What impact does product specification have on a Fisher price index?

Claude Lamboray*

Paper prepared for the 17th Meeting of the Ottawa Group on Price Indices, 7-10 June 2022, Rome, Italy

1 Introduction

The use of scanner data in a Consumer Price Index (CPI) gives rise to an aggregation problem that can be subdivided into three stages.

1. In the first stage, individual transactions are combined into an individual product for which an average price and a total quantity sold can be calculated.
2. In the second stage, the previously specified individual products are further aggregated using, for example, a multilateral method, in order to obtain an elementary price index.
3. In the third stage, the previously calculated elementary price indices are further combined with other price indices in order to obtain price indices for the higher-level aggregates in the CPI.

The second stage has received a lot of attention from a theoretical and practical point of view. There are many studies on the properties, results and implementation issues of different bilateral and multilateral index formulas. Very often these studies rely on the assumption that the individual products together with their prices and quantities are given. However, any index compilation method applied in the second stage is conditional on the product specification applied in the first stage. This will be the focus of this paper.

Product specification has been recognized as a critical step that could jeopardize any gains in bias reduction that we would typically expect from using scanner data. While scanner data

*Eurostat, Unit C4 (Price statistics. Purchasing Power Parities. Housing statistics). The views expressed in this paper are those of the author and do not necessarily reflect the views of the European Commission (Eurostat). The data set included in Annex C was used by A. Chessa in a training session on the MARS method that was conducted during the 2018 Eurostat Workshop on Scanner Data organized by Statistics Norway in Oslo. We are thankful for the permission to reproduce this example in this paper.

helps reducing lower-level substitution bias, other biases can appear because products are specified too tightly or too broadly (see section 3.1.2 in European Central Bank, 2021 [9]). Under certain pricing strategies, products may enter or exit with unusually high or low prices. It is known that this creates biases in traditional matched model methods (see Eurostat, 2021 [10]), and these biases propagate to scanner data and multilateral methods (see Melser and Webster, 2021 [15]). Konny et al., 2019 [14] stresses that multilateral index methods do solve the problem of chain drift but they are not fully satisfactory to cope with life-cycle pricing. For example, the multilateral methods do not solve the downward drift in the price index for new vehicles caused by the downward price trends of a given model year (see Williams et al., 2019 [19]).

Technically, tightly specified products may cause a bias as new and disappearing products in the two comparison periods are not taken into account in a matched price index. Broadly specified products may cause a bias as the underlying transactions that make up the product may not be of the same quality. This trade-off has been referred to as assignment bias versus assortment bias (Von Auer, 2017 [16]). This trade-off between homogeneity and stability over time has also been highlighted by Chessa, 2019 [3] who developed the MARS method as an operational tool for finding a compromise between these two objectives. In this paper, we examine this trade-off in the framework of a Fisher price index and extend the analysis defined in the bilateral context to the multilateral context.

This paper is organized as follows. In section 2, we first discuss the dimensions of product specification. In section 3, we formalize the problem from an index number point of view by considering *matched*, *hybrid* and *imputation* price indices. In section 4, we compare these three types of indices in order to assess the matched-model bias and the unit value bias. The analysis is illustrated on three data sets in section 5. Finally, we draw some conclusions in section 6.

2 Individual product specification

In order to calculate a price index with scanner data, it is necessary to specify beforehand the individual product (see chapter 3 in Eurostat, 2022 [11]). Conceptually, a single transaction specified by the product characteristics, the timing and place of purchase and the terms of supply is the most granular unit for which a price can be observed. In practice, we do not work with single transactions, but with individual products. The individual product is the statistical unit which is tracked over time and which corresponds to the input of, for example, a multilateral method. When specifying individual products, one needs to consider the time, outlet and product dimensions, as shown in figure 1. An average price (unit value) is calculated over days or weeks of the reference period, over outlets and possibly over item codes. An earlier discussion on this topic can be found in Dalèn, 2017 [4].

The individual product can be defined at any level of these successive aggregations. It may

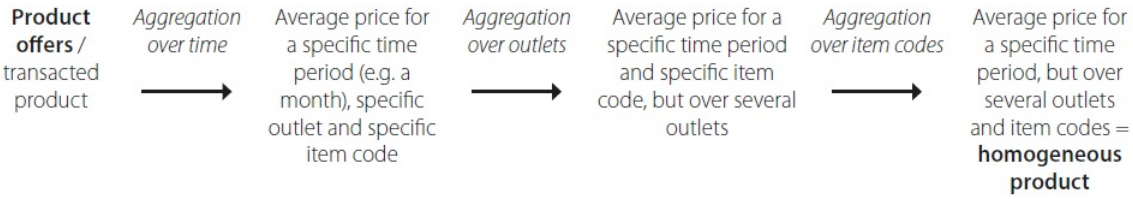


Figure 1: From a transacted product to homogeneous product

be defined in very narrow terms, referring for example to an item code in an outlet for a given time period. Alternatively, it can be defined in broader terms, for example comprising several item codes sold in several outlets for a given time period. The specification of the individual product is a critical step which can have a significant impact on the final index.

The main idea of creating broader individual products is to increase the matching over time. The number of individual products that will be taken into account in the index compilation will decrease when more of the data are grouped together. At the same time, there is a limit to this strategy. In principle, transacted products can only be combined as long as there are no significant quality differences between them. Quality differences must be evaluated with respect to the already mentioned time, outlet and product dimensions. This is the main trade-off which is examined in this paper.

Technically the order of aggregation (first over time, then over outlets and finally over item codes) does not matter. However, we discuss the three dimensions in this order as the decision to combine could be increasingly challenged.

The treatment of the **time dimension** is the least controversial. In general it is appropriate to calculate a unit value when an item is sold at different prices at different times within the same month[1]. Ideally the average price should cover as much as possible of the reference month. Diewert, Fox and de Haan, 2016 [8] showed that aggregation over only one week of the month can be upward biased compared to aggregation over the full month. Production calendar constraints explain that in practice, very often only the first two or three weeks of the reference month are used in a CPI.

The treatment of the **outlet dimension** depends on circumstances. The individual product could be specified at the most detailed outlet level available in the data. Quality differences between outlets can be associated with different opening hours, different assortments, etc. Aggregation across outlets can be envisaged if data is only supplied at a more aggregated level or

[1]For some products, such as accommodation or transport services, the timing is an important quality dimension. Traveling on a Friday evening may be considered as a different product (i.e. a different quality) from traveling on a Wednesday afternoon. The price may also depend on the moment of purchase. In such a context, one could argue to treat differences in the time of supply of the service, and possibly differences in the time of booking of the service, as differences in quality.

if for example price levels are similar in the outlets (for example of the same type and chain). The impact of the outlet dimension can be empirically assessed (see Ivancic and Fox, 2013[13]). The outlet dimension is also examined in Azaircabe, 2022 [1]. This paper looks at unit value aggregation over different providers that offer on the one hand ride sharing services, and on the other hand taxi ride services.

The treatment of the **product dimension** is the most controversial one. The scanner data usually includes an item code such as a Global Trade Identification Number (GTIN) or a slightly broader Stock-Keeping Unit (SKU) code. In general, there is some product churn, meaning that the set of item codes is not stable over time. There are different strategies that can be used to cope with a dynamic product universe. In this paper, we consider the following three strategies.

1. *Matching.* In many cases, it could be satisfactory to define the individual product at the GTIN or SKU level. With such a strategy, item codes are taken into account if they are available in two comparison periods.

2. *Grouping* The item codes are combined into broader products, thereby reducing the lack of matching across time. However, this may create other problems as item codes may be grouped together which are not of the same quality.

3. *Imputation* In order to take into account the item codes that are not available in the two comparison periods, a price is explicitly imputed for these products in the periods in which they are not available. This allows then to estimate a price change for these unmatched item codes.

In principle, the matching approach is the easiest approach to apply. It only requires an appropriate product identifier whereas some kind of product characteristics are usually needed in order to group items together or to estimate a price of a missing item.

The grouping approach is relatively easy to explain. Sometimes the supplied data is already grouped and a more disaggregated approach is not possible. However, often, the supplied data can be grouped in various ways. This will make it possible to compare the index obtained from the item codes with the index obtained after grouping some of the item codes. It leads to the practical question of which one of the two approaches is most suitable. Grouping has been proposed by Chessa as a basis for processing scanner data in the Dutch CPI ([2]).

Finally, the imputation approach can be considered as valid from an index number perspective. The practical challenge with this approach is that we need a method to estimate the missing prices. In this paper, we will use the imputation approach as a benchmark and compare it to the other two approaches.

	p_{t-1}	p_t	Price change
Item A	25	-	
Item B	-	58	
Item C	40	42	+5.0%
Item D	30	33	+10.0%
Item E	20	23	+15.0%
Geo. avg. price	28.8	31.7	+9.9%

Table 1: Matching

Special attention should be given to situations where the pricing strategy depends on the life-cycle of a product. For example, it may happen that the last available price of the an item is a reduced price. This situation can be encountered at the end of a sales period and is especially common in clothing and footwear. Reduced prices can also be observed in situations of inventory clearing or closure of an outlet.

