Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(9), 15–33 https://doi.org/10.59139/ws.2023.09.2
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 3.10.2022, poprawiony/revised: 22.06.2023, zaakceptowany/accepted: 9.08.2023

The use of web-scraped data to analyse the dynamics of clothing and footwear prices[1]

Adam Juszczak[a]

Abstract. Web scraping is a technique that makes it possible to obtain information from websites automatically. As online shopping grows in popularity, it became an abundant source of information on the prices of goods sold by retailers. The use of scraped data usually allows, in addition to a significant reduction of costs of price research, the improvement of the precision of inflation estimates and real-time tracking. For this reason, web scraping is a popular research tool both for statistical centers (Eurostat, British Office of National Statistics, Belgian Statbel) and universities (e.g. the *Billion Prices Project* conducted at Massachusetts Institute of Technology). However, the use of scraped data to calculate inflation brings about many challenges at the stage of their collection, processing, and aggregation. The aim of the study is to compare various methods of calculating price indices of clothing and footwear on the basis of scraped data. Using data from one of the largest online stores selling clothing and footwear for the period of February 2018–November 2019, the author compared the results of the Jevons chain index, the GEKS-J index and the GEKS-J expanding and updating window methods. As a result of the calculations, a high chain index drift was confirmed, and very similar results were found using the extension methods and the updated calculation window (excluding the FBEW method).

Keywords: inflation, web scraping, online shopping, GEKS-J

JEL: C43, C49

Wykorzystanie danych scrapowanych do analizy dynamiki cen odzieży i obuwia

Streszczenie. Web scraping to technika pozwalająca automatycznie pobierać informacje zamieszczone na stronach internetowych. Wraz ze wzrostem popularności zakupów online stała się ona ważnym źródłem informacji o cenach dóbr sprzedawanych przez detalistów. Wykorzystanie danych scrapowanych na ogół nie tylko pozwala znacząco obniżyć koszty badania cen, lecz także poprawia precyzję szacunków inflacji i umożliwia śledzenie jej w czasie rzeczywistym. Z tego względu web scraping jest dziś popularną techniką badań prowadzonych zarówno w ośrodkach statystycznych (Eurostat, brytyjski Office of National Statistics, belgijski Statbel), jak i na uniwersytetach (m.in. *Billion Prices Project* realizowany na Massachusetts

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na konferencji 40th International Conference MSA'2022 joined with MASEP, która odbyła się w dniach 7–9 listopada 2022 r. w Łodzi. / The article is based on a paper delivered at the 40th International Conference MSA'2022 joined with MASEP conference, held on 7–9 November 2022 in Łódź.

[a] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Polska / University of Lodz, Faculty of Economics and Sociology, Poland. ORCID: https://orcid.org/0000-0001-6027-6605.
E-mail: adam.juszczak@eksoc.uni.lodz.pl.

© Adam Juszczak
Artykuł udostępniony na licencji CC BY-SA 4.0 / Article available under the CC BY-SA 4.0 licence

Institute of Technology). Zastosowanie danych scrapowanych do obliczania inflacji wiąże się jednak z wieloma wyzwaniami na poziomie ich zbierania, przetwarzania oraz agregacji. Celem badania omawianego w artykule jest porównanie różnych metod obliczania indeksów cen odzieży i obuwia wykorzystujących dane scrapowane. Na podstawie danych z jednego z największych sklepów internetowych zajmujących się sprzedażą odzieży i obuwia za okres od lutego 2018 r. do listopada 2019 r. porównano wyniki indeksu łańcuchowego Jevonsa, indeksu GEKS-J oraz indeksów GEKS-J z użyciem metod rozszerzenia i aktualizowania okna obliczeń. Potwierdzono wysokie obciążenie dryfem łańcuchowym, a ponadto stwierdzono bardzo podobne wyniki przy użyciu metod rozszerzenia i aktualizowania okna obliczeń (z wyłączeniem metody FBEW).

Słowa kluczowe: inflacja, web scraping, zakupy online, GEKS-J

1. Introduction

Online shopping is becoming increasingly popular. In 2020, 65% of EU citizens made at least one online purchase (Eurostat, 2020). The COVID-19 pandemic accelerated the shift towards a more digitalised world. According to a survey conducted by the United Nations Conference on Trade and Development (UNCTAD) and Netcomm Suisse Observatory in Brazil, China, Germany, Italy, the Republic of Korea, the Russian Federation, South Africa, Switzerland and Turkey, the proportion of active online shoppers who purchased products online at least once in the two months preceding the survey rose by 6–10 p.p. in most product categories, even though online spending on average fell due to the economic crisis (Netcomm Suisse Observatory, UNCTAD, 2020).

Web scraping is a technique that allows researchers to automatically obtain data from websites of online shops. In addition to prices, this method yields information about discounts and product availability, and provides product description. Web scraping enables clients and researchers to see the whole range of products of a given shop and monitor it on a daily basis. As regards the latter, it helps lower the cost of the collection of price data.

However, using web-scraped data entails potential difficulties. It is estimated that goods and services for which prices are not available online constitute approximately 15% of the inflation basket. When creating an index based on scraped data, these COICOP (Classification of Individual Consumption by Purpose) categories cannot be taken into account, even if one tried evenly distributing their weight over other categories located in the same aggregate (Radzikowski & Śmietanka, 2016). However, in the last few years, this problem decreased due to the emergence of new applications and websites offering products and services that previously had not been available online. An example here could be the Booksy application, which provides information on prices of services of hairdressers, barbers, beauty salons, etc. (see Table 1).

The main issue with the use of web-scraped data for the measurement of inflation is the lack of information on consumer behaviour, i.e. the on the quantity of goods sold. This can be observed especially with regard to products which do not have an expiry date, such as books, CDs, etc. As a result, less popular items of which there are only a few in stock have the same weight as bestsellers. Some authors have recommend interesting ways to resolve this issue – for example by approximating the product weight in the final index based on the number of visits to a given website, the number of 'likes' for a given product, or the probability distribution of the consumption levels in a given group of products derived from alternative statistical surveys (Chessa & Griffioen, 2019).

