Using Online Prices to Anticipate Official CPI Inflation

Manuel Bertoloto
PriceStats

Alberto Cavallo
MIT & NBER

Roberto Rigobon
MIT & NBER

June 2014

PRELIMINARY – DO NOT CITE

1 Introduction

The inflation rate a consumer faces should be, in principle, a relatively simple process to measure. Intuitively, a person has a typical consumption basket, and following the monthly average price of such basket should be the inflation rate the individual experiences. This is, however, hard to implement in practice. Complications abound, such as the fact that consumption baskets are rarely stable through time, consumer's substitute products when they face changes in relative prices, and many products are often discontinued and replaced with new version or even entirely new product categories. At the aggregate level, the inflation rate is even a harder process to characterize. Baskets and consumption behavior differ markedly across consumers, and traditional data collection methods are expensive and very limited in the quantity of goods and the frequency with which they can be sampled.

Economists have discussed for many decades what are the most appropriate methodologies to deal with these features. Many important improvements have been made the last decades to the way inflation is measured. Most of the improvements have been methodological. However, recent changes in communication and the rapid penetration of online stores presents the statistical offices with the possibility to improve the collection of the data as well.

Online prices are increasingly being used and considered for the construction of price indices and the measurement of inflation. Cavallo (2010) showed how online data could be used to build alternative inflation indices in countries where official data is unreliable, such as the case of Argentina. Our work at the Billion Prices Project proved that online prices could also be used in other countries, such as the US, to build indices that closely track official statistics. Since 2011, PriceStats, a private company, has been publishing daily inflation series in real-time in over 20

countries. National Statistical Offices (NSOs) have recently started to consider the use on online data in the construction of official CPI series, as discussed in Krsinich(2014) and De Haan, Griffioen and Willenbord (2014).

The objective of this paper is to study whether online price indexes provide information on inflation trends measured by the CPI. We use daily price indices constructed by PriceStats and compare them to equivalent CPIs in the US and the major Eurozone countries. We first explain how the data is collected and discuss some of the main characteristics, focusing on concerns about the representativeness of online prices. We then document the anticipation and information content of the online price indexes in three steps.

First, we estimate simple impulse responses and show how online prices tend to move earlier than offline prices. We find significant anticipation of online prices for several countries in the sample. The anticipation is not limited to the speed with which online prices can be collected and published. In fact, we find that online indices contain information that is gradually incorporated on official CPI data for several months.

Second, we estimate within sample regressions to evaluate the information content (measured as partial R-squares) after controlling for several macro variables – such as gas prices. We show that online price indexes can explain a significant proportion of the unexplained variance left over by other variables.

Lastly, we estimate out of sample forecasting of official CPI, and show that the online price indexes improve CPI inflation forecasts at various horizons. The results are particularly strong when sector-level series are used.

2 Data and Index Construction

The daily price indexes were constructed by PriceStats, a private company that collects online prices for a large number of retailers and uses them to build high-frequency price indices.

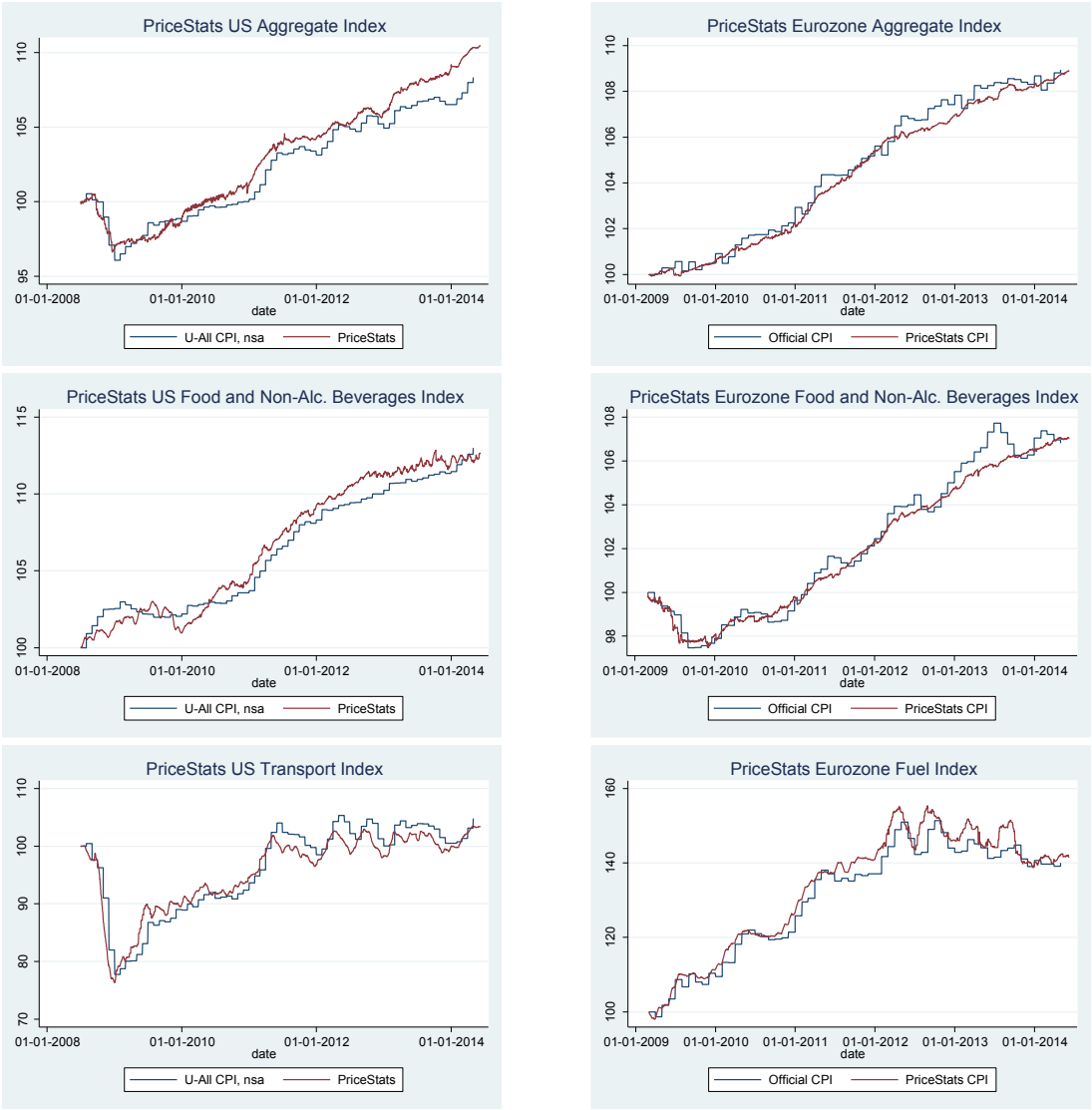
The technology that PriceStats uses to collect online prices is called “web scraping”. The procedure is conceptually simple. A piece of software identifies each product sold in the webpage and collects its price, product description, and any other relevant information. Repeating the process every day, produces a panel database with one record per product per day for a particular retailer. Process can be run for multiple retailers in the same day. After collecting the price information for a website, the products stored in the database are classified using the “classification of individual consumption according to purpose (COICOP)” provided by the UN.

