# METODYKA BADANIA CEN DETALICZNYCH Z WYKORZYSTANIEM ALTERNATYWNYCH ŹRÓDEŁ DANYCH

Jacek Białek, Główny Urząd Statystyczny
Uniwersytet Łódzki

**GUS, 14 czerwiec 2022**

**Agenda:**

- nowe źródła danych w pomiarze inflacji:

  **dane skanowane** vs dane **skrapowane**

- **zalety, ograniczenia i wyzwania** związane z nowymi źródłami danych
- **wypracowane procedowanie danych** skanowanych/skrapowanych
- problem kalkulacji **indeksu cenowego: rekomendacje**
- **problem agregacji** wyników cząstkowych: **rekomendacje**

# NOWE ŹRÓDŁA DANYCH W POMIARZE INFLACJI

**- Dane skanowane**

(kody: **GTIN, EAN, SKU**, inne)


Zgodnie z definicją OECD, przez dane skanowane (scanner data) rozumiemy szczegółowe dane o dobrach konsumpcyjnych uzyskane dzięki skanowaniu ich kodów kreskowych w punktach sprzedaży (CPI Manual, 2004). Technologia użytkowania kodów kreskowych produktów pojawiła się w latach 70-tych XX wieku a ich wykorzystanie do analizy dynamiki cen i poprawy szacunków CPI nabiera szczególnego rozpędu mniej więcej od 20-25 lat. **Są to dane transakcyjne.**

# 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*) lub **SKU** (*Stock Keeping Unit*). Przykładowo, kod GTIN składa się z 8, 12, 13 lub 14 cyfr. Najbardziej popularna jest pełna wersja 13 i 14 cyfrowa. Kod GTIN składa się z: 1 cyfry wskazującej poziom pakowania, 3–cyfrowego kodu organizacji krajowej GS1 (potocznie: "kod kraju", np. 590 – Polska), 4-7 cyfr numeru jednostki kodującej GS1, 2-5 cyfr kodu produktu, 1 cyfry kontrolnej. Kod PLU jest oszczędniejszy w dostarczanych informacjach od kodu GTIN ponieważ jest krótszy, z kolei kod **SKU jest bardziej ogólny niż GTIN** lub równoważnie, GTIN jest bardziej szczegółowy niż SKU i dostarcza więcej detalicznych informacji.

W idealnym przypadku dane skanowane zawierają: **kod sprzedawcy** (określa grupę towarową wg indywidualnej klasyfikacji danej sieci), **kod identyfikujący punkt sprzedaży** w obrębie danej sieci, **etykietę produktu** (dodatkowy opis produktu i jego charakterystyki), **jednostkę sprzedaży** (optymalnie wg ujednoliconego formatu – np. „szt”, „kg”, „paczka”, „500 gr”, „1 litr” itd.) i **gramaturę**, **wartość sprzedaży**, **liczbę sprzedanych jednostek produktu**, **flagę** (np. flaguje się produkty z przecen i promocji), informację o podatku **VAT.**

# Tab. 1. Przykładowa struktura danych skanowanych w gotowej ramce danych

| time | prices | quantities | retID | EAN | coicop | description | grammage | unit | prodID |
|---|---|---|---|---|---|---|---|---|---|
| 2020-12-31 | 10,47 | 8,48 | 26-617 | 5906747171261 | 011111 | ryż długoziarnisty | 0.4 | kg | 1 |
| 2020-12-31 | 12,47 | 5,87 | 40-772 | 5906747171261 | 011111 | ryż długoziarnisty | 0.4 | kg | 1 |
| 2020-12-31 | 11,4 | 15,65 | 70-001 | 5906747171261 | 011111 | ryż długoziarnisty | 0.4 | kg | 1 |
| 2020-12-31 | 13,2 | 16,95 | 85-791 | 5906747171261 | 011111 | ryż długoziarnisty | 0.4 | kg | 1 |
| 2020-12-31 | 11,47 | 85,41 | 01-460 | 5906747171261 | 011111 | ryż długoziarnisty | 0.4 | kg | 1 |
| 2020-12-31 | 11,97 | 7,82 | 05-820 | 5906747171261 | 011111 | ryż długoziarnisty | 0.4 | kg | 1 |

**źródło: opracowanie własne**

**Zalety:**

- relatywnie tanie;

-olbrzymi wolumen;

-olbrzymi przekrój;

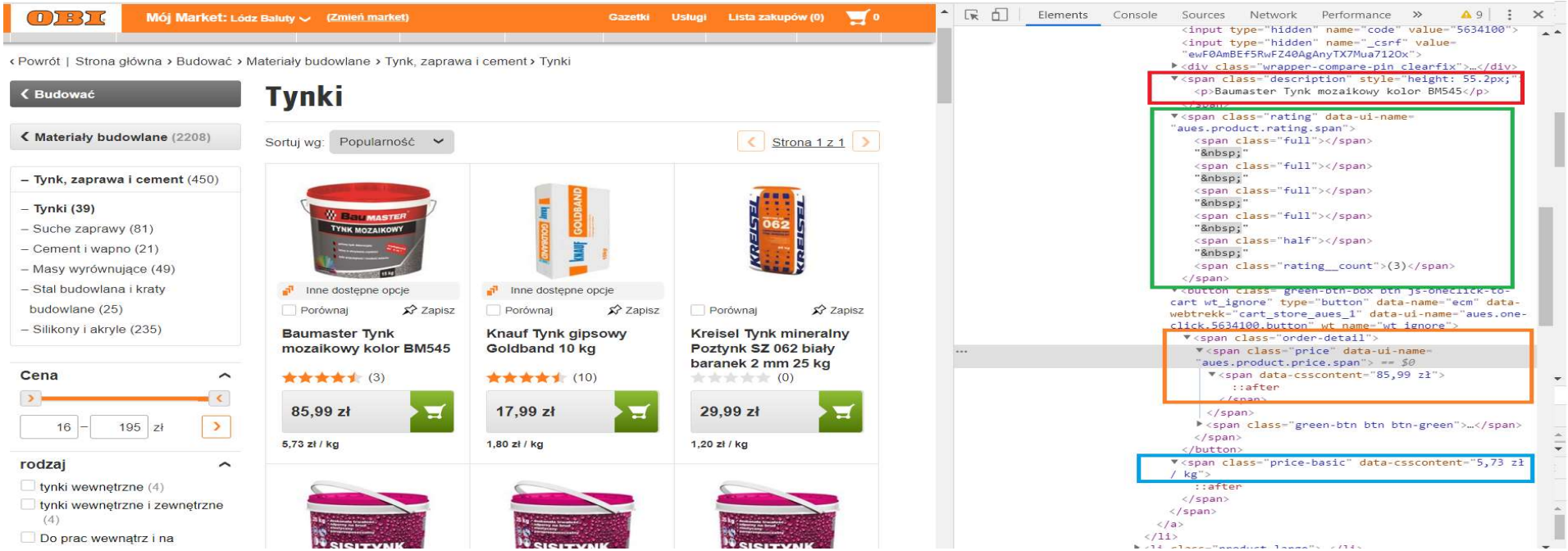
-poziom konsumpcji na EA !!!

