

Measuring food inflation during the COVID-19 pandemic in real time using online data: a case study of Poland

Krystian Jaworski

Warsaw School of Economics, Warsaw, Poland

Abstract

Purpose – The purpose of this study paper is to focus on developing novel ways to monitor an economy in real time during the COVID-19 pandemic. A fully automated framework is proposed for collecting and analyzing online food prices in Poland. This is important, as the COVID-19 outbreak in Europe in 2020 has led many governments to impose lockdowns that have prevented manual price data collection from food outlets. The study primarily addresses whether food price inflation can be accurately measured during the pandemic using only a laptop and Internet connection, without needing to rely on official statistics.

Design/methodology/approach – The big data approach was adopted to track food price inflation in Poland. Using the web-scraping technique, daily price information about individual food and non-alcoholic beverage products sold in online stores was gathered.

Findings – Based on raw online data, reliable estimates of monthly and annual food inflation were provided about 30 days before final official indexes were published.

Originality/value – This is the first paper to focus on measuring inflation in real time during the COVID-19 pandemic. Monthly and annual food price inflation are estimated in real time and updated daily, thereby improving previous forecasting solutions with weekly or monthly indicators. Using daily frequency price data deepens understanding of price developments and enables more timely detection of inflation trends, both of which are useful for policymakers and market participants. This study also provides a review of crucial issues regarding inflation that emerged during the COVID-19 pandemic.

Keywords Nowcasting, Big data, Web scraping, Prices

Paper type Research paper

1. Introduction

The COVID-19 pandemic has highlighted the importance of developing access to data that allow tracking of an economic situation at a much higher frequency than traditional monthly or quarterly indicators. As the economic situation under the pandemic was changing rapidly and subject to significant uncertainty, researchers started using a variety of high-frequency indicators, such as mobile phone data, traffic density, web searches, electricity consumption and credit card transactions, rather than “traditional” economic indicators, to take the pulse of an economy almost in real time (Baker *et al.*, 2020; Carvalho *et al.*, 2020; Cicala, 2020; Kuchler *et al.*, 2020). Real-time monitoring of economic activity and price evolution is essential

JEL Classification — C55, C81, E37

© Krystian Jaworski. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licences/by/4.0/legalcode>

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations of interest: none

Funding: This work was supported by the National Science Centre, Poland – grant number 2016/23/N/HS4/02054.



but is also regarded as a challenge for econometricians. In fact, the economic literature has reached a broad consensus that “forecasting inflation is hard” (Marsilli, 2017).

This study focuses on developing new ways to monitor the economy in real time. The primary focus is on the Consumer Price Index (CPI). The CPI is an important macroeconomic indicator. Its importance stems from its wide usage. It is used to monitor price changes; it impacts government revenues and expenditures and private-sector wage compensation bills; and it is also among the statistical indicators that influence financial markets, particularly interest and exchange rates. Therefore, it is critical that the CPI is based on high-quality data and is measured as timely as possible. Much effort has been made to continually improve the quality and comparability of CPIs within and between countries, according to the well-established international guidelines and methodologies (ILO, IMF, OECD, UNECE, Eurostat, and World Bank, 2004; Graf, 2016). Food prices are crucial when measuring the CPI, owing to their significant weight in the inflation basket as well as their high volatility compared to other CPI categories (National Bank of Poland, 2016). Both characteristics may increase the volatility of the CPI.

In this study, a fully automated framework is proposed for collecting and analyzing online food prices in Poland. This is an important undertaking, as the COVID-19 pandemic has led governments to impose several measures, such as restricting people’s movements and closing outlets, with both direct and indirect effects on household consumption and, thus, CPI. In particular, the situation has negatively affected the collection of prices needed to compile CPI as a measure of inflation. The main research question of this study addresses whether food price inflation can be accurately measured during the pandemic using only a laptop and an Internet connection, without needing to rely on official statistics. This study focuses only on food prices, as tracking the entire consumption basket would be laborious. A thorough search of the relevant literature indicates that this is the first study to measure inflation in real time during the COVID-19 pandemic. Existing studies mainly focus on providing food inflation estimates based on monthly data and do not discuss how the pandemic has affected prices.

This study makes the following contributions. It provides a framework enabling computation of the Polish food CPI with greater timeliness (real-time) and frequency (daily), improved coverage (large sample of products) and more detailed information (single products). It also offers important contributions to three strands of the literature. First, the rapidly growing areas of research into monitoring economies in real time during the COVID-19 pandemic, by providing a way to accurately measure food inflation using online prices. Second, it discusses the policy implications of how web scraping can affect the compilation of official inflation indexes by the Central Statistical Office (CSO) and act as a low-cost data collection method for food price research. Third, it contributes to the literature stream associated with price nowcasting. It provides an original framework to enable estimation of food inflation in real time, a feat that has eluded previous studies. This study also improves on previous research by considering a longer time span of 5 years, compared to the typical period of 1 year or less.

The rest of the paper is organized as follows. Section 2 presents a brief literature review. Section 3 presents the data and methodology. Section 4 discusses a variety of critical issues regarding inflation that have emerged during the COVID-19 pandemic. Section 5 reports the main results of the empirical study. Finally, Section 6 concludes and suggests avenues for further research.

2. Literature review

In recent years, many national statistical offices (NSOs) have experimented with using online data in official CPIs, including the US Bureau of Labor Statistics (Horrigan, 2013), the UK Office of National Statistics (Breton *et al.*, 2015), Statistics Netherlands (Griffioen *et al.*, 2014), Statistics New Zealand (Krsinich, 2015) and Statistics Norway (Nygaard, 2015). Many NSOs have started to incorporate big data for their official statistics (United Nations Statistical

Commission, 2014), such as the use of automatic web scraping of food prices from online stores. Web scraping gathers and copies data from the web, typically into a central local database or spreadsheet, for later retrieval or analysis. In this case, the term refers to automated processes implemented using an algorithm (bot) on supermarkets' websites (e-stores).

NSOs' experience shows that the use of online data to produce food CPI statistics offers multiple advantages over traditional price data collection techniques (Cavallo, 2013; Breton *et al.*, 2015; Cavallo and Rigobon, 2016). The cost of data collection is lower; online data include detailed information of all goods sold by the sampled retailers and not just selected products; there are no gaps in online data: prices are recorded from the first day a product is offered to consumers until the day it is discontinued from the store; and online data can be collected remotely and are available in real time.

Hillen (2019) stated that in agricultural and food economics, web scraping has received little attention as a data collection technique. Indeed, the literature on the use of online data to measure and forecast inflation is limited, albeit rapidly growing. One of the first studies was conducted by Cavallo (2013). For the Latin American countries of Argentina, Brazil, Chile, Colombia and Venezuela, he showed that online prices could be effectively used as an alternative source of price information to construct price indexes. Later, Cavallo (2017) simultaneously collected prices on websites and physical stores for over 24,000 products in 56 of the largest retailers in 10 countries (Argentina, Australia, Brazil, Canada, China, Germany, Japan, South Africa, the UK and the USA). He revealed a high degree of similarity in price levels as well as the frequency and magnitude of price changes between online and offline prices. For NSOs, these results imply that web scraping can be used effectively as an alternative data collection technology to obtain the same prices found offline. Other studies developed the topic of using online prices to forecast inflation even further. The Central Bank of Armenia began collecting big data from 2016 to generate flash estimates of the CPI (Aghajanyan *et al.*, 2017). Hull *et al.* (2017) presented favorable forecasting results for the prices of fruits and vegetables in Sweden using online data. Mustapa *et al.* (2019) evaluated the dependability of online data prices to forecast the inflation of vegetables and fish in Malaysia with promising results. Uriarte *et al.* (2019) implemented a web scraping technique for monitoring prices in a mid-urban area in Argentina and found that web scraping combined with big data techniques enabled estimation of more individualized and efficient metrics, whose quality was comparable to official statistics. Aparicio and Cavallo (2021) and Cavallo and Rigobon (2016) confirmed this result and stated that online-based price indexes were comparable to the traditional CPI despite methodological differences in multiple countries. Aparicio and Bertolotto (2020) developed online price indexes as a useful predictor of the inflation rate in many economies (Australia, Canada, France, Germany, Greece, Ireland, Italy, the Netherlands, the UK and the USA) with a 1-month horizon. Finally, a thorough literature search yielded only one published article on the use of online prices to produce food CPI statistics in Poland. Macias and Stelmasiak (2019) assessed the forecasting accuracy of online prices aggregate alone and within simple linear distributed-lag models and their combinations. Their results showed that using online price data leads to lower forecasting errors than using autoregressive moving average models. Different to the current analysis, their study focused on *forecasting* CPI using online data, and not on *measuring* it. They provided food CPI estimates once a month, whereas this study focuses on updating monthly food inflation estimates in real time.

