Economic commentary

Price information collected online and short-term inflation forecasts

Isaiah Hull, Mårten Löf and Markus Tibblin
The authors work at the Monetary Policy Department of the Riksbank[1]

Forecasting short-term inflation developments (e.g. inflation over the coming months) is important for a central bank. There are certain elements within the published inflation figures that are volatile and inherently hard to forecast even in the short-run. Fruit and vegetable prices, energy prices and air travel prices are examples of product groups within the inflation measure that historically have held a high degree of volatility in Sweden. An automatic internet data collection process was developed to collect sales prices daily for selected fruits and vegetables from a number of Swedish online retailers. The results indicate that the information from the daily data could increase the precision in short-term inflation forecasts in Sweden.
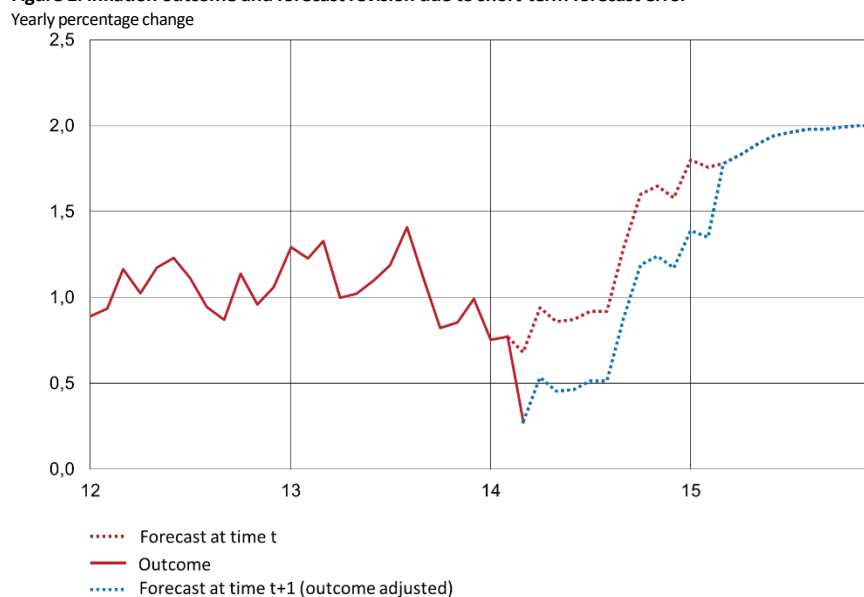
Short-term inflation forecasts matter to central banks

The objective for monetary policy at the Riksbank, like many other central banks, is to maintain price stability. This is interpreted as keeping inflation, i.e. consumer prices measured through an index based on a basket of goods and services households tends to consume, low and stable.[2] The consumer price index (CPI) consists of a range of sub-categories. In Sweden, fruit and vegetables, air travel and fuel prices are examples of sub-categories, which historically have shown a relatively high level of price volatility. This volatility may at times create challenges in forecasting CPI in the short-term (e.g. the coming months) although these sub-categories sum to only a small share of total CPI. Short-term CPI volatility and related short-term forecast errors may seem like a minor issue, given the fact that monetary policy decisions are based on inflation development over a longer time horizon. However, accurate forecasts of future inflation are dependent on precise information regarding current inflation, as well as good forecasts of short-term inflation developments. It is important from a forecasting, and also a monetary policy, perspective to be able to decide to what extent a large deviation between forecast and outcome is due to temporary or more permanent factors. A large forecast error that is due to permanent factors can affect the path of the new inflation forecast over the next twelve months, assuming that no other changes will occur, see Figure 1.

Figure 1. Inflation outcome and forecast revision due to short-term forecast error
Yearly percentage change

.... Forecast at time t
——— Outcome
.... Forecast at time t+1 (outcome adjusted)

Note. The figure illustrates how a forecast error at time t (difference between the solid and the dashed red line) may affect the forecast path for inflation in the coming year (blue dashed line).

New data and new analytical methods

The information published on the Internet is growing very fast. It has been widely discussed how, for example, central banks can make use of new methods to collect and analyse this type of data. The Riksbank arranged a Big Data workshop in September 2015, at which a number of central banks, researchers and private firms laid out examples where new data and methods arising from Big Data may support analysis and decision-making at central banks.[3] One area of growing interest among central banks is the increasing amount of sales price information available online. Leading research in this area has been conducted in the "Billion prices project". This project was initiated by Cavallo and Rigobon, who have in a number of studies shown that a price index produced using online price data follows the official measures of consumer prices such as CPI fairly well (Cavallo and Rigobon, 2014). Also, through discussions with retailers in relation to the Riksbank's Business survey it is clear that there are in general small differences in consumer prices between offline and online prices in Sweden,[4] which is also in line with a large international comparison study conducted by Cavallo (2016). Furthermore, it has been shown that online pricing data performs well in forecasting CPI produced by national statistical offices and for some countries outperform models that include offline data (Aparicio and Bertolotto, 2016).

