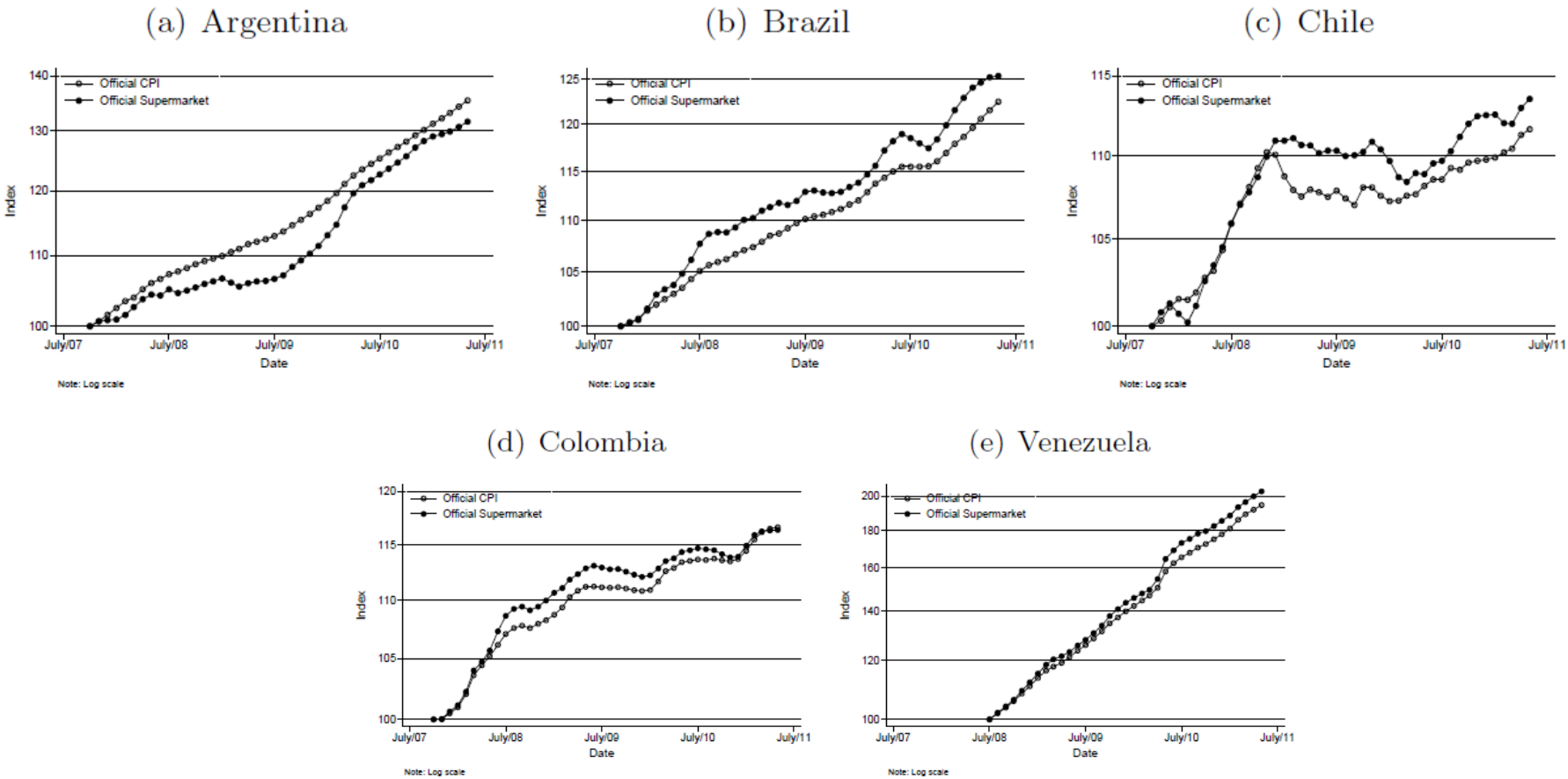


Figure B.3: Official CPI and Official Supermarket Indexes

Notes: The official supermarket index is constructed as a weighted average of the Food and Beverage and Household Product official price indexes in each country.