**Pozyskiwanie:**

- od sieci handlowych (bezpieczny transfer vs API);

- wyspecjalizowane firmy badające rynek (Nielsen, GfK)

# DANE SKRAPOWANE




- **Są to ceny ofertowe;**

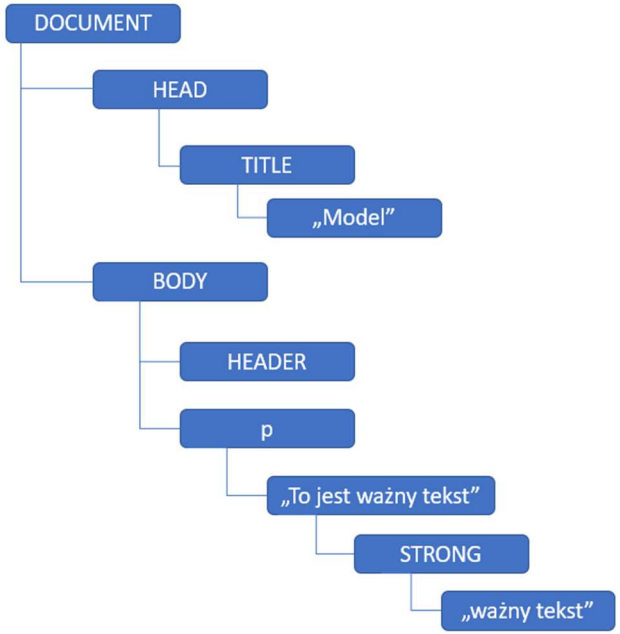
# Rys. 2. Etapy procedowania danych skrapowanych


**Zalety:** niski koszt, olbrzymi wolumen, duża elastyka doboru źródła danych

**Wady:** brak informacji o konsumpcji, nie wiemy czy produkty się sprzedały

# POZYSKIWANIE DANYCH SKRAPOWANYCH:

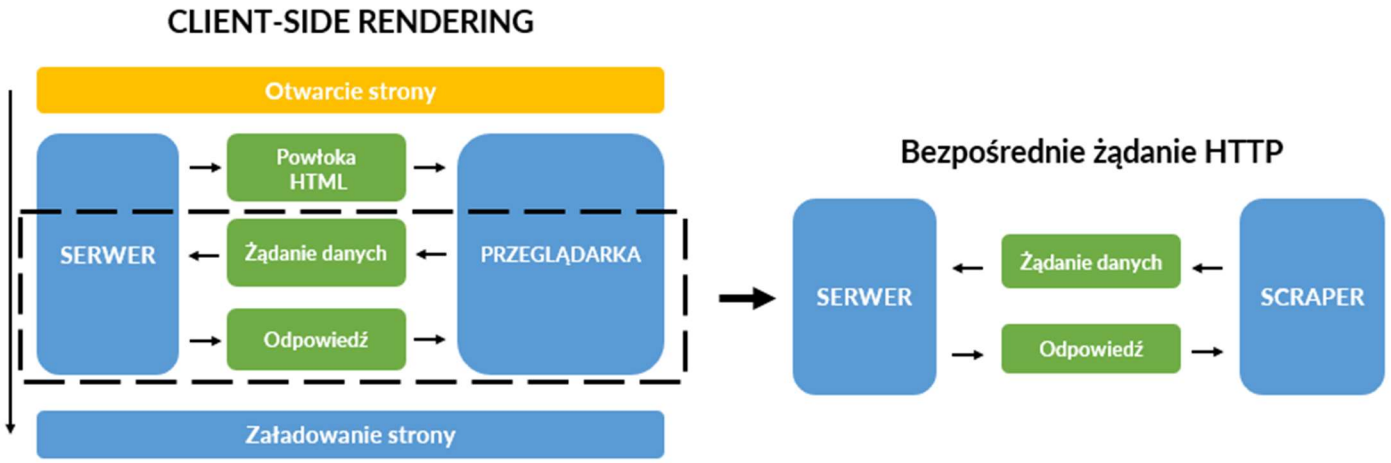
**- analiza DOM** (*Document Object Model:* sposób reprezentacji dokumentów HTML lub XML w strukturze drzewa, które należy przeszukać)**:**




**Rys. 3**

**- bezpośrednie żądania HTPP:**

z myślą o dynamicznych serwisach internetowych powstała kolejna metoda skrapowania danych: **wysyłanie żądań HTTP do endpointów** - punktów końcowych, będących punktami połączenia, w których ujawnione są dane (pliki HTML).




**Rys. 4.**

**Zaleta:** można pobrać bezpośrednio dane już ustrukturyzowane

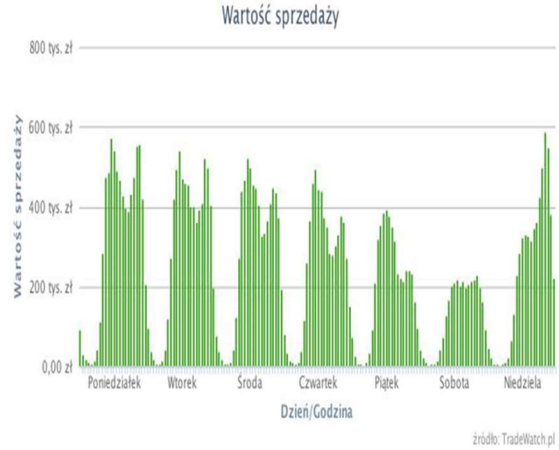
Popularną metodą pozyskiwania danych jest **dopasowywanie tekstu do wzorca** (ang. *text pattern matching*). Program dopasowuje tekst lub ciąg znaków do określonego w kodzie wzorca tzw. **wyrażenia regularnego**. Wzorce mogą zawierać zarówno treść, którą chcemy dopasować jak i znaki specjalne, które zmieniają sposób interpretacji wzorca. Wyrażenia regularne są bardzo przydatne podczas analizy tekstu, pozwalają automatycznie wydobyć z długich ciągów znaków to czego potrzebujemy.

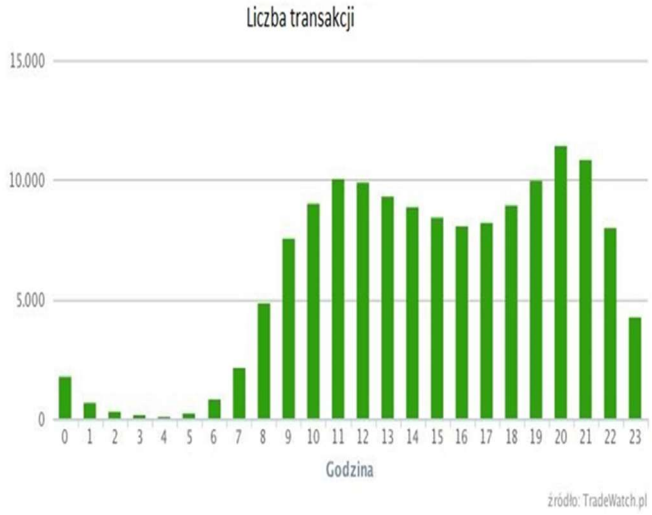
Do *text pattern matching*-u można wykorzystać między innymi język **Python** wraz z bibliotekami bs4 (***Beautiful Soup***) oraz ***requests.*** Inny popularny pakiet to **Selenium**. Można też wykorzystać język **R**, który już w wersji bazowej oferuje liczne funkcje dopasowywania wyrażeń jednak wymaga doinstalowania pakietów odpowiedzialnych za skrapowanie danych takich jak **rvest** czy **Rcrawler**.

**Ograniczenia:** dynamika stron (witryn), blokady po stronie właściciela, aktualizacje na stronie

# WYZWANIA TOWARZYSZĄCE NOWYM ŹRÓDŁOM DANYCH

- umowy z sieciami;
- dobór próby;
- klasyfikacja produktów do grup COICOP;
- dopasowanie produktów (*matching*);
- filtrowanie produktów;
- imputacja cen (ilości);
- budowa systemu IT;
- wybór formuły indeksu;
- system wag?



