University of Economics in Katowice

Journal of Economics and Management

e-ISSN 2719-9975 ISSN 1732-1948 Vol. 43 • 2021

sciendo

Adam Juszczak

https://orcid.org/0000-0001-6027-6605

Department of Statistical Methods
Faculty of Economics and Sociology
University of Lodz, Łódź, Poland
adam.juszczak@edu.uni.lodz.pl

The use of web-scraped data to analyze the dynamics of footwear prices

Accepted by Editor Ewa Ziemba | Received: September 4, 2020 | Revised: April 16, 2021; July 2, 2021 | Accepted: July 9, 2021 | Published: July 28, 2021.

Abstract

Aim/purpose – Web-scraping is a technique used to automatically extract data from websites. After the rise-up of online shopping, it allows the acquisition of information about prices of goods sold by retailers such as supermarkets or internet shops. This study examines the possibility of using web-scrapped data from one clothing store. It aims at comparing known price index formulas being implemented to the web-scraping case and verifying their sensitivity on the choice of data filter type.
Design/methodology/approach – The author uses the price data scrapped from one of the biggest online shops in Poland. The data were obtained as part of eCPI (electronic Consumer Price Index) project conducted by the National Bank of Poland. The author decided to select three types of products for this analysis – female ballerinas, male shoes, and male oxfords to compare their prices in over one-year time period. Six price indexes were used for calculation – The Jevons and Dutot indexes with their chain and GEKS (acronym from the names of creators – Gini–Éltető–Köves–Szulc) versions. Apart from the analysis conducted on a full data set, the author introduced filters to remove outliers.
Findings – Clothing and footwear are considered one of the most difficult groups of goods to measure price change indexes due to high product churn, which undermines the possibility to use the traditional Jevons and Dutot indexes. However, it is possible to use chained indexes and GEKS indexes instead. Still, these indexes are fairly sensitive to large price changes. As observed in case of both product groups, the results provided by the GEKS and chained versions of indexes were different, which could lead to conclusion that even though they are lending promising results, they could be better suited for other COICOP (Classification of Individual Consumption by Purpose) groups.

Cite as: Juszczak, A. (2021). The use of web-scraped data to analyze the dynamics of footwear prices. *Journal of Economics & Management, 43,* 251-269. https://doi.org/10.22367/jem.2021.43.12

Research implications/limitations – The findings of the paper showed that usage of filters did not significantly reduce the difference between price indexes based on GEKS and chain formulas.

Originality/value/contribution – The usage of web-scrapped data is a fairly new topic in the literature. Research on the possibility of using different price indexes provides useful insights for future usage of these data by statistics offices.

Keywords: inflation, CPI, web-scraping, online shopping, big data.
JEL Classification: C43, C49.

1. Introduction

Web-scraping is a technique that enables automatic collection of data from various websites (Cavallo, 2017). After the rise of online shopping (65% of European Union citizens made at least one online purchase in 2020 according to Eurostat (2021)), it allows the acquisition of information about prices of goods sold by retailers, such as supermarkets or internet shops. Alongside with prices, it is often possible to acquire information about discounts, availability of the product, or its description. Generally, web-scraping techniques can be used to acquire data of the entire shop's online offer on a daily basis (which allows us to monitor prices in real-time) for a fraction of costs compared to traditional methods (Macias & Stelmasiak, 2018).

The usage of web-scraping in the measurement of inflation creates numerous challenges. According to various calculations, around 15% of the consumer basket are goods and services whose prices are unavailable online. Building an index based on web-scraped data forces us to skip these products, for example by distributing their weight in the index to other categories in the same aggregate (Radzikowski & Śmietanka, 2016).

With web-scraping, we do not acquire data about the demand for the goods (Cavallo, 2018). It is problematic, especially in sectors where many products do not have an expiration date, so they can be kept in stock for years. A good example could be bookstores and shops with music and movies – bestsellers and the least popular products with only a few copies left are available on the website, hence they have the same weight in building the aggregate.

Even though numerous programs assist in web-scraping, they are flawed in many ways – for example, by not taking the best object class from the code (Juszczak, 2021). Therefore, when the site has a dynamic code or has been the subject of minor corrections, we can experience the gap in the dataset. The most stable constructs for regular data collections are made with a specific code suited for the selected website. The most popular environments for web-scraping are Python and R, which offer many dedicated scripts and libraries. In the Python environment, for example, it is possible to use the website API (Application Programming Interface), simulating the real user movement in Selenium (used originally to automatize application testing) or downloading the whole website and scrape by the *Beautiful Soup* package (Persson, 2019).

Practical usage of web scrapped data provides many challenges. Lack of data about quantity of sold products exclude adoption of many known traditional price indexes used in CPI calculation, such as Laspeyres, Lowe or Young indexes. There are, however, alternatives such as modifications of elementary indexes, for example, Jevons, Törnqvist or Dutot (Van Loon & Roels, 2018). The aim of this paper is to compare price index formulas based on Jevons and Dutot indexes being implemented to the web scrapped data and verify their sensitivity on the choice of data filter type.

The paper is divided into five sections – a literature review where the author briefly analyzes the current usage of web scraping in CPI (electronic Consumer Price Index) calculation; an overview of indexes and formulas used in the study; a data and methodology chapter where the author explains the aggregation of data and used filters; an empirical study where the results are presented; and conclusions from the paper with future studies recommendations.