3. Materials and methods

This section outlines the data-gathering framework (i.e. web scraping) used in this study to prepare a unique dataset of online prices for analysis. Subsequently, the algorithm for calculating the estimates of food inflation using these prices is presented.

3.1 Data and description of web-scraping technique

The prices of food products sold online are not openly available in a pre-prepared dataset. To estimate the inflation of food products in Poland using online prices, these prices had to be gathered by the researchers. To this end, a web scraping technique was applied on the website of one of Poland's major supermarket chains. Hillen (2019) defined web scraping as an automated process of accessing web documents and downloading specific pre-defined information, such as prices, then transforming and saving it in a structured format. A combination of programming languages was used to build a web scraping script, which, in principle, imitates a human web user, navigating websites and extracting the pre-defined information. The automated procedure developed in this study scans the code of the publicly available website of the supermarket chain every day, identifies relevant pieces of information (e.g. product name, price, size and unique ID) and stores these data in a file.

The web scraping algorithm has three steps. First, at a fixed time each day, the software detects all the web pages of individual products available on the retailer's website. These individual pages contain information about products and prices. These pages are individually retrieved every day. Second, the underlying code of the websites is analyzed to locate each piece of relevant information. Special characters in the code identify the start and end of each variable placed by the website's programmers to give the website a particular look. Specifically, the algorithm explores the hypertext markup language format in web pages and extracts and stores the relevant portion of the code. Third, the software stores the scraped information in a database that contains one record per product per day. These variables include the product's price, date, category information and an indicator for whether the item is on sale or not. Online prices include value-added tax and exclude transportation costs to match the prices used in the traditional CPI as closely as possible.

This web scraping procedure collected prices every day from July 2015 to August 2020. This study's database contains the price history of about 20,000 unique food products. Not every product is available every day: prices are recorded from the first day a product is offered to consumers until the day its sale is discontinued. Some information can be missing owing to stock shortages on a given day, seasonal product offers and technical problems (on part of either the supermarket chain or this study). Moreover, during this study, the supermarket chain's website changed several times, necessitating changes in the web crawlers (which have to be specially developed for each website) used in this study to adjust their underlying code. Each redesign of the web crawler takes a day or so, which resulted in some missing data. Nonetheless, there are few missing observations over the five-year study period and these should not distort the results of the analysis.

The prices of all 20,000 products were not used to calculate the online food CPI. The calculation is based on a list of representative items for each product group (i.e. a subcategory at the lowest aggregation level of the weighting system) in the "food and non-alcoholic beverages" category. Detailed information was obtained from the CSO on the kinds of products considered when collecting prices (see the [Appendix](#)). Individual products were selected that represent price changes in each of the elementary groups in the Classification of Individual Consumption According to Purpose classification. In total, 205 individual products were included that cover all 86 elementary groups representing food prices. The main objective was to choose products that were available for most of the study period.

Only the price data collected from the supermarket chain's website was used to calculate food price inflation. There are many other kinds of retail food outlets, such as convenience shops and farmers' markets but these do not offer online access to the prices of their products, which renders web scraping ineffective in gathering their data. The market share of outlets other than big retailers has been rapidly decreasing over the years and now accounts for only a small proportion of the total turnover in Poland. This study analyzes price changes (i.e. inflation) rather than price levels; price changes should follow similar trends in both large and

small outlets. This view is supported by [Kouvavas et al. \(2020\)](#). More importantly, explicit information was found on the supermarket chain's website that it strives to set the same prices in both online and offline trade. This means that the prices gathered are representative of both traditional and online transactions. Existing studies on the similarity of online and offline prices conducted in other countries confirmed the usefulness of web-scraped data to track overall price movements, both in terms of levels and trend dynamics. Following the international experience, one can conclude that web-scraped prices are representative of the evolution of prices captured by official CSO inflation statistics. The high accuracy of this study's approach ([Section 5.1](#)) compared to the official inflation statistics confirms this assessment for Poland.

3.2 Calculation of online food price inflation at monthly frequency

To construct the CPI of different food categories, the standard CSO methodology ([Central Statistical Office, 2019](#)) was used with some modifications to benefit from the advantages of using online price data (larger number of individual products covered, more frequent collection of prices). CPIs are calculated in steps. First, CPI indexes are calculated for each of the 86 elementary groups of food products presented in the [Appendix](#). Subsequently, all of the produced CPI sub-indexes are used to calculate the aggregated CPIs for every food subcategory.

First, to calculate the price indexes for groups of goods at the lowest aggregation levels of the weighting system (86 elementary groups in the “food and non-alcoholic beverages” category), the geometric mean of the daily prices of individual products (x_{ab}^c) is calculated for the 86 elementary food product groups. The price index p_{it} for the i th elementary food group recorded in month t is calculated as follows:

$$p_{it} = N \sqrt[n]{\underbrace{x_{1t}^1 \cdot x_{2t}^1 \cdot \dots \cdot x_{\tau t}^1}_{\text{daily prices for product 1}} \cdot \dots \cdot \underbrace{x_{1t}^n \cdot x_{2t}^n \cdot \dots \cdot x_{\tau t}^n}_{\text{daily prices for product } n}}, \quad (1)$$

where x_{at}^b is the price recorded for product b belonging to the elementary group i on the a th day of the month t ; n indicates the number of products belonging to elementary group i ; τ is the number of daily price quotations in month t (usually close to 30); and $N = n \cdot \tau$ represents the total number of individual prices recorded for all n products during month t .

The price indexes for elementary groups (p_{it}) are aggregated with weights to calculate price indexes at the higher aggregation levels, up to the total CPI of the “food and non-alcoholic beverages” category. The weighting system is based on the expenditure of households on purchasing consumer goods in the year preceding the reference period ($w_{i, t-12}$). Because data on consumer expenditure are derived from the household budget survey conducted by the CSO, the CSO's official weighting system is used.

Price indexes at the higher aggregation levels are calculated according to the Laspeyres' formula:

$$p_{kt} = \sum_i p_{it} \cdot w_{i, t-12}, \quad (2)$$

where p_{kt} represents the CPI of food subcategory k ; p_{it} is the individual price index for elementary group i belonging to k subcategory in the month t ; and $w_{i, t-12}$ is the associated weight.

Based on the price indexes, the relative price changes are calculated compared to the previous month for subcategory k , that is, month-on-month (MoM) inflation as given in the following equation:

$$\pi_{kt}^M = \frac{p_{kt}}{p_{k,t-1}} \cdot 100, \quad (3)$$

and price changes compared to the same period (i.e. month) the year before, that is, year-on-year (YoY) inflation as given in the following equation:

$$\pi_{kt}^Y = \frac{p_{kt}}{p_{k,t-12}} \cdot 100. \quad (4)$$

The inflation for the 86 elementary groups is calculated in the same way as the 10 main subcategories by substituting k with i in [Equations \(3\) and \(4\)](#). The results obtained by following the steps above are inflation estimates for each of the 86 elementary groups and an additional 10 main food subcategories at monthly frequency. The results obtained using online prices can then be compared with the official data (also available at monthly frequency) calculated by the CSO at various levels of aggregation.

3.3 Real-time estimate of online food price inflation

The data collection method used in this study enables the procurement of information about retail food prices in Poland at a daily frequency and to calculate changes in prices (i.e. inflation) every day. However, such results would be misleading at best. First, owing to sudden price changes, such as those caused by price promotions, daily food price inflation is subject to significant volatility and would contain little useful information. Second, one would not be able to say if the results are in line with official statistics, because such official daily measures do not exist. Therefore, although interesting, calculating how prices change every day would not be of much practical use.

Instead, to take advantage of the unique nature of the data used in this study, the food MoM and YoY monthly inflation estimates are presented, updated every day as new data become available. These are real-time estimates of MoM and YoY online food price inflation. This approach means that these estimates can be compared with the official statistics. The leading characteristics of the real-time estimate of online food CPIs are investigated by observing the number of days before the end of the reference period (i.e. the current month), making it possible to produce an accurate nowcast of the official monthly food CPI.

To obtain the estimate of monthly food inflation in daily increments, [Equation \(1\)](#) is modified to calculate the price index p_{it} using only the first d days of the month:

$$p_{it}^d = N \sqrt{\underbrace{x_{1t}^1 \cdot x_{2t}^1 \cdot \dots \cdot x_{\tau t}^1}_{\text{daily prices for product 1}} \cdot \dots \cdot \underbrace{x_{1t}^n \cdot x_{2t}^n \cdot \dots \cdot x_{\tau t}^n}_{\text{daily prices for product 1}}}, \text{ where } \tau \leq d. \quad (5)$$

For example, on the one hand, p_{it}^1 means that the price index is calculated using only the daily prices recorded on the first day of the month. On the other hand, p_{it}^{31} means that all the price points from the whole month are used. p_{it}^{31} essentially equals p_{it} in [Equation \(1\)](#). N is adjusted to correspond to the number of total price points taken into consideration.