# PROCEDOWANIE DANYCH SKANOWANYCH

- proces „czyszczenia” zbioru danych
  i przygotowania wymaganej postaci ramki danych;
- klasyfikacja produktów
- dopasowanie produktów
- filtrowanie produktów
- obliczanie cząstkowych wskaźników cen
- agregacja cząstkowych wskaźników cen

## Klasyfikacja produktów za pomocą słów kluczowych i fraz

- Zastosowanie **metod uczenia maszynowego** do przygotowania zestawu słów kluczowych identyfikujących docelowe homogeniczne grupy produktów;
- Zastosowanie zestawu słów kluczowych lub fraz w sposób **manualny**, tak aby najlepiej identyfikowały docelowe grupy produktów

Sama detekcja słów kluczowych i fraz w etykiecie produktu jest prostym zadaniem analitycznym, wystarczy użyć do tego odpowiedniej komendy języka wysokiego poziomu (np. *str_detect()* z pakietu w R o nazwie *stringr*).

Dokładna klasyfikacja produktów wymaga jednak utworzenia całego zestawu fraz i słów kluczowych (wektorów łańcuchów tekstowych), które alternatywnie **mogą** pojawić się w etykiecie produktu, **muszą** się w niej pojawić lub też **nie mogą** być zawarte w etykiecie. Np. w pakiecie ***PriceIndices*****,** funkcja *data_selecting()*, która realizuje klasyfikację produktów poprzez ich selekcję na podstawie słów kluczowych i fraz, ma dedykowane do tego celu parametry: **include**, **must** i **exclude** (por. Tab. 1).

Tab. 1. Przykładowa klasyfikacja produktów spożywczych poprzez selekcję słów kluczowych i fraz w opisie produktu

| **Homogeniczna grupa produktów** | **Parametr *include*** | **Parametr *must*** | **Parametr *exclude*** |
|---|---|---|---|
| Mąka pszenna | "pszen", "razowa", "uniwersalna", "Uniwer", "Tradyc", "tradycyjna", "krupczatka", "tortowa", "pizz", "Szczec","wroc", "tort", "luks", "pozna", "Zamojska", "Tort","Hetma","chleb", "Wypiek","gdańsk" | „mąka” | „żyt” |
| Ryż długoziarnisty | „ziarn”, „długoziar”, „długo”, „risotto”, „parboiled”, „Basmati”, “Jaśminowy”, „paraboliczny” | „ry” | „płatki”, „platki”, „chrup”, „britta”, „natur”, |
| Chusteczki dla dzieci | „chust” | „dzieci” | „platki”, „patyczki”, „podkład”, „płatki”, „patyc”, „mydło”, „mydlo” |
| Cień do powiek | „cien”, „cień”, „EYESHADOW”, „pow” | „cie” | „baza”, „eyeliner”, "tusz", "podkład", "baza", "pomada" |
| Grzebień | „grzeb” | „grz” | „pędzel”, „pędzel”, „grzywka”, „grzyb” |

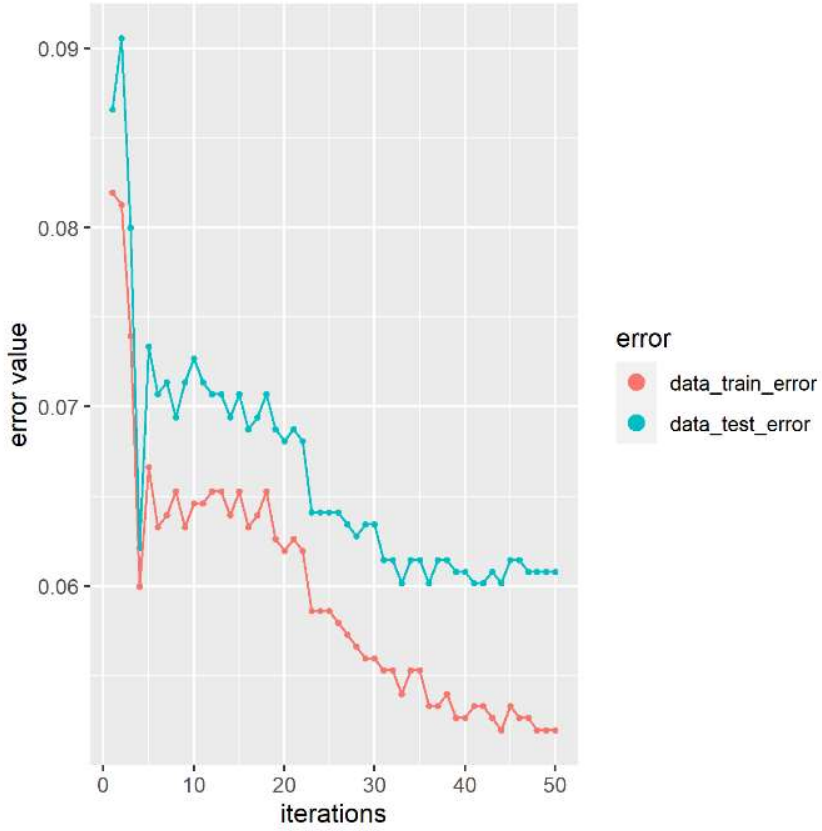
# Klasyfikacja produktów za pomocą metod uczenia maszynowego

- Przygotowanie **zbioru uczącego, walidacyjnego, testowego.**
- **Problem aktualizacji wyuczonego modelu.**

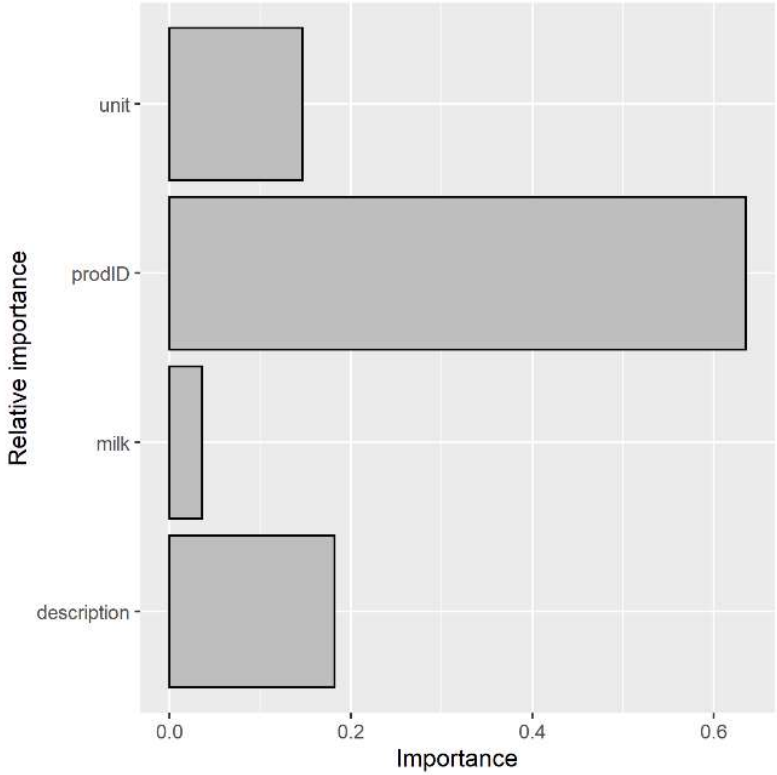
W aplikacji stworzonej przez IPI PAN zaimplementowano metody o największej skuteczności (*accuracy*): **naiwny klasyfikator Bayesowski**, metodę **SVM** oraz **lasy losowe**. Suplementarnie, do pakietu *PriceIndices* (GUS) dołączono algorytm **XGBoost** (Chen i inni, 2001) realizujący klasyfikację produktów za pomocą drzew losowych, dokonując przy tym tzw. obliczeń równoległych z wykorzystaniem rdzeni procesora i tym samym kilkukrotnie przyspieszając proces klasyfikacyjny (funkcja *data_classifying()*).

