Wiadomości Statystyczne. The Polish Statistician, 2021, vol. 66, 9, 32–69 DOI: 10.5604/01.3001.0015.2699
Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges
Zgłoszony/submitted: 14.06.2021, poprawiony/revised: 18.08.2021, zaakceptowany/accepted: 31.08.2021

Porównanie cen i wskaźników cen konsumpcyjnych: tradycyjna metoda uzyskiwania danych a źródła alternatywne[1]

Jacek Białek[a], Alina Dominiczak-Astin[b], Dorota Turek[c]

Streszczenie. Jednym z większych wyzwań stojących przed statystyką publiczną w XXI w. jest wykorzystanie alternatywnych źródeł danych o cenach w celu unowocześnienia statystyki cen konsumpcyjnych, a w rezultacie – zwiększenia dokładności i rzetelności danych o inflacji. Trudności w zbieraniu danych metodą tradycyjną spowodowane przez COVID-19 (obostrzenia dotyczące utrzymywania dystansu, które ograniczyły wyjścia ankieterów w teren, i zamykanie punktów sprzedaży) wpłynęły na zintensyfikowanie prac nad alternatywnymi źródłami danych. W artykule przedstawiono wyniki badania eksperymentalnego, w którym wykorzystano dane o cenach uzyskane metodą tradycyjną (przez ankieterów) oraz dane skanowane i skrapowane, pochodzące z sieci handlowej działającej w Polsce. Głównym celem badania było określenie występowania i oszacowanie wielkości różnic w poziomie cen i wartościach wskaźnika cen wybranych produktów spożywczych obliczonych metodą tradycyjną oraz z wykorzystaniem alternatywnych źródeł danych, czyli danych skanowanych i skrapowanych. Za dodatkowy cel postawiono sobie zidentyfikowanie przyczyn tych różnic w odniesieniu do specyfiki źródeł danych.

Badaniem empirycznym objęto luty i marzec 2021 r. Wyniki otrzymane na podstawie danych z różnych źródeł porównano za pomocą metod graficznych (histogramy, wykresy pudełkowe) oraz wyznaczenia elementarnych indeksów według formuł Dutota, Carliego i Jevonsa. Wyniki wskazały na rozbieżności – niekiedy znaczne – w rozkładach cen uzyskanych z różnych źródeł danych, co skłania do wniosku, że zastosowanie danych skanowanych i skrapowanych może prowadzić do zawyżania lub zaniżania wskaźników cen uzyskanych metodą tradycyjną.

W artykule omówiono również podstawowe aspekty metodologiczne dotyczące uzyskiwania i wykorzystywania danych ze źródeł alternatywnych oraz wskazano prawdopodobne przyczyny różnic, jakie zaobserwowano zarówno w rozkładach cen produktów, jak i w wartościach miesięcznego wskaźnika cen obliczonego przy wykorzystaniu danych z różnych źródeł.

Słowa kluczowe: wskaźniki cen, dane skanowane, dane skrapowane, inflacja

JEL: C43, E31

[1] Artykuł powstał w związku z realizacją projektu „Budowa zintegrowanego systemu statystyki cen (INSTATCENY)", finansowanego przez Narodowe Centrum Badań i Rozwoju (1. edycja GOSPOSTRATEG, No. 1/382525/14/NCBR/2018). / The article has been written in connection with the 'Creation of an integrated system of price statistics (INSTATCENY)' project, financed by the National Centre for Research and Development (1st edition of GOSPOSTRATEG, No. 1/382525/14/NCBR/2018).

[a] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Katedra Metod Statystycznych; Główny Urząd Statystyczny, Departament Handlu i Usług, Polska / University of Lodz, Faculty of Economics and Sociology, Department od Statistical Methods; Statistics Poland, Department of Trade and Services, Poland. ORCID: https://orcid.org/0000-0002-0952-5327. E-mail: jacek.bialek@uni.lodz.pl.

[b] Główny Urząd Statystyczny, Departament Handlu i Usług, Polska / Statistics Poland, Department of Trade and Services, Poland. ORCID: https://orcid.org/0000-0001-5557-3699. E-mail: a.dominiczak@stat.gov.pl.

[c] Główny Urząd Statystyczny, Departament Handlu i Usług, Polska / Statistics Poland, Department of Trade and Services, Poland. ORCID: https://orcid.org/0000-0002-8984-8563. Autor korespondencyjny / Corresponding author, e-mail: d.turek@stat.gov.pl.

Comparison of prices and consumer price indices: traditional data collection and alternative data sources

Abstract. One of the major challenges official statistics is faced with in the 21st century is the use of alternative sources of price data in order to modernise consumer price statistics and, as a result, to improve the accuracy and reliability of inflation data. Data collecting based on the traditional method encountered numerous difficulties caused by COVID-19 (distance-keeping restrictions limiting price collectors' fieldwork, closures of points of sale). As a consequence, the work on alternative data sources intensified. The article presents the results of an experimental study involving the use of prices collected by means of the traditional method (by price collectors), and scanner and web scraped data from one of the retail chains operating in Poland. The aim of the study was to investigate the occurrence of differences in prices and price indices of selected food products and to estimate them, using the traditional method and alternative data sources, i.e. scanner and web scraped data. An additional goal was set to identify source-based reasons for these differences.

The empirical study covered the period of February and March 2021. The results based on data from different sources were compared using both graphical methods (histograms, box plots) and the calculation of elementary price indices according to the Dutot, Carli and Jevons formulas. The findings revealed certain, sometimes serious discrepancies in the distributions of prices obtained from various data sources, which suggests that the application of scanner and web scraped data may lead to the over- and understating of price indices obtained via the traditional method.

The article also discusses the main methodological aspects of obtaining and applying data from alternative sources, and indicates the probable causes of the differences observed both in distributions of product prices and in monthly price indices calculated using data from various sources.

Keywords: price indices, scanner data, web scraped data, inflation

1. Wprowadzenie

Dwie najistotniejsze miary inflacji to wskaźnik cen towarów i usług konsumpcyjnych (Consumer Price Index – CPI) oraz zharmonizowany wskaźnik cen konsumpcyjnych (Harmonised Index of Consumer Prices – HICP)[2]. Przygotowanie koszyka reprezentantów towarów i usług, stanowiącego podstawę obliczania CPI i HICP[3], jest punktem wyjścia badań cen konsumpcyjnych. Produkty w koszyku reprezentują najniższy szczebel agregacji w klasyfikacji spożycia indywidualnego według celu (Classification of Individual Consumption by Purpose – COICOP), która została

[2] Inne miary inflacji to m.in. wskaźnik inflacji bazowej (Core Inflation Rate – CIR), stosowany przez Narodowy Bank Polski w realizacji polityki pieniężnej, deflator PKB, a w zakresie przetwórstwa przemysłowego – wskaźnik cen producentów (Producer Price Index – PPI).

[3] Koszyki towarów i usług stosowane w obliczeniach CPI i HICP różnią się w niewielkim stopniu. Przykładowo w CPI są ujęte gry losowe, a w HICP – nie.

opracowana przez Organizację Narodów Zjednoczonych i zaadaptowana na potrzeby HICP przez Eurostat jako COICOP/HICP. Główny Urząd Statystyczny do 2013 r. stosował czteropoziomową (czterocyfrową) klasyfikację COICOP/HICP w podziale na działy, grupy i klasy. W związku z wdrożeniem przez Eurostat pięciopoziomowej Europejskiej Klasyfikacji Spożycia Indywidualnego według Celu (European Classification of Individual Consumption according to Purpose – ECOICOP) klasyfikacja została uszczegółowiona do piątego, a w zakresie niektórych kategorii – do szóstego szczebla agregacji. Najniższy poziom ECOICOP, dla którego możliwe jest ustalenie wydatków gospodarstw domowych na zakup produktów konsumpcyjnych, to grupa elementarna, przy czym krajowe urzędy statystyczne mogą wprowadzać niższe szczeble agregacji (podgrupy), jeśli dostępne są odpowiednie dane statystyczne. Do każdej grupy elementarnej dobiera się reprezentanty. Przykładowo w badaniach GUS w ramach klasy ECOICOP Pieczywo i produkty zbożowe (01.1.1) wyróżniono podklasę Ryż (01.1.1.1), której reprezentantami są ryż biały i ryż długoziarnisty. Badanie cen towarów i usług konsumpcyjnych prowadzone w GUS w 2021 r. obejmuje 338 grup elementarnych i ok. 1800 reprezentantów.

Tradycyjna metoda uzyskiwania danych polega na zbieraniu przez ankieterów informacji o cenach i cechach reprezentantów w punktach sprzedaży wytypowanych do badania w rejonach notowań (w Polsce jest 207 rejonów notowań oraz ok. 35 tys. punktów sprzedaży odwiedzanych przez ankieterów). Po zebraniu i przeanalizowaniu danych, wprowadzeniu ewentualnych korekt i zatwierdzeniu zbioru dla każdego reprezentanta obliczana jest relacja cen w miesiącu badanym do cen w miesiącu poprzednim. Wskaźnik cen dla grup elementarnych wyznacza się jako średnią geometryczną z wartości wskaźnika cen produktów reprezentantów. Agregacja na wyższe szczeble ECOICOP następuje z uwzględnieniem wag[4] określających udział spożycia poszczególnych grup towarów w wydatkach konsumpcyjnych gospodarstw domowych ogółem.

CPI i HICP są wskaźnikami typu Laspeyresa (1871). Postępująca automatyzacja transakcji rynkowych, dająca coraz większe możliwości tworzenia baz danych o produktach, które potencjalnie mogą stanowić cenne wsparcie w realizacji badań statystycznych, skłoniła statystyków do przeglądu formuł indeksowych i ponownej oceny możliwości ich zastosowania w opracowywaniu wskaźników cen.

Trudności w zbieraniu danych metodą tradycyjną spowodowane przez pandemię COVID-19 – konieczność utrzymywania dystansu ograniczająca pracę ankieterów w terenie i zamykanie punktów sprzedaży – przy równoczesnym zwiększeniu wolu-

[4] Źródło informacji o strukturze wydatków konsumpcyjnych niezbędnych do opracowania CPI jest inne niż stosowane w HICP. Wagi dla CPI ustala się na podstawie badania budżetów gospodarstw domowych, a w przypadku HICP – na podstawie statystyki rachunków narodowych.

menu zakupów konsumenckich w internecie przyczyniły się do zintensyfikowania prac nad uzyskiwaniem danych ze źródeł alternatywnych – danych skanowanych (ang. *scanner data*) i danych skrapowanych (ang. *web scraped data*), które w pomiarze CPI są wykorzystywane od blisko 20 lat i stają się coraz bardziej powszechne. Organizacje międzynarodowe, urzędy statystyczne oraz środowiska naukowe na całym świecie prowadzą prace nad ich właściwym wykorzystaniem[5], a pandemia znacznie przyspieszyła te działania. W 2020 r. zaktualizowano podręcznik do CPI (International Monetary Fund [IMF] i in.)[6], uwzględniając m.in. alternatywne źródła danych. Obecnie trwają prace nad nową wersją podręcznika do HICP (Eurostat, 2018), który docelowo ma dostarczać rekomendacji w zakresie wykorzystania danych skrapowanych i skanowanych do pomiaru inflacji w UE. Warto wspomnieć, że do 2015 r. wśród krajów europejskich tylko Holandia, Norwegia, Szwecja i Szwajcaria wykorzystywały dane skanowane do obliczeń wskaźników cen konsumpcyjnych, a rok później dołączyły do nich Belgia, Dania i Islandia. Obecnie również Luksemburg, Portugalia, Niemcy i Francja prowadzą badania z wykorzystaniem danych skanowanych i skrapowanych dla wybranych podgrup koszyków CPI i HICP.