Table 1. Comparison of the processes of data collection by means of web scraping and traditional interviewers

Web scraping	Price data collection by traditional interviewers
• Data collected automatically • High frequency of data collection (daily frequency is usually preferred) • Data collected only from large shops • Scraper maintenance costs sometimes high (although still lower than traditional price data collection systems) • Some branch websites missing services (approx. 15%) • Need to process large amounts of data • Collection of offer prices, not transaction prices	• Data available with a delay • Lower frequency of data collection and delays • Data collected from both large and small shops • Higher data collection costs • Data adjusted to the inflation basket

Source: Juszczak (2021).

It is also worth noticing that in calculating a price index from scraped data, the term 'product' is broader than 'item'. 'Product' is usually introduced to denote a set of one or more items that share certain characteristics. The definition of the product and its aggregation method, especially regarding the balance between the product match and the homogeneity of a product group, is a big concern in the process of the CPI (Consumer Price Index) measurement (Białek et al., 2022; Chessa, 2021).

Web-scraped data has been tested for use in inflation measurement for nearly two decades. In mid-2000s, Lünnemann and Wintr (2006) noticed the difference in price stickiness between psychical and online shops. In 2008, Alberto Cavallo from Harvard University and Roberto Riggobon from the Massachusetts Institute of Technology founded the *Billion Prices Project*, in which they used prices collected from hundreds of online retailers around the world on a daily basis to conduct research into the macro- and international economics. Despite the project's official conclusion in 2016, they have continued to research the subject, and started a few new initiatives, such as *Inflation Verdadera Argentina*, *Inflation Verdadera Venezuela*, which provides price indices alternative to official government CPI in these countries, and *Pricestats Daily Inflation Indicators*, which calculates daily inflation and PPP (Purchasing Power Parities) indicators from over 20 countries.

In many countries, such as Austria, Canada, Germany, the Netherlands, Norway, Singapore, Italy, or the US, the possibility of including web scraping to the CPI measurement is subject to research by their statistical offices (Auer & Boettcher, 2017; Bosch, n.d.; Chuanyang & Lee Wen Hao, 2016; Polidoro et al., 2015). In 2020, Eurostat gathered some of the European national statistical office's experiences in the field of web scraping in the *Practical guidelines on web scraping for the HICP*. Some case studies featured in the document are presented in this paper.

The purpose of the study is to compare various methods of calculating price indices of clothing and footwear using scraped data.

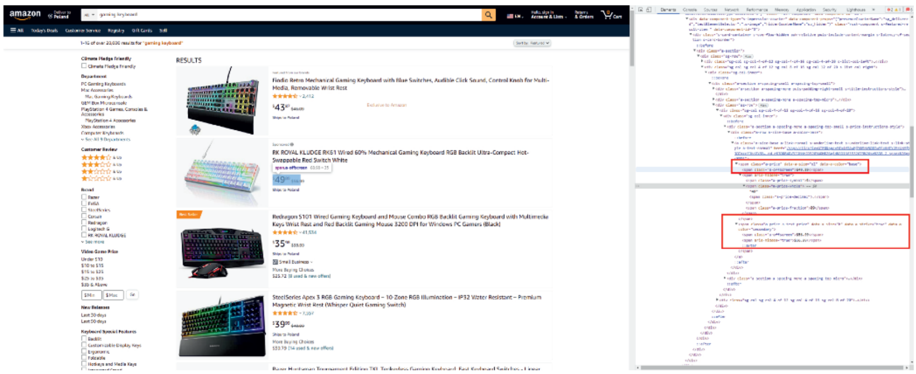
2. Technical aspects of web scraping and practical usage in inflation measurement

As mentioned in the Introduction, web scraping is a technique that makes it possible to automatically obtain data from websites. It is used in many sectors, e.g. in media (collecting comments from social media and discovering trending topics), real estate (collecting potential consumers' profiles), tourism (scraping information on destination spots such as historical sites, hotels, and restaurants to facilitate travel planning), and research.

There is a vast selection of scraping programs and methods. Some of them do not require programming skills from the users. However, these methods are less precise – for example, the programs are less likely to take the best object class from the code (Juszczak, 2021). This could lead to gaps in the collected data if the website has a dynamic code or has been subjected to minor corrections. Therefore, the most stable constructs for regular data collection are made with a specific code suited to the selected website. An example of a code website is presented in Figure 1.

There are many environments for web scraping. One of the most popular tools are Python and R, which offer many dedicated scripts and libraries. In the Python environment, for example, it is possible to use the API (Application Programming Interface) website, simulating the real user movement in Selenium (used originally to automatise application testing), or download the whole website and scrape by the Beautiful Soup package (Persson, 2019). In R, the most popular packages for web scraping are RVest and RCrawler.

Figure 1. Example of websites' code containing product name, description, price and discount



Source: Amazon website.

In 2020, Eurostat compiled the experiences of various European statistical offices in web scraping for the CPI calculation. As it turned out, most European countries had then already introduced web scraping to the process of inflation measurement.

2.1. Calculation of price indices for rail transport

Rail passenger transport prices in France, as well as in many other countries, became dynamic, i.e. the final price that the consumer pays for the ticket depends not only on the destination and the seat class, but also on the occupancy rate and date. For services with volatile prices, Eurostat recommends preparing a sample if prices are representative of the behavior of the consumers.[2]

Thus, the National Institute of Statistics and Economic Studies in France (INSEE) implemented automated collection of data on the prices of train tickets. In 2018 and 2019, the Institute conducted an experiment for high-speed trains. The automated script collected ticket prices with four anticipation dates: 2, 10, 30, and 60 days before the train departure, according to various consumer profiles, for 250 one-way destinations. In 2020, the INSEE extended this kind of price collection to regional and mainland trains.