Rys. 5. Proces uczenia modelu drzew losowych zastosowanego do klasyfikacji produktów mlecznych

A) Oceny błędów klasyfikacji modelu na zbiorze uczącym i testowym




B) Ocena ważności predyktorów w procesie budowy modelu




źródło: opracowanie własne w pakiecie *PriceIndices*

# Dopasowanie produktów w czasie (matching)

W ramach projektu *InstatCeny* wypracowana została procedura dopasowywania produktów oparta zarówno o etykiety produktów jak i ich kody, tj. uwzględniono kod wewnętrzny sieci i kod zewnętrzny (GTIN, EAN lub SKU). Nie wszystkie sieci, z którymi zawarto porozumienia, dostarczają kody zewnętrzne produktu, dlatego procedura postępowania musi być elastyczna i uwzględniać każdy zakres przesyłanych danych. Najogólniej można wypracowaną procedurą przedstawić w następujących krokach:

**Krok 1)** Za dopasowane produkty uznajemy te, które mają ten sam zewnętrzny kod kreskowy (**codeOUT**) oraz ten sam wewnętrzny kod produktu (kod produktu wg sprzedawcy – **codeIN**).

**Krok 2)** W przypadku produktów, które mają równy tylko jeden ze wspomnianych kodów, brane są dodatkowo pod uwagę etykiety tych produktów. Wdrożeniowo do oceny podobieństwa etykiet zastosowano w miarę **Jaccarda.** Podobieństwo graniczne (1-miara odległości), jakie prowadzi do stwierdzenia podobieństwa porównywanych produktów, jest regulowane przez odpowiedni parametr (np. **text_sim_treshold** = 0.95). W pakiecie *PriceIndices* jest to miara **Jaro-Winklera**.

**Krok 3)** W przypadku, gdy produkty mają różny kod sprzedawcy i różny kod zewnętrzny, za dopasowane uznajemy te produkty, które mają identyczna etykietę (odległość tekstowa Jaccarda ich etykiet wynosi 0). Produkty, które nie spełniają tego wymogu są uznawane za niepodobne (niedopasowane).

# Filtrowanie danych skanowanych (skrapowanych)

## - filtr ekstremalnych zmian cen (extreme price filter)

jest to filtr, który eliminuje z próby produkty, których miesięczne zmiany cen były zbyt duże. W praktyce przyjęto, iż takimi zmianami będą miesięczne wzrosty ceny o przynajmniej 200% (oznacza to trzykrotny wzrost ceny) oraz spadek ceny o ponad 75% (czterokrotny spadek ceny).

## - filtr niskich sprzedaży (low sales filter)

to z kolei filtr, który usuwa z próby produkty o relatywnie niskiej sprzedaży na tle innych produktów z tej samej homogenicznej grupy. Warunek dla usunięcia produktu z próby: $\frac{s_i^{t-1} + s_i^t}{2} < \frac{1}{\lambda \cdot n}$ (rekomendacja: $\lambda = 1{,}25$ ).

## - filtr zrzucanych cen (dump price filter)

zgodnie z rekomendacjami, w implementacji tego filtru przyjęto 70% jako graniczną wartość dla spadku ceny i 75% dla spadku sprzedaży.

Tab. 2. Wpływ filtrowania przykładowych danych skanowanych (dane za okres: XII 2020 – XII 2021).

| **Grupa produktów: ryż** | | | | | | |
|---|---|---|---|---|---|---|
| **Rodzaj filtru** | **Liczba wierszy w bazie** | **Liczba produktów w bazie** | **Indeks Jevonsa** | **Łańcuchowy indeks Jevonsa** | **Indeks Fishera** | **Indeks GEKS** |
| Brak filtru | 48072 | 38 | 1,023846 | 1,023784 | 1,03148 | 1,028475 |
| Filtr ekstremalnych zmian cen | 48071 | 38 | 1,020172 | 1,023784 | 1,031476 | 1,028474 |
| Filtr niskich sprzedaży | 27190 | 20 | 1,024338 | 0,9989254 | 1,039426 | 1,034023 |
| Filtr zrzucanych cen | 48071 | 38 | 1,020172 | 1,023784 | 1,031476 | 1,028474 |
| **Grupa produktów: środki papiernicze i higieniczne** | | | | | | |
| **Rodzaj filtru** | **Liczba wierszy w bazie** | **Liczba produktów w bazie** | **Indeks Jevonsa** | **Łańcuchowy indeks Jevonsa** | **Indeks Fishera** | **Indeks GEKS** |
| Brak filtru | 388551 | 553 | 1,019182 | 0,9556525 | 1,080191 | 1,07359 |
| Filtr ekstremalnych zmian cen | 388084 | 506 | 1,017785 | 0,9480464 | 1,081081 | 1,07417 |
| Filtr niskich sprzedaży | 203457 | 190 | 1,050448 | 1,043644 | 1,102698 | 1,097022 |
| Filtr zrzucanych cen | 388088 | 506 | 1,017968 | 0,9556525 | 1,081104 | 1,074226 |

źródło: opracowanie własne w pakiecie *PriceIndices*

# Formuły indeksowe w pomiarze inflacji

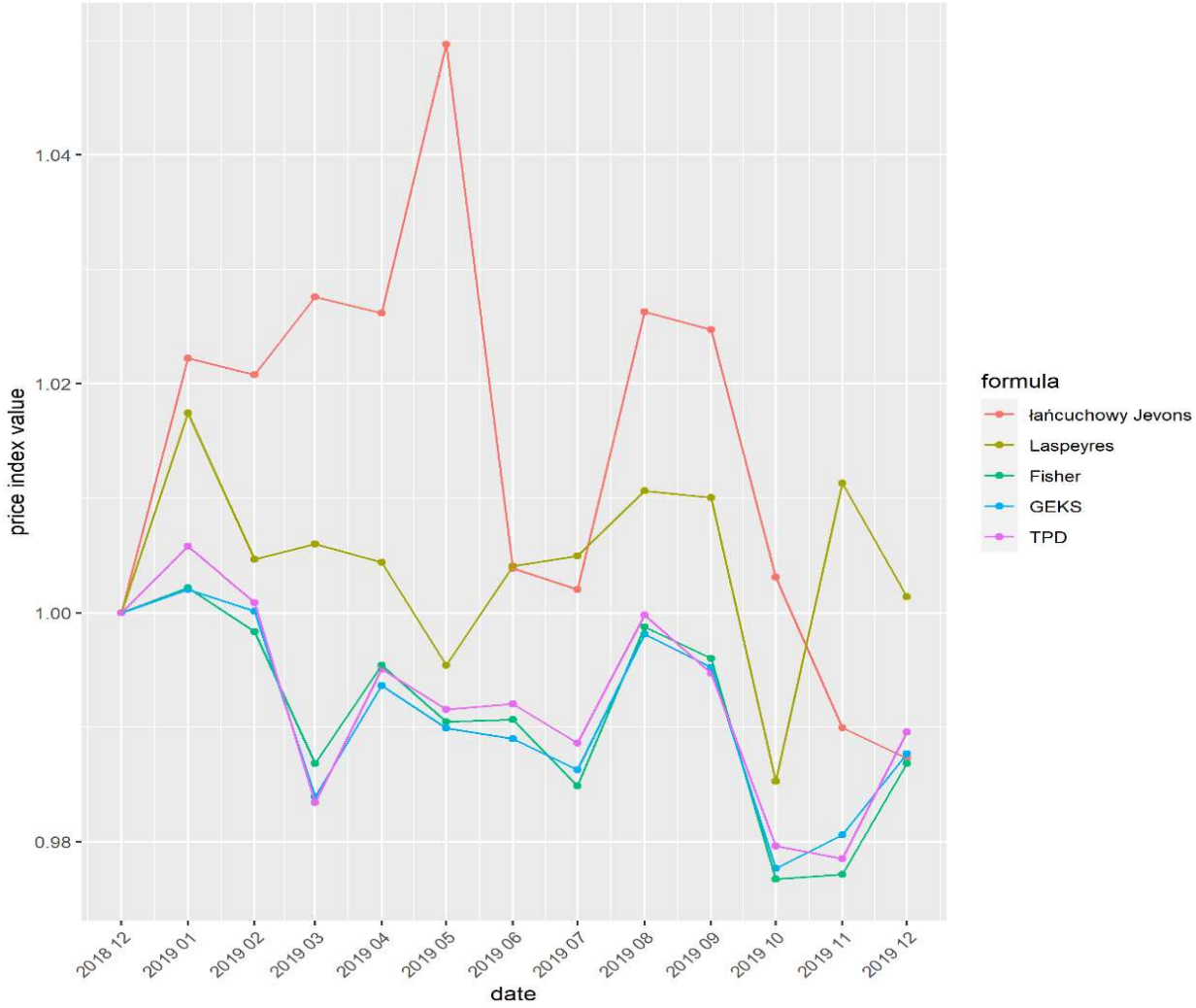
**Indeksy bilateralne**

- **Formuły nieważone**: Dutot, Carli, Jevons, CSWD, BMW;
- **Formuły ważone**: Laspeyres, Paasche, Fisher, Törnqvist, Walsh, Marshall-Edgeworth, Vartia, Sato-Vartia, Young, Lowe, Palgrave, Davies, Stuvel, Banajree, AG, Lloyd-Moutlon, Divisia;
- **Indeksy łańcuchowe** (łańcuchowy Jevons -> “dynamic approach”)

**Indeksy multilateralne**: GEKS, CCDI, GEKS-J, Geary-Khamis, TPD, FBMW, FBEW + metody aktualizacji oknem czasowym (*window splice, movement splice, half splice, mean splice*).

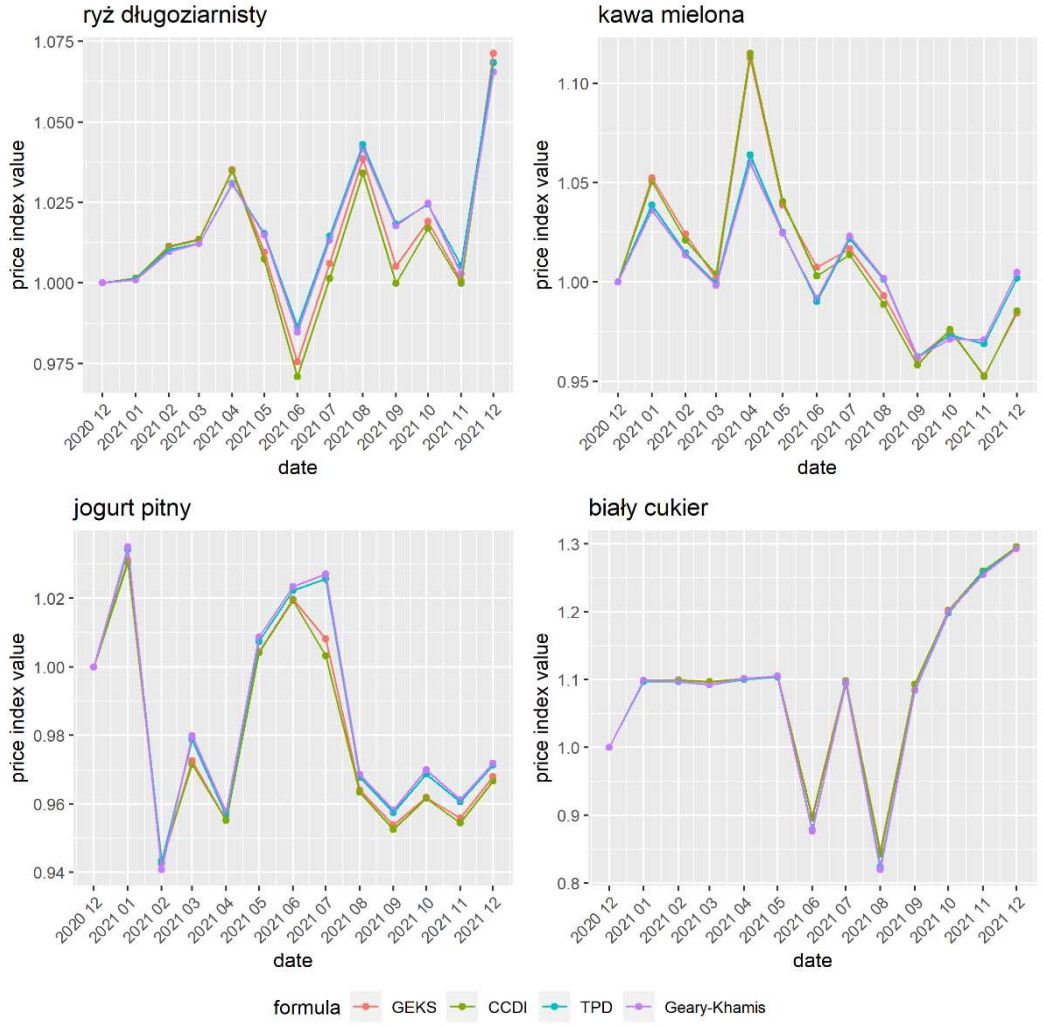
**Nurty w teorii indeksów cen**: podejście aksjomatyczne, ekonomiczne, stochastyczne (tzw. stare i nowe), czynnikowe, addytywne, łańcuchowe, Divisia.

Rys. 6. Porównanie wybranych indeksów cen dla produktów mlecznych (grudzień 2018-grudzień 2019):




źródło: opracowanie własne w pakiecie *PriceIndices*

Rys. 7. Porównanie wybranych indeksów multilateralnych dla 4 grup produktów spożywczych (okres: grudzień 2020 – grudzień 2021).

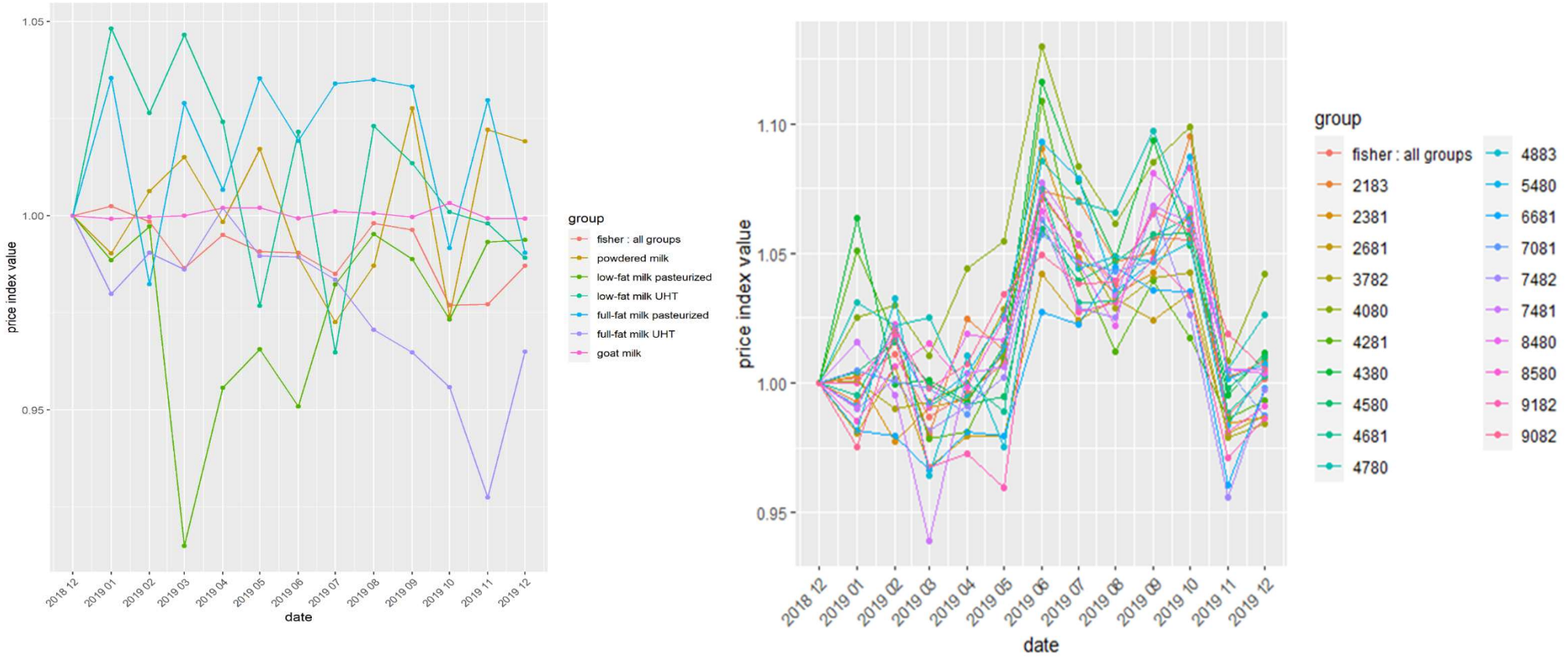



źródło: obliczenia własne w pakiecie *PriceIndices*

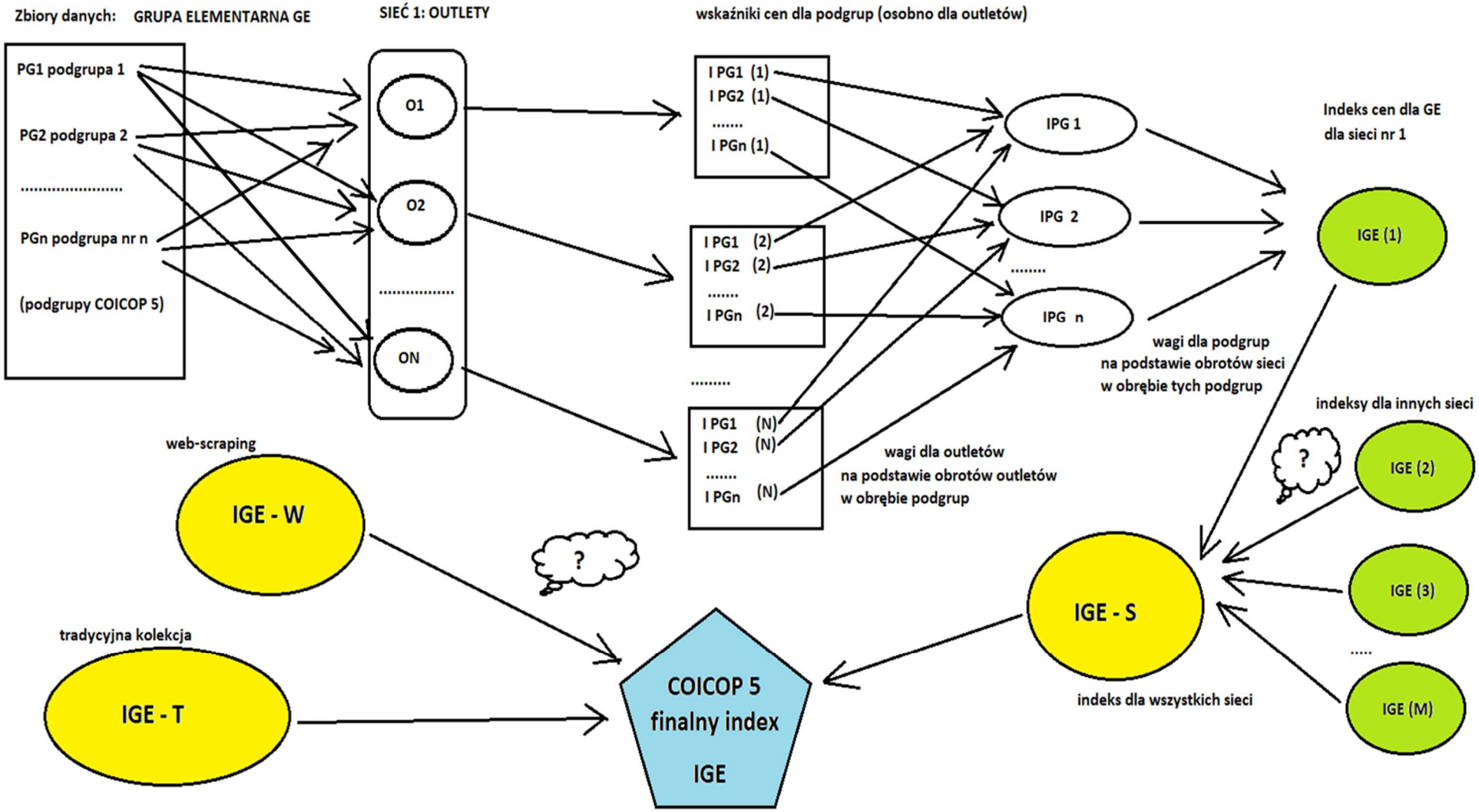
# Agregacja wyników cząstkowych:

- **Agregacja „wewnętrzna” – czemu ją stosujemy?:**

Rys. 8. Porównanie indeksu Fishera dla różnych rodzajów mleka z indeksem Fishera dla całego zbioru *mleka* (agregacja wg Laspeyresa) oraz z indeksami Fishera wyznaczonymi dla różnych outletów:



- **Agregacja „zewnętrzna” – źródła danych spotykają się na poziomie COICOP 5:**

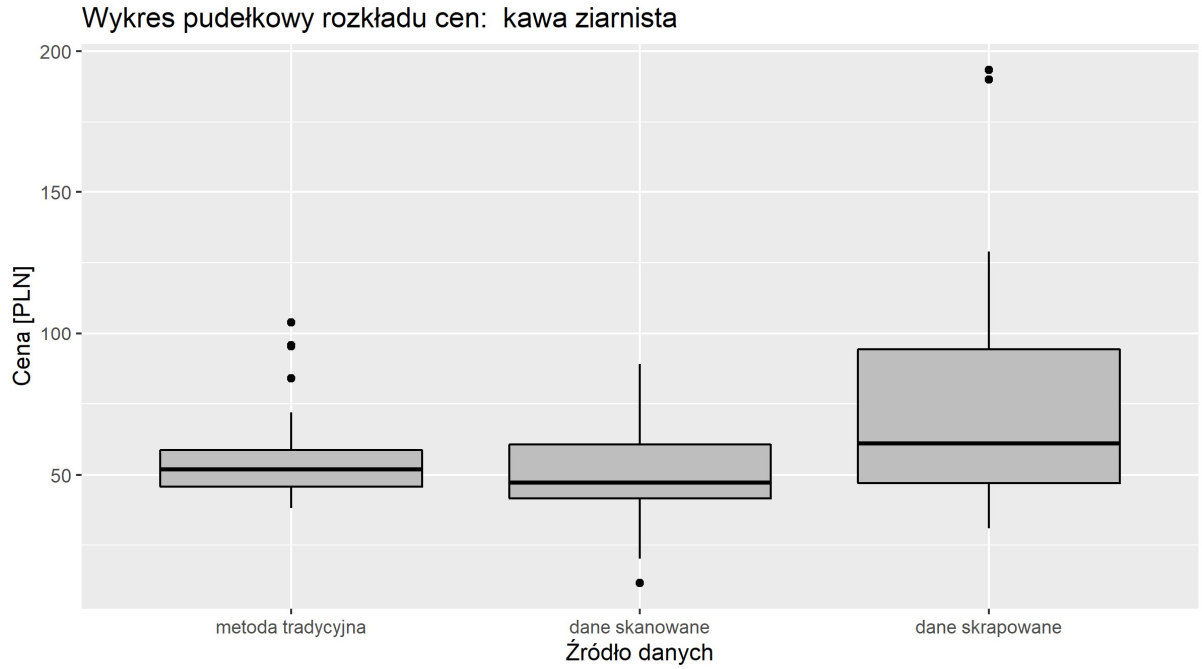


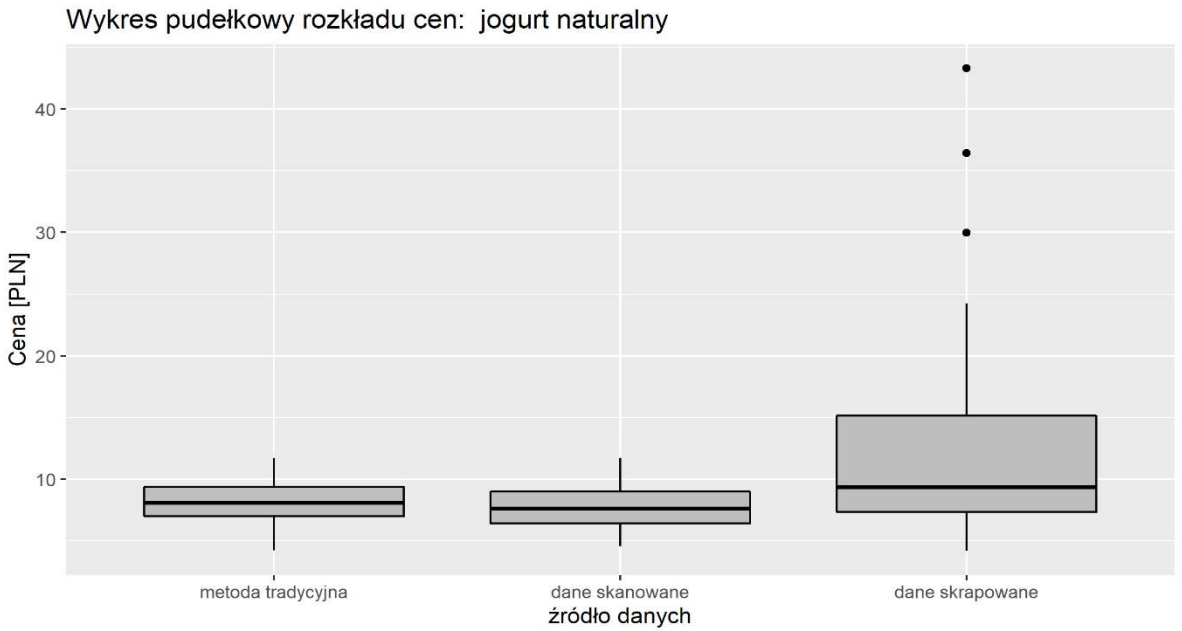
- By wyznaczyć wypadkowe wskaźniki cen poziomie COICOP 5, należy dokonać ważenia wskaźników uzyskanych ze wszystkich źródeł danych wg **formuły Laspeyresa**.

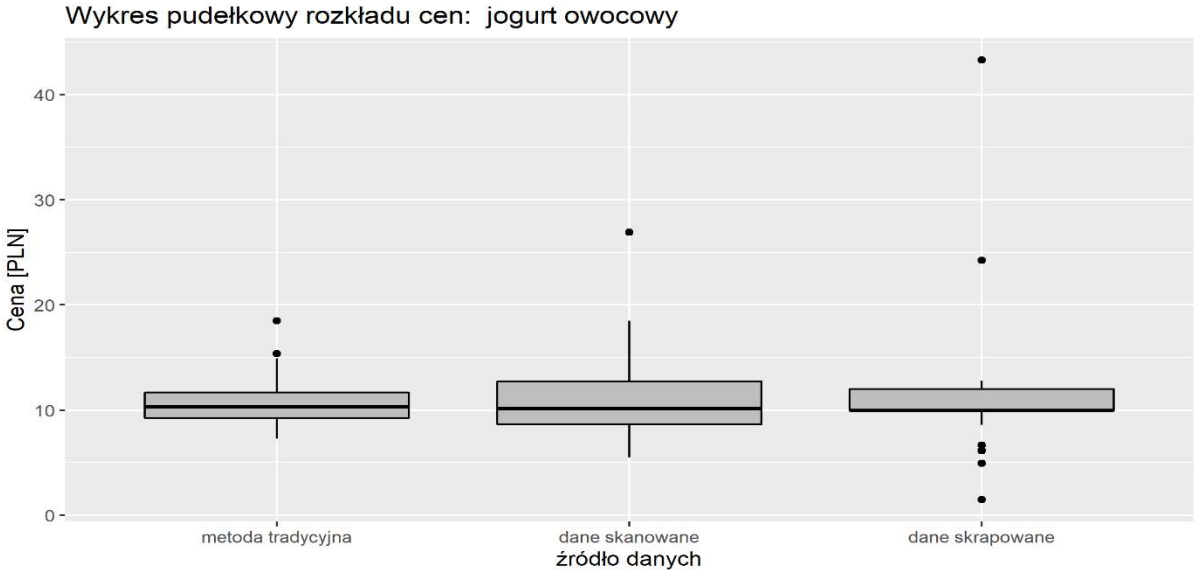
- Aby ustalić poziom wag, w projekcie *InstatCeny* udziały zakupu produktów przez Internet były szacowane na podstawie informacji uzyskiwanych z **badania budżetów gospodarstw domowych** GUS, natomiast udziały zakupów w sieciach handlowych pozyskano z baz danych **Passport GMID, Euromonitor International oraz z badań rynku wewnętrznego** prowadzonych przez GUS.

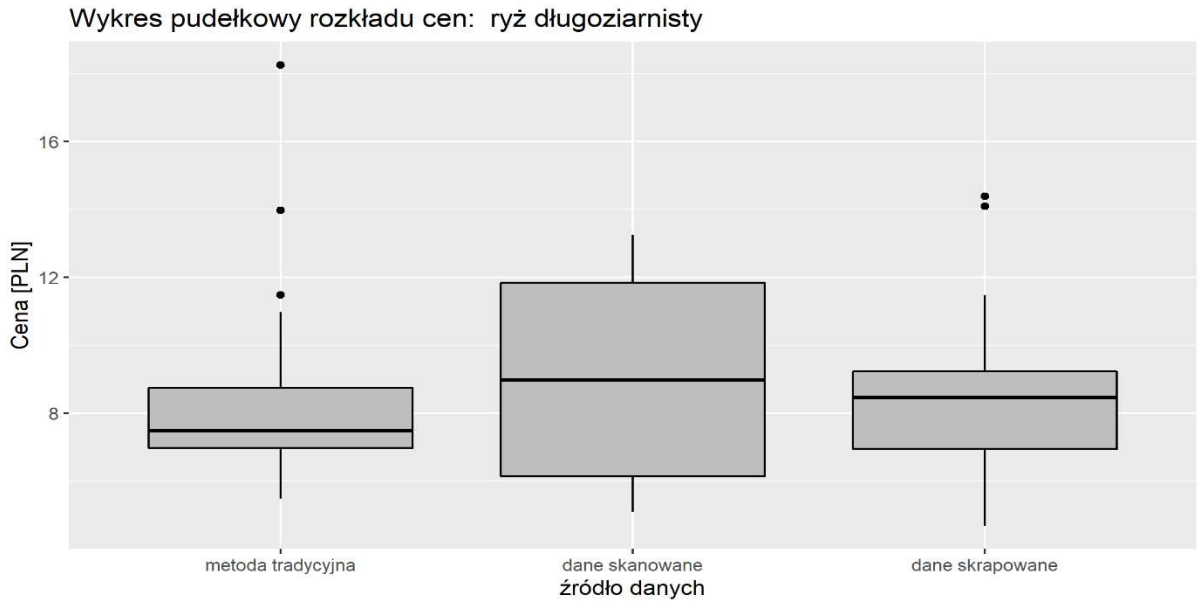
# Czy ceny i wskaźniki otrzymane z różnych źródeł mogą się znacząco różnić?

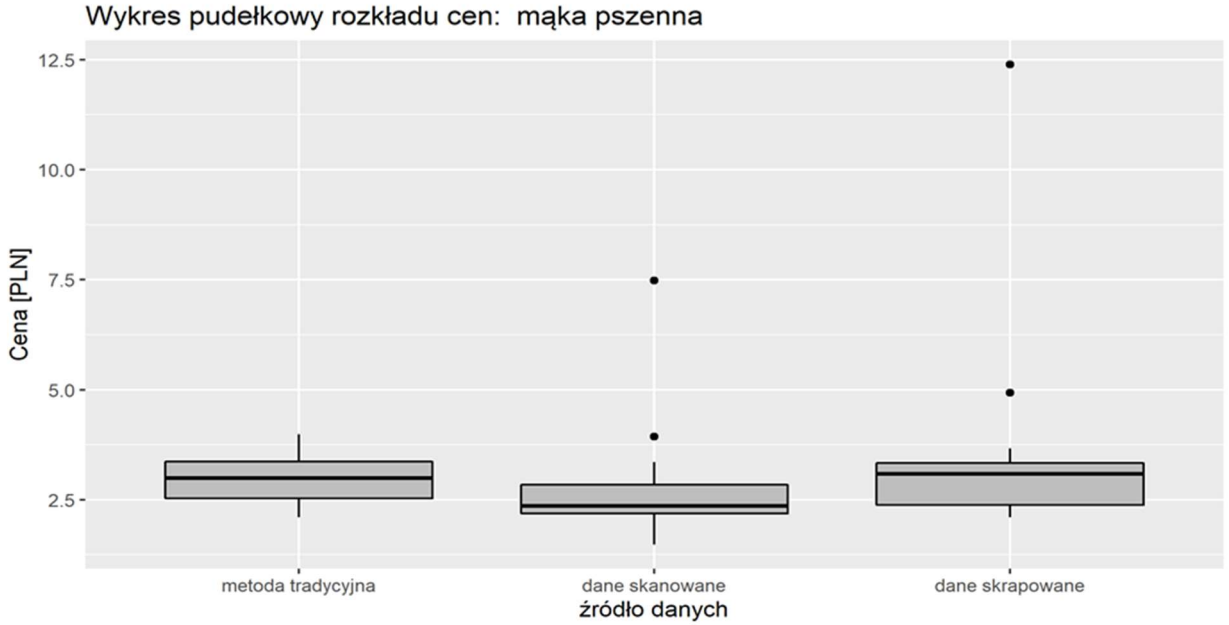
Rys. 9. Wykresy pudełkowe dla cen wybranych reprezentantów według źródła danych (III, 2021)




Wykres pudełkowy rozkładu cen: jogurt naturalny
Cena [PLN]
40
30
20
10
metoda tradycyjna
dane skanowane
dane skrapowane
źródło danych


Wykres pudełkowy rozkładu cen: jogurt owocowy
Cena [PLN]
40
30
20
10
0
metoda tradycyjna
dane skanowane
dane skrapowane
źródło danych


Wykres pudełkowy rozkładu cen: ryż długoziarnisty
Cena [PLN]
16
12
8
metoda tradycyjna
dane skanowane
dane skrapowane
źródło danych


Wykres pudełkowy rozkładu cen: mąka pszenna
Cena [PLN]
12.5
10.0
7.5
5.0
2.5
metoda tradycyjna
dane skanowane
dane skrapowane
źródło danych

**Tabl. 3.** Wskaźniki cen wybranych grup elementarnych według źródła danych, w marcu 2021 r. (luty 2021 = 100)

| Grupy elementarne i podgrupy | Wskaźniki cen | | |
|---|---|---|---|
| | 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 |
| **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 | x | 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 | x | 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 | x | 98,74 | 98,92 |

Pierwsze doświadczenia w zakresie łączenia różnych źródeł danych do obliczania wskaźników cen skłaniają do preliminarnego wniosku, iż wskaźnik cen produktów spożywczych opracowany **na podstawie danych skanowanych może zawyżać** wskaźnik cen na poziomie ECOICOP 5 (najczęściej otrzymano tu największy wskaźnik cen i jednocześnie najrzadziej najmniejszy), a z **kolei wskaźnik cen tych produktów wynikający z cen skrapowanych może dawać wynik niższy** w stosunku wyniku pozyskanego na podstawie danych ankieterów (najczęściej otrzymano tu najmniejszy wskaźnik cen i jednocześnie najrzadziej największy).

Wsparcie dla tego wniosku:

Białek, J., Dominiczak-Astin, A., Turek, D. (2021). *Porównanie cen i wskaźników cen konsumpcyjnych: tradycyjna metoda uzyskiwania danych a źródła alternatywne*, Wiadomości Statystyczne. The Polish Statistician, 66(9), 32-69.

Ueda K, Watanabe, K. Watanabe, T. (2022). *Price Setting in Online and Offline Markets Evidence from Korea*, Paper presented at the 17th Meeting of the Ottawa Group on Price Indices, Rome, Italy (dotyczy danych skrapowanych).

Dziękuję za uwagę!

## BIBLIOGRAFIA


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

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

Białek, J., Beręsewicz, M. (2021). *Scanner data in inflation measurement: From raw data to price indices*, Statistical Journal of the IAOS 37, 1315–1336.

Caves D.W., Christensen, L.R., Diewert, W.E. (1982). Multilateral comparisons of output, input, and productivity using superlative index numbers. *Economic Journal,* 92, 73-86.

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,* 2016(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., 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.

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

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

Dalen, J. (1997). Experiments with Swedish Scanner Data. *Proceedings of the Third Meeting of the International Working Group on Price Indexes*, Balk (ed.), 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, 114-145.

Diewert, W.E., Fox, K.J. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. *Discussion paper,* 17(2), Vancouver School of Economics, The University of British Columbia, Vancouver, Canada.

Eltetö, Ö., Köves, P. (1964). 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*. Boston: Houghton Mifflin.

Geary, R.G. (1958). A Note on Comparisons of Exchange Rates and Purchasing Power between Countries. *Journal of the Royal Statistical Society Series* A, 121, 97-99.

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

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

Inklaar, R., Diewert, W. E. (2016). Measuring Industry Productivity and Cross-Country Convergence. *Journal of Econometrics,* 191, 426-433.

Jevons, W.S. (1865). The variation of prices and the value of the currency since 1782. *J. Statist. Soc. Lond*., 28, 294-320.

Khamis, S.H. (1972). A New System of Index Numbers for National and International Purposes. *Journal of the Royal Statistical Society Series A,* 135, 96-121.

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 Waarenpreissteigerung. *Jahrbücher für Nationalökonomie 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 Borsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, 12, 168—178.

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. (in Polish). *Przeglad Statystyczny,* 3, 239-254.