Wiadomości Statystyczne. The Polish Statistician, 2020, vol. 65, 1, 9–33 DOI: 10.5604/01.3001.0013.9021
Studia metodologiczne / Methodological studies

# Wykorzystanie danych skanowanych do pomiaru inflacji – doświadczenia międzynarodowe i wyzwania metodologiczne[1]

Jacek Białek[a]

**Streszczenie.** Zgodnie z definicją Organizacji Współpracy Gospodarczej i Rozwoju (OECD) dane skanowane to szczegółowe informacje o dobrach konsumpcyjnych uzyskane dzięki skanowaniu ich kodów kreskowych w punktach sprzedaży. Zaletami tego rodzaju danych są: kompletność już na najniższym poziomie agregacji, relatywnie niski koszt ich uzyskania oraz mnogość obserwacji. Niemniej jednak dane skanowane mają też wady i ograniczenia. Celem artykułu jest wskazanie problemów i wyzwań metodologicznych związanych z uzyskiwaniem, przetwarzaniem i agregacją danych skanowanych wykorzystywanych do szacowania wskaźnika towarów i usług konsumpcyjnych (CPI). Jedna z kluczowych decyzji polega na wyborze formuły indeksowej przeznaczonej dla elementarnych, homogenicznych grup produktów. Istotę problemu wraz z rekomendacjami zademonstrowano na przykładzie dwóch zbiorów danych z portalu Allegro za okres 4.12.2015–28.12.2018, uzyskanych za pomocą narzędzia TradeWatch. Badanie wrażliwości wyników pomiaru dynamiki cen ze względu na wybór formuły indeksu objęło dwie grupy produktów: zegarek męski sportowy oraz fotel biurowy.

Najważniejsze spostrzeżenia są następujące: po pierwsze, różnice między indeksami bilateralnymi a ich łańcuchowymi wersjami mogą być znaczne, co wynika zapewne z dynamicznego charakteru danych skanowanych; po drugie, różnice między wskazaniami indeksów multilateralnych mogą wynosić nawet parę punktów procentowych dla rocznego okna obserwacji; po trzecie, różnice pomiędzy wartościami indeksów GEKS i CCDI są nieznaczne, a różnice między indeksem Geary'ego-Khamisa dla pełnego okna czasowego i okna bieżącego (*real time index*) przestają być znaczące już po upływie kilku miesięcy; po czwarte, ceny produktów sprzedawanych na platformie elektronicznej, a także wartość i wielkość ich sprzedaży mogą zależeć od dnia tygodnia, a nawet godziny pomiaru.

**Słowa kluczowe:** wskaźnik cen towarów i usług konsumpcyjnych (CPI), dane skanowane, indeksy cen

**JEL:** C43

# Utilisation of scanner data in the measurement of inflation – international experiences and methodological challenges

**Abstract.** According to the Organisation for Economic Co-operation and Development's definition, scanner data are detailed data on consumer goods obtained by scanning their bar codes at points of sale. The advantages of this kind of data include completeness at the lowest level of aggregation, relatively low cost of acquisition and the fact that they enable a multiplicity of observations. Nevertheless, scanner data also have drawbacks and limitations. The aim of the paper is to identify the problems and methodological challenges related to the acquisition, processing and aggregation of scanner data, which are then used to estimate the Consumer Price Index (CPI). One of the most important decisions in the whole process is the right choice

[1] Artykuł przedstawia wyniki badań sfinansowanych przez Narodowe Centrum Nauki w ramach grantu nr 2017/25/B/HS4/00387.

[a] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny. ORCID: https://orcid.org/0000-0002-0952-5327.

of an index formula dedicated to elementary (homogeneous) product groups. The essence of the problem, along with some recommendations, has been presented on the example of two sets of data from Allegro e-commerce platform for the period of 4 Dec, 2015 to 28 Dec, 2018, obtained through a special tool – TradeWatch. The evaluation of the sensitiveness of the results of the price dynamics measurement according to the chosen index formula has been carried out on the basis of two groups of products: men's sports watches and office chairs.

The most important observations are as follows: firstly, the differences between bilateral price indices and their chained versions are likely to be significant because of the dynamic character of scanner data sets; secondly, the differences between multilateral price indices might amount to several percentage points for a yearly time window; thirdly, the differences between the values of the GEKS and CCDI indices are slight, and the Geary-Khamis index for the full-time window ceases to differ significantly from the real time index just after a few months; and fourthly, the prices of products sold via electronic platforms as well as their quantities and sales volumes might differ depending on which particular day of a week they were sold, and even at which hour.

**Keywords:** Consumer Price Index (CPI), scanner data, price indices

## 1. Wprowadzenie

Zgodnie z definicją Organizacji Współpracy Gospodarczej i Rozwoju (OECD) przez dane skanowane (ang. *scanner data*) rozumie się szczegółowe dane o dobrach konsumpcyjnych uzyskane dzięki skanowaniu ich kodów kreskowych w punktach sprzedaży (ILO, 2004). Technologię użytkowania kodów kreskowych produktów opracowano w latach 70. XX w., a ich wykorzystanie do analizy dynamiki cen i poprawy szacunków wskaźnika cen towarów i usług konsumpcyjnych (Consumer Price Index, CPI) w ciągu ostatnich 20 lat nabierało rozpędu. W tym czasie nie tylko ewoluowały zakres i techniki uzyskiwania danych skanowanych, lecz także poszerzyły się możliwości zdobywania takich informacji i powstała nowa metodologia z zakresu indeksów cen (nowe formuły i nowe sposoby aktualizacji wag).

Pionierem w stosowaniu danych skanowanych są Stany Zjednoczone, zaś niedoścignionym poziom zaawansowania metodologicznego reprezentują Australia i Japonia. Kraje europejskie w niewielkim stopniu korzystają z informacji tego rodzaju, choć ostatnio sytuacja zaczyna się zmieniać. Przykładowo do 2015 r. w Unii Europejskiej (UE) tylko Holandia, Norwegia i Szwecja wykorzystywały dane skanowane do obliczeń indeksów cen detalicznych, a zaledwie rok później dołączyły do nich Belgia, Dania i Islandia. Z literatury przedmiotu opublikowanej do roku 2018 (Guerreiro, Walzer i Lamboray, 2018; Leonard, Sillard, Varlet i Zoyem, 2017; Saraiva dos Santos, Lidonio i Cardoso, 2012) wynika, że obecnie również Luksemburg, Portugalia i Francja eksperymentują z danymi skanowanymi dla wybranych podgrup koszyka CPI. W Polsce konsorcjum Głównego Urzędu Statystycznego (GUS), Instytutu Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN) i Szkoły Głównej Handlowej w Warszawie (SGH) rozpoczyna właśnie projekt *InstatCeny*, ukierunkowany na

wykorzystanie alternatywnych źródeł danych przy kalkulacji CPI, w tym danych skanowanych.

Korzystanie z danych skanowanych jest relatywnie tanie i automatyczne, a do tego operuje się na ich ogromnych wolumenach. Główną zaletą tych danych (czy szerzej: elektronicznych danych transakcyjnych) jest ich kompletność już na najniższym poziomie agregacji. Oznacza to, że dostarczają one informacji zarówno o cenach produktów, jak i o wartości ich sprzedaży na poziomie elementarnym (najbardziej zdezagregowanym). Niemniej jednak dane skanowane mają też wady i ograniczenia, a ich wykorzystanie do pomiaru CPI rodzi wiele trudności metodologicznych. Celem artykułu jest wskazanie problemów i wyzwań związanych z uzyskiwaniem, przetwarzaniem oraz agregacją danych skanowanych. Niezmiernie ważny jest wybór odpowiedniej formuły indeksu cenowego dla elementarnych, homogenicznych grup produktów. Empiryczną ilustrację konsekwencji tego wyboru wraz z rekomendacjami zademonstrowano na przykładzie dwóch dużych zbiorów danych pochodzących z portalu Allegro.

## 2. Zawartość danych skanowanych

Elektroniczne terminale w punktach sprzedaży obsługują najczęściej następujące kody kreskowe: GTIN (Global Trade Item Number) lub jego europejską wersję EAN (European Article Number), PLU (Price Look-Up) i SKU (Stock Keeping Unit). Najbardziej rozpowszechniony jest kod GTIN (EAN), choć funkcjonują też bardzo specyficzne jego wersje, np. UPC (Universal Product Code) czy lokalny APN (Australian Product Number). Przykładowo GTIN zawiera 8, 12, 13 lub 14 cyfr. Najpopularniejsza jest jego pełna wersja – 13- i 14-cyfrowa. Kod GTIN składa się z: 1 cyfry wskazującej poziom pakowania, 3-cyfrowego kodu organizacji krajowej GS1 (potocznie: kodu kraju, np. 590 – Polska), 4–7 cyfr numeru jednostki kodującej GS1, 2–5 cyfr kodu produktu i 1 cyfry kontrolnej. Kod PLU jest krótszy, a SKU – ogólniejszy niż GTIN i dostarcza mniej detalicznych informacji. Paradoksalnie jednak to właśnie zbyt szczegółowy poziom informacji o produkcie GTIN sprawia, że posługując się tym kodem, trudno jest wyłonić homogeniczne grupy produktów (np. ten sam produkt, ale w innym opakowaniu może mieć dwa różne numery GTIN). Niektóre kraje korzystające z danych skanowanych przy szacowaniu CPI używają więc kodu SKU, a nawet podkreśla się, że stosowanie kodów GTIN lub EAN może nie dawać dobrych rezultatów przy pomiarze inflacji (Dalen, 2017).

Poza kodem kreskowym produktu dane skanowane zawierają wiele innych cennych informacji. Ich zasób jest różny w poszczególnych krajach, także w zależności od dostawcy (sieci supermarketów) lub rodzaju produktu. Obejmuje on: kod sprze-

dawcy (określa grupę towarową według indywidualnej klasyfikacji danej sieci), kod identyfikujący punkt sprzedaży w obrębie danej sieci, etykietę produktu (dodatkowy opis produktu i jego charakterystykę), jednostkę sprzedaży (optymalnie według ujednoliconego formatu – np. „szt", „kg", „paczka", „gr" i „litr"), wartość sprzedaży, liczbę sprzedanych jednostek produktu, flagę (np. oznacza się produkty z przecen i promocji) oraz informację o podatku VAT.