2. Theoretical background

2.1. Usage of web scraping in inflation measurement

Alongside the scanner data (for further details about scanner data compare: Białek & Bobel (2019)), scrapped data are being tested for usage in inflation measurement. In the middle of 2000s, Lunneman & Wintr (2006) noticed the difference in price stickiness between physical and online shops. In 2008, Cavallo & Rigobon (2016) founded the Billion Prices Project on MIT. In their research, they compared inflation based on online process to the official inflation rates in South America countries (Cavallo, 2013). They also compared prices in online and physical retailers (Cavallo, 2017), which led to the conclusion that, on average, in 72% of cases they are identical. There is, however, a big difference between various countries – from below 50% in Japan and the U.S. to 91% in Canada or Great Britain.

In many countries, the possibility of including web-scraping for CPI measurement is subject to research by statistical offices in Austria, Canada, Germany, the Netherlands, Norway, Singapore, Italy or the U.S. (Auer & Boettcher, 2017; ten Bosch, n.d.; Chuanyang & Lee, 2016; Polidoro et al., 2015). One of the biggest projects is conducted by the Office of National Statistics in the UK, in which researchers focus on measuring footwear and clothing online price indexes using various indexes, i.e., based on the CLIP method (Clustering Large datasets Into Price indexes) (Office for National Statistics [ONS], 2017).

In Poland, there are three projects focused on the usage of web-scraping in CPI. First, the eCPI project by the National Bank of Poland focuses on forecasting based on current data. Thanks to updating data on the daily basis, methods of nowcasting based on web scrapped data are 11% less biased than the best ARMA models (Macias & Stelmasiak, 2018). Second, Online CASE CPI published by the Center for Social and Economic Research collects data from around 50 retailers, which cover 87% of products in the inflation basket (Radzikowski & Śmietanka, 2016) Third, Gospostrateg by the Polish Statistical Office runs in cooperation with the Warsaw School of Economics and the Polish Academy of Sciences (Bitner & Stech, 2019).

From the technological point of view, web-scraping offers many advantages comparing to traditional data gathering methods. As mentioned above, the data gathering process is fully automated, and much cheaper than gathering data by pollsters in physical shops. It also offers a much higher frequency of collected data, which is available right away (Juszczak, 2021). However, it covers only big retailers (small shops do not usually have a website with published product prices) which requires the processing of large quantities of data and lacks representation of around 15% of the goods and services (Radzikowski & Śmietanka, 2016)

There are different methods tested in price index calculation from web-scrapped data. Most commonly described in the literature are multilateral methods, such as GEKS (Gini–Éltető–Köves–Szulc) based on different base price indexes (such as Jevons or Törnqvist) or Time Product Dummy (Van Loon & Roels, 2018). However, as it has been mentioned, some Statistical Offices undertake a different approach. For example, the Office for National Statistics in the UK is testing CLIP method for clothes and footwear price dynamic measurement (ONS, 2017).

As mentioned in the introduction, web scrapped data present another challenge – we do not acquire information about demand for various goods. Some of the new studies tried a different approach then equal weight for every product.

For example, ONS (2020) suggest approximation of product expenditure based on their page rankings (that is, the order that products appear on a web page when sorted by popularity) via statistical distributions. Chessa & Griffioen (2019) suggested substituting quantities by the total number of web scraped prices for a product in a month, summed over all items. In case of usage of web scrapped data for forecasting Macias & Stelmasiak (2018) aggregated components with official expenditure weights in line with statistical office methodology.

2.2.Price indexes used for the analysis of web-scrapped data

Web-scrapped price data contain only information about prices, not quantities of sold goods. Therefore, it is not possible to use indexes, such as the Fisher or the Laspeyres indexes. Instead, we use indexes which require only information about prices, such as the Jevons and the Dutot indexes.

2.2.1. The Jevons index

The Jevons index is a bilateral index, which compares the current period with a base period fixed in some time in the past. It could be, for example, the first period in the dataset (Jevons, 1865). The bilateral unweighted Jevons price index can be expressed as follows:

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

where:
p_j^t – the price of the product j in period t,
p_j^0 – the price of the product j in period 0,
$N_{0,t}$ – denotes products available in both periods – 0 (base) and t (current).

However, if the market has high product churn, the Jevons index will not work correctly, especially with long time series. It allows, though, an easy comparison of a based period with every next t period.

2.2.2. The chain Jevons index

The standard Jevons index with two adjacent time periods becomes a 'period--to-period' (e.g., month-to-month) index. The chain Jevons index links together these indexes 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 takes into consideration every indirect moment between 0 and t which makes it more suitable for scrapped data analysis. By dividing a long time series into shorter two periods intervals, we can avoid large sample reduction in the case of product churn.

2.2.3. The GEKS-J index

The GEKS index, like each multilateral index, has its origin in the comparison of price levels between countries and regions. Due to transitivity satisfaction, it gave possibility to spatial comparisons with results independent from the base country (Białek & Bobel, 2019). In practice, it is a geometric mean of the chain Jevons indexes between the base period and t period with every intermediate time point ($i = 1,2,3, \dots, t-1$) as a link period in the following manner:

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

The multilateral formula of the GEKS index satisfies the condition $P^{a,b}P^{b,a}$=1 (the time reversal test).