Results: 4 Countries (cont'd)

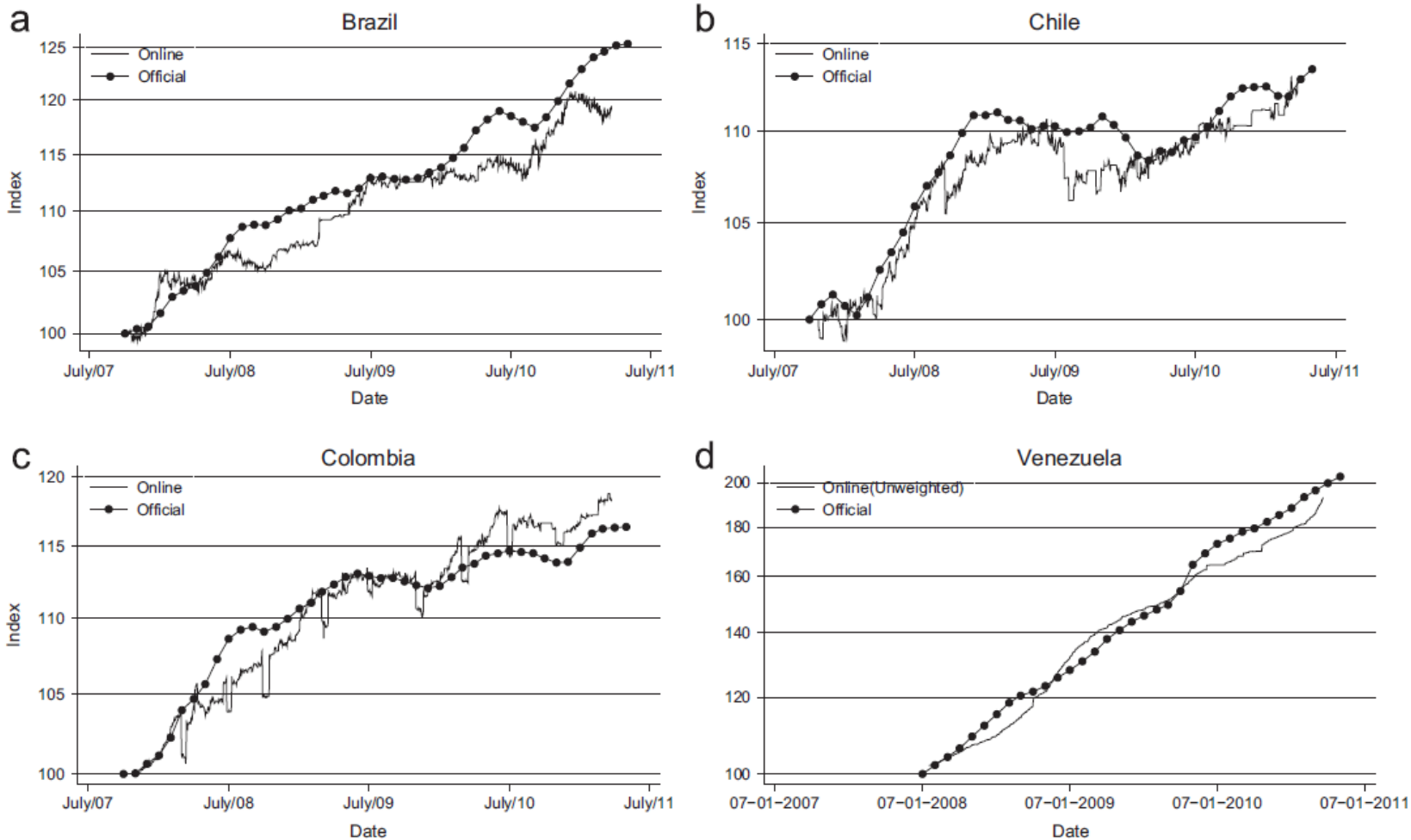


Fig. 1. Online and official price indexes in four Latin American countries. *Notes*: The daily online supermarket index is constructed with an online prices and official CPI category weights. In Venezuela, the online data has no category information and therefore the online index is built as a geometric average of all price changes observed each day. The official supermarket index is an equivalent indicator constructed as a weighted average of the "Food and Beverages" and "Household Products" official price indexes in each country.