Subsequently, the price indexes at higher aggregation levels are calculated according to Laspeyres' formula in the same way as for monthly data [Equation \(2\)](#) as weighted averages:

$$\hat{p}_{kt}^d = \sum_i \hat{p}_{it}^d \cdot w_{i,t-12}. \quad (6)$$

Once a way to update the average level of prices everyday during a month is established, food inflation can be calculated in two ways. The first method compares full information, that is,

the average of all daily observations from the base period (i.e. the last month or year) with all information available on a given day for the current month t , that is, d observations. Therefore, the MoM inflation (π_{kt}^M) calculated on the d^{th} day of the month t for subcategory k using the first method (indicated by superscript 1) is as follows:

$$\pi_{kt}^{M,1,d} = \frac{p_{kt}^d}{p_{k\ t-1}^{31}} \cdot 100 = \frac{p_{kt}^d}{p_{k\ t-1}} \cdot 100. \quad (7)$$

YoY inflation is calculated in the same way, as follows:

$$\pi_{kt}^{Y,1,d} = \frac{p_{kt}^d}{p_{k\ t-12}^{31}} \cdot 100 = \frac{p_{kt}^d}{p_{k\ t-12}} \cdot 100. \quad (8)$$

The second method compares the prices average over the d^{th} first days in the current month with the average over the same days in the base period (i.e. the last month or year).

Using the second method (indicated by superscript 2), the MoM inflation estimated on d^{th} is calculated as follows:

$$\pi_{kt}^{M,2,d} = \frac{p_{kt}^d}{\bar{p}_{k\ t-1}^d} \cdot 100, \quad (9)$$

and the YoY inflation is calculated as follows:

$$\pi_{kt}^{Y,2,d} = \frac{p_{kt}^d}{\bar{p}_{k\ t-12}^d} \cdot 100. \quad (10)$$

If there is a strong and persistent pattern of intra-month price changes, the two methods would lead to different results. This could occur, for example, when prices significantly decrease at the beginning of the month owing to a price promotion and then revert to higher levels at the end of the promotion.

4. Critical issues regarding inflation during the COVID-19 pandemic

This section highlights crucial complexities regarding inflation during the COVID-19 pandemic that are linked to the study method and that have important theoretical implications.

4.1 Overestimation of food price inflation amid inflation expectations

[Akter \(2020\)](#) reported a sharp increase in food prices at the beginning of the pandemic. This price movement appeared simultaneously in multiple economies and the scale of the price hike was positively correlated with the severity of stay-at-home restrictions imposed by the respective governments. She estimated that, on average, the restrictions were conducive to a 1% increase in overall food inflation. This study argues that the scale of food price increase could be *overestimated* in official CSO inflation statistics. Such a view is consistent with [Ebrahimi et al. \(2020\)](#). They showed evidence from advanced and emerging market economies of an increase in food prices, although there was little sign of inflation when considering broader indexes. The authors emphasized that the prices of meat, dairy and canned/frozen fruit and vegetables had spiked early on after the breakout of the pandemic. In March 2020, at the beginning of the pandemic, consumers engaged in panic buying, leading to empty shelves in grocery shops and supermarkets. According to the CSO procedure, if a price collector visiting the outlet cannot record the price of a given representative item, he or she should log the price of a similar product. As a result of household stockpiling activities, most

cheaper products were sold out in grocery shops, while more expensive alternatives were available. For example, [Jaravel and O'Connell \(2020\)](#) documented a fall of around 8% in the number of unique products being available for purchase at the beginning of the pandemic for the UK. This phenomenon is considered to have led to a massive “lack of matching” problem, declining consistency of the inflation time series, and overestimation of actual food inflation at the beginning of the pandemic, as the price collectors were forced to record the prices of alternative – more expensive – sets of items. [Diewert and Fox \(2020\)](#) suggested that web scraping can solve this problem. Indeed, this study method of calculating food inflation is robust to the impact of missing products, as it uses the same sample of items each month. As indicated in [Section 5.3](#), this study's estimate of the MoM food inflation in March 2020 was lower by almost 0.9 percentage points than the official CSO figure (i.e. 0.8%). This suggests that the CSO in Poland overestimated the actual price increase at the beginning of the pandemic owing to missing products. It is believed that this phenomenon could be observed beyond Poland, which has important implications for national statistical offices and researchers interested in the effect of COVID-19 on retail prices.

The abovementioned difficulties in measuring food inflation have wide-ranging implications for the process of forming inflation expectations. [D'Acunto et al. \(2019\)](#) showed that consumers rely on price changes they face in their daily lives while grocery shopping (mostly food products) to form aggregate inflation expectations. Specifically, the frequency and size of price changes, rather than their expenditure share, matter for individuals' inflation expectations. The disconnect between the actual evolution of food prices and food inflation measured by the CSO may produce erroneous conclusions about household inflation expectations and policies based on these expectations may lead to systematic mistakes.

The pandemic led to an immediate and substantial increase in inflation uncertainty. [Armantier et al. \(2020\)](#) observed a sharp polarization in inflation beliefs, with a substantial share of respondents initially believing that the pandemic was going to produce high inflation, and another proportion of respondents believing that the pandemic was going to cause low inflation or even deflation. This result has important policy implications. Indeed, [Kumar et al. \(2015\)](#) considered high-inflation uncertainty to be one of the metrics indicating un-anchored inflation expectations.

Food prices and their correct measurement play an important role in shaping these expectations. [Clark and Davig \(2009\)](#) showed that shocks to food price inflation generate relatively large and persistent responses of both short-term and long-term inflation expectations. Close monitoring of these prices and such expectations is warranted, since they may signal a risk of inflation expectations un-anchoring. [Cavallo \(2020\)](#) emphasized the importance of measuring inflation and expectations and identified the pricing impact of supply shocks as an important area for future research on COVID-19 inflation dynamics. Monitoring how inflation expectations evolve during a crisis is important for anticipating the effectiveness of the transmission of monetary and fiscal policy interventions to the real economy.

4.2 Alternative inflation index

Another problem that became evident during the pandemic is the issue of the weighing system used for calculating CPI. The CPI sub-indexes are aggregated using weights reflecting the previous year's household consumption expenditure patterns. These weights are updated at the beginning of each year and kept constant throughout the year. [Eurostat \(2020\)](#) has issued a list of guidelines for NSOs in the EU to maintain the highest possible quality of CPI statistics during the pandemic. One guideline stipulates that the sub-index weights used to compile the CPI should not be changed during the year, suspending the

standard practice. Thus, the impact of the COVID-19 pandemic on expenditures did not affect the weights during 2020.

However, because of the high demand and low supply and significant shifts in the expenditure distribution, customers' current purchasing activity looks very different compared to the same period last year. Furthermore, some services (hotels, airline travel, etc.) are unavailable owing to enforced lockdown measures. Therefore, significant discrepancies are emerging between the official measure of inflation and economic reality, that is, the price changes of actual consumer baskets.

It is impractical for the CSO to adjust expenditure weights for two main reasons. First, it would not be consistent with the fixed basket concept on which consumer price statistics are based. Second, consistency between countries and overtime must be maintained to enable yearly comparisons. However, there is value in understanding how shifts in the expenditure distribution affect the measures of price change. The UK's [Office for National Statistics \(2020\)](#) is attempting to track the CPI shopping basket at greater than yearly frequency.