2.2.4. The proposition of an alternative indexes based on the Dutot index formula

The Dutot (1738) price index is a commonly accepted index formula (International Labour Organization, International Monetary Fund, Organisation for Economic Co-operation and Development, Statistical Office of the European Communities, United Nations, The International Bank for Reconstruction and Development, The World Bank, 2004) and that is why it is also considered in the paper:

$$P_D^{0,t} = \frac{\sum_{j \in N0,t} P_j^t}{\sum_{j \in N0,t} P_j^0} \tag{4}$$

with its chained version:

$$P_{CH-D}^{0,t} = \prod_{\tau=1}^{t} P_D^{\tau-1,\tau} \tag{5}$$

As a consequence, the modification of the GEKS index, which is based on the Dutot formula, is proposed:

$$P_{GEKS-D}^{0,t} = \prod_{\tau=0}^{t} (\frac{P_D^{\tau,t}}{P_D^{\tau,0}})^{\frac{1}{t+1}} \tag{6}$$

3. Data and methodology

In our calculations, we use the price data scrapped from one of the biggest online shops in Poland. The data were obtained as part of eCPI project conducted by the Economic Analysis Department in the National Bank of Poland.

We decided to select three types of products for our analysis – female ballerinas, male shoes, and male oxfords to compare their prices in over one-year time period. From March 1, 2018 to April 1, 2019, we compared prices of all products in these categories on the first day of every month.

Footwear and clothing tend to have one of the largest price differences in all COICOP (Classification of Individual Consumption by Purpose) categories. It is mostly caused by sales and new collections. However, GEKS indexes are sensitive to sharp price changes. Due to this fact, apart from the analysis conducted on a full data set, we decided to introduce two types of filters to remove outliers.

The first filter is based on sets of constraints. If the absolute relative price change between t and t−1-time moments exceeds 50% then the product is removed from the sample:

$$\left|\frac{p_j^t - p_j^{t-1}}{p_j^{t-1}}\right| > 0.5 \tag{7}$$

i.e., the following condition allows this product to stay in the sample:

$$0.5 < \frac{p_j^t}{p_j^{t-1}} < 1.5 \tag{8}$$

The second filter is based on quantiles. All products with the highest price increase and decrease are removed from the sample:

$$Q_x < \frac{p_j^t}{p_j^{t-1}} < Q_{1-x} \tag{9}$$

where Q_x is the value of x-th quantile of all price changes between t and t−1 time periods.

4. The empirical study results

We decided to calculate the mentioned six price indexes in four cases for every product category – without any filter, removing products below 0.01 and above 0.99th quantiles (filter 1) or 0.1 and 0.9th quantiles (filter 2), and removing every product with month-to-month price change bigger than 50% (filter 3). Results of the filtering are presented in Table 1.

Table 1. Number of unique products in the dataset before and after applying the filters

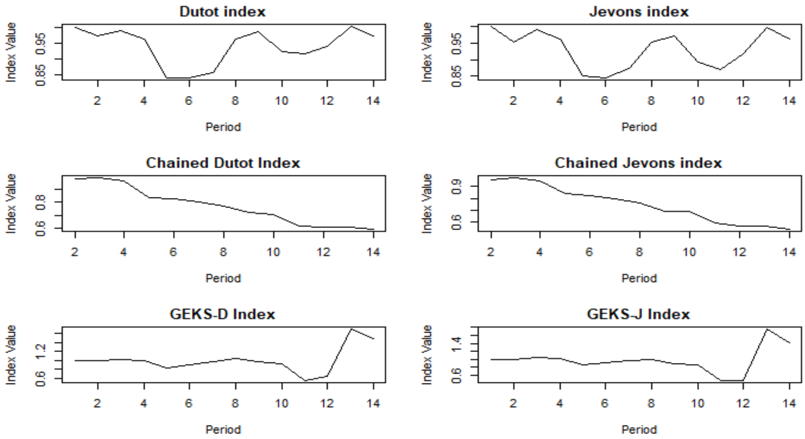
Specification	Oxfords	Ballerinas	Male shoes
Base version (raw data)	15,805	14,836	10,396
Filter 1	15,127	13,887	9,619
Filter 2	13,685	12,684	8,758
Filter 3	15,214	13,958	9,705

Source: Own calculations by using R package.

In the case of ballerinas, there are significant differences between fixed base indexes, chained indexes, and GEKS indexes. There are, however, noticeable similarities between the Dutot and Jevons based formulas.

For fixed based indexes we can notice a big drop in value between the 4th and 5th period, which can be explained by summer sales (Figure 1). This also explains the retraction of the index almost to the previous level between the 7th and 8th period (autumn, the end of sales for this type of footwear). Between the 9th and 10th period we can observe another price drop, which is probably connected to Christmas sales, followed by a price rise again from January.