The pilot project –collecting online fruit and vegetable prices from the internet

Utilising online price data may potentially overcome a number of current issues in short-term inflation forecasts. First, there are no time lags in data collection, as online prices can be collected in real-time i.e. price data are available before official inflation figures are published. Moreover, collecting price data online generates information at a very granular level., This

---

enable a more detailed analysis, which for example could give information on whether an unusual price development is temporary or not. Also, as online data collection can be automated, daily price indices can be produced and included in models at a low cost and with very limited resources.

A small pilot study was initiated to investigate whether prices of fruit and vegetables that are available online could improve the accuracy of short-term inflation forecasts in Sweden. Prices of fruit and vegetables make up 3 percent of the CPI basket. Although these prices constitute a small part of the basket, strong price movements in this sub-index can have a clear impact on the aggregated figures. The pilot project was set up with the aim to:

1. Create a process for automatic online data collection collecting online price data daily for a few selected fruit and vegetables from Swedish retailers with e-commerce. Online prices have been collected for oranges, bananas, peppers, apples, cucumbers, cabbage, grapes, cauliflowers, pears, leeks and tomatoes. The correlation between the subset of selected fruit and vegetable prices and the total index of fruit and vegetables in the CPI is fairly high, see Figure 2 below. Hence, the subset of prices captures the variation in the total index for fruits and vegetables in the CPI quite well. It was therefore assumed that internet data for selected products would be useful when forecasting the overall CPI-index for fruit and vegetables.

2. Create a weighted monthly price index based on observed price changes online for the selected fruit and vegetables mentioned in 1.

3. Test whether the constructed online price index adds value to the existing short-term forecasting models for fruit and vegetables currently used by the Riksbank.

The rest of the paper is outlined as follows: First a brief description of the data collection method is given, then the index construction and data transformations are described. Finally, the results are presented, including a discussion and overall conclusions drawn so far from the pilot project.

Figure 2. Prices of selected products in CPI and the aggregate of fruit and vegetables in the CPI
Monthly percentage change
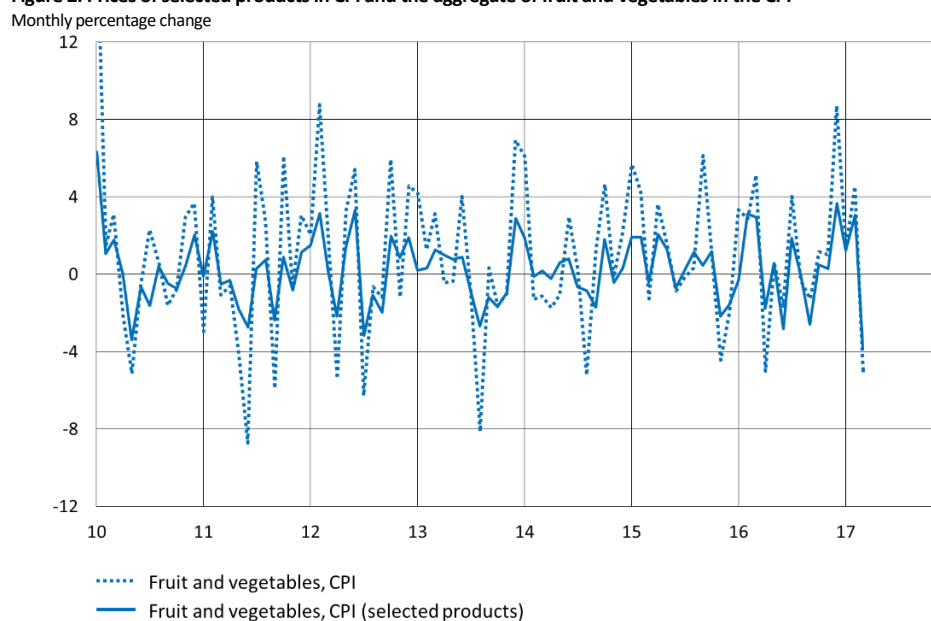
Scraping

We use a technique called scraping to collect the data for this project. Scraping involves sending programmatic requests to a website's server. The server responds by returning the underlying code that would be executed in a browser if you were to visit a given page on its website. The code's structure can then be employed to identify targeted items, such as product names and prices. We use additional tools to schedule the scraping tasks so that they are automatically performed on the targeted sites at the same time each day. We also limit the speed at which we send requests to each website to ensure that no strain is placed on their servers.

