Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

https://www.czasopisma.uni.lodz.pl/foe/
1(374) 2026
DOI: https://doi.org/10.18778/0208-6018.374.02

Natalia Pawelec https://orcid.org/0009-0008-5289-6204
University of Lodz, Faculty of Economics and Sociology, Lodz, Poland, natalia.pawelec@uni.lodz.pl

The Use of Scanner Data for Price Indices Calculation: the Case of Poland

Abstract: In the context of the growing use of scanner data in inflation measurement, a key methodological challenge remains the choice of the appropriate price index formula and data filtering method. This article aims to investigate the impact of these choices on the results of price dynamics measurement. The paper analyses transactional data collected from point-of-sale terminals in Polish supermarkets for selected product groups: coffee and milk. Various price index formulas – including unweighted bilateral (Dutot, Carli, Jevons), weighted bilateral (Törnqvist, Fisher), and multilateral (Geary-Khamis, CCDI) – were compared, and the influence of different scanner data filtering methods was tested. The results show that both the choice of the price index formula and the applied scanner data filtering method have a measurable and significant impact on the assessment of price dynamics. These differences can lead to diverse interpretations of inflation, which underscores the need for conscious and justified methodological choices when using scanner data in price statistics.

Keywords: scanner data, multilateral indices, bilateral indices

JEL: C43

Funding information: University of Lodz, Faculty of Economics and Sociology, Lodz, Poland.
Declaration regarding the use of GAI tools: Not used.
Conflicts of interests: None.
Ethical considerations: The Author assures of no violations of publication ethics and takes full responsibility for the content of the publication.
Received: 2025-02-25. Revised: 2025-11-10. Accepted: 2026-02-25

1. Introduction

The Consumer Price Index (CPI) should reflect actual inflation as accurately as possible for various reasons. One of them is the fact that it is used to index nominal values in the economy, influencing social policy and pricing decisions by business owners. One of the methodological challenges of consumer price research is the necessity to diversify data sources and automate data processing. The increasing availability of electronic transaction datasets for the Consumer Price Index (CPI) presents a significant opportunity for national statistical institutes (NSIs) to elevate the precision of their index numbers. The data facilitate the deployment of more advanced analytical techniques that can account for the fluid nature of consumption habits far more effectively than conventional fixed-basket approaches (Chessa, 2021). One solution is acquiring and processing scanner data. We define scanner data as 'transaction data obtained from retail chains containing data on turnover, quantities per item code based on transactions for a given period and from which unit value prices can be derived at item code level' (Eurostat, 2017).

Scanner data are generated by point-of-sales terminals in shops which log each transaction. Scanner data can be sourced from numerous types of retailers, including supermarkets, pharmacies, hardware stores, electronics shops, clothing stores, and many others. Presently, scanner data are mainly used to substitute price collection in supermarkets, especially for items such as food, beverages, and the personal and household care products available there. As previously mentioned, this guide will focus exclusively on these product categories. Scanner data have several advantages over traditional price collection. They are relatively inexpensive, with their collection and processing being automated. They include information on the actual expenditure for all item codes sold. Integrating scanner data into inflation measurement enhances the data collection process, reduces associated costs, and provides a more accurate representation of shifts in consumer behaviour.

The main objective of this study was to investigate how the choice of a price index formula impacts the final inflation indicator values. Specifically, the aim was to examine the differences that may occur between the values of bilateral and multilateral indices. The analysis was conducted using scanner data from the Polish retail market, covering the period from December 2019 to December 2020. The study focuses on comparing several types of price indices: unweighted bilateral indices (Dutot, Carli, Jevons), weighted bilateral indices (Törnqvist, Fisher), and multilateral indices (Geary-Khamis, CCDI). According to Eurostat (2020b), multilateral indices are the most strongly recommended approach for scanner data analysis, as they better account for product turnover and mitigate chain drift.

The results clearly indicate that the selection of a specific price index formula significantly influences the values of the inflation indicator, which leads to varied interpretations of price dynamics. The differences in the calculation results are presented and discussed in detail, allowing for an assessment of the methodological consequences within the Polish context.

2. Measuring Inflation

2.1. Indices for Measuring Inflation

The CPI is a measure of inflation that is closely observed by consumers, businesses, and policymakers as price movements in the economy reflect changes in the value of money over time. In most countries, inflation measurement is conducted by statistical offices, although central banks (e.g., the National Bank of Poland) are also interested in this measurement. In Poland, Statistics Poland makes calculations based on representative goods and services, including typical products or services most commonly purchased by households (Statistics Poland, 2019).

In official statistics, two price indices are used to measure inflation. The first is the CPI, which represents the inflation rate in a given country. It approximates changes in household consumption costs to maintain a constant utility level (Hałka, Leszczyńska, 2011) and serves as a proxy for the Cost of Living Index (COLI). The CPI measures changes in the prices of goods and services purchased by households for consumption and informs monetary policy, social protection benefits, financial instruments, and the indexing of commercial contracts, wages, pensions, and retirement benefits (Eurostat, 2024).

The second indicator is the Harmonised Index of Consumer Prices (HICP), created by the European Central Bank to ensure a high-quality comparison of consumer price inflation between countries. The HICP aims to deliver a reliable and consistent metric for comparing consumer price inflation. It employs different formulas to aggregate prices and price indices depending on whether products are weighted or not. The HICP is a cost of goods (and services) index, i.e., it measures the changing costs of a fixed basket of products at different sets of prices over time. Prices used to measure the HICP should reflect those actually paid by households and must include costs related to reservations and deliveries often associated with online purchases (Eurostat, 2024).

The CPI is measured monthly. To accurately reflect the current consumption patterns of households, the data must be sourced from various sources. According to Statistics Poland: 'Price data are sourced from price lists, regulations, and decisions regarding uniform prices applicable nationwide or in parts thereof, issued by government and local administration bodies, and economic entities (e.g., prices of selected pharmaceutical products, notarial services, or banking services). Another method of data collection is searching for appropriate data on websites and virtual domains of stores. This method is used to gather data on, among other things, airline ticket fees' (Statistics Poland, 2019).

Statistics Poland collects over 230,000 prices monthly, with the most significant number (72,000) being single prices for the 'food and non-alcoholic beverages' category. Statistics Poland also gathers information on the characteristics of selected products, such as raw material composition, model, and weight. To measure inflation, it is necessary to have information on price notations for identical goods and services in two compared periods and on the expenditure structure of households (Statistics Poland, 2019). The number of price notations collected is increasing yearly due to the growing diversity of goods and services offered, as well as new ways to acquire such information, such as using scanner data.

The weights necessary to calculate the CPI are obtained by grouping expenditures on goods and services collected from household budget surveys. The 'inflation basket' structure considers typical goods and services most commonly purchased by households.

Data used to construct the HICP weights also come from various sources. National accounts data (at different levels of aggregation) are the main source because regulations require this. Their undeniable advantage is that they are structured and organised according to the European Classification of Individual Consumption by Purpose. Other data sources include household surveys, market research, retail trade data, and administrative sources. The traditional method of price recording by interviewers has a well-established methodology and a long history. However, technological advancements, particularly in information technology, enable the use of alternative sources of product price data, such as scanner data (which are primarily obtained from supermarket chains) and scraped data (which are automatically collected from websites). The scope of price notations collected from alternative sources has been increasing yearly (Statistics Poland, 2024).

3. Scanner Data as a Modern Data Source in Inflation Measurement

3.1. Advantages, Disadvantages, and Methodological Challenges

The advent of barcode scanner technology in the 1970s gave rise to scanner data, a new form of information quickly recognised for its value in understanding consumer choices and demands. The data have since fuelled extensive research across various economic behaviours, particularly in industrial organisation, marketing, and increasingly in public, health, and macro/monetary economics (Dubois, Griffith, O'Connell, 2022). Scanner data serve as a modern source of transaction data, providing highly detailed and accurate information due to the immense number of observations obtained by scanning consumer goods barcodes. The most common barcodes are GTIN (Global Trade Item Number) and EAN (European Article Number), the European version of GTIN (Eurostat, 2017).

