First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3133

Online CASE CPI

Radzikowski, Bartosz[a] and Śmietanka, Adam[a]
[a]CASE - Center for Social and Economic Research, Poland

Abstract

Online CASE CPI is an example of using Big data in public statistics. In principle, it is a consumer price index based entirely on online prices: a combination of Central Statistical Office of Poland's methodology and online data sets. An innovative method of data collection – data scrapping – allowed us to substantially reduce a time delay between data collection and a publication of results. A short, nine-month period of data collection has not given rise to make important conclusions, hence the aims of this paper are: to discuss a general framework of measuring consumer inflation online, to present preliminary results for Poland and to highlight the strengths and weaknesses of this approach. Finally, we believe that online consumer price indices have a complementary nature to conventional inflation measurement, but it might be a serious alternative, having in mind a huge growth potential of e-commerce in coming years.

Keywords: *Big data in public statistics; scrapping data; inflation measurement; e-commerce; online vs offline consumption*

1. Introduction

All actors in the economy are interested in accurate and timely information on changes in price levels. A rate of inflation influences households' decision whether to consume or save, enables economists to predict an economy's position in the business cycle, and determines the level of central banks' interest rates.

National statistical offices employ pollsters to visit thousands of retail outlets, restaurants and service units to collect price data across the country. Since the process is time-consuming, final information about prices, e.g. the consumer price index (CPI), is announced with some delay. Moreover, it is hard to establish a precise day for data collection because it lasts a whole month. For instance, Central Statistical Office of Poland (GUS) publishes its CPI index two weeks after the end of the month, meaning the data for a certain month is only available, in the best case, two weeks after the fact. Indeed, having in mind the fact that the data collection lasts almost the whole month, the real delay (between a day of collection and announcement of the results) might be longer than two weeks.

The research embodied in this paper has been performed in order to reduce the time delay and to measure prices volatility online. We developed a methodology to collect the data on prices in real time, along the lines of the innovative project elaborated by Massachusetts Institute of Technology's academics in *The Billion Price Project*. Integrating a model of household consumption used in the CPI index published by GUS, we have created an *online CASE CPI* for Poland that is calculated entirely on the price data from the Internet. What is worth mentioning, our CPI is not a substitute for the GUS's CPI and other official measures of inflation, but rather represents a faster and more frequent estimation of a similar nature. From this standpoint alone, the project is, thus, a unique and innovative study using *Big data* for statistical purposes in Poland.

2. Online vs. offline prices

Researchers, who blazed a trail in terms of analyzing prices via the Internet was Cavallo & Rigobon (2011). They collected individual-product prices in 36 supermarkets across 22 countries and 5 continents using a scraping software. In effect, they received 5-million-observation data sets to test for a price stickiness. Since that time, at least 7 studies have already been performed on these data sets, and some findings are quoted in this paper.

Although Internet shopping is becoming increasingly popular, online traders are still not necessarily representative of the typical consumer. Similarly, prices on the Internet may differ from those in brick-and-mortar retailers. If prices online and offline behave differently in long term, then index based on online prices alone cannot be extrapolated to the whole economy. However, studies comparing prices of the same goods in traditional

stores and online retailers found that those prices are either identical or there is a stable deviation between them. Cavallo *et al.* (2014). Basically, we can distinguish three types of relation between online and offline prices:

a) **Permanent shift**. Online prices may have a tendency to outrun prices in the real world. The shift may be around 2-3 months, like in France.
b) **Cohesion**. Online and offline prices behave in a similar manner, having in mind short-term deviations, like in the USA.
c) **Different strength.** Online prices may react stronger or weaker than those in the real economy, like in Columbia.

Moreover, a relation between prices may also differs due to a reference period. Even though online and offline prices follow similar trend on an annual basis in the USA, monthly indices vary significantly in some points.[1]

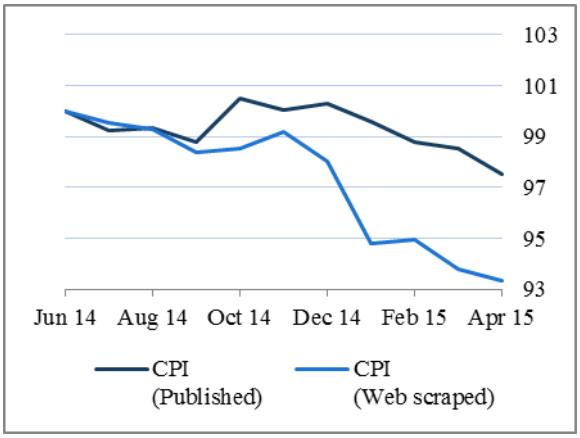

Figure 1. Price of food and drinks in the UK (June 2014 = 100), source: Office for National Statistics in the UK.