Polskie konsorcjum, w skład którego wchodzą: GUS, Instytut Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN) i Szkoła Główna Handlowa w Warszawie (SGH) – od ponad dwóch lat realizuje projekt *InstatCeny*, ukierunkowany na wykorzystanie alternatywnych źródeł danych przy opracowywaniu CPI. Niniejszy artykuł przedstawia wstępne wyniki badania obejmującego produkty z segmentu spożywczego, które uzyskano w ramach tego projektu.

Głównym celem badania omawianego w artykule jest określenie występowania i oszacowanie wielkości różnic w poziomie cen i wartościach wskaźnika cen wybranych produktów spożywczych obliczonych metodą tradycyjną oraz z wykorzystaniem alternatywnych źródeł danych, czyli danych skanowanych i skrapowanych. Za dodatkowy cel postawiono sobie zidentyfikowanie przyczyn tych różnic w odniesieniu do specyfiki źródeł danych. W artykule omówiono również podstawowe aspekty metodologiczne dotyczące uzyskiwania i wykorzystywania danych ze źródeł alternatywnych oraz wskazano prawdopodobne przyczyny różnic pomiędzy źródłami danych, jakie zaobserwowano zarówno w rozkładach cen produktów, jak i w wartościach wskaźnika cen.

[5] Szerzej o genezie danych skanowanych i ich wykorzystaniu do pomiaru CPI przez różne kraje pisali: Bertoloto i in. (2014); Białek (2020a); Białek i Bobel (2019); Chessa (2015, 2016); Diewert i Fox (2018); de Haan (2006); Kalisch (2016) oraz Loon i Roels (2018).

[6] Prace nad aktualizacją podręcznika, trwające kilka lat, prowadzono w Międzynarodowym Funduszu Walutowym, Międzynarodowej Organizacji Pracy, Eurostacie, Europejskiej Komisji Gospodarczej, Organizacji Współpracy Gospodarczej i Rozwoju oraz Banku Światowym.

2. Uzyskiwanie przez GUS danych skanowanych

Zgodnie z definicją przyjętą w podręczniku do CPI przez dane skanowane rozumie się szczegółowe dane o produktach konsumpcyjnych uzyskane dzięki skanowaniu kodów kreskowych w punktach sprzedaży. Do ich wygenerowania dochodzi w elektronicznych terminalach zlokalizowanych w punktach oferujących określone dobra, które dostarczają bardzo szczegółowych informacji zawartych w kodach kreskowych sprzedawanych produktów.

W sieciach handlowych najczęściej stosowane są następujące kody kreskowe: GTIN (Global Trade Item Number) lub jego europejska wersja EAN (European Article Number), PLU (Price Look-Up) albo SKU (Stock Keeping Unit). Najbardziej rozpowszechniony jest kod GTIN (EAN), choć na świecie w użyciu są też specyficzne kody kreskowe, np. UPC (Universal Product Code) czy lokalny APN (Australian Product Number). Kody wraz z etykietami produktów stanowią podstawę zaklasyfikowania produktów do grup ECOICOP 5 oraz niższych szczebli agregacji.

Forma uzyskiwania danych skanowanych i ich zakres różnią się w zależności od dostawców (sieci handlowych). W GUS stosuje się bezpieczne (szyfrowane) transfery, a w przypadku jednej sieci handlowej pobór danych odbywa się bezpośrednio przez udostępnione przez tę sieć API (Application Programming Interface). Poza kodami identyfikującymi produkt pobierana jest ramka danych w formacie csv, zawierająca: etykietę produktu (dodatkowy opis), jednostkę sprzedaży[7] (np. szt., kg, paczka, g, l), datę transakcji, cenę sprzedaży, wartość sprzedaży, liczbę sprzedanych jednostek produktu (opcjonalnie), flagę (flaguje się np. produkty z przecen i promocji) oraz dodatkowo informację o VAT.

Zakres zbieranych danych skanowanych zależy od dostawcy. Jeśli chodzi np. o wolumen, to jedna z sieci współpracujących z GUS dostarcza danych dotyczących jedynie 10 grup produktów z kategorii spożywczej, natomiast inna udostępnia cały swój asortyment. Przeciętnie pojedyncza sieć handlowa, dysponująca ok. 10–15 tys. kodów kreskowych, posiada kilkaset punktów sprzedaży w Polsce i dostarcza od kilku do kilkudziesięciu tysięcy rekordów miesięczne w przypadku każdej elementarnej grupy produktów. Oznacza to, że każdego miesiąca GUS pobiera od każdej sieci 40–700 MB danych, które następnie poddawane są wnikliwej analizie. Ustalenie przeciętnych cen podgrup produktów utworzonych na podstawie listy reprezentantów jest poprzedzane szeregiem czynności przygotowujących dane do wykorzystania w obliczeniach. Po wstępnym wyczyszczeniu zbioru danych (ujednoliceniu nazw, usunięciu błędnych danych i nietypowych cen) produkty są przypisywane do odpowiednich grup ECOICOP 5 i poziomu krajowego ECOICOP 6. Klasyfikacja produk-

[7] W przypadku jednej sieci handlowej współpracującej z GUS niezbędne jest pobieranie informacji o gramaturze i jednostce sprzedaży z opisu produktu znajdującego się na etykiecie.

tów odbywa się przy użyciu oprogramowania dostarczonego przez IPI PAN, przy czym algorytm klasyfikacyjny wykorzystuje metody uczenia maszynowego (tzw. naiwny Bayes[8]), bazując m.in. na etykietach i kodach produktów. Po przyporządkowaniu produktów do odpowiednich kodów klasyfikacji dopasowuje się produkty sprzedawane w porównywanych okresach (ang. *matching*). Za dopasowane produkty uważane są takie, które odpowiadają sobie jakościowo, ale mogą się różnić np. kolorem opakowania czy dodatkowym opisem na opakowaniu. Następnie przeprowadza się filtrowanie produktów w celu wyeliminowania nietypowych lub nieistotnych obserwacji. Obecnie GUS implementuje trzy rodzaje filtrów danych: filtr ekstremalnych cen (ang. *extreme price filter*), filtr niskiego wolumenu sprzedaży (ang. *low sales filter*) i filtr nieistotnych cen (ang. *dump price filter*). Ostatni ma za zadanie wyeliminować z próby te produkty, które w najbliższym czasie najprawdopodobniej znikną z półek sklepowych, ponieważ pomimo spadku ceny ich sprzedaż maleje (zob. Loon i Roels, 2018). Procesy filtrowania danych i obliczania wskaźników cen są przeprowadzane w pakiecie PriceIndices napisanym w języku R (Białek, 2021). Aplikacja dostarczona przez IPI PAN korzysta z niego dzięki zastosowaniu technologii *docker* i konteneryzacji pakietu.

3. Uzyskiwanie przez GUS danych skrapowanych

Na początku 2021 r. GUS rozpoczął skrapowanie witryn dwóch sieci handlowych w celu uzyskania danych o cenach produktów spożywczych[9]. Skrapowanie odbywa się z wykorzystaniem oprogramowania dostarczonego przez IPI PAN, które powstało na bazie pythonowskiej biblioteki Selenium. Zgodnie z protokołem dobrych praktyk właściciele witryn, z których pobierane są dane, zostali o tym fakcie powiadomieni, a skrapowanie odbywa się we wczesnych godzinach rannych, aby nie obciążać serwera sieci handlowej. Programy skrapujące pracują codziennie, a pobrane dane są zapisywane i archiwizowane w postaci plików JSON. Jedna ze skrapowanych sieci dostarcza GUS również dane skanowane, co dało podstawę do porównań prezentowanych w artykule. Okazało się, że za pośrednictwem strony internetowej sieci można nabyć 40–90% produktów tej samej kategorii dostępnych w stacjonarnych punktach sprzedaży. Przykładowo w kategorii ryż na początku 2021 r. wśród danych skanowanych zarejestrowano 33 produkty, a wśród danych skrapowanych – 27. W przypadku kawy relacja była podobna – wynosiła 275 do 152. Brak pełnego

[8] Metoda ta, szerzej opisana np. w pracy Domingosa i Pazzaniego (1997), polega na konstruowaniu prostych klasyfikatorów dla modeli, które przypisują etykiety do badanych klas obiektów na podstawie wektorów wartości cech tych obiektów. Zakłada ona, że każda cecha przyczynia się do klasyfikacji niezależnie od pozostałych cech.

[9] Niezależnie od tego Urząd Statystyczny w Opolu skrapuje witryny aptek i analizuje ceny wyrobów farmaceutycznych (grupa ECOICOP: 06.1.1.0.1).

pokrycia półek sklepowych wynika prawdopodobnie z tego, że na stronach internetowych wystawiane są produkty, które cieszą się największą popularnością lub których sprzedaż sieć chciałaby zwiększyć.

Zestawienia danych pochodzących ze skrapowania i skanowania są do siebie zbliżone. Warto jednak zwrócić uwagę na różnice: z jednej strony GUS skrapuje więcej informacji o produkcie, niż sieć wysyła w ramach podpisanego porozumienia (GUS dysponuje np. dodatkowo informacją o producencie, dostępności towaru, cenie regularnej i ewentualnej zredukowanej w wyniku rabatów i przecen), ale drugiej – nie zna ilości sprzedaży, a skrapowane dane dostarczają informacji jedynie o cenach ofertowych, które niekoniecznie są tożsame z cenami transakcyjnymi (podobnie jak w przypadku tradycyjnej metody pozyskiwania danych). Przed obliczeniem wskaźnika cen dane poddaje się takim samym procesom (przegląd, analiza, ewentualne korekty) jak dane skanowane. Produkty klasyfikuje się według funkcji *data_selecting* z pakietu PriceIndices (Białek, 2021) i/lub stosuje się metodę uczenia maszynowego z wykorzystaniem oprogramowania IPI PAN. Następnie dane są dopasowywane w czasie (funkcja *data_matching* z pakietu PriceIndices) i filtrowane (funkcja *data_filtering*). W GUS zaimplementowano jeden rodzaj filtrowania, a mianowicie filtr ekstremalnych cen (brak danych o konsumpcji wyklucza pozostałe rodzaje filtrów). Z próby usuwa się w szczególności te produkty, których cena z miesiąca na miesiąc wzrosła o więcej niż 200% lub spadła o ponad 75% (zob. Loon i Roels, 2018).

4. Metoda badania

Ze względu na zakres dostępnych danych[10] w badaniu – dla porównania wszystkich trzech źródeł danych – uwzględniono ceny obserwowane w lutym i marcu 2021 r. Wybrano 10 elementarnych grup produktów spożywczych: ryż, mąkę pszenną, pozostałe mąki, mleko pełne świeże, mleko świeże niskotłuszczowe, mleko zagęszczone i w proszku, jogurt, napoje i inne produkty mleczne, cukier oraz kawę.