## 3. Dostawcy danych skanowanych

Wyróżniamy kilka podstawowych źródeł danych skanowanych. Najcenniejszym wydają się bezpośredni dostawcy, a więc punkty sprzedaży, zwłaszcza sieci supermarketów. Supermarkety to potężni potencjalni dostawcy danych – typowy supermarket posiada bazę od 10 tys. do 25 tys. kodów kreskowych sprzedawanych produktów, z których większość stanowią żywność i napoje. Podobnymi dostawcami mogą być mniejsze markety, drobni sprzedawcy, apteki, biura turystyczne, a nawet sklepy internetowe, jeśli tylko archiwizują dane o sprzedaży z uwzględnieniem kodowania produktów.

Drugim źródłem są firmy wyspecjalizowane w badaniu rynku. Niektóre kraje korzystają z danych skanowanych dostarczanych przez firmę A. C. Nielsen lub GfK i włączają je do szacunków krajowego CPI (Krsinich, 2014). Zaletą tego sposobu uzyskiwania danych skanowanych jest możliwość ich szczegółowego filtrowania ze względu na bezpośrednich dostawców, rejony czy uzgodnione z firmą dostawcą charakterystyki homogenicznych grup produktów. Daje to większą swobodę w wyborze dóbr do koszyka CPI i może zwiększyć reprezentatywność próby produktów. Główną wadą tego rozwiązania jest jednak wysoki koszt uzyskiwania danych.

Trzecim, najmniej rozpoznanym w literaturze źródłem danych skanowanych są potężne elektroniczne platformy handlowe (np. działające na polskim rynku serwisy typu Allegro, OLX czy Shumee). Platformy te zrzeszają ogromną liczbę sprzedawców detalicznych i hurtowników, a niektóre z nich (np. Allegro czy Shumee) archiwizują dane o sprzedaży na poziomie kodów EAN. Największą zaletą tego rodzaju źródła danych skanowanych jest ogromny wolumen obrotów dla bardzo wielu produktów. Nie mniej ważny jest bardzo szeroki asortyment sprzedawanych produktów. Za uwzględnieniem tego źródła danych przemawia również możliwość bardzo szczegółowego filtrowania danych o sprzedaży. Na przykład Allegro prowadzi serwis TradeWatch, dzięki któremu można filtrować dane o sprzedaży nie tylko ze względu na kod EAN, lecz także charakterystyki opisowe, kategorie produktów, sprzedawców oraz moment dokonania zakupu (można analizować rozkłady cen danego produktu nawet w ciągu dnia).

## 4. Techniki uzyskiwania i przetwarzania danych skanowanych

Dane skanowane dostarczane są przez dostawców w formacie csv lub rzadziej, ze względu na rozmiar, w formacie xls. Niektóre kraje dokonują też transferu danych przez serwisy sieciowe (format xml) lub stosują swój wewnętrzny format zapisu. Niekiedy niezbędne jest przeformatowanie otrzymanych danych zgodnie z wymaganiami środowiska IT, w którym dokonywane są dalsze analizy lub realizowany jest konkretny skrypt danego środowiska (może to być środowisko R, Python, SAS czy inne akceptujące wymienione formaty zapisu danych). W zależności od kraju częstotliwość uzyskiwania danych skanowanych może być dzienna, tygodniowa lub miesięczna. Dane najczęściej już na wstępie są filtrowane (usunięcie duplikatów i brakujących lub podejrzanych rekordów) oraz agregowane. Stosownie do potrzeb dane skanowane agreguje się do dnia, tygodnia lub (najczęściej) do miesiąca. W tym ostatnim przypadku z reguły uwzględnia się transakcje z trzech środkowych tygodni każdego miesiąca.

Tak przygotowane dane klasyfikuje się do grup COICOP (Classification of Individual Consumption by Purpose); minimalnym wymogiem jest COICOP 5. W tym celu wykorzystuje się kody produktów (EAN, kod dostawcy), a w przypadku ich braku lub niejednoznaczności korzysta się dodatkowo z etykiet produktów za pomocą metod uczenia maszynowego – machine learningu, w tym metody analizy tekstu – text miningu.

Po etapie klasyfikacji następuje dopasowywanie produktów z porównywanych momentów – matching. Chodzi tutaj o obserwowanie ceny tego samego homogenicznego produktu, nawet jeśli zmieni on kolor opakowania i wagę, a co za tym idzie – zostanie mu nadany inny kod, np. EAN. Dysponując kodami kreskowymi (GTIN, EAN, SKU itd.), kodami wewnętrznymi sprzedawców oraz dostarczanymi przez nich charakterystykami ilościowymi i jakościowymi, ustala się zatem zbiór dopasowanych homogenicznych produktów dla porównywanych okresów jednostkowych, np. kolejnych miesięcy.

Zbiory poprawnie sklasyfikowanych i dopasowanych produktów analizuje się ze względu na zachowania ekstremalne. Ten etap filtrowania (czyszczenia) danych najczęściej uwzględnia udział danego produktu w łącznej sprzedaży w przynależnej mu grupie produktów w porównywanych miesiącach czy też zmianę ceny produktu z miesiąca na miesiąc. Produkty o zbyt małym udziale w rynku (np. w porównywanych miesiącach) najczęściej są usuwane z próby, a w przypadku ekstremalnie małej lub dużej zmiany ceny produktu są flagowane i poddawane dalszej analizie (niektóre kraje usuwają je z próby).

Dalsza analiza ma również wiele wariantów w zależności od kraju. Niekiedy ekstremalną cenę traktuje się jako brakującą informację i dokonuje się jej imputacji. Często

także sprawdza się, czy np. nietypowo niskiej cenie towarzyszy nietypowy spadek wolumenu sprzedaży. Jeśli tak jest, to taką cenę uznaje się za zaniżoną (ang. *dump price*) i usuwa z analizy, przyjmując, że reprezentuje produkt wycofywany z rynku.

Tak przygotowane dane stanowią podstawę do wyznaczenia indeksu cenowego. Poszczególne kraje stosują różne rozwiązania – dynamikę cen na podstawie danych skanowanych wyznacza się za pomocą metod bilateralnych (w tym indeksów nieważonych), jak również multilateralnych (Chessa, 2017).

## 5. Doświadczenia międzynarodowe w zakresie wykorzystania danych skanowanych

Skala i zakres uzyskiwanych danych skanowanych, techniki ich filtrowania i imputowania oraz metodologia pomiaru dynamiki cen są odmienne w poszczególnych krajach europejskich, a także poza Europą. Poniżej przedstawiono krótką charakterystykę wybranych krajów ze względu na doświadczenie w wykorzystywaniu danych skanowanych do szacowania inflacji (CPI). Została ona opracowana na podstawie literatury przedmiotu podanej w bibliografii załącznikowej oraz referatów wygłoszonych podczas 16. spotkania grupy ottawskiej (Ottawa Group on Price Indices) w Rio de Janeiro w 2019 r.[2]

### 5.1. Holandia

W Holandii dane skanowane wykorzystano po raz pierwszy w 2001 r. W 2002 r. uzyskiwano dane z dwóch supermarketów, w 2010 r. – z sześciu, a w 2016 r. – z dziesięciu. Od 2017 r. dostawcami danych skanowanych oprócz supermarketów są m.in. sklepy samoobsługowe i biura podróży. Od 2009 r. holenderski urząd statystyczny (Centraal Bureau voor de Statistiek, CBS) otrzymuje dane z częstotliwością tygodniową. Obecnie ponad 20% CPI w Holandii bazuje na danych skanowanych. Od stycznia 2010 r. holenderskie CBS implementuje łańcuchowy indeks Jevonsa (dla miesięcznych podokresów) w modelu dopasowanym (ang. *matched model*).

Holenderscy statystycy uznają kod GTIN za zbyt szczegółowy i preferują SKU. Homogeniczne grupy produktów, poniżej najniższego dostępnego poziomu agregacji ECOICOP, uzyskuje się dzięki holenderskiemu systemowi klasyfikacji produktów według kodów GTIN i ich charakterystykom (Externe Scannerdata Berichtgever Aggregaat, ESBA). W przypadku odzieży uwzględnia się dodatkowe atrybuty produktu: typ ubrania, markę, sezonowość i kolor. Indeksem multilateralnym, który może zastąpić indeks Jevonsa przy kalkulacji holenderskiego CPI, jest indeks Gea-

[2] Elektroniczna wersja wystąpień jest dostępna pod adresem https://eventos.fgv.br/en/ottawa-group-meeting/publications.

ry'ego-Khamisa (Chessa, Verburg i Willenborg, 2017) oraz jego bardzo specyficzna implementacja – indeks czasu rzeczywistego (ang. *real time index*), w której sukcesywnie miesiąc po miesiącu poszerza się okno aktualizacji danych, traktując grudzień poprzedniego roku jako okres bazowy (Chessa, 2016). Holenderscy statystycy eksperymentują ponadto z rolowaną (rolowanie polega na przesuwaniu całego okna czasowego w prawo o miesiąc wraz z każdym nowym miesiącem obserwacji) rocznie wersją indeksu GEKS (Rolling Year GEKS – RYGEKS), który również jest indeksem multilateralnym, ale jego implementacja jest odraczana.

### 5.2. Belgia

W 2015 r. belgijski urząd statystyczny (Direction générale Statistique) zaczął wykorzystywać dane skanowane (otrzymywane z supermarketów). Zaimplementowano wówczas jedną z metod dynamicznych bazującą na kodach SKU i eliminującą produkty powracające na rynek ze zmienionym kodem (ang. *relaunches*). Naczelną formułą przeznaczoną dla danych skanowanych jest łańcuchowy indeks Jevonsa, przy czym od początku testuje się również indeksy multilateralne: GEKS, oparty na indeksie Törnqvista, TPD (Time Product Dummy), Geary'ego-Khamisa oraz quasi-multilateralny indeks Lehra (Loon i Roels, 2018). Zakłada się, że do końca roku 2020 zostanie wybrana docelowa formuła indeksu cen.