The collected prices were aggregated by a geometric average at the elementary level by priority, profile, route, or period (either weekend or weekdays). After that,

[2] Dynamic prices in services tend to change depending on demand. Thus, creating a sample table that represents various consumers (including those who use railways when it is more expensive, e.g. during winter holidays) makes it possible to capture real consumer prices better more accurately than traditional data collection.

the data was aggregated by means of the Laspeyres index using weights calculated from rail traffic data.

The INSEE experiment showed that prices calculated with web-scraped data are much more volatile than prices collected with traditional methods. This applied mainly to school holiday periods, where the difference between the two indices for railway passenger transport reached 7.6 p.p.

2.2. Using web scraping of laptop prices to estimate the hedonic model

Various products due to big product churn have to be replaced in price indices. However, in the case of some products (such as e.g. electronic devices), we can observe quality change in the new models. If this is the case, it is necessary to consider measuring comparable prices, so that we can employ hedonic regression to estimate the difference using product characteristics as variables. In the case of laptops, theses variables would be e.g. the processor, RAM or graphic card.

The INSEE started collecting laptop prices using Python from two French e-commerce platforms: rueducommerce.com and boulanger.com. In addition, the Institute gathered the product's name, brand and technical characterictics. Hedonic model maintenance is easier than traditional web scraping (the script is executed only when the hedonic model coefficient repricing is necessary). However, it creates some problems, such as different names for the same concepts or the use of colours in the model's name. Hence a lot of data 'cleaning' is needed, and the missing values are imputed with the most frequently-occurring values or the mean ones. Only observations with more than 50% of the variables missing are removed from the sample.

2.3. Using Amadeus API to collect flight prices

One of the possible applications of web scraping that complements traditional statistical methods of data collection is gathering prices of flight tickets. Statistics Finland has experimented with web scraping of flight prices using Python Amadeus API with additional packages, e.g. *datetime*, *numpy*, *pandas*, and *sys*.

The process involves sending a request using Python scripts with parameters of a desired flight to the API, which returns the information on the prices with some additional requested parameters, such as the flight operator. What is crucial here is that Statistics Finland uses the mapping table in Excel. It contains a requested sample of selected flights, featuring the destination, seat class (economy, business, etc.), the timespan between the departure and the return dates, and the day of the week of the departure. Using mapping tables allows easy application of changes to samples, without the necessity to alternate the Python web scraping scripts.

2.4. *Use of web scraping in the HICP* – the 2020 European Central Bank research among national statistical offices

In 2020, Eurostat in collaboration with the European Central Bank conducted a survey among national statistical institutes on the use of web scraping in measuring inflation. Twenty countries of those that filled the questionnaire had already been using web scraping, although a few of them were still in the testing phase. Most of the remaining countries were planning to introduce web scraping to the CPI measurement process. The most common web-scraped item categories were garments, shoes and other footwear, household appliances, second-hand motor cars, passenger transport by train and air, mobile telephone equipment, personal computers, books, package holidays, and hotels, motels, inns and similar accommodation services.

Half of the EU national statistical offices do not inform shops about scraping their websites. The remaining NSIs usually send this information by an official letter to the CEO or the person responsible for the company's website. Most of the surveyed statistical offices do not allow the identification of their scrapers. The number of scraped prices differs vastly between the NSIs – from 100 prices per day to 60 thousand prices per month. Usually, one product is scraped from three or fewer shops. There are some exceptions, however – some products like personal computers and mobile phones are often scraped from more than 10 websites.

The most commonly-reported approach to product aggregation is using an unweighted geometric average Jevons index of collected prices. After that, the indexes are usually weighted using Laspeyres aggregation according to their COICOP level. Some NSIs use hedonic methods. Many NSIs have not started calculating indices yet due to still being in the early testing phases of introducing web scraping to the CPI calculation.

2.5. Polish experiences with using web scraping to calculate price indices

In Poland, there have so far been three projects focused on the use of web scraping in the CPI. The *eCPI* project launched by the National Bank of Poland focuses on forecasting the inflation level based on current data. Thanks to updating data on a daily basis, methods of nowcasting drawing upon web-scraped data could be less biased (by 11%) than the best ARMA models (Macias & Stelmasiak, 2018). *Online CASE CPI*, the research carried out by the Center for Social and Economic Research (CASE) in 2016, collected data from around 50 retailers, covering 87% of products in the inflation basket (Radzikowski & Śmietanka, 2016). *Gospostrateg* has been conducted by Statistics Poland in collaboration with the SGH Warsaw School of Economics and the Polish Academy of Sciences (Bitner & Stech, 2019).

3. Research method

In the calculations, the author used price data scraped from one of the largest online retailers in Poland.[3] The data were obtained as a part of the above-mentioned *eCPI* project conducted by the Economic Analysis Department of the National Bank of Poland.

The dataset covered the period from 19th February 2018 to 21st November 2019.[4] It consisted of data on prices of several hundred thousand products (clothes and footwear) collected daily for 22 months. These data was ordered by means of R (for example, a row with the collection date was assigned to each product from the daily dataset). Then various categories of products (e.g. 'oxfords', 'ballerinas', 'suits') were selected, and the data was aggregated from the daily price to the average monthly price by product code using *PriceIndices* Package (Białek, 2021). Due to restrictions of the FBEW method[5] (data has to start in December), the data collection was artificially moved two months back (instead of February 2018 to November 2019, it was changed to December 2017 to September 2019), which did not influence the results of the index comparison.

As web-scraped price data contain only information about prices, i.e. they do not carry information about expenditures (or quantities of sold products), using weighted indices such as the Fisher or the Laspeyres index is not possible with regard to them. Instead, we use indices that require only information about prices, such as the Jevons and the Dutot index. Further in this chapter, the author will introduce price indices used for price indices calculation.