Figure 1. Indexes values calculated using base dataset for ballerinas

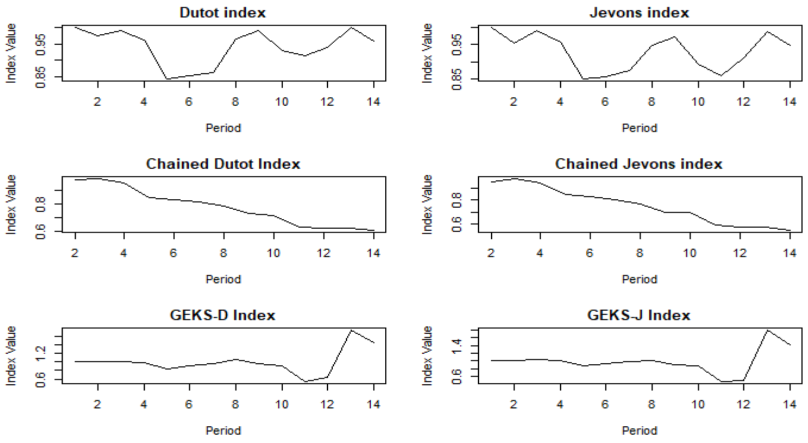


Source: Own calculations by using the R.

Sales can also be noticed in case of chained indexes. However, as this is a chained index, it is harder for a price rise to influence index value after a series of price drops.

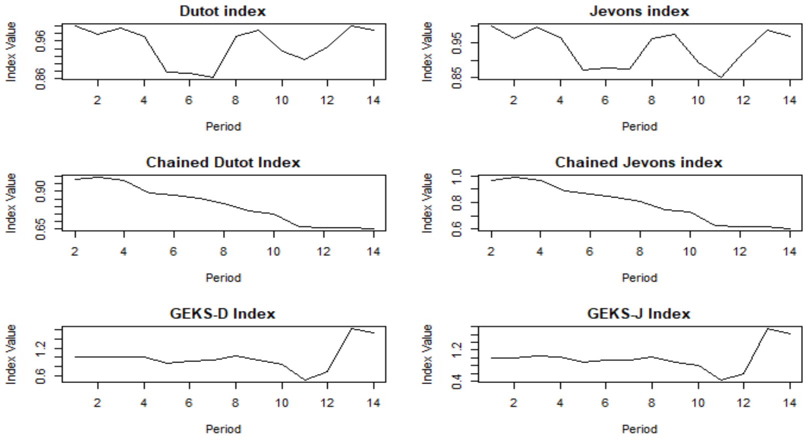
In case of GEKS based indexes, after 12 months of stable behavior (index values between 0.8 and 1), we can notice a sudden price drop in the Christmas period and an unusual rise in the beginning of the year. This can be explained by introducing new collections at the beginning of the spring. The usage of filters (Figures 2-4) mitigated this effect, but even in case of filter 2 (which eliminates 20% of the products, Figure 3) we can notice a great increase of index value.

Figure 2. Indexes values calculated using filter 1 for the ballerinas' dataset



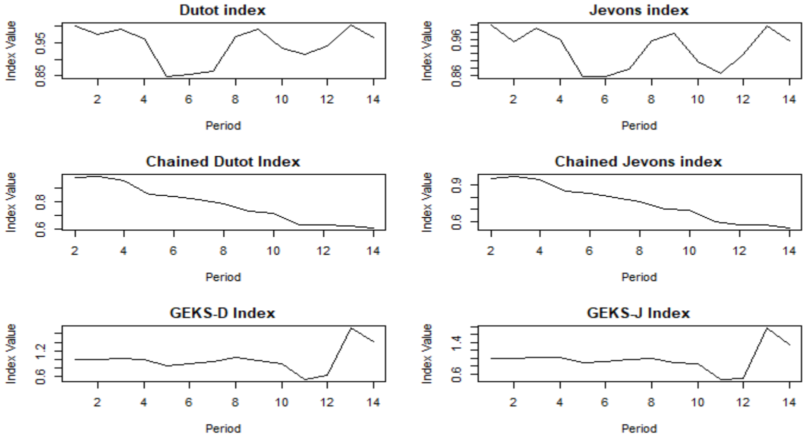
Source: Own calculations by using the R.

Figure 3. Indexes values calculated using filter 2 for the ballerinas' dataset



Source: Own calculations by using the R.

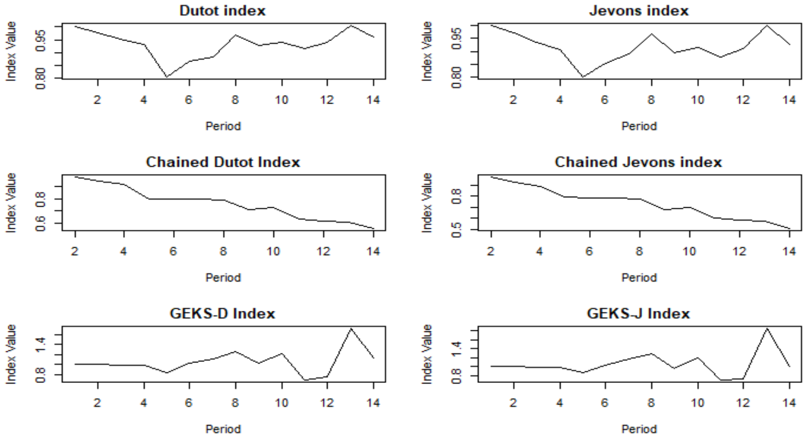
Figure 4. Indexes values calculated using filter 3 for the ballerinas' dataset



Source: Own calculations by using the R.