Obliczanie formuły indeksu poprzedza się podstawowym filtrowaniem danych. Belgowie zaimplementowali następujące filtry: (a) sprawdzający, czy dostępność produktu w homogenicznej grupie jest na zadowalającym poziomie; (b) sprawdzający, czy wartość sprzedaży danego produktu jest powyżej ustalonego poziomu procentowego; (c) eliminujący z homogenicznej grupy produktów te, które charakteryzują się dużym spadkiem ceny i ilości; (d) eliminujący z grupy produkty o dużym spadku wolumenu sprzedaży przy niezmienionej cenie; (e) eliminujący produkty charakteryzujące się ekstremalnymi zmianami cen z miesiąca na miesiąc. Ceny brakujące lub eliminowane imputuje się na podstawie ewolucji cenowej pozostałych produktów danej grupy homogenicznej.

Obecnie prowadzone eksperymenty bazują na danych skanowanych dotyczących odzieży, obuwia, usług hotelarskich i sprzętu elektronicznego. Na koniec 2018 r. dane otrzymywano z trzech największych supermarketów, których udział w rynku wynosił 75%–80%.

Dynamikę cen obliczoną na podstawie danych z supermarketów wyznacza się na dzień bieżący dla następujących grup produktów: żywność i napoje bezalkoholowe (waga 16,9%), napoje alkoholowe i wyroby tytoniowe (2,3%), drobne narzędzia i akcesoria (0,3%), nietrwałe dobra domowe (0,8%), produkty dla zwierząt (0,7%), produkty papierowe (0,1%), produkty do rysowania (0,2%) oraz produkty higieny osobistej (1,4%).

### 5.3. Szwecja

Początki szwedzkich eksperymentów z danymi skanowanymi sięgają lat 90. XX w. W jednej z pierwszych prac na ten temat (Dalen, 1997) można przeczytać, że najwcześniejszym dostawcą tego rodzaju danych dla Szwecji była firma A. C. Nielsen, która stworzyła bazę danych otrzymywanych ze szwedzkich supermarketów obejmującą jedynie cztery homogeniczne grupy produktów: mrożone ryby, płatki śniadaniowe, tłuszcze i detergenty. Początkowo brano pod uwagę jedynie formuły cenowych indeksów Laspeyresa i Fishera wraz z ich wersjami łańcuchowymi (okresem podstawowym był miesiąc). Obecnie dane skanowane w Szwecji pochodzą z: aptek i sklepów farmaceutycznych (od 2010 r.), sklepów monopolowych z alkoholami (od 2016 r.) oraz małych marketów, supermarketów i hipermarketów (których udziały pokrywają 80% rynku – od 2011 r.). Co ciekawe, w przypadku supermarketów losowana jest próba 60 punktów sprzedaży (według schematu losowania proporcjonalnego do udziałów w rynku), a w przypadku aptek oraz sklepów z alkoholami przeprowadzane jest pełne badanie. Generalnie sześć źródeł danych skanowanych daje podstawę do szacowania 19% szwedzkiego CPI. Statystycy skrupulatnie wyliczają, że w Szwecji automatyzacja pobierania i przetwarzania danych skanowanych bez manualnej pomocy ankietera prowadzi do oszczędności rzędu 80 tys. euro rocznie, mimo przynajmniej dwukrotnie większej próby produktów.

Obecnie danych skanowanych używa się w przypadku następujących grup produktów: owoce, warzywa, mięso, wyroby tytoniowe, napoje alkoholowe, lekarstwa, lampy i baterie, środki czystości, żywność dla zwierząt oraz środki higieny osobistej. Od 2001 r. szwedzki urząd statystyczny (Statistika centralbyrån) posługuje się rynkowym indeksem sprzedaży żywności (Food Sales in the Trade), wykorzystującym kody GTIN oraz automatyczne, wewnętrzne kodowanie produktów do grup homogenicznych. Szwedzi szacują, że każdego miesiąca analizuje się w ten sposób 100 tys. cen produktów, przy czym informacje o cenach i ilościach agreguje się do tygodnia. W przypadku supermarketów dane pochodzą z trzech środkowych tygodni miesiąca, a w przypadku aptek ich zbieranie odbywa się między 25. dniem poprzedniego a 24. dniem bieżącego miesiąca. Na najbardziej elementarnym poziomie agregacji danych skanowanych Szwedzi stosują indeks Jevonsa.

### 5.4. Luksemburg

W Luksemburgu współpraca urzędu statystycznego STATEC[3] z dostawcami danych skanowanych trwa od niedawna. W styczniu 2018 r. po opracowaniu metodologii korzystania z tego rodzaju danych rozpoczęła się ich regularna transmisja do syste-

[3] Narodowy Instytut Statystyki i Studiów Ekonomicznych Wielkiego Księstwa Luksemburga (L'Institut national de la statistique et des études économiques du Grand-Duché de Luxembourg).

mu STATEC. W fazie wdrożeniowej założono analizowanie produktów żywnościowych i napojów bezalkoholowych z wyłączeniem produktów sezonowych (świeże owoce i warzywa) i stopniowe poszerzanie asortymentu. Obecnie dane skanowane dostarczane do luksemburskiego urzędu statystycznego zawierają: kod EAN (13-cyfrowy), kod dostawcy, etykietę produktu, kod produktu pochodzący z jego wewnętrznej kategoryzacji, etykietę produktu nadaną przez dostawcę oraz wartość, wielkość i okres sprzedaży z dokładnością do miesiąca. Dane są dostarczane 18. dnia każdego miesiąca, a zamiast cen stosuje się wartości jednostkowe (jest to wartość sprzedaży z danego miesiąca podzielona przez wielkość sprzedaży).

Pierwszym etapem analizy danych skanowanych jest ich klasyfikacja do podgrup znajdujących się poniżej poziomu COICOP 5 (np. żywność i napoje bezalkoholowe z wyłączeniem owoców i warzyw ma w Luksemburgu aż 72 podgrupy). Klasyfikacja odbywa się automatycznie, ale pierwszy zbiór uczący bazował jeszcze na klasyfikacji manualnej. Obecnie system luksemburski uznaje za niedopasowane te produkty, które różnią się kodem EAN lub etykietą, czyli jest wrażliwy na powtórne wprowadzanie na rynek tego samego produktu, ale np. o zmienionej wadze czy w opakowaniu o innym kolorze. Dla wszystkich dopasowanych produktów wyznacza się miesięczną zmianę ceny (zwykły iloraz ceny z bieżącego miesiąca i ceny z poprzedniego miesiąca).

### 5.5. Australia

Australijskie Biuro Statystyczne (Australian Bureau of Statistics, ABS) zaczęło regularnie uzyskiwać dane skanowane w 2014 r. Dotyczyło to zaledwie kilku produktów (m.in. paliw) i kilku dostawców. W ostatnich latach ABS podjęło współpracę z największymi australijskimi sieciami handlowymi i obecnie produkty, w przypadku których wykorzystuje się dane skanowane, stanowią ok. 25% koszyka CPI; są to m.in.: owoce i warzywa, mięso i owoce morza, chleb i produkty zbożowe oraz wyroby tytoniowe.

W 2015 r. Australia rozpoczęła realizację projektu badawczego *Enhancing the Australian CPI: a roadmap*, którego jednym z priorytetowych celów było rozwinięcie metodologii uwzględnienia danych skanowanych w CPI. W pierwszych próbach wyznaczania dynamiki cen dla grup elementarnych na podstawie danych skanowanych posługiwano się bilateralnymi, superlatywnymi formułami Törnqvista i Fishera. Ostatecznie ABS rozpoczęło na szeroką skalę eksperymenty z metodami multilateralnymi. Pod uwagę wzięto m.in. indeks GEKS oparty na formule Törnqvista oraz indeks TPD (Time Product Dummy).

### 5.6. Stany Zjednoczone

W Stanach Zjednoczonych zaczęto korzystać z danych skanowanych ponad 25 lat temu, jeszcze przed erą Internetu. Pierwsze wzmianki na temat współpracy amery-

kańskich urzędów statystycznych z dostawcami danych skanowanych pochodzą z 1993 r., natomiast informacje o kalkulacji indeksów cen bazujących na danych skanowanych – z 1995 r., gdy firma A. C. Nielsen rozpoczęła dostarczanie danych do amerykańskiego Bureau of Labor Statistics (BLS). Początkowo dane te dotyczyły spożycia kawy w Waszyngtonie i Chicago; w 1996 r. Marshall Reinsdorf skonstruował nawet specjalne indeksy cenowe dla kawy oparte na danych skanowanych. Podobne analizy przeprowadzano później z wykorzystaniem danych dotyczących spożycia płatków śniadaniowych sprzedawanych w Nowym Jorku (1998). Dostawcą danych również w tym przypadku była firma A. C. Nielsen. Dane skanowane z amerykańskiego rynku są dostarczane także przez Information Resources, Inc.

Obecnie w przypadku wyrobów medycznych przy każdorazowej kalkulacji indeksu cenowego (aktualizacja miesięczna) bierze się pod uwagę ok. 600 mln rekordów (2,5 TB danych). To znacznie więcej niż przeciętna w Europie, ale trzeba uwzględnić wielkość i liczbę ludności Stanów Zjednoczonych. Obliczenia są wykonywane na specjalnym, przeznaczonym tylko do tego zadania serwerze (Unix Sun server), a mimo to analizy na potrzeby publikacji miesięcznego CPI trwają parę dni.

Metodologia dotycząca danych skanowanych bazuje na kodach UPC, a głównymi formułami indeksów są cenowe indeksy superlatywne, przede wszystkim łańcuchowy indeks Törnqvista.

