STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. xx–XX, https://doi.org/10.59139/stattrans-2026-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Sampling techniques in the CPI measurement

Jacek Białek[1]

Abstract

The procedure used by a National Statistical Office (NSO) for collecting prices to produce the Consumer Price Index (CPI) is based on sample surveys. The universe (or population) of items has three dimensions: product, geographical, and time, all of which are described in the paper. This paper presents and discusses general concepts and techniques of survey sampling that are crucial for the construction of price indices. In particular, both probability and non-probability sampling techniques are discussed and illustrated with the real-world examples. A separate section discusses sampling scanned products. One of the approaches used for such data is the *dynamic approach*, which involves monthly sampling by applying appropriate data filters. This technique can be seen as a special form of *cut-off sampling*. The empirical study investigates the effect of data filtering on the level of price indices. The main pragmatic conclusion is that the low-sales filter has the most significant impact on reducing the size of the scanner dataset. The second important conclusion is that changing the order of data filtering has minimal impact on the value of the price index.

Key words: probability sampling, non-probability sampling, Consumer Price Index, scanner data, dynamic approach, multilateral indices

1. Introduction

The procedure used for price collection by a National Statistical Office (NSO) when producing a Consumer Price Index (CPI) is a *sample survey*. Here, the CPI (or the Harmonised Index of Consumer Prices, HICP) can play a role of a *target quantity* which is defined with respect to (CPI Manual, 2004): (1) *a universe* that comprises finite population of units (e.g., products or outlets); (2) *variables*, which are defined for the units in the universe (e.g., prices and quantities of products or expenditure shares of outlets); (3) *a parameter*, which is a single value obtained on the basis of values of those variables (e.g., the Jevons (1865) or the Laspeyres (1871) price index).

In general, there are three sampling dimensions (HICP Methodological Manual, 2018; CPI Manual: Concepts and methods, 2020): (I) *a product dimensions*, which consists of all purchased products and varieties of products; (II) *a geographical and outlet dimension*, which consists of all places (e.g., small shops, supermarkets, petrol stations, web-pages, etc.) where the product is sold; (III) *a time dimension*, which comprises those days of the month for which the applicable price index is determined.

For each of these dimensions, there is a *general population* from which a sample will be drawn. The population (universe) of products from the CPI basket is divided into COICOP

[1] Department of Statistical Methods, University of Łódź, Łódź, Poland. E-mail: jacek.bialek@uni.lodz.pl & Department of Prices and Services, Statistics Poland, Poland. E-mail: J.Bialek@stat.gov.pl.
ORCID: https://orcid.org/0000-0002-0952-5327.

5-digit sub-classes, although this division can go down to a lower level of data aggregation (e.g., COICOP 6-digit level) for web-scraped data or scanner data (see Section 5). A sample of products is drawn from each product sub-class, with a common practice being to decide on *representative products* in each sub-class. The population (universe) of outlets includes all places that sell consumer products in a given COICOP product group. Since outlets have specific locations on the country map, the outlet universe has a geographical character. For the time dimension, the universe consists of all sub-periods of the month since the consumer may buy products on any day of the month. The CPI Manual (2004) pays less attention to the time dimension because "price variation is usually smaller over a short time span." However, at least for web-scraped datat, this aspect may be more relevant.

Depending on the product group, the above-mentioned dimensions have differing degrees of importance when collecting data to measure inflation (HICP Methodological Manual, 2004). For instance, *fresh fruits* (COICOP 5: 01161) have highly volatile prices within a month, so the price sampling strategy should not focus only on the product and outlet dimension but also on the time dimension. In contrast, prices for *actual rentals paid by tenants* (COICOP 0411) are generally fixed for at least a month; thus, the time dimension is irrelevant in the sampling procedure.

Sampling is an alternative to conducting a full survey on all observations from a population, which is obviously impossible in practice and would be too costly. Additionally, excessive workload for interviewers in the field could substantially reduce their efficiency and the quality of the data collected. However, while this remark is valid for *traditional data collection*, it has limited applicability when dealing with *alternative data sources*. For example, in the case of scanner data, the use of multilateral indices generally does not require any random sampling of products (Eurostat, 2022). Instead, all data from all outlets from a given retail chain are included. However, to ensure the representativeness of the data while simultaneously reducing analysis time, various data filters are then applied. This thread will be discussed in more detail in Section 5.

In the simplest terms, *probability sampling* involves selecting units in such a way that each one (e.g., product, outlet, or day) has a known non-zero probability of being included in the sample. For example, in an outlet draw, we can determine that each outlet has an equal probability of being included in the sample or that this probability is proportional to the number of people employed at that outlet or its sales revenue. In contrast, in *non-probability sampling*, the probability of selecting any particular unit is unknown (and often impossible to determine). Although *probability sampling* is the recommended approach for sampling in statistical surveys, *non-probability sampling* techniques still dominate the CPI measurement in most countries.

The main aim of the paper is to discuss and verify the effectiveness of selected sampling techniques used in the CPI measurement. This article addresses two gaps in the existing literature. First, it illustrates the sampling techniques discussed for constructing an inflation basket with respect to scanner data. To the best of the author's knowledge, there is a lack of studies in the literature that apply these techniques using scanner data; this article not only fills that gap but also provides practical scripts written in the R environment. Second, no existing research examines the impact of the sequence in which data filters are applied to scanner data on the resulting price index. Data filtering is a recommended method for selecting scanned product samples within the so-called dynamic approach (see Section 5.2). Therefore, addressing this second research gap may be of substantial importance to statistical offices that employ the chain Jevons index to estimate inflation based on scanner data.

The structure of the paper is as follows: Section 2 discusses the main probability sampling techniques; Section 3 describes non-probability sampling methods; Section 4 presents the main results obtained when trying to estimate bilateral population price indices; Section 5 discusses the problem of sample selection when using scanner data to compile the CPI, and Section 6 is an empirical study which compares the effectiveness of selected sampling techniques based on real scanner data sets. The main conclusions from the empirical study are discussed in Section 7.

2. Probability sampling techniques

This section presents, discusses and illustrates methods of survey sampling that are implemented when measuring a CPI. In particular, the sub-sections focus on three main probability sampling techniques, i.e., *simple random sampling*, *systematic sampling* (in two variants) as well as *probability proportional to size sampling* (for a broader overview on that topic see Särndal et. al. (2003)). Please note that both Section 2 and Section 3 concern the traditional CPI data collection, and Section 5 discusses sample selection for scanner CPI data.

The survey sampling approach assumes that the universe (population) consists of a finite number N of observational units. The sampling procedure selects a sample S that comprises n units out of N available, where the inclusion probability $\pi_i = P(i \in S)$ is known for each unit $i \in \{1, 2, ..., N\}$.

The universe can be divided into strata, denoted here by $h \in \{1, 2, ..., H\}$. Each stratum, which can be treated as mini-universe with sampling taking place independently in each one, consists of N_h units, where $\sum_{h=1}^{H} N_h = N$. For example, for outlet dimension the universe of outlets can be divided into four disconnected sub-populations: online stores, small neighborhood stores, supermarkets and hypermarkets.

A *sampling frame* is a list of all (or most) of the N units in a given universe. Sampling frames for the outlet dimension could be business registers or any records of local administrations (CPI Manual, 2004). A products list obtained from sellers or a product list obtained from price collectors can be used as sampling frames for the product dimension.

2.1. Simple random sampling and systematic simple sampling

In *simple random sampling* and *systematic simple sampling* each unit is drawn with equal inclusion probability, which means that $\pi_i = \frac{n}{N}$. In simple random sampling, all units are sampled with replacement. Simple random sampling without replacement is not addressed in the CPI Manual (2004), likely because it entails a changing selection probability for each unit as the population size diminishes with successive draws. Therefore, this article only considers simple random sampling with replacement. In systematic sampling, only the first element is drawn randomly in that way, and the remaining units are selected at equal distances from each other in the sampling frame (CPI Manual, 2004).