A proof of different behaviors between online and offline prices came from the UK. The UK's statistical agency compared a conventional consumer price index with an online-based version on data from supermarket websites. During the nine-month period between June 2014 and April 2015, the official CPI for 35 items of food and drink fell 2.5%, while the equivalent built on online prices from three large supermarkets fell by 6.7% (Figure 1) (Giles, 2015).

[1] Figures and data for each country could be found in Cavallo A. (2015)

Short term disparities between online and offline prices result from number of reasons. First, pricing strategies of companies. Since online customers and their behaviors vary in comparison to offline buyers, companies adjust their offer to maximize profits. Second, a number of online transactions is still far beyond those in the real economy. In 2014, an estimated share of online goods in total retail of goods was around 5.9% globally, and 6.4% in Europe. Although this share has increased more than doubled since 2010, it is still a small fraction of the global consumption (Ecommerce Foundation, 2015). Third, online market is more competitive. Due to price-comparison websites, an asymmetry of information about the prices on the market is reduced. In effects, retailers' markups are modest and the prices are more flexible than in bricks-and-mortar stores. Fourth, *menu costs* are extremely low, actually they refer to price data update, which also contributes to a lower price rigidity.

3. GUS's CPI vs. online CASE CPI

A composition of the online CASE CPI's basket of goods and services follows GUS's methodology; thus, it based on 12 main aggregates announced by GUS in February each year (the basket composition for 2016 is presented in the Table 1).

However, those main aggregates do not provide enough precision to take into account the consumption patterns of a typical person. Therefore, each of the main aggregates is divided into smaller categories (93 in total), according to latest available the Household Budget Survey. For instance: according to the Household Budget Survey in 2014, consumption of cereal accounts for 0,59% of consumption of Food and non-alcoholic beverage, which, as an aggregate, accounts for 24,04% of total expenditures, according to the 2016 revised inflation basket (Table 1). A list of representatives, which form each category, come from RAMON Eurostat's Metadata, according to COICOP classification.

Due to missing categories (about 15% of goods and services do not have prices available online) the shares are accordingly scaled. The share of missing categories is split proportionally between other categories inside the same aggregate (the same is done if one or more aggregates is missing – its share is divided accordingly between the remaining ones). For example, services in recreational and cultural category are characterized by high heterogeneity, variability of the offer over time, and dispersion of pricing data. Those factors might influence reliability of the data and as a result that category is omitted. The share of recreational and cultural services in expenditures is proportionally divided between other categories in Recreation and culture aggregate. The idea of scaling missing categories is in line with Cavallo (2012).

Table 1. Official weights of main aggregates of goods and services in inflation basket in 2016

Category	Share
Food and non-alcoholic beverages	24.04%
Alcoholic beverages. Tobacco	6.56%
Clothing and footwear	5.47%
Housing. water. electricity. gas and other fuels	21.04%
Furnishings. household equipment and routine maintenance of the house	4.99%
Health	5.45%
Transport	8.72%
Communication	5.27%
Recreation and culture	6.63%
Education	1.01%
Restaurants and hotels	5.04%
Miscellaneous goods and services	5.78%
Total	100.00%

3.1. Sources of data

To ensure that online CASE CPI is not influenced by one retailer and its pricing strategy, data on prices are collected mostly from the websites which compare prices from online shops. This allows us to easily track prices of certain goods and services from a number of outlets. What is more, those sellers do not have to be predefined - if a new retailer enters the market, price-comparison websites will automatically include its offer in the search results. By using price-comparison websites, online CASE CPI takes into account a dynamically changing market environment. For representatives which are not listed on those websites, we use dedicated websites, like commodity exchange for fresh products or industry portal for petroleum prices. In case of services and utilities data are collected either from private websites that contain listings of prices or public websites publishing current tariffs (for example websites of municipal transportation authorities). In effect, we tracks

prices on about 50 different pages across the Internet, however through the use of price-comparison websites, data from more than 3000 outlets is taken into account. Some of the data sources are updated in real time, some in regular intervals (at least once a month), others only when price of certain commodity has changed (for example prices of electricity are in effect until new tariff is announced).

To gather information on prices, Central Statistical Office of Poland sends out more than 200 agents to collect data from about 35,000 points of sales throughout the country each month. Data for online CASE CPI are scrapped using sophisticated internet robots. They gather approximately 60 000 observations each week, which means 240 000 observations monthly. The main advantage is the speed of collection because the whole process takes a few hours. As a result, we are able to calculate and announce the index almost in the real-time, even CASE's CPI, like the GUS's CPI, tracks approximately 1400 representatives. Main statistics for both indices are presented in Table 2 below.