### 5.7. Japonia

Japonia jest przykładem kraju, który w pełni wykorzystuje swój potencjał technologiczny i podejmuje współpracę z potężnymi instytucjami, takimi jak np. Bank of Japan czy Nikkei Digital Media, aby jak najdokładniej szacować inflację, opierając się na możliwie najszerszej gamie danych skanowanych. Od 1998 r. gromadzi dane skanowane pochodzące z 300 sieci supermarketów z całego kraju, przy czym jest to celowa próba dostawców (w przeciwieństwie np. do Stanów Zjednoczonych, gdzie pobiera się próbę losową).

Obecnie realizowany jest projekt rządowy *UTokyo Daily Price Index*, którego wymierny efekt, jak nazwa wskazuje, stanowi publikacja rocznej stopy inflacji każdego dnia (w Japonii poziom CPI jest podawany do publicznej wiadomości raz na miesiąc). Skala przedsięwzięcia jest ogromna: zbiór danych skanowanych od 1998 r. obejmuje 6 bln transakcji dotyczących ok. 2 mln produktów sklasyfikowanych w 213 homogenicznych grupach produktów. W 2013 r. dane skanowane zebrano z 16008 punktów sprzedaży na łączną kwotę transakcji 0,42 tln jenów. Kodem kreskowym służącym do identyfikacji produktów jest japoński odpowiednik EAN – 13-cyfrowy JAN (Japanese Article Number). Japońskie dane skanowane dotyczą żywności, napojów i artykułów gospodarstwa domowego, a nie obejmują np. artykułów elektronicznych oraz usług, jak w niektórych krajach europejskich. Udział produktów uwzględnionych w *UTokyo Daily Price Index* wynosi 17% CPI.

Wspomniany indeks jest publikowany codziennie i ma tylko trzydniowe opóźnienie. Wskazuje na roczną inflację, co oznacza, że okresem bazowym jest odpowiadający badanemu dniu dzień z roku poprzedniego. Do jego wyznaczenia stosuje się cenowy indeks Törnqvista.

Japonia publikuje również *UTokyo Monthly Price Index*, który obrazuje roczną stopę inflacji i jest aktualizowany raz na miesiąc według takiej samej metodologii co *UTokyo Daily Price Index*.

### 5.8. Polska

Główny Urząd Statystyczny w Polsce ponad trzy lata temu rozpoczął współpracę z kilkoma sieciami supermarketów. Chociaż dane skanowane nie są jeszcze dostarczane regularnie, to jednak ich fragmentaryczny zbiór pozwolił na budowanie pewnych doświadczeń w zakresie importu i przetwarzania danych. W ośrodkach urzędów statystycznych w Poznaniu i Opolu trwają prace nad doskonaleniem metod poprawnej klasyfikacji produktów do odpowiednich grup COICOP (m.in. na podstawie machine learningu). Z początkiem roku 2019 w Departamencie Handlu i Usług GUS podjęto prace eksperymentalne nad wyborem optymalnej formuły multilateralnego indeksu cenowego odpowiedniego dla danych skanowanych.

W tym samym czasie zainicjowano także projekt *InstatCeny* (GUS, IPI PAN, SGH), ukierunkowany na wykorzystanie nowych źródeł danych w pomiarze CPI.

## 6. Problemy i wyzwania metodologiczne

Wykorzystanie danych skanowanych do pomiaru CPI rodzi pewne problemy i przynosi wyzwania metodologiczne. Głównymi są:

- **Wybór dostawcy danych i współpraca z nim.** Doświadczenia Polski i innych krajów wskazują, że nie jest łatwo pozyskać nowego dostawcę danych skanowanych. Sieci marketów są niekiedy niechętnie nastawione do współpracy z urzędami statystycznymi, ponieważ nie widzą w takim działaniu korzyści finansowych i obawiają się osłabienia swojej konkurencyjności (obawa przed niepełną poufnością danych). Doświadczenia różnych krajów pokazują, że od nakłonienia do współpracy supermarketu do sfinalizowania umowy legalizującej jej realizację upływa najczęściej około pół roku.
- **Dobór próby do analizy.** Polityka krajów wykorzystujących dane skanowane w zakresie poboru próby jest różna. Część krajów stosuje próby losowe, a część – celowe, przy czym dużo zależy od tego, z iloma sieciami handlowymi podpisano umowy o współpracy, oraz od zróżnicowania geograficznego i demograficznego danego kraju. Jeżeli sieć marketów różnicuje swoją ofertę i ceny w zależności od regionu, to przy doborze próby niezbędne jest uwzględnienie rejoni-

zacji. Podobnie jeśli w danym rejonie poszczególne punkty notowań należące do tej samej sieci mają np. różne godziny otwarcia czy inne promocje, to dobór próby również powinien to uwzględniać i traktować takie punkty notowań osobno (agregacja do całej sieci jest tu niewskazana).

Problematyczny może być także wybór momentów czasowych dla poboru próby. Rozkład cen w ciągu tygodnia nie jest równomierny, np. blisko weekendu ceny często są zawyżane. Również agregacja danych do tygodnia może budzić wątpliwości, które tygodnie (i ile) z danego miesiąca są najbardziej reprezentacyjne.

- **Budowa środowiska IT.** Środowisko IT odpowiadające potrzebom analizy danych skanowanych musi korzystać z zaawansowanych metod i technik uczenia maszynowego (klasyfikacja, rozpoznawanie tekstu). Dynamiczny charakter zbiorów danych skanowanych w połączeniu z ich ogromnym wolumenem i różnorodnością sprawia, że niezbędne procesy: filtrowanie danych, dopasowywanie produktów, klasyfikacja do grup homogenicznych i obliczenia indeksów cen są dość czasochłonne i wymagają nie tylko szybkich algorytmów, lecz często także uruchomienia dodatkowych serwerów w celu przeprowadzenia obliczeń.
- **Opracowanie metodologii przygotowania i analizy danych skanowanych.** Mimo bogatych doświadczeń urzędów statystycznych w zakresie stosowania danych skanowanych wiele problemów metodologicznych nadal czeka na rozwiązanie. Wśród krajów korzystających z danych skanowanych są zarówno zwolennicy, jak i przeciwnicy filtrowania danych; ci drudzy uważają, że filtrowanie nadmiernie redukuje próbę, co powoduje utratę dynamicznego charakteru zbioru danych). Kraje, które zdecydowały się na filtrowanie i imputację danych, stoją przed wyborem parametrów progowych filtrów (np. jaką zmianę ceny uznać już za nietypową albo jaki udział w rynku uznać za wystarczająco duży, aby włączyć produkt do próby) lub samej techniki imputacji (w tym zakresie dostępne są rekomendacje Eurostatu). Metodologia powinna np. określać, za pomocą jakich technik uczenia maszynowego dokonuje się dopasowywania produktów w porównywanych miesiącach oraz ich klasyfikacji do grup/podgrup COICOP, ponieważ właściwe mapowanie produktów nadal jest dużym wyzwaniem dla większości krajów.

Kolejne wyzwanie stanowi uwzględnienie dóbr sezonowych. Niektóre kraje wykluczają takie produkty z próby (zwłaszcza w przypadku tzw. mocnej sezonowości, polegającej na tym, że w pewnych okresach roku produkty są całkowicie niedostępne). Choć metodologia konstruowania indeksów cen bazuje na pojęciu homogenicznej grupy produktów, to problem właściwego zdefiniowania tego pojęcia oraz stworzenia technik do ich selekcji (np. metoda MARS – zob. Chessa, 2018) jest nadal szeroko dyskutowany w Eurostacie.

Ogromnym wyzwaniem metodologicznym jest wybór formuły indeksu cenowego, który spełniałby określone wymogi formalne i praktyczne (np. powinien mieć dobre

własności aksjomatyczne i redukować zjawisko tzw. dryfu łańcuchowego (ang. *chain drift* – zob. Chessa, 2015).

Problemem, z jakim musi się zmierzyć każdy urząd statystyczny korzystający z danych skanowanych, jest wreszcie wybór systemu wag, dzięki któremu możliwe będzie przechodzenie do wyższych poziomów agregacji i włączanie wyliczonych indeksów cen do publikowanej informacji o poziomie inflacji.

## 7. Formuły indeksów cen

W przypadku danych skanowanych zastosowanie mają indeksy bilateralne (bezpośrednie) i multilateralne. Pierwsza grupa indeksów uwzględnia jedynie okres badany ($t$) oraz bazowy (0), a druga grupa porównuje te dwa momenty i bierze pod uwagę informacje o cenach i ilościach z ustalonego okna czasowego $[0, T] \supset [0, t]$. Najczęściej przyjmuje się 13-miesięczne okno czasowe (Diewert i Fox, 2017), choć niektóre kraje stosują znacznie dłuższe (np. Australia rekomenduje 27-miesięczne). Ponieważ zbiory danych skanowanych są bardzo dynamiczne, chociażby ze względu na sezonowość produktów oraz dobra nowe i znikające z rynku, oznaczmy przez $G_{0,t}$ te produkty z rozważanej homogenicznej grupy produktów, które są dostępne jednocześnie w okresie 0 i $t$. Niech $N_{0,t} = \text{card}\,(G_{0,t})$; przez $p_i^\tau$, $q_i^\tau$ i $s_i^\tau$ oznaczmy odpowiednio cenę, ilość i relatywny udział w sprzedaży $i$-tego produktu w okresie $\tau$. Zakładając, że w momencie $\tau$ dostępnych jest $N_\tau$ produktów, otrzymujemy:

$$s_i^\tau = \frac{p_i^\tau q_i^\tau}{\sum_{k=1}^{N_\tau} p_k^\tau q_k^\tau} \tag{1}$$

Artykuł prezentuje najpopularniejsze formuły indeksów cen używanych w przypadku danych skanowanych, które zostały wykorzystane w badaniu empirycznym. Ich listę otwiera najczęściej stosowany, bilateralny i przy tym nieważony indeks Jevonsa (1865):

$$P_J^{0,t} = \prod_{i \in G_{0,t}} \left( \frac{p_i^t}{p_i^0} \right)^{\frac{1}{N_{0,t}}} \tag{2}$$

przy czym najczęściej korzysta się z jego łańcuchowej wersji (hained Jevons index) w postaci:

$$P_{CH-J}^{0,t} = \prod_{\tau=0}^{t-1} P_J^{\tau,\tau+1} \tag{3}$$

gdzie $P_J^{\tau,\tau+1} = \prod_{i\in G_{0,t}} \left(\frac{p_i^{\tau}}{p_i^{\tau-1}}\right)^{\frac{1}{N_{\tau,\tau-1}}}$, $N_{\tau,\tau-1} = \text{card}\ (G_{\tau,\tau-1})$, gdzie $G_{\tau,\tau-1}$ – produkty z rozważanej homogenicznej grupy produktów, które są dostępne jednocześnie w okresie $\tau$ i $\tau$–1.