2.2. Probability proportional to size sampling

In *probability proportional to size (pps) sampling*, the inclusion probability is proportional to an auxiliary variable x_i (CPI Manual, 2004; HICP Methodological Manual, 2018). This can be expressed as $\pi_i = nx_i / \sum_{j=1}^{N} x_j$. CPI Manual (2004) on page 69 states: "Units for which initially this quantity is larger than one are selected with certainty, whereafter the inclusion probabilities are calculated for the remainder of the universe". For example, when drawing outlets, an auxiliary variable could be the number of people employed at the outlet or the sales volume from the last year of operation (if this information is available).

While it is theoretically possible to consider a fixed or random sample size, when compiling a CPI in practice a fixed sample size is typically considered in each stratum (CPI Manual, 2004, p. 70). Specifically, a statistical office can consider various sampling techniques that provide fixed-size *pps* samples. One such technique is *systematic pps sampling*, which follows a similar concept to *simple sampling*, but the first sample element is drawn in the *pps* scheme. Another techique is *order pps sampling*, which is described below.

Order pps sampling is a commonly accepted technique for selecting *pps* samples, and is widely discussed in Rosén (1997a, 1977b). Once the auxiliary variable x_i is determined, the procedure begins by assigning each $i-$th unit in the population a uniform random number $U_i \in (0,1)$. The units are then assigned a number Q_i as the value of a differential function with arguments x_i and U_i, i.e., $Q_i = f(x_i, U_i)$. The units in the population are then sorted in ascending order relative to the value of Q_i. The n units with the smallest Q_i values are sampled. The CPI Manual (2004) discusses two important cases of the above-mention approach, i.e., *sequential pps sampling* with $Q_i = U_i / z_i$, and *Pareto pps sampling* with $Q_i = (U_i(1-z_i))/(z_i(1-U_i))$, where $z_i = nx_i / \sum_{j=1}^{N} x_j$. Rosen (1997b) showed that for estimating mean and variance, these order sampling techniques are only approximately *pps*. *Pareto pps* is marginally better than *sequential pps* and should therefore be preferred in the price index contex (CPI Manual, 2004, p.71). For more detailed information about *Pareto pps* see, for instance, Lindblom and Teterukovsky (2007), where a case with strata is considered. As it was mentioned above, probability sampling is less commonly used by statistical agencies than non-probability sampling. However, Lindblom (2003) provides details concerning the probabilistic approach used in Sweden, while probability sampling methods used by the US Bureau of Labor Statistics are described in Sections 5.24 - 5.26 in the CPI Manual (2004).

2.3. Empirical illustration

The empirical illustration of probability sampling concerns the selection of outlets of retail chain operating in Poland, with the objective of determining price indices for an elementary group of coffee products. For demonstration purposes, we will use scanner data on coffee sales as available in the *PriceIndices* R package (Białek, 2021). A more extensive discussion of scanner data is given in Section 5, and thus, a detailed description of the structure of this type of data is omitted here.

The 'coffee' dataset contains transaction data of coffee sales in $N = 20$ outlets representing the population for this study. We assume that we need to draw a sample of $n = 4$ outlets. The sales data includes three types of coffee: instant coffee, coffee beans and ground coffee, and we will focus on the period from January 2019 to December 2019. It means that we observed 79 coffee products with a total of 14,392 records. The script that implements the outlet sampling is available at https://github.com/JacekBialek/important_documents/blob/main/SIT_illustration_1.Rmd.

As a result of running the R script, the user receives a table of results on the basis of which the selection of outlets is made. Let us first discuss the columns of this table. The first column (*Outlet ID*) indicates the outlet identification number assigned by the retail chain. The x_i column contains the values of the size variable, which in our illustration is the annual coffee sales revenue of each outlet (in PLN). The next column, x_i^{cum}, contains the cumulative values of the size variable (PLN). Column z_i contains values of the intermediate variable described in Section 2.2, which will be used in the *pps* method. Uniform random values between 0 and 1 are in the column labelled U_i. The values of $Q_i = f(x_i, U_i)$, depending on whether the *sequential pps sampling* or *Pareto pps sampling* technique is implemented, are in columns Q_i^{seq} and Q_i^{Par} respectively. Finally, the last four columns indicate the four outlets drawn, depending on the probabilistic sampling method. Specifically, we have the results for: simple random sampling (*simple*), systematic pps sampling (*systematic*), sequential (order) pps sampling (*seq*) and Pareto (order) pps sampling (*Pareto*).

Table 1. Selection of outlets using the sampling techniques

Outlet ID	x_i	x_i^{cum}	z_i	U_i	Q_i^{seq}	Q_i^{Par}	simple	systematic	seq	Pareto
2183	747848.76	747848.76	0.18	0.81	4.53	19.51	-	-	-	-
2381	859283.40	1607132.16	0.21	0.07	0.34	0.29	-	✓	✓	✓
2681	844018.93	2451151.09	0.20	0.61	3.03	6.23	-	-	-	-
3782	702174.50	3153325.59	0.17	0.13	0.79	0.76	✓	-	-	-
4080	928925.11	4082250.70	0.22	0.42	1.88	2.51	-	-	-	-
4281	938415.01	5020665.71	0.22	0.54	2.42	4.10	-	-	-	-
4380	796774.77	5817440.48	0.19	0.15	0.81	0.78	✓	✓	-	-
4580	1159091.58	6976532.06	0.28	0.08	0.28	0.22	✓	-	✓	✓
4681	807040.59	7783572.65	0.19	0.64	3.33	7.47	-	-	-	-
4780	894942.71	8678515.36	0.21	0.25	1.15	1.20	-	-	-	-
4883	770725.98	9449241.34	0.18	0.51	2.79	4.68	-	-	-	-
5480	826464.99	10275706.33	0.20	0.10	0.52	0.47	-	✓	-	-
6681	809634.89	11085341.22	0.19	0.97	5.01	123.33	✓	-	-	-
7081	728462.60	11813803.82	0.17	0.43	2.48	3.61	-	-	-	-
7481	854912.08	12668715.90	0.20	0.36	1.76	2.18	-	-	-	-
7482	626678.63	13295394.53	0.15	0.05	0.34	0.30	-	-	✓	✓
8480	1153981.94	14449376.47	0.28	0.59	2.13	3.73	-	✓	-	-
8580	846678.61	15296055.08	0.20	0.42	2.07	2.84	-	-	-	-
9082	755509.54	16051564.62	0.18	0.03	0.14	0.12	-	-	✓	✓
9182	712747.31	16764311.93	0.17	0.89	5.21	37.85	-	-	-	-

While *simple random sampling*, *sequential pps sampling* and *Pareto pps sampling* methods have been sufficiently described in Section 2.2, the results in Table 1 on *systematic pps sampling* still require additional clarification. After determining the cumulative values of size variable x_i^{cum}, one integer I_x from the interval $(0, max)$ is drawn (*one_sample_number* in R script), where max is the floor value of the last cumulative size variable value (i.e., the total sum of size variable x_i) divided by $n = 4$. In the next step, the next three numbers are determined non-randomly: $I_x + max$, $I_x + 2max$ and $I_x + 3max$. At this stage we have four threshold values. The final stage involves selecting those outlets for which the cumulative value of the size variable has exceeded the given threshold for the first time. In our empirical illustration we obtain: $max = 4191078$, $I_x = 1469607$, $I_x + max = 5660685$,

$I_x + 2max = 9851763$ and $I_x + 3max = 14042841$, which leads to the following sample of outlets: $\{2381, 4380, 5480, 8480\}$. The sample structure is exactly the same in the case of *sequential (order) pps sampling* and *Pareto (order) pps sampling*, although this in not always guaranteed. In our case these two techniques lead to the following sample of outlets: $\{2381, 4580, 7482, 9082\}$, while *simple random sampling* provides the following sample of outlets: $\{3782, 6681, 4580, 4380\}$ (the outlet with $ID = 4580$ appears in both samples).

3. Non-probability sampling techniques

Probability sampling is more advanced than non-probability sampling and is thus more demanding on the researcher (statistician). Therefore, non-probability sampling is easier to implement, and perhaps this is one of the reasons why this approach is more common in the practices of statistical offices. Another technical reason could be the lack of availability of a sampling frame, especially for the product dimension. An argument for using non-probability sampling may also be the low measurement bias it generates as a result. Moreover, de Haan, Opperdoes and Schut (1999) verified the bias that results from non-probability sampling based on scanner data and found that the mean square error (MSE) was often smaller than that for *pps* sampling. Furthermore, when there is a shortages of interviewers, it may be cheaper to collect prices close to where the interviewers live. Sending interviewers to new locations and training them each time a new sample is drawn is certainly both time-consuming and costly. Finally, in probabilistic sampling, statisticians must often contend with oversampling, such as when the population of individuals is already small at the outlets. Thus, below we will discuss the main approaches commonly used in non-probability sampling.

3.1. Cut-off sampling

Cut-off sampling refers to the situation when the n 'largest' sampling units are selected with certainty, and the remaining units have zero chance of being included in the sample (CPI Manual, 2004). The term 'largest' units refers to units with the highest values of size variable that are highly correlated with the target variable. In general, the *cut-off sampling* method provides biased estimators; however, if we are primarily concerned with reducing MSE, this method may be a good way of sampling. This is because any estimator from *cut-off sampling* has zero variance (de Haan, Opperdoes and Schut, 1999).

A specific case of *cut-off sampling* is the filtering of scanner data using a *dynamic approach*. On the one hand, the automation of the collection of electronic transaction data and its full availability (provided that the retail chain signs an agreement with the statistical office) means that there is no need to sample products, varieties or points in time when using scanner data (CPI Manual, 2004, p. 74). On the other hand, to reduce chain drift bias and account for the impact of clearance sales and dump prices, some statistical offices choose to use the chain Jevons index while first filtering the scanner data (e.g., by eliminating relatively low sales or product with extreme price changes from the dataset). This process is known as a *dynamic approach*, where samples are selected in this way from month to month. It will be discussed in more detail in Section 5.

3.2. Quota sampling

Quota sampling is a non-random selection method for survey samples (Cochran, 1977). The share (number or percentage) of units in the sample is determined in such a way that it is proportional to their actual share in the entire survey population. Although a sample obtained through quota sampling is not selected using a random technique, it can still be representative of the entire population, albeit to a limited extent. This representativeness largely depends on the level of detail available about the population under study.

It is important to note that *quota sampling* requires central management of the whole sampling process, which may limit its usefulness in practice. Additionally, the standard error of any estimate cannot be determined in the case of *quota sampling*, further limiting this method (CPI Manual, 2004).

3.3. The representative item method

The representative item method is a traditional CPI method where the statistical office compiles a list of product types along with their specifications (CPI Manual, 2004). If the product type specification is very precise and, therefore, narrow, interviewers receive exact guidelines on which product should be added to the sample. However, this precision may make searching for a product that is compliant with the specification much more difficult or even impossible in a given area or a given period. Conversely, if the type-specification is relatively broad, interviewers have greater freedom in selecting a sample of the most popular products locally. As a rule, this approach leads to better representativity of the sample compared to the narrow type-specification variant.

3.4. Sampling in time

The CPI (like the HICP) refers to a month. Prices of goods and services usually fluctuate throughout the month; however, in the practice of statistical offices, interviewers collect prices on a specific day of the month. The CPI Manual (2004) gives the 15th day of the month as an example of a reference day for price measurement. In Poland, price quotations are carried out by interviewers from the 5th to the 22nd of each month. As a rule, prices of goods are collected once a month, but for some products, price quotations are more frequent (e.g., prices of fresh fruit in Poland are collected twice a month due to their high price volatility). However, even quoting prices twice a month may be insufficient when the price of a product or service varies substantially and depends, for example, on the day of the week (e.g., cinema tickets are more expensive on weekends). Nevertheless, this practice results more from statistical offices' limited funds and human resources rather than from methodological guidelines.

A separate issue when using scanner data to compile a CPI is *sampling in time*. In this case, the expectations of the statistical office must be confronted with the cooperation offered by the retail chains. However, assuming that the agreement between the retail chain and the statistical office leaves a lot of freedom in selecting the period from which the data should come within a month, the individual product usually covers the first three (or sometimes even four) weeks of the month (Eurostat, 2022). For more details, see Section 5.

3.5. Empirical illustration

This section illustrates the *cut-off sampling* method using scanner data on coffee sales, as described in Section 2. This method will be used to sample $n = 9$ products from $N = 79$ coffees products available for sale in 2019. In this empirical illustration, we will consider two size variables: the value of coffee sales and the number of coffee product sold. The total value of coffee sales in the population is 16,764,311.93 PLN, which corresponds to the total number of coffee products sold: 358575. Each product (coffee) is identified based on an internal code (ID) assigned by the retail chain, which has been verified to have a 1:1 relationship with the EAN (European Article Number) barcode. Additionally, two levels of data aggregation are considered: the higher COICOP 5 level (where, as mentioned above, the population contains 79 units) and the lower level of aggregation, i.e. COICOP 6 level, where sales are divided into three product subgroups: coffee beans (23 IDs), ground coffee (37 IDs) and instant coffee (19 IDs).

Since *cut-off sampling* involves considering the $n = 9$ largest products in terms of sales value and sales volume, let us first look at the histograms for these two size variables (Fig. 1 and Fig. 2). At first glance, noticeable differences can be observed between the distributions of size variables, supporting the initial hypothesis that the choice of size variable may substantially impact the final sample selection using the discussed technique.

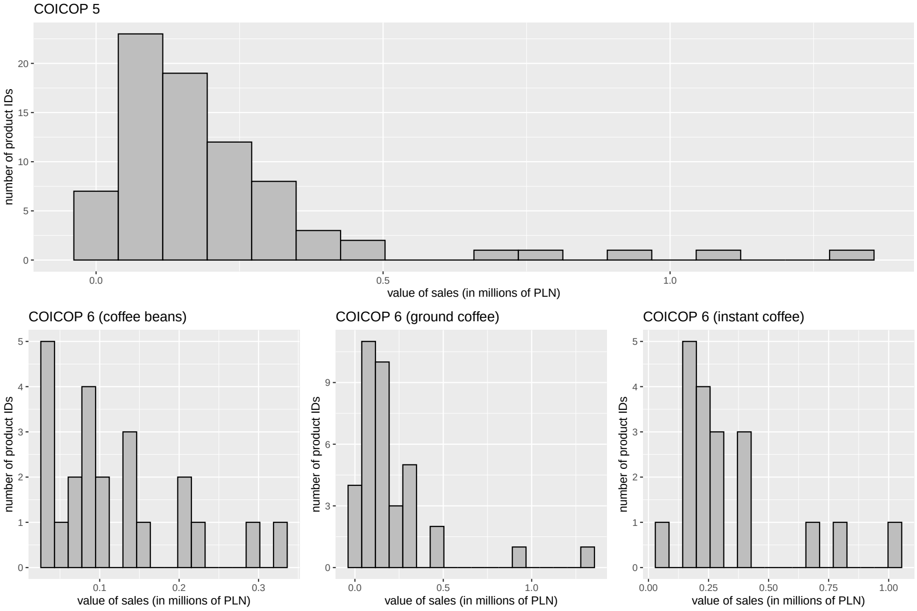


Figure 1. Histogram of sales values determined for the coffee products population, presented at both COICOP 5 and COICOP 6 aggregation levels

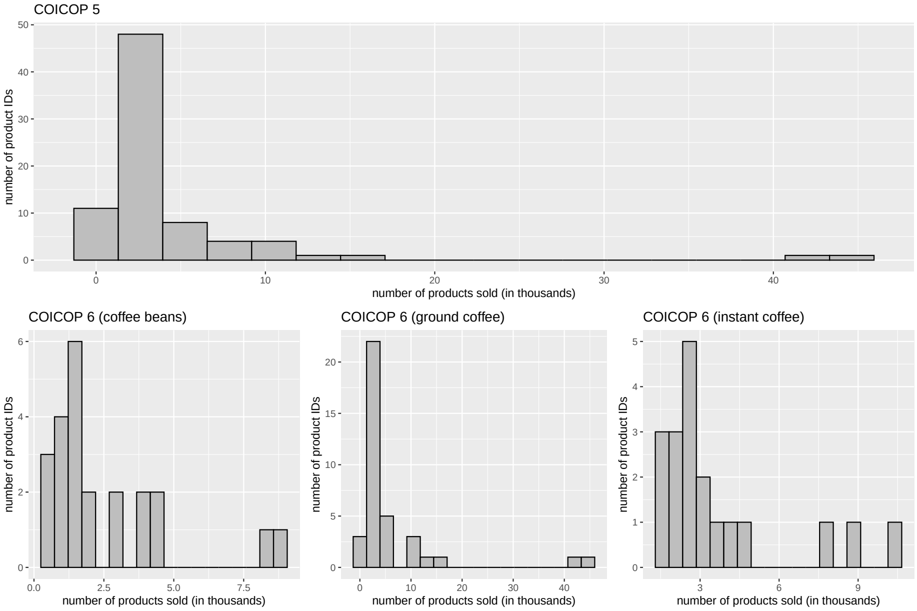


Figure 2. Histogram of the number of products sold determined for the coffee products population, presented at both COICOP 5 and COICOP 6 aggregation levels

Our second hypothesis to be verified is that the samples obtained by selecting three units from each of the three coffee subgroups (based on the given size variable) do not necessarily give the same result as the initial sample of 9 units taken at the COICOP 5 level. Both working hypotheses were confirmed in our empirical illustration, but the first hypothesis holds only at the higher level of data aggregation (COICOP 5). Table 2 and Table 3 highlight the ID numbers of coffee products that appeared in all considered cut-off sampling variants in bold. As one can see, both the selection of the size variable and the choice of the level of data aggregation are important with regard to the structure of the sample obtained using the *cut-off sampling* method.

Table 2. Characteristics of samples of coffee products, presented at both COICOP 5 and COICOP 6 aggregation levels (a size variable is the value of coffee sales)

	COICOP 5	**COICOP 6**		
Characteristics	call coffee products	ccoffee beans	cground coffee	cinstant coffee
sample product IDs	**2401950**, **2401947**, **2402723**, 2401948, **2400379**, **2400915**, **2402453**, 2400368, 32308	33955, 75096, 22687	**2401950**, **2402723**, **2400915**	**2401947**, **2401948**, **2400379**
total sales (PLN)	6436417.44	845251.21	2669982.48	2555225.97
population share (%)	38.39	31.06	34.98	39.84

Table 3. Characteristics of samples of coffee products, presented at both COICOP 5 and COICOP 6 aggregation levels (a size variable is the number of coffee product sold)

	COICOP 5	**COICOP 6**		
Characteristics	all coffee products	coffee beans	ground coffee	instant coffee
sample product IDs	**2402723**, **2401950**, **2400915**, **2402453**, 2400655, 2401380, **2401947**, 2403353, **2400379**	33955, 22687, 89025	**2401950**, **2402723**, **2400915**	**2401947**, **2401948**, **2400379**
no. of sold products	168776	21825	102351	27213
population share (%)	47.06	37.23	44.44	39.05

Coffee bean products were underrepresented at the COICOP 5 level, which is due to the low sales value within this product group. However, at the COICOP 6 level, this group has representatives in the sample (see Table 2 and Table 3). Notably, at the COICOP 6 level there are almost no differences in the sample structure due to the size variable (in fact, the samples differ in only one coffee bean). At the higher level of data aggregation (COICOP 5), samples designated for different size variables overlap in only 2/3 of cases. We encourage the reader to conduct similar experiments for a larger sample size since the script that implements the presented coffee product cut-off sampling is available at https://github.com/JacekBialek/important_documents/blob/main/SIT_illustration_2.Rmd

4. Price indices in the sampling approach

Measurement of the CPI begins at the elementary level, where interviewers note the prices of representatives of each elementary group of products sold in various outlets in the survey regions (e.g., Poland has 207 such regions). At this level, elementary (unweighted) indices are used to determine price dynamics, which are discussed in detail in Section 4.1. At higher levels of aggregation, where information is available on both prices and consumption levels, weighted price indices are used, selectively discussed in Section 4.2. An exception arises with scanner data, where knowledge of consumption levels is already available at the lowest level of data aggregation (the bar-code level), and therefore there are no restrictions on the choice of the price index formula at the elementary level of data aggregation (see Section 5).

Let us suppose we have a population (universe) of N goods and we are interested in estimating a target (population) price index $P^{0,t}$, which compares a current period t with a base one 0. To achieve this aim we collect a sample $S \subset \{1,2,..,N\}$ of goods for which, depending on the information available, we can obtain full observations $\{p_i^0, p_i^t, q_i^0, q_i^t : i \in S\}$ or limited observations $\{p_i^0, p_i^t : i \in S\}$, where p_i^τ and q_i^τ denote the price and quantity of the $i-th$ unit in a period $\tau \in \{0,t\}$, respectively. Based on the drawn sample S of n units we estimate the population price index $P^{0,t}$ using the sample price index $\hat{P}^{0,t}$.

4.1. Population and sample unweighted indices

Well-established elementary price indices include the Dutot, Carli and Jevons indices (von der Lippe, 2007; CPI Manual, 2004; CPI Manual: Concepts and methods, 2020). Chronologically, the first formal proposal of an elementary price index comes from the French economist Nicolas Dutot (1738). The population Dutot price index can be presented as a ratio of unweighted arithmetic means of prices from compared periods, i.e.,

$$P_D^{0,t} = \frac{\frac{1}{N}\sum_{i=1}^{N} p_i^t}{\frac{1}{N}\sum_{i=1}^{N} p_i^0}. \tag{1}$$

In 1764, the Italian economist Gian Rinaldo Carli proposed an elementary index as an unweighted arithmetic mean of price relatives, known as the Carli (1804) index. It can be expressed as follows:

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

However, due to its superior axiomatic properties, the most recommended elementary price index formula is the Jevons (1865) index (Levell (2015)). This index uses an unweighted geometric mean of price relatives and can be written in terms of the natural logarithm of prices as follows:

$$P_J^{0,t} = (\prod_{i=1}^{N} \frac{p_i^t}{p_i^0})^{\frac{1}{N}}. \tag{3}$$

The last elementary index presented is the Balk-Mehrhoff-Walsh (BMW) index, independently obtained by Mehrhoff and Balk as a linear approximation of the Walsh (1901) index (Eurostat (2018), p. 176; Balk (2005), p. 689). It is formulated as:

$$P_{BMW}^{0,t} = \frac{\sum_{i=1}^{N} \sqrt{(\frac{p_i^t}{p_i^0})}}{\sum_{i=1}^{N} \sqrt{(\frac{p_i^0}{p_i^t})}}. \tag{4}$$

The sample counterparts of these formulas are denoted in the paper by $\hat{P}_D^{0,t}$, $\hat{P}_C^{0,t}$, $\hat{P}_J^{0,t}$ and $\hat{P}_{BMW}^{0,t}$ respectively. For instance, the sample Jevons price index can be written as follows:

$$\hat{P}_J^{0,t} = (\prod_{i \in S} \frac{p_i^t}{p_i^0})^{\frac{1}{n}}. \tag{5}$$

Silver and Heravi (2007) compared the population elementary indices. The statistical approach, which treats calculated elementary indices as estimators of population indices, has been discussed in several studies, including Balk (2005), McClelland and Reinsdorf (1999), and Dorfman, Leaver, and Lent (1999). For instance, McClelland and Reinsdorf (1999) highlight the small sample bias associated with the sample Jevons index when used as an estimator of its population counterpart. Białek (2020) extends Silver and Heravi's (2007) findings by considering the case with correlated prices. Specifically, he demonstrates that the Carli population price index is very sensitive to changes in the level of price correlations when prices are log-normally distributed. Furthermore, Białek (2022) uses a very general continuous-time stochastic approach to compare elementary indices. In particular, he compares expected values and variances of sample Dutot, Carli and Jevons indices under the assumption that prices are described by a geometric Brownian motion (GBM).

As the purpose of this paper is not to discuss the properties of sample price indices in detail, the formulas for their variances or Mean Square Errors (MSEs) are omitted. For readers interested in more detail in this area, we recommend Balk (2005). In the following sections, we present only the most important findings regarding sample elementary price indices, which can serve as estimators for the weighted price indices discussed in Section 4.2. These main results are presented in Table 4, where s_i^0 and s_i^t denote the expenditure share of the $i-$th population unit in the base and current periods, respectively.

The term "approximately unbiased" estimator is used when presenting the results in Table 4 and Table 5. Following the CPI Manual (2004), we understand this term to refer to an estimator whose bias is small and decreases as the sample size increases, indicating that the estimator is therefore asymptotically unbiased.

Table 4. Selected estimation finding concerning unweighted sample indices $(*)$

Probability sampling method	Proportionality of weights	Estimation finding
simple random sampling	no weighting scheme	$\hat{P}_D^{0,t}$ is the approximately unbiased estimator of $P_D^{0,t}$
pps sampling method	$p_i^0/\sum_{j=1}^N p_j^0$	$\hat{P}_C^{0,t}$ is the unbiased estimator of $P_D^{0,t}$
pps sampling method	s_i^0	$\hat{P}_J^{0,t}$ is the approximately unbiased estimator of $P_T^{0,t}$
pps sampling method	s_i^0	$\hat{P}_C^{0,t}$ is the unbiased estimator of $P_L^{0,t}$
pps sampling method	q_i^0	$\hat{P}_D^{0,t}$ is the approximately unbiased estimator of $P_L^{0,t}$
pps sampling method	$\sqrt{s_i^0 s_i^t}$	$\hat{P}_{BMW}^{0,t}$ is the approximately unbiased estimator of $P_W^{0,t}$

$*$ *The weighted population indices are described in Section 4.2*

4.2. Weighted population and sample indices

At higher levels of data aggregation, the Laspeyres (1871) index is used to calculate the price dynamics of the CPI basket (see formula (9)). This is due to the fact that consumption data comes from the Household Budget Survey, which is conducted at a certain frequency (e.g., once a year). Consequently, the weighting system based on consumption levels from the base period is, in practice, already outdated in the current period. From both axiomatic and economic perspectives, it would be ideal to use superlative indices (von der Lippe, 2007), which are discussed below. For scanner data (see Section 5), superlative indices can be used even at the lowest data aggregation level.

Superlative price indices, as discussed by Diewert (1976), are the most frequently recommended index formulas for the Cost of Living Index (COLI) approximation. The list of population superlative indices begins with the Walsh (1901) and Törnqvist (1936) price indices, which are given by:

$$P_W^{0,t} = \frac{\sum_{i=1}^{N} \sqrt{q_i^0 q_i^t} \cdot p_i^t}{\sum_{i=1}^{N} \sqrt{q_i^0 q_i^t} \cdot p_i^0}, \tag{6}$$

and

$$P_T^{0,t} = \prod_{i=1}^{N} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}}. \tag{7}$$

where s_i^0 and s_i^t denote the expenditure shares of matched products in months 0 and t.

Another commonly known superlative price index is the Fisher (1922) formula, which can be written as:

$$P_F^{0,t} = \sqrt{P_{La}^{0,t} \cdot P_{Pa}^{0,t}}, \tag{8}$$

where $P_{La}^{0,t}$ and $P_{Pa}^{0,t}$ denote the Laspeyres (1871) price index and the Paasche (1874) price index respectively, given by

$$P_{La}^{0,t} = \frac{\sum\limits_{i \in G_{0,t}} q_i^0 p_i^t}{\sum\limits_{i \in G_{0,t}} q_i^0 p_i^0}, \tag{9}$$

and

$$P_{Pa}^{0,t} = \frac{\sum\limits_{i \in G_{0,t}} q_i^t p_i^t}{\sum\limits_{i \in G_{0,t}} q_i^t p_i^0}. \tag{10}$$

The sample counterparts of the weighted formulas are denoted in the paper by $\hat{P}_{La}^{0,t}$, $\hat{P}_W^{0,t}$, $\hat{P}_F^{0,t}$ and $\hat{P}_T^{0,t}$ respectively. For instance, the sample Walsh price index can be written as follows:

$$\hat{P}_W^{0,t} = \frac{\sum_{i \in S} \sqrt{q_i^0 q_i^t} \cdot p_i^t}{\sum_{i \in S} \sqrt{q_i^0 q_i^t} \cdot p_i^0}. \tag{11}$$

When the target index is a weighted index, the *pps* draw scheme seems to be unnecessary. Table 5 presents the most important findings regarding sample superlative price indices obtained using the *simple random sampling* method (Balk, 2005).

Table 5. Selected estimation finding concerning weighted sample indices

Probability sampling method	Proportionality of weights	Estimation finding
simple random sampling	no weighting scheme	$\hat{P}_T^{0,t}$ is the approximately unbiased estimator of $P_T^{0,t}$
simple random sampling	no weighting scheme	$ln(\hat{P}_F^{0,t})$ is the approximately unbiased estimator of $ln(P_F^{0,t})$
simple random sampling	no weighting scheme	$\hat{P}_W^{0,t}$ is the approximately unbiased estimator of $P_W^{0,t}$

4.3. Empirical illustration

This section illustrates the selected sampling methods for drawing scanner products to calculate sample unweighted and weighted indices, as described in Sections 4.1 and 4.2. The demonstration is based on the scanner data on milk sales, which is implemented in the *PriceIndices* R package (Białek, 2021). The *milk* data set contains $N = 61$ milk products observed over the time interval between Dec, 2018 - Dec, 2019. The following methods are used to sample $n \in \{10, 20, 30\}$ products out of the $N = 61$ milk products available for sale: *cut-off sampling* using total sales value as the size variable, *simple random sampling*, and *pps sampling* with weights proportional to base period expenditure shares.

Table 6 presents the above-discussed population and sample indices for the three selected sampling methods. The columns labelled *cut_off_10, cut_off_20* and *cut_off_30* present sample index results obtained after using the *cut-off sampling* procedure and for sample sizes: $n = 10$, $n = 20$ and $n = 30$, respectively. The columns labelled *simple_10, simple_20* and *simple_30* present sample index results obtained after using the *simple sampling* procedure for the same sample sizes. The columns labelled *pps_10, pps_20* and *pps_30* present sample index results obtained after using the *pps sampling* procedure for the same sample sizes. For the last two probabilistic sampling techniques, the presented index numbers are the results of the simulation experiment in which the sampling procedure was repeated $K = 200$ times, and the mean of the obtained index values was taken.

Table 6. Population indices and mean values of sample indices for the three selected sampling methods

Index name	population_index	cut_off_10	cut_off_20	cut_off_30	simple_10	simple_20	simple_30	pps_10	pps_20	pps_30
Dutot	0.951437	0.988638	1.001703	1.022526	0.992398	0.966445	0.965991	1.002938	0.999994	1.005274
Carli	1.041709	0.986144	0.995187	1.006313	1.041499	1.038971	1.045865	1.060286	1.062067	1.062032
Jevons	1.024937	0.981871	0.992935	1.004036	1.026928	1.023023	1.028455	1.038357	1.038215	1.039196
BMW	1.025366	0.981873	0.992930	1.004035	1.027230	1.023403	1.028890	1.039075	1.039077	1.040043
Laspeyres	1.001400	0.998832	0.999979	1.001354	1.000120	0.999190	1.002136	1.007905	1.006967	1.009558
Paasche	0.972483	0.959394	0.969410	0.972084	0.980302	0.971805	0.975483	0.994187	0.992919	0.992916
Fisher	0.986835	0.978915	0.984576	0.986611	0.990075	0.985339	0.988682	1.000974	0.999887	1.001177
Tornqvist	0.986757	0.978668	0.984461	0.986539	0.989994	0.985244	0.988525	1.000652	0.999512	1.000721
Walsh	0.985306	0.976565	0.982900	0.985069	0.989185	0.983949	0.986995	0.999171	0.997832	0.998695

Table 7. Biases of the sample indices for the three selected sampling methods

Index name	cut_off_10	cut_off_20	cut_off_30	simple_10	simple_20	simple_30	pps_10	pps_20	pps_30
Dutot	0.037200	0.050266	0.071088	0.040960	0.015007	0.014553	0.051501	0.048557	0.053837
Carli	-0.055565	-0.046522	-0.035396	-0.000210	-0.002738	0.004156	0.018577	0.020358	0.020323
Jevons	-0.043066	-0.032002	-0.020902	0.001991	-0.001914	0.003518	0.013420	0.013278	0.014259
BMW	-0.043493	-0.032436	-0.021331	0.001864	-0.001963	0.003524	0.013709	0.013711	0.014677
Laspeyres	-0.002568	-0.001421	-0.000046	-0.001280	-0.002210	0.000736	0.006505	0.005567	0.008158
Paasche	-0.013088	-0.003073	-0.000399	0.007819	-0.000678	0.003001	0.021704	0.020437	0.020433
Fisher	-0.007921	-0.002259	-0.000225	0.003239	-0.001496	0.001846	0.014139	0.013052	0.014342
Tornqvist	-0.008089	-0.002297	-0.000219	0.003237	-0.001513	0.001768	0.013895	0.012754	0.013964
Walsh	-0.008741	-0.002405	-0.000236	0.003880	-0.001356	0.001690	0.013865	0.012527	0.013390

Table 7 presents biases of the sample indices, i.e., differences between their mean values (expected values) and the corresponding population indices. Using the *cut-off method*, there was no simulation procedure, and the sample was taken once (the sales value is fixed for a given period). Table 8 presents standard deviations of the sample indices obtained

in a simulation study, i.e., it concerns only the *simple sampling* and *pps sampling* procedures. The R script that implements the discussed sampling methods in the context of price index estimates is available at:
https://github.com/JacekBialek/important_documents/blob/main/SIT_illustration_3.Rmd

Table 8. Standard deviations of the sample indices obtained using probabilistic sampling methods for two selected sampling methods

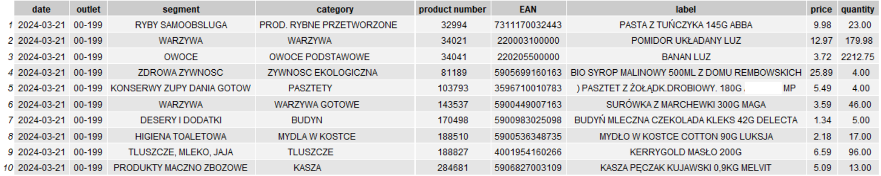
index name	simple_10	simple_20	simple_30	pps_10	pps_20	pps_30
Dutot	0.086655	0.077570	0.064017	0.033060	0.031947	0.024423
Carli	0.065156	0.041950	0.033380	0.074991	0.042128	0.023385
Jevons	0.052950	0.034015	0.026580	0.055243	0.031691	0.018728
BMW	0.053305	0.034293	0.026819	0.055927	0.032093	0.018902
Laspeyres	0.039094	0.028847	0.021635	0.028870	0.020125	0.016812
Paasche	0.038158	0.028326	0.022333	0.014513	0.010779	0.010118
Fisher	0.036257	0.026305	0.020186	0.020425	0.013983	0.011862
Tornqvist	0.036046	0.025999	0.019899	0.020035	0.013574	0.011484
Walsh	0.035144	0.024953	0.018960	0.018347	0.012007	0.010101

As shown in Table 7, *cut-off sampling* works much better for weighted sample price indices (when the target indices are their population counterparts) than for unweighted sample indices. Surprisingly, when using this method, the measurement bias of the weighted sample price index is considerably less than the bias generated by the weighted sample index obtained when using probabilistic techniques to draw products. *Simple sampling* works well for both categories of sample indices, although for weighted sample indices, it works worse than *cut-off sampling* but better than *pps sampling*. Nevertheless, similar to the *pps sampling*, increasing the sample size does not lead to a clear reduction in the sample index bias (Table 7). However, the standard deviation - and thus the variance - of the estimators for both unweighted and weighted sample indices noticeably decreases as the sample size increases (see Table 8).

Please note that the population Dutot price index is the most difficult to estimate (Table 7). Perhaps this is due to the fact that this index - as recommended by the CPI Manual (2004) and Eurostat (2018) - should only be used for highly homogeneous product groups. However, in the this study, the *milk* collection contains clearly disjointed subgroups of milk group, e.g., goat's and cow's milk, UHT and pasteurized milk, and low-fat and high-fat milk products. Therefore, the homogeneity condition may be weakened here, which consequently generates an additional bias in measuring the Dutot price index.

5. Sample selection for scanner data

Scanner data refer to electronic transaction data that specify product prices and expenditures obtained from supermarket IT systems by scanning product bar codes, such as the Global Trade Item Number (GTIN), European Article Number (EAN) or Stock Keeping Unit (SKU). Scanner data are a relatively new and cheap data source for calculating the Consumer Price Index (CPI) and the main advantage of using these data is that they provide full information about products, even at the lowest data aggregation level (see Figure 3).



	date	outlet	segment	category	product number	EAN	label	price	quantity
1	2024-03-21	00-199	RYBY SAMOOBSLUGA	PROD. RYBNE PRZETWORZONE	32994	7311170032443	PASTA Z TUŃCZYKA 145G ABBA	9.98	23.00
2	2024-03-21	00-199	WARZYWA	WARZYWA	34021	220003100000	POMIDOR UKŁADANY LUZ	12.97	179.98
3	2024-03-21	00-199	OWOCE	OWOCE PODSTAWOWE	34041	220205500000	BANAN LUZ	3.72	2212.75
4	2024-03-21	00-199	ZDROWA ZYWNOSC	ZYWNOSC EKOLOGICZNA	81189	5905699160163	BIO SYROP MALINOWY 500ML Z DOMU REMBOWSKICH	25.89	4.00
5	2024-03-21	00-199	KONSERWY ZUPY DANIA GOTOW	PASZTETY	103793	3596710010783	) PASZTET Z ŻOŁĄDK.DROBIOWY. 180G. MP	5.49	4.00
6	2024-03-21	00-199	WARZYWA	WARZYWA GOTOWE	143537	5900449007163	SURÓWKA Z MARCHEWKI 300G MAGA	3.59	46.00
7	2024-03-21	00-199	DESERY I DODATKI	BUDYN	170498	5900983025098	BUDYŃ MLECZNA CZEKOLADA KLEKS 42G DELECTA	1.34	5.00
8	2024-03-21	00-199	HIGIENA TOALETOWA	MYDLA W KOSTCE	188510	5900536348735	MYDŁO W KOSTCE COTTON 90G LUKSJA	2.18	17.00
9	2024-03-21	00-199	TLUSZCZE, MLEKO, JAJA	TLUSZCZE	188827	4001954160266	KERRYGOLD MASŁO 200G	6.59	96.00
10	2024-03-21	00-199	PRODUKTY MACZNO ZBOZOWE	KASZA	284681	5906827003109	KASZA PĘCZAK KUJAWSKI 0,9KG MELVIT	5.09	13.00

Figure 3. Sample scanner data frame from a Polish supermarket

Processing scanner data poses a number of challenges, including the automatic classification of products into COICOP groups, matching products over time, data filtering, as well as the selection of a price index formula and the aggregation of partial results (e.g., over outlets). These processes are described in detail by Białek and Beręsewicz (2021). However, the issue of selecting a sample of products for determining a price index on the basis of scanner data is often overlooked. In practice, we can consider the *time dimension*, the *outlet dimension* and the *product dimension* when using CPI scanner data, and each of these aspects can play a measurable role in shaping the final price index (see our empirical study presented in Section 6). These above-mentioned dimensions are described in Section 5.1, while Section 5.2 discusses two main approaches in scanner sample selection. Section 5.3 describes the most popular multilateral indices that are considered in the empirical study.

Multilateral price indices designed for scanner data are much more complex than the bilateral indices discussed in Sections 4.1 and 4.2. Perhaps this is why the literature lacks theoretical results on population and sample-based multilateral indices that are analogous to the results presented in Table 4 and Table 5.

5.1. The time, outlet and product dimensions

The time dimension. According to Eurostat (2022, p. 10) we can read: "If all points in time during a certain period are equivalent to the consumer and there are no price level differences between weekdays and hours of the day, then the whole time period (month or week) can be considered as homogeneous for the purpose of price aggregation". It recommends aggregating data across a period that covers as much of the reference month as possible. In practice, however, statistical offices are limited by the terms of data transmission established with specific retail chains: for example, contracts may stipulate that the data are aggregated from the 5th to the 20th day of the month. A commonly used approach involves collecting scanner data that cover the first three weeks of sales from subsequent months of the retail chain's operations.

The outlet dimension. When working with CPI scanner data, its is generally recommended to specify individual products at the level of a single outlet. Retail chain often have different pricing policies in different outlets, depending on local conditions (e.g., demand for products or competitors' prices). However, determining the price index for each outlet separately is a time-consuming task. In some cases, there are reasons to aggregate scanner data across outlets, e.g., when the chain has an identical pricing policy within a specific region. This strategy can effectively reduce the computation time needed for multilateral price index calculations.

The product dimension. Typically, barcodes are used to identify products at the lowest level of aggregation, e.g., GTIN (Global Trade Item Number), EAN (European Article Number) or SKU (Stock Keeping Unit). However, the problem with disaggregated data is that over a longer period, we can observe *product churn*), i.e., a large number of products emerging and disappearing from the market. This means that the life cycle of a given product code may last a few months. The second problem observed at the bar-code level is identifying *relaunches*. Relaunches may occur when there are changes in the size or colour of the packaging. A change in size requires quality correction and price standardization, while the latter case does not affect product quality but may mean a change in its bar-code. Both scenarios should be detected automatically, which can be achieved by the procedure of matching products in time based not only on the bar-code, but also using the code assigned by the retail chain or the product description. The detection procedure (*data_matching*) is implemented in the *PriceIndices* R package (Białek, 2021).

If homogeneous product are defined too broadly, there is a risk of unit value bias. Conversly, operating at the bar-code level or defining homogeneous products too tightly may lead to problems with detecting relaunches (Eurostat, 2022). The MARS methods can be seen as a solution of this problem since it is a compromise between the above-mentioned two objectives (Chessa, 2021).

5.2. Static vs dynamic approach

There are two approaches that have emerged for using scanner data in the CPI measurement, i.e., *static* and *dynamic*. The *static* approach aligns with traditional data collection methods based on field surveys, while the *dynamic* approach uses the concept of monthly matched samples with the chain Jevons index as a target index.

In the *static* approach, a sample of items is selected at the beginning of each year and these items are monitored and maintained over time. Every month, prices for the selected products are taken from the scanner data files. Similar to the practices of price collectors from the field, if a particular item becomes unavailable, a replacement item is selected and used for further price index calculations.

Due to the high dynamics of scanner data related to product rotation and product seasonality, implementing a *dynamic* approach seems to be a better choice. This approach involves selecting the best-selling items available in two consecutive months each month to measure a monthly price changes. In practice, sample selection is carried out using the *cut-off method*, which is implemented by applying data filters.

The dynamic basket is determined using turnover figures of individual products in two adjacent months, i.e., the product is included in the sample if its turnover is above a fixed threshold determined by the number of products in a given product group. Van Loon and Roels (2018) provided the following condition for the above mentioned rule, which indicates whether the i-th product is taken into consideration when comparing months $t-1$ and t:

$$\frac{s_i^{t-1}+s_i^t}{2} > \frac{1}{n\lambda}, \tag{12}$$

where n is the number of considered products and λ is a fixed parameter (usually set to

1.25). This kind of data filter can be called a *low sale filter*. Proponents of using filters also believe that products displaying extreme price changes from one month to another should also be excluded from the sample (*extreme price filter*). For example, Statistics Poland uses the *extreme price filter* to remove products from the sample whose price has increased more than threefold or decreased more than fourfold. The list of possible data filters is extensive, e.g. Statistics Belgium implements a filter for dump prices (Van Loon and Roels, 2018). With this *dump price filter*, products are eliminated from the sample if a simultaneous, clear decrease in price and sales value is observed. These products will most likely be withdrawn from sale in the near future and, therefore, they are no longer representative.

Data filtering can also be considered when using multilateral indices, which are, in fact, specifically designed for scanner data cases (see Section 5.3). For instance, the *low sales filter* and *dump price filter* are mentioned as a part of *data pre-processing* serving as an initial step before computing multilateral price indices (see Eurostat (2022), p. 4). In particular, the aforementioned document recommends using *dumping filters* together with the CCDI multilateral index (p. 25). It seems that the same remark concerns the GEKS and GEKS-W price indices, since they give more weight to the price decrease of the dumped products (see Sections 5.3 and our *Empirical study*).

5.3. Multilateral indices

As it was mentioned above, multilateral indices are recommended for statistical offices to determine the dynamics of scanner prices (Eurostat, 2020). Commonly known and accepted methods include the GEKS method (Gini, 1931; Eltetö and Köves, 1964), the Geary-Khamis method (Geary, 1958; Khamis, 1972), the CCDI method (Caves et al., 1982), or the Time Product Dummy Methods (de Haan and Krsinich, 2018). Multilateral indices operate on a time window $[0,T]$ and therefore take into account phenomena such as product rotation or product seasonality. Moreover, due to the *transitivity* property, multilateral indices eliminate *chain drift bias* (Eurostat, 2022). The chain drift effect occurs when prices and quantities of products sold return to their original values (e.g., after the season) but the index deviates from the expected value of one. The most commonly used multilateral indices can be also found in Eurostat (2022).

6. Empirical study

This section examines the impact of scanner data sampling methods (under the *dynamic approach*) on the value of the multilateral price index. For this purpose, we will use the data filters discussed in Section 5.2 and the full-window multilateral price indices discussed in Section 5.3. The empirical study is based on the basis of scanner data collection on sales of *cleaning and preservatives* (COICOP: 056111) and *cosmetics and hygiene products* (COICOP: 121321) obtained from a Polish retail chain. The data covers the period: Dec, 2022 - Dec, 2023. The author of the study has not received receive permission to share these datasets, so the R script without the data is available for download from: https://github.com/JacekBialek/important_documents/blob/main/SIT_Empirical%20 study.Rmd.

In particular, the study will consider the following data filtering variants: (**v1**) data sets without filtering, (**v2**) the *low sales filter* (**f1**) used with $\lambda = 1.25$; (**v3**) the *extreme price filter* (**f2**) with thresholds: $lower = 0.25$ for price decrease and $upper = 3$ for price increase; (**v4**) the *dump price filter* (**f3**) with thresholds: $lower1 = 0.25$ for price decrease and $lower2 = 0.3$ for sales decrease; (**v5**) all data filters **{f1, f2, f3}** working independently; (**v6**) data filters implemented in order (**f1, f2, f3**); (**v7**) data filters implemented in order (**f1, f3, f2**); (**v8**) data filters implemented in order (**f2, f1, f3**); (**v9**) data filters implemented in order (**f2, f3, f1**); (**v10**) data filters implemented in order (**f3, f1, f2**) and (**v11**) data filters implemented in order (**f3, f2, f1**). The results concerning these variants - specifically regarding dataset reduction and its impact on multilateral price index levels - are presented in Tables 9 and 10. In particular, columns labelled *sample size* and *normalized sample size* describe the number of different products after applying the given type of filter, with the first row in these tables indicating the situation with no filtering.

Table 9. Different variants of data filtering and their impact on sample size and multilateral index values (*cleaning and preservatives*)

Filter variant	Sample size	Normalized sample size	Chain Jevons	Geary-Khamis	GEKS	CCDI	TPD
v1	2078	100	1.05733	1.14268	1.13548	1.13480	1.14140
v2	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v3	1914	92.11	1.05452	1.14287	1.13465	1.13404	1.14233
v4	1915	92.16	1.05733	1.14287	1.13463	1.13399	1.14233
v5	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v6	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v7	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v8	903	43.45	1.11527	1.15401	1.14328	1.14414	1.15401
v9	903	43.45	1.11527	1.15401	1.14328	1.14414	1.15401
v10	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15405
v11	903	43.45	1.11527	1.15401	1.14328	1.14414	1.15401

Table 10. Different variants of data filtering and their impact on sample size and multilateral index values (*cosmetics and hygiene products*)

Filter variant	Sample size	Normalized sample size	Chain Jevons	Geary-Khamis	GEKS	CCDI	TPD
v1	5966	100	0.97274	1.09909	1.09594	1.09465	1.09829
v2	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v3	5393	90.39	0.96662	1.09945	1.09649	1.09540	1.09830
v4	5395	90.43	0.97458	1.09949	1.09654	1.09544	1.09743
v5	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v6	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v7	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v8	1994	33.42	1.08567	1.11877	1.11542	1.11598	1.11984
v9	1994	33.42	1.08567	1.11877	1.11542	1.11598	1.11984
v10	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v11	1994	33.42	1.08567	1.11877	1.11542	1.11598	1.11984

7. Conclusions

The ultimate aim of CPI sampling techniques is to obtain the most accurate estimate of inflation. A general conclusion from the empirical illustrations presented is that the sample structure depends not only strongly on the sampling technique adopted (particularly on the choice between random and non-random sampling), but also on the level of data aggregation (see the empirical illustration in Section 3.5). The considered sampling technique may turn out to be better than other techniques at COICOP level 5, but worse at COICOP level 6. Further, if we take the bias of the final estimated price index as an evaluation criterion, it may turn out that the considered method performs better or worse depending on whether we estimate a weighted or unweighted index. For instance, in Section 4.3 we found that *cut-off sampling* works much better than *simple random sampling* and *pps sampling* for estimating weighted population price indices.

An important practical conclusion of the empirical study (see Section 6) is that the *low sales filter* has the greatest impact on reducing the size of the scanner dataset. In both analyzed scanner datasets, the product sample size was reduced by more than 55% after applying this filter. In contrast, the other two types of data filters (i.e., the *extreme price filter* and the *dump price filter*) reduced the sample size in a similar yet smaller way (by less than 10%), although they substantially affected the price index value. We can also conclude that the order in which the scanner data filters are applied has no effect on either the sample structure or the value of the price index (see Tables 9 and 10). In other words, changing the order of data filtering has little impact on the value of the price index. As a consequence, each of the filters can be applied independently.

Finally, as expected, the chain Jevons index proved to be much more sensitive to the choice of the data filter than multilateral indices. It is important to note that data filtering is essential if a statistical office intends to use the chain Jevons index as part of a *dynamic approach*. With weighted multilateral indices, while data filtering may not seem necessary, it can effectively reduce the sample size and, thus, the time needed to estimate the index.

References

Balk, B. M., (2005). Price indexes for elementary aggregates: the sampling approach. *Journal of Official Statistics*, 4(21), pp. 675–679.

Bialek, J., (2020). Comparison of elementary price indices. *Communications in Statistics - Theory and Methods*, 49(19), pp. 4787–4803.

Białek, J., (2021). PriceIndices – a new R package for bilateral and multilateral price index calculations. *Statistika – Statistics and Economy Journal*, 36(2), pp. 122–141.

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

Bialek, J., (2022). Elementary price indices under the GBM price model. *Communications in Statistics - Theory and Methods*, 51(5), pp. 1232–1251.

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(365), pp. 73–86.

Chessa, A. G., (2021). A Product Match Adjusted R Squared Method for Defining Products with Transaction Data. *Journal of Official Statistics*, 37(2), pp. 411–432.

Cochran, W. G., (1977). *Sampling techniques*. New York: John Wiley.

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), pp. 757–776.

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

Diewert, W. E., (2004). *On the Stochastic Approach to Linking the Regions in the CPI*. Department of Economics, University of British Columbia.

Diewert, W. E., Fox, K. J., (2018). *Substitution bias in multilateral methods for CPI construction using scanner data*. UNSW Business School Research Paper (2018–13).

Dorfman, A. H., Leaver, S. and Lent, J., (1999). *Some observations on price index estimators*. Bureau of Labor Statistics working paper no. 324, Washington DC.

de Haan, J., Opperdoes, E. and Schut, C. ,(1999). *Item Sampling in the Consumer Price Index: A Case Study using Scanner Data*. Research Report, Statistics Netherlands, Voorburg.

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

Eurostat, (2018). *Harmonised Index of Consumer Prices (HICP) Methodological Manual*. Luxembourg: Publications Office of the European Union.

Eurostat, (2022). *Guide on Multilateral Methods in the Harmonised Index of Consumer Prices*. Luxembourg: Publications Office of the European Union.

Fisher, I., (1922). The making of index numbers: a study of their varieties, tests, and reliability. *Number 1. Houghton Mifflin*.

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), pp. 97–99.

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

International Labour Office, (2004). *Consumer price index manual: Theory and practice*, Geneva.

International Monetary Fund, (2020). *Consumer Price Index manual: Concepts and methods*. Washington, D.C.

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), pp. 96–121.

Levell, P., (2015). Is the Carli index flawed?: assessing the case for the new retail price index RPIJ. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(2), pp. 303–336.

Lindblom, A., Teterukovsky, A., (2007). *Coordination of Stratified Pareto πps Samples and Stratified Simple Random Samples at Statistics Sweden.* Papers presented at the ICES-III, Montreal, Quebec, Canada.

Laspeyres, K., (1871). Ix. Die berechnung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, 16(1), pp. 296–318.

Lindblom, A., (2003). AMU - The system for coordination of frame populations and samples from the Business Register at Statistics Sweden, *Background Facts on Economic Statistics*, 2003:3, Statistics Sweden.

McClelland, R., Reinsdorf. M., (1999). *Small sample bias in geometric mean and seasoned CPI component indexes.* Bureau of Labor Statistics working paper no. 324, Washington DC.

Paasche, H., (1874). Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, pp. 168–178.

Rosén, B., (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62(2), . 135–158.

Rosén, B., (1997b). On sampling with Proportionality Proportional to Size. *Journal of Statistical Planning and Inference*, 62(2), pp. 159–191.

Särndal, C.E., Swensson, B. and Wretman, J., (2003). *Model Assisted Survey Sampling.* Springer Science and Business Media, Berlin, Heidelberg.

Silver, M., Heravi, S., (2007). Why elementary price index number formulas differ: Evidence on price dispersion. *Journal of Econometrics*, 140(2), pp. 874–883.

Törnqvist, L., (1936). *The bank of Finland's consumption price index.* Bank of Finland Monthly Bulletin, pp. 1–8.

van Loon, K. V. Roels, D., (2018). *Integrating big data in the Belgian CPI.* Paper presented at the meeting of the group of experts on consumer price indices, 8-9 May 2018, Geneva, Switzerland.

von der Lippe, P., (2007). *Index Theory and Price Statistics.* Peter Lang, Berlin, Germany.

Walsh, C. M., (1901). *The Measurement of General Exchange Value.* Macmillan and Co.