NOWCASTING INFLATION USING PRICES FROM THE WEB

Mirko Đukić, Iva Krsmanović, Miodrag Petković

Available at www.nbs.rs

Economic Research and Statistics Department

NATIONAL BANK OF SERBIA

Belgrade, 12 Kralja Petra Street

Telephone: (+381 11) 3027 100

Belgrade, 17 Nemanjina Street

Telephone: (+381 11) 333 8000

www.nbs.rs

Nowcasting inflation using prices from the web

Mirko Đukić, Iva Krsmanović, Miodrag Petković

Abstract: The paper presents the methodology which the National Bank of Serbia uses to nowcast inflation in real time, based on prices from the web, downloaded automatically using web scraping. A specific feature of the method used by the National Bank of Serbia is that it is based not only on prices for online shopping, but on every relevant data on the prices, including those displayed on the web merely informatively. The intention of the NBS was to cover as many items in the CPI as possible (around 90% at the time of writing this paper), in an endeavour to acquire a more reliable nowcast of the inflation central tendency. In the first year of applying this method, nowcasting performance has been encouraging – on average, inflation nowcasts were at the level of the official figures (nowcasts are not biased), the mean forecasting absolute error was 0.20 pp, and the median was 0.13 pp, which is not significant given that the observed period was characterized by relatively high and volatile inflation.

Keywords: inflation forecasting, web prices, web scraping, big data
[JEL Code]: C53, E17, E58

Non-Technical Summary

The main goal of the National Bank of Serbia is to achieve and maintain price stability. The headline inflation measure, a percentage change of the consumer price index, is published by the Statistical Office of the Republic of Serbia on the 12th day of a month for the previous month. There are multiple reasons why monetary policy makers would find it useful to have a reliable estimate of current inflation before the actual figure is published, and one of the increasingly popular methods to achieve that is by using prices from the web downloaded automatically through so-called web scraping. This method is supported by the abundance of prices on the web, which, in combination with the ever-increasing computing power, makes it possible to scrape a large amount of data from the web. Nowcasting the CPI means using a large number of prices of products comparable to products in the CPI basket.

One of the main advantages of this approach is that it allows price monitoring on a weekly, or even daily basis, which is cost and time efficient. The main flaw is a lack of guaranties that websites and firms will continue to publish prices in the same volume, quality, and form. Changes in prices, acquired in this way, are not a substitute, but rather a complement to the official methodology of producing CPI inflation, and they are an important input in forecasting inflation.

For the nowcasting purpose, the NBS uses all available information that can improve the nowcast, including websites of retailers, websites that display prices only informatively, or ones that compare the prices of various retailers, as well as government websites that list the prices of its services. By using these various sources, we managed to cover around 90% of the CPI basket. This is what distinguishes our method from comparable ones, which mainly nowcast subcategories of prices, such as food prices. The prices for which we did not manage to find a reliable source on the web were imputed using some of the methods.

The performance of nowcasts in the first year of applying this method has been encouraging – nowcast inflation in the observed period on average was at the level of actual inflation, the absolute deviation was 0.20 pp on average, while the median was 0.13 pp which is not significant, taking into account that the average m-o-m inflation in the same period was 1.2%. Besides, the correlation between nowcast and actual rates is relatively high. In the coming period we intend to look for and include new sources from the web, in order to improve the precision of nowcasts.

Contents

1 Introduction

The abundance of price data on various products and services on the web can be used as a good source of information for estimating inflation movements in real time. A precondition for this is to set the system to take those large amounts of data automatically and process them in an adequate way. The purpose of this paper is to present the way this is done in the National Bank of Serbia.

The current inflation figure is an important input in medium-term projections. Besides being used for analysing current inflation pressures, it also affects the projection directly, through the calculation of y-o-y rates for the next twelve months, and indirectly, through the persistency in monthly/quarterly inflation rates.

The problem central banks are faced with in reality is that official data on current inflation are available with a time lag, typically after the forecasting process is finished. For instance, official inflation data in Serbia are published around three weeks after most of the prices from the CPI basket were recorded by the Statistical Office. For this reason, a number of techniques and models have been developed – econometric models, surveys, experimental research… – that give an initial estimate of prices and inflation (nowcast) before the official figure is published.

Availability of an enormous number of prices on the web, with an ever-increasing power of computers, gave rise to the use of web scraping for nowcasting purposes. This method is one of the cheapest and most efficient ones timewise. Besides prices, the set of data can also include information on products, discounts, and availability, and the frequency can be not only monthly, but also weekly, or even daily (Macias &Stelmasiak, 2019). The relevance of prices collected through this method is very important, with an advantage that their dynamics can be monitored in real time (Powell et al., 2018). Unlike survey methods, the data are available immediately and can be acquired without an intermediary. Depending on the specific needs, prices collected this way can be a separate category of prices on the web, an approximation of one category (for instance, food within the CPI), or an approximation of the entire CPI.