Ważonymi indeksami bilateralnymi są tzw. indeksy superlatywne (Diewert, 1976) – najczęściej indeks Törnqvista (1936):

$$P_T^{0,t} = \prod_{i\in G_{0,t}} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{s_i^0+s_i^t}{2}} \tag{4}$$

lub indeks Fishera (1922), stanowiący średnią geometryczną z indeksów Laspeyresa (1871) i Paaschego (1874), oznaczonych odpowiednio przez $P_{La}^{0,t}$ i $P_{Pa}^{0,t}$:

$$P_F^{0,t} = \sqrt{P_{La}^{0,t} \cdot P_{Pa}^{0,t}} \tag{5}$$

gdzie:

$$P_{La}^{0,t} = \frac{\sum_{i\in G_{0,t}} p_i^t q_i^0}{\sum_{i\in G_{0,t}} p_i^0 q_i^0} \tag{6}$$

$$P_{Pa}^{0,t} = \frac{\sum_{i\in G_{0,t}} p_i^t q_i^t}{\sum_{i\in G_{0,t}} p_i^0 q_i^t} \tag{7}$$

W ostatnich latach szczególną uwagę w literaturze przedmiotu poświęca się indeksom multilateralnym, m.in. dlatego, że są one tranzytywne, czyli niewrażliwe na wybór referencyjnego okresu, oraz niwelują zjawisko dryfu łańcuchowego. Problem z dryfem łańcuchowym polega na tym, że w przypadku gdy ceny i ilości powracają do wartości wyjściowych, wartość indeksu cenowego nie wraca do 1 (chociaż powinna). Jest to charakterystyczne dla dóbr sezonowych i występuje nawet w przypadku łańcuchowych wersji indeksów superlatywnych.

W niniejszej pracy wykorzystano następujące indeksy multilateralne:
a) indeks GEKS (zob. Gini, 1931; Eltetö i Köves, 1964; Szulc, 1964):

$$P_{GEKS}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{P_F^{\tau,t}}{P_F^{\tau,0}}\right)^{\frac{1}{T+1}} \tag{8}$$

gdzie $P_F^{\tau,t} = \sqrt{P_{La}^{\tau,t} \cdot P_{Pa}^{\tau,t}}$, $P_{La}^{\tau,t} = \frac{\sum_{i \in G_{0,t}} p_i^t q_i^\tau}{\sum_{i \in G_{0,t}} p_i^\tau q_i^\tau}$, $P_{Pa}^{\tau,t} = \frac{\sum_{i \in G_{0,t}} p_i^t q_i^t}{\sum_{i \in G_{0,t}} p_i^\tau q_i^t}$

b) indeks CCDI (Caves, Christensen i Diewert, 1982; Inklaar i Diewert, 2016):

$$P_{CCDI}^{0,t} = \prod_{\tau=0}^{T} \left( \frac{P_T^{\tau,t}}{P_T^{\tau,0}} \right)^{\frac{1}{T+1}} \tag{9}$$

gdzie $P_T^{\tau,t} = \prod_{i \in G_{\tau,t}} \left( \frac{p_i^t}{p_i^\tau} \right)^{\frac{s_i^\tau + s_i^t}{2}}$

c) uogólniony indeks Lehra, traktowany niekiedy jako quasi-multilateralny (Loon i Roels, 2018):

$$P_{Lehr}^{0,t} = \frac{\sum_{i \in G_t} p_i^t q_i^t \,/\, \sum_{i \in G_0} p_i^0 q_i^0}{\sum_{i \in G_t} v_i^L q_i^t \,/\, \sum_{i \in G_0} v_i^L q_i^0} \tag{10}$$

gdzie:

$$v_i^L = \frac{\sum_{\tau=0}^{T} p_i^\tau q_i^\tau}{\sum_{\tau=0}^{T} p_i^\tau} \tag{11}$$

d) indeks Geary'ego-Khamisa (Geary, 1958; Khamis, 1972):

$$P_{GK}^{0,t} = \frac{\sum_{i \in G_t} p_i^t q_i^t \,/\, \sum_{i \in G_0} p_i^0 q_i^0}{\sum_{i \in G_t} v_i^{GK} q_i^t \,/\, \sum_{i \in G_0} v_i^{GK} q_i^0} \tag{12}$$

gdzie:

$$v_i^{GK} = \sum_{z=0}^{T} \varphi_{i,GK}^z \frac{p_i^z}{P_{GK}^{0,z}} \tag{13}$$

$$\varphi_{i,GK}^z = \frac{q_i^z}{\sum_{\tau=0}^{T} q_i^\tau} \tag{14}$$

Formuły (12), (13) i (14) wyznaczają układ wzajemnie powiązanych równań nieliniowych, którego rozwiązanie uzyskuje się jedynie symultanicznie (pewne iteracyjne metody można znaleźć w pracach: Chessa, 2016; Diewert i Fox, 2017; Maddison

i Rao, 1996). Z technicznego punktu widzenia jest to indeks najtrudniejszy do wyznaczenia. Cen i ilości ze wszystkich okresów $T$+1 można użyć dopiero w ostatnim okresie (przyjmijmy, że w miesiącu $T$) rozpatrywanego okna czasowego. Ta niedogodność została wyeliminowana w indeksie czasu rzeczywistego[4] (*real time index* $P_{RT}^{0,t}$). Jego konstrukcja bazuje na koncepcji indeksu GEKS, przy czym za pomocą specjalnego algorytmu (omawia go szczegółowo np. praca Chessy, 2016) aktualizuje się system wag występujących w formule (12) dla bieżącego, analizowanego momentu $t$, tzn.:

$$v_i^{RT} = \sum_{z=0}^{t} \varphi_{i,RT}^{z} \frac{p_i^z}{P_{GK}^{0,z}} \tag{15}$$

$$\varphi_{i,RT}^{z} = \frac{q_i^z}{\sum_{\tau=0}^{t} q_i^{\tau}} \tag{16}$$

Dla ostatniego okresu z okna czasowego $[0, T]$ zachodzi $P_{GK}^{0,T} = P_{RT}^{0,T}$. W ilustracji empirycznej wykorzystano także indeks JGEKS, którego formuła odpowiada wzorowi (8), lecz zamiast bazowego indeksu Fishera stosuje się indeks Jevonsa (zob. (2)).

## 8. Przykład empiryczny

Departament Handlu i Usług GUS prowadzi prace eksperymentalne nad formułami indeksów cen i sposobami agregacji danych skanowanych, korzystając m.in. z danych otrzymywanych z sieci supermarketów. Pierwsze wyniki omawiają m.in. Białek i Bobel (2019) oraz Białek i Roszko-Wójtowicz (2019). W niniejszym artykule wykorzystano jednak dane pochodzące z portalu Allegro, który może stanowić alternatywne źródło elektronicznych danych transakcyjnych przy pomiarze inflacji (podobnie jak OLX, Shumee i in.). Dane tego rodzaju mają odmienną naturę niż dane skanowane pochodzące z supermarketów. Po pierwsze, charakteryzuje je bardzo duża liczba podmiotów wystawiających dany produkt, często pojedynczo (wiele osób wystawia bardzo dużo produktów; czasem są to pojedyncze sztuki danego modelu, a czasem jest to sprzedaż hurtowa). Wolumen sprzedaży produktów jest ogromny (tablica).

[4] Autor artykułu nie spotkał się z polskim tłumaczeniem nazwy tego indeksu. Nazwa *real time index*, wprowadzona przez jego pomysłodawcę (Chessa, 2015, 2016), jest powszechnie stosowana przez ekspertów od indeksów multilateralnych (np. należących do tzw. grupy ottawskiej), choć nie została oficjalnie wprowadzona do literatury przedmiotu. Koncepcji, zgodnie z którą szacowany jest ten indeks, nadano nazwę FBEW (Fixed Base Expanding Window).

Po drugie, fluktuacje cenowe na elektronicznych platformach handlowych są znacznie dynamiczniejsze niż w supermarketach. Ceny produktów na Allegro często zmieniają się w ciągu doby, a histogram rozkładu ceny (także wielkości sprzedaży) jest najczęściej powiązany z dniem tygodnia (wykr. 1 i 2). W przypadku tego rodzaju danych ważne jest zatem, aby do kalkulacji indeksów używać nie cen z konkretnego momentu czasowego, lecz wartości jednostkowych (ang. *unit values*), stanowiących uśrednienie ceny zagregowane do ustalonego okresu (np. tygodnia czy miesiąca).

W niniejszej analizie wykorzystano dane dotyczące fotela biurowego oraz męskiego zegarka sportowego (pobrane za pomocą wyspecjalizowanego narzędzia, jakim jest TradeWatch), a więc dane z bardzo niskiego szczebla agregacji, znacznie poniżej COICOP 5. Te dwie homogeniczne grupy produktów mogłyby uzupełniać listę reprezentantów w grupach odpowiednio COICOP 05111 (meble do mieszkania i domu) oraz COICOP 12312 (zegary i zegarki). Tablica przedstawia ogólną charakterystykę sprzedaży tych produktów na Allegro w okresie 4.12.2015–28.12.2018 oraz przykładowe kody EAN w tych grupach. Wykresy 1 i 2 ukazują fluktuacje cen i wielkości sprzedaży tych produktów w różnych ujęciach (m.in. rozkład według godzin i dni tygodnia).