Data selected for the analysis was at the lowest aggregation level (e.g. 'suits', 'business shoes', 'oxfords') apart from the individual product ID. Due to that and the lack of additional information, i.e. the product description, the author could not use a higher aggregation level of sub-categories to avoid the product churn price-linking problem.[6]

3.1. The Jevons index

The Jevons index is a bilateral index, which compares the current period with a base period fixed at some time in the past. It could be, for example, the first period

[3] Good practices of web scraping suggest that unless it is necessary, authors should not provide specific names of shops to avoid price comparisons. In our case, as the aim of the article is the comparison of price indices, the author decided not to disclose the name of the retailer.
[4] November 2019 is the last month that the author took part in the *eCPI* project.
[5] The FBEW method is further explained in chapter 4.4.5.
[6] Usually, if a product is sold out from a shop, the statistical office uses product description to find a replacement with the same quality to link the prices.

in the dataset (Jevons, 1865). The bilateral unweighted Jevons price index can be expressed in the following way:

$$P_J^{0,t} = \prod_{j \in G_{0,t}} \left(\frac{p_j^t}{p_j^0}\right)^{\frac{1}{N_{0,t}}}, \quad t = 1, 2, \dots, T, \tag{1}$$

where p_j^t is the price of product j in period t, p_j^0 is the price of product j in period 0, and $N_{0,t} = Card(G_{0,t})$, $G_{0,t}$ are the products from the analysed homogenic group available in both periods (0,t).

However, if there are vast product rotations on the market, the Jevons index will not work correctly, especially with long time series.

3.2. The chain Jevons index

The standard Jevons index with two adjacent time periods becomes a 'period-to-period' (for example, 'month-to-month') index. The chain Jevons index links together these indices with successive multiplication in the following manner:

$$P_{CH-J}^{0,t} = \prod_{\tau=1}^{t} P_J^{\tau-1,\tau} = P_J^{0,1} P_J^{1,2} \dots P_J^{t-1,t}. \tag{2}$$

The chain Jevons index, which takes into consideration every indirect moment between 0 and t , is more suitable for the scraped-data analysis. By dividing a long time series into two shorter intervals, we avoid big sample reduction in case of product rotation.

3.3. The GEKS-J index

The GEKS-J index, like each multilateral index, has its origins in the comparison of price levels between countries and regions (Eltetö & Köves, 1964; Gini, 1931; de Haan et al., 2016; Szulc, 1964). Due to the transitivity satisfaction, it allows spatial comparisons with the results independent from the base country (Białek & Bobel, 2019). In practical terms, it is a geometric mean of the chain of ratios of Jevons indices between the base period and period t with every intermediate time point ($i = 1,2,3, \dots, t-1$) as a link period, and can be expressed as

$$P_{GEKS-J}^{0,t} = \prod_{\tau=0}^{t} \left(\frac{P_J^{\tau,t}}{P_J^{\tau,0}}\right)^{\frac{1}{t+1}}. \tag{3}$$

3.4. Additional window updating methods

Updating the multilateral index with new information may change the values of previously calculated indices. Changes like that should be avoided in the CPI measurement. One of the possible ways of dealing with data revisions is 'rolling window' methods, which shift the estimation window (often totalling 13 months) forward one period (here: a period = one month) and then splicing the new indices onto the existing time series. The methods of this type which were tested here include the window splice method, half splice method, mean splice method and movement splice method.

Not all window-updating methods require splicing the time window. Chessa (2016) proposed a method that uses a time window with a fixed base month every year (December) that is expanded every month – a Fixed Base Monthly Expanding Window (FBEW). In 2017, Lamboray proposed a Fixed Base Moving Window (FBMW), which is a mix of a FBEW and movement splice using a rolling window where the last month plays the role of the fixed base (like in the FBEW method).

3.4.1. Movement splice method

First introduced by de Haan and van der Grient (2011), this method makes calculations by chaining a month-to-month index for the last month of the shifted window to the index of the previous month in the following way:

$$P_{MS}^{0,t} = P_{MS}^{0,t-1} P_{t-T+1,t}^{t-1,t}, \tag{4}$$

where T is the length of the window.

Thus, the movement splice method takes a month-to-month index $P_{t-T+1,t}^{t-1,t}$ in respect to $t-1$ of the index series calculated on $t-T+1, t$, which is linked to the published index in $t-1$. The result is a published index in month t.

3.4.2. Window splice method

This method calculates the price index for the new month by chaining the current index to the one calculated 12 months before. The formula is defined in the following way:

$$P_{WS}^{0,t} = P_{WS}^{0,t-1} \frac{P_{t-T+1,t}^{t-T+1,t}}{P_{t-T,t-1}^{t-T+1,t-1}}. \tag{5}$$

3.4.3. Half splice method

The half splice method shifts the rolling window one period splicing the chain in the middle of window length ($t = \frac{T+1}{2}$ if T is odd and $t = \frac{T}{2}$ if T is even). If the window length is 13 months, the splicing will take place in the 7th consecutive month (Van Loon & Roels, 2018).

$$P_{HS}^{0,t} = P_{HS}^{0,t-1} \frac{P_{t-T+1,t}^{t-\frac{T+1}{2}+1,t}}{P_{t-T,t-1}^{t-\frac{T+1}{2}+1,t-1}}. \tag{6}$$

3.4.4. Mean splice method

This method uses the geometric mean of all possible choices of splicing (months from both the current and the previous window). The formula can be defined in the following way:

$$P_{GMS}^{0,t} = P_{GMS}^{0,t-1} \prod_{l=t-T+1}^{t-1} \left(\frac{P_{t-T+1,t}^{l,t}}{P_{t-T,t-1}^{l,t-1}}\right)^{\frac{1}{T-1}}. \tag{7}$$

3.4.5. Fixed base monthly expanding window (FBEW)

As suggested by Chessa (2016), a fixed-base monthly-expanding window starts with a base period (December), and is enlarged by one month at a time. This means that in January, the window consists of two months, in February of three months, etc., until it reaches a 13-months length (next December). After that, the next December is taken as the base month:

$$P_{FBEW}^{0,t} = P_{b-T,b}^{b-T,b} P_{b,t}^{b,t}, \tag{8}$$

where b is the base month.