On the other hand, it is necessary to consider the shortcomings of this approach to inflation assessment, which arise primarily from the very nature of the data, which is a direct input. Additionally, there is the problem of data availability – there is no guarantee that websites or private companies will continue to publish prices in the same volume, quality, or form in which they do now (Kapetanios&Papailias, 2018). Clearly, the main advantage of using data obtained from the internet is their high frequency, but it would be incorrect to assume that these data can be a substitute for traditional data collection and analysis. It would be far more accurate to say that this is a complement to the official methodology of creating the CPI.

One of the first authors to assess current inflation using internet prices was Cavallo (2013), who in his first paper attempted to obtain an inflation estimate in five Latin American countries using prices related to food, non-alcoholic beverages, and household goods, and in addition to the usual reasons, he stated that such an index can be used to obtain alternative measures of inflation when official statistics have lost their credibility. His results show that online indices managed to approximate well the levels and dynamics of official inflation when it comes to Brazil, Chile, Colombia, and Venezuela. On the other hand, when it comes to Argentina, a

statistically significant result was obtained indicating a discrepancy between online prices and official statistics, as this was a period when the Argentine government was criticized over doubts concerning the validity of official inflation measures.

An important source of information for this type of research is the Billion Prices Project, an academic initiative at MIT and Harvard, one of whose creators is Cavallo. Prices from hundreds of online stores around the world were collected daily in the period 2008–2016. Cavallo's work was continued by Carvalho (2020) with his research on the Brazilian inflation measures, in which he expanded the methodology with a dynamic factor model for assessing current inflation. Model results show a statistical significance and improvement in projections using dynamic variables versus using raw data.

In their work, Macias and Stelmasiak (2019) from the National Bank of Poland used web scraping to improve food inflation projections. Given that food accounted for about a quarter of total household expenditure in Poland at the time the paper was published, and that food prices are characterized by high volatility and seasonality, improving the accuracy of these price estimates proved to be significant. As a potential obstacle, they cited the fact that a small number of basic food purchases are made over the internet, hence the question of comparability of retail and internet prices arises. The prices obtained by the web scraping method were aggregated into the food price index, which, despite this, proved to be statistically significant and with a smaller prediction error when it comes to future values of the official index compared to the benchmark ARMA model. The performance of the model in estimating food inflation improves even more when the lags of the used variables are included.

The Central Bank of Armenia uses web scraping as a methodological approach, in order to monitor the dynamics of goods and services prices on the market (food, non-food, services) in real time and to obtain quick estimates of current inflation, but also to monitor housing prices for real estate market research (Aghajanyan et al., 2017). In addition, a significant number of statistical offices have their own measure of current inflation based on online prices, including the statistical offices in the USA (Horrigan, 2013), the United Kingdom (Breton et al., 2015), as well as in the Netherlands, New Zealand, and Norway.

Jaworski (2021) also assessed food inflation in Poland using online prices during the crisis caused by the coronavirus pandemic. It must be underlined that this way of collecting prices is extremely important in a period when it is not possible to read the prices physically, as was the case during the lockdown, as well as when the number of online purchases is on the rise. Using a similar methodology, Jaworski showed that it is possible to obtain a reliable estimate of food inflation one month before the officially published indices. A similar result was obtained in relation to food inflation in Turkey, where Soybilgen et al. (2021) showed that online inflation indices successfully assess the price of food.

Of the papers that dealt with the assessment of overall inflation, Bertolotto et al. (2014) tried to improve inflation projection using the VAR model, relying on the realized inflation indices and online prices in the past six months as variables. In the online price index, prices that cover a large part of the basket were used for food, soft drinks, and transport. For other subgroups within the CPI that are not included in these categories (such as services), approximations based on the group dynamics were used. It has been shown that online prices

are more responsive than retail prices for comparable products, and that triggers for price changes on the internet are incorporated into retail prices more slowly and gradually.

Relative to these papers, our approach has some specific features. For one, our nowcasts are based not only on prices for online shopping, but also on every relevant information on prices on the web, including those that are displayed only informatively. For this purpose, in addition to websites of retailers, we also used those that compare prices of different retailers, as well as websites of government institutions and public companies which list the prices of their services. Besides, while methods described in other papers deal with inflation partially, nowcasting only some components, mainly food, our intention was to cover as many items from the CPI as possible. At the time of writing this paper, around 90% of the CPI basket is covered with sources from the web, while the missing part is imputed with some of the methods.

The remainder of the paper is structured as follows. First, we explained how the Statistical Office of the Republic of Serbia monitors prices and calculates inflation, and what kinds of sources from the web we use to record the prices of various products and services. This is followed by the technical explanations of methods we use to scrape data from the web and link it to the CPI basket. After that, we provided a detailed insight into the inflation nowcasting process based on web prices that we apply in the NBS, ending the paper with nowcasting performance of this method since its introduction a year ago.