PriceStats methodology to build the indexes follows standard CPI techniques as closely as possible, but there are significant differences in the collection and treatment of the data. First, since PriceStats collects data for all goods available for purchase at the chosen retailers and categories, the price contain information from thousands of different goods varieties, even at the lowest levels of aggregation. Second, there are no forced price substitutions, hedonic quality adjustments, or seasonal adjustments in the data. Third, despite the absence of online prices for some category of products (eg. services) the online series account for all the categories typically surveyed in the CPI through to a proprietary methodology that incorporates inflation estimates for each of these sub-sectors.

This paper uses a set of PriceStats inflation series for the US and the Eurozone area. In each case, we use the aggregate inflation series as well as the sector series for food and non-alcoholic beverages, and for transportation. The Eurozone series during the sample period are constructed with country-level series from France, Germany, Ireland, Italy, Netherlands, Spain, and Greece, weighted using relative household consumption expenditures. The data covers the period July 2008 to February 2014, with the exception of the PriceStats Eurozone series that start in April 2009.

Figure 1 plots each of these series next to the comparable (all items, not seasonally adjusted) official CPI in each of the countries or regions.

Figure 1: PriceStats Inflation Series and CPIs in the US and the Eurozone



Note: The PriceStats price series are computed on a daily basis and published every day with a three-day lag. In the US graphs, the blue line is CPI is the urban, all-items, non-seasonally adjusted CPI published on a monthly basis by the Bureau of Labor Statistics. The Eurozone series are constructed by computing a weighted arithmetic average of the country-level PriceStats series for Germany, France, Ireland, Italy, Netherlands, Spain, and Greece. As weights we use the relative level of household consumption expenditures in each country. The Eurozone CPI series shown in these graphs are constructed in the same way using official CPI series.

In additional to CPIs and PriceStats's inflation series, we also use publicly available data for fuel prices in each location. For the US, we use the Weekly Retail Gasoline and Diesel Prices (WRGDP) in dollars per gallon that are collected and released every Monday by the Energy Information Administration. We average four types of gas and one type of diesel prices and keep the last observation (the last Monday) of each month to calculate monthly variation.[1] For the Eurozone and the member countries we use the weekly consumer prices released by the Market Observatory for Energy of the European Commission. We average two prices, Euro-super 95 and Automotive Gas Oil prices, and keep the last available observation for each month. [2]

2.1 Are Online Prices Representative?

One important and common concern with online data is the representativeness of the prices observed online. The answer depends on the sector – and more importantly – on the behavior of each individual store sampled in the data.

PriceStats focuses exclusively on large offline retailers that also sell (or simply show prices) online. It does not use data from online-only retailers. Still, there may be concerns that the retailers included may have different prices online and offline, so the concerns about representativeness of the online data remains. In particular, there are two important aspects to consider for each retailer: the share of transactions that take place in the online store, and the differences in pricing behaviors across stores.

Some stores such as Apple, for instance, are actually online stores that incidentally have offline stores. In the US, Apple sells more than half through the online store. And more importantly, the prices online are identical to the prices offline. So, collecting the information of one is not only representative but identical to the census of the offline stores. Another example would be Walmart. It has identical prices across the US for all non-perishable products in their web pages. Again, there is no discrepancy between the online and offline stores at all. However, Walmart has products that are locally sold, whose prices are indeed different across stores, and the products cannot be found elsewhere. Hence, online prices from Walmart have different degrees of representativeness depending on the pricing practices it follows.

The second question is about how relevant are the online stores when a small share of the revenues occur online. Walmart, again, is a very good example. In the US it

[1] The fuel prices we use are the diesel fuel price, and three categories of gas: midgrade all formulations, premium all formulations and regular all formulations. In the choice we follow Modugno (2011).

[2] In Modugno (2011) Heating Oil is also included in the average. We chose to exclude it since the fuel series for the USA includes only fuels for transportation purposes.

sells approximately 8 percent online. How should we think about that share? On one hand, that would appear to be a small share of the total. On the other hand, compared to any individual store, the 8 percent is in fact a huge share of sales. The answer depends on whether our view about price dispersion across stores.

If there is significant price discrimination in offline pricing, it means that each store is different because they are highly segmented. That implies that the online business should be compared to one single store. In other words, if every store is different their representativeness depends on their share. In the case of Walmart, it has 4759 stores in the US. The median store sells a little less than 0.02 percent. Which means the online store is 400 times bigger than the typical store most people frequent. So, even though the prices are not representative the online store is huge relative to the price dispersion in the offline business.

Alternatively, if the price dispersion is small, meaning that all the offline stores have identical prices for almost all their products, then the online store should be compared to the total offline business. Interestingly, retailers that exhibit this behavior tend to have offline and online prices identical – meaning that collecting the online price is quite representative of the retailer. Burger King, for example, sells nothing online. It has more than 45 thousand establishments in the US. Each one probably has a tiny proportion of the revenues of Burger King. However, the menus are online, and not surprisingly, the prices are identical across more than 97 percent of them. Hence, collecting the price of the most minuscule of the stores is extremely representative of the whole. Of course, not all stores behave like this. Gap is an example of a store where the online and offline prices differ dramatically; not only in the posted prices but also in the discounts and sales. And even though the online store is orders of magnitude larger than any individual store (like the case from Walmart) it is impossible to argue that those prices are representative. There are sectors where the online business is large and stores tend to align online and offline prices. This makes the online data a very good quality data point to measure inflation for that retailer.