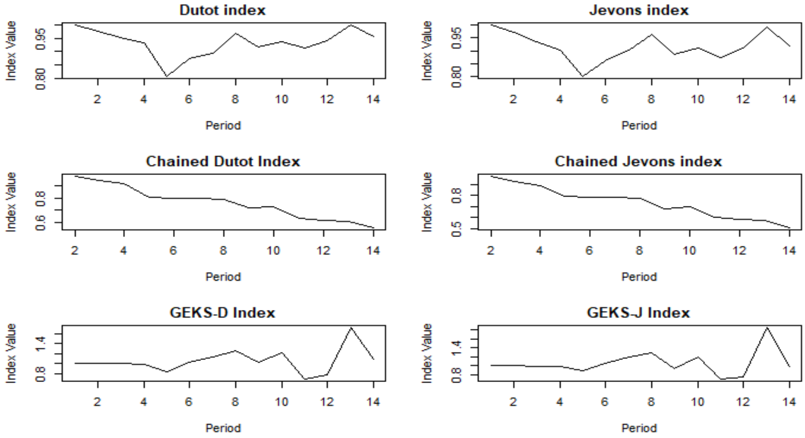
In the case of male oxfords (Figures 5-8), we can notice more significant price changes than for ballerinas. For the index of both Dutot and Jevons fixed base indexes, the value drops to 0.8 during summer sales. This time, however, prices lower gradually from the 1st period. Lower index values can also be observed for the chained Jevons index for male oxfords' dataset (0.5) in comparison to the ballerinas' dataset (0.6)

Figure 5. Indexes values calculated using base dataset for male oxfords



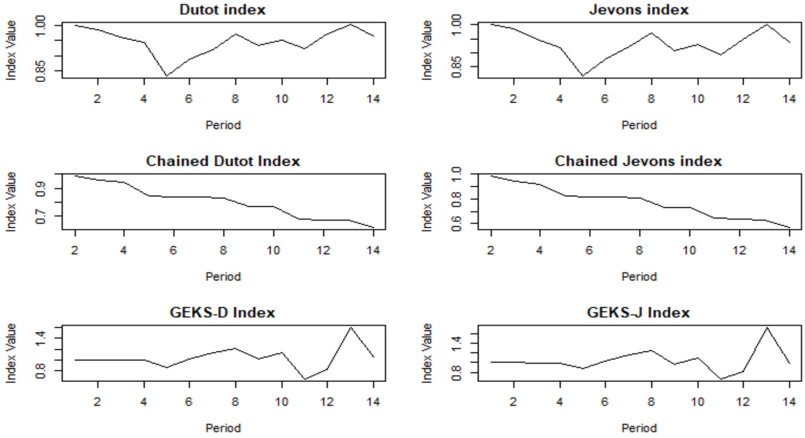
Source: Own calculations by using the R.

Figure 6. Indexes values calculated using filter 1 for the male oxfords' dataset



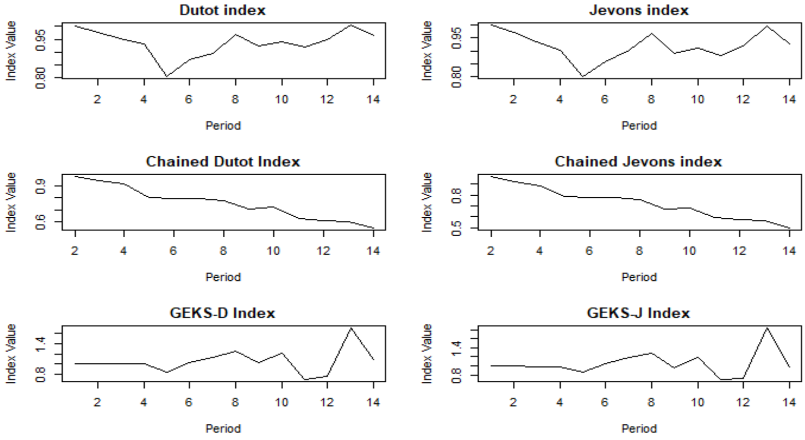
Source: Own calculations by using the R.

Figure 7. Indexes values calculated using filter 2 for the male oxfords' dataset



Source: Own calculations by using the R.

Figure 8. Indexes values calculated using filter 3 for the male oxfords' dataset

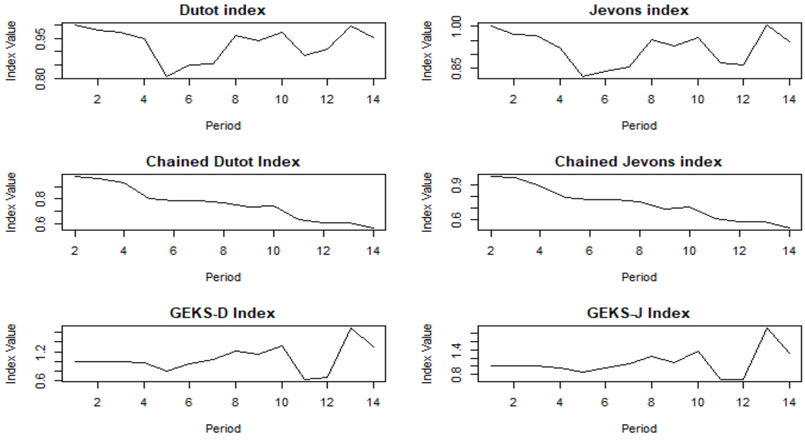


Source: Own calculations by using the R.

With reference to the ballerinas' dataset for GEKS based indexes, we can observe a sharp rise of the index value between the 12th and 13th period, which persists in the case of both the raw and the filtered dataset. Contrary to the ballerinas' dataset, we can notice another price drop of the index values to around 1 in the last period.