The impact of a reduced last price is best explained by an example[2]. Suppose that an item A is available up to period $t-1$, while its successor item B is available from period t onwards. We suppose that the two items are of similar quality. The other items C, D and E are available in the two comparison periods. There are different outcomes:

- Under the *matching* strategy (see Table 1), the price change between items A and B is not taken account, but only the price changes of items C, D and E will be used in the calculations.
- Under the *grouping* strategy (see Table 2), we would combine item codes A and B into a single homogenous product and thereby capturing the price change between these two item codes.
- Under the *imputation* strategy (see Table 3), we would impute a price for item A in period t, or alternatively impute a price for item B in period $t-1$. The imputations are based on linking the two items A and B. This means that the imputed price for item A is equal to the observed price of item B in that period. Similarly, the imputed price for item B in period $t-1$ is equal to the observed price of item A in that period.

This example shows that with clearance prices, the matching strategy will lead to an index that is downward biased compared to an index calculated with a grouping or imputing strategy. Note that the direction of the bias depends on the pricing strategy[3].

[2]This example is taken from Eurostat, 2021 [10] which includes recommendations for the Bridged Overlap method.

[3]For example, if the first observed price of a new item is unusually high, the index obtained with a matching strategy will have an upward bias. This may happen at the beginning of the life cycle of a product under a price skimming strategy (setting a high initial price that a subset of customers is willing to pay in order to maximise profit (see Eurostat, 2021 [10]).

	p_{t-1}	p_t	Price change
Hom. prod. (Items A and B)	25	58	+132.0 %
Item C	40	42	+5.0%
Item D	30	33	+10.0%
Item E	20	23	+15.0%
Geo. avg. price	27.7	36.9	+32.5%

Table 2: Grouping

	p_{t-1}	p_t	Price change
Item A	25	*58 (imp.)*	+132.0 %
Item C	40	42	+5.0%
Item D	30	33	+10.0%
Item E	20	23	+15.0%
Geo. avg. price	27.7	36.9	+32.5%

	p_{t-1}	p_t	Price change
Item B	*25 (imp.)*	58	+132.0 %
Item C	40	42	+5.0%
Item D	30	33	+10.0%
Item E	20	23	+15.0%
Geo. avg. price	27.7	36.9	+32.5%

Table 3: Imputation

The strategies discussed in this paper can also be related to product descriptions that are commonly used in a CPI in order to make sampling, replacement and quality adjustment decisions. Some exploratory work has been recently conducted on defining product characteristics for mobile phones in the Harmonised Index of Consumer Prices (HICP). Instead of simply matching mobile phone models across time (matching strategy), the main price-determining product characteristics such as operating system, processor speed, memory, and screen size could be used for imputing the missing prices (imputation strategy). Another option would be to group together mobile phone models that have the same values for these main price-determining product characteristics (grouping strategy).

In the next section we are going to formalize the three strategies in the context of a Fisher index which is based on both prices and quantities and investigate how much the three strategies can differ from each other.

3 Matched, imputation and hybrid indices

3.1 Definitions

If both prices and quantities are available, a price index formula should be used that relies on the weights in the two comparison periods. In this study we will focus on a Fisher index. A Fisher index has good axiomatic properties and is consistent with a basket approach. In fact, it is defined as an average of two basket indices that rely on either base or current period quantities. Finally, from an analytical point of view, the Fisher index can be more easily related to and combined with unit values, which will be a key element in the analysis. The Fisher index is also the basis for some of the multilateral methods, such as GEKS.

As a starting point, we calculate a Fisher index on the matched item codes. This means that the aggregate price change is only derived from the set of items that are available in the two comparison periods. Let N_t be set of items available in period t $(t = 0, 1)$. The set of items in the two comparison periods 0 and 1 is denoted by $M_{01} = N_0 \cap N_1$. Moreover, we denote by p_i^t and q_i^t the price, and the quantity of the item i in period t $(t = 0, 1)$. The matched Laspeyres, Paasche and Fisher indices between periods 0 and 1 are defined as follows:

$$P_{ML}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 q_i^0}{\sum_{i \in M_{01}} p_i^0 q_i^0} \tag{1}$$

$$P_{MP}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 q_i^1}{\sum_{i \in M_{01}} p_i^0 q_i^1} \tag{2}$$

$$P_{MF}^{01} = \sqrt{P_{ML}^{01} P_{MP}^{01}} \tag{3}$$

One issue with the matched Fisher index is that items that are available in only one of two

comparison periods are ignored. In order to overcome this limitation, we could estimate a price for an item in the period in which it is not available. Formally, let $N_{01} = N_1 \setminus N_0$ be the set of items available in period 1 but not in period 0. Moreover, let $D_{01} = N_0 \setminus N_1$ be the set of items available in period 0 but not in period 1. We denote by $\hat{p}_i^t$ an estimated (i.e. not observed) price of an item in period t ($t = 0, 1$). The imputation Laspeyres, Paasche and Fisher indices between periods 0 and 1 are defined as follows (see de Haan, 2001 [5]):

$$P_{IL}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 q_i^0 + \sum_{i \in D_{01}} \hat{p}_i^1 q_i^0}{\sum_{i \in M_{01}} p_i^0 q_i^0 + \sum_{i \in D_{01}} p_i^0 q_i^0} \tag{4}$$

$$P_{IP}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 q_i^1 + \sum_{i \in N_{01}} p_i^1 q_i^1}{\sum_{i \in M_{01}} p_i^0 q_i^1 + \sum_{i \in N_{01}} \hat{p}_i^0 q_i^1} \tag{5}$$

$$P_{IF}^{01} = \sqrt{P_{IL}^{01} P_{IP}^{01}} \tag{6}$$

The imputation indices solve the lack of matching from which matched indices may suffer but it requires an estimation of the prices. An alternative strategy to increase the matching would be to first combine the initial items. The idea is to group similar items together and create broader individual products.

Formally, let H_k be set of items that belong to the broader product group k. The average price and total quantity of the product H_k in period t ($t = 0, 1$) can be derived from the initial data of items as follows:

$$\bar{p}_k^t = \frac{\sum_{i \in H_k} p_i^t q_i^t}{\sum_{i \in H_k} q_i^t} \tag{7}$$

$$Q_k^t = \sum_{i \in H_k} q_i^t \tag{8}$$

The index formula is then applied to these broader individual products, instead of applying it to the initial, tighter, individual products. Following the terminology used in Diewert, 2010 [7], we will refer to this as hybrid indices. The hybrid Laspeyres, Paasche and Fisher indices are defined as follows[4]:

$$P_{HL}^{01} = \frac{\sum_k \bar{p}_k^1 Q_k^0}{\sum_k \bar{p}_k^0 Q_k^0} \tag{9}$$

$$P_{HP}^{01} = \frac{\sum_k \bar{p}_k^1 Q_k^1}{\sum_k \bar{p}_k^0 Q_k^1} \tag{10}$$

$$P_{HF}^{01} = \sqrt{P_{HL}^{01} P_{HP}^{01}} \tag{11}$$

[4]Technically, these hybrid indices are defined for 'matched' product groups that are available in the two comparison periods.

3.2 Imputation method

In order to calculate an imputation index, a price must be estimated for the items that are only available in one of the two comparison periods. Recall that the missing item i could be grouped together with other, similar, items. Let $\kappa(i)$ be the group to which item i belongs ($i \in H_{\kappa(i)}$).The price of the missing item i is then set equal to the average price of $H_{\kappa(i)}$ in that period.

$$\hat{p}_i^t = \bar{p}_{\kappa(i)}^t \tag{12}$$

The quantity for the missing item is zero as no purchase took place for item i in that period.

This imputation method makes use of the information that some items can be grouped together and can therefore be assumed to be relatively 'similar'. From a practical point of view, no additional information is needed apart from the assignment of the items into groups[5].

This imputation method can also be formalized with a regression in which the dependent variable is the price and the independent variables are dummy variables for the groups. Formally, let G_{ki} be a dummy variable that is set to 1 if the item i belongs to group k, and that is set to 0 otherwise. Consider the following model to be estimated in period t.

$$p_i^t = \alpha + \sum_{k \neq 1} \beta_k G_{ki} + \epsilon_{it} \qquad \forall i \in N_t \tag{13}$$

If each item i in this regression is weighted by its quantity q_i, it can be shown that the estimated price $\hat{p}_i^t$ for an item obtained from model 13 corresponds to the average price defined in equation 12.