2 CPI inflation and prices from the web

The headline measure of inflation in Serbia is the CPI inflation, published by the Statistical Office of the Republic of Serbia (SORS) on the 12th day of a month for the previous month. At the lowest level of aggregation, the CPI basket consists of 660 products and services (structure from 2023), whose prices are monitored in 15 Serbian cities, in different sales points. For different groups of the CPI basket, prices are monitored in different periods of a month:

- Non-food products – 1–10th
- Non-agriculture food products – 10–14th
- Agriculture products – first and third week in a month
- Services – 14–23rd
- Car fuels – every Wednesday in a month.[1]

Note that the official inflation data come weeks, or even more than a month after certain categories of prices were monitored. This lag provides an opportunity to produce a nowcast before the official data, which is very useful in the medium-term forecasting process.

The prices of numerous items can be found on the internet. For nowcasting purposes, we collect all relevant information on the prices of items in the CPI basket. The sources can be websites of retailers, with prices displayed for online shopping or as information for

[1]https://data.stat.gov.rs/Metadata/03_Cene/Html/030103_ESMS_G0_2018_2.html

consumers; websites that collect prices from various markets or retailers; or (general) government websites with information on administered prices. Websites that we use as a source cover from one to more than a hundred of products in the CPI basket.

In producing nowcasts we try to mimic what the SORS does as much as possible: we collect web prices for certain categories in the periods when the SORS does the same, typically on one day in the later part of the monitoring period, most often on the last working day; in line with the SORS methodology, we exclude the prices of temporary discounts, and we use the SORS weights from the CPI to calculate aggregate measures of web-inflation based on individual price changes.

On the other hand, unlike the SORS, we do not scrape prices by cities because: most often it is not even possible to find data by cities (big retailers publish unique prices for the entire country), it is reasonable to assume that changes in prices across cities are similar, and also because having too many sources would make the whole process more difficult to follow and maintain.

To keep things as simple as possible, we prefer to use websites that collect prices from various retailers (rather than their individual websites). For instance, we use: a website that compares daily prices of more than 6000 items (mostly food) of seven major retailers, the Ministry of Trade website (STIPS) that collects and displays prices of fruit, vegetables, and other food items on markets in 20 major Serbian cities, or a webpage that compares car fuel prices from six major fuel retailers. The first two cover most of the food items (both agriculture and non-agriculture), which is very convenient for nowcasting purposes.

When it comes to non-food prices, things tend to be more complicated. The easy part is administered prices, which have a significant share in the Serbian CPI basket (approx. one-fifth). These prices can be found on numerous government websites, such as those of government agencies, local municipalities, public enterprises, or in government bulletins. For instance, the price of electricity is displayed on the website of Public Enterprise Electric Power Industry of Serbia, the prices of cigarettes in the government's official bulletin (Official Gazette), administered prices of medicines on the website of the Serbian Health Insurance Fund, prices of municipal services (public transport, utilities, kindergartens…) on local government websites, and so on. In case of municipal services, while the SORS monitors them in 15 Serbian cities, we do this for three major cities (Belgrade, Novi Sad, Niš), covering 55% of the SORS sample in order to avoid making the process too cumbersome.

Picture 1 **Table with vegetable prices from STIPS website**

R.Br.	Proizvod	Jed.mere	Beograd(Kalenić)	Beograd(Skadarlija)	Čačak	Kragujevac
1	Blitva (sve sorte)visrednja pistandardno	veza	40.00	40.00	50.00	
2	Brokola (sve sorte)visrednja pistandardno	kg	400.00	450.00	350.00	
3	Celer (sve sorte)visrednja pistandardno	kg	300.00	300.00	250.00	220.00
4	Cvekla (sve sorte)visrednja pistandardno	kg	120.00		60.00	70.00
5	Dinja (sve sorte)visrednja pistandardno	kg	500.00			
6	Karfiol (sve sorte)visrednja pistandardno	kg	350.00	350.00	250.00	
7	Kelj (sve sorte)visrednja pistandardno	kg	150.00	150.00		
8	Kelj pupčar (sve sorte)visrednja pistandardno	kg	400.00	300.00		
9	Krastavac (Kornišon)visrednja pistandardno	kg				
10	Krastavac (salatar)visrednja pistandardno	kg	350.00	300.00	300.00	230.00
11	Krompir (beli)visrednja pistandardno	kg	150.00	130.00	80.00	100.00
12	Krompir (crveni)visrednja pistandardno	kg	150.00	130.00	80.00	100.00
13	Kupus (mladi)visrednja pistandardno	kg	200.00			
14	Kupus (sve sorte)visrednja pistandardno	kg	70.00	70.00	30.00	60.00
15	Luk beli (mladi)visrednja pistandardno	veza	40.00			
16	Luk beli (sve sorte)visrednja pistandardno	kg	900.00	700.00	500.00	450.00
17	Luk crni (mladi)visrednja pistandardno	veza	40.00	50.00		40.00
18	Luk crni (sve sorte)visrednja pistandardno	kg			70.00	80.00
19	Paprika (Babura)visrednja pistandardno	kg	500.00	450.00		
20	Paprika (ljuta)visrednja piposebno	kg				
21	Paprika (ljuta)visrednja pistandardno	kg				
22	Paprika (ostala)visrednja pistandardno	kg	500.00	500.00	350.00	
23	Paprika (šilja)visrednja pistandardno	kg	450.00	450.00	350.00	
24	Paradajz (chery)visrednja piposebno	kg	600.00	600.00		
25	Paradajz (sve sorte)visrednja pistandardno	kg	400.00	350.00	300.00	260.00
26	Pasulj (beli gradištanac)visrednja pistandardno	kg	400.00	350.00		
27	Pasulj (beli tetovac)visrednja pistandardno	kg	500.00	400.00		
28	Pasulj (beli)visrednja pistandardno	kg			350.00	320.00

Source: https://www.stips.minpolj.gov.rs/stips/nacionalni.