Dealing with the problem of representativeness is quite important, but it is not the only issue; sales, seasonality, and marketing in general can generate a gap between the online and offline prices, especially in the short run. As we discuss later these gaps imply that indexes based on online prices have quite surprising properties – one being anticipation – which can be used to construct better forecasts of the official CPI.

3 Information Content in Online Price Indexes

We evaluate the information content of online prices with three different methodologies. First, we compute simple VAR's where only the official CPI and the Online index are included in the regression. We estimate impulse responses and

evaluate their significance. This very simple procedure shows the anticipation in online prices as well as its significance. Second, because the VAR does not control for other important variables that could be measured in the economy, the second exercise is a within sample R-square estimation. The idea is to include variables that should be correlated with the official CPI and determine the explanatory power of the online indexes. In this subsection the purpose is to give to other variables – such as gas prices – the highest chance to explain the data, and only afterwards evaluate the explanatory power of the online indexes. Finally, we perform an out-of-sample forecasting exercise to evaluate the importance of online prices.

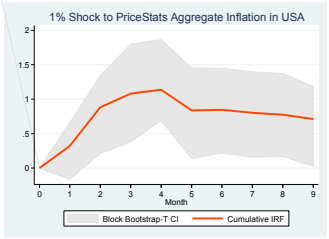
The results in this section are as follows: First, we find significant anticipation of online prices for several countries in the sample. Second, the online price indexes explain a significant proportion of the unexplained variance after controlling for several macro variables. Lastly, online price indexes significantly improve out-of-sample forecasting of official inflation, particularly when sector-level series are used.

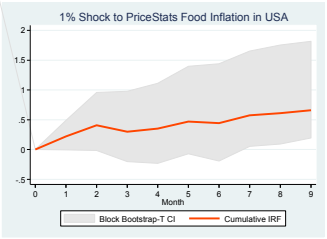
3.1 Impulse Responses

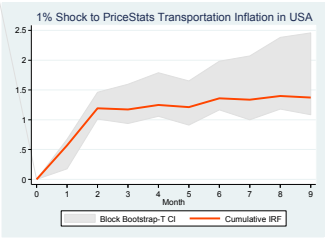
The first exercise we perform is to compute simple impulse responses. We estimate a VAR using the official CPI and the online price index (OPI) where we assume that the OPI is the exogenous variable. This is of course giving the OPI the highest chance to explain the observed variation. However, there is no clear way of identifying the system given that, under the null hypothesis, both are valid measures of the underlying inflation. The regression includes 6 lags of official monthly CPI inflation and 6 lags of the OPI monthly inflation, plus the contemporaneous value of OPI inflation. By including the "current" information of the OPI this specification reflects the earlier availability with respect to the time release of official inflation. For each month *t*, the specification is as follows:

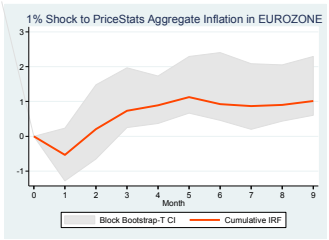
$$CPI_t = a + OPI_t + \sum_{i=1}^{6} OPI_{t-i} + \sum_{i=1}^{6} CPI_{t-i} + u_t$$

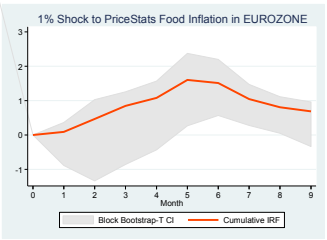
The confidence bands are computed by bootstrapping in blocks. We present the 95 % percentile-t confidence intervals.

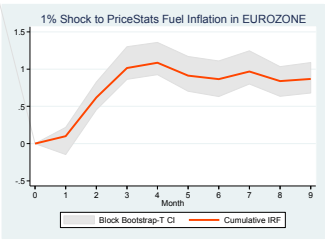

1% Shock to PriceStats Aggregate Inflation in USA
Month
Block Bootstrap-T CI
Cumulative IRF

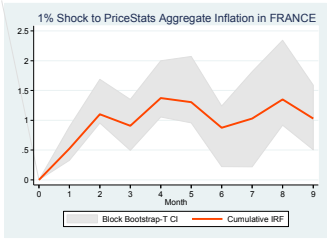

1% Shock to PriceStats Food Inflation in USA
Month
Block Bootstrap-T CI
Cumulative IRF

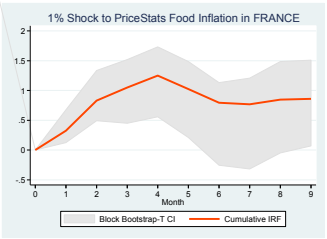

1% Shock to PriceStats Transportation Inflation in USA
Month
Block Bootstrap-T CI
Cumulative IRF

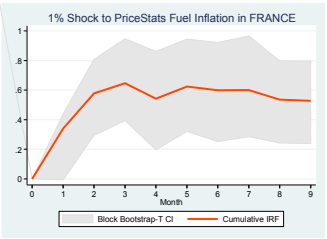

1% Shock to PriceStats Aggregate Inflation in EUROZONE
Month
Block Bootstrap-T CI
Cumulative IRF

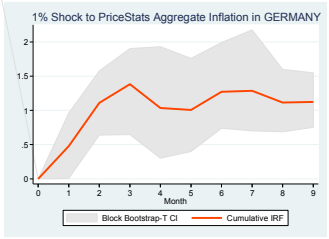

1% Shock to PriceStats Food Inflation in EUROZONE
Month
Block Bootstrap-T CI
Cumulative IRF

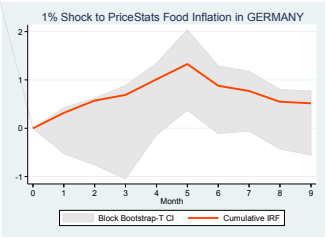

1% Shock to PriceStats Fuel Inflation in EUROZONE
Month
Block Bootstrap-T CI
Cumulative IRF