We can also relate the hybrid index to this model. Recall that in the imputation index, only missing prices are imputed. This is sometimes referred to as a *single* imputation approach. We now consider a *full* imputation index in which prices for all items are estimated.

$$P_{FIL}^{01} = \frac{\sum_{i \in M_{01} \cup D_{01}} \hat{p}_i^1 q_i^0}{\sum_{i \in M_{01} \cup D_{01}} \hat{p}_i^0 q_i^0} \tag{14}$$

$$P_{FIP}^{01} = \frac{\sum_{i \in M_{01} \cup N_{01}} \hat{p}_i^1 q_i^1}{\sum_{i \in M_{01} \cup N_{01}} \hat{p}_i^0 q_i^1} \tag{15}$$

$$P_{FIF}^{01} = \sqrt{P_{FIL}^{01} P_{FIP}^{01}} \tag{16}$$

If prices are estimated with model 13, it is straightforward to check that $P_{FIL}^{01} = P_{HL}^{01}$ and

[5]If $H_{\kappa(i)}$ is empty in period t, then no price can be imputed with this method for the missing item i in that period. In fact, in such circumstances, the item i will not be included in the matched index (because it is missing in one period), nor in the imputation index (because a price cannot be estimated in the missing period), nor in the hybrid index (because the group $\kappa(i)$ is missing in one period).

that $P_{FIP}^{01} = P_{HP}^{01}$, and that $P_{FIF}^{01} = P_{HF}^{01}$. In other words, the hybrid Fisher index can be seen as a full imputation index whereas the imputation Fisher index can be seen as a single imputation index, assuming that the prices are estimated with model 13.

3.3 Decompositions

In order to compare the matched, imputation and hybrid indices, we introduce additional notations. We denote by s_i^t the quantity share in period t of item i within its grouping $\kappa(i)$ $(t = 0, 1)$.

$$s_i^t = \frac{q_i^t}{Q_{\kappa(i)}^t} \tag{17}$$

Moreover, we denote by σ_i^t the quantity share in period t of grouping $\kappa(i)$ to which the item i belongs $(t = 0, 1)$. Note that σ_i^t is the same for all items that belong to the same grouping.

$$\sigma_i^t = \frac{Q_{\kappa(i)}^t}{\sum_k Q_k^t} \qquad \forall i \in N_t \tag{18}$$

Following the result presented in de Haan, 2001 [5], the difference between an imputation Fisher index and a matched Fisher index can be decomposed according to the impact of new and disappearing items (see Annex A):

$$\frac{P_{IF}^{01}}{P_{MF}^{01}} = \frac{\nu_{11}^1}{\nu_{11}^0} \cdot \frac{\delta_{00}^1}{\delta_{00}^0} \tag{19}$$

The terms ν_{11}^1 and ν_{11}^0 refer to the impact of new items and are defined as follows:

$$\nu_{11}^1 = \left(1 + \frac{\sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^1}\right)^{0.5} \tag{20}$$

$$\nu_{11}^0 = \left(1 + \frac{\sum_{i \in N_{01}} \hat{p}_i^0 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^1 \sigma_i^1}\right)^{0.5} \tag{21}$$

The terms δ_{00}^1 and δ_{00}^0 refer to the impact of disappearing items and are defined as follows:

$$\delta_{00}^1 = \left(1 + \frac{\sum_{i \in D_{01}} \hat{p}_i^1 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0}\right)^{0.5} \tag{22}$$

$$\delta_{00}^0 = \left(1 + \frac{\sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^0}\right)^{0.5} \tag{23}$$

If the set of new and disappearing items is small, then all of these terms would be close to 1. As a consequence, there would be no big difference between the matched and the imputation Fisher index. More interestingly, the equation can also be examined from the perspective of the life-cycle of a product. Suppose that there are no new items but only disappearing items.

For the disappearing items, it can be shown that:

$$\frac{\delta_{00}^{1}}{\delta_{00}^{0}} > 1 \iff \frac{\sum_{i \in D_{01}} \hat{p}_i^1 s_i^0 \sigma_i^0}{\sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^0} > \frac{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^0} \tag{24}$$

Suppose now that an item is sold at a low (exit) price in period 0 and disappears in period 1. It is therefore likely that $\hat{p}_i^1 >> p_i^0$,and as a consequence, the right-hand inequality in 24 is likely to hold and therefore $\frac{\delta_{00}^{1}}{\delta_{00}^{0}} > 1$. This implies that P_{IF}^{01} will be larger than P_{MF}^{01} (i.e. the matched index has a downward bias).

Similarly, the difference between an hybrid Fisher index and an imputation Fisher index can be decomposed according to the impact of the matched, new and disappearing items (see Annex B).

$$\frac{P_{HF}^{01}}{P_{IF}^{01}} = \mu \cdot \nu_{10}^{1} \nu_{11}^{0} \cdot \frac{1}{\delta_{00}^{1} \delta_{01}^{0}} \tag{25}$$

The term ν_{11}^{0} is defined in equation 21. The term δ_{00}^{1} is defined in equation 22. The remaining terms are defined as follows:

$$\mu = \left(\frac{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0} \right)^{0.5} \times \left(\frac{\sum_{i \in M_{01}} p_i^0 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^1} \right)^{0.5} \tag{26}$$

$$\nu_{10}^{1} = \left(1 + \frac{\sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0} \right)^{0.5} \tag{27}$$

$$\delta_{01}^{0} = \left(1 + \frac{\sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^1} \right)^{0.5} \tag{28}$$

Suppose that there are no new or disappearing items, so that all factors except μ are equal to one. As a consequence, the matched and imputation Fisher indices are the same. Moreover, suppose that the items that make up a grouping have the same price in each period, which is therefore identical to the average price for that group. Under such circumstances, it can be shown that $\mu = 1$, and hence the hybrid index will be identical to the matched index. This could be a theoretical justification for grouping together items based on similar price levels.

We now consider the following decomposition that brings together the matched, imputation and hybrid Fisher index:

$$\frac{P_{HF}^{01}}{P_{MF}^{01}} = \frac{P_{IF}^{01}}{P_{MF}^{01}} \cdot \frac{P_{HF}^{01}}{P_{IF}^{01}} \tag{29}$$

We first decomposed the difference between the matched and imputation Fisher index according to the impact of new and disappearing items (equation 19). We then decomposed the he difference between the imputation and hybrid Fisher index into the impact, of matched, new and disappearing items (equation 25). It is now possible to combine both decompositions

	$\frac{P_{HF}^{01}}{P_{MF}^{01}}$	$=$	$\frac{P_{IF}^{01}}{P_{MF}^{01}}$	$\times$	$\frac{P_{HF}^{01}}{P_{IF}^{01}}$
	$\|\|$		$\|\|$		$\|\|$
	$\mu\frac{\nu_{11}^1\nu_{10}^1}{\delta_{00}^0\delta_{01}^0}$	$=$	$\frac{\nu_{11}^1\delta_{00}^1}{\nu_{11}^0\delta_{00}^0}$	$\times$	$\mu\frac{\nu_{10}^1\nu_{11}^0}{\delta_{00}^1\delta_{01}^0}$
	$\|\|$		$\|\|$		$\|\|$
Matched	μ	$=$	1	$\times$	μ
	$\times$		$\times$		$\times$
New	$\nu_{11}^1\nu_{10}^1$	$=$	$\frac{\nu_{11}^1}{\nu_{11}^0}$	$\times$	$\nu_{10}^1\nu_{11}^0$
	$\times$		$\times$		$\times$
Disappearing	$\frac{1}{\delta_{00}^0\delta_{01}^0}$	$=$	$\frac{\delta_{00}^1}{\delta_{00}^0}$	$\times$	$\frac{1}{\delta_{00}^1\delta_{01}^0}$

Table 4: Decomposition according to matched, new and disappearing items.

with equation 29, as shown in table 4. The price change between an hybrid and matched index is decomposed into a price change between an imputation and a matched index and a price change between an hybrid and an imputation index. Each of these components can be further decomposed into an impact stemming from the matched, from the new and from the disappearing items. This decomposition depends on the imputed prices. In other words, a different imputation method for the missing prices leads to a different result for this decomposition.

3.4 Quality adjusted unit value indices

A quality adjusted unit value index is defined as follows.

$$P_{QU}^{01} = \frac{\sum_{i\in N_1} p_i^1 q_i^1}{\sum_{i\in N_1} v_i q_i^1} / \frac{\sum_{i\in N_0} p_i^0 q_i^0}{\sum_{i\in N_0} v_i q_i^0} \tag{30}$$

The v_i factors are quality adjustment factors. There are different ways to obtain these factors (see for example Von Auer, 2014 [18]). In the Geary-Khamis method, these factors are defined as the average deflated price over a set of periods.