The most difficult task at this stage was to cover the non-administered section of non-food prices (hereinafter we will refer to this group simply as "non-food"). This group consists of 427 products and services, with a total weight of 45% in the CPI, covering various items like clothes, footwear, alcoholic beverages, medicines, furniture, personal care products, appliances, cars (new and used), car repair services, car insurance, rents, dental services, medical services, gym fees, cinema tickets, and so on. This heterogenous group of items required a large number of sources. Initially, we looked for websites that cover a large number of non-food products, and then went into more details, finding sources for one product at a time, prioritizing the ones with a higher weight. If available, we included multiple sources for individual items.

In total, we have managed to cover 90.5% of the CPI basket directly from the web. Aside from car fuel that consists of only two products (petrol and diesel), the best covered group from the CPI is fruit and vegetables (98%), followed by food excl. fruit and vegetables (94%), administered prices (91%), and non-food (85%).

3 Web scraping data and linking it to the CPI

R software was used for web scraping data on the prices of products and services. Web scraping is based on downloading data for an entire range of products from a particular online store, since the product range is not constant and the position of some items in CPI baskets can change over time in the online store.

For this purpose, our code automatically accesses the online store and in the first step it collects the web addresses with all products in the store. Next, the prices and characteristics of all products are collected from each web address of the online store.

The Rvest package is the first choice for web scraping, as it enables extremely fast data downloading in HTML format, without needing direct access to the website via a web browser. However, most online stores are created in a dynamic environment which requires interaction with the user, therefore we have to create additional codes based on the Rselenium package. This package allows the automation of web browsers in order to satisfy different requirements necessary for the presentation of products in online stores. Among other things, Rselenium allows selection from a drop-down list and automatic page scrolling to load all products available at a unique web address. It should be emphasized that web scraping based on the Rselenium package is significantly slower, since it requires loading the entire page content in the web browser.

A structural change of websites requires a correction of our code and adjustment to the new website setting. Therefore, we try to avoid using an excessive number of websites; otherwise, the functionality of the whole process would become an extremely demanding job, with the possibility of frequent errors.

In the process of collecting data, we respect the data-availability policy of the websites. Before scraping data, we check if the website explicitly prohibits the use of their data for purposes other than online shopping. We also do not scrape data more often than needed for nowcasting, so as not to put too much burden on the website.

After scraping, raw data are processed automatically in order to express prices in the pure number format with a dot as a decimal separator, and without any textual characters. For instance, if the price is expressed as “1.099,99 RSD/kg”, it needs to be changed to “1099.99”. The names of products in the database have to be unique, at least for the ones selected to be representative of CPI products.

Finally, before the data are used in the later stages of nowcasting, they have to be converted to a proper format. A single database of data from the web contains a column for names and a column for prices of products from one or multiple web pages, for any given day. In some cases, there are multiple columns for prices from different retailers, markets, or cities.

These databases are stored separately by month. In the last nowcasting process prior to the publication of this paper, we used 28 databases containing data from 130 web pages, with a total of roughly 30,000 items. These figures are occasionally changed as sources of data are added or left out.

The data we scrape can come in various formats. Most often they are in a form that enables automatic scraping of names and prices of products/services, ideally directly on the web page. In some cases, the table is given as a PDF file which makes scraping particularly challenging. Sometimes the price we are interested in is located in a text.

There are cases when data are unscrapable for some reason, most typically when they are in a PDF format that cannot be converted into a text or a table, or when a website itself restricts access to scraping its data. In those cases, we resort to manual scraping, meaning we simply type the data manually in our database. This is done only exceptionally, in cases when the

price is not changed frequently (once or twice a year) and when the source of the data is 100% reliable, as is the case with some administered prices on (general) government websites.

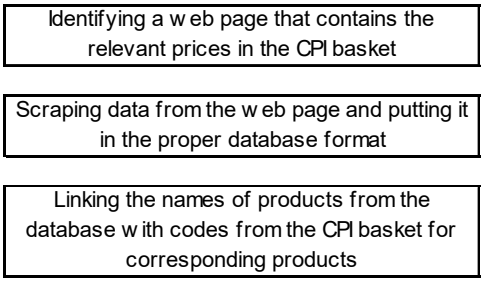
3.1 Selecting representative products

Once a new database is included in the system, it is necessary to decide which products/services from it represent items in the CPI basket. Technically, this linking is done through codes that the SORS uses in its calculations. When data from a website are scraped for the first time and transformed into a suitable database, we go through the data, item by item, and once we conclude that data are representative of the CPI items, they are assigned an appropriate SORS code.