Another advantage of the *scrapped data* is that prices of newly introduced products and services can be collected from the first day of their online appearance without need for any administrative adjustments in the basket. There is no need to decide whether new products should be included in the index - they automatically are. Moreover, GUS tracks prices for chosen representatives until they stop being offered. Then they look for the closest substitute in order to remain continuity. Official CPI measures control and "divide" price between attributes of certain products - if such product is discontinued (for example when a new technology is introduced) it can be compared to a new product with slightly different attributes[2].

Table 2. Comparison of CPI conducted by GUS and CASE

	CASE CPI	**GUS CPI**
Number of representatives	~1400	~1400
Number of observations	~240 000	~260 00
Frequency of data collection	Weekly	1-3 times a month
Time of availability	Real-time	Up to 4 weeks delay

[2] For more information please refer to: IWGPS: Consumer Price Index Manual: Theory and Practice (2004), chapter 7

We decided to track prices for chosen representatives only in a few categories, including processed food and apparel. These goods or services are characterized by relatively large heterogeneity in terms of price range and attributes and thus may disrupt the final result (for example a price reduction of an upscale dress does not mean that cost of living of a typical person fell to any degree). In general, online CASE CPI does not adjust goods and services for quality changes, as all items are treated independently. Because of use of a number of representatives for each product type, it could simply omit discontinued products.

4. Preliminary results

Online CASE CPI has been calculated since August 2015. The project is ongoing and we expect significant conclusions after 2 years of observation. At this point, we present a graph based on 9-month observation period. Green bar on Figure 2 presents an absolute value (in percentage point) of a difference between online CASE CPI and GUS's CPI in a given month. A negative value means online CASE CPI noted stronger deflation or weaker inflation. On the other hand, a positive value means GUS CPI index noted weaker deflation or stronger inflation. As we can see the differences between both indices did not exceed 0.3 p.p. apart from October 2015, in which we noted a significant decreases in online price in the three categories: Food and non-alcoholic beverages, Clothing and footwear, and Restaurants and hotels.

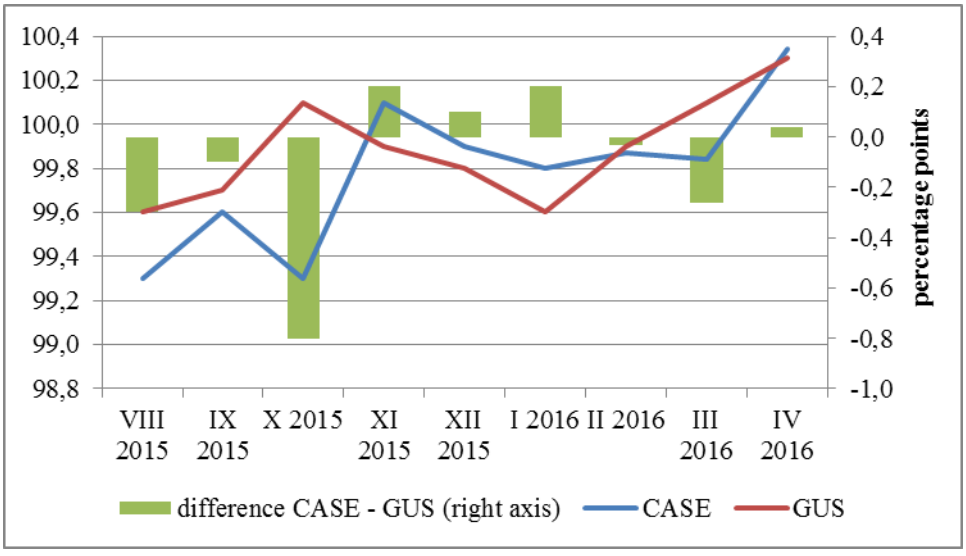

Figure 2. Comparison between GUS's and CASE's indices since August 2015

Moreover, since November 2015 we have published online CASE CPI weekly[3]. For this purposes, we compute an average price level for the last 4 week and compare it with the same average for the 4 week before. In other words, we analyze 4-week moving average each week. GUS estimates CPI on a monthly basis, so a direct comparison is impossible but we can expect some relations that would outrun the changes in prices earlier. Figure 3 presents GUS's and CASE's weekly indices.