3.4.6. Fixed base moving window (FBMW)

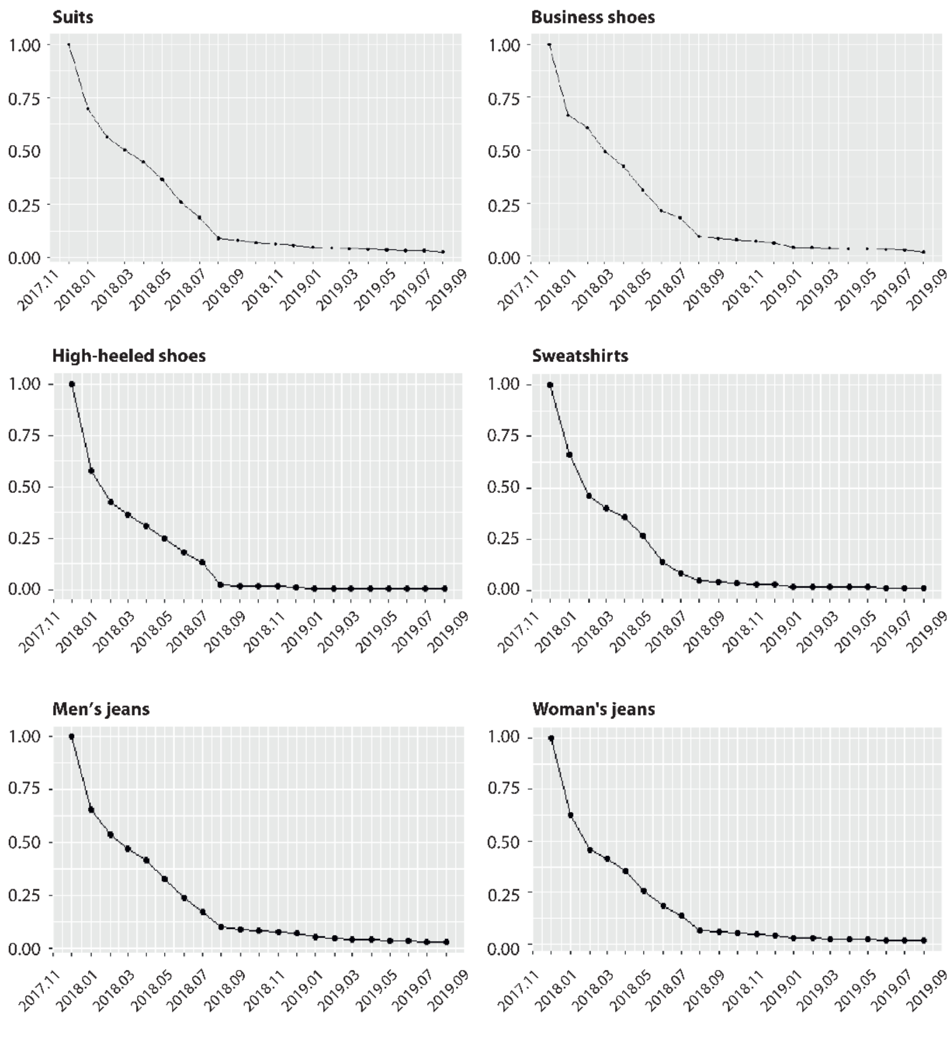
This method, proposed by Lamboray (2017), starts in a similar way to the FBEW method. However, it also uses a rolling window (the last month of the window is compared to the base):

$$P_{FBMW}^{0,t} = P_{b-T,b}^{b-T,b} P_{t-T,t}^{b,t}. \tag{9}$$

4. Study results

Fast fashion as an industry is an especially difficult object for the CPI calculation due to high product churn. As can be seen in Figure 2, the proportion of products still available after nine months dropped below 10%. In some tested product categories, none of the products from the beginning of the calculation window were still available after 20 months. In those categories where there were still products from 20 months before, the percentage varied from less than 0.5% to less than 7% (see Table 2).

Figure 2. Proportion of ID codes of products in various categories from December 2017 available in the following months



Source: author's work in R based on the analysed dataset of clothing and footwear prices.

Table 2. Availability of products from the first month of the calculation window after 20 months

Category	Percentage of products
Suits	6.19
Men's jeans	5.80
Oxfords	5.77
Business shoes	5.65
Women's jeans	3.41
Shirts	2.82
Sweatshirts	2.38
Pumps	1.34
Skirts	1.06
High-heeled shoes	1.02
Dresses	0.57
Jackets	0.00
Coats	0.00
Sandals	0.00

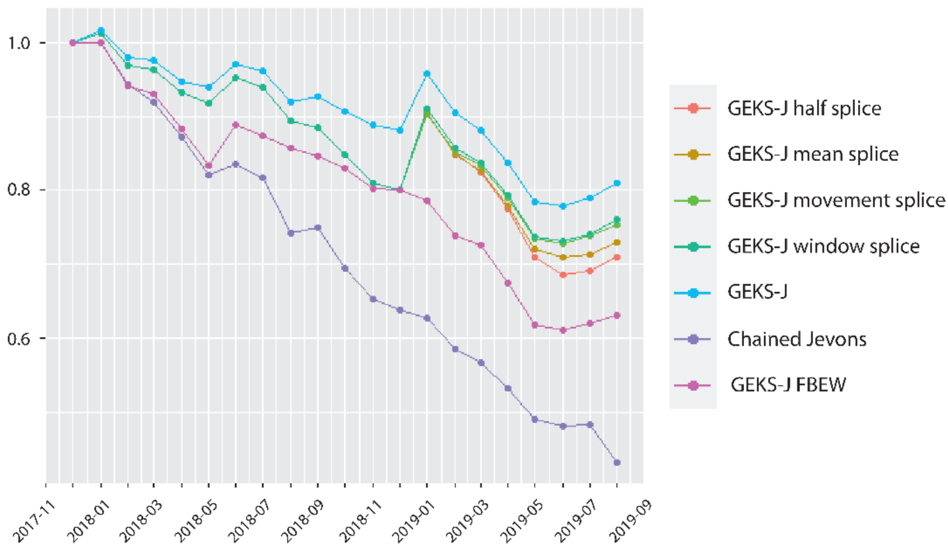
Source: author's work in R based on the analysed dataset of clothing and footwear prices.