Przeanalizowano ceny uzyskane za pomocą tradycyjnej metody zbierania danych pochodzące z 207 rejonów notowań dla każdego reprezentanta wymienionych grup, przy czym ceny – podobnie jak w przypadku cen skanowanych i skrapowanych – zostały przeskalowane do ustalonej jednostki miary (np. dla mleka był to litr, a dla ryżu – kilogram). Należy nadmienić, że rejonem notowań mogły być: miasto, część dużego miasta, gmina lub dzielnica[11]. Wybrano tylko te notowania cen, których dokonano w punktach sprzedaży detalicznej poza sieciami handlowymi. Nie zastosowano żadnych dodatkowych filtrów dla rejestrowanych cen, gdyż ankieter – na podstawie wywiadu i obserwacji – z założenia zapisuje ceny najbardziej typowe

[10] Jak już wspomniano, GUS skrapuje ceny produktów spożywczych dopiero od początku 2021 r.
[11] Rejony badania cen są ustalane i aktualizowane przez GUS we współpracy z urzędami statystycznymi.

w danym rejonie notowania. Dla każdego reprezentanta obliczono średnią (arytmetyczną) cenę w danym miesiącu oraz cząstkowy wskaźnik cen, wyznaczony jako iloraz średniej ceny z marca w stosunku do średniej ceny z lutego. Następnie, zgodnie z metodologią, wskaźnik cen dla grup elementarnych został obliczony jako średnia geometryczna z wartości wskaźników wyznaczonych dla reprezentantów tych grup.

W przypadku cen skanowanych i skrapowanych listę reprezentantów, na podstawie której utworzono podgrupy grup elementarnych, rozszerzono o trzy nowe pozycje: jogurt czekoladowo-owocowy, cukier puder oraz kawa mielona. Te podgrupy były na tyle licznie reprezentowane, przy wyraźnej homogeniczności pod względem jakości i zmienności cen produktów wchodzących w ich skład, że mimo rozbieżności w stosunku do obowiązującej listy reprezentantów postanowiono uwzględnić je przy ocenie odpowiadających im grup ECOICOP 5. W przypadku obu źródeł alternatywnych klasyfikację produktów do grup elementarnych i ich podgrup, jak również ich dopasowanie w czasie przeprowadzono na podstawie etykiet produktów oraz utworzonych wcześniej słowników słów kluczowych i fraz identyfikujących przynależność do tych grup[12].

Następnie próba produktów skanowanych została poddana filtrowaniu – w ten sposób usunięto zarówno ekstremalne miesięczne zmiany cen[13] (3% przypadków), jak i produkty o relatywnie niskiej sprzedaży[14] (w zależności od grupy nawet do 25% produktów). W przypadku danych skrapowanych zaimplementowano jedynie filtr ekstremalnych cen, z progami odcięcia omówionymi wcześniej, co właściwie nie wpłynęło na wielkość próby (usunięto zaledwie dwa produkty z grupy jogurtów). Należy zaznaczyć, że pojęcie *miesięcznej ceny* w przypadku danych skanowanych i skrapowanych nie jest tożsame oraz odbiega od ceny reprezentanta, którą ankieter notuje danego dnia w wybranym do badania punkcie sprzedaży. Za średnią miesięczną cenę produktu uzyskaną z danych skanowanych przyjmuje się wartości stanowiące iloraz łącznej wartości sprzedaży danego produktu i sumarycznej ilości jego sprzedaży z analizowanego miesiąca (ang. *unit value*). W przypadku danych skrapowanych, które są pobierane każdego dnia miesiąca (bez względu na to, czy produkt został sprzedany, czy nie), wyznacza się średnią arytmetyczną ze wszystkich obserwacji uzyskanych w danym miesiącu.

[12] Zastosowano funkcje *data_selecting* oraz *data_matching* z pakietu PriceIndices w środowisku R. Etykiety tekstowe porównywano miarą odległości Jaro-Winklera (Jaro, 1989; Winkler, 1990), przy czym ustaloną graniczną odległością, powyżej której uznawano dwie etykiety za różne, było 0,02.

[13] Miesięczną zmianę ceny uznawano za ekstremalną, jeśli oznaczała wzrost ceny przynajmniej o 300% lub spadek o ponad 75%.

[14] Graniczny, relatywny udział sprzedaży danego produktu względem sprzedaży ogółem określono za pomocą odwrotności z liczby produktów w danej grupie przemnożonej przez stałą o wartości 1,25.

W badaniu porównano średnie miesięczne ceny reprezentantów oraz miesięczne wskaźniki cen reprezentantów i odpowiadających im grup ECOICOP 5 z uwzględnieniem wszystkich omawianych źródeł danych. Mimo że obowiązująca metodologia obliczania wskaźnika cen dla grup elementarnych bazuje na średniej geometrycznej ze wskaźników cząstkowych (Jevons, 1865), analizę porównawczą dopełniono wyznaczeniem dwóch innych znanych z literatury elementarnych indeksów cen (ang. *elementary price indices*) według formuły Dutota (1738) oraz Carliego[15] (1804).

Wprowadzono następujące oznaczenia:

N – liczba dopasowanych produktów w porównywanych miesiącach (ang. *matched products*),

$\tau = 0$ – okres bazowy dla wskaźnika cen (luty 2021 r.),

$\tau = t$ – okres badany (marzec 2021 r.),

p_i^τ – cena *i*-tego produktu w okresie τ.

Indeks Dutota, opracowany jako pierwszy, stanowi iloraz średnich cen z miesiąca badanego i bazowego. Można go wyrazić jako:

$$P_D = \frac{\frac{1}{N}\sum_{i=1}^{N} p_i^t}{\frac{1}{N}\sum_{i=1}^{N} p_i^0}. \tag{1}$$

Z kolei indeks Carliego wyraża średnią arytmetyczną z indeksów cząstkowych:

$$P_C = \frac{1}{N}\sum_{i=1}^{N} \frac{p_i^t}{p_i^0}. \tag{2}$$

Formuła Jevonsa, najbardziej rekomendowana w literaturze do obliczania wskaźnika cen grup elementarnych (IMF i in., 2020), stanowi średnią geometryczną z cząstkowych indeksów cen (sprawia to m.in., że ich wartości zawsze są mniejsze od lub równe wartości wskaźników obliczonych według formuły Carliego):

$$P_J = \frac{\prod_{i=1}^{N}(p_i^t)^{\frac{1}{N}}}{\prod_{i=1}^{N}(p_i^0)^{\frac{1}{N}}} = \prod_{i=1}^{N}\left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{N}}. \tag{3}$$

[15] Eurostat nie rekomenduje formuły Carliego przy pomiarze HICP, gdyż nie spełnia ona testu odwracalności w czasie i wykazuje tendencję do przeszacowywania wyników badania (tj. poziomu inflacji). Z kolei stosowanie formuły Dutota dopuszcza się tylko w zakresie produktów homogenicznych i wyłącznie dla grup elementarnych, w których wariancja cen jest niska. Porównanie formuł indeksów elementarnych i omówienie ich wpływu na uzyskane wyniki można znaleźć m.in. w pracy Białka (2020b).

5. Wyniki

Zestawienie średnich miesięcznych cen reprezentantów wybranych 10 elementarnych grup produktów spożywczych przedstawiono w tabl. 1. Ceny z lutego i marca 2021 r. podano z uwzględnieniem wszystkich omawianych w pracy źródeł danych, czyli metody tradycyjnej (z udziałem ankieterów), danych skanowanych i danych skrapowanych.

Tabl. 1. Średnie miesięczne ceny produktów spożywczych w wybranych grupach elementarnych według metody zbierania danych

Reprezentanty grup elementarnych	Metoda tradycyjna		Dane skanowane		Dane skrapowane	
	luty 2021	marzec 2021	luty 2021	marzec 2021	luty 2021	marzec 2021
	w zł					
Ryż						
Ryż długoziarnisty	8,29	8,24	9,12	9,04	8,69	8,66
Ryż biały	4,64	4,62	5,55	5,81	9,98	9,98
Mąka pszenna						
Mąka pszenna	2,91	2,94	2,70	2,70	3,61	3,63
Pozostałe mąki						
Mąka żytnia	3,88	3,83	3,05	3,08	4,54	3,73
Mleko pełne świeże						
Mleko pełne UHT	3,23	3,23	2,69	2,67	3,01	3,03
Mleko pełne pasteryzowane	2,98	2,96	2,62	2,69	3,25	3,33
Mleko świeże niskotłuszczowe						
Mleko niskotłuszczowe UHT	3,02	3,04	2,63	2,64	2,86	2,86
Mleko kozie	11,22	11,33	7,95	7,95	9,37	9,37
Mleko niskotłuszczowe pasteryzowane	2,82	2,85	2,55	2,58	4,08	4,08
Mleko zagęszczone i w proszku						
Mleko zagęszczone i w proszku	10,03	10,04	12,10	11,92	18,84	18,75
Jogurt						
Actimel[a]	17,83	18,41	13,80	14,65	14,36	14,39
Jogurt owocowy	10,79	10,74	11,22	11,27	12,09	12,16
Jogurt czekoladowy i orzechowy	.	.	24,17	22,31	14,68	14,50
Jogurt pitny	8,94	8,86	7,97	7,94	8,94	8,97
Jogurt naturalny	7,92	7,99	7,82	7,81	13,53	13,39
Napoje i inne produkty mleczne						
Kefir	5,81	5,92	4,80	4,87	4,97	4,97
Maślanka	3,42	3,49	3,33	3,40	3,85	3,90
Monte[b]	19,78	19,99	15,84	17,57	12,40	12,40
Serek homogenizowany	13,28	13,26	12,47	12,54	12,94	12,83

a Napój mleczny. b Marka deserów mlecznych.

Tabl. 1. Średnie miesięczne ceny produktów spożywczych w wybranych grupach elementarnych według metody zbierania danych (dok.)

Reprezentanty grup elementarnych	Metoda tradycyjna		Dane skanowane		Dane skrapowane	
	luty 2021	marzec 2021	luty 2021	marzec 2021	luty 2021	marzec 2021
	w zł					
Cukier						
Cukier trzcinowy	10,39	10,34	9,32	9,40	10,66	10,78
Cukier biały	2,87	2,83	2,73	2,67	2,99	2,78
Cukier puder	.	.	4,76	4,76	9,11	8,92
Kawa						
Kawa rozpuszczalna	137,81	137,49	90,62	91,68	119,56	115,45
Kawa ziarnista	56,05	55,80	50,55	49,67	72,86	71,49
Kawa mielona	.	.	40,01	39,51	62,52	61,84