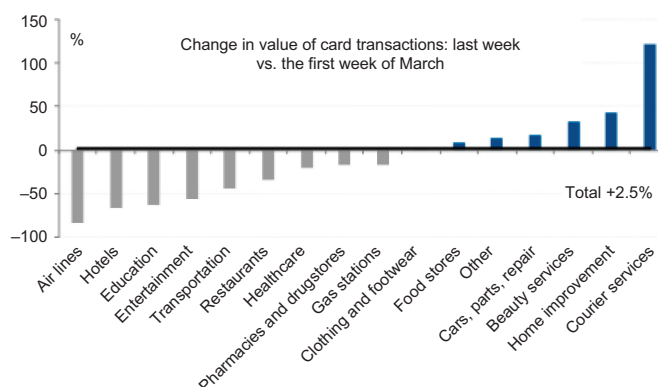
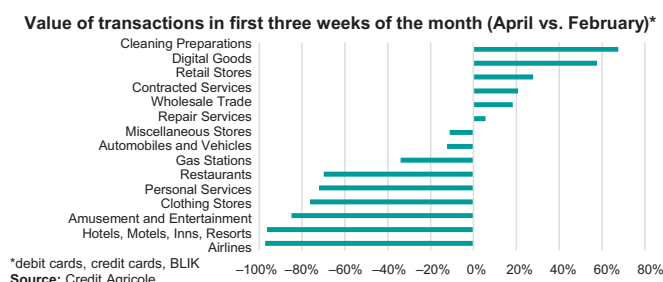
To adopt this approach, Poland must first obtain a reliable source of current expenditure data, which can be obtained via cooperation between the CSO and commercial banks, which could share information about their clients' transactions. There have been attempts to track the structure of households' consumption via bank transactions in Spain ([Carvalho et al., 2020](#)). This approach is also possible in Poland. Some commercial banks in Poland already publish aggregated information about their clients' transactions, either in public reports or on Twitter (see the examples in [Figure 1](#)). This information evidently offers insight into significant changes in the expenditure structure during the COVID-19 pandemic.

Accessing information about households' transactions would allow reweighting of the inflation basket in real time, allowing improved measurement of inflation in view of a significant shock to household consumption patterns. Admittedly, inflation indexes that use weights based on transactions are not consistent with the fixed basket concept on which consumer price statistics are based, and thus, cannot be incorporated as part of official time series. However, they could be a useful supplementary measure alongside the official CPI and could more accurately reflect prices changes during extraordinary times, which would be useful for policymakers, including the central bank, interested in correctly tracking inflation. This is an important avenue for further research.

Moreover, accurately measuring food inflation will be of utmost importance, as the weight of the "food and non-alcoholic beverages" category doubled during the lockdown period in some countries and constituted almost half of total consumer expenditures ([Huynh et al., 2020](#)). Consumption of food items has increased because households are spending more time at home (effectively switching away from food served in restaurants and bars).

4.3 Implications for long-term inflation forecasting

Based on an extensive literature review, [Knotek and Zaman \(2017\)](#) reported that inflation is difficult to forecast accurately using econometric models. These difficulties extend to contemporaneous forecasting (nowcasting) of the inflation rate in the current month or quarter. This study's approach, which accurately measures food inflation in real time, provides valid nowcasts. Moreover, improving inflation nowcasting is not an end in itself. [Del Negro and Schorfheide \(2013\)](#) and [Faust and Wright \(2013\)](#) emphasized that inflation forecasts at longer horizons benefit by using more accurate conditioning via nowcasts. [Modugno \(2013\)](#) showed that higher frequency data, which are more timely than lower frequency data, are necessary for more accurate inflation forecasts. [Woodford \(2003\)](#) added that timely update of macroeconomic projections is essential for modern monetary policy based on market expectations. Thus, this study's model, which produces accurate nowcasts of food inflation, has broad applications for academic economists and professional forecasters.



Source: @PKO_Research

Source(s): Upper chart – MACROmap, *Weekly Economic Comentary of Credit Agricole*, https://static.credit-agricole.pl/asset/m/a/c/macromap-20200427_18539.pdf [accessed: 08.06.2020]. Lower chart – Twitter profile of PKO Research, https://twitter.com/PKO_Research/status/1264949419092385792/photo/1 [accessed: 08.06.2020]

Figure 1.
Examples of data
about transactions of
commercial banks'
clients

4.4 Price-setting behavior of firms

This research offers important insights into the price-setting mechanisms used by supermarkets during the onset of the COVID-19 pandemic. Financial distortions create an incentive for firms to raise prices in response to adverse financial or demand shocks (Gilchrist *et al.*, 2017). This reaction reflects the firms' decision to preserve internal liquidity. However, the supermarket chain analyzed in this study did not increase prices at the beginning of the pandemic – food prices declined by 0.1% between February and March 2020 (Section 5.3).

The COVID-19 pandemic shock is in many ways similar to the shock of a natural disaster. Cowen (2017) argued that the reluctance to raise prices in the aftermath of Hurricane Sandy was especially pronounced for nationally branded stores. He explained that a high reputation is associated with being a national brand (vs a local outlet). A local entrepreneur might not care much if consumers are concerned about price gouging, but major companies fear damage to their national reputations. This is a possible explanation for the lack of significant price increases observed in the supermarket chain used in this analysis at the beginning of the pandemic.

Because of the high granularity of the data, the day-by-day relationship between the COVID-19 impact and price-setting behavior could be observed. Although, overall food prices did not change, the scale of promotions significantly decreased. According to calculations performed on this study's dataset, the number of products with discounted prices declined by

31.3% between the end of February and mid-March, when the state of epidemic was officially declared in Poland. The promotions were partially reintroduced before the end of the month: the number of discounted products in this study's dataset was 20.3% higher than that in mid-March.

On the one hand, the price-setting behavior of the supermarket chain indicates is favorable for its reputation, as it shows a lack of price gouging (the price recorded on each day has the same weight when calculating inflation). On the other hand, it raises profit, owing to the lower number of promotions during the period of panic buying (i.e. increased turnover). Other firms (i.e. farmers markets and grocery stores), by having access to the real-time pricing data of a major supermarket chain, would be able to properly adjust their own prices to maximize profits. This contributes to the limited studies on the firm-level impact of COVID-19 (cf. [Cabral and Xu, 2020](#)).

4.5 Future of web-scraped data in measuring food prices

During the COVID-19 pandemic, customers shopped more online, and thus, online prices could reflect price trends more accurately than those collected from traditional shops. Even after the pandemic ends, analyzing online food prices will likely gain importance owing to the increasing popularity of the Internet as a sales channel. Online food is purchased by 28% of Internet users ([Mobile Institute, 2017](#)). [E-Grocery \(2019\)](#) reported that 16% of respondents in Poland regularly buy food online. The potential for developing online food trade is enormous – penetration of this category in Poland is estimated at just 0.7% of the market for fast-moving consumer goods while, according to Euromonitor International data, the average annual growth rate of e-grocery sales is 15–20%. It is believed that in the future, most price collection will still occur in traditional shops, but NSOs will likely intensify the use of the big data approach, including web scraping. A larger range of products, more frequent recording and more coverage of the reporting month are the three main ways in which the web-scraped data will affect the compilation of inflation statistics. This prediction is consistent with trends outlined by the [European Central Bank \(2019\)](#).

5. Results

This section examines the accuracy of this study's estimates of food inflation compared to the official statistics and then discusses their leading characteristics. Finally, their application during the COVID-19 pandemic is analyzed.

5.1 Accuracy of monthly price food inflation estimates

Using the algorithm outlined in [Section 3](#), the price indexes for all 86 elementary groups can be calculated. To save space, the results of this study's calculations for overall food inflation (i.e. "food and non-alcoholic beverages") and the 10 most commonly used subcategories are presented.

The accuracy of the estimates is measured using two standard statistics: the root mean square error (RMSE, [Equation \(11\)](#)) and the correlation coefficient between the official measure and this study's estimate of food inflation ([Equation \(12\)](#)):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\pi}_{kt} - \pi_{kt})^2}, \quad (11)$$

where $\hat{\pi}_{kt}$ is the estimated food price inflation for subcategory k using this study's approach and π_{kt} represents the food price inflation for the corresponding subcategory officially provided by the CSO. The lower this measure, the higher the accuracy of the estimates.

The second way to assess the quality of the estimates is to compare their correlation with the official measure:

$$r = \text{corr}(\hat{\pi}_{kt}, \pi_{kt}). \quad (12)$$

The superscripts for $\hat{\pi}_{kt}$ and π_{kt} have been omitted in Equations (11) and (12) to maintain their universality when referring to the assessment of the accuracy of this study's estimates. When discussing the accuracy of this method, symbols $\hat{\pi}_{kt}$ and π_{kt} represent the version of the food inflation estimate (MoM or YoY) and the frequency of estimates (monthly or daily) according to the approach currently under evaluation.

First, the RMSE and correlation statistics for both the MoM and YoY versions of food inflation are calculated. Using this study's notation, $\hat{\pi}_{kt}^M$ is formally compared with π_{kt}^M , and $\hat{\pi}_{kt}^Y$ with π_{kt}^Y . Table 1 presents the results.

According to Table 1, the RMSE criterion shows that this study method provides accurate estimates of food price inflation in all cases. In particular, the RMSE is markedly lower for the broad "food and non-alcoholic beverages" category compared to the 10 subcategories. This means that errors corresponding to levels of aggregation effectively cancel each other out. Such a phenomenon appears for both MoM and YoY food inflation. For example, in the case of MoM headline food inflation, the RMSE is 0.006, whereas, for the "bread and cereals" subcategory, it is at 0.013; for "meat," it is 0.017; and for "fruit," it is 0.041.