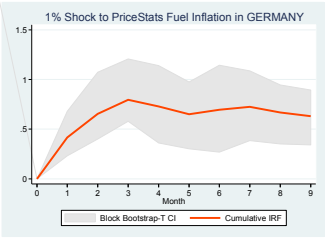

1% Shock to PriceStats Aggregate Inflation in FRANCE
Month
Block Bootstrap-T CI
Cumulative IRF

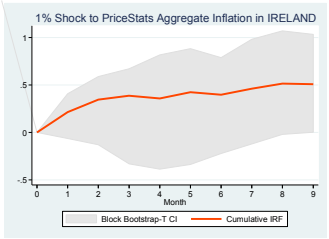

1% Shock to PriceStats Food Inflation in FRANCE
Month
Block Bootstrap-T CI
Cumulative IRF

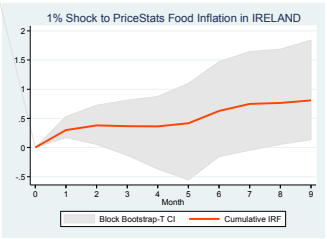

1% Shock to PriceStats Fuel Inflation in FRANCE
Month
Block Bootstrap-T CI
Cumulative IRF

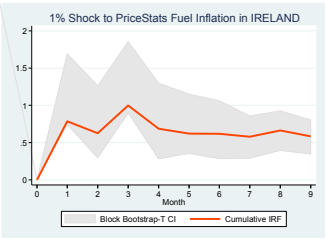

1% Shock to PriceStats Aggregate Inflation in GERMANY
Month
Block Bootstrap-T CI
Cumulative IRF


1% Shock to PriceStats Food Inflation in GERMANY
Month
Block Bootstrap-T CI
Cumulative IRF


1% Shock to PriceStats Fuel Inflation in GERMANY
Month
Block Bootstrap-T CI
Cumulative IRF


1% Shock to PriceStats Aggregate Inflation in IRELAND
Month
Block Bootstrap-T CI
Cumulative IRF


1% Shock to PriceStats Food Inflation in IRELAND
Month
Block Bootstrap-T CI
Cumulative IRF


1% Shock to PriceStats Fuel Inflation in IRELAND
Month
Block Bootstrap-T CI
Cumulative IRF

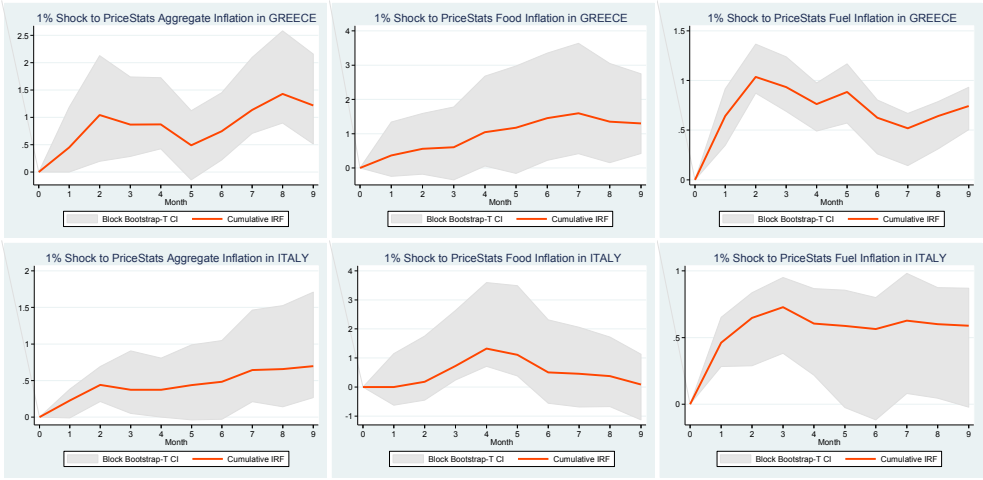


Table 1. Magnitude, Timing and Significance of Peak Response

	USA	Eurozone	France	Germany	Ireland	Greece	Italy
PS Aggregate	1.14*	1.13*	1.37*	1.38*	0.52	1.43*	0.44
	month 4	month 5	month 4	month 3	month 8	month 8	month 2
PS Food	0.41	1.60*	1.25*	1.33*	0.81	1.60*	1.32*
	month 2	month 5	month 4	month 5	month 9	month 7	month 4
PS Fuel or Transport	1.19*	1.09*	0.65*	0.80*	1.00*	1.04*	0.73*
	month 2	month 4	month 3	month 3	month 3	month 2	month 3

* significant at 5% level

As can be seen from these results not only the OPI is statistically significant, but more importantly, there is anticipation in the sense that the shocks to the OPI are incorporated slowly into the CPI. This result holds even when the contemporaneous effect of the OPI is dropped from the regression. In other words, in the two extremes of the identification space there is anticipation.

We do not know exactly why the anticipation occurs. Our conjecture is that most of the behavior derives from the fact that online consumers are less sensitive to price changes than offline customers. Two reasons are behind this conjecture, one is the lack of customer anger when individuals have no memory about the price history, and the second one is the mixture of stores that exist in offline versus online.

Online customers search quite extensively at the moment of the purchase. In fact, they tend to assume that because the found at that instance the lowest price means that they are purchasing at the best possible price. That actually is incorrect in general. Prices change significantly through time but online customers can rarely see such behavior. In fact, if one has ever purchased anything online try to answer the following two questions: First, can you remember what you purchased recently and its price? Some might be able to answer that question. Then the second question

is to ask if the person knows how much the exact same item cost a month before the purchase. This second question will be answered usually in two instances: if the product is very expensive (like a DSLR camera) or if the item is something the person purchases repeatedly (like iTunes songs). In fact, the very expensive items on the web exhibit significant price stickiness. This is exactly the case, because if they change their prices repeatedly customers might get upset. The second item, on the other hand, it is not that the person knows that the song is purchasing today was at 1.29 a month ago. Not at all! The customer assumes it was at 1.29 because "all songs" are at 1.29. Actually, those individuals might be surprised. Especially with new songs, sometimes they are offered free in iTunes (two songs a week, for those interested). And if the band is new the songs could be at a discount (0.69 for those also interested). So, the customer does not get upset not because it has evidence there was no price discrimination, mostly because he/she assumes there is none.