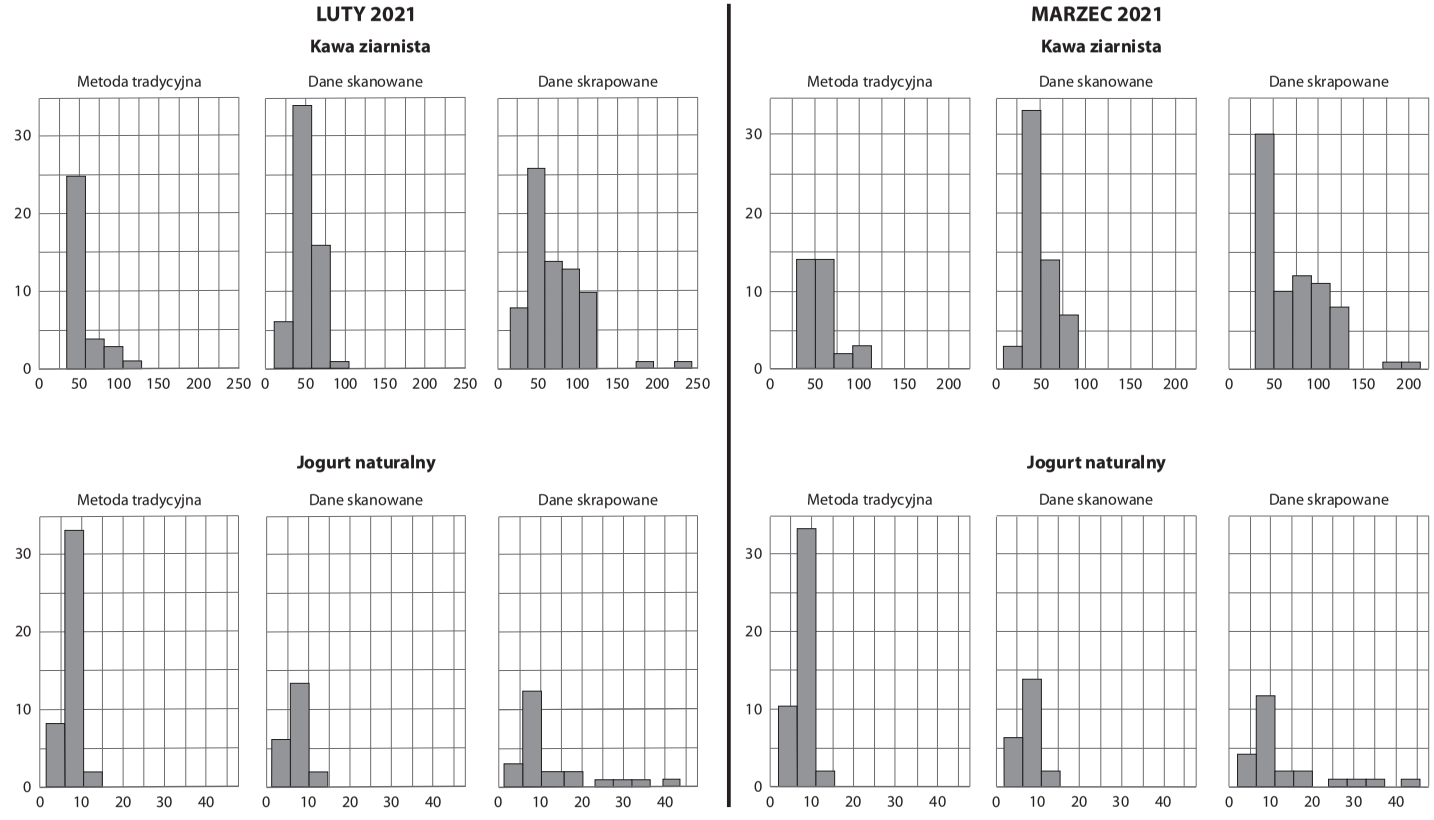
a Napój mleczny. b Marka deserów mlecznych.

Źródło: obliczenia własne w środowisku R na podstawie danych GUS.

Z porównania wynika, że w przypadku takich produktów, jak: ryż biały, mąka pszenna i żytnia, mleko pełne pasteryzowane, mleko zagęszczone i w proszku, jogurt naturalny czy kawa ziarnista uśrednione ceny skrapowane były zdecydowanie wyższe niż w przypadku pozostałych źródeł. Średnie ceny takich reprezentantów, jak mleko kozie, Actimel, kefir, Monte czy kawa rozpuszczalna uzyskane metodą tradycyjną były wyższe od przeciętnych cen tych produktów uzyskanych ze źródeł alternatywnych. Zasadniczo różnice pomiędzy średnimi cenami uzyskanymi z różnych źródeł danych były duże, szczególnie widoczne w przypadku takich produktów, jak: mleko zagęszczone i w proszku, mleko kozie, Actimel, Monte czy kawa ziarnista (chociaż w przypadku części reprezentantów, np. mleka niskotłuszczowego UHT, jogurtu pitnego czy cukru białego – niewielkie). Na podstawie badanego zestawu danych trudno ustalić uniwersalną relację pomiędzy średnimi cenami wskazującą to źródło danych, które determinuje ceny jednoznacznie wyższe lub niższe. Konieczne są dalsze prace nad tym zagadnieniem, w tym uwzględnienie szerszego okna czasowego analizy.

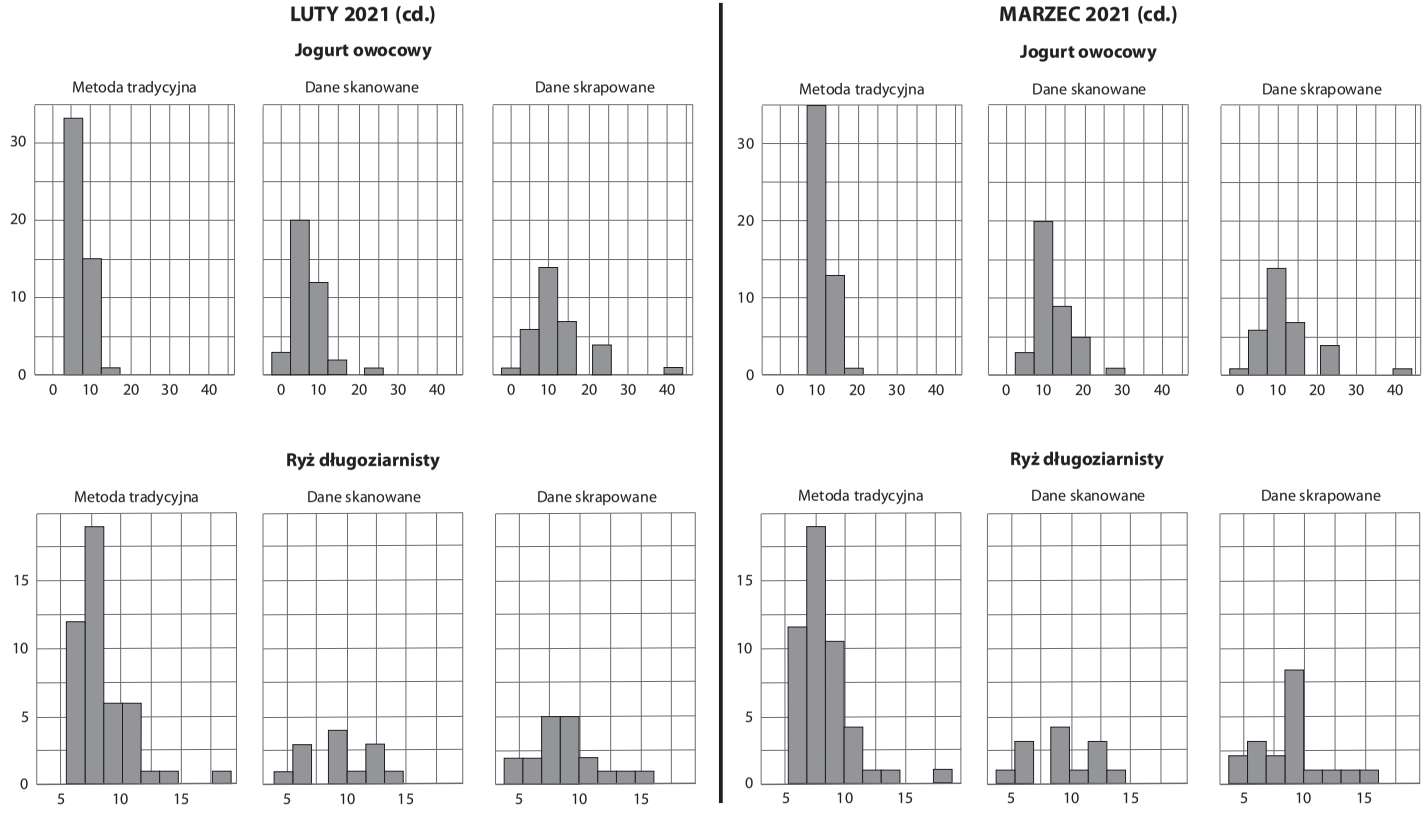
Dla porównania rozkładów cen reprezentantów występujących najliczniej w alternatywnych źródłach danych (kawa ziarnista, jogurt naturalny, jogurt owocowy, ryż długoziarnisty i mąka pszenna) opracowano histogramy (wykr. 1) i wykresy pudełkowe (ang. *box plots*) (wykr. 2). Wykresy pudełkowe sporządzono dla miar klasycznych, czyli prezentujących typowy obszar zmienności badanej cechy (średni poziom cechy +/– odchylenie standardowe z obserwacji).

Wykr. 1. Histogramy rozkładu cen wybranych reprezentantów według metody zbierania danych



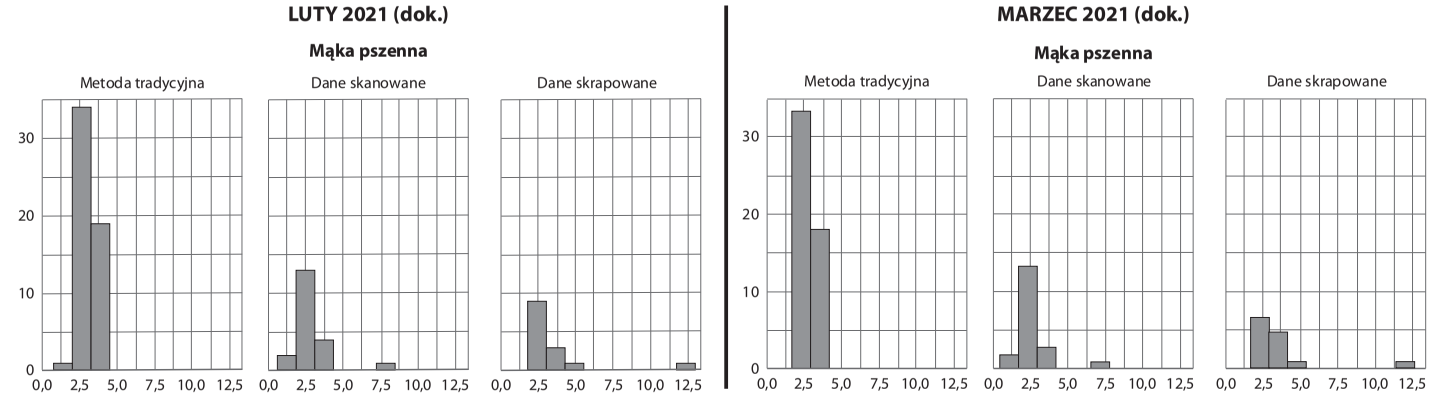
oś *y* – liczba obserwacji, oś *x* – cena w zł

Wykr. 1. Histogramy rozkładu cen wybranych reprezentantów według metody zbierania danych (cd.)



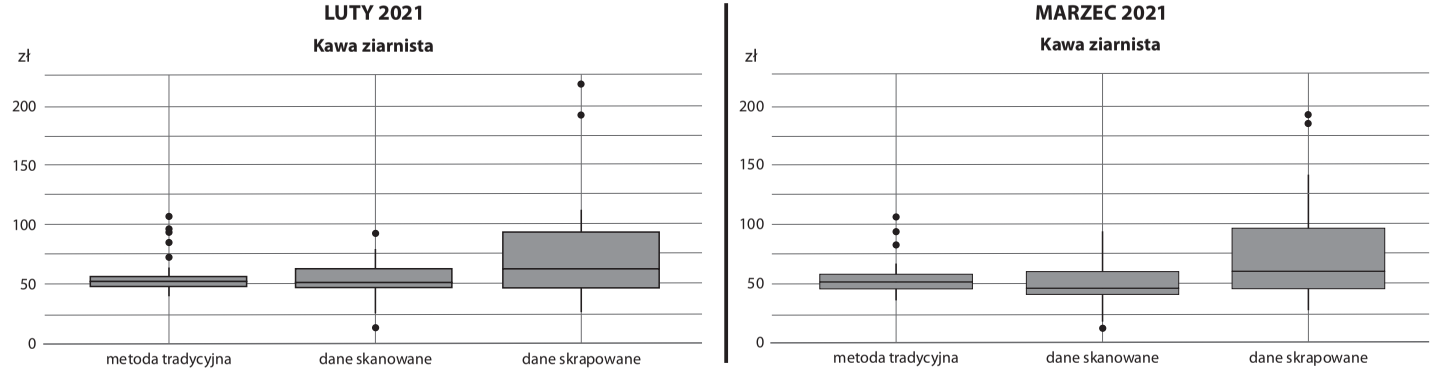
oś *y* – liczba obserwacji, oś *x* – cena w zł

Wykr. 1. Histogramy rozkładu cen wybranych reprezentantów według metody zbierania danych (dok.)