Note that the QU index is based on the tight product specification. The QU index cannot accommodate imputed prices as items with no quantities are zeroed out. However, a 'hybrid' variant of the QU index can be defined. This index is based on the average price and aggregated quantities as defined in 7 and 8.

$$P_{QU-H}^{01} = \frac{\sum_k \bar{p}_k^1 Q_k^1}{\sum_k \bar{v}_k Q_k^1} / \frac{\sum_k \bar{p}_k^0 Q_k^0}{\sum_k \bar{v}_k Q_k^0} \tag{31}$$

Note that the hybrid variant must have some quality adjustment factors $\bar{v_k}$ for each product grouping k. It can be shown that the hybrid variant compares to the initial variant as follows:

$$\frac{P_{QU-H}^{01}}{P_{QU}^{01}} = \frac{\sum_k \bar{v}_k Q_k^1}{\sum_{i \in N_1} v_i q_i^1} / \frac{\sum_k \bar{v}_k Q_k^0}{\sum_{i \in N_0} v_i q_i^0} \tag{32}$$

Suppose that the adjustment factors v_i for all the items that belong to the group k are the same and are identical to the adjustment factor $\bar{v}_k$ of that group. In fact, a group should be composed of items of the same quality and therefore it could make sense to assume that these items have the same quality adjustment factors. In that case, it follows from equation 32 that P_{QU-H}^{01} and P_{QU}^{01} are equivalent.

There is no imputation variant for the QU index. It can be shown that QU index compares to the imputation Fisher index as follows.

$$\begin{aligned}\frac{P_{QU}^{01}}{P_{IF}^{01}} = &\left(\frac{\left(\sum_{i \in M_{01}} p_i^1 q_i^1 + \sum_{i \in N_{01}} p_i^1 q_i^1\right)^{0.5} \left(\sum_{i \in M_{01}} p_i^0 q_i^1 + \sum_{i \in N_{01}} \hat{p}_i^0 q_i^1\right)^{0.5}}{\sum_{i \in M_{01} \cup N_{01}} v_i q_i^1} \right) \\ &/ \left(\frac{\left(\sum_{i \in M_{01}} p_i^1 q_i^0 + \sum_{i \in D_{01}} \hat{p}_i^1 q_i^0\right)^{0.5} \left(\sum_{i \in M_{01}} p_i^0 q_i^0 + \sum_{i \in D_{01}} p_i^0 q_i^0\right)^{0.5}}{\sum_{i \in M_{01} \cup D_{01}} v_i q_i^0} \right)\end{aligned} \tag{33}$$

Let us suppose that the quality-adjustment factors are defined as the average of the observed or estimated price in the two comparison periods.

$$v_i = 0.5 * (p_i^0 + p_i^1) \qquad \forall i \in M_{01} \tag{34}$$

$$v_i = 0.5 * (p_i^0 + \hat{p}_i^1) \qquad \forall i \in D_{01} \tag{35}$$

$$v_i = 0.5 * (\hat{p}_i^0 + p_i^1) \qquad \forall i \in N_{01} \tag{36}$$

It has been noted by Von Auer [18] that a QU index with such factors is equivalent to the Banerjee index. With such quality adjustment factors, the two terms on the right-hand side of equation 33 may approximate unity: in each fraction, the nominator is a geometric average and the denominator is an arithmetic average of the same two terms.

Equation 33 can be rearranged in the following way by separating the impact of matched, new and disappearing items:

$$\begin{aligned}\frac{P_{QU}^{01}}{P_{IF}^{01}} &= \left(\left(\sum_{i\in M_{01}} p_i^1 q_i^1\right)^{0.5}\left(\sum_{i\in M_{01}} p_i^0 q_i^1\right)^{0.5}\right) / \left(\sum_{i\in M_{01}} v_i q_i^1\right) \\ &\times \left(\sum_{i\in M_{01}} v_i q_i^0\right) / \left(\left(\sum_{i\in M_{01}} p_i^1 q_0^1\right)^{0.5}\left(\sum_{i\in M_{01}} p_i^0 q_i^0\right)^{0.5}\right) \\ &\times \left(\left(1+\frac{\sum_{i\in N_{01}} p_i^1 q_i^1}{\sum_{i\in M_{01}} p_i^1 q_i^1}\right)^{0.5}\left(1+\frac{\sum_{i\in N_{01}} \hat{p}_i^0 q_i^1}{\sum_{i\in M_{01}} p_i^0 q_i^1}\right)^{0.5}\right) / \left(1+\frac{\sum_{i\in N_{01}} v_i q_i^1}{\sum_{i\in M_{01}} v_i q_i^1}\right) \\ &\times \left(1+\frac{\sum_{i\in D_{01}} v_i q_i^0}{\sum_{i\in M_{01}} v_i q_i^0}\right) / \left(\left(1+\frac{\sum_{i\in D_{01}} p_i^0 q_i^0}{\sum_{i\in M_{01}} p_i^0 q_i^0}\right)^{0.5}\left(1+\frac{\sum_{i\in D_{01}} \hat{p}_i^1 q_i^0}{\sum_{i\in M_{01}} p_i^1 q_i^0}\right)^{0.5}\right)\end{aligned} \tag{37}$$

Note that in equation 37, the nominator of the third term is equal to $\nu_{11}^1\nu_{11}^0$, whereas the denominator of the fourth term is equal to $\delta_{00}^0\delta_{00}^1$. We now combine equation 37 with equation 19 in order to derive a relationship between the QU index and the matched Fisher.

$$\begin{aligned}\frac{P_{QU}^{01}}{P_{MF}^{01}} = \frac{P_{QU}^{01}}{P_{IF}^{01}}\frac{P_{IF}^{01}}{P_{MF}^{01}} &= \left(\left(\sum_{i\in M_{01}} p_i^1 q_i^1\right)^{0.5}\left(\sum_{i\in M_{01}} p_i^0 q_i^1\right)^{0.5}\right) / \left(\sum_{i\in M_{01}} v_i q_i^1\right) \\ &\times \left(\sum_{i\in M_{01}} v_i q_i^0\right) / \left(\left(\sum_{i\in M_{01}} p_i^1 q_0^1\right)^{0.5}\left(\sum_{i\in M_{01}} p_i^0 q_i^0\right)^{0.5}\right) \\ &\times \left(1+\frac{\sum_{i\in N_{01}} p_i^1 q_i^1}{\sum_{i\in M_{01}} p_i^1 q_i^1}\right) / \left(1+\frac{\sum_{i\in N_{01}} v_i q_i^1}{\sum_{i\in M_{01}} v_i q_i^1}\right) \\ &\times \left(1+\frac{\sum_{i\in D_{01}} v_i q_i^0}{\sum_{i\in M_{01}} v_i q_i^0}\right) / \left(1+\frac{\sum_{i\in D_{01}} p_i^0 q_i^0}{\sum_{i\in M_{01}} p_i^0 q_i^0}\right)\end{aligned} \tag{38}$$

In the end, the difference between the QU index and the matched and imputation Fisher indices depend on the way that the quality adjustment factors compare to the observed or estimated prices.

4 Matched-model bias and unit value bias

We consider two nested options for specifying the individual product. On the one hand we have a tight product specification. On the other hand, we have a broad product specification which is obtained by grouping together the initial items. Tightly specified products may cause a bias as new and disappearing items in the two comparison periods are not taken into account in a matched price index. Broadly specified products may cause a bias as the underlying items that are grouped together may not be of the same quality. Our objective is to evaluate the two product specifications and find out which one works best. To do so, we will estimate matched-

model bias and unit value bias.

We quantify the matched-model bias by comparing the matched index with an imputation index.

$$b_{MM}^{01} = ln(\frac{P_{MF}^{01}}{P_{IF}^{01}}) \approx \frac{P_{MF}^{01}}{P_{IF}^{01}} - 1 \tag{39}$$

We quantify the unit value bias by comparing the hybrid index with an imputation index.

$$b_{UV}^{01} = ln(\frac{P_{HF}^{01}}{P_{IF}^{01}}) \approx \frac{P_{HF}^{01}}{P_{IF}^{01}} - 1 \tag{40}$$

It follows from the decomposition 29 that the difference between the matched and hybrid index can be explained by these two biases:

$$ln(\frac{P_{HF}^{01}}{P_{MF}^{01}}) = b_{UV}^{01} - b_{MM}^{01} \tag{41}$$

These biases have the following practical implications on the specification of the individual product.

- Suppose that the matched model-bias is close to 0 but the unit value bias is very different from 0. In that case, we would prefer the initial (tight) product specification over the grouped (broad) product specification. This is because the matching of the P_{MF} seems to be sufficient while P_{HF} is subject to some unit value bias.

- Suppose that the matched-model bias is very different from 0 but the unit value bias is close to 0, we would prefer the the grouped (broad) product specification over the initial (tight) product specification. This is because the matching in the P_{MF} seems not to be sufficient. The matching problem is solved by the P_{HF}, without creating unit value bias.

- Suppose that both matched model-bias and unit value bias are different from 0. Neither P_{MF} nor P_{HF} is fully satisfactory. In such a case, P_{IF} might be the best solution.

One could argue to use by default the imputation index as it acts as a benchmark index. However, both the matched and hybrid indices do not rely on imputations. From this point of view, we may prefer to apply these approaches if results remain satisfactory. The imputation index is subject to an additional uncertainty as it depends on a specific imputation method. In our framework, the imputation is based on the groups used in the hybrid index. This imputation method should help the compiler to identify any biases with either the tight or the brought product specification.

