Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
1(352) 2021

DOI: https://doi.org/10.18778/0208-6018.352.02

Adam Juszczak
Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Katedra Metod Statystycznych
adam.juszczak2@gmail.com

Zastosowanie danych scrapowanych w pomiarze dynamiki cen

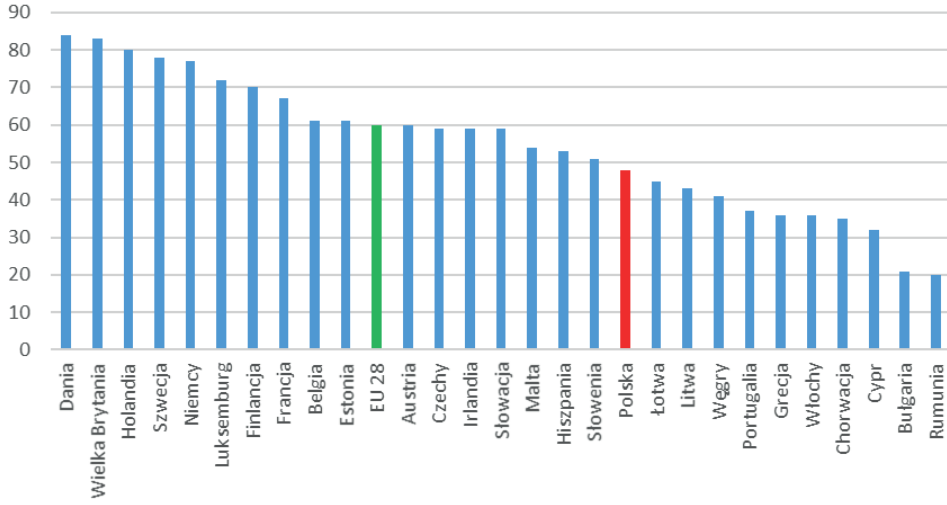
Streszczenie: Web-scraping to technika, którą można wykorzystać do automatycznego pozyskiwania danych zamieszczonych na stronach internetowych. Wraz ze wzrostem popularności zakupów on-line coraz więcej sklepów i usługodawców zainwestowało w strony WWW z ofertą cenową. Przekłada się to na możliwość automatycznego ściągania przez badaczy cen detalistów z wielu branż, m.in. odzieżowej czy spożywczej. Wykorzystanie danych scrapowanych skutkuje nie tylko znaczącym obniżeniem kosztów badania cen, ale także poprawia precyzję szacunków inflacji i daje możliwość śledzenia jej w czasie rzeczywistym. Z tego względu web-scraping jest dziś popularnym obiektem badań zarówno ośrodków statystycznych (Eurostat, brytyjski Office of National Statistics, belgijski Statbel), jak i uniwersytetów (m.in. Billion Prices Project prowadzony w Massachusetts Institute of Technology). Zastosowanie danych scrapowanych do liczenia inflacji wiąże się jednak z wieloma wyzwaniami na poziomie ich zbierania, przetwarzania oraz agregacji. Celem artykułu jest zbadanie możliwości wykorzystania danych scrapowanych do analizy dynamiki cen zabawek, a w szczególności porównanie wyników uzyskanych za pomocą różnych formuł indeksowych. W opracowaniu przedstawiono wynik badania empirycznego na podstawie danych pochodzących z czterech sklepów (z 53 wybranych produktów sprzedawanych w Amazonie, Wallmarcie, Smarterkids oraz KBkids).

Słowa kluczowe: inflacja, CPI, *web-scraping*, Jevons, Dutot, GEKS-J, GEKS-D, łańcuchowy Jevons, łańcuchowy Dutot, zakupy on-line, *big data*

JEL: C43, C49

1. Wprowadzenie

Web-scraping to technika, którą można wykorzystać do automatycznego pozyskiwania danych zamieszczonych na stronach internetowych. Wzrost popularności zakupów on-line (średnio w Unii Europejskiej 60% obywateli w 2018 roku dokonało przynajmniej jednego zakupu – wynik ten najwyższy był w Danii, a najniższy w Rumunii – por. wykres 1) przełożył się na większą liczbę sklepów oferujących sprzedaż swoich produktów w sieci. Zwiększyły się tym samym możliwości wydobycia informacji o cenach dóbr sprzedawanych przez detalistów, na czele z dużymi marketami i sklepami wysyłkowymi. Poza cenami podstawowymi często ze stron internetowych sklepów można uzyskać takie informacje, jak przeceny, opis produktu czy jego dostępność. Wykorzystanie technik web-scrapingowych daje także możliwość pozyskania znacznie większej ilości informacji w postaci całości dostępnej oferty sklepu. Umożliwia ono także obniżenie kosztów pomiaru i monitorowanie cen w czasie rzeczywistym, w praktyce pozwalając na tworzenie indeksów nawet codziennej częstotliwości zmian cen.