For further calculations, three categories were picked to compare various methods of window expanding to the results obtained from the raw GEKS-J and chained Jevons indices. These were the categories of products of which more than 5% of the amount from the base month were still in stock after 20 months. In all three cases, it was not possible to calculate the fixed base moving window GEKS-J due to data churn.

In the case of 'business shoes' (5.65 % of ID codes available in the end of calculation window), we can observe significant differences between the values of indices. The result yielded by the Chain Jevons was below 0.5; the highest value, yielded by GEKS-J index, was slightly above 0.8. Indices with splice methods varied between 0.7 and 0.77, whereas movement splice and window splice methods yielded very similar results (Figure 3).

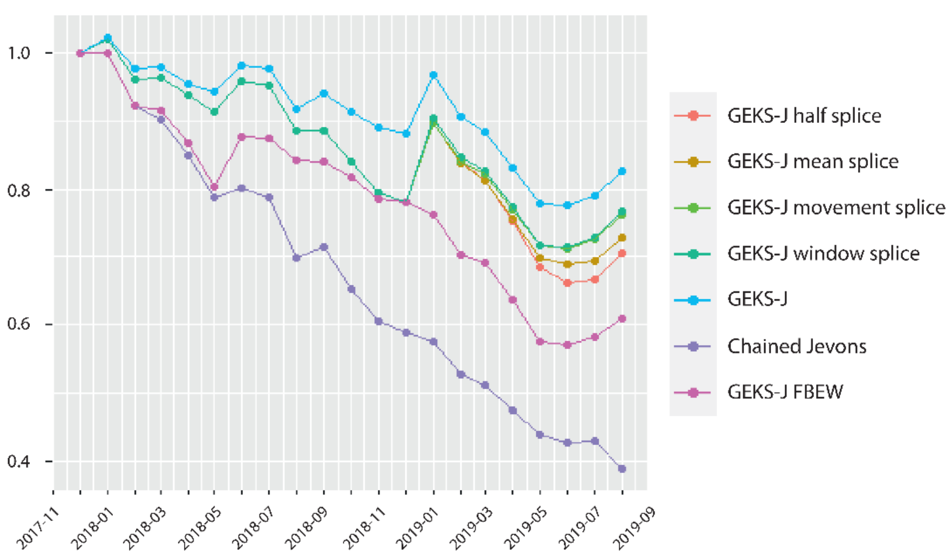
As regards 'oxfords' (5.77% of ID codes still available at the end of the calculation window), the observed trends were similar to those observed for the previous category. However, we can see that the drop in the value of the chain Jevons index was even more drastic in their case, i.e. after 20 months its value fell to below 0.4. Like in the case of 'business shoes', the splice and window splice methods yielded results similar to each other, and the results of the fixed-base expanding window were similar to those yielded by the chain Jevons index until the 6th month of the analysis (Figure 4).

Figure 3. Price index values for 'business shoes' calculated by various indices



Source: author's work in R based on the analysed dataset of clothing and footwear prices.

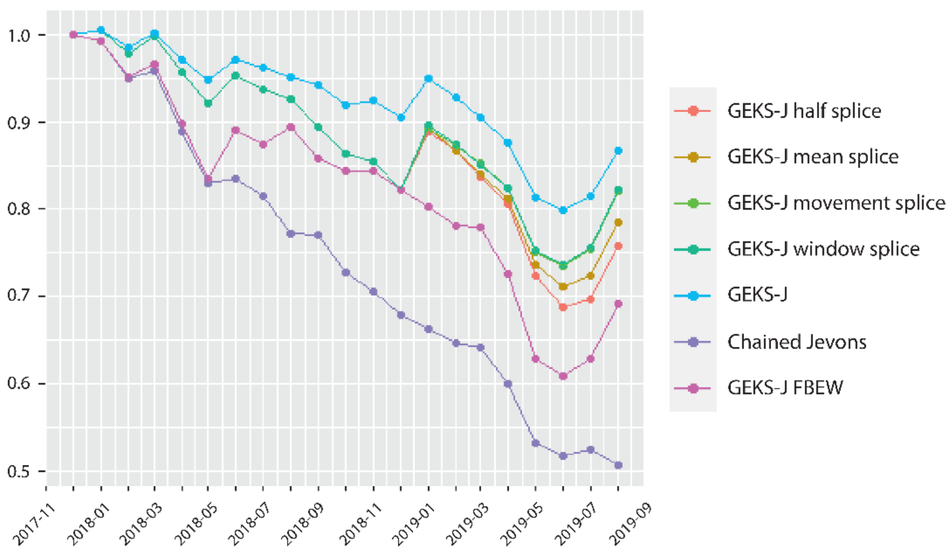
Figure 4. Price index value for 'oxfords' calculated by various indices



Source: author's work in R based on the analysed dataset of clothing and footwear prices.

In the case of 'suits' (6.19% of ID codes available at the end of the calculation window), the results of index calculation proved similar to those for 'oxfords' and 'business shoes'. However, the surge in the prices of clothing from this category in the last months of the analysis was larger than in the case of the two previous categories (Figure 5).

Figure 5. Price index value for 'suits' calculated by various indices



Source: author's work in R based on the analysed dataset of clothing and footwear prices.

The author also decided to try the imputation of missing prices. If a missing product had a previous price, then that previous price was carried on until the next real observation. If there was no previous price, the next real observation was found and carried backwards. The quantities for the imputed prices were set to zero, and then the prices were aggregated to monthly data. In the case of our three product categories, if there were no missing data, all unweighted price indices yielded the same result (see Table 3 for the 'business shoes' example). The bilateral index equalled the chained index and multilateral indices (circular test):

$$P_{t3/t1} = P_{t3/t2} * P_{t2/t1}. \tag{10}$$

where $P_{t2/t1}$ means price index between second and first period.

Table 3. The comparison of price indices and imputed price indices for 'business shoes'

Date	Jevons	Jevons imputet data	GEKS-J	GEKS-J imputed data	Chain Jevons	Chain Jevons imputed Data
2017-12	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
2018-01	0.9993063	0.9998492	1.0164081	0.9998492	0.9993063	0.9998492
2018-02	0.9410338	0.9864399	0.9788936	0.9864399	0.9429992	0.9864399
2018-03	0.9365177	0.9812515	0.9756036	0.9812515	0.9184874	0.9812515
2018-04	0.8890974	0.9684778	0.9465781	0.9684778	0.8716938	0.9684778
2018-05	0.8435621	0.9550507	0.9397652	0.9550507	0.8200934	0.9550507
2018-06	0.9081258	0.9598211	0.9703735	0.9598211	0.8353832	0.9598211
2018-07	0.8968886	0.9545321	0.9616312	0.9545321	0.8175919	0.9545321
2018-08	0.9332794	0.9330078	0.9186296	0.9330078	0.7426566	0.9330078
2018-09	0.8631369	0.9344008	0.9271442	0.9344008	0.7491014	0.9344008
2018-10	0.9070468	0.9144304	0.9058772	0.9144304	0.6943588	0.9144304
2018-11	0.8988572	0.8962291	0.8881066	0.8962291	0.6531729	0.8962291
2018-12	0.8894936	0.8883962	0.8809325	0.8883962	0.6374272	0.8883962
2019-01	0.9930019	0.8849442	0.9584619	0.8849442	0.6266572	0.8849442
2019-02	0.9500899	0.8702992	0.9044228	0.8702992	0.5845006	0.8702992
2019-03	0.9092937	0.8648703	0.8818521	0.8648703	0.5678037	0.8648703
2019-04	0.8774195	0.8460481	0.8377611	0.8460481	0.5316831	0.8460481
2019-05	0.8246801	0.8225365	0.7848102	0.8225365	0.4907581	0.8225365
2019-06	0.8271393	0.8165895	0.7790649	0.8165895	0.4809424	0.8165895
2019-07	0.8408488	0.8180891	0.7898819	0.8180891	0.4832437	0.8180891
2019-08	0.8970931	0.7953924	0.8091767	0.7953924	0.4327452	0.7953924

Source: author's work in R based on the analysed dataset of clothing and footwear prices.

5. Conclusions

The main purpose of the research presented in this paper was to compare various price indices used for web-scraped data along with methods of rolling and extending the calculation window. The author used data for the period from 19th February 2018 to 21st November 2019, obtained from one of the largest clothing and footwear internet shops.

Clothing and footwear are believed to be among the most difficult categories of goods to measure their price change indexes. It is due to high product churn and high frequency of discounted sales that impact price stability. Imputation of prices with high product churn and low level of matching data could greatly impact the results of the analysis. It should also be remembered that full imputation of data for unweighted price indices results in identical outcome for every index.

As the analysed dataset from the fast fashion retailer demonstrated, the proportion of products still available after nine months dropped below 10% in all the analysed categories. After 20 months, the percentage of most of the products still available ranged between 0.5% and 7%.

Product churn limits the possibility of using the traditional Jevons index with longer time series. However, it is possible to use chained indexes and the GEKS indexes instead. Unfortunately, when applied to longer periods, the chained index is prone to abnormal value drop connected with the chain drift.[7] In the case of GEKS-J index, updating a multilateral index with new information may change the values of previously calculated indices. One of possible solutions to this problem could be to use window-moving or window-expanding methods. It is worth mentioning that various window splice methods, and likewise FBEW and FBMW methods, yield different results, which leaves open the debate which extension method for GEKS-J index are preferable.

Analysing window expanding methods, we can observe that apart from FBEW, indices yield results varying up to 10 p.p. What is worth noticing is that the results of the movement splice and the window splice methods were almost identical.

Using web-scraped data in the calculation of inflation is still a relatively new idea. The main contribution of this paper to the existing body of research is showing that Jevons-based indexes are suitable for calculating price indexes for product categories with high product churn, such as footwear and clothing, and demonstrating the extent to which window-expanding and window-moving methods can impact the results.

All the analysed data came from one online retailer, so there might be differences in results when we compare them with the results from others. Further study is necessary in order to compare the GEKS and chained indexes with alternative approaches for footwear and clothing, as well as to use filters on results.

It would be advisable that in future studies data is analysed by means of data aggregation to sub-categories in order to avoid price-linking problem due to high product churn.

Acknowledgements

I would like to thank Prof. Jacek Białek for his help and valuable remarks at every stage of writing this paper.

The data were acquired during my internship at the National Bank of Poland, where I worked on the *eCPI* project. I would also like to extend special thanks to Paweł Macias and Damian Stelmasiak, the main authors of the project.

[7] The chain drift occurs if the index differs from unity when prices and quantities revert to their base level.

References

Auer, J., & Boettcher, I. (2017). *From price collection to price data analytics. How new large data sources require price statisticians to re-think their index compilation procedures. Experiences from web-scraped and scanner data*. Ottawa Group. https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/From%20price%20collection%20to%20price%20data%20analytics%20-Josef%20Auer,%20Ingolf%20Boettcher%20-Paper.pdf.

Białek, J. (2021). PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations. *Statistika. Statistics and Economy Journal*, (2), 122–141. https://www.czso.cz/documents/10180/143550797/32019721q2_bialek.pdf/3cd5bf11-22f4-4ee5-b294-1d7d5909e4b4?version=1.1.

Białek, J., & Bobel, A. (2019). *Comparison of Price Index Methods for the CPI Measurement Using Scanner Data*. https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+16/$FILE/Comparison%20of%20Price%20Index%20Methods%20paper.pdf.

Białek, J., Kłopotek, M., & Panek, T. (2022). *Nowoczesne technologie i nowe źródła danych w pomiarze inflacji*. Główny Urząd Statystyczny. https://bws.stat.gov.pl/BWS/gus_bws_70_nowoczesne_technologie_i_nowe_zrodla_danych_w_pomiarze_inflacji.pdf.

Bitner, T., & Stech, G. (2019, April 9). *GUS: Big Data to nasz priorytet*. https://www.computerworld.pl/wywiad/GUS-Big-Data-to-nasz-priorytet,412891.html.

Bosch, O. (n.d.). *Uses of web scraping for official statistics. ESTP course on Big Data Sources – Web, Social Media and Text Analytics*. https://circabc.europa.eu/sd/a/5e250346-44a9-471b-87f1-5b5ddb59aa77/1_Big%20Data%20Sources%20part3-Day%201-A%20Use.pdf.

Chessa, A. G. (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review of National Accounts and Macroeconomic Indicator*, (1), 49–69. https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2016-art2.pdf.

Chessa, A. G. (2021). A Product Match Adjusted R Squared Method for Defining Products with Transaction Data. *Journal of Official Statistics*, *37*(2), 411–432. https://doi.org/10.2478/jos-2021-0018.

Chessa, A. G., & Griffioen, R. (2019). Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data. *Economie et Statistique. Economics and Statistics*, (509), 49–68. https://doi.org/10.24187/ecostat.2019.509.1984.

Chuanyang, F., & Lee Wen Hao, J. (2016). Experiences with the Use of Online Prices in Consumer Price Index. *Statistics Singapore Newsletter*, (2), 1–4. https://www.singstat.gov.sg/-/media/files/publications/economy/ssnsep16-pg1-4.ashx.

Eltetö, Ö., & Köves, P. (1964). Egy nemzetközi összehasonlításoknál fellépö indexszámításl. *Statisztikai Szemle*, *42*, 507–518. http://real-j.mtak.hu/21330/5/Statisztikai_szemle_1964_05.pdf.

Eurostat. (2020). *Practical guidelines on web scraping for the HICP*. https://ec.europa.eu/eurostat/documents/272892/12032198/Guidelines-web-scraping-HICP-11-2020.pdf/.

Gini, C. (1931). On the Circular Test of Index Numbers. *Metron*, *9*(9), 3–24.

de Haan, J., & van der Grient, H. A. (2011). Eliminating chain drift in price indexes based on scanner data. *Journal of Econometrics*, *161*(1), 36–46. https://doi.org/10.1016/j.jeconom.2010.09.004.

de Haan, J., Willenborg, L., & Chessa, A. (2016). *A Review of Price Index Methods for Scanner Data*. https://doi.org/10.13140/RG.2.2.27350.88643.

Jevons, W. S. (1865). On the Variation of Prices and the Value of the Currency since 1782. *Journal of the Statistical Society of London*, *28*(2), 294–320. https://doi.org/10.2307/2338419.

Juszczak, A. (2021). Zastosowanie danych scrapowanych w pomiarze dynamiki cen. *Acta Universitatis Lodziensis. Folia Oeconomica*, *1*(352), 25–37. https://doi.org/10.18778/0208-6018.352.02.

Lamboray, C. (2017). *The Geary Khamis index and the Lehr index: how much do they differ?*. 15th Meeting of the Ottawa Group, Elville am Rhein. https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/The%20Geary%20Khamis%20index%20and%20the%20Lehr%20index%20how%20much%20do%20they%20differ%20-%20Claude%20Lamboray%20-%20Presentation.pdf.

van Loon, K., & Roels, D. (2018, May 7–9). *Integrating big data in the Belgian CPI*. Meeting of the Group of Experts on Consumer Price Indices, Geneva. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf.

Lünnemann, P., & Wintr, L. (2006). *Are Internet Prices Sticky?* (ECB Working Paper No. 645). https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp645.pdf.

Macias, P., & Stelmasiak, D. (2019). *Food inflation nowcasting with web scraped data* (NBP Working Papers No. 302). https://static.nbp.pl/publikacje/materialy-i-studia/302_en.pdf.

Netcomm Suisse Observatory, United Nations Conference on Trade and Development. (2020). *COVID-19 and E-commerce Findings from a survey of online consumers in 9 countries*. https://unctad.org/system/files/official-document/dtlstictinf2020d1_en.pdf.

Persson, E. (2019). *Evaluating tools and techniques for web scraping* [Second cycle degree project, KTH School of Electrical Engineering and Computer Science]. https://www.diva-portal.org/smash/get/diva2:1415998/FULLTEXT01.pdf.

Polidoro, F., Giannini, R., Conte, R. L., Mosca, S., & Rosetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, *31*(2), 165–176. https://doi.org/10.3233/sji-150901.

Radzikowski, B., & Śmietanka, A. (2016). *Online CASE CPI*. 1st International Conference on Advanced Research Methods and Analytics, València. http://dx.doi.org/10.4995/CARMA2016.2016.3133.

Szulc, B. (1964). Indices for multiregional comparisons. *Przegląd Statystyczny*, (3), 239–254.