A bilateral Fisher index can be applied as a fixed base index or as a chained index. None of these two strategies is satisfactory in the context of scanner data. A fixed base Fisher index compares prices in a fixed base period with prices in the current period. The choice of the base period may have too much influence on the resulting index. Moreover, by moving away

from the base period, the overlap of products declines, which makes the calculation of price comparisons more difficult. One way of increasing the overlap of products is to update the base period each month and chain link the resulting month-on-month Fisher indices. However, it has been found that a chained Fisher index can be subject to chain drift because the Fisher index is not transitive.

In order to overcome these limitations, transitive index formulas can be used. Transitivity is an index number property in which an index that compares periods a and b indirectly through period c is required to be identical to one that compares periods a and b directly. Several transitive index formulas have been proposed as a solution when using scanner data (see Chapter 10 in [12], and [11]). These index formulas are part of the family of multilateral methods. In a multilateral method, the aggregate price change between two comparison periods is obtained from prices and quantities observed in multiple periods, not only in the two comparison methods.

One specific example of a multilateral method is the Gini–Eltetö–Köves–Szulc (GEKS) method. This method is based on the bilateral Fisher indices calculated between any two periods of a given time window. These bilateral price comparisons are then averaged in order to obtain the GEKS price index. It can be shown that the GEKS index is the transitive index that is closest to its underlying bilateral indices.

In our context, we define the following GEKS indices based on the matched, imputation and hybrid Fisher indices. Let us consider a time window consisting of periods $0, 1, \ldots, T$ over which the GEKS index is applied. The different GEKS indices are then defined as follows:

$$P^{0,t}_{GEKS-M} = \prod_{k \in 0..T} \left(P^{0k}_{MF} \cdot P^{kt}_{MF} \right)^{\frac{1}{T+1}} \quad \forall t \in 0, 1, \ldots, T \tag{42}$$

$$P^{0,t}_{GEKS-I} = \prod_{k \in 0..T} \left(P^{0k}_{IF} \cdot P^{kt}_{IF} \right)^{\frac{1}{T+1}} \quad \forall t \in 0, 1, \ldots, T \tag{43}$$

$$P^{0,t}_{GEKS-H} = \prod_{k \in 0..T} \left(P^{0k}_{HF} \cdot P^{kt}_{HF} \right)^{\frac{1}{T+1}} \quad \forall t \in 0, 1, \ldots, T \tag{44}$$

Note that all three GEKS indices are transitive. As a consequence, these indices do solve the problem of 'chain drift' caused by the bouncing of prices and quantities. This type of chain drift has been examined in Von Auer, 2019 [17]. However, the GEKS indices are not necessarily exempted from the matched-model bias and unit value bias. As in the bilateral case, we can now distinguish, on the one hand, the matched-model bias for a GEKS index based on the tight product specification from, on the other hand, the unit value bias for a GEKS index based on the broad product specification. Note that these 'multilateral' biases can be defined as a GEKS-type average of the biases observed in the bilateral case.

$$b^{0,t}_{GEKS-MM} = ln(\frac{P^{0t}_{GEKS-M}}{P^{0t}_{GEKS-I}}) = ln\left(\prod_{k\in 0..T}\left(\left(\frac{P^{0k}_{MF}}{P^{0k}_{IF}}\right)\cdot\left(\frac{P^{kt}_{MF}}{P^{kt}_{IF}}\right)\right)^{\frac{1}{T+1}}\right) = \frac{1}{T+1}\sum_{k=0..T}\left(b^{0,k}_{MM} + b^{k,t}_{MM}\right) \tag{45}$$

$$b^{0,t}_{GEKS-UV} = ln(\frac{P^{0t}_{GEKS-H}}{P^{0t}_{GEKS-I}}) = ln\left(\prod_{k\in 0..T}\left(\left(\frac{P^{0k}_{HF}}{P^{0k}_{IF}}\right)\cdot\left(\frac{P^{kt}_{HF}}{P^{kt}_{IF}}\right)\right)^{\frac{1}{T+1}}\right) = \frac{1}{T+1}\sum_{k=0..T}\left(b^{0,k}_{UV} + b^{k,t}_{UV}\right) \tag{46}$$

We can naturally extend the decomposition 29 defined in a bilateral context to a multilateral context as follows:

$$\frac{P^{0t}_{GEKS-H}}{P^{0t}_{GEKS-M}} = \frac{P^{0t}_{GEKS-I}}{P^{0t}_{GEKS-M}}\cdot\frac{P^{0t}_{GEKS-H}}{P^{0t}_{GEKS-I}} \tag{47}$$

It follows from the decomposition 47 that the difference between the matched and hybrid GEKS indices can be explained by these two biases:

$$ln(\frac{P^{0t}_{GEKS-H}}{P^{0t}_{GEKS-M}}) = b^{0t}_{GEKS-UV} - b^{0t}_{GEKS-MM} \tag{48}$$

The decomposition applied to the GEKS cannot easily be extended to other multilateral methods such as the Weighted Time Product Dummy or the Geary-Khamis (see Chapter 10 in [12] for the formal definitions of these methods). This is because these two multilateral methods are not sensitive to imputed prices [6]. It is possible to use either the tight or the broad product specification with these methods, but there is not a third option based on imputation. At the same time, the Weighted Time Product Dummy or the Geary-Khamis are not necessarily subject to the same type of matched-model bias and unit value bias. Compared to the GEKS, product churn is treated differently in these two methods. Moreover, these methods are by definition more closely related to unit value calculations. For example, the Geary-Khamis method can be seen as a special case of a quality adjusted unit value index.

5 Examples

We illustrate the analysis on three data sets. The first data set (milk) is included in the *IndexNumR* package [7]. The second data set (T-shirts) is included in Annex C. The third data set is included in Annex D. The first example is used to show either matched-model bias or unit value bias. The second example is about a situation with both matched-model and unit value bias occurring together. The third example illustrates life-cycle pricing.

[6] Technically these two methods do not pass the responsiveness test (see for example section 4.4.2 in Eurostat, 2022 [11]).

[7] See `hiips://CRAN.R-project.org/package=PriceIndices`

5.1 Example 1

The data set covers 21 periods. There are 75 item codes that are sold in 5 outlets. According to the tight specification, the individual product is specified as an item code in a specific outlet. There are 275 such tightly defined products. According to the broad specification, the individual product is simply defined as an item code. In order to derive the broad specification, the same item code sold in different outlets is combined in order to obtain an average price and a total quantity for the item code.

In this example, the two product specifications give very similar results, and both matched-model bias (of the tight product specification) and and unit value bias (of the broad product specification) is very small. This is because the prices of an item in the different outlets are relatively similar. In the end, we could be indifferent between both product specifications.

In order to illustrate matched-model bias, we will create some missingness in the data set by randomly removing tightly defined individual products. Four scenarios are considered. We randomly remove 10%, 20%, 30% or 40% of the individual products. The price indices and bias decomposition are then calculated for each adjusted data set. The results included in figure 2 show that a larger share of missing products logically increases matched-model bias. Note that the direction of the matched-model bias is undefined, and may even cancel out on average over the 21 time periods. As expected, unit value bias remains small in all scenarios as prices. Still, in order to treat the missingness in this case, we would prefer the broad product specification over the tight product specification.

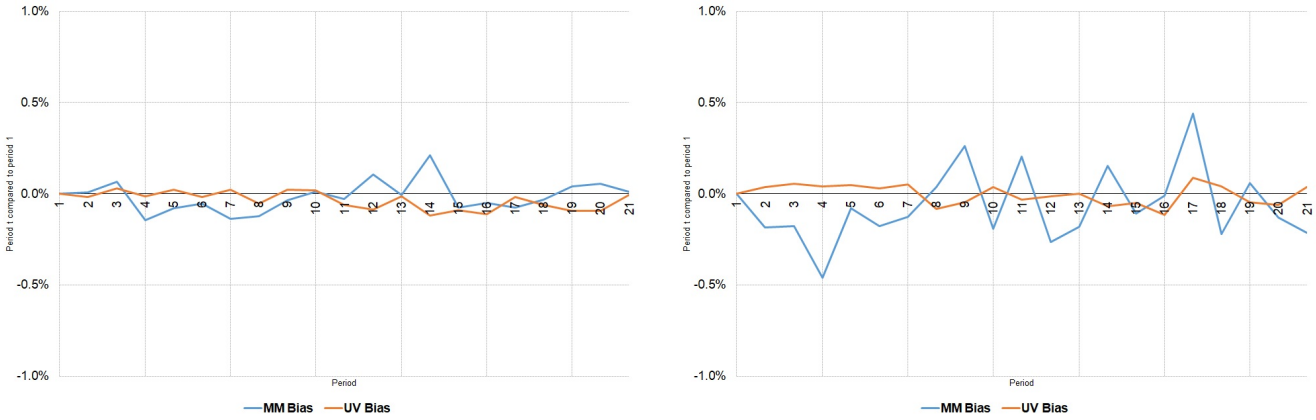
Instead of randomly removing products, we now conduct another modification on the initial milk data set. In 2 out of the 5 outlets[8], we will increase the prices by respectively 10%, 20%, 30 % and 40 %. We assume that these outlets charge a premium for additional benefits (for example larger opening hours, convenient location, etc.) while selling the same products. The results included in figure 3 show that a higher price level in 2 out of 5 outlets increases unit value bias. The matched-model bias remains small in all four scenarios. The matched, imputation and hybrid GEKS indices under the fourth scenario (prices increased by 40% in 2 out of 5 outlets) are shown in figure 4. In such a circumstance, we would prefer the tight product specification over the broad product specification[9].