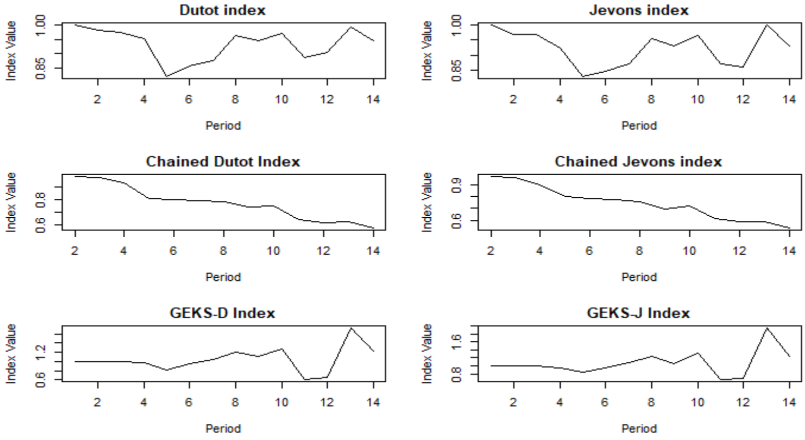
In case of male shoes (Figures 9-12), the values of dataset indexes for fixed based indexes and chained indexes are similar to the ballerinas' and male oxfords' datasets. As in the previous two cases, filters mildly altered the results, but the differences between various indexes formulas (fixed based, chained, and GEKS) are significant.

Figure 9. Indexes values calculated using base dataset for the male shoes' dataset



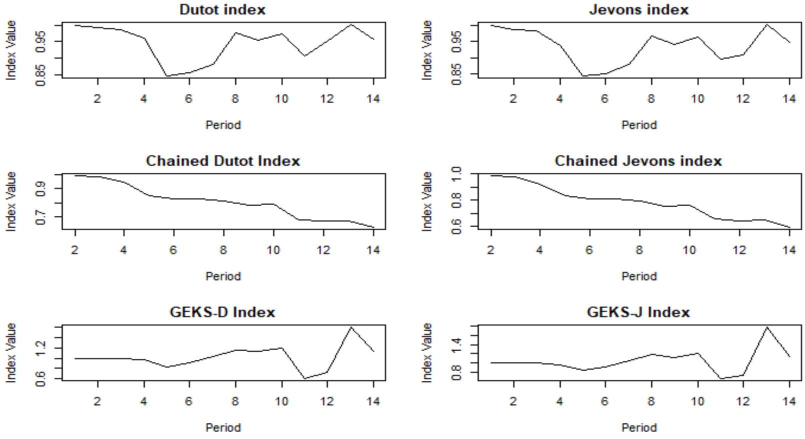
Source: Own calculations by using R.

Figure 10. Indexes values calculated using filter 1 for the male shoes' dataset



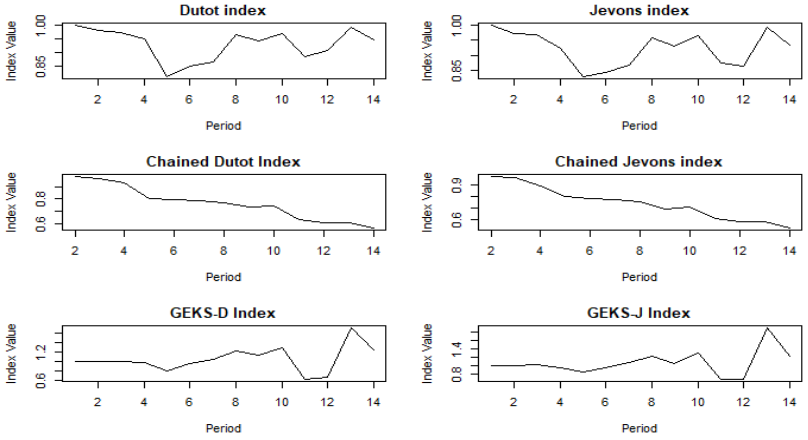
Source: Own calculations by using R.

Figure 11. Indexes values calculated using filter 2 for the male shoes' dataset



Source: Own calculations by using R.

Figure 12. Indexes values calculated using filter 3 for the male shoes' dataset



Source: Own calculations by using R.

To study the differences between the values of indexes for different filters more precisely, we present tables with the comparison of values between the first and last analyzed period.

In the case of base versions of the Dutot and Jevons indexes for ballerinas' dataset (Table 2), there are noticeable but not particularly marked differences between raw data and data with applied filters. The differences are more significant in the case of chained indexes, especially for data with applied filter 2 (Figure 3).

Table 2. Indexes values between first (t = 1) and last (t = 14) time period calculated for the ballerinas' base data and data with applied filters

Index	Base data	Filter 1	Filter 2	Filter 3
Jevons Index	0.9625463	0.9494611	0.9702390	0.9550619
Dutot Index	0.9756015	0.9590124	0.9879700	0.9643163
Chained Jevons Index	0.5414944	0.5528503	0.6030146	0.5447422
Chained Dutot Index	0.5961454	0.6098417	0.6532415	0.6065576
GEKS-J Index	1.4183390	1.4098770	1.6185490	1.3503640
GEKS-D Index	1.4862050	1.4404860	1.5456170	1.4178890

Source: Own calculations by using R.