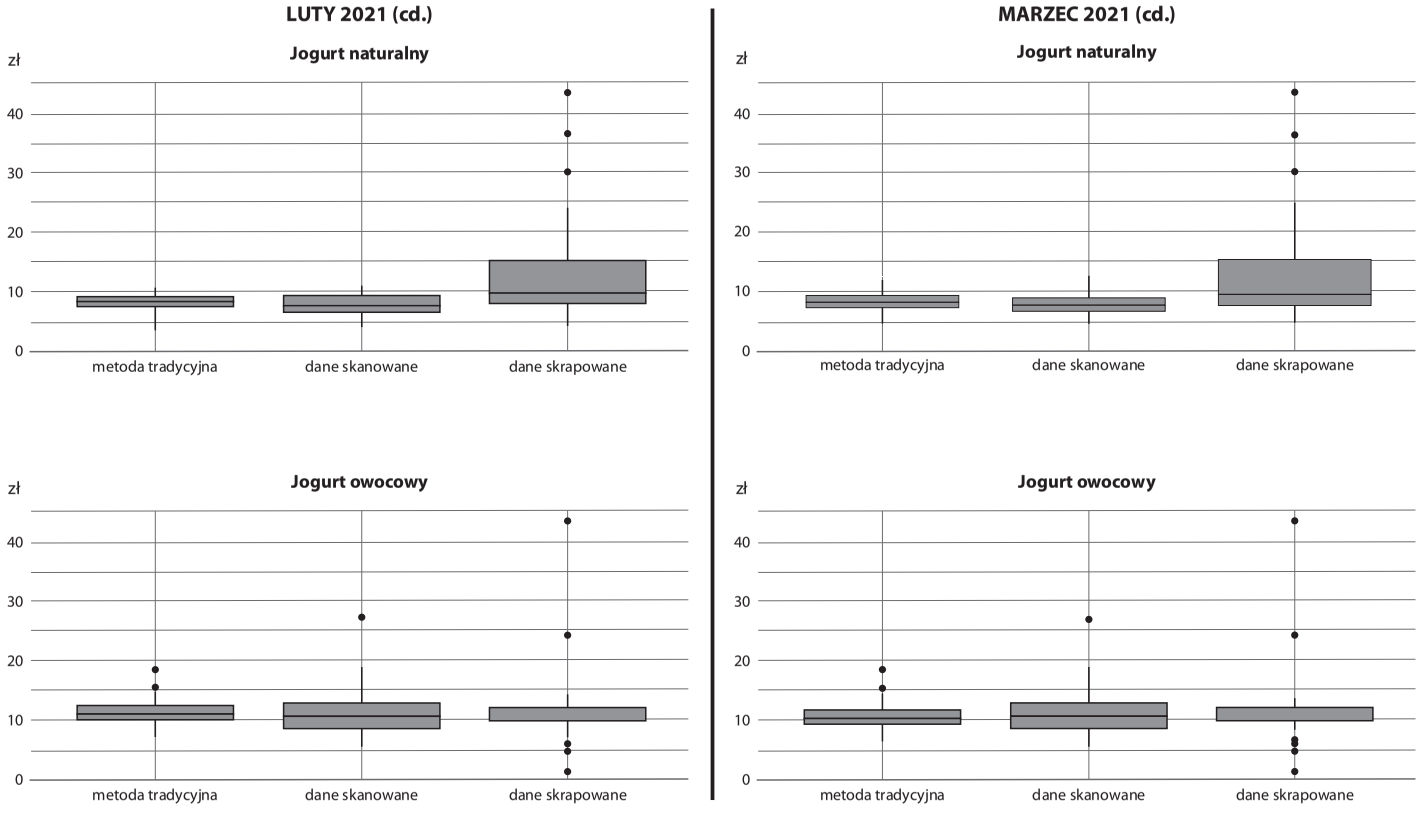


oś *y* – liczba obserwacji, oś *x* – cena w zł
Źródło: opracowanie własne w środowisku R na podstawie danych GUS.

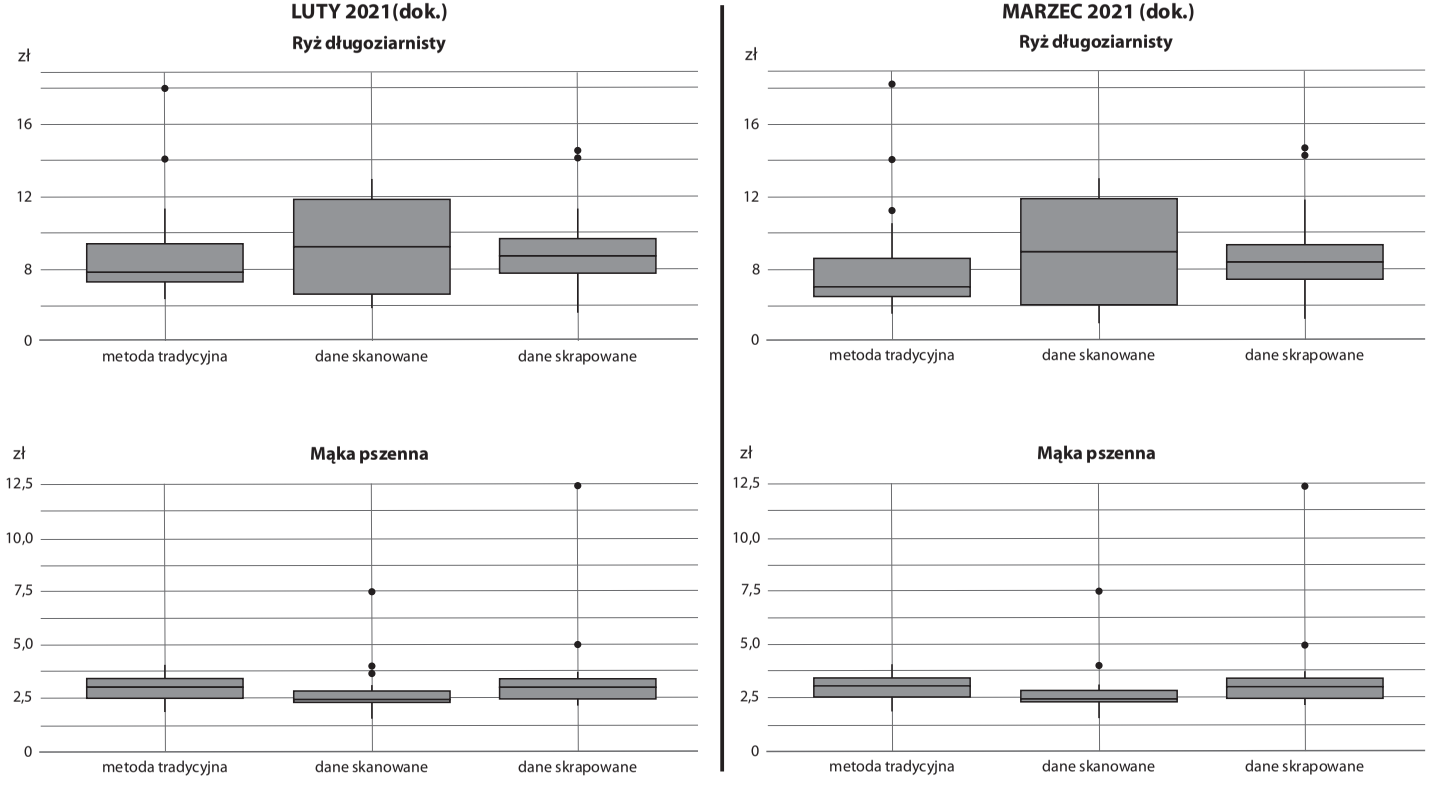
Wykr. 2. Wykresy pudełkowe rozkładu cen wybranych reprezentantów według metody zbierania danych



Wykr. 2. Wykresy pudełkowe rozkładu cen wybranych reprezentantów według metody zbierania danych (cd.)



Wykr. 2. Wykresy pudełkowe rozkładu cen wybranych reprezentantów według metody zbierania danych (dok.)



Źródło: opracowanie własne w środowisku R na podstawie danych GUS.

Analizując wykr. 1 i 2, można dostrzec kilka prawidłowości. Po pierwsze rozkłady cen z lutego i marca 2021 r. są do siebie zbliżone w ramach poszczególnych źródeł danych, ale pomiędzy źródłami – różne. Po drugie, ugólniając, zmienność cen jest najmniejsza w przypadku danych uzyskiwanych metodą tradycyjną, a największa – danych skrapowanych. Jednak np. ceny ryżu długoziarnistego wykazują największe fluktuacje w przypadku danych skanowanych. Po trzecie relatywnie najmniej nietypowych wartości cen (ang. *outliers*) zaobserwowano w odniesieniu do danych skanowanych, ponieważ te dane przechodziły opisane wcześniej potrójne filtrowanie. Największe zaszumienie danych, mimo zastosowania filtru ekstremalnych cen na poziomie kodu GTIN (najbardziej zdezagregowanego), dotyczyło cen skrapowanych. Wyjątek stanowił ryż długoziarnisty – wśród danych zebranych metodą tradycyjną zarejestrowano dwie wyraźne wartości odstające w lutym i trzy w marcu.

Różnice między średnimi cenami, nawet stosunkowo duże, nie muszą się przenosić na różnice w pomiarze dynamiki cen w ujęciu miesiąc do miesiąca, dlatego tę część analizy uzupełniono porównaniem wskaźników cen grup elementarnych. W tabl. 2 zestawiono wartości wskaźnika cen reprezentantów 10 wybranych grup elementarnych obejmujących produkty spożywcze oraz wartości wskaźnika cen obliczone dla całych grup elementarnych. Ponadto opracowano histogramy (wykr. 3) i wykresy pudełkowe (wykr. 4) przedstawiające relacje cen (ang. *price relatives*) tych reprezentantów[16], dla których uzyskano największą liczbę danych ze źródeł alternatywnych (kawa ziarnista, jogurt naturalny, jogurt owocowy, ryż długoziarnisty i mąka pszenna). Wykresy pudełkowe, tak jak w przypadku poziomu cen, sporządzono dla miar klasycznych.

Tabl. 2. Wskaźnik cen wybranych grup elementarnych według metody zbierania danych w marcu 2021 r. (luty 2021 = 100)

Grupy elementarne[a] i podgrupy	Metoda tradycyjna	Dane skanowane	Dane skrapowane
Ryż	99,54	101,84	99,88
Ryż długoziarnisty	99,48	99,14	99,71
Ryż biały	99,60	104,61	100,06
Mąka pszenna	100,98	101,49	96,83
Mąka pszenna	100,98	99,77	100,7
Pozostałe mąki	98,55	101,03	82,11
Mąka żytnia	98,55	101,03	82,11

a Oznaczone pogrubioną czcionką.