Turning to the results obtained from the correlation analysis, for the MoM headline food inflation, the correlation with the official CSO figures is 0.71, which indicates that this method accurately estimates the direction of monthly price changes. In the case of YoY food inflation, the accuracy is even higher. The correlation coefficient is 0.96, indicating almost perfect information about the direction of price changes.

The RMSE and correlation results mentioned above indicate that this study method provides accurate estimates of food price inflation using both the MoM and YoY approaches. Not only can it procure information for the headline indicator but also subcategories at lower aggregation levels. It is worth noting that MoM changes of estimated CPI are not well correlated with the official CPI series for some of the 10 subcategories (e.g. "non-alcoholic beverages," "food products not elsewhere classified" and "fish and seafood"), while this issue does not occur in the case of YoY changes. This issue has also been reported by other researchers (e.g. Cavallo, 2013). The cause of these deviations is that a retailer can adjust its

Category	COICOP code	MoM changes of food CPI		YoY changes of food CPI	
		$\hat{\pi}_{kt}^M$ vs π_{kt}^M	Corr	$\hat{\pi}_{kt}^Y$ vs π_{kt}^Y	Corr
Food and non-alcoholic beverages	01	0.006	0.71	0.024	0.96
Bread and cereals	01.1.1	0.013	0.42	0.040	0.90
Meat	01.1.2	0.017	0.39	0.047	0.95
Fish and seafood	01.1.3	0.017	0.17	0.034	0.84
Milk, cheese and eggs	01.1.4	0.011	0.68	0.041	0.93
Oils and fats	01.1.5	0.017	0.65	0.021	0.97
Fruits	01.1.6	0.041	0.58	0.039	0.94
Vegetables	01.1.7	0.031	0.80	0.064	0.82
Sugar, jam, honey, chocolate and confectionery	01.1.8	0.019	0.34	0.047	0.36
Food products not elsewhere classified	01.1.9	0.016	0.34	0.062	0.62
Non-alcoholic beverages	01.2	0.026	0.03	0.010	0.60

Note(s): The table presents RMSE and correlation coefficients for main categories in line with Equations (11) and (12)

Table 1.
RMSE and correlation
coefficients for the
main food inflation
categories

prices slower or faster than those of the entire economy in the short run. In the longer run, such discrepancies with official data tend to be corrected, improving the correlation in the case of YoY food inflation.

One should remember that the primary aim of this method is to measure, not forecast, food price inflation. Therefore, the small errors that occur do not indicate weakness in this approach. They mean only that the food prices observed on the Internet evolve slightly differently from those observed in traditional shops. This is an important conclusion that should be taken into consideration when measuring food inflation in periods when an increasing share of food purchases occur via the Internet. Such a tendency was especially pronounced during the COVID-19 pandemic in 2020.

5.2 Leading characteristics of real-time online food price inflation

The results suggest that accurate estimates of food inflation can be obtained on the last day of the reference period. This is a helpful result, as the final official inflation data are usually published about 2 weeks after the end of the reference period (in March 2018, the Polish CSO started releasing “flash” food price inflation estimates on the last working day of the reference period, revising them 2 weeks later). Therefore, this study method provides a timelier way to analyze inflationary trends. Can one estimate monthly food inflation earlier (i.e. before the end of the reference period) without sacrificing too much accuracy?

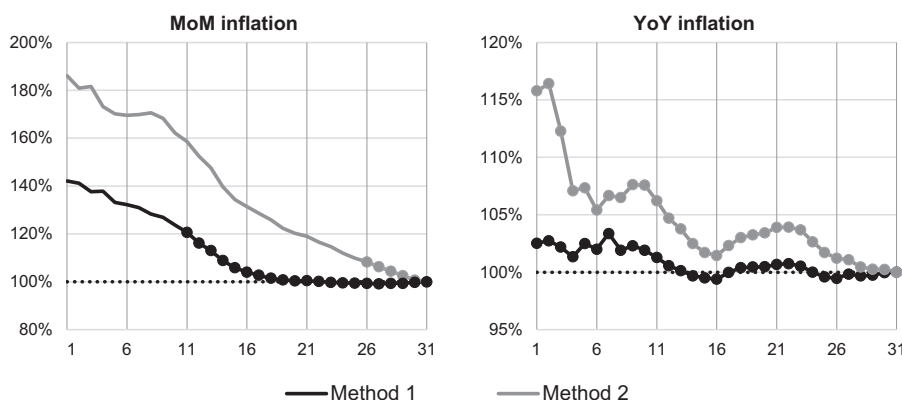
To answer this question, the two methods outlined in [Section 3.3](#) are used. Method 1 is captured by [Equations \(7\) and \(8\)](#) and Method 2 by [Equations \(9\) and \(10\)](#).

[Equations \(11\) and \(12\)](#) are used to evaluate the accuracy of the real-time estimates of the monthly and annual food inflation, because these daily estimates correspond to the official monthly or annual food inflation figures published by the CSO (i.e. π_{kt}^M or π_{kt}^Y). Formally, for Method 1, $\hat{\pi}_{kt}^{M,1,d}$ is compared with π_{kt}^M and $\hat{\pi}_{kt}^{Y,1,d}$ with π_{kt}^Y and, for Method 2, $\hat{\pi}_{kt}^{M,2,d}$ is compared with π_{kt}^M and $\hat{\pi}_{kt}^{Y,2,d}$ with π_{kt}^Y .

Common sense dictates that if monthly food inflation estimates are updated later in the month (i.e. d is higher), their accuracy is higher, as more information about prices becomes available. Therefore, one can assume that the estimate prepared on the last day of the month, using all available information from that reference period, should have the lowest possible RMSE and the highest possible correlation with the official statistics. The RMSEs of Methods 1 and 2 obtained using daily data only up to a certain day in the month are reported as divided by the RMSEs calculated using data for the whole month. Values above 100% indicate that real-time estimates of food inflation using only information up to a given day have lower accuracy than the estimate calculated using full monthly information. Naturally, the ratio of real-time and monthly RMSEs is 100% on the last day of the month (i.e. $d = 31$). [Figure 2](#) shows the relative RMSE values for MoM and YoY food inflation estimates as the number of days in the reference period are increased.

Using all the sample data (i.e. from July 2015 to August 2020), the daily estimates of MoM food inflation are calculated spanning 61 months, and YoY food inflation spanning 50 months. For the MoM food inflation, if data from only the first day of the month is used for the calculation, the RMSE when using Method 1 is approximately 142% of the RMSE using full-month data and 186% when using Method 2. As the data from the first 10 days of the month are included, the RMSE ratios gradually decrease for both methods, which indicate the improved accuracy of the real-time estimate of food inflation. After 20 days, Method 1 provides estimates with an accuracy similar to the full-month estimate (ratio of 100% for RMSE). For Method 2, this is achieved closer to 30 days.

For YoY food inflation, the ratios of RMSEs using data from only the first day of the month are lower, equal to 103% and 116% for Methods 1 and 2, respectively. Consequently, the convergence to the ratio of 100% occurs faster, with Method 1 falling below 101% after



Note(s): Each line represents the ratio of RMSE from a given method using daily data available up to a given day of the month (daily estimate of inflation) to RMSE using all the available monthly data (monthly inflation). The horizontal axis shows the days of the month. Values above 100% indicate the inferior accuracy of the daily inflation methods. Markers (dots) denote horizons and methods, for which the null of the Diebold and Mariano (1995) test, stating that the RMSE from the daily inflation method is not significantly different from the RMSE of the monthly inflation method, has not been rejected at the 5% significance level

Figure 2.
Ratio of RMSEs
corresponding to
monthly food inflation
and its daily estimate
version for “Food and
non-alcoholic
beverages” category
(by day)

including data from the first 12 days of the month and Method 2 after the inclusion of the first 28 days.