Scanner data are a modern source of information about prices and consumption levels. Specifically, they provide data on prices and quantities of purchased products at the lowest level of data aggregation, namely at the individual barcode level. One of the many advantages of using scanner data is that they contain information at the barcode or GTIN level, enabling products to be 'recognised' and classified into appropriate COICOP groups.

A clear benefit of scanner data is that they provide additional product information and information related to the seller's code, point-of-sale identifier, product label, unit of sale (e.g., 'pcs,' 'kg'), sales value, number of units sold, flag (containing information, e.g., about discounted products), and VAT information. This is useful for aggregating elements into homogeneous groups. Scanner data are collected automatically and relatively cheaply (Białek, 2020a). Moreover, they provide information about actual expenditures for all sold items, excluding unsold products. These data typically cover 2–4 weeks of each month and are sometimes transferred with daily or weekly frequency. All these advantages suggest that using scanner data offers hope for improving the quality of the HICP (Eurostat, 2017).

However, one problem that arises from using scanner data is the overly detailed level of information about the product contained in the GTIN code. The GTIN code comprises 13 or 14 digits, including: 1 digit to indicate the packaging level, a 3-digit GS1 national organisation code (country code), a 4–7-digit GS1 coding unit number, a 2–5-digit product code, and a check digit. This makes it challenging to identify homogeneous product groups using this code because different codes may represent the same product with only packaging variations. In addition to GTIN and EAN barcodes, other barcodes, such as PLU (Price Look-Up) and SKU (Stock Keeping Unit), also exist. The PLU code is shorter, and the SKU is more general than the GTIN, and in practice, further complicating product identification.

The choice of the time frame can be problematic when aggregating scanner data because many products change periodically, sometimes every quarter, month, week, or even day. Prices are often higher on Saturdays, and sales value, transaction numbers, and prices peak in the afternoon or evening (Białek, 2020a).

The advantages and disadvantages of using scanner data are also linked to methodological challenges that must be addressed during scanner data processing. One challenge is the need to process and prepare the acquired data, which requires an appropriate IT structure and qualified personnel. Scanner data can be sourced from points of sale, such as supermarkets, pharmacies, and travel agencies. However, the most valuable suppliers are supermarkets with a large number of barcodes (typically between 10,000 and 25,000 barcodes, mainly for food and beverages). Another source of scanner data is market research companies and electronic trading platforms (e.g., Allegro or OLX, which operate in the Polish market). This results in greater dependence on the supplier than in the past when it was only necessary to obtain permission to visit the sales point. Nowadays, maintaining good relationships with retailers and having legal agreements with retailers to define the cooperation details is crucial. Another significant methodological challenge is selecting the appropriate price index formula for elementary homogeneous product groups.

4. Choice of Price Index Formula

The choice of formula is one aspect of the broad spectrum of decisions to be made when working with scanner data. When describing bilateral and multilateral methods (indices), it is important to acknowledge that using new data sources (including scanner data) is highly demanding for statistical offices. This is due to the vast volume of data and the high turnover of products associated with these sources (Statistics Poland, 2024).

In traditional data collection, elementary indices are used for the lowest level of data aggregation, i.e., where data on only the prices of goods and services are collected by interviewers in the field. Weighted indices are used at higher data aggregation levels, where the statistical office has knowledge of the level of consumption obtained through the Household Budget Survey (Eurostat, 2024). Nevertheless, the exception to this rule is scanner data. Scanner data, even at the lowest level of data aggregation (i.e., at the barcode level), provide information on the value of product sales; thus, it is then possible to use both weighted and unweighted indices (Eurostat, 2017). Some EU countries use the **dynamic approach**, whereby a sample of products is taken from month to month by appropriate data filtering (e.g., a low sales filter is applied), and a chained Jevons index (unweighted formula) is calculated (Van Loon, Roels, 2018). However, there are also countries that implement weighted multilateral indices, e.g., Statistics Netherlands has implemented the Geary-Khamis index into regular production at the barcode level (Chessa, 2018).

The economic literature offers a wide range of price indices, each with its own specific applications as well as theoretical and empirical justifications. This analysis focuses on the indices selected by the author. A broader set of measures could be considered; however, to maintain clarity and transparency of the discussion, it was necessary to make a selection.

4.1. Bilateral Indices

Most National Statistical Offices use bilateral indices to measure the CPI, particularly in the traditional data collection framework, using the Jevons index at the lowest aggregation level and the Laspeyres index at the highest level (Białek, 2020b). These methods compare current prices and quantities of goods with corresponding prices and quantities from a base (fixed) period, making them transparent and easy to explain to users.

Unweighted Indices

The three most known elementary formulas are the Dutot index (Dutot, 1738), the Carli index (Carli, 1804) and the Jevons index (Jevons, 1865), all of which are based on unweighted average prices or price ratios (Eurostat, 2020b). Eurostat prohibits the use of the Carli index because it generates charges that overestimate inflation. The Dutot and Jevons indices, which are the two elementary aggregate formulas laid down in HICP regulations, do not suffer from this problem (Eurostat, 2024). According to EUROSTAT, the most recommended formula is the Jevons index (Eurostat, 2020a)

The Dutot index is defined as the ratio of the average prices of goods from the current and base months:

$$P_D^{0,t} = \frac{\frac{1}{N_{o,t}} \sum p_i^{\mathrm{t}}}{\frac{1}{N_{o,t}} \sum p_i^0}, \tag{1}$$

where:

$N_{o,t}$ – the number of matched products in the compared months;

τ = 0 – the base period;

$\tau = t$ – the current period;

p_i^τ – the price of the *i*-th product in the period τ.

The second formula discussed in the article is the Carli index, which represents the arithmetic mean of the partial indices:

$$P_c = \frac{1}{N_{o,t}} \sum_{i=1}^{N} \frac{p_i^t}{p_i^0}. \tag{2}$$

The last unweighted index discussed is the Jevons price index, which is calculated as the geometric mean of the partial price indices:

$$P_J^{0,t} = \prod_{i \in G_{0,t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{N_{0,t}}}, \tag{3}$$

where:

p_i^τ – the price of the *i*-th product at the time $\tau \,\epsilon \{0, t\}$;

$G_{o,t}$ – products observed simultaneously in the months *0* (base) and *t* (current);

$N_{o,t} = \mathrm{card}\, G_{o,t}$.

Weighted Indices

The second group of indices are weighted indices, including superlative indices firstly proposed by Diewert (1976), such as the Törnqvist index (Törnqvist, 1936):

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

where:

s_i^0 and s_i^t represent the expenditure shares of the i-th in the base and current periods, respectively.

The Fisher index (Fisher, 1922) is the geometric mean of the Laspeyres (1871) ($P_{\mathrm{La}}^{0,t}$) and Paasche ($P_{\mathrm{Pa}}^{0,t}$) indices (Paasche, 1874):

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

where:

$P_{\mathrm{La}}^{0,t}$ – the Laspeyres index,

$P_{\mathrm{La}}^{0,t}$ – the Paasche index.

4.2. Multilateral Indices

Multilateral indices are particularly useful for scanner data and were initially introduced to compare price levels between countries or regions (Gini, 1931; Éltetö, Köves, 1964; Chessa, 2021). They provide transitive price comparisons, which is a desirable property because the results are independent of the choice of the base country. As such, multilateral indices are considered a solution to the problems encountered with bilateral indices (Eurostat, 2020b). The price index measures the aggregate price change in the current period compared to the base period. Multilateral indices include all products available in at least two periods of a time window, usually 13 months, ensuring that the seasonality of products is considered. Multilateral indices weigh each product according to its significance in each period and help avoid issues related to chain drift found in bilateral indices (Eurostat, 2020b). Numerous studies emphasise that the use of multilateral indices in scanner data analysis reveals their advantages over simple bilateral formulas (see, e.g., Vartia, 2018; Diewert, 2022).

Multilateral indices are transitive, meaning their result for any two periods within the time window does not depend on the choice of the base period. This transitivity is crucial for eliminating chain drift. 'In case of promotional sales with reduced prices, the quantities purchased often increase substantially. But when the prices return to their original level, the quantities purchased of storable goods may not return to their 'normal' level. This type of asymmetric behaviour can cause chain drift in superlative price indices, which is typically downward' (Eurostat, 2020b). Well-known multilateral methods include the GEKS index (Gini, 1931), the Geary-Khamis (GK) index (Geary, 1958), and the CCDI (Caves Christensen and Diewert) index (Caves, Christensen, Diewert, 1982; Eurostat, 2020b).

The standard formula of the GEKS index (Gini, 1931; Éltetö, Köves, 1964), which was originally used for international comparisons between countries or regions, is based on the superlative Fisher index (Fisher, 1922). The GEKS index between the period 0 and period t is calculated as the unweighted geometric mean of T + 1 bilateral price indices $P^{t,t}$ and $\mathrm{P}^{\mathrm{t},0}$, which are based on the same formula. Bilateral indices should pass the time reversal test, i.e., they should satisfy $P^{a,b} * P^{b,a} = 1$. The formula for the GEKS index can be expressed as:

$$P_{\mathrm{GEKS}}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{p_F^{\tau,t}}{P_F^{\tau,0}} \right)^{\frac{1}{T+1}}. \tag{6}$$

The Geary-Khamis index (Geary, 1958; Khamis, 1972) can be expressed as:

$$P_{\mathrm{GK}}^{0,t} = \frac{\sum_{\mathrm{i} \in \mathrm{G_t}} \mathrm{p_i^t q_i^t} / \sum_{\mathrm{i} \in \mathrm{G_0}} \mathrm{p_i^0 q_i^0}}{\sum_{\mathrm{i} \in \mathrm{G_t}} \mathrm{v_i^{GK} q_i^t} / \sum_{\mathrm{i} \in \mathrm{G_0}} \mathrm{v_i^{GK} q_i^0}}, \tag{7}$$

where:

$$\mathrm{v_i^{GK}} = \sum_{\mathrm{z}=0}^{\mathrm{T}} \varphi_{\mathrm{i,GK}}^{\mathrm{z}} \frac{\mathrm{p_i^z}}{\mathrm{p_{GK}^{0,z}}}, \tag{8}$$

$$\varphi^{\mathrm{z}}_{\mathrm{z,GK}} = \frac{\mathrm{q}^{\mathrm{z}}_{\mathrm{i}}}{\sum_{\tau=0}^{\mathrm{T}} \mathrm{q}^{\tau}_{\mathrm{i}}}. \tag{9}$$

The CCDI index (Caves, Christensen, Diewert, 1982) is essentially the GEKS index based on the Törnqvist formula, replacing the 'standard' Fisher index in GEKS. It is expressed as follows:

$$P^{0,t}_{\mathrm{CCDI}} = \prod_{\tau=0}^{T} \left(\frac{p^{\tau,t}_{T}}{P^{\tau,0}_{T}} \right)^{\frac{1}{T+1}}, \tag{10}$$

where:

$$P^{\tau,t}_{T} = \prod_{i \in G_{\tau,t}} \left(\frac{p^{t}_{i}}{p^{\tau}_{i}} \right)^{\frac{s^{\tau}_{j}+s^{t}_{j}}{2}},$$

s^{τ}_{j} – the expenditure share of the i-th product at the time τ;

s^{t}_{j} – the expenditure share of the i-th product at the time t.

Despite their complexity, multilateral indices are recommended for compiling price indices from transaction data. Due to their valuable axiomatic properties and the fact that they operate over the full-time window, multilateral indices are well-suited for index calculations when working with scanner data. Research is still ongoing to determine the best multilateral indices for compiling elementary groups using scanner data. Scanner data offer many possibilities for new research and development.

5. Selection of Samples: Static and Dynamic Approaches

Processing scanner data involves several key stages. One of them is to select the appropriate price index, which was discussed earlier. Another crucial step is to determine the sampling approach for products, which directly affects how the sales structure is represented in price calculations. In this context, two primary approaches are distinguished: the static approach and the dynamic approach (Van Loon, Roels, 2018).

5.1. The Static Approach

The static approach closely resembles traditional price collection methods and relies on a fixed product sample. Product codes are selected at the end of the year and remain in use for the following 12 months. This selection is carefully designed to ensure that the sample accurately represents the sales structure for the entire year while minimising the impact of seasonal sales fluctuations, such as those typical in December.

Maintaining this sample requires balancing representativeness and feasibility. If a product becomes less representative or is discontinued, it must be replaced with an equivalent item. While this method follows conventional sampling procedures, it benefits from full transaction data, allowing for a more informed selection of products and adjustments as necessary throughout the year.

The static approach is particularly useful when scanner data are used on a limited scale and need to be integrated with traditionally collected price data. In such cases, it may be practical to manually select product codes that best match the descriptions used in conventional price collection. However, this method also has drawbacks – it is labour-intensive and does not fully exploit the wealth of available scanner data.

5.2. The Dynamic Approach

The dynamic approach involves automatically updating the product sample on a monthly basis. This method selects product codes that appear in both the current and previous month and that account for a significant share of turnover in a given elementary aggregate. This ensures that important new products are included while less relevant ones are removed. Research shows that while a wide range of goods is available on the market, only a small portion of them generate the majority of turnover. The dynamic method takes this into account.

This method resembles a continuous replenishment system and chain-linked index calculation, making it well-suited for handling large volumes of scanner data. Its key advantage lies in its automation, which greatly simplifies data processing. However, frequent product relaunches or replacements must be handled separately to ensure appropriate quality adjustments.

The dynamic basket utilises a set of filters and an algorithm to select a matched sample. The dump filter allows for the removal of products that experience a simultaneous significant drop in price and sales value, indicating that the product will soon be withdrawn and will no longer be representative.

The low-sales filter is designed to eliminate product codes with very low sales or, conversely, to ensure that the selected products account for a sufficiently large share of total turnover (ranging from 50% to 80%). The low-sales algorithm is defined as:

$$\frac{s_{m-1}+s_m}{2} > \frac{1}{(\mathrm{nx}\lambda\,)},$$

where:

s_{m-1} and s_m – the turnover share in the month m – 1 and m;

n – the number of matched items in the elementary aggregate;

λ = 1.25.

The value 1.25 for λ is empirically determined so that the selected item codes represent about 80% of turnover. Since the filter is based on turnover value, certain product categories, such as low-value or unbranded goods, are automatically excluded from the sample, as their low prices contribute to a low turnover.

6. Empirical Example

This chapter uses a sample set of scanner data to present the behaviour of various price indices, distinguishing between bilateral indices (weighted and unweighted) and multilateral indices. In the subsequent section, the dynamics of these indices are analysed using selected filters: the low sales filter and the dump prices filter.

The following empirical study uses daily scanner data from a retail chain in Poland for the period December 2018 to December 2019. The collections come from the PriceIndices package in R (Białek, 2021), which is publicly available and free of charge. The study analyses two datasets: coffee (which contains 42,561 observations (rows)) and milk (4,386 rows). The datasets contain six columns and include the following variables:

- time – transaction dates (year-month-day);
- prices – prices of sold products (PLN);
- quantities – quantities of sold products (milk in litres, coffee in kilograms);
- prodID – unique product codes;
- retID – unique identifiers for retail outlets;
- description – descriptions of sold products.

The milk dataset contains six different product descriptions that correspond to **subgroups** of the milk group, i.e., powdered milk, low-fat pasteurised milk, low-fat UHT milk, full-fat pasteurised milk, full-fat UHT milk, and goat milk. The coffee dataset includes three different homogeneous products: instant coffee, coffee beans, and ground coffee.

Below is an example of a scanner dataset used to calculate indices:

Table 1. Data frame with the first six milk harvest observations

	time	prices	quantities	prodID	retID	description
1	2018–12–01	8.78	9.0	14 215	2 210	powdered milk
2	2019–01–01	8.78	13.5	14 215	2 210	powdered milk
3	2019–02–01	8.78	0.5	14 215	1 311	powdered milk
4	2019–02–01	8.78	8.0	14 215	2 210	powdered milk
5	2019–03–01	8.78	0.5	14 215	1 311	powdered milk
6	2019–03–01	8.78	1.5	14 215	2 210	powdered milk

Source: PriceIndices package (Białek, 2021).

6.1. Unweighted and Weighted Bilateral Indices

Figures 1 and 2 compare selected elementary chain price indices calculated for the product groups. Figure 1 compares selected unweighted indices, while Figure 2 shows weighted bilateral indices. Tables 2 and 3 present the values of individual indices for each product category.

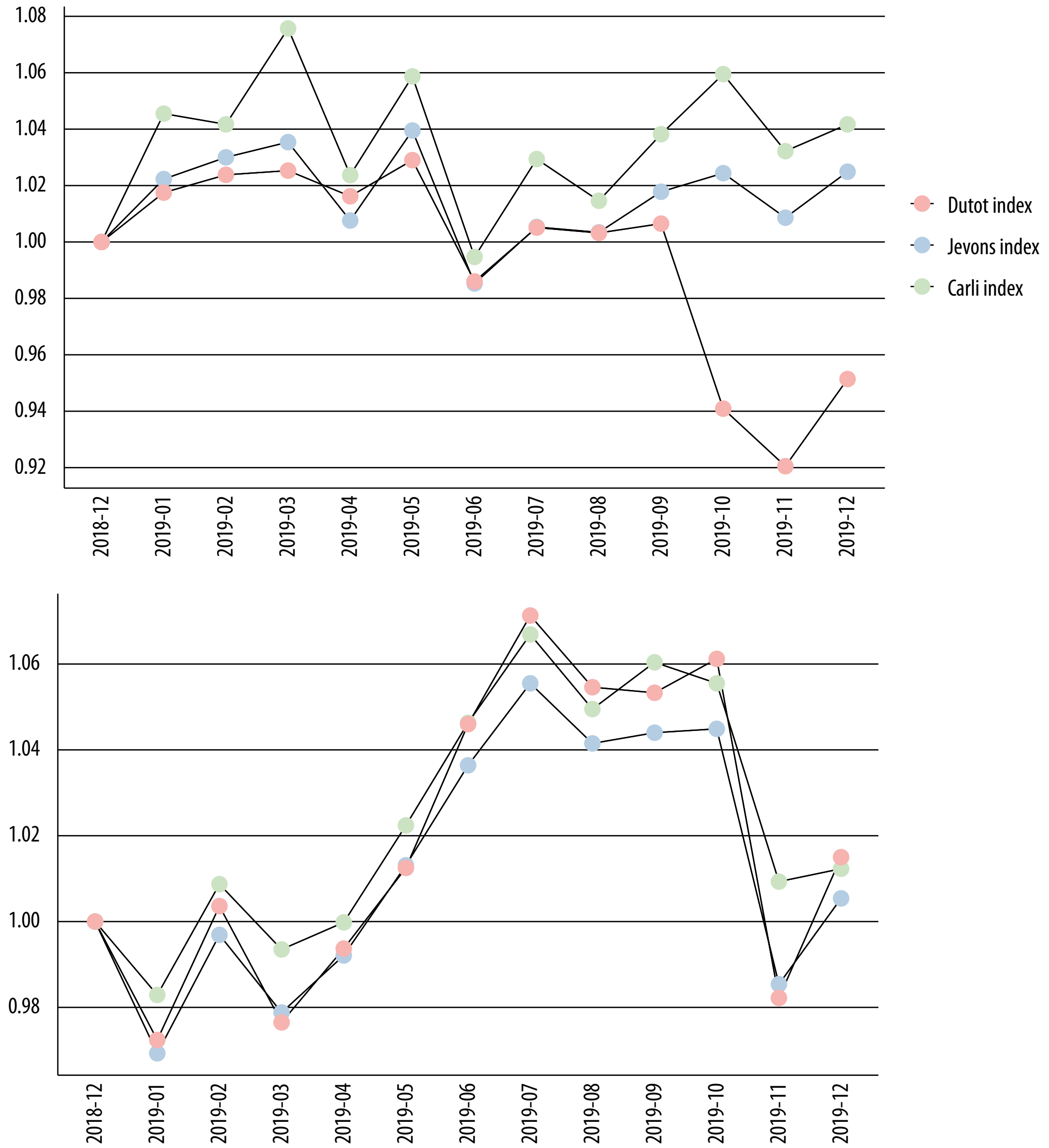

Figure 1. Unweighted bilateral indices

Source: own elaboration based on data from the PriceIndices package.

As shown in Figure 1, the values of individual unweighted indices differ slightly. For the coffee dataset, the indices generate similar values. The milk dataset shows more significant differences between the indices. Looking at the milk and coffee charts, it is not possible to unequivocally state which index generates the lowest or highest values as these relationships differ between datasets. Table 1 presents the values of the indices for each month discussed.

Table 2. Unweighted bilateral index values for milk products

Data	Dutot index	Jevons index	Carli index
2018–12	1.0000	1.0000	1.0000
2019–01	1.0175	1.0223	1.0455
2019–02	1.0238	1.0300	1.0417

Data	Dutot index	Jevons index	Carli index
2019–03	1.0253	1.0354	1.0757
2019–04	1.0162	1.0076	1.0237
2019–05	1.0290	1.0395	1.0587
2019–06	0.9860	0.9853	0.9947
2019–07	1.0051	1.0053	1.0294
2019–08	1.0032	1.0034	1.0146
2019–09	1.0065	1.0178	1.0382
2019–10	0.9409	1.0244	1.0595
2019–11	0.9205	1.0086	1.0322
2019–12	0.9514	1.0249	1.0417

Source: own elaboration based on data from the PriceIndices package in R.

Examining the values of the individual indices shows that the lowest values were recorded for the Dutot index and the highest for the Carli index. The most significant differences occurred in October, November, and December between the Dutot index and the Carli index, with discrepancies of −0.1186 pp, 0.1117 pp, and 0.0903 pp, respectively.

Table 3. Unweighted bilateral index values for coffee products

Data	Dutot index	Jevons index	Carli index
2018–12	1.0000	1.0000	1.0000
2019–01	0.9724	0.9693	0.9829
2019–02	1.0036	0.9969	1.0087
2019–03	0.9765	0.9788	0.9935
2019–04	0.9937	0.9921	0.9998
2019–05	1.0125	1.0131	1.0224
2019–06	1.0460	1.0364	1.0463
2019–07	1.0713	1.0555	1.0669
2019–08	1.0546	1.0415	1.0495
2019–09	1.0533	1.0440	1.0604
2019–10	1.0612	1.0449	1.0555
2019–11	0.9822	0.9854	1.0093
2019–12	1.0150	1.0054	1.0123

Source: own elaboration based on data from the PriceIndices package in R.

The unweighted bilateral indices calculated for coffee are characterised by smaller differences than for milk. In this case, the most significant differences in values between the indices occur in November, with a difference of 0.0270 pp. between the Dutot and Carli indices and −0.0239 pp. between the Dutot and Jevons indices.

In the following analysis, weighted bilateral indices for the products were examined. Figure 1 presents a graph comparing the indices, and Tables 3 and 4 show the values for each chart.

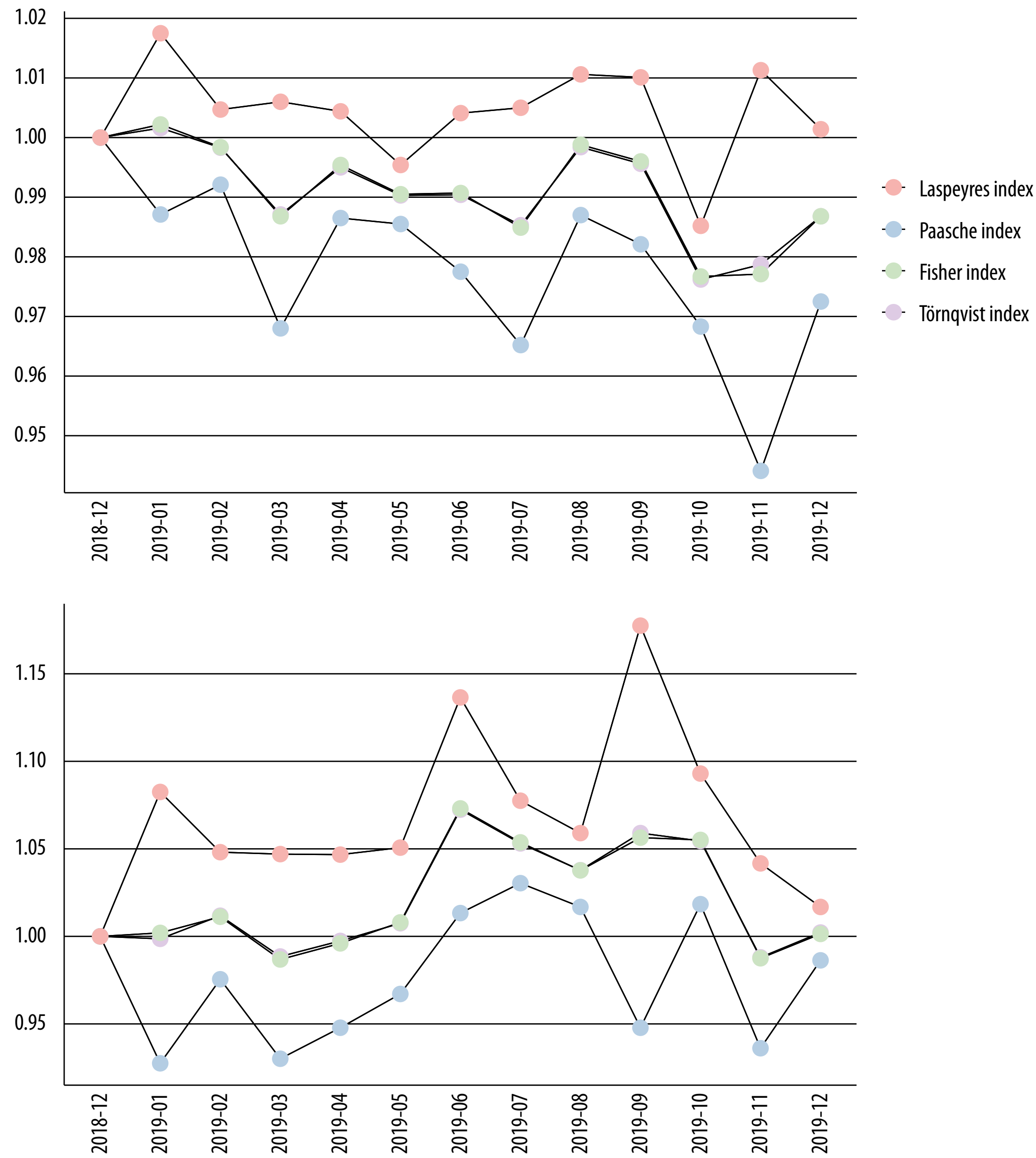

Figure 2. Weighted bilateral indices

Source: own elaboration based on data from the PriceIndices package in R.

For the weighted indices, a consistent pattern emerges across both datasets. The Fisher and Törnqvist indices have very close values, which is consistent with the analytically demonstrated mutual approximation of superlative indices (Białek, Beręsewicz, 2021). Conversely, the Laspeyres index generates the highest values, while the Paasche index generates the lowest. It is not clearly established whether a fixed-base approach or a chain version of these indices is better for the Fisher and Törnqvist indices.

Table 4. Weighted bilateral index values for milk products

Data	Laspeyres index	Paasche index	Fisher index	Törnqvist index
2018–12	1.0000	1.0000	1.0000	1.0000
2019–01	1.0175	0.9871	1.0022	1.0016
2019–02	1.0047	0.9921	0.9984	0.9983
2019–03	1.0060	0.9680	0.9868	0.9871
2019–04	1.0044	0.9865	0.9954	0.9950
2019–05	0.9954	0.9855	0.9905	0.9903
2019–06	1.0041	0.9775	0.9907	0.9904
2019–07	1.0050	0.9652	0.9849	0.9853
2019–08	1.0106	0.9870	0.9988	0.9984
2019–09	1.0101	0.9821	0.9960	0.9956
2019–10	0.9852	0.9683	0.9767	0.9762
2019–11	1.0113	0.9441	0.9771	0.9787
2019–12	1.0014	0.9725	0.9868	0.9868

Source: own elaboration based on data from the PriceIndices package in R.

For the weighted bilateral indices for milk, the most significant differences can be observed in November, with a difference of 0.0673 pp between the Laspeyres and Törnqvist indices. In March and July, significant differences were also noted for these indices, amounting to 0.0380 pp and 0.0398 pp, respectively.

Table 5. Weighted bilateral index values for coffee products

Data	Laspeyres index	Paasche index	Fisher index	Törnqvist index
2018–12	1.0000	1.0000	1.0000	1.0000
2019–01	1.0825	0.9275	1.0020	0.9986
2019–02	1.0481	0.9755	1.0112	1.0118
2019–03	1.0470	0.9301	0.9868	0.9885
2019–04	1.0467	0.9478	0.9960	0.9974
2019–05	1.0507	0.9671	1.0080	1.0075
2019–06	1.1365	1.0133	1.0731	1.0724
2019–07	1.0775	1.0304	1.0537	1.0531
2019–08	1.0590	1.0168	1.0377	1.0378
2019–09	1.1775	0.9478	1.0564	1.0590
2019–10	1.0930	1.0184	1.0551	1.0545
2019–11	1.0417	0.9361	0.9875	0.9880
2019–12	1.0168	0.9863	1.0014	1.0022

Source: own elaboration based on data from the PriceIndices package in R.

The average values for the Laspeyres index are 1.0168, for the Paasche index 0.9863, for the Fisher index 1.0014, and for the Törnqvist index 1.0021. The most significant differences were noted in January and September between the Laspeyres index and the Paasche index, amounting to 0.1550 pp and 0.2298 pp.

6.2. Comparison of Selected Bilateral and Multilateral Indices

Multilateral methods are increasingly used when working with scanner data because they address problems such as chain drift. Despite this, some countries still use the chain Jevons index to develop price indices for scanner data, although it is not the optimal choice. Some statistical institutions also use chain Törnqvist indices (e.g., the USA and Japan), which are weighted formulas.

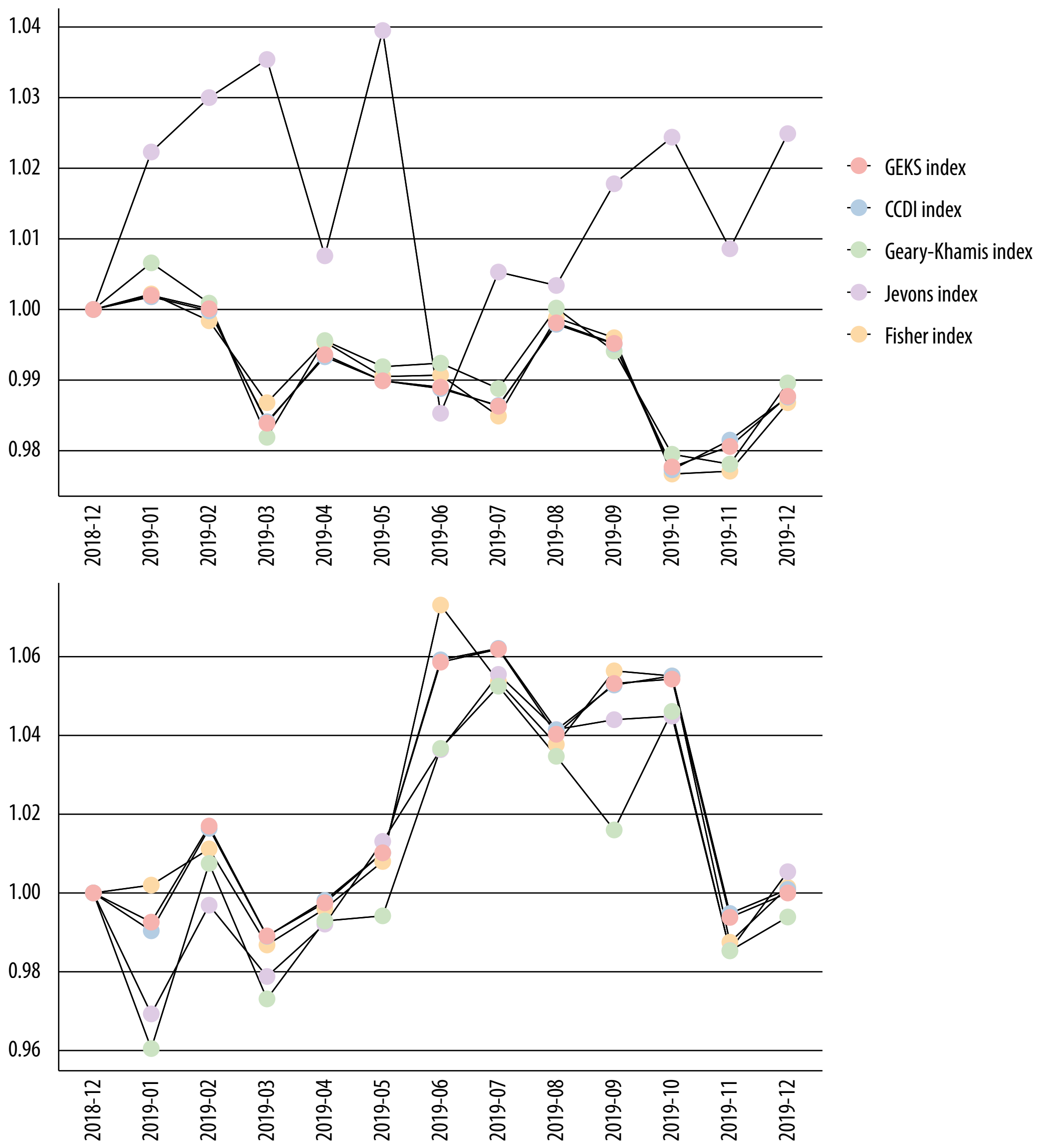

Figure 3. Comparison of bilateral and multilateral indices
Source: own elaboration based on data from the PriceIndices package.

As an alternative to bilateral indices, Figure 3 presents multilateral indices recommended for scanner data. Selecting the unweighted Jevons formula for the dataset results in significantly higher values than other indices. While its application to classically obtained prices at the lowest level of data aggregation is justified (as interviewers only note product prices), it is not the best choice for scanner data because it does not account for consumption volumes.

Table 6. Comparison of bilateral and multilateral indices for milk products

Data	GEKS index	CCDI index	Geary-Khamis index	Jevons index	Fisher index
2018–12	1.0000	1.0000	1.0000	1.0000	1.0000
2019–01	1.0020	1.0018	1.0066	1.0223	1.0022
2019–02	1.0001	0.9998	1.0009	1.0300	0.9984
2019–03	0.9839	0.9841	0.9819	1.0354	0.9868
2019–04	0.9936	0.9933	0.9956	1.0076	0.9954
2019–05	0.9899	0.9899	0.9919	1.0395	0.9905
2019–06	0.9890	0.9888	0.9924	0.9853	0.9907
2019–07	0.9863	0.9864	0.9888	1.0053	0.9849
2019–08	0.9981	0.9979	1.0002	1.0034	0.9988
2019–09	0.9952	0.9951	0.9941	1.0178	0.9960
2019–10	0.9777	0.9773	0.9795	1.0244	0.9767
2019–11	0.9806	0.9815	0.9781	1.0086	0.9771
2019–12	0.9877	0.9876	0.9896	1.0249	0.9868

Source: own elaboration based on data from the PriceIndices package in R.

When looking solely at the multilateral indices, the differences in values appear from the second or third decimal place. It is also not possible to clearly state which index consistently has the highest or lowest values. Adding bilateral indices to the analysis reveals that their values are higher compared to the multilateral indices, especially in the case of the Jevons index.

Table 7. Comparison of bilateral and multilateral indices for coffee products

Data	GEKS index	CCDI index	Geary-Khamis index	Jevons index	Fisher index
2018–12	1.0000	1.0000	1.0000	1.0000	1.0000
2019–01	0.9926	0.9904	0.9605	0.9693	1.0020
2019–02	1.0170	1.0164	1.0075	0.9969	1.0112
2019–03	0.9891	0.9890	0.9731	0.9788	0.9868
2019–04	0.9975	0.9981	0.9929	0.9921	0.9960
2019–05	1.0102	1.0102	0.9942	1.0131	1.0080
2019–06	1.0586	1.0592	1.0367	1.0364	1.0731

Data	GEKS index	CCDI index	Geary-Khamis index	Jevons index	Fisher index
2019–07	1.0619	1.0621	1.0525	1.0555	1.0537
2019–08	1.0403	1.0414	1.0347	1.0415	1.0377
2019–09	1.0532	1.0528	1.0160	1.0440	1.0564
2019–10	1.0543	1.0551	1.0461	1.0449	1.0551
2019–11	0.9938	0.9948	0.9853	0.9854	0.9875
2019–12	1.0000	1.0009	0.9939	1.0054	1.0014

Source: own elaboration based on data from the PriceIndices package in R.

In the case of the coffee dataset, no index stands out in terms of value, unlike the situation with the milk dataset. Here, the index values are more similar to each other. Among the multilateral indices, the Geary-Khamis index stands out slightly in comparison to the other two indices. When comparing the multilateral indices with the bilateral ones, the Fisher index stands out more, usually exhibiting the highest values.

6.3. Static and Dynamic Approaches

Figure 4 presents the comparison of selected bilateral indices (the Jevons index and the Carli index) and multilateral indices (the GEKS index and the CCDI index). They show how index values change after applying the low sales filter and the dump prices filter. The results are presented separately for the milk and coffee datasets.

Figure 4. GEKS index values for the milk dataset depending on the applied data filter (low sales filter example)

Source: own elaboration based on data from the PriceIndices package in R.

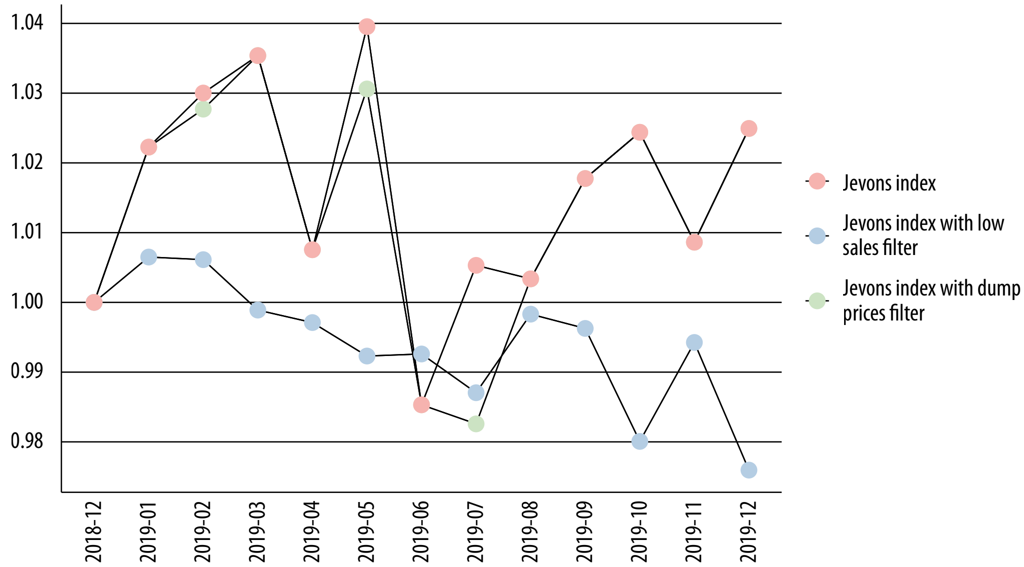

Figure 5. Jevons index values for the milk dataset depending on the applied data filter (low sales filter example)

Source: own elaboration based on data from the PriceIndices package in R.

Figure 6. Dutot index values for the milk dataset depending on the applied data filter (low sales filter example)

Source: own elaboration based on data from the PriceIndices package in R.

For the milk dataset and across the three different indices (GEKS, Jevons, Dutot), the standard index and the index after applying the dump prices filter display similar trends. Conversely, the indices with the low sales filter applied tend to exhibit lower values. The multilateral GEKS index exhibits the least variability among the analysed differences. The differences between index variants are relatively small, with the low sales filter only significantly deviating from the other variants in a few instances. The Jevons index (Figure 5) shows greater variability than GEKS, especially in certain periods (e.g., the first half of 2019). Clear differences between the indices can be seen, particularly for the version with the low sales filter which deviates from the index without filters. In the case of the Dutot index (Figure 6), the indexes without filters and with the dump prices filter record a decrease in value in 09–2019 while the index with the low sales filter over the entire observation remains relatively the same. In the other two cases of the above-mentioned indices, there is no such relationship.

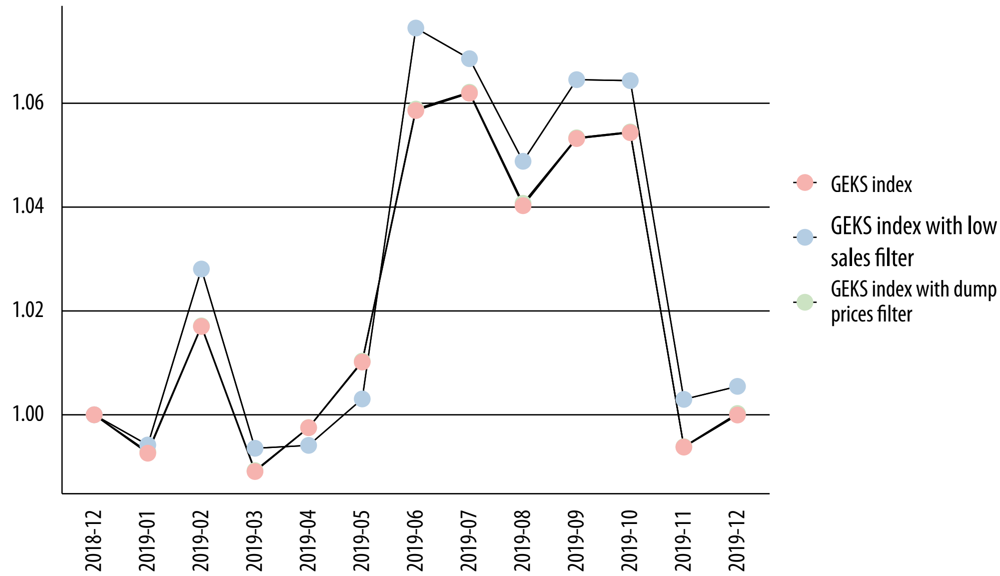

Figure 7. GEKS index values for the coffee dataset depending on the applied data filter (low sales filter example)

Source: own elaboration based on data from the PriceIndices package in R.

Figure 8. Jevons index values for the coffee dataset depending on the applied data filter (low sales filter example)

Source: own elaboration based on data from the PriceIndices package in R.

Figure 9. Dutot index values for the coffee dataset depending on the applied data filter (low sales filter example)

Source: own elaboration based on data from the PriceIndices package in R.

For the coffee dataset, three types of analysis are also presented: the indexes without applying filters, with the low sales filter and the dump prices filter for three indices (GEKS, Jevons, Dutot). In each case, the indices without filters and those after applying the dump prices filter follow a similar pattern, while the index with the low sales filter shows greater deviations. The Jevons index (Figure 8) and the Dutot index (Figure 9) exhibit high variability among the analysed methods. The index with the low sales filter shows stronger deviations, particularly in mid-2019, where it significantly exceeds the values of the other indices. The Dutot index (Figure 9) behaves similarly to the Jevons index, with moderate variability. The use of low sales filter also increases the index values. The multilateral GEKS index shows the smallest differences between the filtered and unfiltered variants. However, it is still evident that the low sales filter leads to higher index values, although this effect is less pronounced than in the case of the Jevons or Dutot indexes.

7. Conclusions

Scanner data hold enormous potential for refining and modernising inflation measurement. They provide information on consumption at the lowest level of aggregation, making the quality of scanner data incomparable to traditional datasets. However, it is important to recognise the many problems and challenges associated with using such data. As mentioned in Section 3, one of the main challenges is processing and preparing scanner data. The detailed level of information, the choice of time frame, and the large volume of data (typically ranging between 10,000 and 100,000 individual barcodes) require specialist knowledge in machine learning and statistics, as well as an appropriate IT environment for process automation.

This paper focused on comparing price indices and determining the scale of differences that arise from using individual indices. The choice of index formula is not straightforward, as the difference between indices for a single homogeneous product group can be significant. In the context of measuring inflation, such differences between the actual and measured price index are very dangerous and lead to serious financial consequences, such as excessive government expenditure indexing and taxes (Boskin et al., 1996). Other studies also show that at the elementary level, differences may result not only from the applied formula but also from the choice of data source (Białek, Panek, Zwierzchowski, 2022).

Multilateral indices, characterised by transitivity and the absence of chain drift, are a rational choice for scanner data. To formulate a recommendation for the most appropriate price index for scanner data, further research on index properties – in both theoretical and empirical terms – is needed, along with analysing the experiences of countries that already use this methodology.

The results obtained from scanner data analysis provide several important insights. Firstly, they confirm relationships that are already well-established in traditional price collection: superlative indices such as Fisher and Törnqvist tend to approximate each other, while the Laspeyres index systematically overestimates inflation and the Paasche index underestimates it. Secondly, however, the divergence between the Laspeyres and Paasche indexes is notably larger in scanner data than in traditional price collection. In conventional surveys, this difference in Poland, at the COICOP 2 level, usually does not exceed a few promilles. In contrast, when working with scanner data at a very low level of aggregation (barcode level), the differences between bilateral indices become significant, which clearly illustrates the substitution effect between goods. This finding highlights the methodological importance of carefully selecting the appropriate index formula in scanner data applications.

When applying the low sales and dump prices filters, it is evident that bilateral indices show greater variation in values compared to multilateral indices. In both datasets, the low sales filter deviated from the index without filters, while the dump prices filter displayed similar values.

The choice between the static and dynamic approaches depends on the characteristics of the data, their availability, and the analytical objectives. The static approach aligns more closely with traditional methodologies and is advantageous when integrating scanner data with conventional price collection. In contrast, the dynamic approach offers greater flexibility and makes better use of large datasets, particularly in rapidly changing markets.

References

Białek J. (2020a), *Remarks on Price Index Methods for the CPI Measurement Using Scanner Data*, “Statistics and Economy Journal Statistika”, vol. 100(1), pp. 55–70.

Białek J. (2020b), *Wykorzystanie danych skanowanych do pomiaru inflacji – doświadczenia międzynarodowe i wyzwania metodologiczne*, “Wiadomości Statystyczne. The Polish Statistican”, vol. 65(01), pp. 9–33.

Białek J. (2021), *PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations*, “Statistika – Statistics and Economy Journal”, vol. 101(2), pp. 121–141.

Białek J., Beręsewicz M. (2021), *Scanner data in inflation measurement: From raw data to price indices*, “Statistical Journal of the IAOS”, vol. 37(4), pp. 1315–1336

Białek J., Panek T., Zwierzchowski J. (2022), *Assessing the effect of new data sources on the consumer price index: a deterministic approach to uncertainty and sensitivity*, “Statistics in Transition New Series”, vol. 23(3), pp. 1–25.

Boskin M.J., Dulberger E.R., Gordon R.J., Griliches Z., Jorgenson D.W. (1996), *Toward a More Accurate Measure of the Cost of Living. Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index*, https://www.ssa.gov/history/reports/boskinrpt.html [accessed: 18.12.2025].

Carli G.R. (1804), *Del valore e della proporzione de'metalli monetati*, “Scrittori Classici Italiani di Economia Politica”, vol. 13, pp. 297–336.

Caves D.W., Christensen L.R., Diewert W.E. (1982), *Multilateral comparisons of output, input, and productivity using superlative index numbers*, “Economic Journal”, vol. 92(365), pp. 73–86.

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

Chessa A. (2021), *A comparison of index extension methods for multilateral methods*, https://eventos.fgv.br/sites/eventos.fgv.br/files/arquivos/u161/index_extension_methods_compared_chessa_og19.pdf [accessed: 18.12.2025].

Diewert W.E. (1976), *Exact and superlative index numbers*, “Journal of Econometrics”, vol. 4(2), pp. 115–145.

Diewert W.E. (2022), *Scanner data, elementary price indexes and the chain drift problem*, [in:] K.J. Fox, E.R. Lawrence (eds.), *Advances in Economic Measurement: A Volume in Honour of DS Prasada Rao*, Palgrave Macmillan, Singapore, pp. 445–606.

Dubois P., Griffith R., O'Connell M. (2022), *The use of scanner data for economics research*, “Annual Review of Economics”, vol. 14(1), pp. 723–745.

Dutot C. (1738), *Reflexions Politiques sur les Finances et le Commerce*, Les freres Vaillant et Nicolas Prevost, The Hague.

Éltetö O., Köves P. (1964), *On a problem of index number computation relating to international comparison*, “Statisztikai Szemle”, vol. 42(10), pp. 507–518.

Eurostat (2017), *Harmonised Index of Consumer Prices. Practical Guide for Processing Supermarket Scanner Data*, https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/practical-guide-supermarket-scanner-data-september-2017.pdf [accessed: 19.12.2025].

Eurostat (2020a), *Consumer Price Index Manual – Concepts and Methods*, https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@stat/documents/publication/wcms_761444.pdf [accessed: 20.12.2025].

Eurostat (2020b), *Guide on multilateral methods in the Harmonised Index on Consumer Prices (HICP) – 2022 edition*, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-21-020 [accessed: 20.12.2025].

Eurostat (2024), *Harmonised Index of Consumer Prices (HICP) Methodological manual – 2024 edition*, https://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/w/ks-gq-24-003 [accessed: 21.12.2025].

Fisher I. (1922), *The making of index numbers: a study of their varieties, tests, and reliability*, Houghton Mifflin, Boston–New York.

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)”, vol. 121(1), pp. 97–99.

Gini C. (1931), *On the circular test of index numbers*, “Metron”, vol. 9(9), pp. 3–24.

Hałka A., Leszczyńska A. (2011), *Wady i zalety wskaźnika cen towarów i usług konsumpcyjnych – szacunki obciążenia dla Polski*, „Gospodarka Narodowa", no. 9, pp. 51–75.
Jevons W.S. (1865), *On the variation of prices and the value of the currency since 1782*, "Journal of the Statistical Society of London", vol. 28(2) pp. 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 (General)", vol. 135(1), pp. 96–121.
Laspeyres K. (1871), *Ix. die berechnung einer mittleren waarenpreissteigerung*, "Jahrbücher für Nationalökonomie und Statistik", vol. 16(1), pp. 296–318.
Paasche H. (1874), *Über die Preisentwicklung der letzten Jahre nach den Hamburger Börsennotirungen*, "Jahrbücher für Nationalökonomie und Statistik", vol. 12, pp. 168–178.
Statistics Poland (2019), *Co warto wiedzieć o inflacji?*, Główny Urząd Statystyczny, Warszawa, https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5464/18/1/1/co_warto_wiedziec_o_inflacji.pdf [accessed: 21.12.2025].
Statistics Poland (2024), *Co warto wiedzieć o inflacji? cz. II*, Główny Urząd Statystyczny, Warszawa, https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5464/18/2/1/co_warto_wiedziec_o_inflacji_cz.ii.pdf [accessed: 21.12.2025].
Törnqvist L. (1936), *The bank of Finland's consumption price index*, "Bank of Finland Monthly Bulletin", vol. 10, pp. 1–8.
Van Loon K., Roels D. (2018), *Integrating big data in the Belgian CPI. Meeting of the Group of Experts on Consumer Price Indices*, UNECE, Geneva, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf [accessed: 21.12.2025].
Vartia Y., Suoperä A., Nieminen K., Montonen K. (2018), *Circular Error in Price Index Numbers Based on Scanner Data. Preliminary Interpretations*, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Finland_main.pdf [accessed: 29.12.2025].

Zastosowanie danych skanowanych do obliczania inflacji: na przykładzie Polski

Streszczenie: W kontekście rosnącego wykorzystania danych skanowanych w pomiarze inflacji kluczowym wyzwaniem metodologicznym pozostaje wybór odpowiedniej formuły indeksu cenowego oraz metody filtrowania danych. Celem artykułu jest zbadanie wpływu tych wyborów na wyniki pomiaru dynamiki cen. W pracy przeanalizowano dane transakcyjne zebrane z terminali punktów sprzedaży w polskich supermarketach dla wybranych grup produktów: kawy oraz mleka. Porównano różne formuły indeksów cenowych – w tym bilateralne nieważone (Dutot, Carli, Jevons), bilateralne ważone (Törnqvist, Fisher) oraz multilateralne (Geary-Khamis, CCDI) – oraz przetestowano wpływ różnych metod filtrowania danych skanerowych. Wyniki pokazują, że zarówno wybór formuły indeksu cenowego, jak i zastosowana metoda filtrowania danych skanerowych mają mierzalny i istotny wpływ na ocenę dynamiki cen. Różnice te mogą prowadzić do odmiennych interpretacji inflacji, co podkreśla potrzebę świadomego i uzasadnionego wyboru metodologicznego przy wykorzystaniu danych skanerowych w statystyce cen.

Słowa kluczowe: dane skanowane, indeksy multilateralne, indeksy bilateralne