[16] Przez *relację cen* danego produktu rozumie się iloraz jego średniej ceny z marca i średniej ceny z lutego 2021 r.

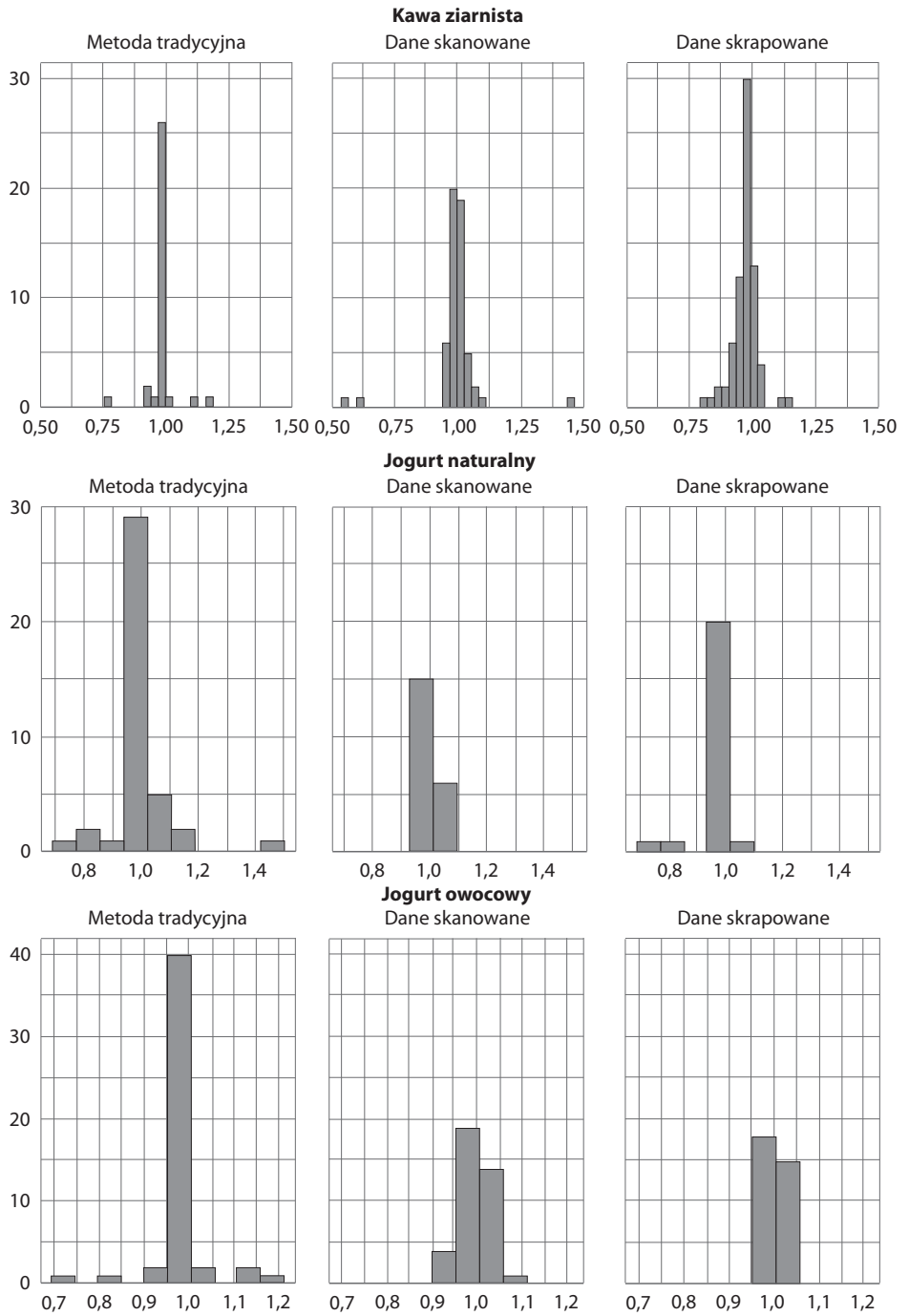
Tabl. 2. Wskaźnik cen wybranych grup elementarnych według źródła danych w marcu 2021 r. (luty 2021 = 100) (dok.)

Grupy elementarne[a] i podgrupy	Metoda tradycyjna	Dane skanowane	Dane skrapowane
Mleko pełne świeże	99,71	100,80	101,52
Mleko pełne UHT	100,07	99,14	100,8
Mleko pełne pasteryzowane	99,36	102,49	102,25
Mleko świeże niskotłuszczowe	100,95	100,56	100,06
Mleko niskotłuszczowe UHT	100,84	100,57	100,00
Mleko kozie	100,95	100,00	100,00
Mleko niskotłuszczowe pasteryzowane	101,05	101,11	100,18
Mleko zagęszczone i w proszku	100,02	98,50	99,57
Mleko zagęszczone i w proszku	100,02	98,50	99,57
Jogurt	100,69	99,58	99,76
Actimel	103,25	106,16	100,15
Jogurt owocowy	99,61	100,49	100,59
Jogurt czekoladowy i orzechowy	.	92,31	98,75
Jogurt pitny	99,04	99,62	100,34
Jogurt naturalny	100,92	99,79	98,97
Napoje i inne produkty mleczne	101,20	103,69	100,10
Kefir	101,90	101,45	100,01
Maślanka	102,01	102,14	101,31
Monte	101,08	110,91	100,00
Serek homogenizowany	99,84	100,59	99,11
Cukier	99,08	99,50	97,32
Cukier trzcinowy	99,54	100,83	101,17
Cukier biały	98,63	97,81	93,02
Cukier puder	.	99,88	97,94
Kawa	99,66	99,39	97,87
Kawa rozpuszczalna	99,77	101,17	96,56
Kawa ziarnista	99,55	98,27	98,13
Kawa mielona	.	98,74	98,92

a Oznaczone pogrubioną czcionką.

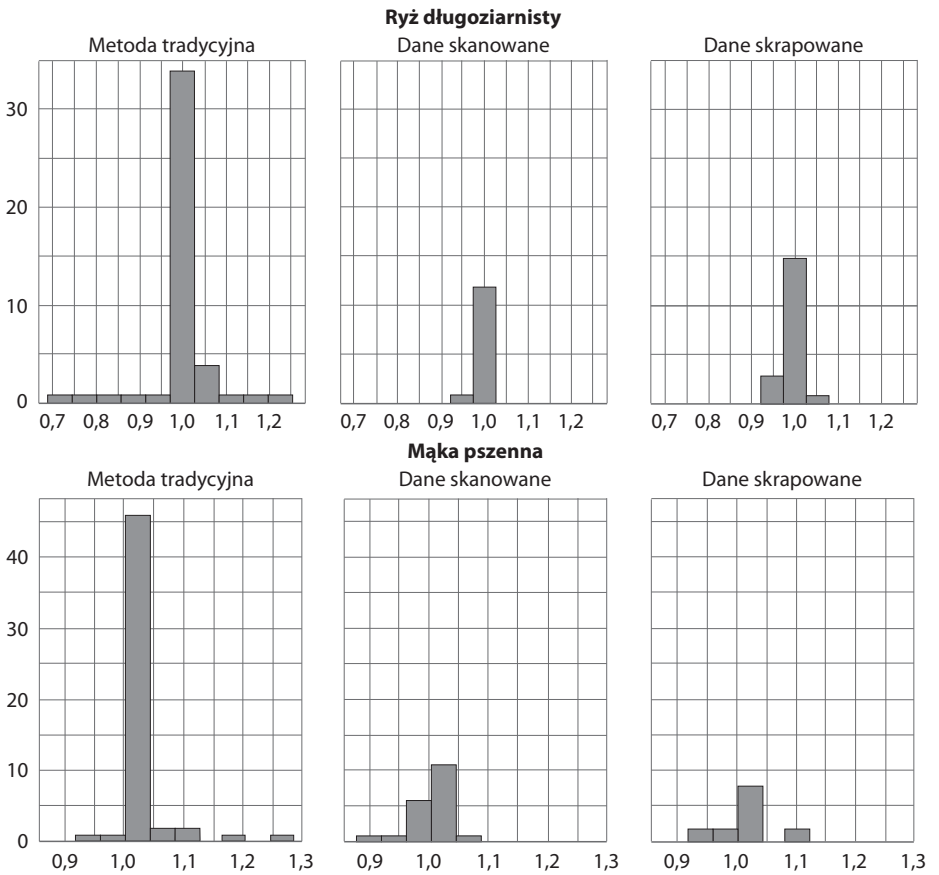
Źródło: obliczenia własne w środowisku R na podstawie danych GUS.

Wykr. 3. Histogramy rozkładu relacji cen wybranych reprezentantów według metody zbierania danych



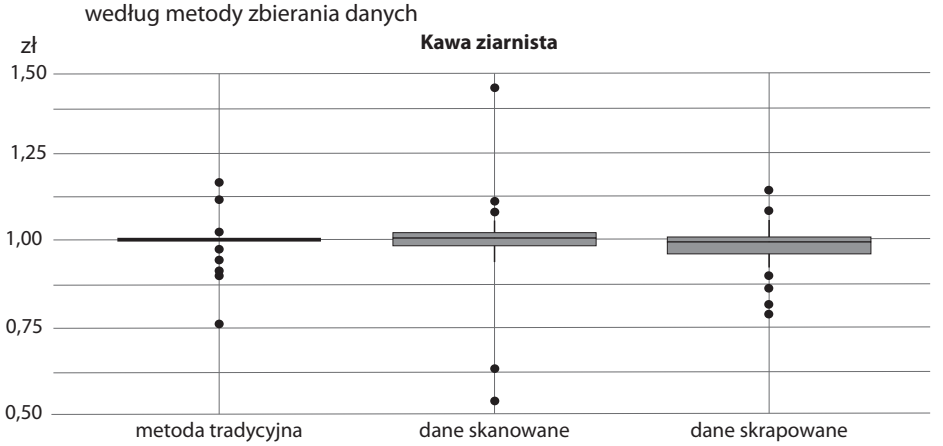
oś *y* – liczba osberwacji, oś *x* – relacja cen

Wykr. 3. Histogramy rozkładu relacji cen wybranych reprezentantów według metody zbierania danych (dok.)