Most CPI items have more than one representative product in these databases. For instance, in one of the databases "sunflower oil" is represented with "Dijamant", "Iskon" and "Vital" brands, therefore all of them are assigned the relevant product code.

If we were unable to find an identical online match for a product, we can choose the most similar one as the representative product. For instance, a 40-inch TV set can be represented with a 43-inch set, assuming that prices of such similar products co-move.

Picture 2 **The process of linking web prices with products in the CPI basket**



Source: authors' calculation.

Some of the databases contain as many as several thousand items, therefore this stage of the process can indeed be time-consuming. However, it is usually a one-off task, i.e. it is done only when the database is first downloaded. If items on websites that are elements of our databases are changed frequently, then a revision of the codes is required from time to time. In order to minimize this stage of work, we tend to choose websites that do not change their products frequently.

The end result of this stage is a codebook – with the names of items in the first column, and codes in the second – that is used in the nowcasting process to connect items from the web with items in the CPI basket.

4 The nowcasting process

The nowcasting process begins with scraping data from selected websites. Following the SORS dynamics, we do this according to the following schedule:

- Non-food – 10th
- Non-agriculture food products – 14th
- Agriculture products – end of the third week in a month
- Services – 17th
- Car fuels – every Wednesday in a month

If a date falls on a weekend, we scrape the data on the Friday prior to that weekend.

This way we created databases containing data from one or multiple websites for a given day. At the time of writing this paper, we used 35 databases for nowcasting, which includes databases from one source recorded on different dates. For instance, STIPS was scraped twice, once for nowcasting fruit and vegetables, and the other time for nowcasting food excl. fruit and vegetables.

Calculating inflation nowcast consists of: 1. Nowcasting price growth for CPI items that have representative items on the web, 2. Imputing missing prices, and 3. Calculating weighted price change (for both recorded and imputed prices) based on CPI weights used by the SORS.

Nowcasting the percentage price change for individual items that have representatives on the web is done in several steps:

- Calculating the percentage change in prices relative to the previous month for each product from every database.
- Calculating the percentage change in the prices of CPI items from individual databases as a geometric mean of representative products from the database. For instance, the percentage price change in the price of sunflower oil is calculated as an average of percentage price change of the prices of "Dijamant", "Vital" and "Iskon" brands (from the database). The CPI basket and databases are at this stage linked via the codebook described in the previous section.
- If a database contains multiple sales points (cities, markets, retail stores), rates by products are first calculated separately for sales points (as described in the previous bullet), and then the aggregate rate for a product from the database is calculated as a weighted average of rates from sales points. For instance, the percentage price change in the price of tomato from the STIPS database is calculated as a weighted average of rates from individual markets.
- Finally, if some of the products in the CPI basket have representatives in multiple databases, the final nowcast is calculated as an average of rates from these databases (acquired in the previous step).

It is important to add that for products that are monitored more than once during a month, as is the case with the price of fuel, their price for a given month is calculated as an average of the monitored prices, and then the procedure described above is applied.

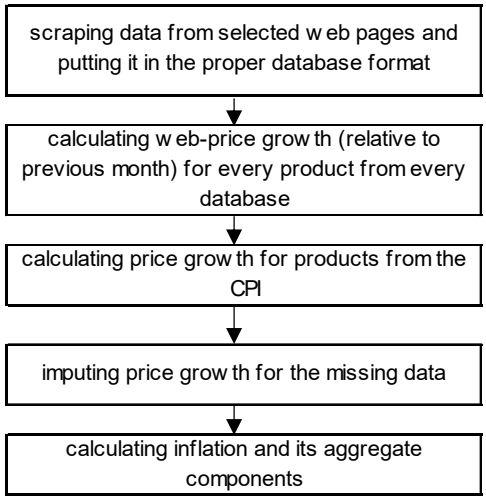
In the next phase, the prices of products for which, temporarily or permanently, we do not have representatives on the web, are imputed in some way. Here we distinguish between imputation that is also applied by the SORS from the one that we apply ourselves for items for which we do not have a source on the web.

First, in case of seasonal fruit and vegetables, we apply the method of imputation that the SORS uses in months in which these products are seasonally absent from the market. These prices are imputed based on the rest of the group in the month or based on their own prices in previous months. For instance, from January to May the percentage price change of peppers is imputed with the percentage price change of fresh vegetables in these months, while in the first month of the imputation period (December), its price is imputed as a weighted average of its own price for previous months, when it was available on the market.