In the end what happens is that only in very few items online customers have memory of the price history. People already have small memory on prices offline, they have even less in online purchases. Even though most customers do not realize they are memory-less, stores know it. And online prices tend to shift faster than offline prices – especially for essential products like bread and nutella in France, or pasta in Italy, or beer in the US.

In the downturn (price declines) online business is so competitive that they also tend to go down much faster than offline prices.

The second reason for anticipation is the mixture of store. Offline prices are collected to be representative, so in countries in which there is a significant proportion of citizens living in poverty stores that attend the poor – which usually are underrepresented in online stores – are included in the official data. Low income customers are quite price sensitive. So, if there is going to be inflation, those stores tend to increase prices slowly, adjusting margins in the short run. Furthermore, because in the upswing of inflation prices are sluggish, then stores tend to be equally slow when there is price deflation. Because the official statistic collects from this type of stores, in the end the official inflation rate moves slower than the online prices.

Notice that the reason of the anticipation is not due to the speed at which the data is collected, nor to the outcome of an identification assumption, but it is coming from the store's behavior and the mixture of stores collected in the CPI and the OPI.

3.2 Within Sample Information

The second exercise is to compare the explanatory power of the OPI when other variables are included. The idea of the exercise is to incrementally include variables into the regression and observe the evolution of the R-square from. This is an identical transformation of the Ftest for parameter significance, but instead of

concentrating on the joint significance of the coefficients this allows us to discuss what share of the variance is explained by certain variable – it is equivalent to a variance decomposition when the variables are correlated.

The idea is to estimate

$$CPI_t = \alpha + \sum_{i=1}^{n} CPI_{t-i} + \varepsilon_{cpi,t} \ (1)$$

This first regression gives the CPI lags the best chance to explain itself. We then take the residuals from this regression and regress them against monthly gas prices (for example)

$$\hat{\varepsilon}_{cpi,t} = \alpha + \sum_{i=0}^{n} P_{t-i}^{oil} + \varepsilon_{cpi,oil,t} \ (2)$$

In this case, the R-square of the second regression indicates what share of the unexplained variance from the CPI on lags is explained by oil prices. We then estimate a further regression taking the residuals from the second regression and using them as dependent variables with the OPI as regressors:

$$\hat{\varepsilon}_{cpi,oil,t} = \alpha + \sum_{i=0}^{n} OPI_{t-i} + \varepsilon_{cpi,oil,opi,t} \ (3)$$

This procedure measures the explanatory power of OPI after we have given the CPI and Gas prices the chance to explain most of the variation. For instance, assume that the OPI were a perfect linear combination of CPI and Gas prices, then after the first two regressions are run, there is no variation that OPI could explain. The OPI will only have explanatory power if it provides information beyond the variables previously included.

Furthermore, we also run a special case of a horse race between Gas prices and the OPI. The reason why we perform this race is because gas prices is the only price that the market observes daily and in the short run the OPI and Gas prices are indeed highly correlated. We run the following two regressions

$$\hat{\varepsilon}_{cpi,t} = \alpha + \sum_{i=0}^{n} P_{t-i}^{oil} + \varepsilon_{cpi,oil,t} \ (2)$$
$$\hat{\varepsilon}_{cpi,t} = \alpha + \sum_{i=0}^{n} OPI_{t-i} + \varepsilon_{cpi,opi,t} \ (4)$$

The horse race here is to compare the explanatory power of oil prices versus the online price index. The reason why we have chosen to follow this methodology is because inflation rates are highly multicolinear. The null hypothesis in all these tests is that the OIP explains nothing – meaning that the other macro variables included in the regression are sufficient statistics for the online prices. The alternative, of course, is that online prices have information content. By partialing out the macro variables we are given them the highest chance to explain the official inflation – hence, the explanatory power we derive from online prices is actually the lower bound of their information content.

Table 2 presents the results of these tests using the official inflation in the USA. For the OPI we use the PriceStats Aggregate and Transport series together; that is *n* lags of the aggregate online prices index plus *n* lags of the transport sector online prices index. We set the number of lags at *n=9.*

Table 2. Within-sample regressions for the USA

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	38.92%	0.26%
(2)	$\hat{\varepsilon}_{cpi,t}$	Gas prices	60.60%	0.17%
(4)	$\hat{\varepsilon}_{cpi,t}$	PS Aggregate and PS Transport	71.60%	0.16%
(3)	$\hat{\varepsilon}_{cpi,oil,t}$	PS Aggregate and PS Transport	22.33%	0.17%

Table 3 repeats the analysis but for the Eurozone official CPI, and instead of the PriceStats Transport series we use the Fuel series. Then tables 4 to 8 do the same for France, Germany, Ireland, Greece and Italy.

Table 3. Within-sample regressions for the Eurozone

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	22.79%	0.13%
(2)	$\hat{\varepsilon}_{cpi,t}$	Gas prices	48.47%	0.09%
(4)	$\hat{\varepsilon}_{cpi,t}$	PS Aggregate and PS Fuel	55.12%	0.10%
(3)	$\hat{\varepsilon}_{cpi,oil,t}$	PS Aggregate and PS Fuel	33.45%	0.09%

Table 4. Within-sample regressions for France

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	19.90%	0.14%
(2)		Gas prices	25.66%	0.12%
(4)		PS Aggregate and PS Fuel	49.78%	0.11%
(3)		PS Aggregate and PS Fuel	49.49%	0.10%

Table 5. Within-sample regressions for Germany

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	15.13%	0.18%
(2)		Gas prices	39.74%	0.14%
(4)		PS Aggregate and PS Fuel	51.82%	0.14%
(3)		PS Aggregate and PS Fuel	51.89%	0.11%

Table 6. Within-sample regressions for Ireland

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	17.69%	0.22%
(2)		Gas prices	15.49%	0.21%
(4)		PS Aggregate and PS Fuel	40.53%	0.20%
(3)		PS Aggregate and PS Fuel	32.01%	0.20%

Table 7. Within-sample regressions for Greece

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	19.16%	0.36%
(2)		Gas prices	46.96%	0.27%
(4)		PS Aggregate and PS Fuel	57.52%	0.31%
(3)		PS Aggregate and PS Fuel	49.69%	0.24%

Table 8. Within-sample regressions for Italy

Equation number	Dependent variable	Regressors	R2	RMSE
(1)	Official CPI	Official CPI	31.67%	0.15%
(2)		Gas prices	29.31%	0.12%
(4)		PS Aggregate and PS Fuel	56.99%	0.12%
(3)		PS Aggregate and PS Fuel	47.42%	0.11%

3.3 Out of Sample Information

The final exercise is to estimate the out-of-sample performance of a simple forecasting model. The purpose is to evaluate how much does the OPI helps reduce the forecasting errors.