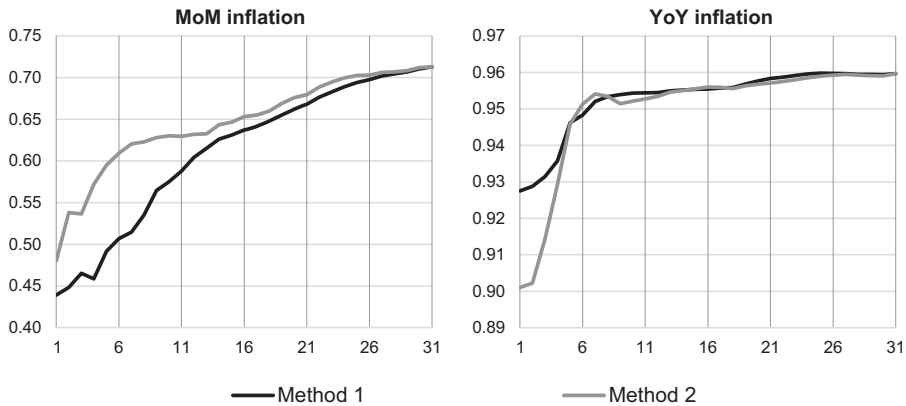
To determine how fast (i.e. on what day of the month) one can obtain a reliable estimate of food inflation, one could specify an arbitrary number (ratio), treated as a satisfactory error compared to the full-month estimate. However, each user may have a different view of what an acceptable error is. To avoid subjective assumptions, the Diebold–Mariano (DM) test (Diebold and Mariano, 1995) is used. The null hypothesis of equal forecast accuracy with the two-sided DM test between the RMSEs of the daily estimate of food inflation (i.e. $\hat{\pi}_{kt}^{M,1,d}$ or $\hat{\pi}_{kt}^{Y,1,d}$ for Method 1, and $\hat{\pi}_{kt}^{M,2,d}$ or $\hat{\pi}_{kt}^{Y,2,d}$ for Method 2) and the full-month estimate (i.e. $\hat{\pi}_{kt}^M$ or $\hat{\pi}_{kt}^Y$) are tested. If the null hypothesis cannot be rejected at the 5% confidence level, it implies that both the daily estimate and full-month estimate have the same accuracy. The tests are performed separately for daily estimates of food inflation obtained on every day of the month.

The results of the DM test are presented in Figure 2. Dots indicate the days of the month, for which the daily estimate of monthly food inflation provides the same accuracy as the full-month estimate. For MoM food inflation, the estimates display no statistical difference for Method 1 starting from day 11, and for Method 2 starting from day 26. In the earlier part of the month, the daily estimates are significantly less accurate than the full-month estimate.

For YoY food inflation, the estimates prepared even on the first day of the month are satisfactory. There is no statistical difference between their RMSE and the RMSE using data for the whole month. Estimates prepared on the consequent days display the same properties.

This study also compares how the directional accuracy of the daily estimates of monthly food inflation changes throughout the month. To do so, the correlation coefficient of the daily estimates and the official food inflation figures (Figure 3) are presented. The raw correlation coefficients (not divided by the full-month value) are used, as they are easier to interpret. As expected, the correlation of daily estimates improves as more data become available throughout the month. The correlation converges quicker to the maximum value (i.e. that

Figure 3.
Correlation coefficient
of daily inflation
estimates for “Food
and non-alcoholic
beverages” category
with the official
measure



Note(s): The correlation coefficients are presented in the raw form (not as ratios vs. the monthly estimates). Horizontal axis show the days of the month

observed at the month end) for YoY food inflation than for MoM food inflation. Nevertheless, 20 days before the month’s end, the calculated estimate is correlated with the official CSO statistic over 0.60 for MoM food inflation and 0.95 for YoY food inflation. The same results as shown in [Figures 2 and 3](#) are also presented in [Table 2](#).

To save space, the results of each food subcategory are not presented separately in the main text. For more details, see the [Appendix](#). Generally, the accuracy and leading characteristics of the proposed method hold for all 10 main subcategories. Accurate nowcasts for food inflation in subcategories are provided a few days before the end of the reference period. There is an improved correlation with the official statistics and the RMSE decreases as the number of days to prepare the food inflation estimates increases. Interestingly, for some subcategories (“sugar, jam, honey, chocolate, and confectionery,” “food products not elsewhere classified” and “non-alcoholic beverages”), the daily estimates prepared before the end of the month have lower RMSE values than those calculated using full-month data. This phenomenon may be related to the timing of the price collection by the CSO when prices change significantly during the month. This study considers more price data, whereas the CSO captures the price only at a certain point in time.

5.3 Application of this framework during the COVID-19 pandemic

The proposed method is a valid approach for measuring food inflation without the need for manual price collection. This is especially important during the COVID-19 pandemic for two reasons. First, manual price collection is difficult, as it increases the risk of price collectors contracting the virus. Second, during the pandemic, customers are doing more online shopping. According to CSO data ([Central Statistical Office, 2020b](#)), online sales of food products almost doubled between January and April 2020 due to the epidemic spread. Importantly, this involved only sales of specialized food, beverages and tobacco shops: it did not cover food sales in non-specialized stores, such as supermarkets, which are even more likely to have increased online sales.

Therefore, apart from the overall usability of this approach to measure food prices, particular attention is also paid to the period of the COVID-19 pandemic in 2020 (the lockdown in Poland started in mid-March). [Figure 4](#) shows the estimate of MoM food inflation and the official CSO figures.

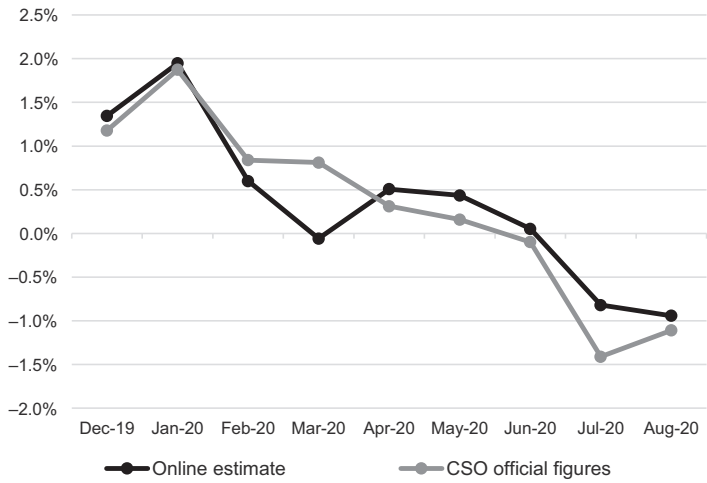
Day of the month (d)	MoM changes of food CPI				YoY changes of food CPI			
	Method 1 ($\hat{\pi}_t^{M,1,d}$ vs π_t^M)		Method 2 ($\hat{\pi}_t^{M,2,d}$ vs π_t^M)		Method 1 ($\hat{\pi}_t^{Y,1,d}$ vs π_t^Y)		Method 2 ($\hat{\pi}_t^{Y,2,d}$ vs π_t^Y)	
	RMSE	Corr	RMSE	Corr	RMSE	Corr	RMSE	Corr
1	142.1%	0.439	186.0%	0.480	102.5%*	0.928	115.8%*	0.901
2	141.2%	0.448	180.9%	0.538	102.7%*	0.929	116.4%*	0.902
3	137.6%	0.465	181.7%	0.537	102.2%*	0.931	112.3%*	0.914
4	137.8%	0.458	173.2%	0.572	101.3%*	0.936	107.1%*	0.929
5	133.1%	0.492	170.2%	0.595	102.5%*	0.946	107.3%*	0.946
6	132.2%	0.507	169.6%	0.609	102.0%*	0.948	105.4%*	0.951
7	131.0%	0.515	169.9%	0.620	103.4%*	0.952	106.7%*	0.954
8	128.3%	0.535	170.6%	0.623	101.9%*	0.953	106.5%*	0.953
9	126.9%	0.565	168.3%	0.628	102.3%*	0.954	107.6%*	0.951
10	123.7%	0.575	162.1%	0.630	101.9%*	0.954	107.6%*	0.952
11	120.7%*	0.588	158.7%	0.629	101.3%*	0.954	106.2%*	0.953
12	116.2%*	0.604	152.5%	0.632	100.6%*	0.954	104.7%*	0.953
13	113.0%*	0.615	147.6%	0.633	100.1%*	0.955	103.8%*	0.955
14	108.9%*	0.626	139.8%	0.643	99.7%*	0.955	102.5%*	0.955
15	105.9%*	0.631	134.3%	0.647	99.5%*	0.955	101.7%*	0.955
16	104.0%*	0.637	131.4%	0.653	99.4%*	0.955	101.5%*	0.956
17	102.7%*	0.641	128.6%	0.655	100.0%*	0.956	102.3%*	0.956
18	101.5%*	0.648	125.9%	0.660	100.4%*	0.956	103.0%*	0.956
19	100.7%*	0.655	122.4%	0.669	100.4%*	0.957	103.2%*	0.956
20	100.4%*	0.662	120.2%	0.676	100.5%*	0.958	103.4%*	0.957
21	100.4%*	0.668	119.1%	0.679	100.7%*	0.958	103.9%*	0.957
22	100.1%*	0.676	116.5%	0.689	100.7%*	0.959	103.9%*	0.958
23	99.8%*	0.683	114.7%	0.695	100.5%*	0.959	103.7%*	0.958
24	99.5%*	0.689	112.0%	0.700	100.0%*	0.960	102.6%*	0.959
25	99.5%*	0.694	109.9%	0.703	99.6%*	0.960	101.7%*	0.959
26	99.4%*	0.697	108.3%*	0.703	99.5%*	0.960	101.2%*	0.959
27	99.1%*	0.702	106.3%*	0.706	99.8%*	0.960	101.1%*	0.959
28	99.3%*	0.704	104.4%*	0.707	99.7%*	0.959	100.4%*	0.959
29	99.4%*	0.707	102.5%*	0.708	99.7%*	0.959	100.3%*	0.959
30	99.7%*	0.710	100.7%*	0.712	100.0%*	0.959	100.2%*	0.959
31	100.0%*	0.713	100.0%*	0.713	100.0%*	0.96	100.0%*	0.960