Results: 4 Countries (cont'd)

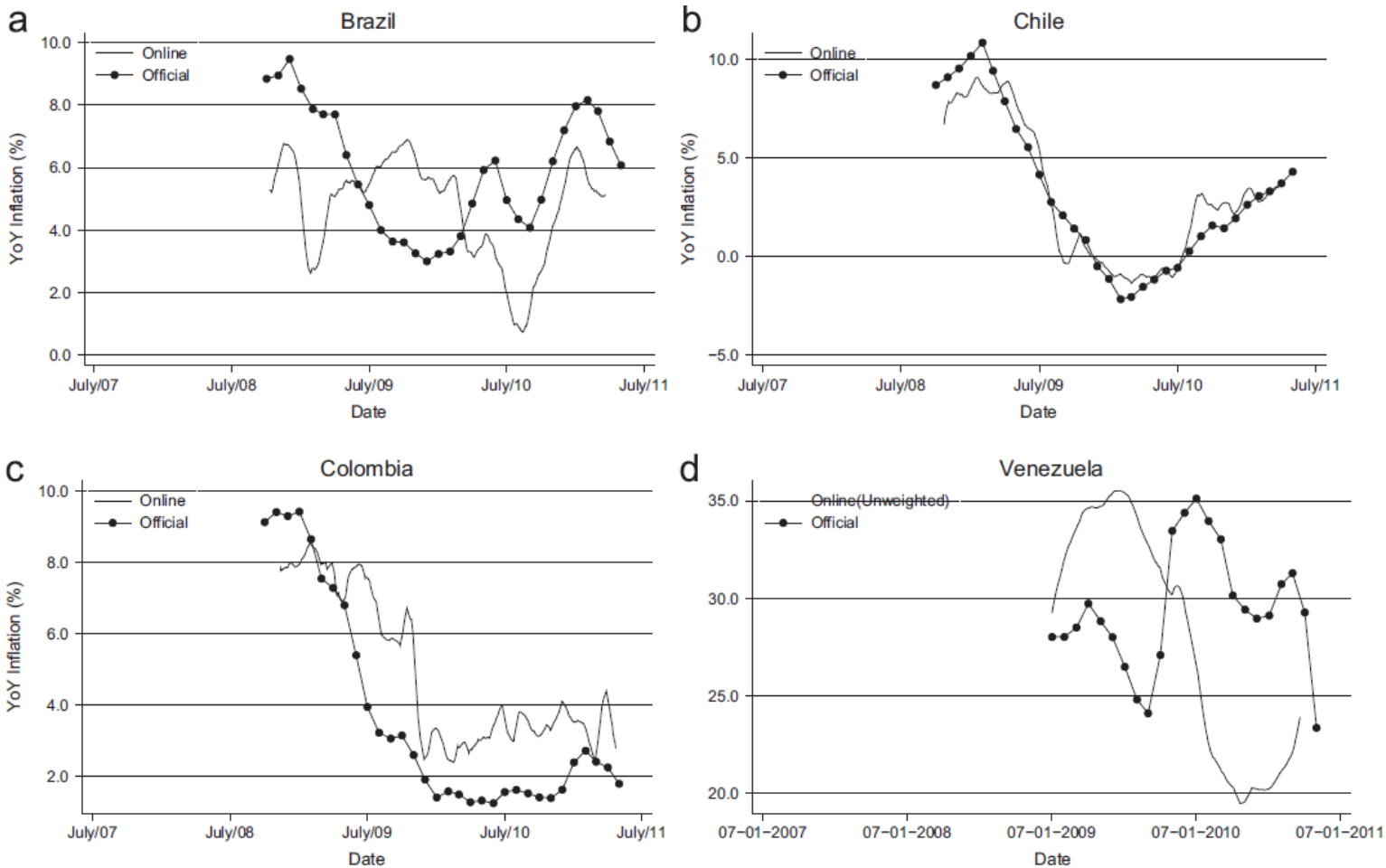


Fig. 2. Online and official annual inflation rates. *Notes*: The annual online inflation rate is a daily time series computed as the percentage change in the average of the index during the previous 30 days with respect to the average of the index in the same period a year before. The annual official inflation rate is a monthly time series computed as the percentage change in the index in the previous 12 months.

Results: 4 Countries (cont'd)

Table 2
Online vs. official series.