All data collection tasks are performed on a Linux virtual private server (VPS). The server executes three scripts in sequence at the same time each day, as illustrated by Figure 3. The first visits the websites of four large grocery retailers. It extracts the code from all pages related to fruit and vegetables. A list of all product prices and names is then identified in the code and saved in a .csv file for each location and day. The raw code is also saved in .txt format for 90 days, allowing us to correct errors discovered later. The script then uses regular expressions, which identify patterns in text to filter the data, creating a second .csv file for each location and day that consists only of targeted fruits and vegetables.

After the scrape is finished, the server executes the second script, which merges the filtered location and day files with past data. It first identifies matches between products in the new data and products identified in previous scrapes. Items that were not identified in previous scrapes are assigned new location-product IDs and products that already have IDs are merged with existing location-product time series.

Finally, the server executes the third script, which checks for errors. This script identifies the number of files produced, the size of those files, and the data types used in those files. The server then delivers an error report via SMS, allowing us to quickly identify and correct errors.

Figure 3. The daily scraping process
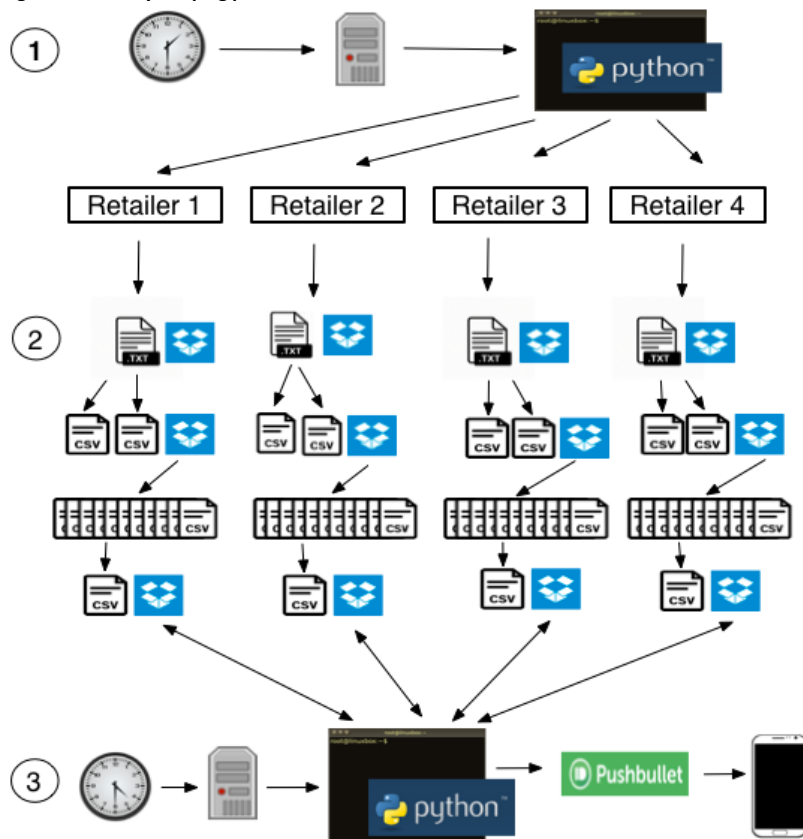
Forming indices

In a first step the collected online prices are sorted into different groups, orange prices in one group, and apple prices in another group and so on. The dataset is then truncated so that it exactly matches the measurement weeks in the Swedish CPI survey.[5] In the next step a geometric mean is calculated for each product and month, i.e. an average price of oranges in January, one for orange prices in February and so on.

Figure 4 shows the monthly percentage changes in prices for oranges, cucumbers, peppers, and tomatoes from May 2015 to January 2017. The blue lines show monthly percentage price changes according to the CPI, while the red lines show the corresponding changes based on information from the internet. The correlation between the CPI prices and prices available online are relatively high. The correlation is highest for prices on cucumbers and tomatoes. The lowest correlations are measured for prices on oranges.