Wykres 1. Odsetek osób deklarujących w 2018 roku dokonanie przynajmniej jednego zakupu przez internet w ciągu ostatnich 12 miesięcy

Źródło: Eurostat

Należy jednak pamiętać, że wykorzystanie web-scrapingu stawia przed badaczami wiele wyzwań. Szacuje się, że około 15% koszyka inflacyjnego to dobra i usługi, których ceny nie są dostępne on-line. Przy budowaniu indeksu opartego na danych scrapowanych zmusza to do nieuwzględniania tych kategorii COICOP, na przykład poprzez równomierne rozdzielenie wagi brakującej kategorii na inne znajdujące się w tym samym agregacie (Radzikowski, Śmietanka, 2016).

Istotnym problemem w kontekście web-scrapingu jest także to, że nie otrzymujemy tu danych dotyczących wolumenu sprzedaży danego produktu (Cavallo, 2018). Widoczne jest to zwłaszcza w sektorach, w których produkty nie mają daty ważności. W efekcie towar popularny wśród klientów ma taką samą wagę przy budowaniu agregatu jak ten, którego sklep ma jedynie kilka sztuk w magazynie. W literaturze przedmiotu można jednak znaleźć prace proponujące sposoby ominięcia tego problemu. Na przykład niektórzy autorzy polecają ciekawe metody aproksymacji brakujących wag lub informacji o poziomie sprzedaży, oparte na liczbie odwiedzin danej witryny, liczbie tzw. polubień danego produktu lub na podstawie rozkładu prawdopodobieństwa poziomów konsumpcji w danej grupie produktów pochodzących z alternatywnych badań statystycznych (Chessa, Griffioen, 2019; Zhang 2020).

Dostępnych jest wiele programów służących do web-scrapingu, jednakże ze względu na specyfikacje kodu stron internetowych ciągle za najtrwalszą[1] metodę uznawane jest ręczne pisanie kodu do web-scrapingu konkretnej witryny. Najpopularniejszymi środowiskami programistycznymi do web-scrapingu cen produktów są zdecydowanie R i Python, które oferują wiele dedykowanych skryptów i bibliotek. W środowisku Python możliwe jest więc wykorzystanie API witryny WWW, symulowanie działań użytkownika na stronie przy użyciu pakietu Selenium lub też ściągniecie kodu strony i działanie off-line, jak to ma miejsce w przypadku pakietu Beautiful Soup.

2. Dotychczasowe badania

Dane scrapowane są jednym z dwóch najczęściej rozpatrywanych źródeł danych wspomagających proces liczenia inflacji (drugie to dane skanowane – patrz Białek, 2019). Temat web-scrapingu w mierzeniu inflacji poruszany był od połowy pierwszej dekady XXI wieku. Lunnemann i Wintr (2006), porównując ceny internetowe z cenami ze sklepów fizycznych, zauważyli różnicę między lepkością cen[2] w obu przypadkach. W 2008 roku uruchomiony został Billion Prices Project, który do tej pory działa w Massachusetts Institute of Technology (Cavallo, Rigobon, 2016). W ramach prac mierzono między innymi inflację cen on-line i porównywano do oficjalnej miary podawanej przez urzędy statystyczne krajów w Ameryce

1 Najtrwalszą, tj. odporną na zmiany w budowie kodu strony, z której zbieramy dane. Strony internetowe sklepów podlegają zmianom cały czas. Niektóre mają także dynamiczną budowę. Ręczne pisanie kodu pozwala uodpornić go na mniejsze zmiany (chociażby poprzez zastosowanie komend takich jak *try* i *except*, dzięki którym skrypt nie zakończy działania, gdy nie uda się pobrać danych konkretnego edytowanego produktu).