	Argentina	Brazil	Chile	Colombia	Venezuela
Mean annual inflation (%)					
Official CPI index	8.53	5.28	2.44	3.79	27.37
Official supermarket index	8.38	5.91	3.19	3.73	29.38
Online supermarket index	20.14	4.72	3	4.88	27.43
Correlations between online and official supermarket series					
Price index	0.98	0.96	0.97	0.95	0.92
Annual inflation	0.84	0.09	0.97	0.89	−0.08
Monthly inflation	0.6	0.5	0.38	0.43	0.18
Regression – official supermarket on online monthly inflation rates (12 lags)					
Constant	0.84	−0.54	0.14	0.03	−1.96
Constant *p*-value	0.000	0.19	0.17	0.86	0.23
R2	0.9	0.55	0.66	0.59	0.66
Monthly inflation rate volatility (standard deviation)					
Official supermarket index	0.57	0.51	0.58	0.48	0.98
Online supermarket index	1.11	0.73	0.62	0.76	0.86

Note: The top panel shows that Argentina is the only country where online data does not approximate the average official annual inflation rates. However, the second panel shows that the correlation between the monthly inflation rates is higher in Argentina than in any of the other countries. The discrepancy is therefore in the level of inflation reported, not its dynamic behavior over time. The third panel reinforces this idea with a simple OLS regression of the official monthly rate and 12 lags of the online monthly rate. The R2 is highest in Argentina, which is also the only country where the constant is statistically significant. The fourth panel shows that the official monthly inflation rate is surprisingly stable in Argentina compared to both the online index and the volatility observed in a high-inflation country like Venezuela.

Results: Argentina

- Government intervention in the statistical office
  - Since 2003, Argentina's inflation grew steadily as a result of an expansionary monetary policy designed to stimulate consumption and avoid an appreciation of the currency.
  - Inflation became a **politically sensitive issue** in 2006, when the annual inflation rate increased above 12%.
  - The Argentine government has intervened the National Statistics Institute (INDEC) since January 2007.

- Large differences with official data
  - In Argentina's case, the online price index follows a completely different trend than the official index, as shown in Figure 3.
  - Between October 2007 and March 2011, the online index increased over 100%, while the official index grew only 35%.

- Main findings
  1. First, online inflation has been consistently between **2 and 3 times higher** than official inflation.
  2. Second, the online and official estimates share **a surprisingly similar pattern** over time.

Results: Argentina (cont'd)