The data used is the following:

- Variable to be predicted: CPI-all NSA
- Weekly Retail Gasoline and Diesel Prices (WRGDP) in dollars per gallon are collected and released every Monday by the Energy Information Administration. We average four types of gas and one type of diesel prices and keep the last observation (the last Monday) of each month to calculate monthly variation. This is publicly available information.
- PriceStats US Daily Aggregate Inflation Index (PS Aggregate). We keep the index value for the last day of each month.
- Two PriceStats Daily Sector Series: Food and Non Alcoholic Beverages (PS100), Transport (PS 700). We keep the index value for the last day of each month.

Our inflation figure for PS series for any given month will be the variation between index´s value at the last day of the month and the value at the last day of the previous month. For the publicly available gas prices the logic is the same, but the variation is between the average prices the last Monday of the month and the average prices at the last Monday of the previous month.

The series for all variables used are then seasonally adjusted. We compute the logarithmic difference of the CPI index series and then we regress this $\Delta\text{logCPI}_{t,t-1}$ on month dummies. Next we keep the computed residual $\hat{\varepsilon} = \Delta\text{logCPI}_{t,t-1} - \widehat{\Delta\text{logCPI}_{t,t-1}}$ as our seasonally adjusted time series. We do the same using $\Delta\text{logPSAll}_{t,t-1}$, $\Delta\text{logPS700}_{t,t-1}$, $\Delta\text{logPS100}_{t,t-1}$ and $\Delta\text{logGAS}_{t,t-1}$ as dependent variable.

We estimate rolling regressions using 24 months back starting at January 2014. The rolling regressions are used for the forecasting out of sample.

For example, Let´s assume we are forecasting January 2014 inflation.

First we choose an estimation window $W \in \{24, 36, 48, 60\ months\}$ (say for example W=24) and then a number of lags $l \in \{1, 2, 3\}$. Once we have that, we can generate the forecast:

1) We estimate the following model by OLS for the W months before the month we are forecasting (and excluding that month). In our example the regression will run over data from December 2013 and the 23 preceding months. [3]

[3] In this model we are "living" at the last day of month t.

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

HFD (for high frequency data) in the previous notation can be Gas prices or any of the PS series. We will make predictions using more than one PS series, for example PS All and PS 700 (transport). In that case each the terms including HFD in the previous notation will be repeated for PS All and PS 700.

2) We compute the fitted value for $\Delta\log\widehat{\text{CPI}}_{t,t-1}$ in February 2014 using the coefficients from the previous regression and now including the high frequency data from January 2014. That fitted value is our forecasted inflation.

By doing this for all values of W and of *l* we will have three forecasts with different number of lags each for each of the 4 windows of estimation W. That is 12 forecasts for January 2014 inflation. However in the first stages of the project we observed that averaging forecasts tends to improve accuracy.[4] Then for many of the presentations of the results the forecasts with up to 1, 2 or 3 lags are averaged into one.

To perform forecasting exercises with longer horizons than the forecast we "move back" in time the lagged regressors as well as the information used to fit the models. The aim is to simulate a situation when we are standing farther apart in time from the month we want to forecast.

We perform forecasts for bi-monthly inflation, quarterly inflation, 4 months inflation, 5 months inflation and 6 months inflation. We can focus first on the 2 months inflation.

Now the forecast for January 2014 will be for the bi-monthly inflation from the end of November 2013 to the end of January 2014. We will only use the information available at the end of December 2013.

1) We run the regression for the W months before the two months that compound our bi-monthly inflation (and excluding those months). In our example we run the regression with data up to (and including) November 2013. The model is:[5]

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

[4] See Bates & Granger (1969), "The Combination of Forecasts". Also Hendry & Clemens (2004), "Pooling of forecasts".

[5] In this model we are "living" at the last day of month t-1.

If we fix the number of lags to *l=1*, it looks like

$$\Delta\text{logCPI}_{t,t-2} = c_0 + \alpha_0\Delta\text{logHFD}_{t-1,t-2} + \alpha_1\Delta\text{logHFD}_{t-2,t-3} + \beta\Delta\text{logCPI}_{t-2,t-3}$$

2) We then fit the model with the information up to the first month that compounds our bi-monthly inflation, and compute the predicted dependent variable that is our forecast. In our example we fit the model with data up to (and including) December 2013.

As we did for the previous case, we repeat the estimation for all values of W and of *l* we will have three forecasts with different number of lags each for each of the 4 windows of estimation W. That is 12 forecasts for December 2013 - January 2014 inflation. We also pool forecasts across different lags into one for many of the reported results.

For quarterly inflation forecasts the logic is the same. The forecast will be for the inflation between October 2013 and January 2014. We will now use only the information available at the end of November 2013.

1) We run the following regression for the previous W months excluding the three months that compound our quarterly inflation.[6]

$$\Delta\text{logCPI}_{t,t-3} = c_0 + \alpha_0\Delta\text{logHFD}_{t-2,t-3} + \sum_{i=3}^{l+2}\alpha_i\Delta\text{logHFD}_{t-i,t-i-1} + \sum_{i=3}^{l+2}\beta_i\Delta\text{logCPI}_{t-i,t-i-1}$$

Which looks like this if we fix the number of lags to *l=1*:

$$\Delta\text{logCPI}_{t,t-3} = c_0 + \alpha_0\Delta\text{logHFD}_{t-2,t-3} + \alpha_1\Delta\text{logHFD}_{t-3,t-4} + \beta\Delta\text{logCPI}_{t-3,t-4}$$

2) We fit the model using information up to the first month that compounds the quarterly inflation we are forecasting, in the example November 2013.