At this stage, products and services for which we found representatives on the web and those imputed using SORS methodology cover 90.5% of the CPI basket. The remaining part is approximated using some of the following methods:

- Imputation with similar products – assuming that the price of the product for which we do not have a source on the web, changes at a similar pace as the price of a similar product for which we do;
- Imputation with the group – applied for food excl. fruit and vegetables, where for missing prices we assume that they move at the same pace as the rest of the group;
- Imputation with previous growth – applied for non-food prices, given that they tend to behave more persistently than other groups;
- Imputation with zero growth – applied for the missing administered prices, as they are not changed frequently (most often once a year).

Picture 3 **The process of estimating current inflation based on web prices**



Source: authors' calculation.

We apply imputation in order to keep the shares of different CPI groups unchanged. If we did not do that, components of inflation with a higher web coverage, such as fruit and vegetables, would give a higher contribution to inflation, with all the consequences that would have on the precision of the nowcast.

In the last step, we calculated inflation and other aggregate rates as weighted averages of (recorded and imputed) individual rates, applying weights that the SORS uses in its calculations.

In order to produce the final nowcast based on all information available, we would have to wait almost until the end of the month (the price of fuel is recorded until the last Wednesday). For practical reasons, however, it is often useful to produce the nowcast earlier in the month. Although this means losing some relevant information for the nowcast, this does not have to affect its precision to a large degree.

Thus, if a nowcast is produced around the 20th day in a month, basically the only missing information is the price of fuel for one or two last Wednesdays of the month. In that case we assume that the price will remain unchanged, at the level of the last recorded price. Given that the monthly price of fuel is calculated as an average of weekly prices, a deviation in the last week is not likely to affect the final nowcast significantly.

If a preliminary nowcast is produced in mid-month, the recording period is to a large degree covered, except for two Wednesdays for fuels, one week for fruit and vegetables, and a couple of days for services. The only problem that could occur here is with fruit and vegetables in the peak of the agriculture season, when their prices are the most volatile, which is taken into account when evaluating the risk of the nowcast.

Finally, it should be noted that the process of nowcasting also includes the logical control of nowcast rates. A too high or too low growth rate could indicate a typing error on a web page (a wrong number or a decimal separator). If that indeed is the case, the figure is corrected or eliminated.

5 Nowcasting performance

The described, all-encompassing method of nowcasting has been used in the NBS since March 2022 (partial pilot nowcasts were applied before, during 2021), with a caveat that in the first two months we imputed an expert judgement for non-food, given that at the time we did not have this group sufficiently covered.

Though one year is an insufficient period for a final judgement about this method, we can say that the results so far have been encouraging. In the period between March 2022 and January 2023, our nowcasts were on average fairly close to the official SORS data, suggesting no systematic bias.

Chart 1 **Nowcasted and actual CPI**

Sources: SORS and authors' calculation.

The mean absolute nowcasting error[2] in the analysed period was 0.20 pp, while the median absolute error was 0.13 pp. Out of 11 nowcasts, the deviation from the official figure was: 0.1 pp six times (rounded to one decimal point), 0.2 pp three times, and 0.3 pp once, while the only significant deviation was recorded in July 2022, when inflation based on web prices was 0.6 pp higher than the official SORS figure, as the strong fall in the prices of fruit and vegetables was recorded by the SORS, but not by the websites that monitor these prices.

Here it is important to take into account that these nowcasts were produced in the most turbulent period in the past decade when it comes to inflation, with the average m-o-m inflation amounting to 1.2%, which makes nowcasting particularly challenging, regardless of the method.

Table 1 **Statistics of web prices deviations from SORS data**

	Mean error	Mean absolute error (in pp)	Mean absolute error (in stat. dev.)	Median absolute error (in pp)	Coefficient of correlation of individual products
Inflation	0.03	0.20	0.37	0.13	0.77
Fuel	0.08	0.23	0.05	0.05	1.00
Food excl. f. and v.	0.20	0.45	0.56	0.43	0.56
Fruit and vegetables	0.52	1.57	0.33	1.04	0.82
Administered prices	-0.13	0.15	0.27	0.09	0.65
Non-food	-0.07	0.39	0.87	0.21	0.61

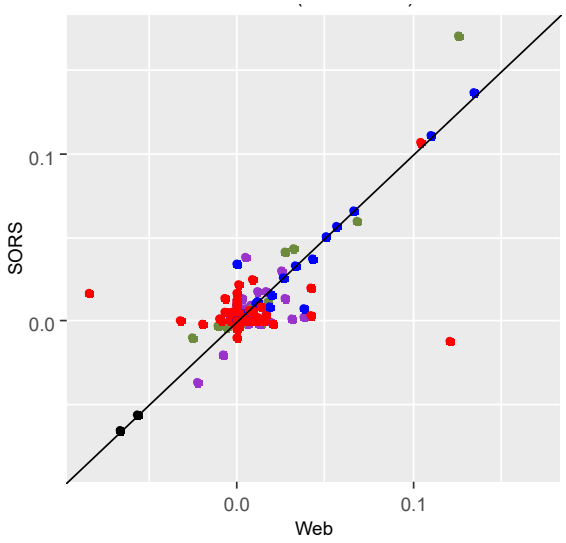
Source: authors' calculation.