5.2 Example 2

In this data set, there are 13 periods and 30 item codes, which can be grouped together into 6 homogeneous products. The grouping is derived from the following attributes of the T-shirts: Fabric (cotton or organic), Sleeves (long or short), Number of items (1, 2 or 3). The data set is

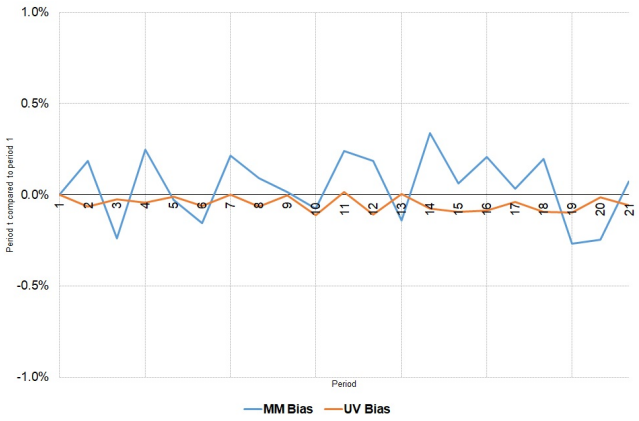
[8] We augment the prices of the outlets coded as 2210 and 1311 in the data set.

[9] Another, intermediate, solution would be to group together the 2 outlets that have a higher price level, and the 3 outlets that have a lower price level.

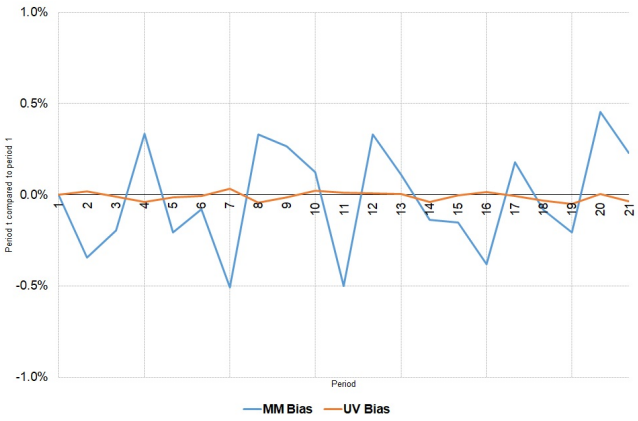


10% of products deleted

20% of products deleted

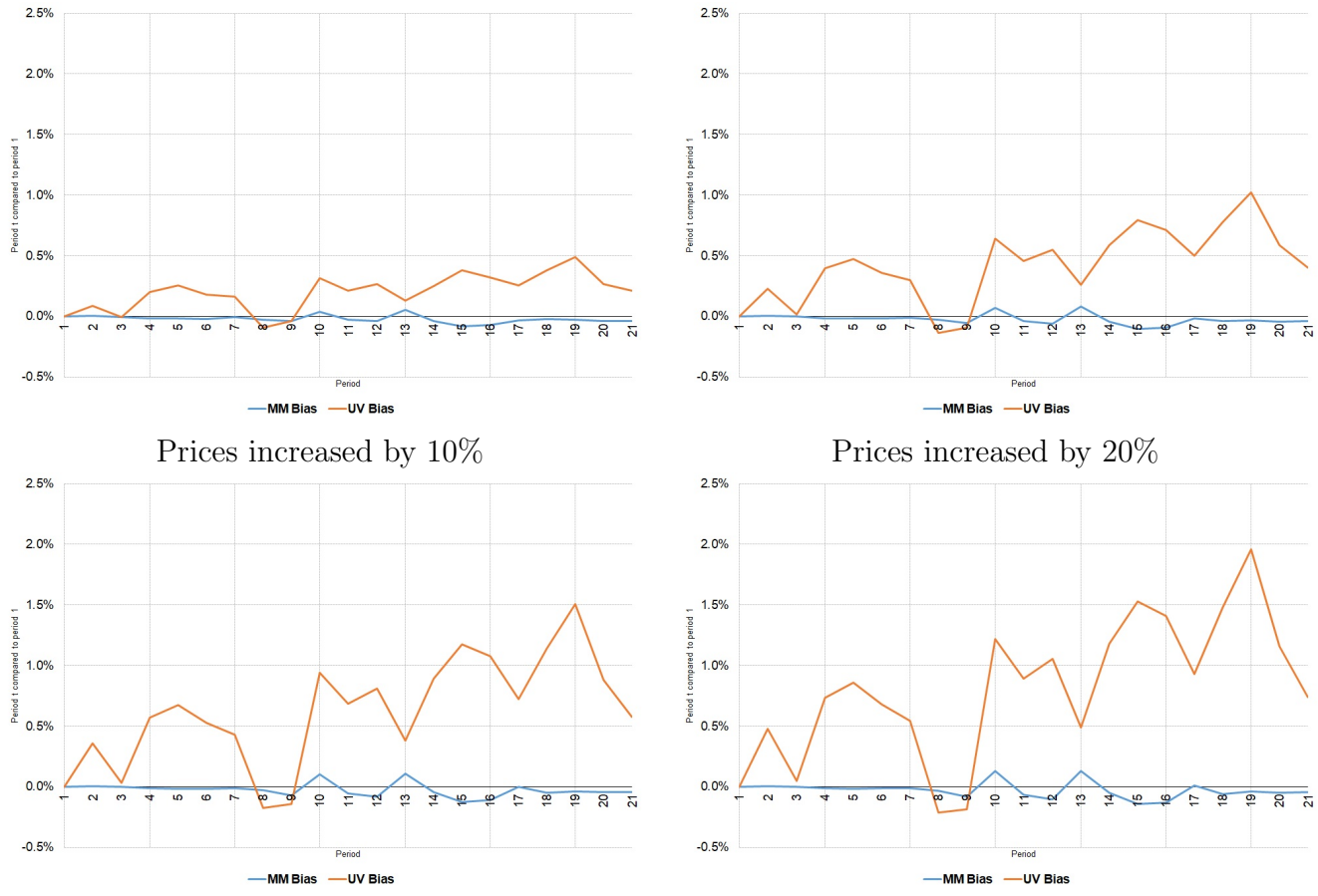


30% of products deleted



40% of products deleted

Figure 2: Bias decomposition for the milk data set after randomly deleting tightly defined products.



Prices increased by 10%

Prices increased by 20%

Prices increased by 30%

Prices increased by 40%

Figure 3: Bias decomposition for the milk data set after increasing the price level in 2 out of 5 outlets.

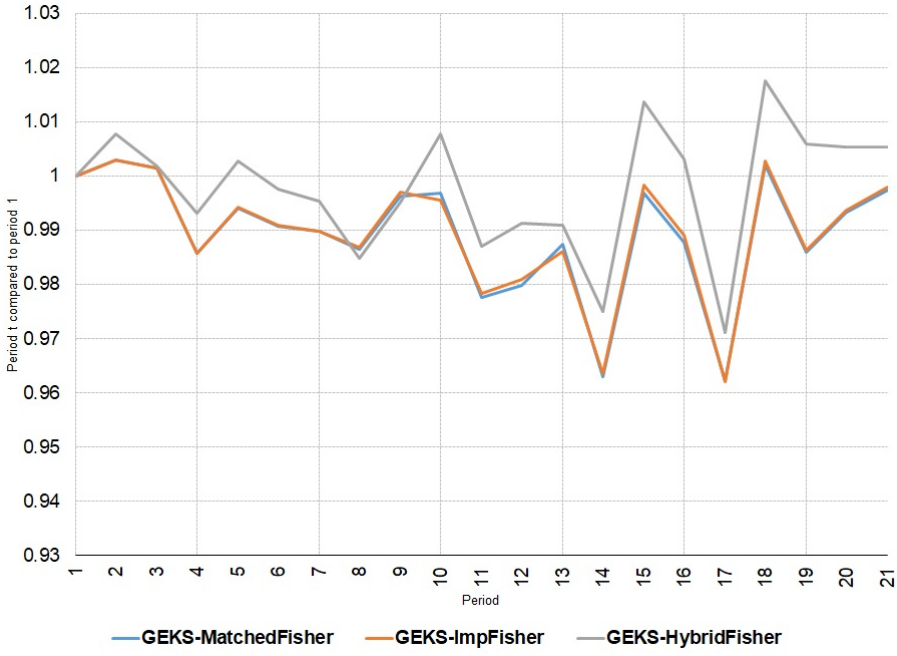


Figure 4: Matched, imputation and hybrid GEKS indices for the milk data set (Prices increased by 40% in 2 out of 5 outlets).