The procedure described for 2 month and quarterly inflation is then generalized for forecasts of 4, 5 or 6 months inflation.[7]

The best way to present the results of the out-of-sample forecasting is to show the cumulative distributions of average deviations. The purpose is not only to show that the average deviation decreases but also to pay attention to the "extremes" of the distribution.

We present results only for the US. (Results for the EuroZone will be ready soon)

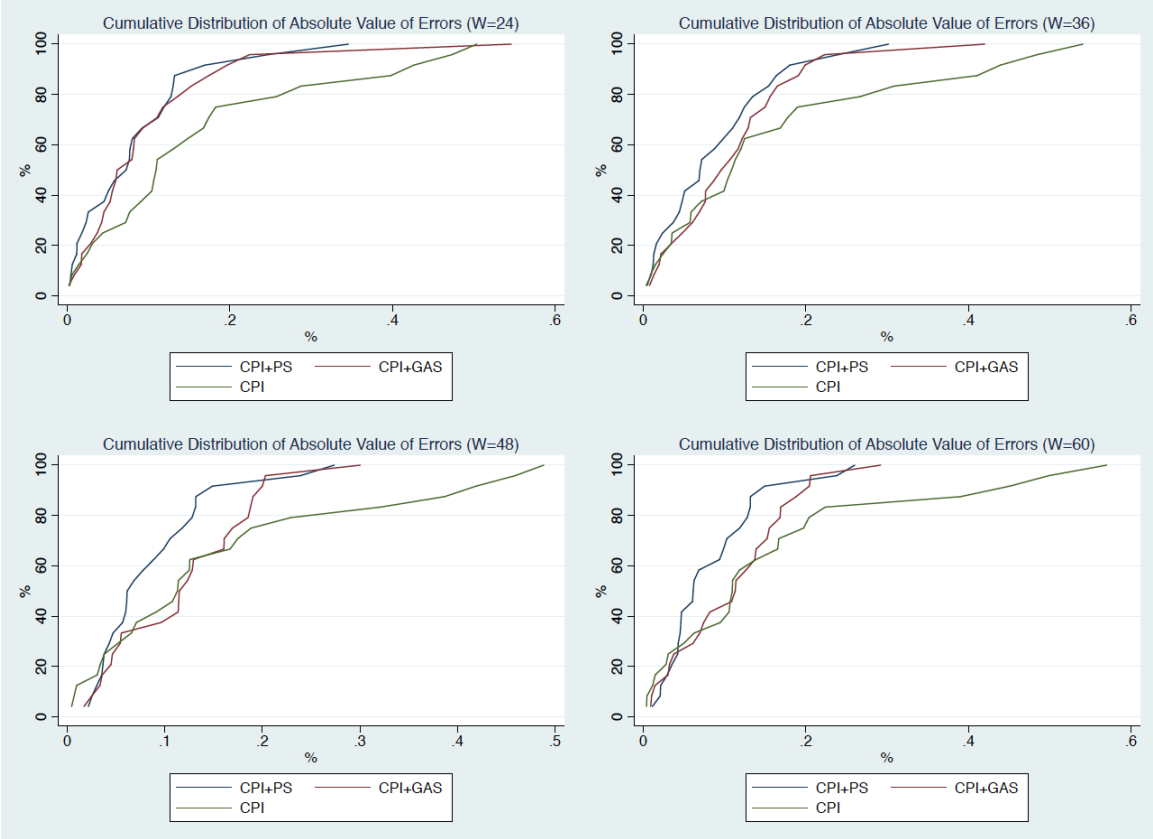
---

[6] In this model we are "living" at the last day of month t-2.

[7] For the 6-month inflation forecast the model will be $\Delta\text{logCPI}_{t,t-6} = c_0 + \alpha_0\Delta\text{logHFD}_{t-5,t-6} + \sum_{i=6}^{l+5}\alpha_i\Delta\text{logHFD}_{t-i,t-i-1} + \sum_{i=6}^{l+5}\beta_i\Delta\text{logCPI}_{t-i,t-i-1}$. Again, if *l=1*: $\Delta\text{logCPI}_{t,t-6} = c_0 + \alpha_0\Delta\text{logHFD}_{t-5,t-6} + \alpha_1\Delta\text{logHFD}_{t-6,t-7} + \beta\Delta\text{logCPI}_{t-6,t-7}$

For example, for the case of the US forecasting the contemporaneous CPI the absolute out-of-sample errors are distributed as follows (for the different estimation windows).