2 Także sztywność cen. Jest to powszechne zjawisko, które polega na opóźnionym dostosowywaniu się cen do sił podaży i popytu.

Południowej (Cavallo, 2013). Porównywano też ceny on-line z cenami w dużych fizycznych sklepach detalicznych (Cavallo, 2017) – okazuje się, że w 72% przypadków były one identyczne, jednak wykazano duże różnice między krajami. W Kanadzie i Wielkiej Brytanii aż 91% cen on-line odpowiadało cenom off-line, natomiast w Japonii i Brazylii odsetek zgodności wynosił poniżej 50%.

Możliwości włączenia danych scrapowanych do pomiaru CPI są też badane przez pracowników urzędów statystycznych. Prace w tej dziedzinie prowadzą między innymi badacze z urzędów statystycznych w Kanadzie, Niemczech, Holandii, Norwegii czy USA. Jeden z największych projektów tego typu prowadzony jest przez brytyjski urząd statystyczny (Office of National Statistics). Skupiając się zwłaszcza na – problematycznej z punktu widzenia liczenia inflacji – grupie odzieży i obuwia, rozwinął on między innymi indeksy oparte na metodzie CLIP (*Clustering large datasets into price indices*) (Office for National Statistics, 2017).

W Polsce temat zastosowania danych scrapowanych badany jest przez Narodowy Bank Polski w ramach projektu e-cpi, skupiającego się na prognozowaniu inflacji na podstawie bieżących danych. Dzięki szybkiemu pozyskiwaniu danych z web-scrapingu metody oparte na nowcastingu okazują się 11% mniej obciążone błędami od najlepszych modeli ARMA (Macias, Stelmiasiak, 2018).

Swoje badanie prowadzi w Polsce Centrum Analiz Społeczno-Ekonomicznych (CASE). Publikuje ono Online CASE CPI, zbierając dane z około 50 sklepów internetowych i pokrywając około 87% koszyka inflacyjnego (Radzikowski, Śmietanka, 2016).

Web-scraping to technologia mająca wiele zalet w porównaniu do tradycyjnego zbierania danych przez ankieterów. Jak wspomniano we wprowadzeniu, dane te zbierane są w sposób zautomatyzowany, z dużą częstotliwością, co pozwala na niemal natychmiastowe ich wykorzystanie. Proces ten jest także znacznie tańszy. Należy jednak pamiętać, że obejmuje on jedynie dużych detalistów mających swoje strony internetowe i ofertę on-line (zbierane są ceny ofertowe, a nie transakcyjne). Ciągle wiele zakupów dokonuje się w sklepach mniejszych, do których dotrzeć mogą jedynie ankieterzy (patrz Tabela 1).

Tabela 1. Porównanie cech pozyskiwania danych za pomocą web-scrapingu i przez tradycyjnych ankieterów

Web-scraping	Ankieterzy
– Dane zbierane automatycznie – Duża częstotliwość (najlepiej codzienna) – Tylko duże sklepy – Brak stron internetowych niektórych gałęzi usług (ok. 15%) – Konieczność przetworzenia dużych ilości danych – Zbieranie cen ofertowych, a nie cen transakcyjnych	– Reprezentacja zarówno dużych, jak i małych sklepów – Niższa częstotliwość zbierania danych i opóźnienia – Wyższy koszt – Dane dostępne z opóźnieniem – Dane dostosowane do koszyka inflacyjnego

Źródło: opracowanie własne

3. Indeksy cen wykorzystywane do analizy danych scrapowanych

Ze względu na to, że w danych scrapowanych mamy dostępne jedynie ceny, nie można użyć indeksów wymagających także danych o ilości kupowanych dóbr (czyli tzw. indeksów ważonych, takich jak m.in. indeks Laspeyresa, Fishera itp.). Z tego powodu do obliczeń używa się indeksów bazujących na indeksach Jevonsa oraz Dutot.

