Dane skanowane w pomiarze CPI

Nowe możliwości i wyzwania

Jacek Białek, Uniwersytet Łódzki, GUS
Anna Bobel, GUS

18-19.03.2019, Łódź

Plan prezentacji

1. Dane skanowane jako nowe źródło danych o cenach
2. Dostawcy danych skanowanych
3. Zalety i wady danych skanowanych
4. Wybrane problemy i wyzwania
5. Metodologia indeksów cen dla danych skanowanych
6. Dotychczasowe doświadczenia GUS - innowacyjność
7. Przykład empiryczny
8. Wnioski

1. Dane skanowane jako nowe źródło danych o cenach

Definicja

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*).

Przegląd literatury przedmiotu skłania jednak do wniosku, iż pojęcie danych skanowanych jest dzisiaj nieco szersze i obejmuje **elektroniczne dane dotyczące transakcji** (przynajmniej tak drobiazgowe, jak drobiazgowy jest kod kreskowy sprzedawanych produktów) uzyskane od sieci handlowych.

1.Dane skanowane jako nowe źródło danych o cenach (cd.)

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*).

Najbardziej rozpowszechniony jest kod GTIN (EAN), choć na świecie funkcjonują też bardzo specyficzne szczególne jego przypadki, np. **UPC** (*Universal Product Code*), czy lokalny **APN** (*Australian Product Number*). Przykładowo, kod GTIN składa się z 8, 12, 13 lub 14 cyfr. Najbardziej popularna jest pełna wersja 13 cyfrowa.

Kod GTIN składa się z:

- 1 cyfry wskazującej 'poziom pakowania'
- 3–cyfrowego kodu organizacji krajowej GS1 ("kod kraju", np. 590 – Polska)
- 4-7 cyfr numeru jednostki kodującej GS1 (potocznie: "numer firmy")
- 2-5 cyfr kodu produktu
- 1 cyfry kontrolnej

1. Dane skanowane jako nowe źródło danych o cenach (cd.)

Geneza

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 18-20 lat. Przykładowo, do 2015 roku w Unii Europejskiej tylko **Holandia** (od 2001 roku!)**, Norwegia, Szwecja i Szwajcaria** używały danych skanowanych, a zaledwie rok później, dołączyły do nich kolejne kraje: **Belgia, Dania i Islandia**. Z publikacji do roku 2018 wynika, że obecnie również **Luxemburg, Portugalia i Francja** eksperymentują z danymi skanowanymi dla wybranych podgrup koszyka CPI.

Polska (GUS, IPI PAN, SGH) rozpoczyna projekt *InstatCeny*, który ukierunkowany jest na wykorzystanie alternatywnych źródeł danych przy kalkulacji CPI, w tym także danych skanowanych.

2. Dostawcy danych skanowanych

- Najcenniejszym źródłem tego typu danych wydają się być **bezpośredni dostawcy**, a więc punkty sprzedaży ze szczególnym uwzględnieniem sieci supermarketów. **Supermarkety** to potężni potencjalni dostawcy danych skanowanych – typowy supermarket posiada bazę 10 000 – 25 000 kodów kreskowych sprzedawanych produktów, z których większość stanowi żywność i napoje.
- Teoretycznie podobnymi dostawcami danych skanowanych (czy też szerzej - transakcyjnych) mogą być również mniejsze **markety, drobni sprzedawcy** lub nawet **sklepy internetowe**, o ile tylko archiwizują dane o sprzedaży uwzględniając wspomniane kody produktów. Mogą to być **apteki** czy **biura podróży**.
- Kolejnym, alternatywnym źródłem danych skanowanych mogą być **firmy wyspecjalizowane w badaniu rynku**. Np. niektóre kraje korzystają z danych skanowanych dostarczanych przez firmę **Nielsen** lub **GfK** i włączają je do szacunków krajowego CPI (są to najczęściej dane płatne).
- Potężne **elektroniczne platformy handlowe** (OLX, Allegro).
- E-paragon(?).

3. Zalety i wady danych skanowanych

Doświadczenia międzynarodowe

Zalety:

- relatywnie niski koszt,
- olbrzymi wolumen danych,
- niemal pełna automatyzacja procesu kolekcji danych,
- niższa agregacja niż ECOICOP6,
- informacja o cenach i ilościach na poziomie elementarnym,

Wady:

- wtórne wykorzystanie danych zbieranych w innym celu,
- wymagają stworzenia nowego IT,
- wymagają sporządzenia (czasochłonnych) umów z dostawcami,
- wymagają zaawansowanych technik statystyczno-informatycznych,
- metodologia obarczona mnogością wciąż otwartych problemów,

4. Wybrane problemy i wyzwania

- wybór dostawcy i współpraca z nim,
- losowy dobór próby,
- wybór formuły indeksowej na najniższym poziomie agregacji (patrz pkt. 5),
- klasyfikacja produktów do homogenicznej grupy,
- dopasowywanie produktów,
- produkty sezonowe,
- filtrowanie danych,
- imputacja danych,
- wagi przy przejściu do wyższych poziomów agregacji,
- największe wyzwanie: IT.

5. Metodologia indeksów cen dla danych skanowanych

- **Cenowe indeksy bilateralne (bezpośrednie)**

  a) **Nieważone indeksy (+ wersje łańcuchowe)**

Indeks Jevonsa (1865)

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

Indeks Carli (1804)

$$P_C^{0,t} = \frac{1}{N_{0,t}} \sum_{i \in G_{0,t}} \frac{p_i^t}{p_i^0}$$

b) Ważone indeksy (+ wersje łańcuchowe)

Indeks Fishera (1922)

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

Indeks Törnqvista (1936)

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

Indeks Walsha (1901)

$$I_W^{0,t} = \frac{\sum_{i \in G_{0,t}} p_i^t \cdot \sqrt{q_i^0 q_i^t}}{\sum_{i \in G_{0,t}} p_i^0 \cdot \sqrt{q_i^0 q_i^t}}$$

- **Cenowe indeksy multilateralne**

Indeks GEKS

Gini (1931), Eltetö and Köves (1964), Szulc (1964)

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

Indeks CCDI

Caves, Christensen and Diewert (1982), Inklaar and Diewert (2016)

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

Indeks Geary-Khamisa (GK)

Geary (1958), Khamis (1972)

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

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

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

(rozwiązanie układu równań symultanicznie)

Inne metody: TPD, Real Time Index, CM, inne…

Metody aktualizacji wag (window updating methods)

The movement splice method $$P_{MS}^{0,t} = P_{MS}^{0,t-1} \cdot P_{t-T,t}^{t-1,t}$$

The window splice method $$P_{WS}^{0,t} = P_{WS}^{0,t-1} \cdot \frac{P_{t-T,t}^{t-T,t}}{P_{t-T-1,t-1}^{t-T,t-1}}$$

The half splice method $$P_{HS}^{0,t} = P_{HS}^{0,t-1} \cdot \frac{P_{t-T,t}^{t-t_0,t}}{P_{t-T-1,t-1}^{t-t_0,t-1}}$$

The mean splice method $$P_{GMS}^{0,t} = P_{GMS}^{0,t-1} \cdot \prod_{t_0=1}^{T} \left(\frac{P_{t-T,t}^{t-t_0,t}}{P_{t-T-1,t-1}^{t-t_0,t-1}}\right)^{\frac{1}{T}}$$

6. Dotychczasowe doświadczenia GUS – innowacyjność

- Podstawy współpracy z sieciami handlowymi – Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/792 z dnia 11 maja 2016 r., Program Badań Statystyki Publicznej, konieczne **indywidualne porozumienia** z gestorami danych,
- Nowy model współpracy przy pozyskaniu danych oparty na negocjacjach – gestorzy oczekują **minimalizacji obciążenia**, raportowania **informacji zwrotnej** i **integracji obowiązków** sprawozdawczych,
- Zakres danych zróżnicowany zależnie dostawcy danych, różne metody dla poszczególnych grup produktów – wyzwanie dla pełnej kontroli nad procedurami analitycznymi, automatyzacji i skalowalności systemu,

6. Dotychczasowe doświadczenia GUS – innowacyjność (cd.)

- **Wieloaspektowe zastosowanie** danych – na potrzeby CPI (nie tylko jako źródło danych do pomiaru zmian cen, ale także podstawa doskonalenia założeń definiowania reprezentantów dla notowań ankieterów), statystyki rachunków narodowych, sprzedaży detalicznej,
- **Możliwość współpracy** z jednostkami naukowymi, urzędami statystycznymi, Centrum Informatyki Statystycznej - symulacje i transfer do praktyki badania nowych rozwiązań, np. algorytmów z obszaru sztucznej inteligencji,
- **Wzmocnienie kompetencji** analitycznych,
- Aktywny **udział w tworzeniu nowych i doskonaleniu istniejących międzynarodowych wytycznych** metodologicznych dedykowanych danym skanowanym - Grupa Ekspertów ds. Danych Skanowanych UE, Międzynarodowa Grupy Roboczej ds. Wskaźników Cen (Ottawa Group).

7. Przykład empiryczny

Produkt: fotelik dziecięcy samochodowy, 15-36 kg.

(ECOICOP5– 12.3.2.2)

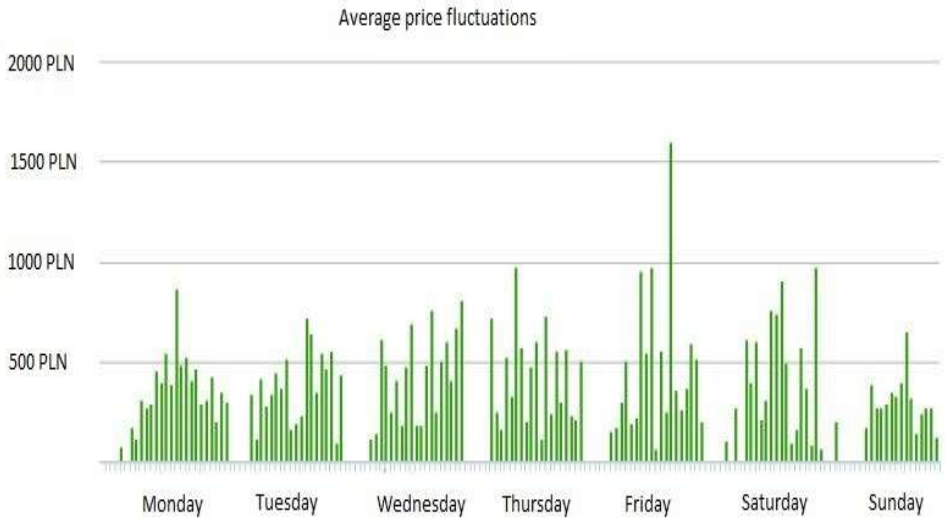
Dane: allegro.pl, okres: 04.12.2016-28.12.2018

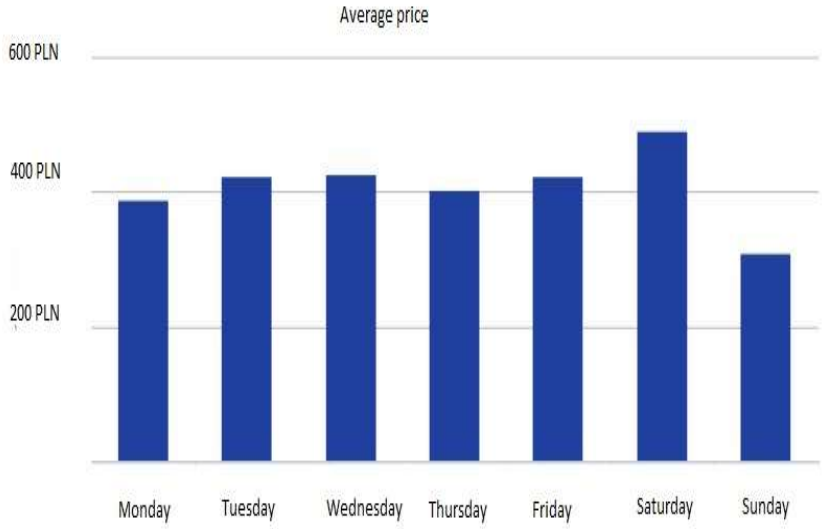
Przykładowy produkt z homogenicznej grupy fotelików:

EAN: 3660730036211
FOTELIK SAMOCHODOWY GRACO ISOFIX 15-36 KG

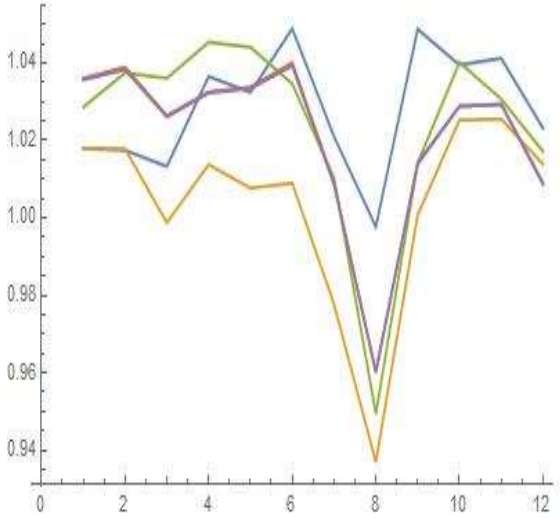
Sprzedaż łączna	3 603 355,87 zł
Śr. sprzedaż dzienna	4 772,66 zł
Śr. cena 1 szt.	193,22 zł
Śr. wartość 1 transakcji	207,35 zł
Sprzedane sztuki	18 649
Zawarte transakcje	17 378

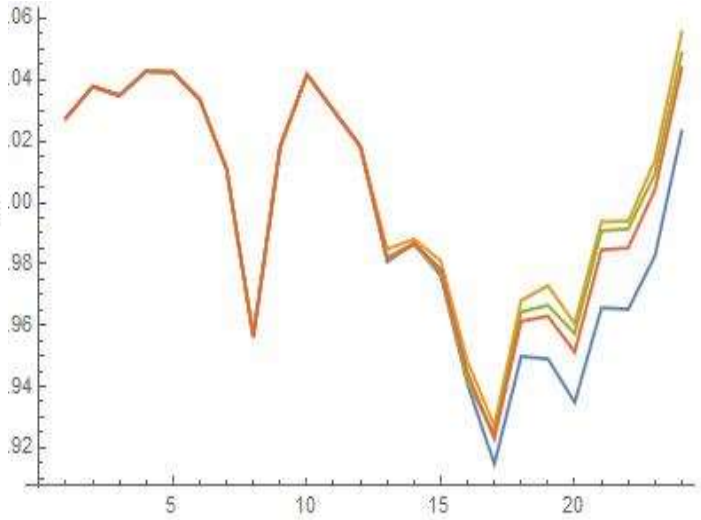
Możliwości *TradeWatch*:

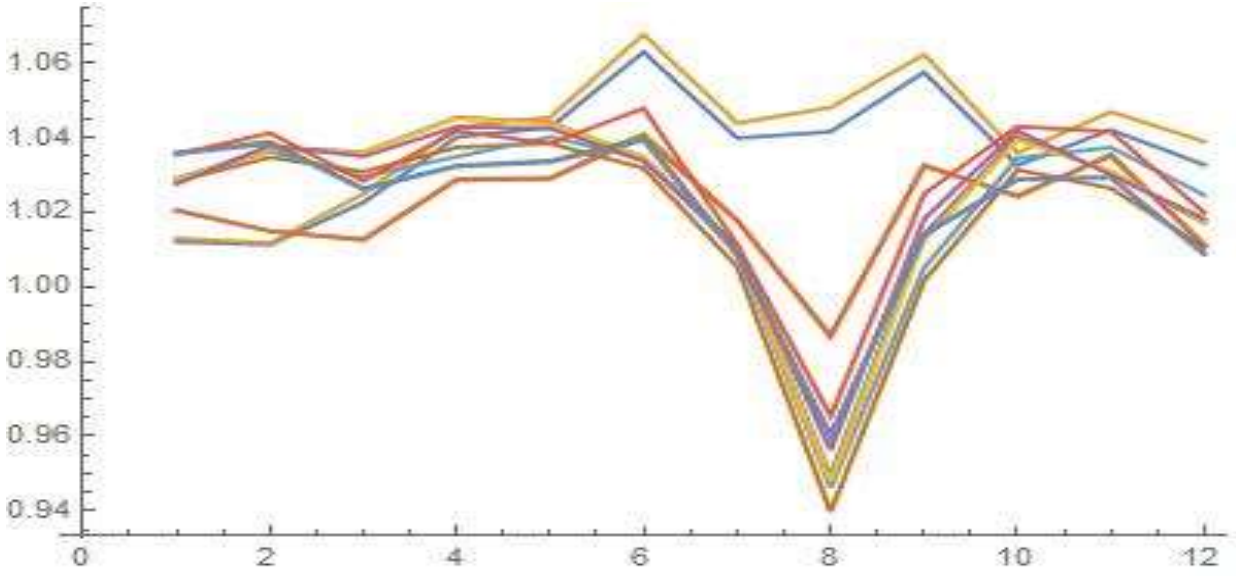


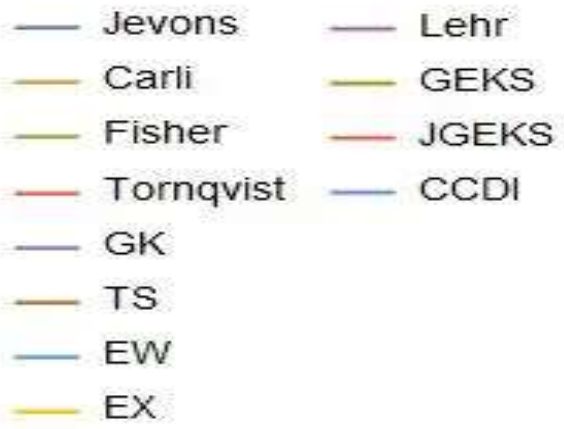


Przykładowa analiza porównawcza formuł indeksów









8. Wnioski

- Nadal jest dużo pytań i **otwartych problemów** (ale mamy już pierwsze doświadczenia z danymi skanowanymi);
- Konieczne są **prace eksperymentalne** np. nad wyborem optymalnej formuły indeksu;
- Konieczna jest **współpraca** środowiska naukowego, jak i praktyków badania cen konsumpcyjnych;
- Niezbędne są:
  a) kontynuacja **gromadzenia danych** od dotychczasowych dostawców oraz (?) nowych;
  b) stworzenie metodologii **„obróbki danych”** oraz odpowiednie definiowanie **homogenicznych grup produktów** (stopień agregacji ?);
  c) budowa **środowiska IT**.

Literatura

(ABS) Australian Bureau of Statistics. (2016). Making Greater Use of Transactions Data to Compile the Consumer Price Index, Information Paper 6401.0.60.003, 29 November 2016, Canberra.

Carli, G. (1804). Del valore e della proporzione de'metalli monetati. In: Scrittori Classici Italiani di Economia Politica, 13, 297-336.

Caves D.W., Christensen, L.R. and 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, 20-22 May 2015, Urayasu City, Japan.

Chessa, A.G. (2016), "A New Methodology for Processing Scanner Data in the Dutch CPI", Eurona 1/2016, 49-69.

Chessa, A.G., and Griffioen, R. (2016). Comparing Scanner Data and Web Scraped Data for Consumer Price Indices. Report, Statistics Netherlands.

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, 14-15 March 2017. Statistics Netherlands.

Chessa, A.G., Verburg, J., and 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, 10-12 May 2017

Consumer Price Index Manual. Theory and practice. (2004). ILO/IMF/OECD/UNECE/Eurostat/The World Bank, International Labour Office (ILO), Geneva.

Diewert, W. E., 1976. Exact and superlative index numbers. Journal of Econometrics 4, 114-145.

Diewert, W. E. (2012). Consumer price statistics in the UK. Office for National Statistics, Newport.

Diewert, W.E., and Fox, K.J. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. Discussion paper 17-02, Vancouver School of Economics, The University of British Columbia, Vancouver, Canada.

de Haan, J. (2006). The re-design of the Dutch CPI. Statistical Journal of the United Nations Economic Commission for Europe, 23, 101-118.

de Haan, J., van der Grient, H.A. (2011). Eliminating chain drift in price indexes based on scanner data. Journal of Econometrics, 161, 36-46.

de Haan, J., and Krsinich, F. (2014). Time Dummy Hedonic and Quality-Adjusted Unit Value Indices: Do They Really Differ? Paper presented at the Society for Economic Measurement Conference, 18-20 August 2014, Chicago, U.S.

de Haan J.(2015). A Framework for Large Scale Use of Scanner Data in the Dutch CPI. Paper presented at the 14th Ottawa Group meeting, Tokyo, Japan.

de Haan, J., Willenborg, L., and Chessa, A.G. (2016). An Overview of Price Index Methods for Scanner Data. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 2-4 May 2016, Geneva, Switzerland.

Eltetö, Ö., and Köves, P. (1964), "On a Problem of Index Number Computation Relating to International Comparisons", (in Hungarian), Statisztikai Szemle 42, 507-518.

FEENSTRA, R. C., MA, H., and RAO, D. S. PRASADA. (2009). "Consistent comparisons of real incomes across time and space". Macroeconomic Dynamics, 13(S2), pp.169-193.

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.

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

Griffioen, A.R., and ten Bosch, O. (2016). On the Use of Internet Data for the Dutch CPI. Paper presented at the UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices, 2-4 May 2016, Geneva, Switzerland.

Hill, R.J. (2000), 'Measuring substitution bias in international comparisons based on additive purchasing power parity methods', European Economic Review, 44, pp. 145–162.

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

Ivancic, L., Diewert, W.E., and Fox, K.J. (2011). Scanner Data, Time Aggregation and the Construction of Price Indices. Journal of Econometrics, 161 (1), 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, 26-28 May 2014, Geneva, Switzerland.

Lamboray C. (2017). The Geary Khamis index and the Lehr index: how much do they differ?. Paper presented at the 15th Ottawa Group meeting, 10-12 May 2017, Elville am Rhein, Germany.

Laspeyres E. (1871), Die Berechnung einer mittleren Waarenpreissteigerung, Jahrbücher für Nationalökonomie und Statistik, 16, 296—314.

Levell, P. (2015). Is the Carli index flawed?: assessing the case for new retail price index RPIJ. J. R. Statist. Soc. A, 178(2), 303-336.

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, 8-9 May 2018, Geneva, Switzerland.

Maddison, A., and 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,

Summers, R. (1973). International Price Comparisons Based Upon Incomplete Data. Review of Income and Wealth, 19, 1-16.

Szulc, B.J. (1964), "Indices for Multiregional Comparisons", (in Polish), Przeglad Statystyczny 3, 239-254.

Törnqvist, Leo. 1936. The Bank of Finland's Consumption Price Index, Bank of Finland Monthly Bulletin 10, 1-8.

Walsh C. M. (1901). The Measurement of General Exchange Value, The MacMillan Company: New York.

Willenborg, L. (2010). Chain Indexes and Path Independence. Report, Statistics Netherlands.

Willenborg, L., and van der Loo, M. (2016). Transitivizing Price Index Numbers Using the Cycle Method: Some Empirical Results. Report, Statistics Netherlands.

Willenborg, L. (2017). Transitivizing Elementary Price Indexes for Internet Data using the Cycle Method. Discussion Paper, Statistics Netherlands

Van der Grient, H.A., de Haan, J. (2010). The use of supermarket scanner data in the Dutch CPI. Paper presented at the Joint ECE/ILO Workshop on Scanner Data, 10 May 2010,Geneva.

Von Auer L. (2017). Processing scanner data by an augmented GUV index. Eurostat Review of National Accounts and Macroeconomic Indicators, 1/2017, 73-91.

Dziękujemy za uwagę

Jacek Białek, Uniwersytet Łódzki, GUS
J.Bialek@stat.gov.pl

Anna Bobel, GUS
A.Bobel@stat.gov.pl