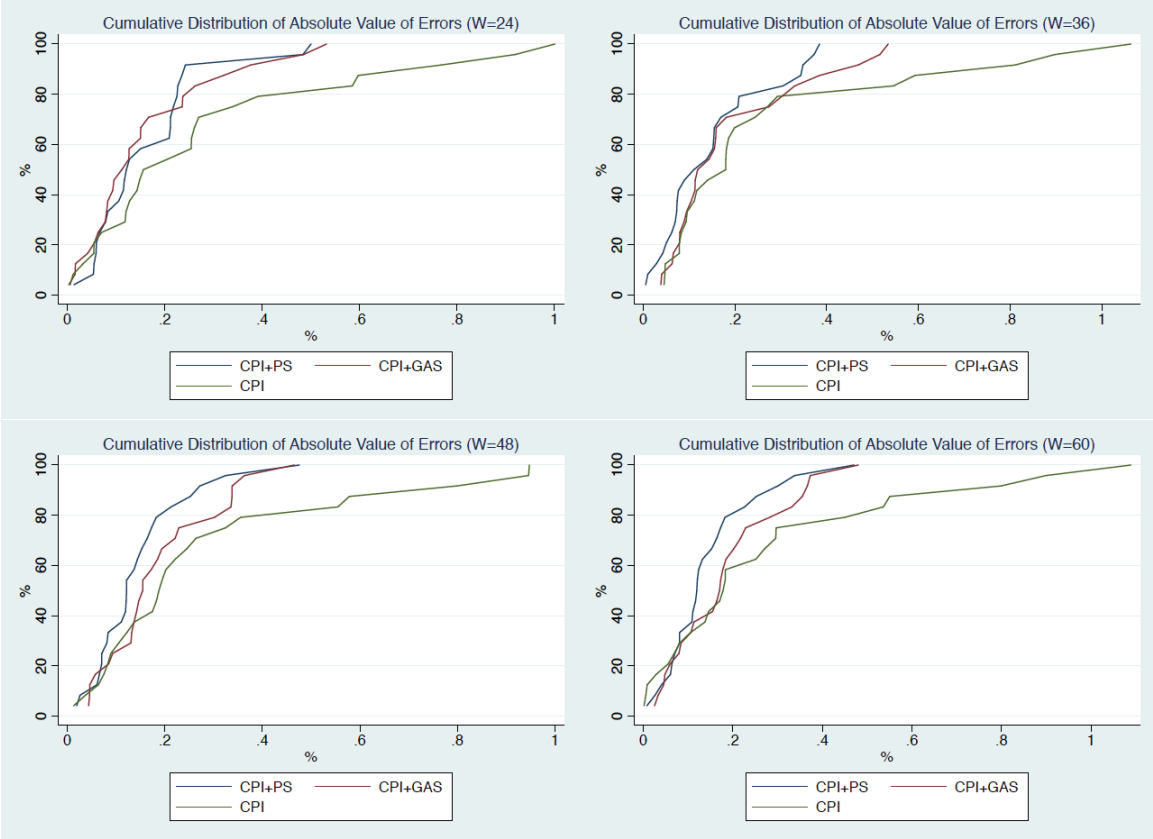
We present three regressions. The green is the CPI being predicted by lag CPI realizations and a constant. What we denote as CPI+GAS show the out-of-sample errors of estimating using not only the CPI but also Gas prices. Finally, the CPI+PS are the estimates including CPI and the OPI from pricesats



Notice that the Green lines consistently experiences large absolute errors. In fact, it is dominated by both the red and blue lines. So, including either gas prices or OPI improves the out-of-sample error. Furthermore, the errors from the PS regressions is significantly better than just using Gas prices. It is almost stochastically dominated!

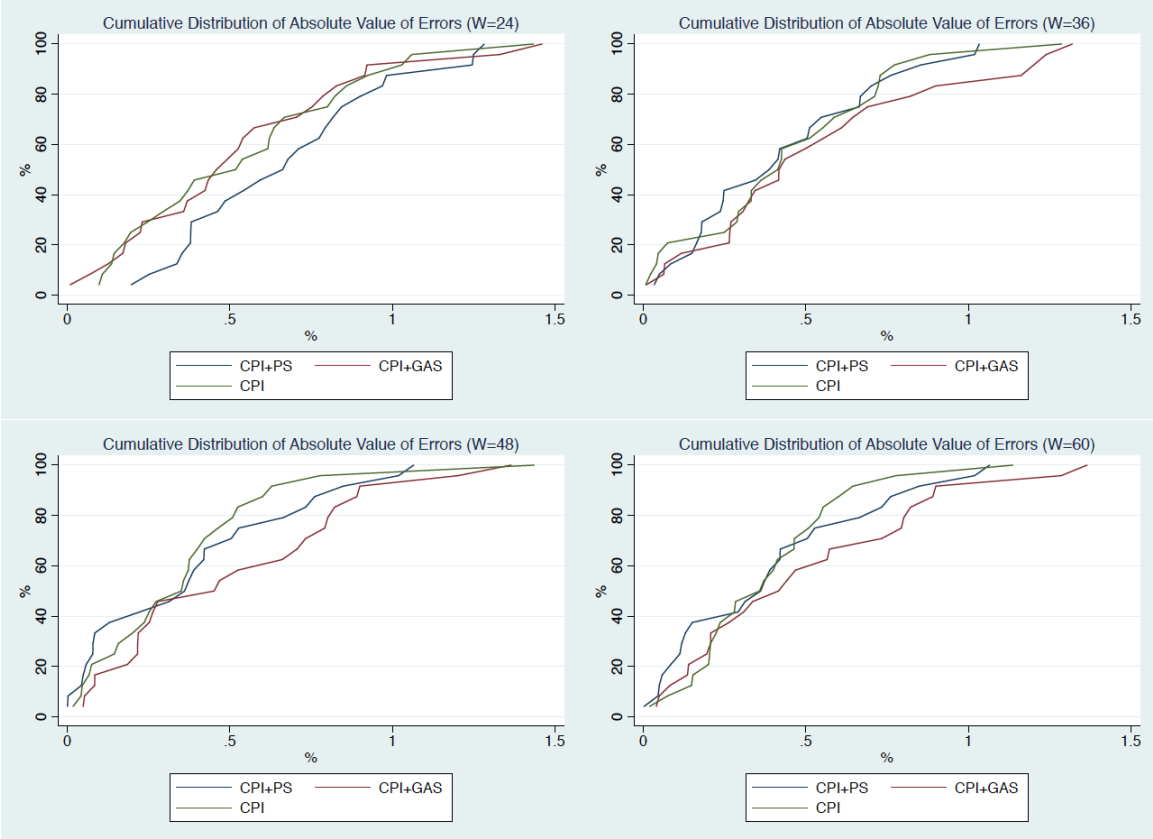
This is also the case for longer horizons. The errors for two months ahead are:

Two months ahead forecast errors



However, as should be expected, the advantage of OPI should disappear through time. When we estimate 6 months ahead the forecast errors are

Forecast Errors 6 months ahead



4 Discussion

Online Price Indexes are highly correlated with official CPI. Not only they are congruent in the long run, but in the short run the OPI's provide information about the CPI trends. This is measured by significant impulse responses. Moreover, because the peaks in the impulse responses occur several month later this means that inflation online tends to anticipate or occur before ofline inflation. Finally we prove that the OPIs provide information both within a sample and out-of-sample.

5 References

Cavallo (2010), “Online vs Official Price Indexes: Measuring Argentina’s Inflation”, Journal of Monetary Economics. December 2012.

De Haan, Griffioen and Willenbord (2014), “Collecting clothing data from the Internet”, Statistics Netherlands, Working Paper.

Krsinich (2014), “The FEWS index: Fixed effects with a window splice: Non-revisable quality-adjusted price indexes with no characteristic information”, Statistics New Zealand, Working Paper.