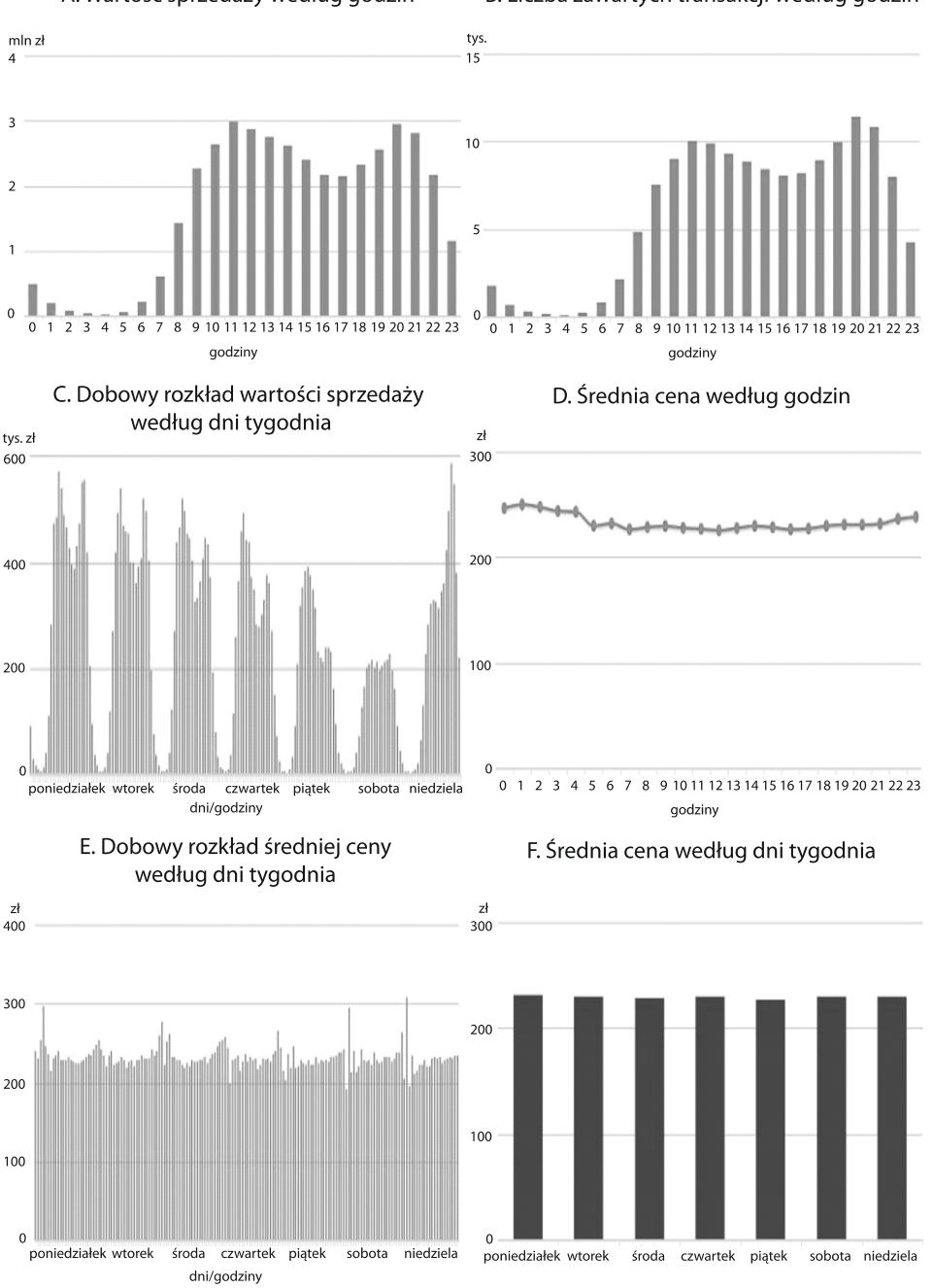
**Tablica.** Charakterystyka sprzedaży fotela biurowego i męskiego zegarka sportowego na Allegro w okresie 4.12.2015–28.12.2018

| Wyszczególnienie | Fotel biurowy | Męski zegarek sportowy |
|---|---|---|
| Łączna sprzedaż w zł | 40588787,48 | 6829768,26 |
| Średnia dzienna sprzedaż w zł | 36207,66 | 6092,57 |
| Średnia cena 1 sztuki w zł | 229,91 | 56,19 |
| Średnia wartość 1 transakcji w zł | 278,47 | 63,43 |
| Liczba sprzedanych sztuk | 176541 | 121551 |
| Liczba zawartych transakcji | 145758 | 107666 |
| Przykładowe kody EAN | 5902759970007 | 4971850907152 |
| | 5902767100175 | 4971850948377 |
| | 5908239692667 | 4971850984191 |

Źródło: TradeWatch.pl.

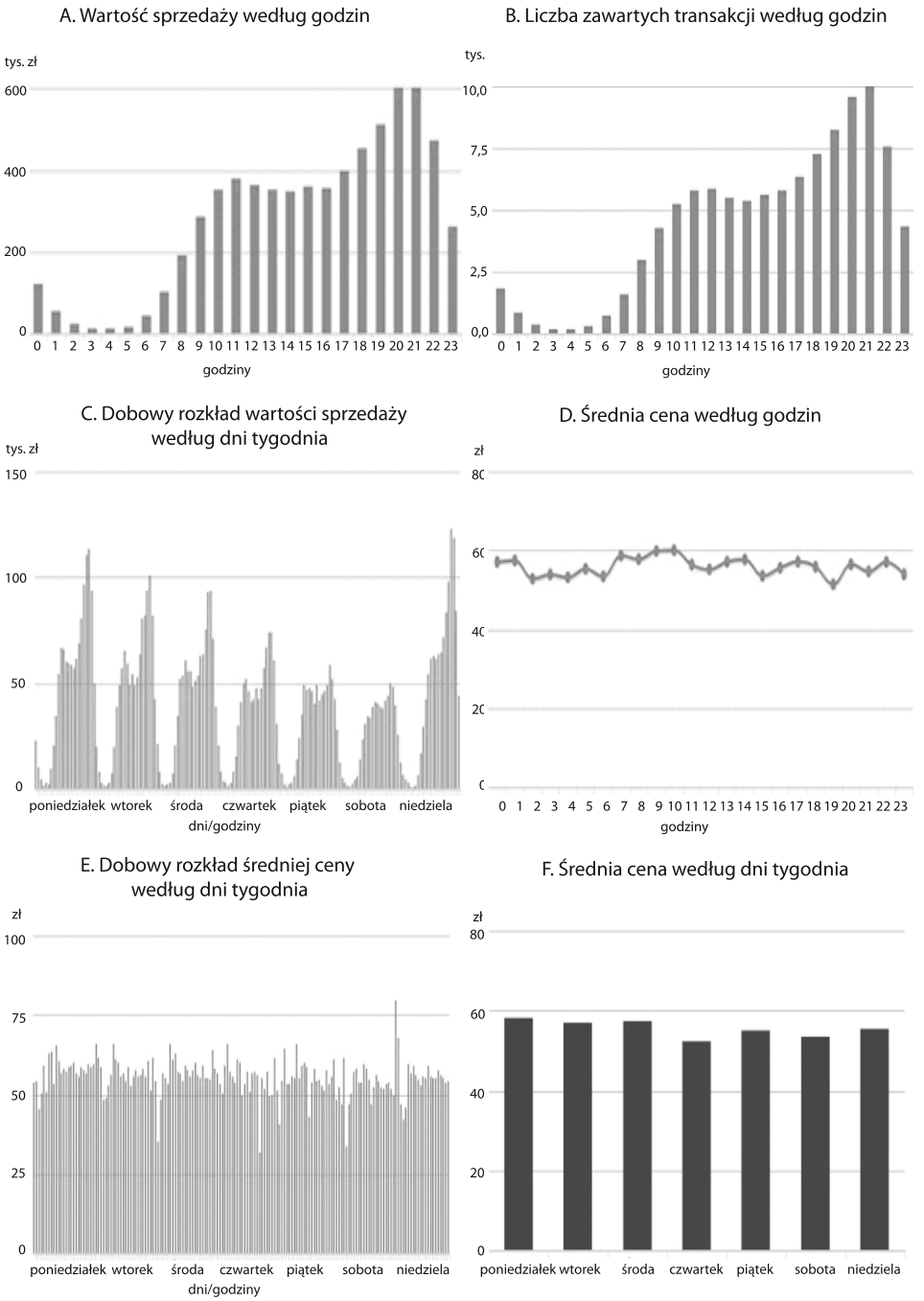
W dalszej analizie ograniczono się do ostatnich 12 miesięcy rozpatrywanego okresu w celu wyznaczenia rocznej dynamiki cen. Dane zagregowano do miesiąca, a następnie przefiltrowano w celu usunięcia ekstremalnych zmian cen (np. ponad 50-procentowy wzrost ceny z miesiąca na miesiąc) oraz produktów o relatywnie niskiej wartości sprzedaży (udział poniżej 1% w rynku). W ten sposób usunięto ok. 30% produktów w obu grupach. Wyniki dla indeksów cen prezentują wykr. 3 i 4.

**Wykr. 1.** Fluktuacje cen i wielkości sprzedaży fotela biurowego w ujęciu godzinowym i dziennym: transakcje na Allegro w okresie 4.12.2015–28.12.2018




Źródło: TradeWatch.pl.