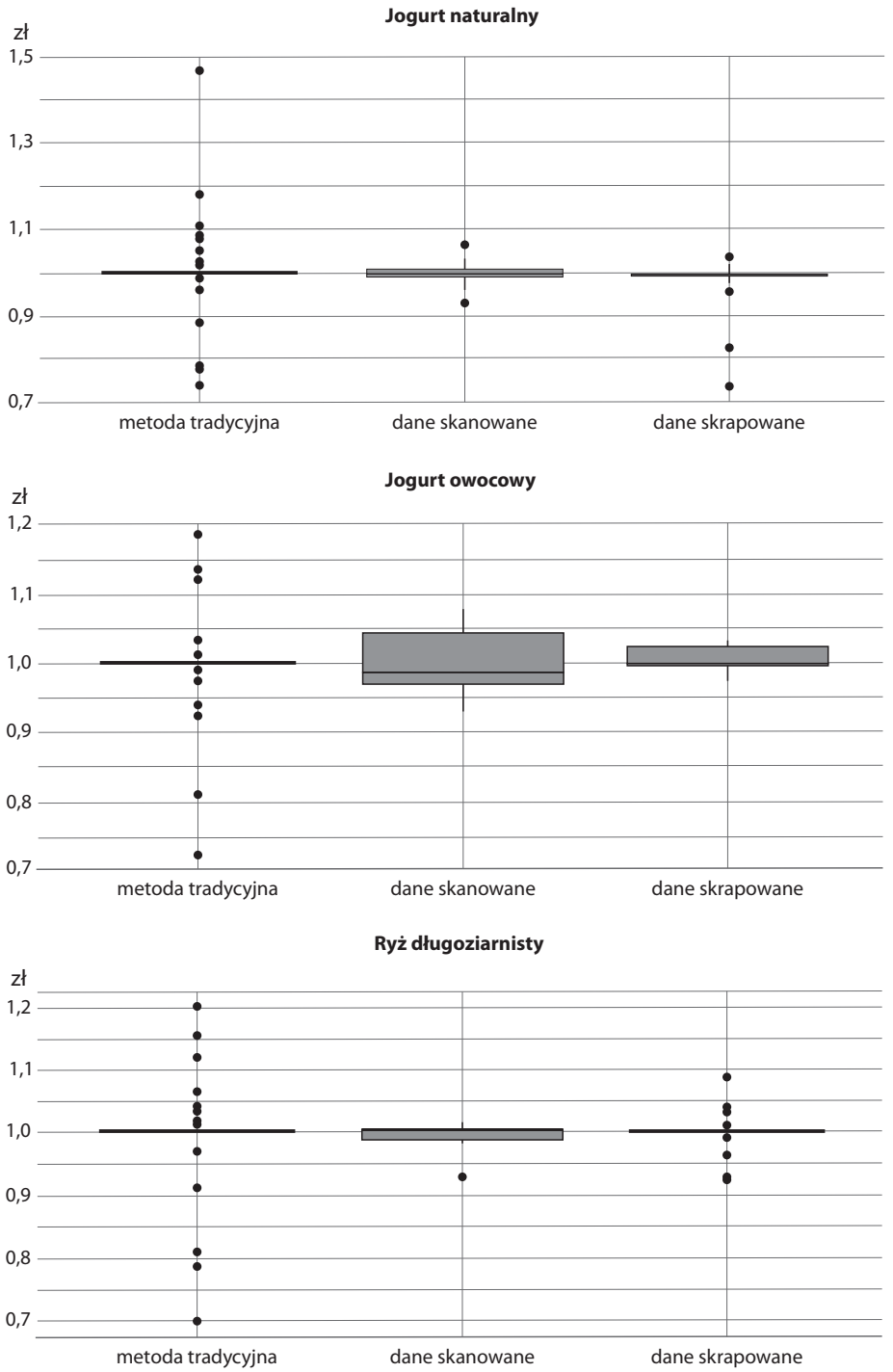


oś *y* – liczba obserwacji, oś *x* – relacja cen
Źródło: opracowanie własne w środowisku R na podstawie danych GUS.

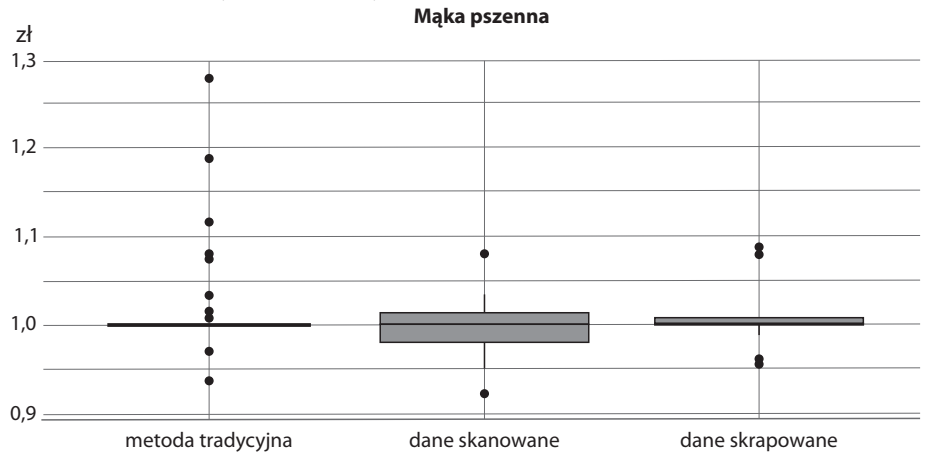
Wykr. 4. Wykresy pudełkowe rozkładu relacji cen wybranych reprezentantów według metody zbierania danych



Wykr. 4. Wykresy pudełkowe rozkładu relacji cen wybranych reprezentantów według metody zbierania danych (cd.)



Wykr. 4. Wykresy pudełkowe rozkładu relacji cen wybranych reprezentantów według metody zbierania danych (dok.)



Źródło: opracowanie własne w środowisku R na podstawie danych GUS.

Na podstawie analizy wartości wskaźnika cen reprezentantów i odpowiadających im grup elementarnych – podobnie jak w przypadku cen produktów – trudno wskazać ogólne prawidłowości i relacje zachodzące pomiędzy nimi. Okres badania wynosił jedynie dwa miesiące, co utrudnia wysnuwanie ogólnych wniosków. Znamienne jest jednak to, że gdy porówna się wszystkie źródła danych, okazuje się, że w przypadku danych skanowanych najwyższy wskaźnik cen dla danej grupy produktów został uzyskany tylko raz (mleko pełne świeże), podczas gdy dla danych zebranych metodą tradycyjną najwyższy wskaźnik cen wystąpił czterokrotnie, a dla danych skrapowanych – pięciokrotnie. Najmniejszą wartość wskaźnika cen otrzymano w przypadku danych skrapowanych aż pięciokrotnie, w przypadku danych zebranych metodą tradycyjną – trzykrotnie, a w przypadku danych skanowanych – tylko dwukrotnie. Pierwsze doświadczenia w zakresie łączenia różnych źródeł danych w celu obliczenia wskaźnika cen skłaniają do wniosku, że wskaźnik cen produktów spożywczych obliczony na podstawie danych skanowanych może zawyżać wynik na poziomie ECOICOP 5 (dla tych danych najczęściej otrzymywano najwyższy, a najrzadziej – najniższy wskaźnik cen), natomiast wskaźnik cen tych produktów obliczony na podstawie cen skrapowanych może go zaniżać (najczęściej otrzymywano najniższy, a najrzadziej – najwyższy wskaźnik cen). Wskaźnik cen obliczony na podstawie danych uzyskanych metodą tradycyjną mieścił się pomiędzy wartościami wskaźnika otrzymanymi ze źródeł alternatywnych.

Mimo że wskaźnik cen obliczono dla krótkiego okresu, przeciętna różnica jego wartości wyliczonych dla porównywanych źródeł danych oscylowała w przedziale 2–4 p.proc. To dość znacząca różnica, tym bardziej że skrapowano tę samą sieć, któ-

ra dostarczała danych skanowanych. I choć zdarzały się podgrupy produktów utworzone z reprezentantów, dla których różnice pomiędzy wartościami wskaźnika cen były niewielkie (np. mleko niskotłuszczowe UHT, mleko kozie czy mleko niskotłuszczowe pasteryzowane), to jednak zarejestrowano również duże rozbieżności. Wskaźnik cen ryżu białego był np. o blisko 5 p.proc. wyższy od analogicznego wskaźnika obliczonego na podstawie danych zebranych metodą tradycyjną, a w przypadku reprezentanta pod nazwą „pozostałe mąki" zanotowano prawie 19 p.proc. różnicy pomiędzy wartością wskaźnika obliczoną na podstawie danych skanowanych i na podstawie danych skrapowanych. Analiza wykresów pudełkowych obrazujących relacje cen (wykr. 4) prowadzi do wniosku, że najmniejsza amplituda zmian występuje w przypadku danych uzyskanych metodą tradycyjną (mimo licznych niekiedy wartości odstających), a największa – w przypadku danych skanowanych.

Ostatnim etapem badania było porównanie elementarnych wskaźników cen obliczonych dla podgrup produktów omówionych na wykr. 1–4 ze względu na źródło danych. W tabl. 3 przedstawiono wartości wskaźnika cen za marzec 2021 r. w stosunku do lutego 2021 r. wyliczone według trzech formuł indeksowych: Dutota, Carliego i Jevonsa[17]. Choć zasadniczo, bez względu na wybór formuły indeksu, wartości odpowiadających sobie indeksów wyznaczonych dla różnych źródeł danych są podobne (różnice zazwyczaj nie przekraczają 0,5 p.proc.), to jednak zdarzają się wyjątki. Na przykład wartości wskaźnika cen dla kawy ziarnistej są niemal o 1,4 p.proc. większe w przypadku metody tradycyjnej niż w przypadku pozostałych źródeł. Względnie duża różnica, przekraczająca 1,5 p.proc., dotyczy też jogurtu naturalnego.

Tabl. 3. Porównanie wartości wskaźnika cen dla wybranych grup elementarnych w marcu 2021 r. (luty 2021 = 100)

Grupy elementarne	Formuła obliczeń		
	Dutota	Carliego	Jevonsa
Metoda tradycyjna			
Ryż długoziarnisty	99,47	99,83	99,50
Mąka pszenna	100,97	101,43	101,32
Jogurt naturalny	100,92	101,51	100,80
Jogurt owocowy	99,61	99,75	99,53
Kawa ziarnista	99,55	99,53	99,36
Dane skanowane			
Ryż długoziarnisty	99,14	99,35	99,32
Mąka pszenna	99,76	99,72	99,67
Jogurt naturalny	99,79	99,87	99,84
Jogurt owocowy	100,48	100,31	100,24
Kawa ziarnista	98,27	98,81	97,81

[17] W omawianym badaniu formuły Dutota i Carliego zastosowano w celach analitycznych i poznawczych. W praktyce statystyki cen konsumpcyjnych GUS, zgodnie z zaleceniem Eurostatu, stosuje tylko formułę Jevonsa.

Tabl. 3. Porównanie wartości wskaźnika cen dla wybranych grup elementarnych w marcu 2021 r. (luty 2021 = 100) (dok.)

Grupy elementarne	Formuła obliczeń		
	Dutota	Carliego	Jevonsa
Dane skrapowane			
Ryż długoziarnisty	99,71	99,75	99,69
Mąka pszenna	100,70	100,73	100,66
Jogurt naturalny	98,96	97,93	97,69
Jogurt owocowy	100,59	100,73	100,72
Kawa ziarnista	98,12	98,22	98,08

Źródło: obliczenia własne w środowisku R (pakiet PriceIndices) na podstawie danych GUS.

W obrębie jednego źródła danych najmniejsze wartości przyjmował najczęściej wskaźnik cen obliczony według formuły Jevonsa, a największe (w większości przypadków) – według formuły Carliego, co wynikało z atrybutów tych formuł. Analizując wartości wskaźnika cen obliczone według formuły Jevonsa dla grup elementarnych objętych badaniem, należy zwrócić uwagę nie tylko na skalę różnic w wynikach pomiędzy poszczególnymi źródłami danych, lecz także na ich kierunek. Dotyczy to trzech grup: mąki pszennej, jogurtu naturalnego i jogurtu owocowego. Ceny mąki pszennej zanotowane metodą tradycyjną wykazywały wzrost o 1,32%, a skanowane – o 0,66%, natomiast ceny skrapowane wskazywały na spadek o 0,33%. Według wskaźnika cen zebranych w sposób tradycyjny jogurt naturalny podrożał w marcu 2021 r. w stosunku do miesiąca poprzedniego o 0,80%, a według danych skanowanych i skrapowanych jego cena spadła odpowiednio o 0,16% i 2,31%. Ceny jogurtu owocowego zanotowane przez ankieterów były w marcu 2021 r. o 0,47% niższe niż w lutym, a dane skanowane i skrapowane wskazywały, że nastąpił ich wzrost, odpowiednio o 0,24% i 0,72%.