Note(s): The table presents RMSE and correlation coefficients for daily food inflation estimates in line with Equations (11) and (12). RMSE obtained using data up to a given day of the month (d) are reported as divided by the RMSE of the full-month estimate. * denote days of the month, for which the null hypothesis of the Diebold and Mariano (1995) test, stating that the RMSE for a given daily estimate are not significantly different from the RMSE for the full-month estimate, cannot be rejected at 5% significance level

Table 2.
RMSE and correlation
coefficients for the
daily food inflation
estimates

The approach used in this study could track official food inflation quite accurately during the COVID-19 epidemic in Poland. The RMSE (between $\hat{\pi}_{kt}^M$ and π_{kt}^M) of 0.005 for the period of March–August 2020 is similar to the RMSE observed for the full 2015–2020 sample (i.e. 0.006). This comparison indicates that the error of measurement was even lower during that time than in the past few years. It is worth noting that the biggest discrepancies between this study's estimate and the official statistics occurred in March and July 2020. This study's estimate missed the official figure of MoM food inflation by 0.87 percentage points in March and by 0.59 percentage points in July. March was the month when the CSO reported difficulties in collecting price information in traditional outlets owing to the start of the lockdown in Poland and tried to use alternative data sources (Central Statistical Office, 2020a), whereas July was the month when the CSO ceased gathering price data remotely and switched back to fully manual collection, which likely reduced the consistency of the official

Figure 4.
Comparison of online
estimate and official
statistics of MoM food
inflation



Note(s): *Own calculations*

time series. The discrepancies between this study’s estimates and the official statistics do not mean that this method could not properly capture price changes during the pandemic. On the contrary, one can argue that such a discrepancy makes the case for using online data even stronger. After the CSO adjusted its price collection methods, the discrepancy between estimates using this method and the official statistics dropped to 0.21 percentage points from April to June 2020. Considering that this method entails a significantly lower workload, these results can be deemed satisfactory.

6. Conclusions

This study demonstrates that food inflation can be accurately measured during the COVID-19 pandemic using only a laptop and an Internet connection, without the need to rely on official statistics. More importantly, these CPI estimates can be provided in a timely manner. Using this study’s approach, a monthly index similar to the official food CPI can be obtained 2 weeks before the end of the reference period and about 30 days before the official final CSO release. Furthermore, this study method does not require manual price collection from the outlets, which eliminates the risk of price collectors contracting the virus. During the COVID-19 pandemic, customers shopped more online, and thus, online prices may reflect price trends more accurately than those collected from traditional shops. This study contributes to the fast-growing body of literature focused on developing novel methods to monitor the economy during the COVID-19 pandemic.

This study offers some important theoretical implications. High-frequency inflation data are useful for detecting the impact of a variety of events (e.g. policy announcements by central banks, changes in the exchange rate and commodity price shocks) on retailers’ pricing behavior. Moreover, it is shown that the official CSO statistics may have overestimated the food inflation spike at the beginning of the pandemic, leading to elevated inflation uncertainty and possible un-anchoring of inflation expectations. These distortions observed during the pandemic contributed to errors in estimating the actual cost of living, interpreting inflation and conducting economic policy based on inflation indexing. This study also offers insights into the price-setting mechanism of supermarkets during the pandemic. Moreover, smaller

firms could benefit from using the framework given herein to optimize their own pricing strategy in real time.

Although the usefulness of web-scraped data for a variety of inflation-related issues is proven, further work is suggested as follows. Online data enable calculation of how prices change from day to day, and thus, instead of traditional monthly or annual growth rates, one can obtain more granular information about inflation. Furthermore, it is possible to monitor other categories of consumer goods besides food. In future, one may be able to track most of the items in the inflation basket using only web-scraped data. Such information could then be provided in an open repository in real time for all interested parties. The challenges of tracking the changing structure of the inflation basket at higher than yearly frequency would need to be addressed. Conducting such research would require cooperation between academics and the private sector (e.g. commercial banks) to provide information about consumer purchases in real time, for example, via credit card transactions. Finally, future research could compare the degree of price stickiness in offline and online stores to better understand how prices are set.

References

- Aghajanyan, G., Baghdasaryan, T. and Lazyan, G. (2017), "The use of big data in Central Bank of Armenia", *Paper Presented at the IFC-Bank Indonesia Satellite Seminar on 'Big Data', ISI Regional Statistics Conference*, Bali, Indonesia, 21 March 2017.
- Akter, S. (2020), "The impact of COVID-19 related 'stay-at-home' restrictions on food prices in Europe: findings from a preliminary analysis", *Food Security*, Vol. 12 No. 4, pp. 719-725.
- Aparicio, D. and Bertolotto, M.I. (2020), "Forecasting inflation with online prices", *International Journal of Forecasting*, Vol. 36 No. 2, pp. 232-247.
- Aparicio, D. and Cavallo, A. (2021), "Targeted price controls on supermarket products", *The Review of Economics and Statistics*, Vol. 103 No. 1, pp. 60-71.
- Armantier, O., Koşar, G., Pomerantz, R., Skandalis, D., Smith, K.T., Topa, G. and Van der Klaauw, W. (2020), *How Economic Crises Affect Inflation Beliefs: Evidence From the COVID-19 Pandemic*, Federal Reserve Bank of New York Staff Report 949.
- Baker, S.R., Farrokhnia, R., Meyer, S., Pagel, M. and Yannelis, C. (2020), "How does household spending respond to an epidemic? Consumption during the 2020 COVID-19 pandemic", NBER Working Papers 26949, National Bureau of Economic Research.
- Breton, R., Clews, G., Metcalfe, L., Milliken, N., Payne, C., Winton, J. and Woods, A. (2015), *Research Indices Using Web Scraped Data*, Office for National Statistics.
- Cabral, L. and Xu, L. (2020), "Seller reputation and price gouging: evidence from the COVID-19 pandemic", *Covid Economics: Vetted and Real-Time Papers*, Vol. 12, pp. 1-20.
- Carvalho, V.M., Hansen, S., Ortiz, Á., Ramón García, J., Rodrigo, T., Rodríguez Mora, S. and Ruiz, J. (2020), *Tracking the COVID-19 Crisis with High-Resolution Transaction Data (DP14642)*, CEPR Discussion Papers.
- Cavallo, A. (2013), "Online and official price indices: measuring Argentina's inflation", *Journal of Monetary Economics*, Vol. 60 No. 2, pp. 152-165.
- Cavallo, A. (2017), "Are online and offline prices similar? Evidence from large multi-channel retailers", *American Economic Review*, Vol. 107 No. 1, pp. 283-303.
- Cavallo, A. (2020), "Inflation with covid consumption baskets", NBER Working Papers 27352, National Bureau of Economic Research.
- Cavallo, A. and Rigobon, R. (2016), "The billion prices project: using online prices for measurement and research", *Journal of Economic Perspectives*, Vol. 30, No. 2, pp. 151-178.
- Central Statistical Office (2019), *Prices in the National Economy in 2014–2018*, Statistics Poland Warsaw, available at: <https://stat.gov.pl/download/gfx/portalinformacyjny/en/defaulttakt>