Looking at the groups of CPI, we can note that nowcasts are mostly not biased, with the exception of fruit and vegetables, where the average nowcasting error is 0.5 pp, mainly driven by the above-mentioned large deviation in July 2022. For the same reason, the largest mean

[2]Deviation of the nowcast from official data.

absolute error was recorded for fruit and vegetables. This is to some degree expected since this group is characterized by the strongest volatility, even at a weekly level, therefore a relatively small discrepancy from the SORS monitoring period can lead to a significant difference between recorded prices from the two sources (web and SORS).

Chart 2 **Estimated and actual contributions to inflation of individual products from the CPI basket in December 2022** (in pp)



Sources: SORS and authors' calculation.

While on the aggregate level the deviation of the nowcast from SORS figures is rather small, on the individual-item level there can be significant differences. Chart 2 shows a comparison of contributions to inflation of individual products and services (x-axis) with the ones registered by the SORS (y-axis) for December 2022, where the deviation from the 45-degree line shows the nowcasting error for the individual item. These charts for other months in the observed period are presented in the Appendix.

The correlation between nowcast and actual contributions (without the imputed prices) by months moved between 0.5 and 0.9. Aside from fuel, which consists of only two products, looking at the groups, the highest correlation was observed for fruit and vegetables (approx. 0.8), while for other groups it is roughly the same, at 0.6. At a first glance, this is in contradiction with the finding that this group has the highest nowcasting error. This can be explained by the fact that fruit and vegetables have the highest variability within the group (rates sometimes go from -50% to +150%), and also over time. Although the web sources that we use pick up relative changes of prices within the group (leading to high correlation), absolute deviations (nowcast from actual) can still be substantial.

If, when comparing errors, we take into account the variability of groups, we will get a more consistent story (with correlations). Thus, when we put the nowcasting error in relation to standard deviations, the most stable group (non-food) becomes the one with the largest error, while this indicator is much lower for fruit and vegetables and administered prices.

For some of the more complex methods of evaluating nowcasting performance, such as the comparison with the out-of-sample forecasting errors of some of the inflation models, we have to wait for the time period to become sufficiently long for this type of analysis.

6 Conclusion

The purpose of this paper was to provide an overview of the methodology of nowcasting using prices from the web, which has been applied in the NBS since March 2022.

Our methodology, unlike most others described in papers on this topic, is based not only on prices for online shopping, but on every relevant data on prices, including those displayed on the web merely informatively. For that purpose, in addition to the websites of retailers, we also use those that compare the prices of different retailers, as well as websites of government institutions and public companies listing the prices of their services.

Another specific feature of our method is an all-encompassing approach to nowcasting. While methods described in other papers deal with inflation partially, by nowcasting only some components, mainly food, our intention was to cover as many items from the CPI as possible (around 90% at the time of writing this paper), which requires a large number of sources from the internet – 130 websites to be specific. This is done because for forecasting purposes we need nowcasts of various components of inflation, but also because by including more items in the sample we are supposed to achieve a more precise measure of the central tendency of inflation.