6. Podsumowanie

Badanie przeprowadzone na podstawie wybranych elementarnych grup produktów spożywczych prowadzi do wniosku, że zarówno ceny, jak i wskaźniki cen obliczone z wykorzystaniem danych z trzech porównywanych źródeł mogą się znacząco różnić. Chociaż większe różnice w średnim poziomie i skali zmienności są obserwowane w przypadku rozkładów cen, to jednak – jak widać na przykładzie ryżu białego i mąki – rozbieżności pomiędzy miesięcznymi wskaźnikami cen obliczonymi na podstawie danych zebranych metodą tradycyjną i uzyskanych ze źródeł alternatywnych mogą przekraczać kilka punktów procentowych. To duża różnica, biorąc pod uwagę, że badano dynamikę cen w krótkim okresie, zaledwie w stosunku do miesiąca poprzedniego. Mimo że jest to zjawisko naturalne, zastanawiające są różnice obser-

wowane pomiędzy wartościami wskaźnika cen obliczonymi na podstawie danych skrapowanych i danych skanowanych w ramach jednej sieci handlowej. Analiza cen i wskaźnika cen wybranych produktów spożywczych skłania do wstępnego wniosku, że wskaźnik cen obliczony na podstawie danych skanowanych może zawyżać wskaźnik cen na poziomie ECOICOP 5, a wskaźnik cen wyznaczony na podstawie cen skrapowanych może go zaniżać. Ten wątek wymaga dalszych badań, uwzględniających kolejne elementarne grupy produktów i szersze okno czasowe. Wstępne obserwacje wskazują również, że ceny skrapowane zdają się charakteryzować największą zmiennością wśród cen z porównywanych źródeł danych, a najmniejsza zmienność cen dotyczy danych zebranych metodą tradycyjną. Dane skanowane, filtrowane potrójnie, zawierały najmniej nietypowych obserwacji cen, co uwiarygadnia wyniki uzyskane na ich podstawie.

Osobną kwestią pozostaje próba identyfikacji przyczyn rozbieżności w poziomie cen i wartościach wskaźnika cen uzyskanych z różnych źródeł. Lista potencjalnych powodów jest długa – od różnic w lokalizacji rejonów notowań i pomiędzy punktami sieci handlowej do rozbieżności dotyczących zakresu oferowanego i obserwowanego asortymentu. Ten ostatni czynnik wydaje się mieć większe znaczenie niż pozostałe przyczyny, ponieważ w przypadku danych skanowanych asortyment produktów jest z reguły dużo bogatszy niż w przypadku danych zbieranych przez ankieterów w terenie. Natomiast skrapowanie pozwala uzyskać informację jedynie o produktach flagowych (na stronie internetowej wystawia się do sprzedaży dwu-, a nawet trzykrotnie mniej produktów niż jest dostępnych w sklepach stacjonarnych).

Warto dodać, że ceny skanowane – zgodnie z przyjętą metodologią – poddaje się miesięcznej agregacji i oblicza się ich średnią cenę w miesiącu (definiowaną jako iloraz wartości i ilości sprzedaży). Nie bez znaczenia pozostaje także uwzględnianie danych o ilości sprzedaży, ponieważ mogą one mieć wpływ na wahania średniej ceny mimo nieobserwowania zmian poszczególnych cen. Ceny skrapowane również są uśredniane dla miesiąca (z zastosowaniem średniej arytmetycznej nieważonej), z wykorzystaniem informacji zebranych przez programy skrapujące (skrapery) pracujące każdego dnia. Ceny skrapowane – podobnie jak ceny niektórych produktów notowane przez ankieterów – są cenami ofertowymi, natomiast w przypadku danych skanowanych mamy pewność, że są to ceny zapłacone przez nabywców.

Zgodnie z tradycyjną metodą gromadzenia danych ceną miesięczną jest cena ściśle opisanego produktu reprezentanta zanotowana przez ankietera danego dnia w punkcie sprzedaży wybranym do badania. Zebrane w ten sposób dane nie zawierają informacji o ilości sprzedaży, co stanowi jedną z potencjalnych przyczyn różnic pomiędzy średnimi cenami pochodzącymi z różnych źródeł i obliczonymi na ich podstawie wartościami wskaźnika cen. Omówione w artykule różnice pomiędzy danymi z różnych źródeł wynikały głównie z:

- zakresu asortymentu;
- innych formuł obliczania średniej miesięcznej ceny oraz wskaźnika cen;
- wyboru różnych punktów sprzedaży, z których pochodziły dane o cenach (i ilości sprzedaży w przypadku danych skanowanych).

Różnice pomiędzy cenami i wartościami wskaźnika cen są uzasadnione, a pełne włączenie w przyszłości danych skanowanych i skrapowanych do bieżących obliczeń wskaźnika cen konsumpcyjnych wzbogaci bazę danych zarówno pod względem asortymentu, jak i ilości przetwarzanych danych, zwiększając tym samym dokładność danych o inflacji.

Przedstawione w artykule wyniki badania dotyczyły kilku artykułów spożywczych i grup elementarnych, a analiza, ze względu na dostępność danych, obejmowała stosunkowo krótki okres – dwa kolejne miesiące, w związku z czym wyniki te mają charakter wstępny. Planowane są dalsze prace badawcze nad rozważanym zagadnieniem – za wzór posłużą doświadczenia innych krajów i rekomendacje organizacji międzynarodowych – m.in. rozszerzenie analizy na inne grupy produktów oraz stopniowe zwiększanie okna czasowego analizy (ostatnie badania przeprowadzone przez Eurostat, a także np. w Holandii wskazują na 25-miesięczne okno czasowe jako optymalne). Udoskonalenia wymaga także metoda filtrowania i dopasowywania produktów, w tym uwzględnienie większej liczby parametrów identyfikujących produkt, co w dużej mierze zależy od szczegółowości i zakresu danych udostępnianych przez sieci handlowe i uzyskiwanych ze skrapingu. Ponadto na arenie międzynarodowej opracowywana jest metoda łączenia danych z różnych źródeł oraz badane są możliwości zastosowania bardziej zaawansowanych formuł obliczania wskaźników (np. wskaźników multilateralnych). Należy również pamiętać, że obliczanie HICP w krajach UE, w tym w Polsce, musi być oparte na metodyce umocowanej w prawie unijnym. Eurostat prowadzi prace nad wykorzystaniem nowych źródeł danych w obliczeniach HICP. Powołano specjalny zespół, którego zadaniem jest m.in. przygotowanie zaleceń w zakresie możliwości zastosowania wskaźników multilateralnych, a także aktualizowania wytycznych odnośnie do wykorzystywania danych skanowanych i skrapowanych. Eksperymentalne badanie prowadzone w GUS stanowi polski wkład do tej dyskusji.

Dotychczas wyniki pomiaru dynamiki cen uzyskane na podstawie danych skanowanych oraz danych skrapowanych służyły w GUS przede wszystkim do realizacji prac eksperymentalnych dotyczących nowych źródeł danych i oceny możliwości ich wykorzystania w praktyce statystycznej. Od kilku miesięcy, w sytuacji braku danych spowodowanego przez pandemię COVID-19, GUS używa danych ze źródeł alternatywnych również w celu bardziej efektywnego wdrażania metody imputacji. Regularne obliczanie wskaźnika cen towarów i usług konsumpcyjnych (zarówno CPI, jak

i HICP) na podstawie danych pochodzących z różnych źródeł, w tym alternatywnych, powinno być jednak poprzedzone pracami nad rozstrzygnięciem kwestii metodologicznych (Białek, 2020a) związanych głównie z systemem ważenia oraz wypracowaniem adekwatnej formuły indeksowej uwzględniającej atrybuty danych ze źródeł alternatywnych.

Bibliografia

Bertoloto, M., Cavallo, A., Rigobon, R. (2014). *Using Online Prices to Anticipate Official CPI Inflation* (UTokyo Price Project Working Paper No. 049). https://www.centralbank.e.u-tokyo.ac.jp/wp-content/uploads/2018/08/p_wp049.pdf.

Białek, J. (2020a). Wykorzystanie danych skanowanych do pomiaru inflacji – doświadczenia międzynarodowe i wyzwania metodologiczne. *Wiadomości Statystyczne. The Polish Statistician*, *65*(1), 9–33. https://doi.org/10.5604/01.3001.0013.902.

Białek, J. (2020b). Comparison of elementary price indices. *Communications in Statistics – Theory and Methods*, *49*(19), 4787–4803. https://doi.org/10.1080/03610926.2019.1609035.

Białek, J. (2021). PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations. *Statistika: Statistics and Economy Journal*, *101*(2), 122–141.

Białek, J., Bobel, A. (2019, 8–10 maja). *Comparison of Price Index Methods for CPI Measurement using Scanner Data* [referat]. 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro.

Carli, G. (1804). Del valore e della proporzione de'metalli monetati. W: *Scrittori Classici Italiani di Economia Politica: 13* (s. 297–336). Milano: G. G. Destefanis.

Chessa, A. (2015, 20–22 maja). *Towards a generic price index method for scanner data in the Dutch CPI* [referat]. 14th meeting of the Ottawa Group on Price Indices, Tokyo.

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

Diewert, W. E., Fox, K. J. (2018). Substitution bias in multilateral methods for CPI construction using scanner data (UNSW Economics Working Paper No. 2018-13). http://research.economics.unsw.edu.au/RePEc/papers/2018-13.pdf.

Domingos, P., Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2–3), 103–130. https://doi.org/10.1023/A:1007413511361.

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

Eurostat. (2018). *Harmonised Index of Consumer Prices (HICP): Methodological Manual*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/documents/3859598/9479325/KS-GQ-17-015-EN-N.pdf/d5e63427-c588-479f-9b19-f4b4d698f2a2.

de Haan, J. (2006). The re-design of the Dutch CPI. *Statistical Journal of the United Nations Economic Commission for Europe*, *23*(2–3), 101–118. https://doi.org/10.3233/SJU-2006-232-302.

International Monetary Fund, International Labour Organization, Statistical Office of the European Union (Eurostat), Organisation for Economic Co-operation and Development, The World Bank. (2020). *Consumer Price Index Manual: Concepts and Methods*. Geneva.

https://www.ilo.org/global/statistics-and-databases/publications/WCMS_761444/lang--en/index.htm.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785.

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

Kalisch, D. W. (2016). *Making Greater Use of Transactions Data to Compile the Consumer Price Index, Australia* (ABS Information Paper No. 6401.0.60.003). https://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/FE1AE4B7443728E5CA258079000EAF99/$File/6401060003_2016.pdf.

Laspeyres, E. (1871). IX. Die Berechnung einer mittleren Warenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, *16*(1), 296–318. https://doi.org/10.1515/jbnst-1871-0124.

Loon, K. V., Roels, D. (2018, 7–9 maja). *Integrating big data in the Belgian CPI* [referat]. Meeting of the Group of Experts on Consumer Price Indices, Geneva.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. W: *Proceedings of the Section on Survey Research Methods* (s. 354–359). Alexandria: American Statistical Association. http://www.asasrms.org/Proceedings/y1990f.html.