- ualnosci/3284/2/15/1/prices_in_the_national_economy_in_2014-2018.pdf (accessed 25 September 2020).
- Central Statistical Office (2020a), "Consumer price indices in April 2020", available at: <https://stat.gov.pl/en/topics/prices-trade/price-indices/consumer-price-indices-in-april-2020,2,96.html> (accessed 8 June 2020).
- Central Statistical Office (2020b), "Retail sales index—April 2020", available at: <https://stat.gov.pl/en/topics/prices-trade/trade/retail-sales-index-april-2020,11,28.html> (accessed 8 June 2020).
- Cicala, S. (2020), "Early economic impacts of COVID-19 in Europe: a view from the grid", Working Paper, University of Chicago, 8 April.
- Clark, T.E. and Davig, T. (2009), *The Relationship between Inflation and Inflation Expectations*, FOMC Secretariat.
- Cowen, T. (2017), "Price gouging can be a type of hurricane aid", *Bloomberg Opinion*, 5 September, available at: <https://www.bloomberg.com/opinion/articles/2017-09-05/price-gouging-can-be-a-type-of-hurricane-aid>.
- Del Negro, M. and Schorfheide, F. (2013), "DSGE model-based forecasting", *Handbook of Economic Forecasting*, Elsevier, Vol. 2, pp. 57-140.
- Diebold, F.X. and Mariano, R. (1995), "Comparing predictive accuracy", *Journal of Business and Economic Statistics*, Vol. 13, No. 3, pp. 253-265.
- Diewert, W.E. and Fox, K.J. (2020), "Measuring real consumption and CPI bias under lockdown conditions", NBER Working Papers 27144, National Bureau of Economic Research.
- D'Acunto, F., Malmendier, U., Ospina, J. and Weber, M. (2019), "Exposure to daily price changes and inflation expectations", NBER Working Papers 26237, National Bureau of Economic Research.
- E-Grocery (2019), "E-Grocery w Polsce. Zakupy online [E-Grocery in Poland. Online shopping]", available at: https://www.ecommercepolska.pl/files/4415/1775/0535/E-grocery_w_Polsce_Zakupy_spozywcze_online_raport.pdf (accessed 12 August 2020).
- Ebrahimy, E., Igan, D. and Peria, S.M. (2020), *The Impact of COVID-19 on Inflation: Potential Drivers and Dynamics*, International Monetary Fund.
- European Central Bank (2019), "New features in the Harmonised Index of Consumer Prices: analytical groups, scanner data and web-scraping", *Economic Bulletin* No. 2, pp. 53-56.
- Eurostat (2020), "Guidance on the compilation of the HICP in the context of the COVID-19 crisis", available at: https://ec.europa.eu/eurostat/documents/10186/10693286/HICP_guidance.pdf (accessed 8 June 2020).
- Faust, J. and Wright, J.H. (2013), "Forecasting inflation", *Handbook of Economic Forecasting*, Elsevier, Vol. 2, pp. 2-56.
- Gilchrist, S., Schoenle, R., Sim, J. and Zakrajšek, E. (2017), "Inflation dynamics during the financial crisis", *American Economic Review*, Vol. 107 No. 3, pp. 785-823.
- Graf, B. (2016), "Updating of the 2004 CPI manual", *UNECE-ILO Group of Experts Meeting on Consumer Price Indices*, Geneva, 2-4 May, available at: <https://www.unece.org/index.php?id=46772> (accessed 26 September 2020).
- Griffioen, R., De Haan, J. and Willenborg, L. (2014), "Collecting clothing data from the Internet", *Proceedings of Meeting of the Group of Experts on Consumer Price Indices*, Geneva, 26-28 May, UNECE, available at: <http://www.unece.org/stats/documents/2014.05.cpi.html#/> (accessed 27 November 2018).
- Hillen, J. (2019), "Web scraping for food price research", *British Food Journal*, Vol. 121 No. 12, pp. 3350-3361.
- Horrigan, M.W. (2013), "Big data: a perspective from the BLS", *Amstat News*, 1 January, available at: <http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/> (accessed 27 November 2018).

- Hull, I., Löf, M., Tibblin, M. and Riksbank, S. (2017), "Price information collected online and short-term inflation forecasts", *Paper Presented at IFC-Bank Indonesia Satellite Seminar on "Big Data" at the ISI Regional Statistics Conference*, Bali, Indonesia, 22–24 March 2017, .
- Huynh, K., Lao, H., Sabourin, P. and Welte, A. (2020), *What Do High-Frequency Expenditure Network Data Reveal About Spending and Inflation During COVID-19?*, Staff Analytical Note No. 2020-20, Bank of Canada.
- ILO, IMF, OECD, UNECE, Eurostat and World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, ILO Publications, Geneva, available at: https://www.ilo.org/wcmsp5/groups/public/-dgreports/-stat/documents/presentation/wcms_331153.pdf (accessed 26 September 2020).
- Jaravel, X. and O'Connell, M. (2020), *Inflation Spike and Falling Product Variety during the Great Lockdown*, CEPR Discussion Papers No. 14880.
- Knotek, E.S. and Zaman, S. (2017), "Nowcasting US headline and core inflation", *Journal of Money, Credit and Banking*, Vol. 49 No. 5, pp. 931-968.
- Kouvavas, O., Trezzi, R., Eiglsperger, M., Goldhammer, B. and Gonçalves, E. (2020), "Consumption patterns and inflation measurement issues during the COVID-19 pandemic", *Economic Bulletin Boxes*, European Central Bank No. 7, available at: https://www.ecb.europa.eu/pub/economic-bulletin/focus/2020/html/ecb.ebbox202007_03~e4d32ee4e7.en.html.
- Krsinich, F. (2015), "Price indices from online data using the fixed-effects window-splice (FEWS) index", *Paper Presented at the Ottawa Group*, Tokyo, Japan, 20–22 May, available at: http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s2p7_pap.pdf (accessed 27 November 2018).
- Kuchler, T., Russel, D. and Stroebel, J. (2020), "The geographic spread of COVID-19 correlates with structure of social networks as measured by facebook", NBER Working Papers 26990, National Bureau of Economic Research.
- Kumar, S., Afrouzi, H., Coibion, O. and Gorodnichenko, Y. (2015), "Inflation targeting does not anchor inflation expectations: evidence from firms in New Zealand", NBER Working Papers 21814, National Bureau of Economic Research.
- Macias, P. and Stelmasiak, D. (2019), "Food inflation nowcasting with web scraped data", NBP Working Papers 302, Narodowy Bank Polski, Economic Research Department, Warsaw.
- Marsilli, C. (2017), "Nowcasting US inflation using a MIDAS augmented Phillips curve", *International Journal of Computational Economics and Econometrics*, Vol. 7 Nos 1-2, pp. 64-77.
- Mobile Institute (2017), "Polish wallet report", available at: <https://www.slideshare.net/CEOPOLSKA/raport-eizby-portfel-polaka-o-oszczdzaniu-inwestycjach-efinansach> (accessed 9 September 2020).
- Modugno, M. (2013), "Now-casting inflation using high frequency data", *International Journal of Forecasting*, Vol. 29 No. 4, pp. 664-675.
- Mustapa, M., Ponnusamy, R.R. and Kang, H.M. (2019), "Forecasting prices of fish and vegetable using web scraped price micro data", *International Journal of Recent Technology and Engineering*, Vol. 7 No. 5, pp. 251-256.
- National Bank of Poland (2016), *Metodyka Obliczania Miara Inflacji Bazowej Publikowanych Przez Narodowy Bank Polski*, Instytut Ekonomiczny NBP, Warsaw.
- Nygaard, R. (2015), "The use of online prices in the Norwegian Consumer Price Index", *Statistics Norway, Paper Presented at 14th meeting of the Ottawa Group*, Tokyo, Japan, 20–22 May 2015, available at: [https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/d012f001b8a1cf6cca257eed008074c9/\\$FILE/Ragnhild%20Nygaard%20\(Statistics%20Norway-%20The%20use%20of%20online%20prices%20in%20the%20Norwegian%20Consumer%20Price%20Index.pdf](https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/d012f001b8a1cf6cca257eed008074c9/$FILE/Ragnhild%20Nygaard%20(Statistics%20Norway-%20The%20use%20of%20online%20prices%20in%20the%20Norwegian%20Consumer%20Price%20Index.pdf) (accessed 26 September 2020).
- Office for National Statistics (2020), "Coronavirus and the effects on UK prices", available at: <https://www.ons.gov.uk/economy/inflationandpriceindices/articles/coronavirusandtheeffectsonukprices/2020-05-06> (accessed 8 June 2020).

United Nations Statistical Commission (2014), *Big Data and Modernization of Statistical Systems*, Report of the Secretary-General, Presented at the 45th Session, 4–7 March 2014.

Uriarte, J.I., Ramírez Muñoz de Toro, G.R. and Larrosa, J. (2019), “Web scraping based online consumer price index: the ‘IPC Online’ case”, *Journal of Economic and Social Measurement*, Vol. 44 Nos. 2-3, pp. 141-159, 2019.

Woodford, M. (2003), *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton University Press, Princeton, New York, NY.

Appendix

The Appendix file is available online for this paper.

Corresponding author

Krystian Jaworski can be contacted at: kjawor@sgh.waw.pl