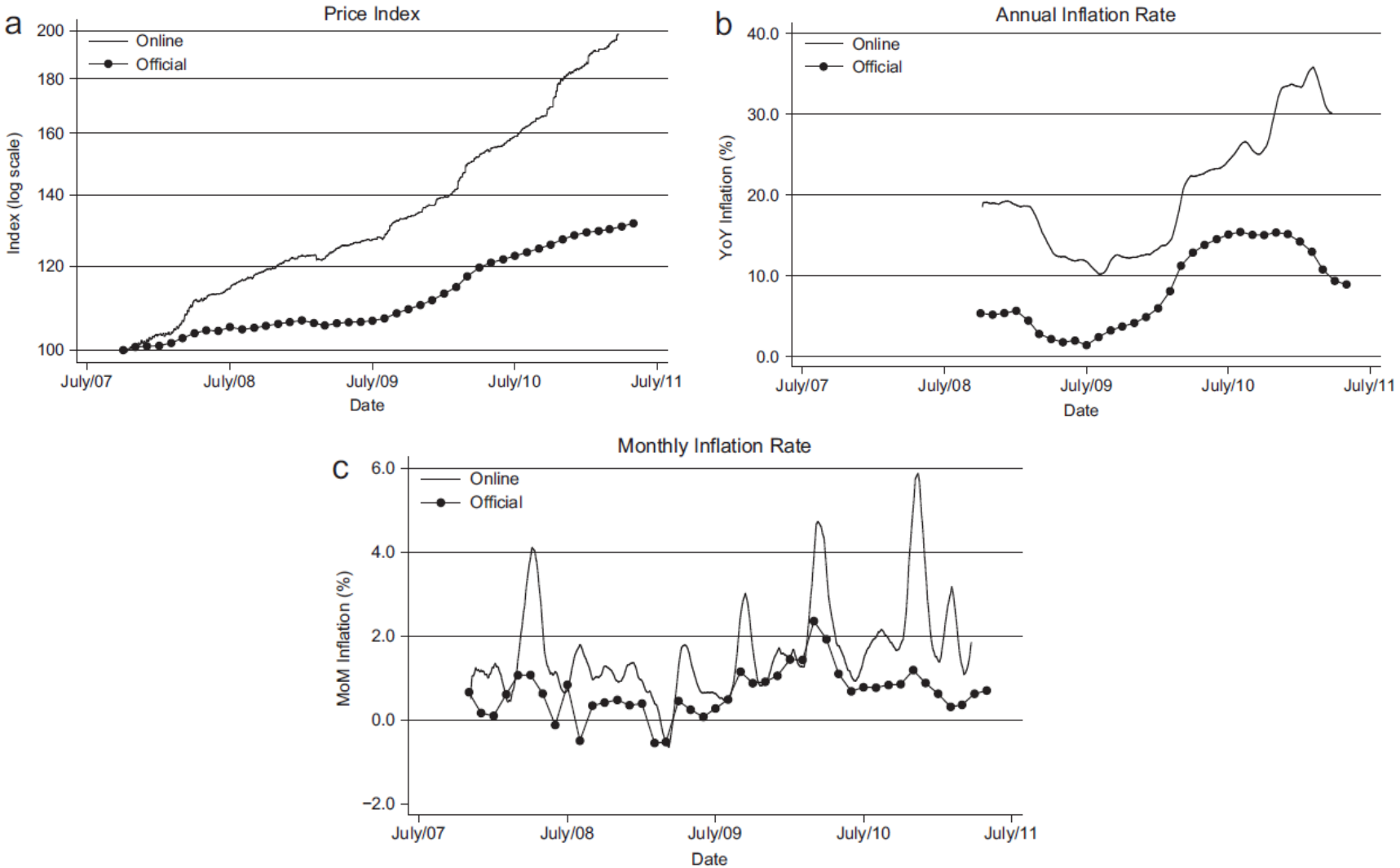


Fig. 3. Online supermarket index in Argentina. *Notes*: The monthly online inflation rate is a daily time series computed as the percentage change in the average of the index in the last 30 days with respect to the average of the index in the same period a month before. The monthly official inflation rate is a monthly time series computed as the percentage change in the index over the previous month.

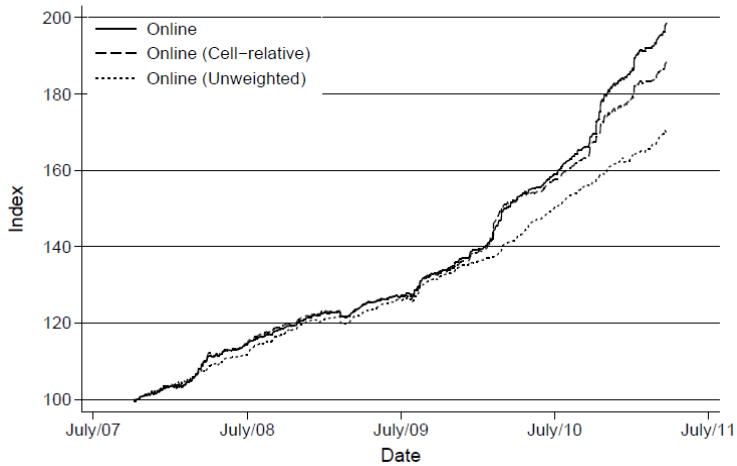
Robustness: Cell-Relative Imputation and Unweighted Index

- The first exercise uses a different approach to impute missing values within price series.
  - The method used in the paper is to fill missing prices with the last available price for each product. This approach is reasonable because the gaps in the online data last only a few days.
  - However, official statistical offices deal with missing values in their monthly series in different ways. The standard approach, also used in Argentina, is to impute missing prices with the average price change of similar products.