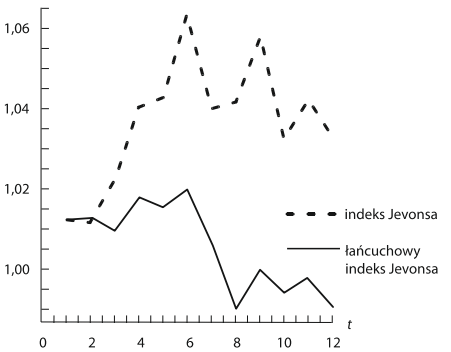
**Wykr. 2.** Fluktuacje cen i wielkości sprzedaży męskiego zegarka sportowego w ujęciu godzinowym i dziennym: transakcje na Allegro w okresie 4.12.2015–28.12.2018



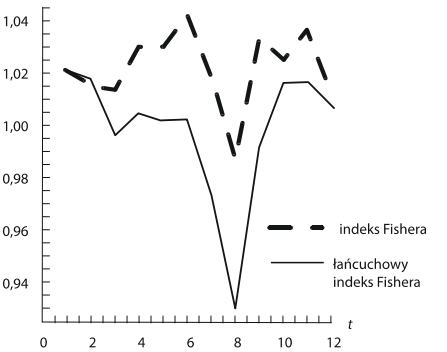

Źródło: jak przy wykr. 1.

**Wykr. 3.** Porównanie indeksów cen dla fotela biurowego: transakcje na Allegro, okno: 13 miesięcy ($t = 0, 1, ...,12$) w okresie grudzień 2017 – grudzień 2018

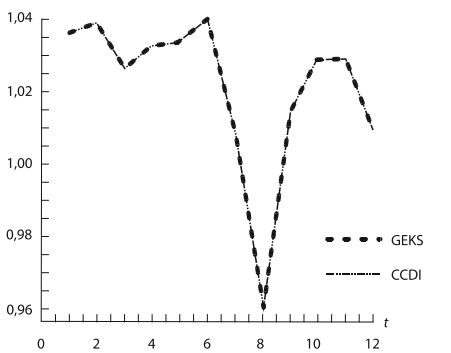
A. Indeks Jevonsa i łańcuchowy indeks Jevonsa



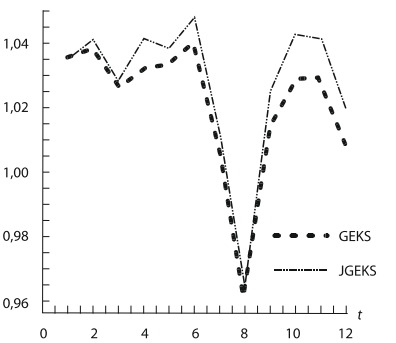

B. Indeks Fishera i łańcuchowy indeks Fishera



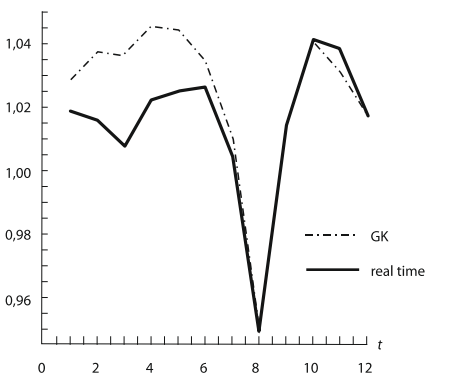

C. Indeksy GEKS i CCDI



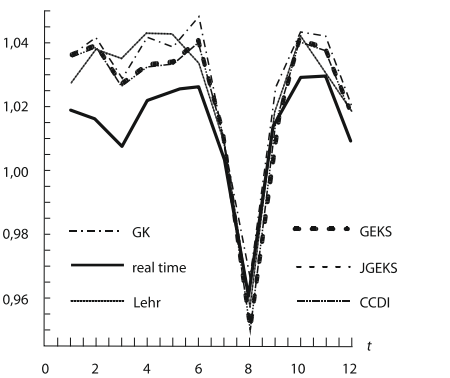

D. Indeks GEKS i jego wersja oparta na indeksie Jevonsa (JGEKS)




E. Indeksy Geary'ego-Khamisa i real time

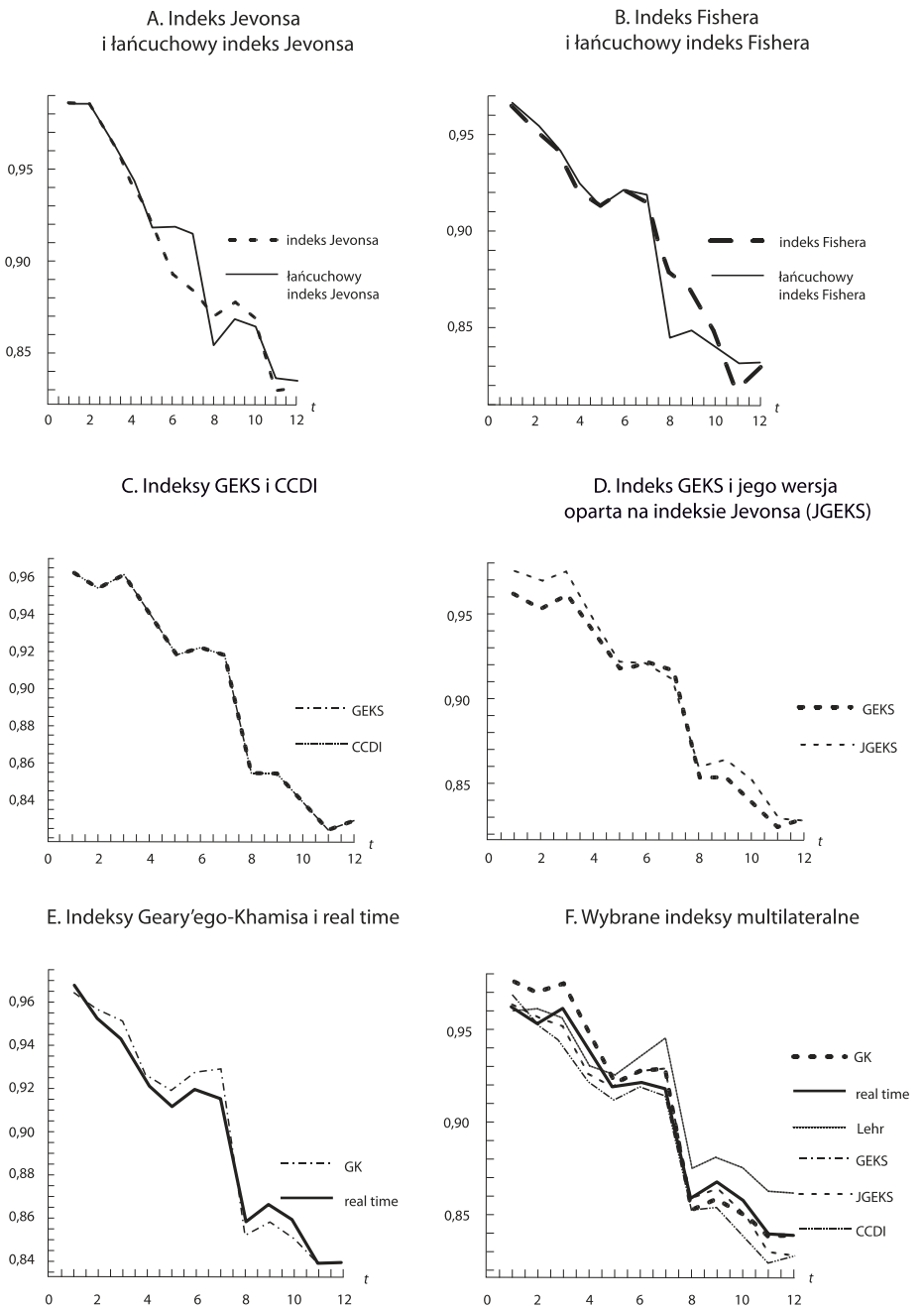



F. Wybrane indeksy multilateralne




Źródło: jak przy wykr. 1.

**Wykr. 4.** Porównanie indeksów cen dla zegarka męskiego sportowego: transakcje na Allegro, okno: 13 miesięcy ($t$ = 0,1,...12) w okresie grudzień 2017 – grudzień 2018




Źródło: jak przy wykr. 1.

Na podstawie analizy produktów sprzedawanych na platformie elektronicznej można zaobserwować, że ich ceny i wielkość sprzedaży mogą się zmieniać nie tylko ze względu na dzień tygodnia (zob. wykr. 1E, 1F, 2E i 2F), lecz także w ciągu doby (wykr. 1A–D, 2A–D). Dlatego – na co zwrócono uwagę w artykule i co jest podkreślane w literaturze przedmiotu – przy kalkulacji indeksów cen nie powinno się używać cen dla konkretnego momentu czasowego, lecz cen uśrednionych (ang. *unit values*) za dany okres, najczęściej miesiąc.

Dane z Allegro charakteryzują się ogromnym wolumenem zawieranych transakcji (tablica), jest to zatem cenne źródło informacji o zachowaniach konsumentów. Co jednak za tym idzie, różnice wskazań nieważonych indeksów cen w stosunku do wersji ważonych mogą być znaczne, gdyż obserwuje się duże różnice w udziałach w sprzedaży nawet wśród najlepiej sprzedawanych produktów w danej grupie (wykr. 3).

Zarówno dane skanowane pochodzące z supermarketów, jak i dane transakcyjne z elektronicznych platform handlowych cechuje bardzo duża dynamika z uwagi na sezonowość produktów, trendy sprzedaży oraz politykę dostawców. Na Allegro występuje znaczna rotacja produktów i sprzedawców, nawet po odfiltrowaniu sprzedaży incydentalnych, z czego wynikają zauważalne różnice wskazań indeksów bilateralnych (ważonych lub nie) oraz ich wersji łańcuchowych (wykr. 3A i 4A oraz wykr. 3B i 4B). W przeprowadzonym badaniu różnice te sięgały nawet kilku punktów procentowych w ciągu roku.

