ARTICLES

CENTRAL EUROPEAN REVIEW OF ECONOMICS & FINANCE
Vol. 44. No 3 (2023) pp. 5-20
DOI https://doi.org/10.24136/ceref.2023.011

Jacek Białek[1]

SCANNER DATA AND THE PROBLEM OF SELECTING A PRICE INDEX FORMULA

Abstract
Scanner data are electronic transaction data most often from retail chains and obtained from electronic retail terminals. The identification of products takes place after scanning their characteristic barcode (e.g. EAN or GTIN), thus in the case of scanner data, we have full product information (price, sales volume, weight, description, etc.) at the most disaggregated level. In the cases of many countries, as well as Poland, this type of data is a valuable alternative source of information when estimating inflation. This paper discusses the main advantages but also the challenges of using scanner data in the CPI measurement. The main purpose of the paper, however,is to discuss the problem of selecting an optimal price index formula that would be appropriate for the highly dynamic (in terms of product rotation) scanner data. The considerations, supported by examples of empirical studies, will be demonstrated using the *PriceIndices* package in the R environment.

Keywords: scanner data, Consumer Price Index, bilateral indices, multilateral indices.

JEL Classification: C43

[1] Dr hab., associate professor of the University of Lodz, Department of Statistical Methods.

Introduction

The following definition of scanner data can be found in the literature: "Scanner data mean transaction data that specify turnover and numbers of items sold by barcodes, e.g. GTIN, formerly known as the EAN code (International Labour Office, 2004)."

These data can be obtained from a wide variety of retailers (supermarkets, home electronics, Internet shops, etc.). 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 at the barcode level, i.e. information about prices and quantities. As a rule, scanner data sets have huge volume and may provide some additional information about products (such as the following attributes: size, grammage, sale unit, colour, package quantity, etc.). These attributes may be useful in aggregating items into the homogeneous groups or when product matching over time (Białek and Beręsewicz, 2021). The use of scanner data in the assessment of inflation leads to the improvement of the data collection process, its costs reduction and a better reflection of changes that occur in consumer behaviour. The form of the sample scanner dataset is presented in Tab.1.

Table 1. Sample scanner data frame obtained from one of retail chain in Poland

Time	Prices	Quantities	Retid	EAN	COICOP	Label	Grammage	Unit	Prodid
2022-12-31	10.47	8.48	26-617	590674717126	011111	long grain rice	0.4	kg	1
2022-12-31	12.47	5.87	40-772	590674717126	011111	long grain rice	0.4	kg	1
2022-12-31	11.40	15.65	70-001	590674717126	011111	long grain rice	0.4	kg	1
2022-12-31	13.20	16.95	85-791	590674717126	011111	long grain rice	0.4	kg	1
2022-12-31	11.47	85.41	01-460	590674717126	011111	long grain rice	0.4	kg	1
2022-12-31	11.97	7.82	05-820	590674717126	011111	long grain rice	0.4	kg	1

Source: Białek et. al. (2022), p. 71.

This paper, however, focuses not on the advantages but on the challenges that accompany the implementation of scanner data for the CPI measurement. One of the major challenges facing statistical offices in this case is the choice of the price index formula (Chessa, 2015). The purpose of this article is to point out potential problems related to this aspect and also to demonstrate possible differences in measuring the dynamics of scanner prices that may arise when using different price indexes.

1. Challenges when using scanner data

Using scanner data in the context of CPI measurement poses a number of challenges, both technological and methodological. First, the processing of scanner data generates an IT challenge, as it is a huge volume of data that needs to be automatically cleaned, classified into appropriate product segment groups (COICOP) and then matched over time (Białek and Beręsewicz, 2021).

From a methodological point of view, the challenge in turn is the appropriate filtering of the data. This stage requires selection of the type of data filter and its thresholds, e.g. *extreme price filter* or *low sales filter* can be applied – see van Loon and Roels (2018). Sometimes the statistical office is forced to build a completely new IT environment to handle the processes mentioned earlier. However, some statistical offices choose to implement functioning packages or programs dedicated to scanner data and price indexes. The R packages such as the *IndexNumR* or *PriceIndices* package (Białek, 2022a) should be mentioned here, the latter of which has been implemented at Statistics Poland (Białek, et, al. 2022).