3.1. Indeks Jevonsa

Indeks Jevonsa jest tzw. indeksem bilateralnym, który porównuje okres bieżący z wybranym okresem poprzedzającym okres badany (Jevons, 1865). Okresem poprzedzającym może być na przykład grudzień poprzedzający rok analizy lub pierwszy dostępny okres w analizowanym zbiorze danych. Jest to wariant z ustalonym okresem bazowym (tzw. *fixed base approach*):

$$P_J^{0,t} = \prod_{j \in N0,t} \left(\frac{p_j^t}{p_j^0} \right)^{\frac{1}{card N_{0,t}}}, \; t = 1, 2, \ldots, T, \tag{1}$$

gdzie:

p_j^t – cena produktu j w okresie t,

p_j^0 – cena produktu j w okresie 0,

$N_{0,t}$ – produkty dostępne jednocześnie w okresie 0 (bazowym) i t (badanym).

Podstawową wadą indeksu Jevonsa jest to, że nie będzie on działać dobrze na rynkach, na których występuje duża rotacja produktów (a więc zbiór $N_{0,t}$ będzie się zmniejszał w czasie). Zjawisko to występuje zwłaszcza w zestawach danych zawierających długi szereg czasowy, gdyż im dalej od okresu bazowego, tym mniejsze prawdopodobieństwo, że dany produkt dalej będzie występował w ofercie sprzedażowej. Natomiast zaletą indeksu Jevonsa jest to, że w przeciwieństwie do dwóch pozostałych indeksów elementarnych, czyli indeksu Carliego oraz indeksu Dutot, jest on indeksem opartym na średniej geometrycznej. Z tego względu daje niższe wyniki niż indeksy oparte na średnich arytmetycznych z relatywnych cen. Spełnia on wiele wymaganych postulatów (tzw. testów – por. Balk, 1995), w tym kryterium współmierności i odwracalności w czasie (Lewel, 2015). Z tego względu jest on najpowszechniej używaną formułą elementarną, wykorzystywaną między innymi przez GUS do liczenia ogólnopolskich indeksów cen produktów (Białek, 2019).

3.2. Łańcuchowy indeks Jevonsa

To indeks, który uwzględnia wszystkie momenty czasowe z okna [0, *t*], tj. 0, 1, 2, 3, …, *t* – 1, *t*, przy czym stanowi iloraz wszystkich indeksów Jevonsa wyznaczonych dla sąsiadujących ze sobą okresów, tj.:

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

gdzie $P_J^{\tau-1,\tau}$ – indeks Jevonsa pomiędzy okresem analizowanym a okresem go poprzedzającym.

Łańcuchowy indeks Jevonsa, uwzględniając wszystkie momenty pośrednie między okresami 0 i *t*, jest bardziej adekwatny do analizy danych scrapowanych. Wynika to z faktu, że dane scrapowane charakteryzują się dużą rotacją produktów (dobra nowe i znikające) i dzieląc długie okno czasowe na serię dwuokresowych interwałów, dokonujemy o wiele mniejszej redukcji próby niż w przypadku indeksu Jevonsa o ustalonej podstawie $P_J^{0,t}$.

3.3. Indeks GEKS-J

Indeksy multilateralne pierwotnie wykorzystywano do porównań cenowych między krajami i regionami, ze względu na spełnianie aksjomatu przechodniości, co uniezależnia wynik obliczeń od wyboru kraju lub regionu służącego za podstawę (Białek, Bobel, 2019). Jednym z najpopularniejszych indeksów, obok indeksu Geary-Khamisa oraz CCDI, jest indeks GEKS, którego nazwa pochodzi od nazwisk jego

twórców – C. Giniego (1931), O. Eltetö i P. Kövesa (1964) oraz B. Szulca (1964). Indeks GEKS-J, będący odmianą indeksu GEKS, został zaproponowany stosunkowo niedawno, bo w 2009 roku (Ivancic, Fox, Diewert, 2011). W praktyce jest on geometryczną średnią łańcuchowych indeksów Jevonsa między okresem bazowym i okresem t z każdym pośrednim punktem ($i = 1, \ldots, t - 1$) w następujący sposób:

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

3.4. Propozycje indeksu alternatywnego

Można rozważyć akceptowany przez teoretyków (*Consumer Price Index Manual*..., 2004) inny indeks elementarny, mianowicie indeks Dutot (1738):

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

oraz jego łańcuchową wersję:

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

W artykule proponuje się modyfikację indeksu GEKS opartą na formule Dutot, tzn. indeks GEKS-D postaci:

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

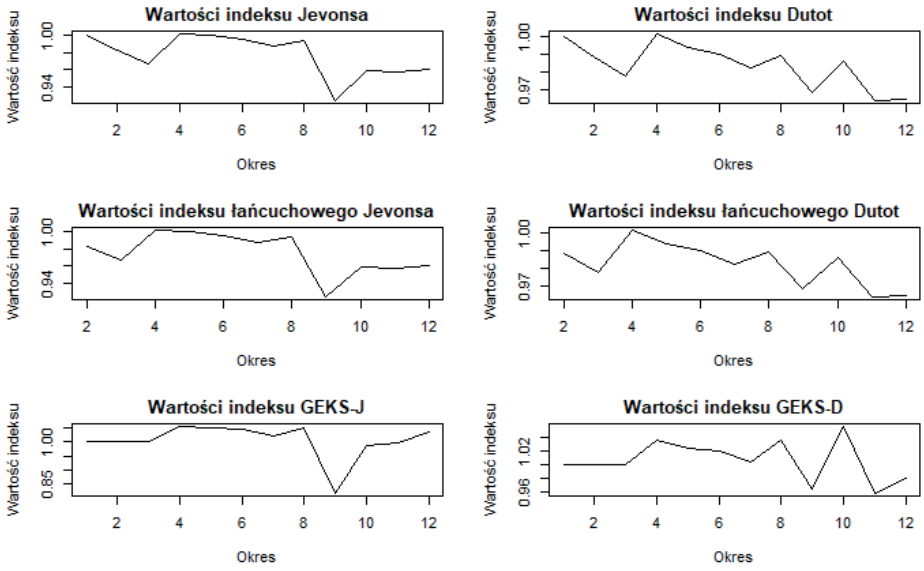
4. Badanie empiryczne

4.1. Opis źródeł danych

Dane wykorzystane do obliczeń zostały pozyskane ze źródła zewnętrznego (Yang, Gan, Tang, 2010). Są to dane scrapowane między sierpniem 2003 oraz styczniem 2004 roku w dwutygodniowych odstępach. W każdym przypadku ściągano ceny 53 ustalonych wcześniej zabawek ze stron internetowych czterech sklepów – Amazona, Wallmarta, Smarterkids oraz KBkids. Co istotne, w każdym analizowanym przypadku mamy pełne obserwacje dla wszystkich 12 okresów (brak luk w danych powszechnych dla danych scrapowanych). Wszystkie wartości analizowane są w stosunku do pierwszego okresu.

4.2. Rezultaty badania empirycznego

Za pomocą programu R wygenerowano wykresy zmiany wartości poszczególnych indeksów w stosunku do okresu bazowego (mającego wartość 1).

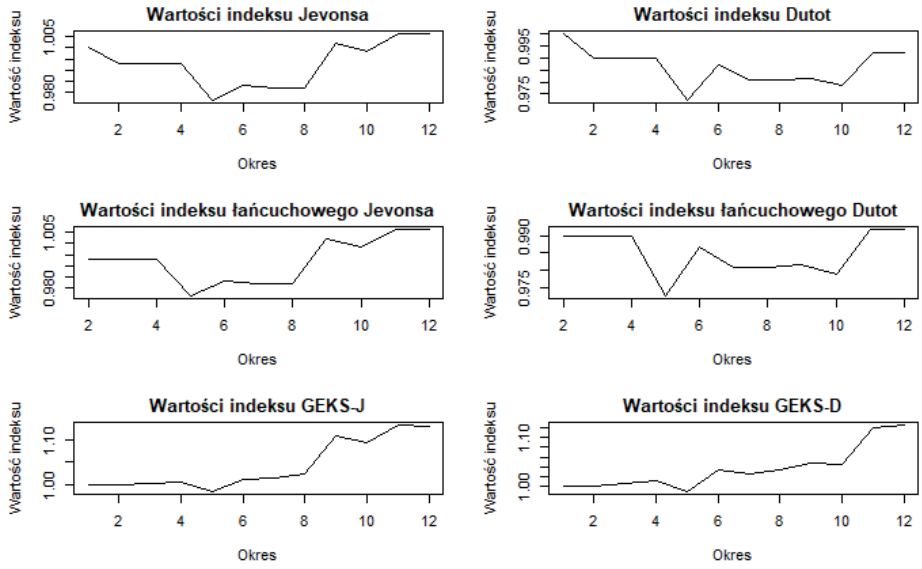


Wykres 2. Zestawienie wartości indeksów dla danych w sklepie Amazon

Źródło: opracowanie własne w programie R

Towary w sklepie Amazon wykazują dość wysoką niestabilność cenową. Po początkowym spadku widać powrót do poziomu wyjściowego. W dziewiątym okresie można zauważyć wysoki jednorazowy spadek cenowy związany prawdopodobnie z sezonowymi przecenami.