Jako alternatywę dla indeksów bilateralnych w badaniu uwzględniono indeksy multilateralne, szczególnie rekomendowane w literaturze do analizy danych skanowanych. W przypadku indeksu GEKS zastąpienie bazowej formuły, jaką jest indeks Fishera, indeksem Törnqvista (w ten sposób uzyskuje się indeks CCDI) zdaje się nie mieć znaczenia (wykr. 3C i 4C), ale już powrót do bazowego indeksu Jevonsa powoduje zauważalne różnice wartości (wykr. 3D i 4D). Z kolei różnice między indeksem Geary'ego-Khamisa szacowanym dla 13-miesięcznego okna czasowego i indeksem czasu rzeczywistego stają się nieznaczne już po upływie 6–7 miesięcy (wykr. 3E i 4E). Jest to dobra wiadomość dla krajów, które dopiero rozpoczynają uzyskiwanie i gromadzenie danych skanowanych i nie posiadają jeszcze informacji dla pełnego (np. rocznego) okna analizy.

Podsumowując, nawet wybór pomiędzy indeksami multilateralnymi nie jest oczywisty, a różnica ich wskazań dla jednej homogenicznej grupy produktów może wynosić kilka punktów procentowych (wykr. 3F i 4F).

## 9. Podsumowanie

Dane skanowane powinny być traktowane w szczególny sposób. Po pierwsze, już na najniższych poziomach agregacji dostarczają informacji o konsumpcji produktów. Jest to podstawowa różnica w stosunku do tradycyjnego zbioru danych, a także da-

nych scrapowanych (zbieranych ze stron internetowych). Jakość danych skanowanych jest zatem nieporównywalnie lepsza niż danych scrapowanych, tym bardziej że zawierają informacje o cenach transakcyjnych, a nie ofertowych. Elektroniczne platformy handlowe są potencjalnie bardzo dobrym źródłem danych elektronicznych, które noszą wszelkie znamiona danych skanowanych (wolumen jest duży, ceny są cenami transakcyjnymi, wielkość sprzedaży i kod produktu są znane).

Po drugie, gromadzenie, przetwarzanie, filtrowanie, klasyfikowanie i dopasowywanie (matching) danych skanowanych wymaga zaawansowanych technik – machine learningu i text miningu – oraz budowy odpowiedniego środowiska IT w celu automatyzacji tych procesów.

Po trzecie, mimo wieloletniego doświadczenia niektórych krajów w posługiwaniu się danymi skanowanymi, wciąż istnieje wiele wyzwań metodologicznych. W artykule wskazano na potencjalne problemy i utrudnienia w stosowaniu danych skanowanych, związane m.in. z wyborem ich dostawcy, doborem próby produktów czy procedurą postępowania z danymi. Jak nadmieniono, nawet właściwe zdefiniowanie i zlokalizowanie homogenicznych grup produktów nadal rodzi wiele problemów, takich jak np. wybór właściwej formuły indeksu cen. Wiąże się to z ogromną dynamicznością zbiorów danych skanowanych, wynikającą z sezonowości produktów, trendów sprzedaży i polityki dostawców.

W pracy pominięto analizę głównych problemów, takich jak dryf łańcuchowy, silna sezonowość czy efekt substytucji. Skoncentrowano się na określeniu skali różnic, jakie może spowodować zmiana sposobu obliczania dynamiki cen. Wybór nieważonej formuły Jevonsa, jakkolwiek zasadny przy klasycznie uzyskiwanych cenach z najniższych poziomów agregacji, wydaje się tu mniej racjonalny, gdyż nie wykorzystuje się wówczas informacji o wielkości konsumpcji. Nie zmienia to jednak faktu, że większość krajów stosuje łańcuchowy indeks Jevonsa do analizy danych skanowanych, ponieważ indeks ten stwarza znacznie mniej trudności aplikacyjnych niż indeksy multilateralne. Niektóre kraje, np. Stany Zjednoczone czy Japonia, sięgnęły po ważone indeksy superlatywne (Fishera, Törnqvista), ale i w ich wypadku nie ma zgodności co do tego, czy lepiej przyjąć podejście z ustaloną podstawą (ang. *fixed base approach*), czy też stosować łańcuchowe wersje tych indeksów. Wiadomo, że to drugie rozwiązanie może prowadzić do efektu dryfu łańcuchowego, czyli braku powrotu wartości indeksu do 1 w przypadku dóbr sezonowych, których ceny wracają do poziomu wyjściowego.

Rozsądnym wyborem w przypadku danych skanowanych wydają się zatem indeksy multilateralne, które „nadążają" za dynamizmem homogenicznych grup produktów i cechują się tranzytywnością oraz brakiem dryfu łańcuchowego. Do sformułowania ostatecznej rekomendacji indeksu cen odpowiedniego dla danych skanowanych potrzebne są jednak kolejne doświadczenia krajów w tym zakresie oraz pogłębione badania własności indeksów zarówno na gruncie teoretycznym, jak i empirycznym.

## Bibliografia

Białek, J., Bobel, A. (2019). Comparison of Price Index Methods for CPI Measurement using Scanner Data. W: *Paper presented at the 16th Meeting of the Ottawa Group on Price Indices* (s. 1–40). Rio de Janeiro, Brazil, 8–10 May 2019. Rio de Janeiro: FGV.

Białek, J., Roszko-Wójtowicz, E. (2019). The Impact of the Price Index Formula on the Consumer Price Index Measurement. *Statistika – Statistics and Economy Journal*, *99*(3), 246–258.

Caves, D. W., Christensen, L. R., Diewert, W. E. (1982). Multilateral comparisons of output, input, and productivity using superlative index numbers. *Economic Journal*, *92*(365), 73–86. DOI: 10.2307/2232257.

Chessa, A. G. (2015). Towards a generic price index method for scanner data in the Dutch CPI. *Room document for Ottawa Group Meeting*. Urayasu City, Japan.

Chessa, A. G. (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurona, 1*, 49–69.

Chessa, A. G. (2017). Comparisons of QU-GK Indices for Different Lengths of the Time Window and Updating Methods. *Paper prepared for the second meeting on multilateral methods organised by Eurostat*. Luxembourg: Statistics Netherlands.

Chessa, A. G. (2018). Product definition and index calculation with MARS-QU: Applications to consumer electronics. *Report Statistics Netherlands.*

Chessa, A. G., Verburg, J., Willenborg, L. (2017). A comparison of price index methods for scanner data. *Paper presented at the 15th Meeting of the Ottawa Group on Price Indices*. Eltville am Rhein, Germany.

Dalen, J. (1997). Experiments with Swedish Scanner Data. *Proceedings of the Third Meeting of the International Working Group on Price Indexes*. Research Paper no. 9806. Statistics Netherlands, Division Research and Development, Department of Statistical Methods.

Dalen, J. (2017). Unit values in scanner data and some operational issues. *Paper presented at the fifteenth Ottawa Group Meeting*. Eltville am Rhein, Germany.

Diewert, W. E. (1976). Exact and superlative index numbers. *Journal of Econometrics*, *4*(2), 115–145. DOI: 10.1016/0304-4076(76)90009-9.

Diewert, W. E., Fox, K. J. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. *Discussion paper*, *17*(2), 1–62. DOI: 10.2139/ssrn.3276457.

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, 42*, 507–518.

Fisher, I. (1922). *The Making of Index Numbers: A Study of Their Varieties, Tests, and Reliability*. Boston, New York: Houghton Mifflin Company.

Geary, R. C. (1958). A Note on the Comparisons of Exchange Rates and Purchasing Power Between Countries. *Journal of the Royal Statistical Society. Series A (General)*, *121*(1), 97–99. DOI: 10.2307/2342991.

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

Guerreiro, V., Walzer, M., Lamboray, C. (2018). The use of Supermarket Scanner data in the Luxembourg Consumer Price Index. *Working papers du STATEC*, *Economie et Statistiques*, *97*, 1–18.

ILO. (2004). *Consumer Price Index Manual. Theory and practice*. Geneva: International Labour Office.

Inklaar, R., Diewert, W. E. (2016). Measuring industry productivity and cross-country convergence. *Journal of Econometrics*, *191*(2), 426–433. DOI: 10.1016/j.jeconom.2015.12.013.

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

Khamis, S. H. (1972). A New System of Index Numbers for National and International Purposes. *Journal of the Royal Statistical Society. Series A (General), 135*(1), 96–121. DOI: 10.2307 /2345041.

Krsinich, F. (2014). The FEWS Index: Fixed Effects with a Window Splice – Non-Revisable Quality-Adjusted Price Indices with no Characteristic Information. *Paper presented at the meeting of the group of experts on consumer price indices* (s. 26–28). Geneva, Switzerland.

Laspeyres, E. (1871). Die Berechnung einer mittleren Waaren-preissteigerung. *Jahrbücher für Nationalöko-nomie und Statistik*, *16*, 296–314.

Leonard, I., Sillard, P., Varlet, G., Zoyem, J. P. (2017). Scanner data and quality adjustment. *Serie des Documents de Travail*. Working Paper No. F1704, INSEE, 1–31.

Loon, K. V., Roels, D. (2018). Integrating big data in the Belgian CPI. *Paper presented at the meeting of the group of experts on consumer price indices* (s. 8–9). Geneva, Switzerland.

Maddison, A., Rao, D. S. P. (1996). A Generalized Approach to International Comparison of Agricultural Output and Productivity. Research memorandum GD-27. *Groningen Growth and Development Centre*. Groningen, The Netherlands.

Paasche, H. (1874). Über die Preisentwicklung der letzten Jahre nach den Hamburger Börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, *23*(2/4), 168–178. DOI: 10.1515/jbnst-1874-0113.

Saraiva dos Santos, P., Lidonio, F., Cardoso, C. (2012). Scanner Data Project: the experience of Statistics Portugal. *Paper presented at the Workshop on Scanner Data* (s. 1–13). Stockholm.

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

Törnqvist, L. (1936). The Bank of Finland's Consumption Price Index. *Bank of Finland Monthly Bulletin*, *16*(10), 27–34.