From a methodological perspective, scanner data also provides many opportunities but also challenges. Due to the highly detailed nature of this data, opportunities open up for statisticians to accurately model probability distributions of product prices (Białek and Sulewski, 2022) and thus study the stochastic properties of price indices (Silver and Heravi, 2007; Białek, 2020, 2022b). Nevertheless, the high turnover of scanner products (so-called *product churn*) makes the choice of index formula not at all an obvious choice. The choice of a price index for estimating inflation on the basis of the scanner data should take into account the *weak* and *strong seasonality* of products (CPI Manual, 2004), and should eliminate the measurement bias resulting from the *substitution effect* of goods and *chain drift*. This thread will be developed in the next Section, which is devoted to the selection of the price index formula.

2. Scanner data and price index selection

In the case of traditional data collection, where interviewers collect information about prices from the field and the consumption level is evaluated via household budget surveys, statistical agencies use bilateral index numbers (von der Lippe, 2007; Białek and Roszko-Wójtowicz, 2021). In practice, at the lowest level of data aggregation, where only prices are available, the unweighted Jevons (1865) price index formula is used to calculate price indexes, which is due to the good axiomatic properties of this formula and also to the fact that it is anchored within the so-called *stochastic approach* (von der Lippe, 2007). At higher levels of data aggregation, where the statistical office has knowledge of the level of consumption of specific product groups, the Laspeyres-type formula (1871) is most often used. For a set of prices and quantities of goods observed in the base (0) and current (t) period, the Jevons formula can be expressed as:

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

and the Laspeyres price index can be written as

$$P_{La}^{0,t} = \frac{\sum_{i=1}^{N_{0,t}} q_i^0 p_i^t}{\sum_{i=1}^{N_{0,t}} q_i^0 p_i^0},$$

where $N_{0,t}$ denotes number of products available in the periods 0 and t, p_i^τ means a price of the $i-th$ product observed at the time τ, q_i^τ means a quantity of the $i-th$ product observed at the time $\tau \in \{0,t\}$. The use of the Laspeyres index at higher levels of data aggregation and ultimately at the COICOP 2 level is dictated by the lag with the household budget survey providing information on the level of consumption of goods and services.

However, in the case of scanner data, there is no contraindication to using weighted price indexes that also use current period consumption data. Scanner data are complete already at the lowest levels of aggregation and we have information on both prices and quantities of products for the selected moment of transaction. Thus, the "ideal" Fisher (1922) index seems to be the best choice from the perspective of the axiomatic and economic approach in the index theory. The Fisher index is a geometric mean of the Laspeyres and Paasche (1874) indices, i.e.:

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

where the Paasche price index is as follows:

$$P_{Pa}^{0,t} = \frac{\sum_{i=1}^{N_{0,t}} q_i^t p_i^t}{\sum_{i=1}^{N_{0,t}} q_i^t p_i^0}.$$

However, some problems arise with the use of the bilateral indexes. The Jevons index does not take full advantage of the information because it does not take into account knowledge of product consumption. On the other hand, the use of weighted bilateral indexes does not take into account intermediate periods between the base period and the current period,

i.e. it can generate measurement bias due to the high turnover of scanner products. Unfortunately, even the use of chain weighted indexes (such as the chain Fisher index) does not guarantee an unbiased measurement. It can be shown (Chessa, 2015) that frequently chained weighted indices lead to the *chain drift* bias. The *chain drift* can be formalised in terms of the violation of the *multi period identity test* (Białek, 2022c). The above-mentioned test states, that when all prices and quantities in the current period return to their values from the base period, then the index should equal one.

The most promising group of indexes in the context of scanner data appear to be multilateral indexes. The multilateral price index is calculated for a given time window consisting of $T+1$ consecutive months, which we number $0,1,2,...,T$ (typically $T=12$). Multilateral indices use all prices and quantities of individual products which are available in a set time window. Multilateral indices are *transitive* (Australian Bureau of Statistics, 2016), which eliminates the chain drift problem. The known and popular multilateral methods are the GEKS method (Gini, 1931; Eltetö and Köves, 1964), the Geary-Khamis (GK) method (Geary, 1958; Khamis, 1972), the CCDI method (Caves et al., 1982) or the Time Product Dummy Methods (de Haan and Krsinich, 2018). For instance, the popular GEKS formula which is based on the Fisher price index, can be expressed as follows:

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

3. Potential differences between price indices while using scanner data

The first aim of our empirical study is indicating potential differences between bilateral indices and full-time window multilateral indices. In the study, scanner data from one retail chain in Poland were used, i.e. monthly data on *ground coffee* (subgroup of COICOP 5 group: 012111) and *white sugar* (subgroup of COICOP 5 group: 011811) sold in 212 outlets during the period from December 2019 to December 2020. Before price index calculations, the database was carefully prepared. First, after deleting missing and duplicated data, the sold products were classified first into the relevant elementary groups and their subgroups (COICOP 6 level). Product classification was done via *data_selecting()* and *data_classification()* functions from the *PriceIndices* R package (Białek, 2022a). The first function required manual preparation of dictionaries of phrases and keywords which are able to identify individual product groups. The *data_classification()* function, which is based on *machine learning* techniques, was used for problematic, previously unclassified products and required manual

preparation of learning data sets. This step of classification was based on random trees and the *XGBoost* algorithm (Tianqi and Carlo, 2016). Next, the product matching was carried out on the basis on the available GTIN bar codes, internal retailer codes, and product labels. To match products over time we run the *data_matching()* function from the *PriceIndices* package. All products with two identical codes or one of the codes identical and an identical label were automatically matched. Products were also matched if they had identical one of the codes and the Jaro-Winkler (1989) distance of their labels (descriptions) was smaller than the fixed precision value: 0.02. In the last step before calculating indices, two data filters were applied to remove unrepresentative products from the data set, i.e. the *data_filtering()* function from the *PriceIndices* package was applied. The *extreme price filter* (Białek and Beręsewicz, 2021) was used to eliminate items with more than three-fold price increase or more than double price drop from period to period. The *low sale filter* (van Loon and Roels, 2018) was run to roll out products with relatively low sales (almost 35% of products were removed).

Fig. 1 presents a comparison of bilateral indices (unweighted and weighted) prepared for these two above-mentioned scanner data sets. Fig. 2 presents a comparison of the selected multilateral index methods (GEKS, Geary-Khamis and TPD indices) with the chain Jevons and chain Fisher indices. As one can see, the bilateral Jevons index clearly lags behind bilateral weighted price indexes, whereby it can overestimate or underestimate the “ideal” Fisher index by as much as more than 10 percentage points (see Fig. 1). The chain Jevons index also differs substantially from the chain Fisher index and multilateral indexes (the difference is as extreme as 23 percentage points for August 2020 for *white sugar*). As it was above-mentioned, the Fisher chain index is subject to the chain drift effect, and it can be seen that its values differ somewhat from those of the multilateral indexes, which are free of the chain drift problem.

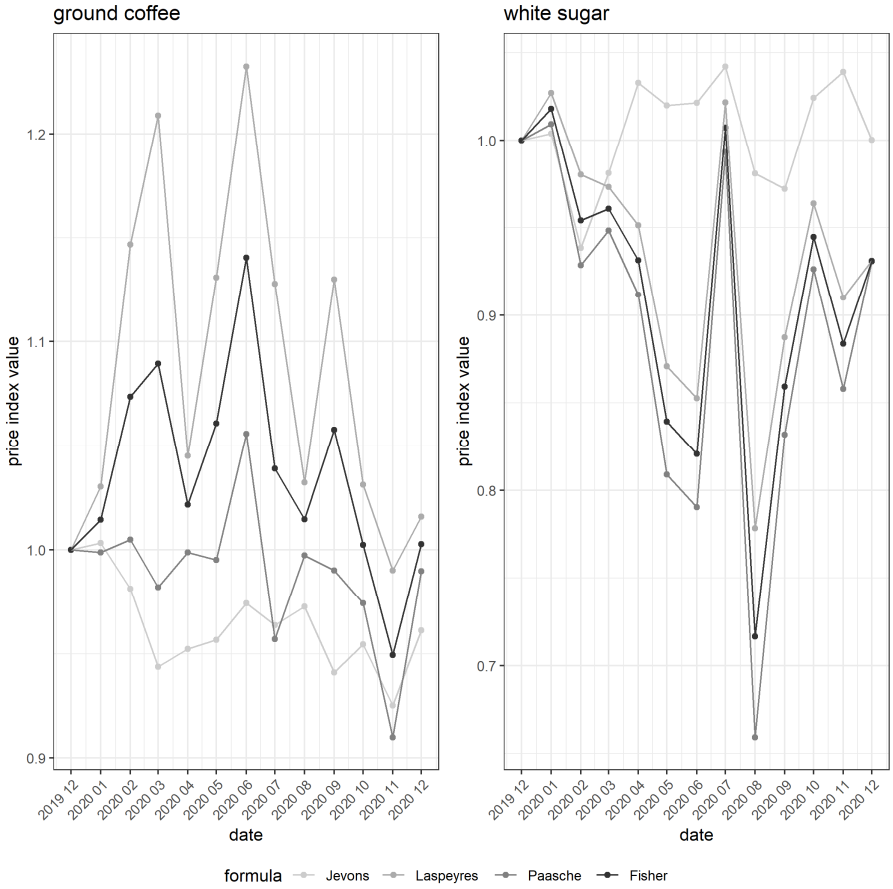

Figure 1. Comparison of bilateral indices for data on ground coffee and white sugar

Source: Own calculations in the *PriceIndices* R package

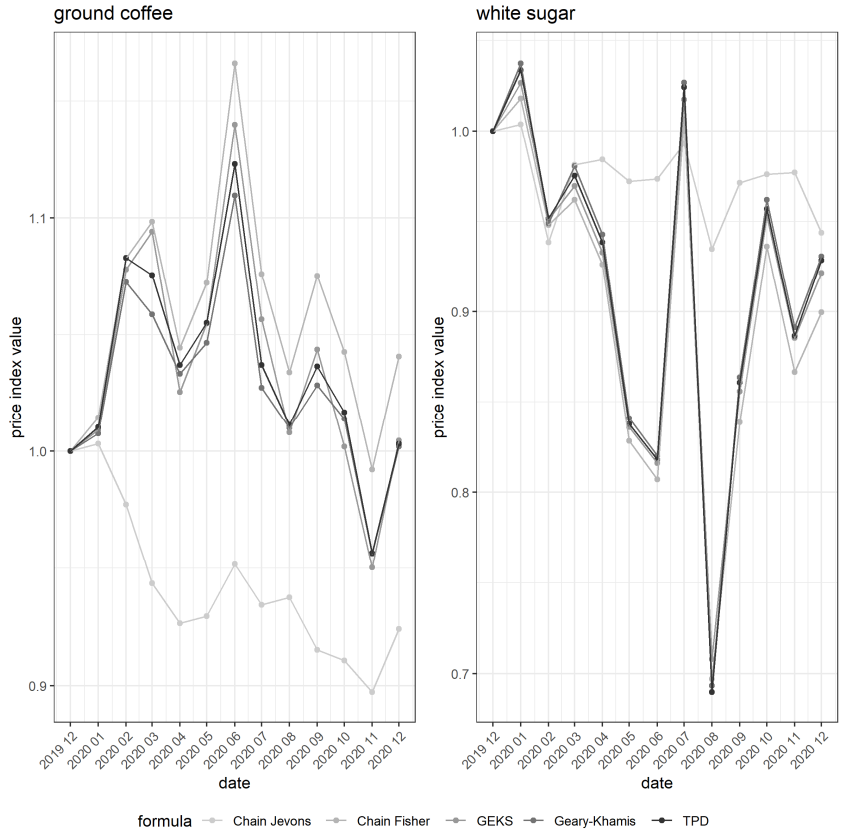

Figure 2. Comparison of the multilateral index methods with the chain Jevons and chain Fisher indices
Source: Own calculations in the *PriceIndices* R package

4. Comparison of price indices due to their time-consuming

As previous work has shown (Bialek and Beręsewicz, 2021), price indexes vary widely due to the timing of the calculations. Thus, the cited work proposes the *time-consuming criterion* for evaluating multilateral indexes applied to the case of scanner data. The choice of an index whose calculation time is relatively small is of practical importance, since scanner data are generally very large data sets and the final calculation time is proportional to the number of outlets of the retail chain (provided the retail chain has a regional pricing policy). Figure 3 presents a comparison of the calculation times of the considered price indexes for the two product groups analysed.

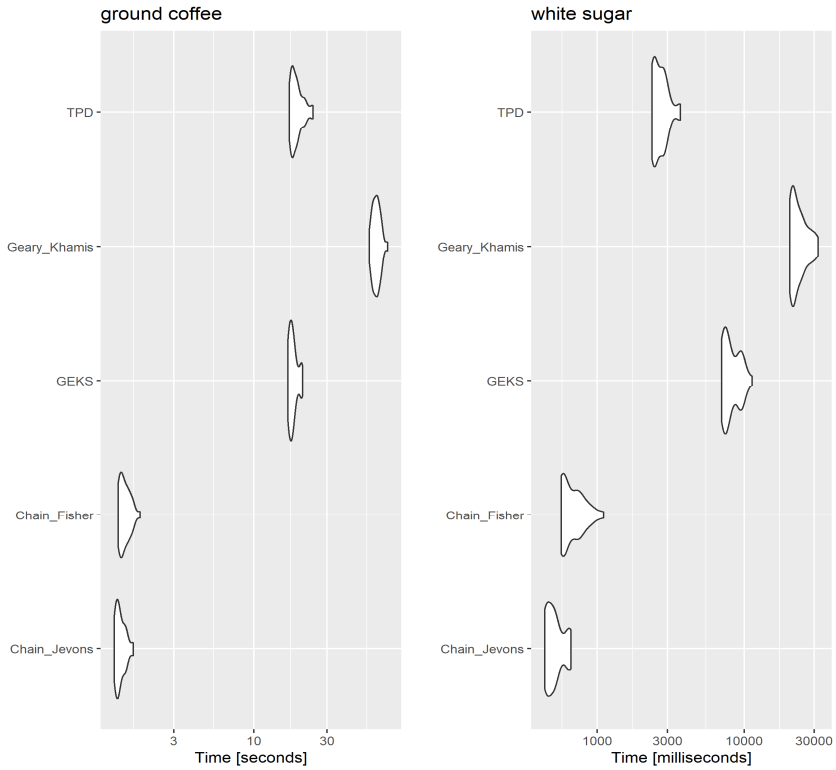

Figure 3. Comparison of the calculation times of selected price indices
Source: Own calculations in the *PriceIndices* R package

As can be seen, the computation of chain indices is relatively faster than the computation of multilateral indices due to the greater complexity of the latter. Among multilateral indexes, the calculation time of the TPD index is relatively attractive as long as the dataset is small (white sugar). For larger datasets (ground coffee), the GEKS index can be calculated the fastest. The Geary-Khamis index, due to its iterative procedure (usually 4-6 iterations are needed), is relatively time-consuming. A similar time-consuming comparison of multilateral indexes which includes additional price indexes, can be found in the paper of Białek (2022d).

5. Problem of aggregation of partial indices over outlets

As it was above-mentioned, the time-consumption of the price index is proportional to the number of outlets of a given retail chain. In other words, if the retail chain has a regional pricing policy, it is necessary to calculate price indices for each outlet separately and then aggregate the partial results into one resultant price index. This arises the natural question of whether the possible aggregation of the results relative to the outlets makes practical

sense, i.e. whether there is any substantial difference between the price index calculated without this aggregation and the index that takes this aggregation into account. For a traditional data collection, the only aggregation formula is the Laspeyres formula, since knowledge of consumption of goods and services is unavailable for the current period. In the case of scanner data, any aggregation formula can be considered. Figures 4, 5 and 6 show the effect of aggregation on the value of the price index while considering different aggregation methods, i.e. the aggregation by using the Laspeyres, Paasche and Fisher formulas. For example, when aggregating over outlets by using the Fisher's formula, the impact of results from an outlet will be proportional to the relative shares of sales revenue from that outlet in the base and current periods. Based on Figures 4-6, it can be concluded that the Jevons chain index is particularly sensitive to the decision to perform sub-score aggregation over outlets. The Fisher chain index and the multilateral GEKS index are marginally sensitive to the choice of aggregation formula and also to the abandonment of aggregation (see Fig. 5 and 6). However, this conclusion requires further research, as it may be due to the very similar pricing policies of the given retail chain across all outlets.

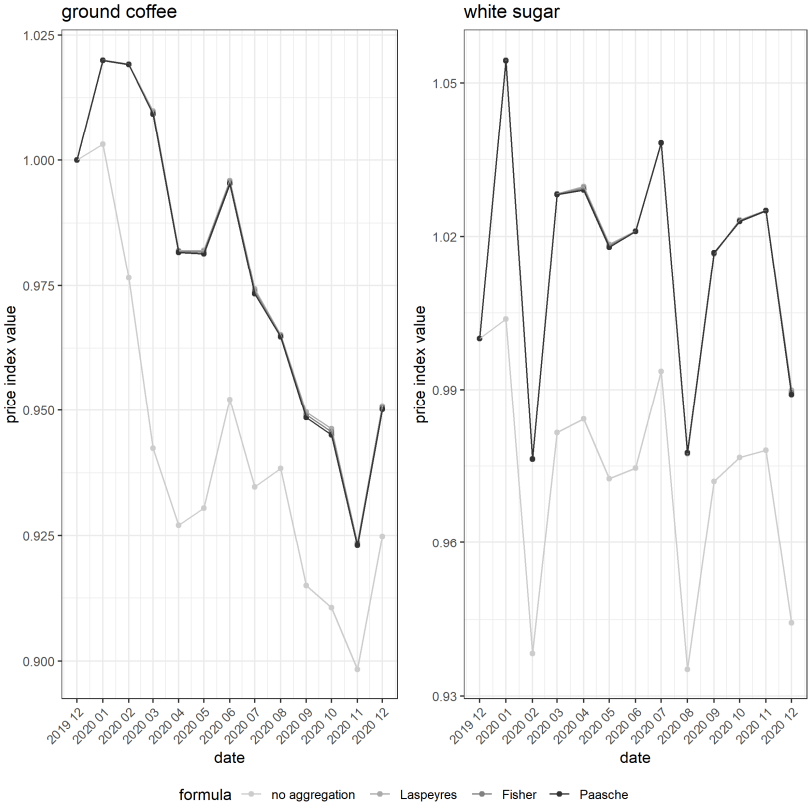

Figure 4. Impact of an aggregation method on the chain Jevons index
Source: Own calculations in the *PriceIndices* R package

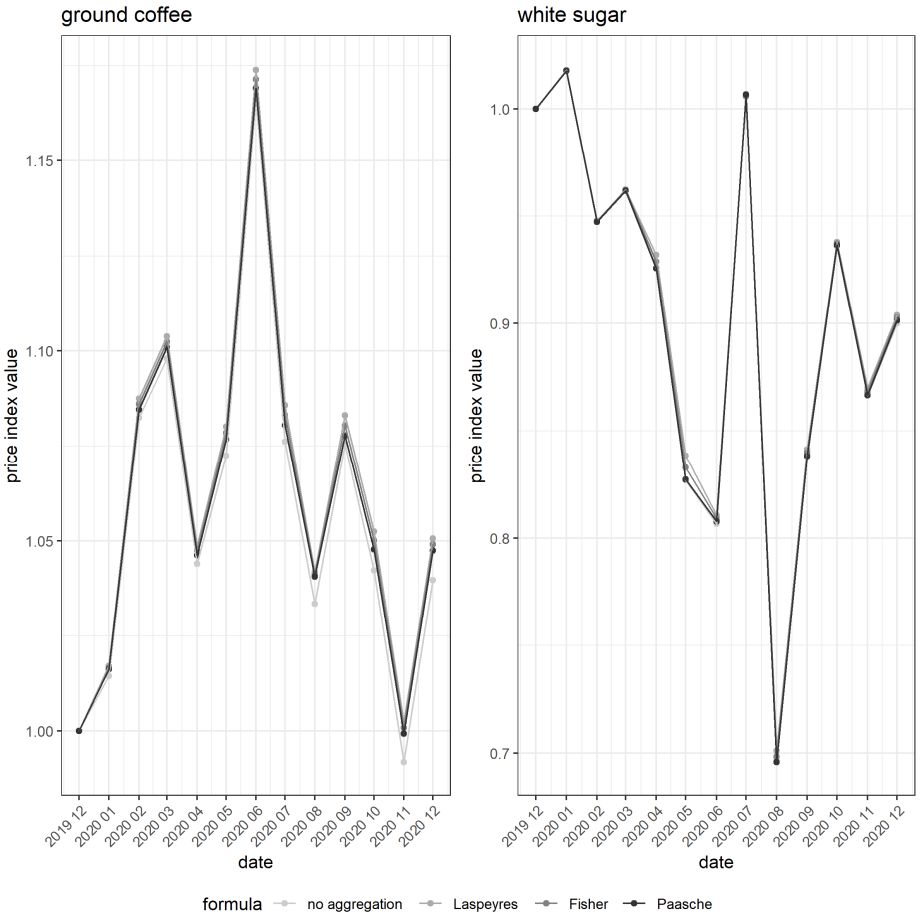

Figure 5. Impact of an aggregation method on the chain Fisher index
Source: Own calculations in the *PriceIndices* R package

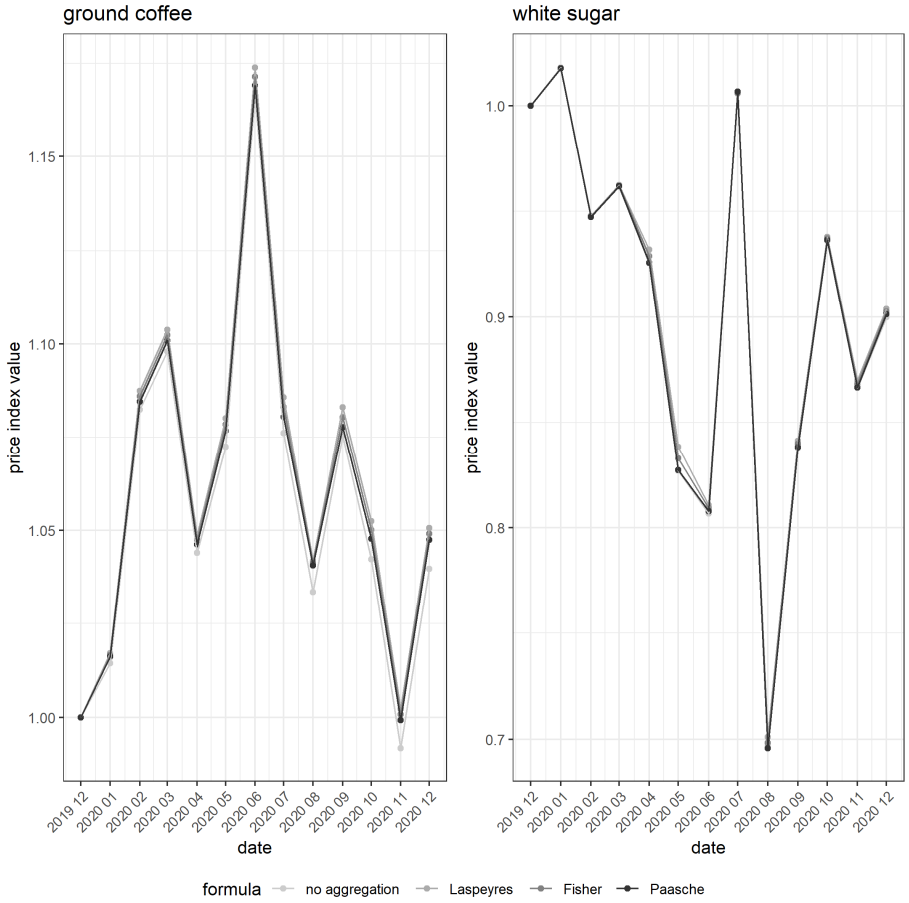

Figure 6. Impact of an aggregation method on the GEKS index
Source: Own calculations in the *PriceIndices* R package

Conclusions

It seems that from a methodological point of view, the problem of selecting the price index for scanner data case is one of the key problems. This is due to the fact that, firstly, the value of measuring the dynamics of scanner prices strongly depends on the choice of the index itself, and secondly, there are many criteria on the basis of which the choice of an optimal price index can be made (for example, the axiomatic criterion, the economic criterion, the stochastic criterion or the time-consuming criterion). However, it seems that the main conclusion from the study is that there is no recommendation for the chain Jevons index, which, as an unweighted formula, completely fails to reflect the differentiation of products by scale of their sales.

References

1. Australian Bureau of Statistics (2016), *Making Greater Use of Transactions Data to Compile the Consumer Price Index*.
2. Białek J., (2020), Basic Statistics of Jevons and Carli Indices under the GBM Price Model*, Journal of Official Statistics*, 36 (4), 737-761.
3. Białek J., Beręsewicz M., (2021), Scanner data in inflation measurement: from raw data to price indices, *The Statistical Journal of the IAOS*, 37, 1315-1336.
4. Białek J., (2022a), Scanner data processing in a newest version of the PriceIndices package, *Statistical Journal of the IAOS*, 38 (4), 1369-1397.
5. Białek J., (2022b), Elementary price indices under the GBM price model*, Communications in Statistics – Theory and Methods,* 51(5), 1232-1251.
6. Białek J., (2022c), The general class of multilateral indices and its two special cases, *Paper presented at the 17th Meeting of the Ottawa Group on Price Indices*, Rome, Italy.
7. Białek J., (2022d), Improving quality of the scanner CPI: proposition of new multilateral methods, *Quality and Quantity,* https://doi.org/10.1007/s11135-022-01506-6.
8. Białek J., Roszko-Wójtowicz E., (2021), Dynamics of price level changes in the Visegrad group: comparative study, *Quality and Quantity,* 55, 357-384.
9. Białek J., Panek T., Kłopotek M. (ed), (2022), *Nowoczesne technologie i nowe źródła danych w pomiarze inflacji*, GUS, Warsaw.
10. Białek J., Sulewski P., (2022), Probability Distribution Modelling of Scanner Prices and Relative Prices*, Statistika – Statistics and Economy Journal*, 3/2022, 282-298, Czech Statistical Office, Prague.
11. 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.
12. Chessa A., (2015), Towards a generic price index method for scanner data in the Dutch CPI. *In: 14th Meeting of the Ottawa Group*, Tokyo, 20-22.
13. de Haan J., Krsinich F., (2018), Time dummy hedonic and quality-adjusted unit value indexes: Do they really differ? *Review of Income and Wealth,* 64(4), 757-776.
14. Eltetö O., Köves P., (1964), On a problem of index number computation relating to international comparison, *Statisztikai Szemle,* 42(10), 507-518.
15. Fisher I., (1922), *The making of index numbers: a study of their varieties, tests, and reliability. Number 1*, Houghton Mifflin.
16. Geary R. C., (1958), A note on the comparison of exchange rates and purchasing power between countries, *Journal of the Royal Statistical Society. Series A (General)*, 121(1), 97-99.
17. Gini C., (1931), On the circular test of index numbers, *Metron,* 9(9), 3-24.

18. International Labour Office (2004), *Consumer Price Index Manual: Theory and Practice*, Geneva.
19. Jevons W. S., (1865), On the variation of prices and the value of the currency since 1782, *Journal of the Statistical Society of London,* 28(2), 294-320.
20. 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.
21. Laspeyres K., (1871), Ix. die berechnung einer mittleren waarenpreissteigerung, *Jahrbücher für Nationalökonomie und Statistik,* 16(1), 296-318.
22. Paasche H., (1874), Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen, *Jahrbücher für Nationalökonomie und Statistik*, 23, 168-178.
23. Silver H., Heravi S., (2007), Why elementary price index number formulas differ: Evidence on price dispersion, *Journal of Econometrics*, 140 (2007), 874-883.
24. Tianqi C., Carlo G., (2016), Xgboost: A scalable tree boosting system, *In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 785- 794.
25. Van Loon K.V., Roels D., (2018), Integrating big data in the Belgian CPI, *In: Paper Presented at the Meeting of the Group of Experts on Consumer Price Indices*, 8-9 May 2018, Geneva, Switzerland.
26. Von der Lippe P., (2007), *Index Theory and Price Statistics*, Peter Lang, Germany.
27. Winkler W., (1990), String comparator metrics and enhanced decision rules in the fellegisunter model of record linkage, *In Proceedings of the Section on Survey Research Methods. American Statistical Association*, 354-35.