Wyraźnie wyższe różnice w dynamice cen zauważyć można w przypadku indeksów bazujących na formułach Jevonsa (w okresie spadku do wartości poniżej 0,85). Z kolei jedynie indeksy multilateralne GEKS uzyskują wartości wyższe niż 1, czyli wskazują na wzrost cen w stosunku do początkowego okresu.

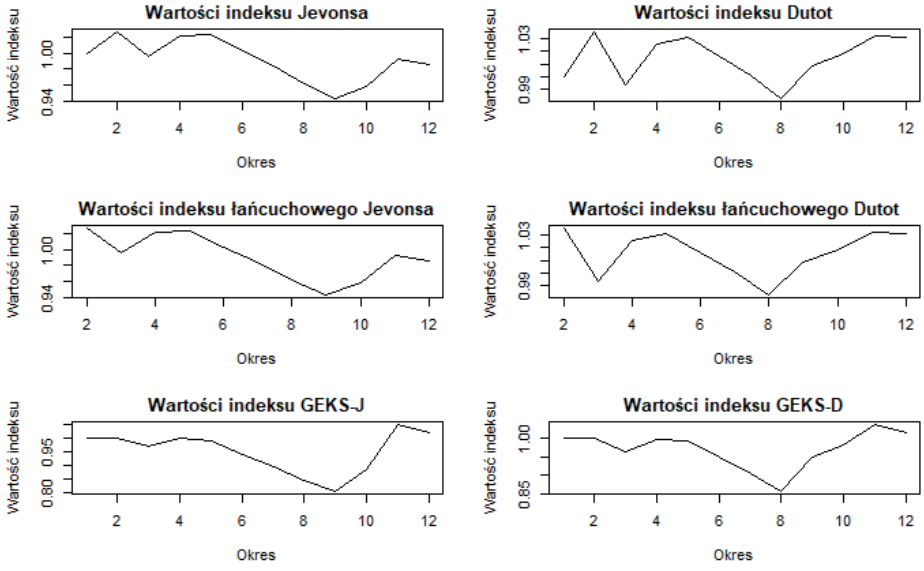


Wykres 3. Zestawienie wartości indeksów dla danych w sklepie Wallmart

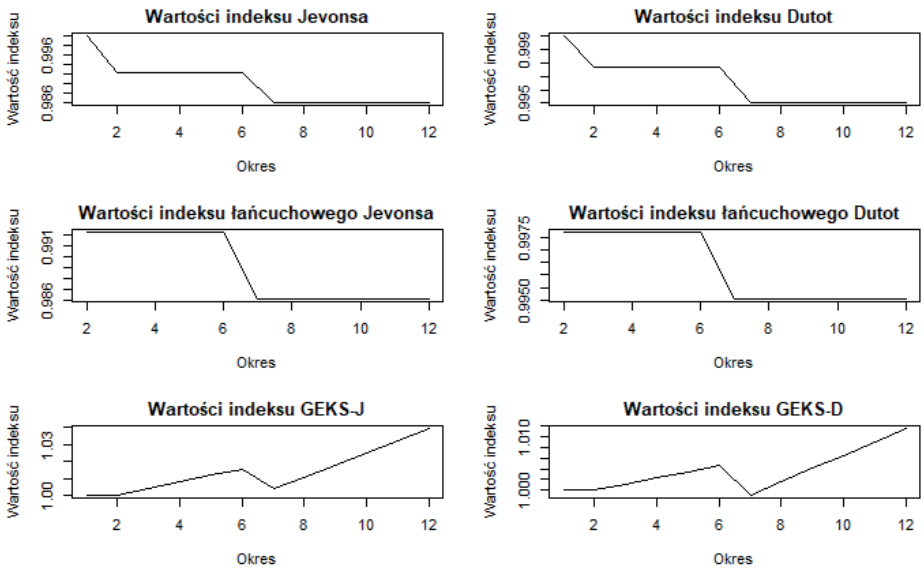
Źródło: opracowanie własne w programie R

Inną politykę cenową można zaobserwować dla zabawek oferowanych przez sklep Wallmart. W przypadku większości analizowanych indeksów jest ona znacznie bardziej konsekwentna. Po początkowym spadku widać znaczący wzrost w okresie okołoświątecznym. Warto zauważyć, że dla części indeksów (wszystkich indeksów opartych na formule Jevonsa oraz indeksu GEKS-D) wartości w końcowych okresach są wyższe od okresu bazowego nawet o 0,05–0,10.