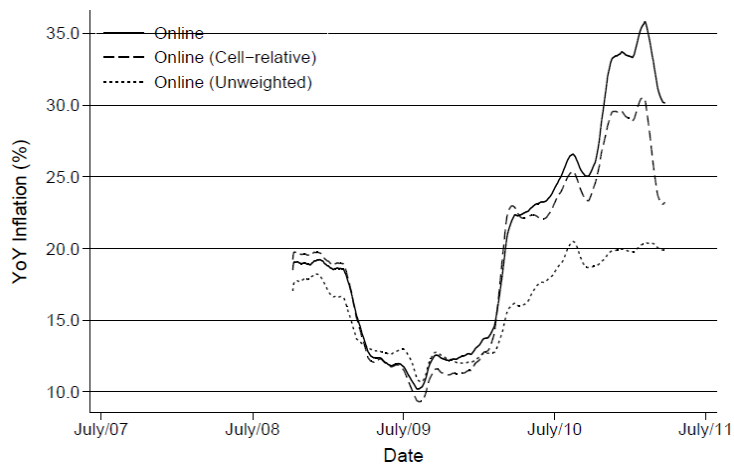
- Instead, we follow the "cell-relative" approach used by the Bureau of Labor Statistics (BLS): if a product is missing on a particular day, we do not use that product for the calculation of that day's inflation, but impute a price for it equal to the previous price times the average price change for products in the same category that day.

Robustness: Cell-Relative Imputation and Unweighted Index (cont'd)

(a) Daily Index



(b) Annual Inflation



(c) Monthly Inflation

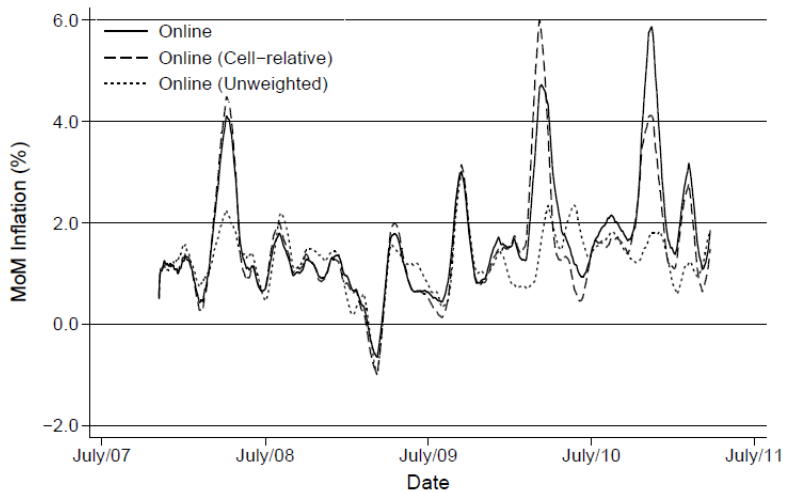


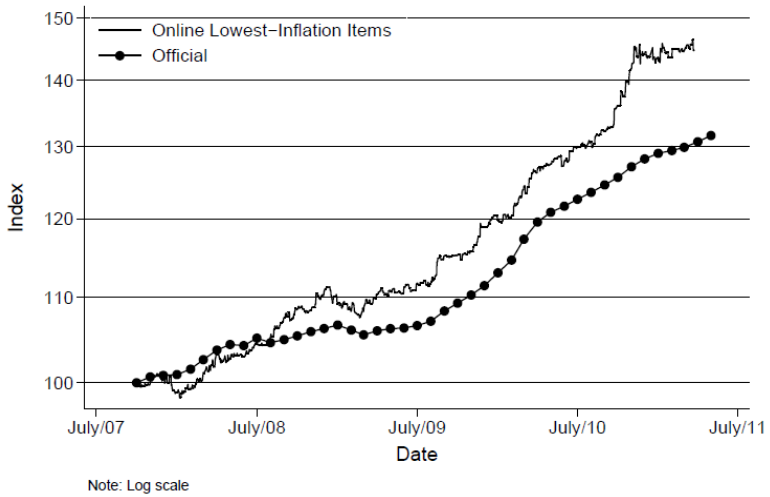
Figure B.1: Robustness: Cell-Relative Imputation