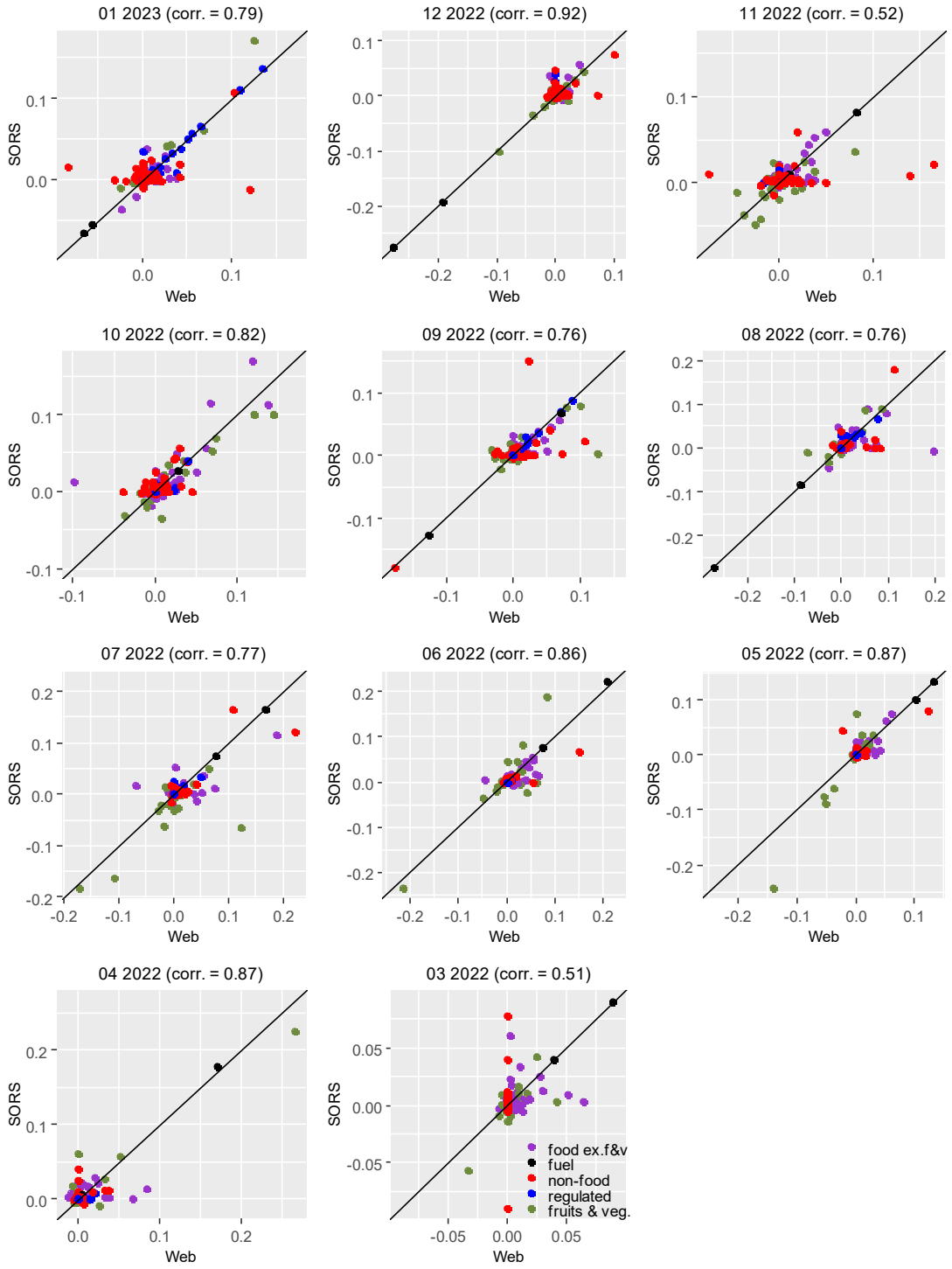
The missing part of the prices is covered with various methods of imputation – with similar products, the price change of the group, previous price change, or zero – depending on the characteristics of price movements of certain groups. By doing so, we avoid the possibility that CPI components with a higher web coverage give a higher contribution to inflation with all the consequences that would have on the precision of the nowcast.

The performance of nowcasts to this point has been encouraging. Nowcast inflation in the observed period on average was at the level of actual inflation (nowcasts are not biased), the mean absolute nowcasting error was 0.20 pp, while the median was 0.13 pp, which is not significant, taking into account that the average m-o-m inflation in the same period was 1.2%. For individual items in some cases there are some observed significant deviations, but the correlation between nowcast and actual rates is relatively high (0.8 on average).

After a year of applying this method, however, it is still too early to draw a final conclusion about its performance. Once the method is applied long enough, future research studies may involve a comparison of web-based nowcasts with out-of-sample forecasting errors from some of the inflation models. In the meantime, as we have done in the past, we will include new sources from the web in the process, and perhaps exclude some of the existing ones, depending on their reliability, availability and timeliness, in an endeavour to improve the precision of our nowcasts.

Appendix

Chart A1 **Estimated (web) and actual (SORS) contributions to inflation of individual products from the CPI basket from March 2022 to January 2023 (Correlation coefficient in parentheses).**



Sources: SORS and authors' calculation.

Bibliography

Aghajanyan, G., Baghdasaryan, T., & Lazyan, G. (2017). The use of Big Data in Central Bank of Armenia.

Bertolotto, M., Cavallo, A., & Rigobon, R. (2014). Using Online Prices to Anticipate Official CPI Inflation. UTokyo Price Project Working Paper Series.

Breton, R., Clews, G., Metcalfe, L., Milliken, N., Payne, C., Winton, J., & Woods, A. (2015). Research indices using web scraped data. Office for National Statistics (ONS).

Carvalho, I. (2020). Nowcasting CPI using online retail prices: Forecasting combination of dynamic factor models.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, 60(2).

Horrigan, M. W. (2013). Big data: A perspective from the BLS. http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/

Jaworski, K. (2021). Measuring food inflation during the COVID-19 pandemic in real time using online data: A case study of Poland. British Food Journal.

Kapetanios, G., & Papailias, F. (2018). Big Data & Macroeconomic Nowcasting Methodological Review.

Macias, P., & Stelmasiak, D. (2019). Food inflation nowcasting with web scraped data. 302.

Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J., & Winton, J. (2018). Tracking and modelling prices using web-scraped price microdata: Towards automated daily consumer price index forecasting. Journal of the Royal Statistical Society: Series A (Statistics in Society).

Soybilgen, B. ş, Kaya, H., & Yazgan, M. E. (2021). Nowcasting Turkish Food Inflation Using Daily Online Prices.

Tissot, B. (2019). The Role of Big Data and Surveys in Measuring and Predicting Inflation.

https://data.stat.gov.rs/Metadata/03_Cene/Html/030103_ESMS_G0_2018_2.html