W sklepie KBKids, podobnie jak w przypadku sklepu Wallmart, dla części analizowanych indeksów można zauważyć w miarę stabilną politykę cenową w początkowych analizowanych okresach. Następnie w okolicach końca października i listopada widać dość istotny spadek cen, który jest bardziej dynamiczny dla indeksów multilateralnych niż dla pozostałych.



Wykres 4. Zestawienie wartości indeksów dla danych w sklepie KBKids
Źródło: opracowanie własne w programie R



Wykres 5. Zestawienie wartości indeksów dla danych w sklepie SmarterKids
Źródło: opracowanie własne w programie R

Najstabilniejsza polityka cenowa zauważalna jest w przypadku sklepu SmarterKids. Dla indeksów łańcuchowych oraz zwykłych widoczne są jedynie dwie obniżki – w drugim oraz siódmym okresie. Inaczej sytuacja przedstawia się w przypadku indeksów multilateralnych, które wykazują raczej tendencję wzrostową.

5. Podsumowanie i rekomendacje

Między analizowanymi sklepami widać dużą różnicę w prowadzonej polityce cenowej. Zdecydowanie najbardziej skoordynowana wydaje się polityka sklepu SmarterKids, natomiast największą zmienność cen wykazuje sklep Amazon.

Indeksy multilateralne oparte na indeksach Dutota i Jevonsa wykazują się znacząco większą zmiennością niż reszta indeksów. Podczas gdy podstawowe i łańcuchowe wersje obu badanych indeksów wykazywały odchylenia względem wartości bazowej (czyli 1) o kilka punktów procentowych, w przypadku indeksów w formule GEKS wynosiła ona nawet do 20 punktów procentowych. Wynika to najprawdopodobniej z wysokiej wrażliwości multilateralnych indeksów nieważonych na szoki cenowe oraz wartości nietypowe.

Autor planuje kontynuowanie badań wyżej wspomnianych indeksów przy ręcznym lub automatycznym sterowaniu szokami cenowymi oraz wartościami nietypowymi, a także filtrów danych. Ponadto w kolejnych badaniach warto sprawdzić wyżej wymienione formuły multilateralne na mniej jednorodnej grupie oraz w szerszym oknie czasowym, w którym można zaobserwować pojawianie się i znikanie kolejnych dóbr, z którym to zjawiskiem, wedle literatury, radzą one sobie znacznie lepiej od wersji bazowych oraz łańcuchowych.

Podziękowania

Praca prezentowana na konferencji MSA 2019 „Organizacja międzynarodowej konferencji Multivariate Statistical Analysis 2019 (MSA 2019)" – zadanie finansowane w ramach umowy 712/P-DUN/202019 ze środków Ministra Nauki i Szkolnictwa Wyższego przeznaczonych na działalność upowszechniającą naukę.

Autor chciałby podziękować prof. Jackowi Białkowi za wartościowe uwagi na poszczególnych etapach tworzenia artykułu.

Bibliografia

Balk B.M. (1995), *Axiomatic Price Index Theory: A Survey*, „International Statistical Reviews", vol. 63, s. 69–93.

Białek J. (2019), *Remarks on Geo-Logarithmic Price Indices*, „Journal of Official Statistics", vol. 35, no. 2, s. 287–317.

Białek J., Bobel A. (2019), *Comparison of Price Index Methods for the CPI Measurement Using Scanner Data*, 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro.

Cavallo A. (2013), *Online vs Official Price Indexes: Measuring Argentina's Inflation*, „Journal of Monetary Economics", vol. 60, no, 2, s. 152–165.

Cavallo A. (2017), *Are Online and Offline Prices Similar? Evidence from Large Multi-channel Retailers*, „American Economic Review", vol. 107, s. 283–303.