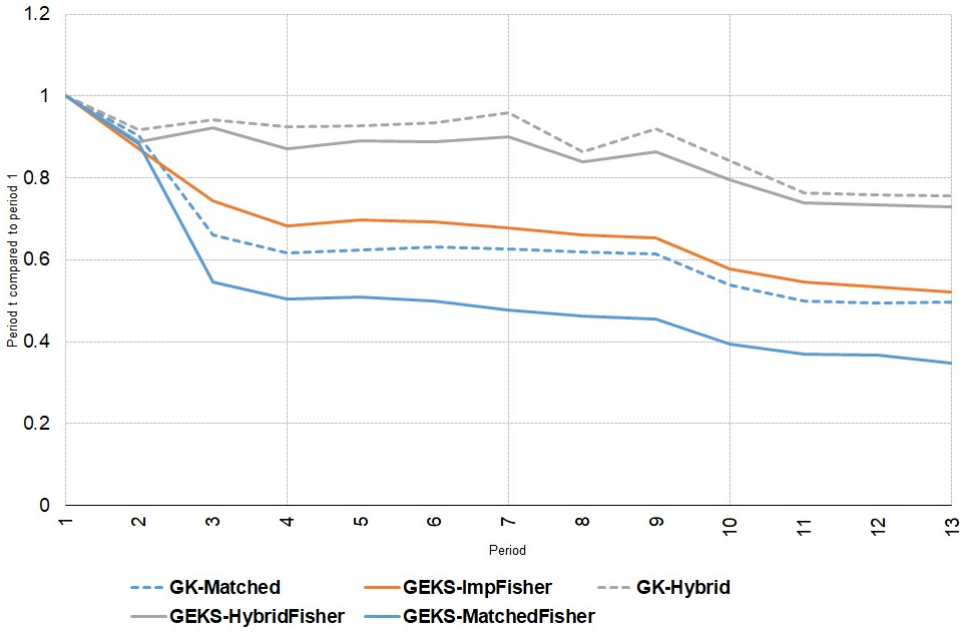


Figure 5: Matched, imputation and hybrid indices for the T-shirt data set.

characterized by an overall downward trend in prices and significant item churn. All the data are included in Annex C [10].

In this example, there are significant differences between the matched, imputation and hybrid indices (see Figures 5). In fact, the matched GEKS index sits below the GEKS imputation index, which means that there is some negative matched-model bias. The hybrid GEKS index sits above the imputation GEKS index, which means that there is some positive unit value bias. As a consequence, an imputation index could be the preferred solution.

The matched-model and unit value biases become larger starting with period 3 (see figure 6). This is because there are several items that are not available in periods 0 and 1, but available thereafter. These items lead to both matched-model bias because of their non-inclusion in the matched approach and to unit value bias once they are grouped with other items. This example is also useful to illustrate the impact of using a multilateral approach instead of a bilateral approach. The biases in the bilateral case are increasing in magnitude from period 3 onwards, while in the multilateral case they are approximately stable at an average level.

The Geary-Khamis is also applied to this data set. There are two variants. The first variant is based on the individual products defined at the GTIN level (GK-Matched). The second variant is based on the homogeneous products (GK-Hybrid). There is no variant based on imputed prices. On this example, the GK-Matched index is in fact very similar to the imputation GEKS index, wheres the GK-Hybrid index is very similar to the hybrid GEKS index.

[10] A related data set has also been used in De Haan, 2021 [6] to show the impact of new and disappearing items on Time-Product Dummy and Time Dummy Hedonic Indexes.

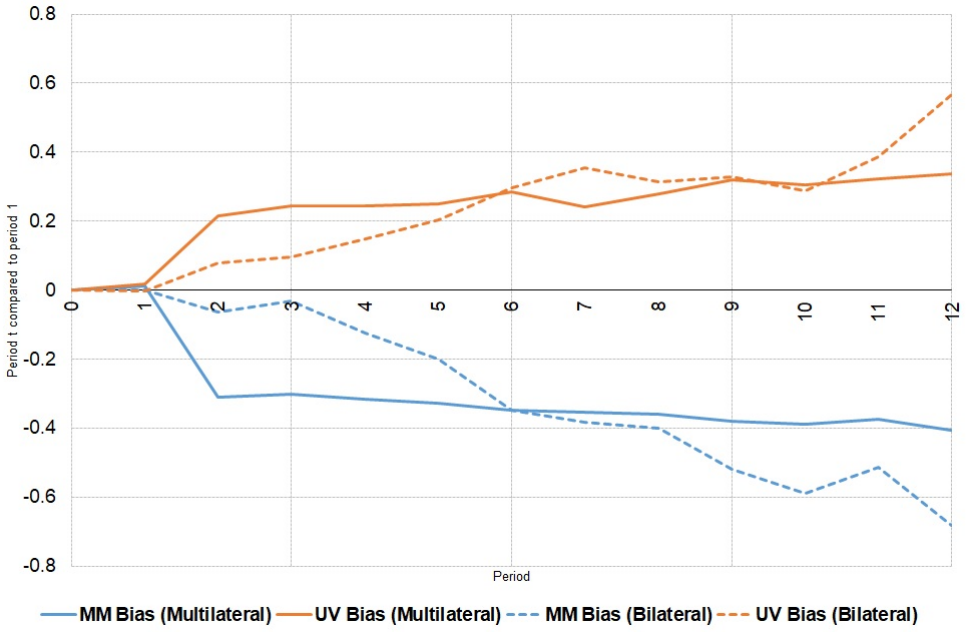


Figure 6: Bias decomposition for the T-shirt data set.

5.3 Example 3

This data set illustrates products that are subject to life-cycle pricing. The data set is composed of the following models:

- Model 1a is available in periods 1-7, its successor model 1b is available in periods 6-11, and its successor model 3c is available in periods 11-12.
- Model 2a is available in periods 2-9, its successor model 2b is available in periods 8-12, and its successor model 3c is available in periods 11-12.
- Model 3a is available in periods 1-12 and its successor model 3b is available in periods 11-12.

For each model, prices continuously decrease during its life-cycle. The successor model has a slightly higher initial price than the initial price of the previous model. The matched indices are calculated by treating each model as a different individual product. In the hybrid indices, the different variants of a model are grouped together.

This example shows (see figure 7) that the matched GEKS index is downward biased compared to the imputation GEKS index. In other words, the matched GEKS index has some negative matched-model bias. The matched GEKS index does not capture well the rebound caused by high initial prices of the successor models. The imputation GEKS index slightly sits

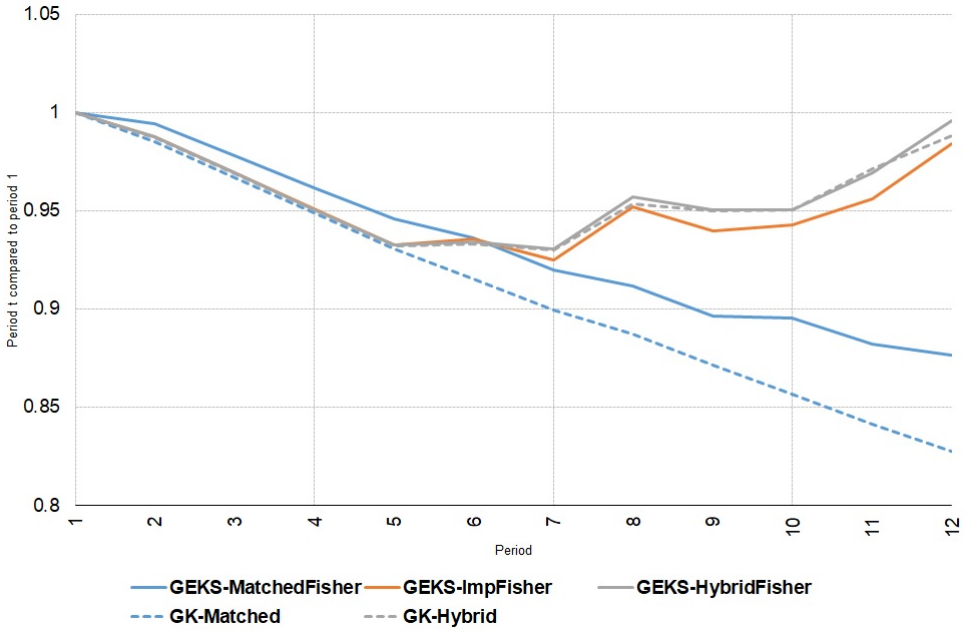


Figure 7: Matched, imputation and hybrid indices with life-cycle pricing data.

below the hybrid GEKS index, pointing to some small unit value bias of the hybrid GEKS index.

The matched and hybrid variants of the GK are also applied to this data set. In this case, the matched GK has a downward trend that is even more pronounced than the one observed for the matched GEKS, whereas the hybrid variant of the GK is similar to the hybrid GEKS.