For Fixed based indexes, the largest index value can be noticed for filter 2 (Figure 3), and the lowest for filter 1 (Figure 2). In the case of chained indexes, the most considerable price drop can be noticed for data with applied filter 3 (Figure 4), and the lowest for filter 2 (Figure 2).

In the case of GEKS indexes, the most significant price rise in the last period can be observed for filter 2 (Figure 3) and the lowest for filter 3 (Figure 4).

Table 3. Indexes values between first (t = 1) and last (t = 14) time period calculated for the male oxfords' base data and data with applied filters

Index	Base data	Filter 1	Filter 2	Filter 3
Jevons Index	0.9289061	0.9200919	0.9373305	0.9288719
Dutot Index	0.9597300	0.9575648	0.9655342	0.9641742
Chained Jevons Index	0.5056607	0.5074591	0.5709610	0.4970819
Chained Dutot Index	0.5562005	0.5594460	0.6197308	0.5516299
GEKS-J Index	1.0023470	0.9701082	0.9841282	0.9686279
GEKS-D Index	1.1317280	1.0938870	1.0023470	1.0914520

Source: Own calculations by using R.

Similar to the ballerinas' dataset, in the case of male oxfords (Table 3), we can notice that for Dutot and Jevons indexes we can achieve the highest index value for filter 2 (Figure 7) and the lowest for filter 1 (Figure 6). For chained indexes, the sharpest drop in price dynamics can be observed in the case of filter 3 (Figure 8). For the GEKS-D index, the lowest price index value can be noticed for data with the usage of filter 2 (Figure 7).

For the male shoes' dataset (Table 4), as in the two previous product types, filter 2 (Figure 11) provides the lowest results for GEKS indexes. This is the opposite effect compared to the ballerinas' dataset when GEKS based formulas provided much higher results for filter 2 (Figure 7).

Table 4. Indexes values between first (t = 1) and last (t = 14) time period calculated for the male shoes' base data and data with applied filters

Index	Base data	Filter 1	Filter 2	Filter 3
Jevons	0.9460867	0.9323798	0.9461031	0.9344793
Dutot	0.9532173	0.9461723	0.9589734	0.9479962
Chained Jevons	0.5302576	0.5357866	0.5957619	0.5256066
Chained Dutot	0.5668681	0.5743239	0.6279515	0.5650978
GEKS-J	1.3041550	1.2267400	1.1440670	1.2363870
GEKS-D	1.3112510	1.2326160	1.1338170	1.2524090

Source: Own calculations by using R.

For base Jevons and Dutot version of indexes, the highest value can be observed in the case of filter 2 (Figure 11) and the lowest for filter 1 (Figure 10).

To sum up, in the case of GEKS indexes calculated on ballerinas' dataset, we can observe that for the first ten months the results were stable with one drop in value around the 4th month for the same reason as in traditional indexes.

A sudden drop can be noticed between the 10th and 11th period, which can be explained by Christmas and New Year's sales. After next two time periods, we can observe a sudden rise of values of indexes – to a 1.5 or even 1.7 level. This can be explained by introducing new collections at the beginning of spring. Similar results occurred in the case of male oxfords and male shoes. For this footwear, we can, however, observe additional value growth of GEKS indexes from the 5th period (July) to the 8th period (October).

For traditional versions of the Dutot and the Jevons indexes, we can observe that index values tend to stay between the 0.8 and 1 value with one sharp price drop between the 4th and 5th time period, which can be caused by summer sales. Chained indexes tend to show a stable decline in value with sharper drops in the 4th and 10th period due to sales mentioned in the case of mentioned indexes.

Even though filtered data provided results with fewer sharp changes in values of indexes, the trends in all the cases are similar. Therefore, we can say that outliers did not have a decisive influence that disrupted the overall trend in any of these cases.

5. Conclusions

Clothing and footwear are considered one of the most difficult groups of goods to measure price change indexes due to high product churn and many sales which impact the price stability. Product churn undermines the possibility to use the traditional Jevons and Dutot indexes. However, it is possible to use chained indexes and the GEKS indexes instead.

These indexes, however, are sensitive to big price changes. Price volatilities always lead to substantial differences between price index formulas. As observed in the case of both product groups, the results provided by the GEKS and chained versions of indexes were different. It is worth noticing that chained indexes exhibit downward drift similar to Australian Bureau of Statistics research (Australian Bureau of Statistics [ABS], 2018). Even though usage of filters altered the results mildly the issues are still visible, and we can confirm that the outliers did not had significant impact on the overall trend of price indexes. These findings could lead to the conclusion that even though both GEKS and chained indexes are giving promising results, they could be better suited for other COICOP groups with less price volatility and product churn. Therefore, it is recommended to proceed to other methods for calculating price dynamics for footwear and clothing.

Usage of web – scrapped data in calculation of inflation is still a new topic. The main contribution of this paper is showing the suitability of Jevons and Dutot based indexes for calculating price indexes for category with high product churn, such as footwear and clothing, and how eliminating outliers impacts the results. This this paper is one of the first to compare well known GEKS-Jevons index with an alternative GEKS index based on Dutot formula.

All analyzed data comes came from one online retailer so there might be differences in results when we compare it with different retailers. Further study is needed to compare GEKS and chained indexes with alternative approaches for footwear and clothing, such as CLIP method.