---

[5] Price data on fruit and vegetables in the CPI were collected for three weeks in the middle of each month (a total of 21 days) until December 2016. Thereafter, Statistics Sweden has made some changes.
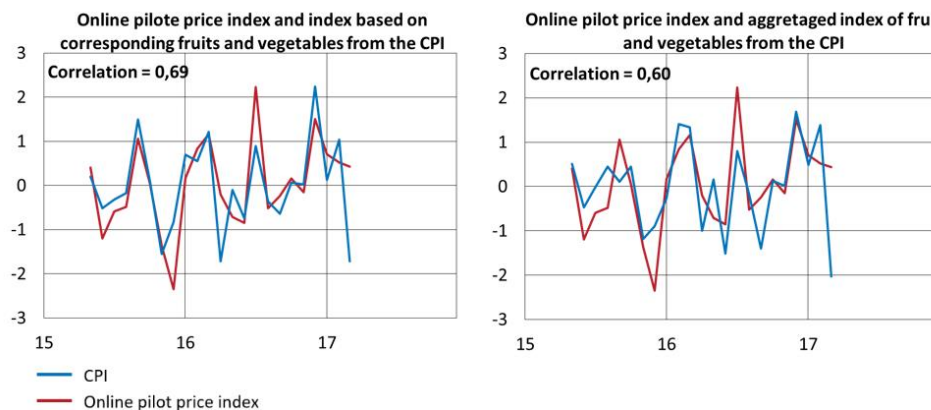
Figure 4. Price changes according to internet data and corresponding price changes according to the CPI
Monthly percentage price changes. May 2015 – March 2017

CPI
Online pilot price index
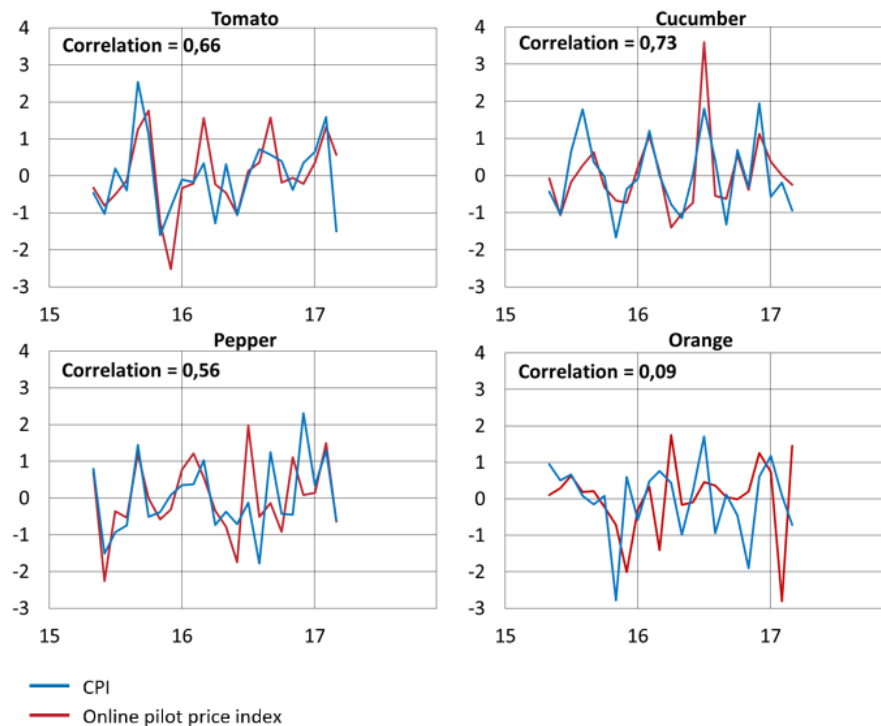
Note. The red lines indicate online data while the blue lines indicate data from official CPI.

These price changes are then merged in to an overall index (henceforth denoted the online pilot price index) using the CPI weights according to CPI for the fruit and vegetable prices collected online. The left panel of Figure 5 compares the online pilot price index (red line) with a weighted CPI-index for the corresponding fruit and vegetable prices (blue line). The right panel in Figure 5 shows the online pilot price index together with the total price index for fruit and vegetables in the CPI (i.e. not only the fruit and vegetables collected in the pilot study). Naturally, the correlation decreases when comparing the online pilot price index with the total index for fruit and vegetables (left panel compared to right panel in figure 5). However, it is still fairly high and the online pilot price index captures most of the volatility in the official fruit and vegetables index.

Figure 5. Comparison between an aggregated index based on internet data and two different indexes with CPI data
Monthly percentage price changes. May 2015 – March 2017.

Note. The figure show aggregated price changes according to the data from the internet (red lines). Blue lines show price changes according to CPI. The blue line in the left hand panel show price changes if one use the same products as in the internet collection. The blue line in the right panel shows price changes according to the index for fruit and vegetables in CPI.

Forecast evaluation

In this section we want to test more formally whether the internet prices can be used when forecasting the index for fruit and vegetables in the CPI.

Here we compare the forecasts from the Riksbank's current models for fruit and vegetables with the forecasts generated using the internet prices. A variety of indicators are included in the models now used by the Riksbank for short-term forecasting. The models for prices of fruit and vegetables include one indicator at a time, together with a moving average of an exchange rate index (KIX). In addition, lags of the dependent variable, dummies for outliers and in some cases moving average terms are included in the models. The results are summarized as the mean forecast from all of these specifications. This approach is denoted MEAN in the evaluation below. Principal component analysis is also used to summarise the information from the indicators in the first step. The resulting Principal component indices can be seen as a weighted averages of all the indicators. In that case the models include one or more of these summary indices instead of the individual indicators. This approach is denoted PC in the evaluation below. These model-based forecasts are compared with an approach using the online price index. Here we simply use the online price index as a forecast for the aggregate of fruit and vegetables in the CPI. We denote this second approach (OP).

The evaluation period is May 2015 to March 2017. The root mean square error (RMSE) is used to compare the forecasting ability between the approaches. The RMSE summarises the standard deviation in the forecast errors and their systematic deviation. The lower the estimated RMSE, the better the forecasting ability. A forecast that is always correct has a zero RMSE.

Table 1 below summarizes the results. During this short evaluation period the best approach has been OP, where the online price index are used directly. Hence, it seems like online price information could add value in forecasting price changes for fruit and vegetables in the coming month. However, the evaluation period is very short, and longer time series are required to draw clearer conclusions.

Table 1. RMSE for different approaches, May 2015-March 2017.

Horizon	Nowcasting procedures		Online prices
	Mean	PC	OP
1 month	2.0	2.0	1.6

Discussion

The work on developing scrapers for automatically collecting and analysing online price data has been a bit of a trial-and-error exercise. The codes developed have had to be tailored to handle a range of different, changing website layouts. However, it has been possible to build scrapers coping with changing layouts, thus making it possible to maintain a collection process with very limited resources.

The journey of analysing online micro data has just started during this pilot project. A fair amount of time has been devoted to investigating whether further transformation of the online data could increase forecasting ability. For example, an index only based on online prices showing high correlation with corresponding official CPI prices was constructed. The online price information has also been summarized using Principal Component Analysis. These types of indices have at times performed very well, but also shown large variability and overall not been as good as just the simple online index described above. Thus, no transformation is currently performed on the data. This could potentially be an area of further investigation.

Additionally, the collected daily data could also be analysed from the perspective of firms' pricing behaviour. Detailed micro data may be a source of understanding when, why and how firms change their prices. This could also be a potential area of further investigation.

Conclusions

The pilot project has revealed a number of insights regarding online price collection and the analysis of such data in relation to inflation forecasting.

First of all, it has been proved possible to consistently scrape online retail prices from e-commerce websites held by retailers in Sweden. On the one hand, creating scripts and IT-processes ensuring a stable data collection takes time and requires programming competence not traditionally available at central banks. On the other hand, once scrapers have been put in place, only minimum maintenance and development have been required.

Furthermore, the result so far indicates that online pricing data add some value when forecasting short-term developments of consumer prices for fruit and vegetables. Given this outcome, there may be scope for further expanding the collection of prices available online and using them as input in short-term forecasting models.

The aggregated time series of online prices is still short and further analysis is required to ensure that the online price index have forecasting ability also in the future. The result also indicates that there may be room for increasing the collection of prices available on the internet and using them in short-term forecast models. However, there are still many questions that should be investigated, such as how to use the information in the best possible way.

References

Aparicio, Diego, & Manuel Bertolotto. (2016). "Forecasting Inflation with Online Prices." Working Paper - MIT.

Bernanke, B. S. & Boivin, J. (2003), "Monetary policy in a data-rich environment," Journal of Monetary Economics 50 (3), 525-546.

Bertoloto, M. & Cavallo, A. & Rigobon, R. (2014), "Using Online Prices to Anticipate Official CPI Inflation," UTokyo Price Project Working Paper Series 031, University of Tokyo, Graduate School of Economics.

Cavallo, F. Alberto, (2016), "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers," NBER Working Papers 22142, National Bureau of Economic Research, Inc.

Stock, J. H. & Watson, M.W. (2002), "Forecasting using principal components from a large number of predictors," Journal of the American Statistical Association 97:460, 1167-1179.

Stock, J. H. & M. W. Watson, (2004), "Combination forecasts of output growth in a seven-country data set," Journal of Forecasting 23 (Issue 6), 405–430.