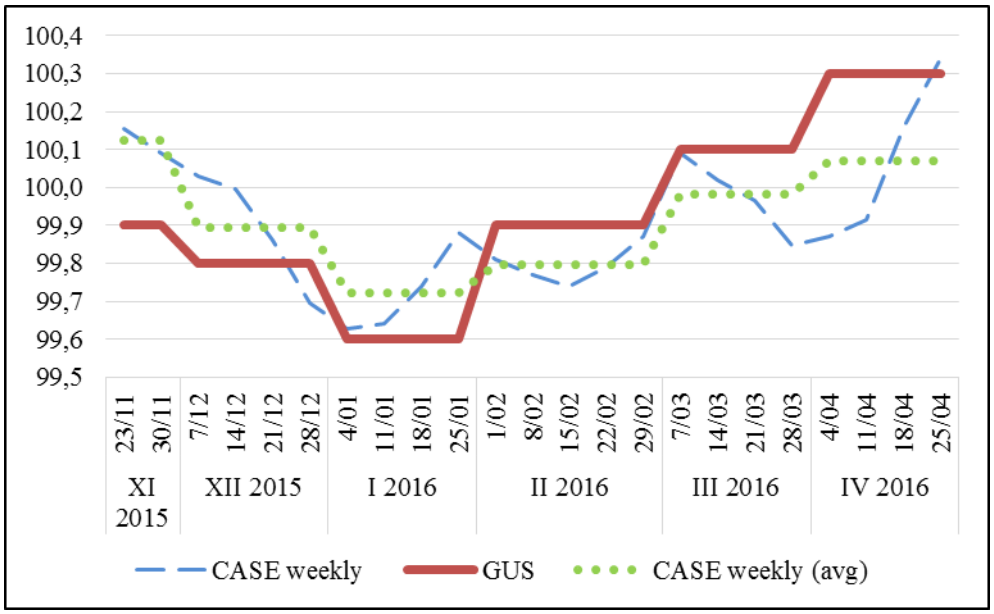

Figure 3. Comparison between GUS's monthly and CASE's weekly indices since November 2015

A blue dashed line shows online CASE CPI counted each week. After a decrease in the end of 2015, the online prices fluctuated rather regularly and local maximums were reflecting an upward trend. A green dotted line is an average of weekly estimates obtained for each month. Both lines are presented in reference to GUS's CPI (red line).

[3] http://www.case-research.eu/

5. Findings and further research

Summarizing, due to a short observation period it is too early to conclude on a correlation over time between both indices. Testing Cavallo's types of relationships: Permanent shift, Cohesion or Different strength is also premature at this phase of the project.

At present, we can highlight the main findings that appeared during our research. First, in terms of time and money spend on data collection, online based indices are more economical. In effect, online CPIs can be measured and published more frequently, with a small time delay. Second, an availability of certain goods and services in the Internet. According to Cavallo (2012), about 60% of the basket that makes-up the CPI is available online. The proportion is different for each country and depends on the distribution of CPI basket and how well-developed the online market is. CASE's CPI includes about 87% of goods and services of a typical household basket (77 out of 93 weighted categories are covered). Third, compared to the traditional CPI measure, more goods than services are available online, or at least they have had a price in the Internet so far. For instance, services like hairdressing, construction services, or a dentist's appointment have a better representation out of the Internet - it mean an overrepresentation of goods in online indices. Fourth, since profiles of online and offline customers are not the same, both online sellers and buyers behave differently in comparison to offline counterparts. Therefore, to analyze online consumption exclusively, the basket of goods and services should stem from virtual markets instead of the official weights from the offline surveys. In effect, online prices are measured according to habits of conventional consumers, instead of using a structure of consumption occurring in the Internet. Such surveys have not been performed in Poland yet.

As previously explained, a consumer price index based on online prices may precede classical measures of consumer prices. This allows one to predict the level of officially announced measures, as well as, to predict the possible reactions of financial and public institutions which use inflation measures in their decision making process. In this sense, the online CASE CPI could be used as an inflation expectations measure. The micro scale of collected data may also allow for investigating different properties of price adjustments in Poland such as price stickiness, frequency, and scope of price changes. Furthermore, we are able to disaggregate our index up to 93 subcategories, which allows us to make analyses of price trends in certain product groups and branches of economy.

Online CASE CPI will be constantly analyzed and developed in order to cover goods and services which will be entering the virtual markets. As an example of so-called *Big Data* this exercise will also allow for the permanent storage of big datasets relating to prices in Poland. Furthermore, alongside e-commerce development around the world, we can expect increasingly importance of inflation statistics based on the online data.

References

Cavallo, A., & R. Rigobon (2010). The Distribution of the Size of Price Changes NBER Working Paper, w16760. Link

Cavallo, A. (2012). The Billion Prices Project: Building Economic Indicators From Online Data, MIT Sloan, Geneva, May 31st, 2012. Link

Cavallo, A., Cruces G., & Perez-Truglia R. (2014). Inflation Expectations, Learning, and Supermarket Prices: Evidence from Field Experiments, NBER Working Paper 20576, November 2014. Link

Cavallo A. (2015). The Billion Prices Project and PriceStats. AEI Conference: The federal statistical system in a Big Data world, March 2015. Link

Ecommerce Foundation (2015). Global B2C E-commerce Report 2015, Ecommerce Foundation. Link

Giles, C. (2015). Supermarket prices fall faster than official inflation measure, Financial Times 2015. Link