Acknowledgements

The author would like to thank Prof. Jacek Białek for valuable remarks on every stage of this paper's creation.

The data were acquired during the internship in the National Bank of Poland, where the author worked on the eCPI project. Special thanks to Paweł Macias and Damian Stelmasiak – main authors of the project.

References

Australian Bureau of Statistics [ABS]. (2018). *Web scraping in the CPI Australian Bureau of Statistics*. Retrieved from https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Australia_-_poster.pdf

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*. Paper presented on Ottawa Group Meeting. Retrieved from 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., & Bobel, A. (2019). *Comparison of price index methods for CPI measurement using scanner data*. Paper presented at the 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro, Brazil. Retrieved from https://eventos.fgv.br/sites/eventos.fgv.br/files/arquivos/u161/bialek_bobel_paper_2.pdf

Bitner, T., & Stech, G. (2019). *GUS: Big Data to nasz priorytet*. Wywiad z Dominikiem Rozkrutem, prezesem GUS [CSO: Big Data is our priority. An interview with Dominik Rozkrut, president of Central Statistical Office in Poland]. Retrieved from https://www.computerworld.pl/wywiad/GUS-Big-Data-to-nasz-priorytet,412891.html ten Bosch, O. (n.d.). *Uses of web scraping for official statistics ESTP course on big data sources – web, social media and text analytics*. Retrieved from https://circabc.europa.eu/sd/a/5e250346-44a9-471b-87f1-5b5ddb59aa77/1_Big%20Data%20Sources%20part3-Day%201-A%20Use.pdf

Cavallo, A. (2013). *Online vs official price indexes: Measuring Argentina's inflation* (Research Paper, No. 4975-12). Cambridge: MA: MIT Sloan. https://doi.org/10.2139/ssrn.1906704

Cavallo, A. (2017, January). Are online and offline prices similar? Evidence from large multi-channel retailers. *American Economic Review, 107*(1), 283-303. https://doi.org/10.1257/aer.20160542

Cavallo, A. (2018, March). Scraped data and sticky prices. *The Review of Economics and Statistics, 100*(1), 105-119. https://doi.org/10.1162/REST_a_00652

Cavallo, A., & Rigobon, R. (2016, Spring). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives, 30*(2), 151-178. https://doi.org/10.1257/jep.30.2.151

Chessa, A. G., & Griffioen, R. (2019). Comparing price indices and footwear for scanner data and web scraped data. *Economie et Statistique, 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*. Singapore: Singapore Department of Statistics. Retrieved from https://www.singstat.gov.sg/-/media/files/publications/reference/newsletter/ssnsep2016.pdf

Dutot, C. F. (1738). *Reflexions politiques sur les finances et le commerce* (tome 1). The Hague: Les Freres Vaillant et Nicolas Prevost.

Eurostat. (2021). *Internet purchases by individuals* [Data base]. Retrieved from https://ec.europa.eu/eurostat/web/digital-economy-and-society/data/database

International Labour Organization, International Monetary Fund, Organisation for Economic Co-operation and Development, Statistical Office of the European Communities, United Nations, The International Bank for Reconstruction and Development, The World Bank. (2004). *Consumer Price Index Manual: Theory and practice*. Retrieved from https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf

Jevons, W. S. (1865, June). On the variation of prices and the value of the currency since 1782. *Journal of the Statistical Society of London, 28*, 294-320. Retrieved from https://archive.org/details/jstor-2338419/mode/2up

Juszczak, A. (2021). Usage of scraped data in price dynamic measurement. *Acta Universitatis Lodziensis. Folia Oeconomica, 1*(352), 25-37. https://doi.org/10.18778/0208-6018.352.02

Lunnemann, P., & Wintr, L. (2006). *Are internet prices sticky?* (ECB Working Paper, No. 645). Frankfurt am Main: European Central Bank. Retrieved from https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp645.pdf

Macias, P., & Stelmasiak, D. (2018). Food inflation nowcasting with web scraped data (Working Paper, No. 302). Warsaw: NBP. Retrieved from https://www.nbp.pl/publikacje/materialy_i_studia/302_en.pdf

Office for National Statistics [ONS]. (2017). *Research indices using web scraped price data: August 2017 update*. Retrieved June 20, 2020, from https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedprice data/august2017update

Office for National Statistics [ONS]. (2020). *Using statistical distributions to estimate weights for web-scraped price quotes in consumer price statistics*. Retrieved March 11, 2021 from https://www.ons.gov.uk/economy/inflationandpriceindices/articles/using statisticaldistributionstoestimateweightsforwebscrapedpricequotesinconsumerprice statistics/2020-09-01

Persson, E. (2019). *Evaluating tools and techniques for web scraping*. Retrieved from https://www.diva-portal.org/smash/get/diva2:1415998/FULLTEXT01.pdf

Polidoro, F., Giannini, R., Lo Conte, R., 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*. Paper presented at the First International Conference on Advanced Research Methods and Analytics, Universitat Politecnica de València, València, Spain, July 6-7, 2016. https://doi.org/10.4995/CARMA2016.2016.3133

Van Loon, K., & Roels, D. (2018). *Integrating big data in the Belgian CPI*. Paper presented at Meeting of the group of experts on Consumer Price Indices in Geneva, Switzerland 7-9 May. Brussels: StatBel Belgium in Figures. Retrieved from https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf