SCANNER DATA, PRICE INDICES AND THE CPI CHAIN DRIFT BIAS

Jacek Białek – Elżbieta Roszko-Wójtowicz

Abstract

Scanner data mean transaction data that specify product prices and their expenditures obtained from supermarkets' IT systems by scanning bar codes (i.e. GTIN or SKU). Scanner data are a relatively new and cheap data source for the calculation of the Consumer Price Index (CPI) and the main advantage of using scanner data is the fact that they provide full information about products even on the lowest data aggregation level. One of main challenges while using scanner data is the choice of the appropriate price index formula. The list of potential price indices, which could be used in the scanner data case, is quite wide, i.e. bilateral and multilateral indices are used in practice. For instance, some countries use the chain Jevons price index formula while some other countries prefer the multilateral GEKS index or the Geary-Khamis method. One of the most important criterions in selecting index formula for scanner data case is the potential reduction of the chain drift bias. The chain drift occurs if the index differs from unity when prices revert back to their base level. In the paper we present some simulation results which show the situations on the market leading to the serious chain drift bias.

Key words: scanner data, Consumer Price Index, price indices, multilateral indices, chain drift

JEL Code: C43, E31

Introduction

Scanner data are a quite new data source for statistical agencies and the availability of electronic sales data for the calculation of the Consumer Price Index (CPI) has increased over the past 18 years. Scanner data can be obtained from a wide variety of retailers (supermarkets, home electronics, Internet shops, etc.). Scanner data mean transaction data that specify turnover and numbers of items sold by GTIN (barcode, formerly known as the EAN code). Scanner data have numerous advantages compared to traditional survey data collection because such data sets are much bigger and cheaper than traditional ones and they contain complete transaction information, i.e. information about prices and quantities. In other words, scanner data contain expenditure information at the item level, which makes it possible to use expenditure shares of items as weights for calculating price indices at the lowest (elementary) level of data aggregation. Scanner data sets are huge and may provide some additional information about products (such as the following attributes: size, colour, package quantity, etc.). These attributes may be useful in aggregating items into homogeneous groups. Nevertheless, there are lots of challenges while using scanner data sets.

The first challenge connected with scanner data concerns item codes. The following codes may be used: global trade article number GTIN, price look-up (PLU) and stock-keeping units (SKUs). PLU codes are shorter than GTINs and SKUs can be slightly more generic than GTINs. The next challenge is detecting items which were returned within the given period after the purchase. Since typically, 10000-25000 item codes are used in the supermarket, a huge challenge is to create the appropriate, preferably automatic (or at least almost fully automatic) IT system which is able to go through with the above-mentioned detections and which takes into consideration seasonal goods, replacements, as well as disappearing and appearing item codes in the sample. Finally, one of new challenges connected with scanner data is the choice of the index formula which should be able to reduce the chain drift bias and the substitution bias.

1 Price index formulas

In the traditional data collection, if expenditure information is not available, the European Commission recommends the unweighted Jevons price index. In the case of scanner data, the expenditure information is available but most of countries still use this formula for the scanner data sets. Nevertheless, many countries, having data from retail chains, experiment with weighted bilateral or multilateral price indices (de Haan, 2015; Chessa, 2016, Białek and Bobel, 2019). Superlative price indices, firstly proposed by Diewert (1976), are the most recommended index formulas for the scanner data case among bilateral, weighted price index formulas. In this case, NSIs as a rule decide on the Törnqvist or the Fisher price index or their chain versions. The number of possible price index formulas is reaaly big (CPI Manual, 2004) but in the paper we focus on the most popular, Fisher price index, which can be written as

$$P_F^{0,t} = \sqrt{P_{La}^{0,t} \cdot P_{Pa}^{0,t}} \qquad (1)$$

where $P_{La}^{0,t}$ and $P_{Pa}^{0,t}$ denote the Laspeyres price index and the Paasche price index respectively (CPI Manual, 2004). On the other hand, some NSIs started experiments with multilateral price index formulas which seem to be the most appropriate for the scanner data case. Multilateral index methods have their genesis in comparisons of price levels across countries or regions. These methods satisfy the *transitivity*, which is a desirable property for spatial comparisons due to the fact that the results are independent of the choice of base country (region). Commonly known methods are the GEKS method (Gini, 1931; Eltetö and Köves, 1964; Szulc, 1964), the Geary-Khamis (GK) method (Geary, 1958; Khamis, 1972), the CCDI method (Caves, Christensen and Diewert, 1982), Inklaar and Diewert (2016)) or the *real time* index method (Chessa (2015)).

In the paper we use the GEKS index to demonstrate the chain drift effect. Let us consider a time interval $[0,T]$of observations of prices and quantities which will be used for the GEKS index construction. The GEKS price index between months 0 and t is an unweighted geometric mean of $T+1$ ratios of bilateral price indices $P^{\tau,t}$ and $P^{\tau,0}$ which are based on the same price index formula. The bilateral price index formula should satisfy the time reversal test, i.e. it should satisfy the condition $P^{a,b} \cdot P^{b,a} = 1$. Typically, the GEKS method uses the superlative Fisher price index and in such case the GEKS formula can be written as follows

$$P_{GEKS}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{P_F^{\tau,t}}{P_F^{\tau,0}}\right)^{\frac{1}{T+1}}. \qquad (2)$$

2 Chain drift bias

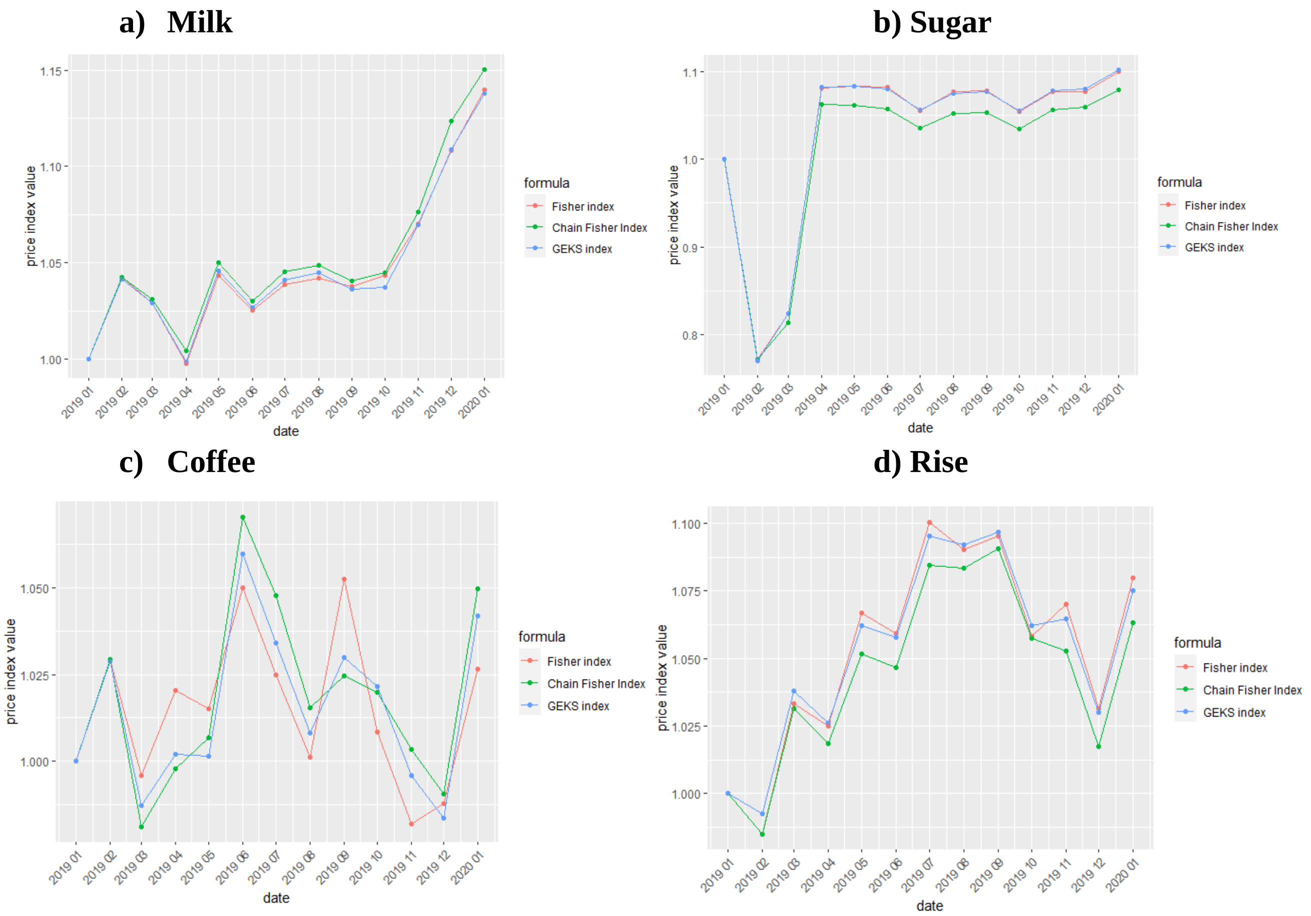
Chain drift occurs when an index does not return to unity when prices in the current period return to their levels in the base period. For instance, Szulc (1983) demonstrated how big the chain problem could be with chained Laspeyres indices but also, as it is commonly known, chain drift can also be a problem with chained superlative indices. Some authors consider the chain drift problem more narrowly, i.e. they assume that only when both prices and quantities in the current period revert back to their levels in the base period, a corresponding price index should indicate that no price change occurred (von Auer (2019)). Potentially, multilateral methods should deal with the chain problem in this “narrow” sense. Nevertheless, even multilateral indices may not return to unity when prices revert back to the levels in the base period but quantities do not.

3 Empirical illustration

As it was above-mentioned, the multilateral GEKS index is free from the chain drift in the narrow sense. Let us determine for a moment, that this formula is a benchmark in our empirical study. In the empirical illustration we use scanner data from one of retail chains in Poland, i.e.

monthly data from over 200 outlets on *milk*, *sugar*, *coffee* and *rise* sold during the period: Jan, 2019 – Jan, 2020. The collected data are classified automatically to rigtht COICOP 5 product groups using original text-mining procedure in R software, i.e the procedure is based on product labels. Products are matched in time by using *Reclin* R package and before index calculation the *low sale filter* and the *extreme price* filters are used (Van Loon and Roels, 2018), i.e. the most extreme 20% product prices are rulled out. Results for the GEKS method with a 13-month time window vs the Fisher price index and its chain version are presented on Fig. 1.

Fig. 1: Comparison of the GEKS, Fisher and chain Fisher indices for 4 product groups



a) Milk

b) Sugar

c) Coffee

d) Rise

Source: own calculations in the original *PriceIndices* R package

As one can see the direct (bilateral) Fisher index and the chain Fisher index may generate the chain drift bias. To be more precise: the differences between the multilateral GEKS formula calculated for the whole time window (thus being free from the chain drift problem in the narrow sense) and the rest of indices may exceed 2-3 percentage points (see Fig. 1b or Fig. 1c). The natural question arises whether and when the splicing GEKS index (see Section 4) can also generate chain drift bias in the considered sense. Section 5 of the paper focuses on that question.

4 Window updating methods

In the case of bilateral methods, a fixed base month (period) is used and the current period is shifted each month. In monthly chained index methods, the base and the current month are both moved one month. The problem with proceeding with the next month arises in the case of multilateral index methods. Adding information from a new month may influence the values of quality adjustment parameters and values of the corresponding multilateral indices. In the literature we can meet the following window updating methods (or splicing methods): a) *movement splice method* (MS), where a price index for the new month is calculated by chaining the month-on-month index for the last month of the shifted window to the index of the previous month; b) *window splice method* (WS), which calculates the price index for the new month by chaining the indices of the shifted window to the index of T months ago. c) *half splice method* (HS), where the splicing period is chosen to be in the middle of the previous time window, d) *mean splice method* (GMS), which uses the geometric mean of all possible choices of splicing, i.e. all months $\{1,2,...,T\}$ which are included in the current window and the previous one.

5 Simulations study

In our experiment we consider 10 products observed during 25 months. We take into consideration the situation when both prices and quantities return to their levels in the base period. We generate price and quantity processes by using p_k^t and q_k^t functions built in *Wolfram Mathematica* software, where $k \in \{1,2,...,10\}$ and $t \in \{1,2,...,25\}$. We focus on the following two scenarios (a parameter Δ reflects the delay in consumers reaction on price changes):

- **Case 1**, where prices and quantities are negatively correlated (price function is concave, quantity function is convex) and the reaction of consumers is delayed, i.e.

$$p_k^t = \begin{cases} 150 + kt : t \leq 12 \\ 150 + k(24 - t) : t > 12 \end{cases} \tag{3}$$

$$q_k^t = \begin{cases} 100 - kt : t \leq 12 + \Delta \\ 100 - k(12 + \Delta) + k\dfrac{12 + \Delta}{12 - \Delta}(t - 12 - \Delta) : t > 12 \end{cases} \tag{4}$$

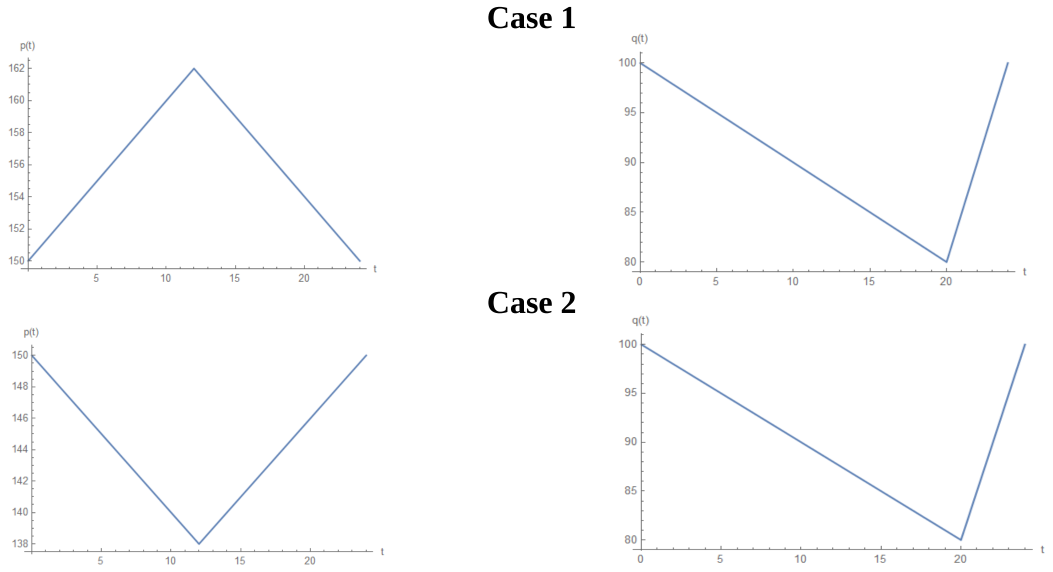
- **Case 2**, where prices and quantities are positively correlated (price function is convex, quantity function is convex) and the reaction of consumers is delayed, i.e.

$$p_k^t = \begin{cases} 150 - kt : t \leq 12 \\ 150 - k(24 - t) : t > 12 \end{cases} \tag{5}$$

and quantities q_k^t are defined as in (4). Please note that Δ is expressed in months.

Figure 2 presents the sample realization of price and quantity processes for $\Delta = 8$ obtained for considered cases. Please note that $p_k^0 = p_k^{24}$ and $q_k^0 = q_k^{24}$ for any $k \in \{1,2,...,10\}$. Figures 3 and 4 present the differences between the GEKS index calculated for the whole available 24-month time window (GEKS FULL) and splicing GEKS indices calculated for a 13-month time window (for $\Delta = 8$). The full-window GEKS plays a role of benchmark.

Fig. 2: Sample realization of price and quantity processes for $\Delta = 8$ (Case1 and Case 2)

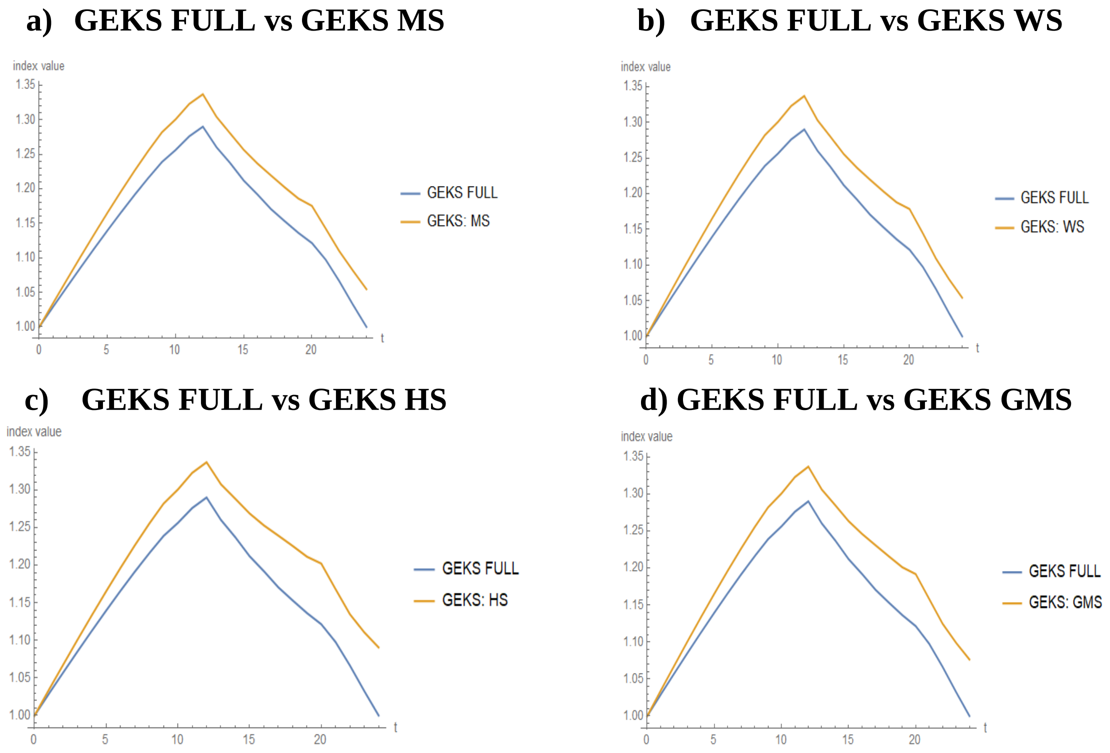


Case 1

Case 2

Source: own calculations in the Mathematica 11

Fig. 3: Differences between the full-window GEKS and splicing GEKS indices (Case 1)



a) GEKS FULL vs GEKS MS

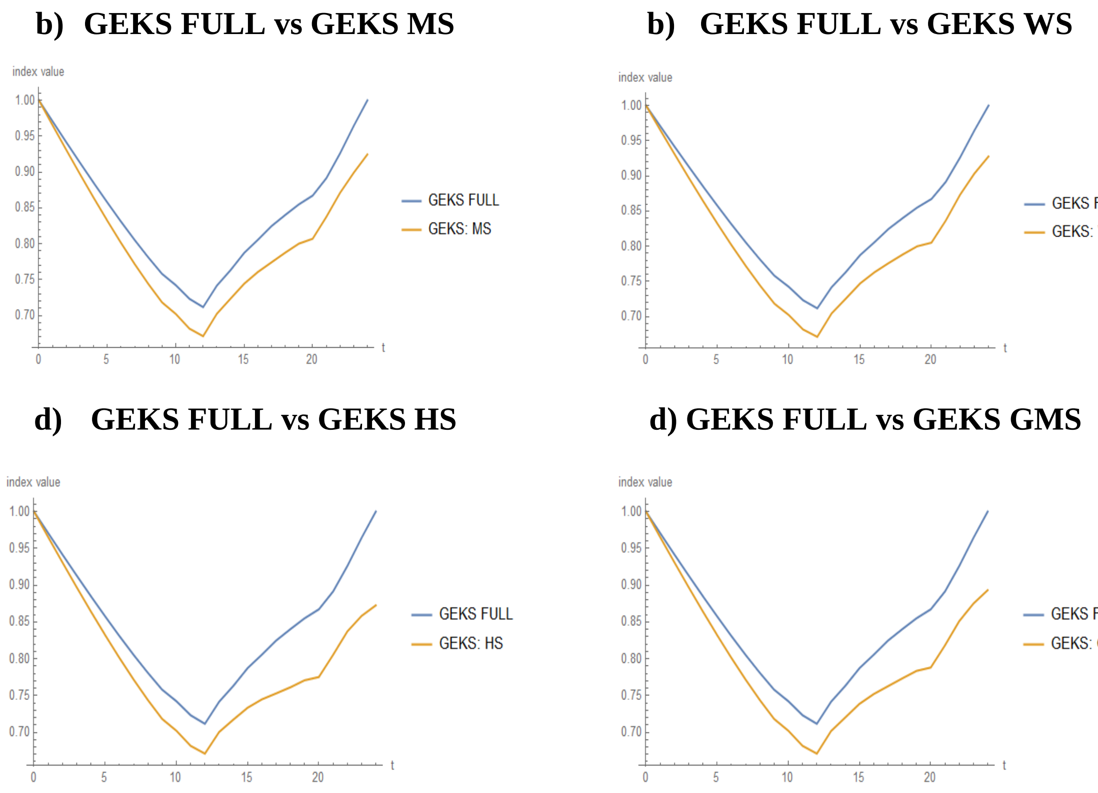
b) GEKS FULL vs GEKS WS

c) GEKS FULL vs GEKS HS

d) GEKS FULL vs GEKS GMS

Source: own calculations in the Mathematica 11

Fig. 4: Differences between the full-window GEKS and splicing GEKS indices (Case 2)



b) GEKS FULL vs GEKS MS

b) GEKS FULL vs GEKS WS

d) GEKS FULL vs GEKS HS

d) GEKS FULL vs GEKS GMS

Source: own calculations in the Mathematica 11

Conclusions

In two simple experiments (see Case 1 and Case 2 in Section 5) we show that although the full-time GEKS index (and any multilateral price index) is free from the chain drift problem, its splicing extensions may lead to the chain drift bias. This fact is commonly known, nevertheless the nature of the chain drift is still unrecognized. The general conclusion from our experiments if the fact that the substantial chain drift bias may occur when the reaction of consumers on price changes is delayed. In fact, there is always some delay in consumers reaction in reality. To show the differences between considered price indices, the value of the parameters Δ is set to 8, but substantial price index differences will be also observed for smaller values of this parameter, e.g. when $\Delta = 3$. It is quite interesting that the correlation between prices and quantities does matter in the chain drift problem. The most typical situation on the market is when the correlation between prices and quantities is negative and it is considered in Case 1 in the Simulation Study. In the above-mentioned Case 1, each splicing GEKS index seems to overestimate the real price change (see Fig. 3). When the analogous correlation is positive (Case 2), the splicing methods lead to underestimation of the real price dynamics (see Fig. 4).

Acknowledgment

The author would like to thank the National Science Centre in Poland for financing this publication (grant no. 2017/25/B/HS4/00387).

References

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.

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.

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.

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.

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

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.

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

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.

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

Szulc, B. (1983). Linking Price Index Numbers. In: *Price Level Measurement*, W. E. Diewert and C. Montmarquette (eds.), 537 – 566.

Van Loon, K., Roels, D. (2018). *Integrating big data in the Belgian CPI*. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 7-9 May, Geneva, Switzerland.

Von Auer L. (2019). *The Nature of Chain Drift*. Paper presented at the 17th Meeting of the Ottawa Group on Price Indices, 8 – 10 May, Rio de Janerio, Brasil.

Contact

Jacek Białek
University of Lodz, Department of Statistical Methods
Rewolucji 1905 r. No. 41/43, 90-214 Łódź, Poland
jacek.bialek@uni.lodz.pl
Statistics Poland, Department of Trade and Services
Aleja Niepodległości 208, 00-925 Warsaw, Poland
J.Bialek@stat.gov.pl

Elżbieta Roszko-Wójtowicz, Department of Social and Economic Statistics
Rewolucji 1905 r. No. 41/43, 90-214 Łódź, Poland
eroszko33@gmail.com