Cavallo A. (2018), *Scraped Data and Sticky Prices*, „The Review of Economics and Statistics", vol. 100, s. 105–119.

Cavallo A., Rigobon R. (2016), *The Billion Prices Project: Using Online Prices for Measurement and Research*, „Journal of Economic Perspectives", vol. 30, no. 2, s. 151–178.

Chessa A.G., Griffioen R. (2019), *Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data*, „Economics and Statistics: Big Data and Statistics", no. 509, s. 49–69.

Consumer Price Index Manual. Theory and practice (2004), International Labour Office, Geneva.

Dutot C.F. (1738), *Reflexions Politiques sur les Finances et le Commerce*, vol. 1, Les Freres Vaillant et Nicolas Prevost, The Hague.

Eltetö Ö., Köves P. (1964), *Egy nemzetközi összehasonlításoknál fellépő indexszámítási problémáról. On a Problem of Index Number Computation Relating to International Comparisons (in Hungarian)*, „Statisztikai Szemle", no. 42, s. 507–518.

Eurostat, https://ec.europa.eu/eurostat/web/digital-economy-and-society/data/database (dostęp: 10.02.2020).

Gini C. (1931), *On the Circular Test of Index Numbers*, „Metron", no. 9, s. 3–24.

Ivancic L., Fox K.J., Diewert W.E. (2011), *Scanner Data, Time Aggregation and the Construction of Price Indexes*, „Journal of Econometrics", vol. 151, s. 24–35.

Jevons W. (1865), *The Coal Question*, Macmillan & Co., London.

Lewel P. (2015), *Is the Carli index flawed? Assessing the case for the new retail price index RPIJ*, „Journal of the Royal Statistical Society Series A (Statistics in Society)", vol. 178, no. 2, s. 303–336.

Lunnemann P., Wintr L. (2006), *Are Internet Prices Sticky?*, ECB Working Paper, no. 645.

Macias P., Stelmasiak D. (2018), *Food inflation nowcasting with web scraped data*, NBP Working Paper, no. 302.

Office for National Statistics (b.r.), *ONS methodology working paper series number 12 – a comparison of index number methodology used on UK web scraped price data*, https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber12acomparisonofindexnumbermethodologyusedonukwebscrapedpricedata (dostęp: 1.02.2020).

Office for National Statistics (2017), *Research indices using web scraped price data: clothing data*, https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clothingdata (dostęp: 1.02.2020).

Radzikowski B., Śmietanka A. (2016), *Online CASE CPI*, First International Conference on Advanced Research Methods and Analytics, València.

Szulc B. (1964), *Indices for Multiregional Comparisons*, „Przegląd Statystyczny", nr 3, s. 239–254.

Yang Z., Gan L., Tang F. (2010), *A Study of Price Evolution in the Online Toy Market. Economics*, „Open-Assessment E-Journal", vol. 4, no. 28, s. 1–29.

Zhang L. (2020), *Proxy expenditure weights for Consumer Price Index: audit sampling inference for big-data statistics*, „Journal of the Royal Statistical Society: Series A (Statistics in Society)", https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/rssa.12632 (dostęp: 10.02.2020).

Usage of scraped data in price dynamic measurement

Abstract: Web-scraping is a technique used to automatically extract data from websites. After the rise-up of on-lines shopping (which results in more shops posting their full price offer on their websites) it allows to acquire information about prices of goods sold by the retailers such as supermarkets or internet shops. Usage of web-scraped data allows to lower the costs, improve the measurement quality and monitor the price change in real time. Due to before mentioned reasons this method became the object of research studies from both statistical offices (Eurostat, British Office of National Statistics, Belgium Statbel) and universities (for ex. Billion Prices Project conducted on MIT). However, usage of scrapped data for the CPI calculation entails with multiple challenges with their collection, processing and aggregation. The purpose of this article is to examine the possibility of using scrapped data in toy price dynamic analysis. Especially the purpose is to compare the results from different index formulas. In this article the empirical study based on data from 4 different shops is presented (53 chosen products sold in Amazon, Wallmart, Smarterkids and KBKids).

Keywords: inflation, CPI, Web-scraping, GEKS-J, Jevons, Dutot, GEKS-D, Chained Jevons, Chained Dutot, online shopping, Big data

JEL: C43, C49

Received: 2020-03-22; verified: 2020-10-02. Accepted: 2021-03-01