Robustness: Lowest-Inflation items

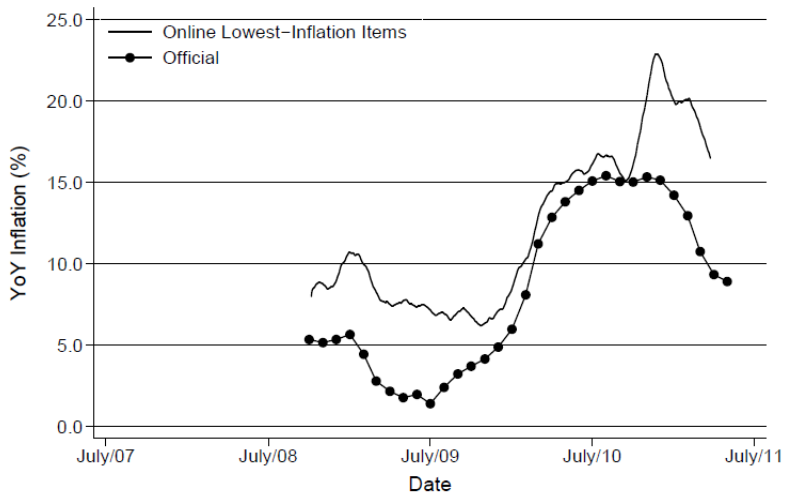
- This section uses an extreme assumption: that the government is able to select, within narrowly defined categories, only the goods that have **the lowest inflation rate** over this whole period.
    - This would not be a realistic alternative for the INDEC, given that it is hard to know ex-ante which goods will have the lowest inflation rates, but at least it can provide a lower-bound inflation rate for a strategy that favors lower-inflation brands or items.

Robustness: Lowest-Inflation items (cont'd)

(a) Daily Index



(b) Annual Inflation



(c) Monthly Inflation

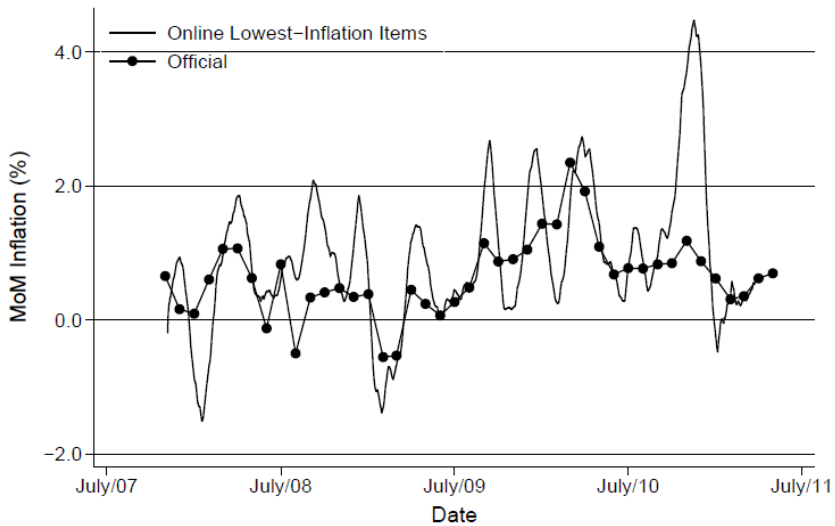


Figure B.2: Lowest-Inflation Items within URLs

Implications for Poverty

- The bias in inflation estimates also affects other statistics, such as **the poverty and real GDP estimates**.

- Every quarter, INDEC uses the cost of the CBA basket to see how many individuals are in extreme poverty conditions.Taking the cost of the official basket in the first quarter of 2008, and adjusting it with the CBA inflation rate observed online, the basket in July 2011 would have a cost of $259.5 Argentine pesos.

- Using INDEC's income survey, this implies that **6.69%** of the population was under extreme poverty at the time, compared to only 2.5% reported in official statistics.
  - Similarly, after adjusting the CBA to obtain the broader "Canasta Basica Total" (CBT), which adds non-food items to the basket, the level of poverty becomes **25.9%** compared to the 9.9% officially reported.

Implications for Real GDP

- To estimate the impact on real GDP, we start by looking at how the CPI and the GDP deflator have behaved in the past decade.

- The data from 1994 and 2006 show that both series were closely correlated, which is the expected behavior under normal conditions.
  - However, since 2007, the GDP deflator has increased significantly faster than the CPI. This means that government has recognized higher inflation in the GDP deflator.
  - An annual growth rate with an "Adjusted Real GDP" using the online index is **significantly lower** than in official estimates. (Note that the growth rate has been obtained by assuming the deflator had increased at the same rate as the online index.)

Implications for Real GDP (cont'd)

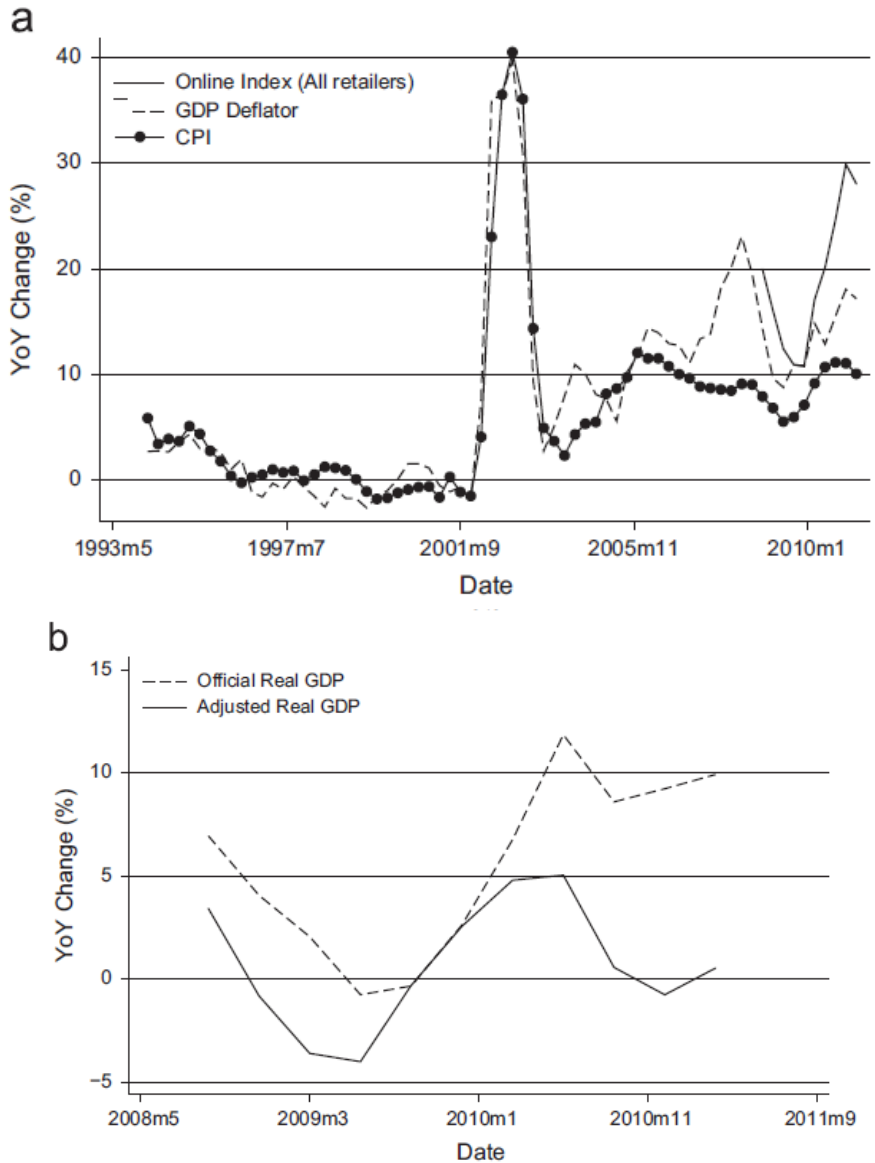


Fig. 7. Implications for real GDP growth: (a) GDP deflator, CPI, and online index – annual change and (b) Real GDP – annual growth rate. *Notes*: The GDP deflator and CPI series in (a) co-moved closely together from 1994 to 2006, but started to deviate from 2007 onwards. Although higher than the CPI, the GDP deflator still has less inflation than the online index in the past 3 years. Assuming the deflator had increased at the same rate as the online index, then we can compute an "Adjusted Real GDP" with a growth rate that is significantly lower than in official estimates.

Conclusion

- Online price indexes, constructed using a combination of online data and official methods, are capable of **matching both the level and main dynamics of official inflation** in Brazil, Chile, Colombia, and Venezuela.

- The matching is best at annual frequencies and improves when the data come from supermarkets with large market shares and cities that are **more representative of the country as a whole**.

- For Argentina, Indeed, online inflation has been consistently between **two and three times larger** than in official estimates for over 3 years.

- Online price indexes can provide a **good approximation to the real inflation rate** in the country.