6 Conclusions

The individual product that enters the price index can often be specified in different ways. In order to provide some guidance on which product specification to use, we compare a matched price index based on a tight product specification to a hybrid index based on a broad specification. We assess both unit value bias and matched-model bias by comparing the resulting matched and hybrid indices to an imputation index. One contribution of this paper is to clarify the relationship between these three indices.

Our practical conclusion is to prefer the broad product specification over the tight product specification if the matched-model bias of the tight product specification outweighs the unit value bias of the broad product specification, and vice-versa. Imputation should best be considered if both matched-model bias and unit value bias are large.

While the analysis in this paper is an attempt to formalize the problem of product specification, there are several limitations that should be further investigated.

- In the imputation index, a price is estimated for the unmatched individual products based on the average price of similar individual products in the same period. This may not be the best imputation method. Other imputation methods could lead to other conclusions. More work is needed on imputation methods.
- While we focus on unit value bias, the real objective is to measure the degree of quality differences of products that are grouped together. Unit value bias is a only proxy measure and other, more targeted, measures should be developed.
- In practice, there can be different ways how items can be grouped together. As a consequence, there may be more than two different product specification to chose from. The model in this paper needs to be made more operational in order to help compilers select one out of many product specifications.
- The hybrid index is based on two stages. A Fisher index is calculated from previously calculated unit values. This could be generalized by calculating a Fisher index from quality adjusted unit values. It should be investigated how such an index would compare to the imputation Fisher index.
- The framework is derived in the context of a GEKS index. It may also be applicable to other multilateral methods that are related to the Fisher index. It should be examined how the concepts of matched-model bias and unit value bias can be extended to other multilateral methods such as the Weighted Time Product Dummy or the Geary-Khamis.

Annex A: Comparison between the imputation index and the matched index

Following the definitions of a matched Laspeyres index (see equation 1) and an imputation Laspeyres index (see equation 4), we have the following:

$$\frac{P_{IL}^{01}}{P_{ML}^{01}} = \frac{1 + \frac{\sum_{i \in D_{01}} \hat{p}_i^1 q_i^0}{\sum_{i \in M_{01}} p_i^1 q_i^0}}{1 + \frac{\sum_{i \in D_{01}} p_i^0 q_i^0}{\sum_{i \in M_{01}} p_i^0 q_i^0}} = \frac{1 + \frac{\sum_{i \in D_{01}} \hat{p}_i^1 q_i^0 \times \frac{1}{\sum_r Q_r^0}}{\sum_{i \in M_{01}} p_i^1 q_i^0 \times \frac{1}{\sum_r Q_r^0}}}{1 + \frac{\sum_{i \in D_{01}} p_i^0 q_i^0 \times \frac{1}{\sum_k Q_k^0}}{\sum_{i \in M_{01}} p_i^0 q_i^0 \times \frac{1}{\sum_k Q_k^0}}} \tag{49}$$

We also know from equations 17 and 18 that:

$$s_i^0 \times \sigma_i^0 = \frac{q_i^0}{Q_{\kappa(i)}^0} \times \frac{Q_{\kappa(i)}^0}{\sum_k Q_k^0} = \frac{q_i^0}{\sum_k Q_k^0} \tag{50}$$

Combining equations 49 and50, we finally obtain that:

$$\frac{P_{IL}^{01}}{P_{ML}^{01}} = \frac{1 + \frac{\sum_{i \in D_{01}} \hat{p}_i^1 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0}}{1 + \frac{\sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^0}} \tag{51}$$

With a similar reasoning, we obtain the following relationship between a matched Paasche index (see equation 2) and an imputation Paasche index (see equation 5)

$$\frac{P_{IP}^{01}}{P_{MP}^{01}} = \frac{1 + \frac{\sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^1}}{1 + \frac{\sum_{i \in N_{01}} \hat{p}_i^0 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^1 \sigma_i^1}} \tag{52}$$

The decomposition in 19 is based on a Fisher index and can be obtained by taking the square root of the equations 51 and 52.

Annex B: Comparison between the hybrid index and the imputation index

Following the definition of a hybrid Laspeyres index (see equation 9), we get the following:

$$P_{HL}^{01} = \frac{\sum_k \bar{p}_k^1 Q_k^0}{\sum_k \bar{p}_k^0 Q_k^0} = \frac{\sum_k \left(\sum_{i \in H_k} \frac{p_i^1 q_i^1}{\sum_{i \in H_k} q_i^1} \right) Q_k^0}{\sum_k \left(\sum_{i \in H_k} \frac{p_i^0 q_i^0}{\sum_{i \in H_k} q_i^0} \right) Q_k^0} \tag{53}$$

Taking into account the definitions for quantity shares as given in equations 17 and 18, the previous equation 53 can be rewritten as follows:

$$P_{HL}^{01} = \frac{\sum_k \left(\sum_{i \in H_k} p_i^1 s_i^1\right) Q_k^0}{\sum_k \left(\sum_{i \in H_k} p_i^0 s_i^0\right) Q_k^0} = \frac{\sum_k \sum_{i \in H_k} p_i^1 s_i^1 \sigma_i^0 \left(\sum_k Q_k^0\right)}{\sum_k \sum_{i \in H_k} p_i^0 s_i^0 \sigma_i^0 \left(\sum_k Q_k^0\right)} = \frac{\sum_k \sum_{i \in H_k} p_i^1 s_i^1 \sigma_i^0}{\sum_k \sum_{i \in H_k} p_i^0 s_i^0 \sigma_i^0} \tag{54}$$

The set of items will now be separated into matched, new and disappearing products.

$$P_{HL}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0 + \sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^0 + \sum_{i \in D_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^0 + \sum_{i \in N_{01}} p_i^0 s_i^0 \sigma_i^0 + \sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^0} \tag{55}$$

Note that for a new item, the period 0 quantity share s_i^0 must be zero, whereas for a disappearing item he period 1 quantity share s_i^1 must be zero. Hence, the previous equation can be further simplified:

$$P_{HL}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0 + \sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^0 + \sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^0} \tag{56}$$

Using again the quantity shares defined equations 17 and 18, the imputation Laspeyres index (equation 4)can be rewritten as follows:

$$P_{IL}^{01} = \frac{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0 + \sum_{i \in D_{01}} \hat{p}_i^1 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^0 + \sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^0} \tag{57}$$

Dividing the hybrid Laspeyres index as defined in equation 56 with an imputation Laspeyres index as defined in equation 57, we obtain the following:

$$\frac{P_{HL}^{01}}{P_{IL}^{01}} = \frac{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0 + \sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0 + \sum_{i \in D_{01}} \hat{p}_i^1 s_i^0 \sigma_i^0} = \left(\frac{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0}\right) \left(\frac{1 + \frac{\sum_{i \in N_{01}} p_i^1 s_i^1 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^1 \sigma_i^0}}{1 + \frac{\sum_{i \in D_{01}} \hat{p}_i^1 s_i^0 \sigma_i^0}{\sum_{i \in M_{01}} p_i^1 s_i^0 \sigma_i^0}}\right) \tag{58}$$

With a similar reasoning, we can obtain the following for the Paasche indices.

$$\frac{P_{HP}^{01}}{P_{IP}^{01}} = \left(\frac{\sum_{i \in M_{01}} p_i^0 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^1}\right) \left(\frac{1 + \frac{\sum_{i \in N_{01}} \hat{p}_i^0 s_i^1 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^1 \sigma_i^1}}{1 + \frac{\sum_{i \in D_{01}} p_i^0 s_i^0 \sigma_i^1}{\sum_{i \in M_{01}} p_i^0 s_i^0 \sigma_i^1}}\right) \tag{59}$$

The decomposition in 25 is based on a Fisher index and can be obtained by taking the square root of the equations 58 and 59.

Annex C: Data example 2

ProdID	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
GTIN-01	6.92	6.33	6.16	6.37	5.67	5.67	4.00	3.33	4.00	2.00	2.00	4.00	2.00
GTIN-02			8.88	7.34	8.65	8.33	8.31	8.84	8.21	6.05	5.01	4.92	5.04
GTIN-03			14.29	13.93	14.38	14.00	12.60	12.33	12.56	13.33	14.52	14.74	15.11
GTIN-04	6.97	5.93	4.50	3.00	2.00	2.00	3.00	2.00	1.00				2.00
GTIN-05			4.90	4.88	4.79	4.97	5.22	5.02	5.00	5.02	4.94	4.90	5.10
GTIN-06				7.24	8.39	8.04	8.05	8.67	7.96	5.88	5.01	4.92	4.99
GTIN-07			12.39	11.05	11.87	12.65	12.57	12.06	10.43	10.94	12.32	12.03	12.52
GTIN-08	9.00	6.25	4.00	2.00	1.00	1.00	2.00	1.00				1.00	
GTIN-09			7.93	7.89	7.59	7.98	7.91	8.00	7.84	7.25	7.85	7.74	8.00
GTIN-10	9.92	9.85	8.75	8.33	6.80	8.50	7.50	5.50	8.00	4.00	6.00	5.00	4.00
GTIN-11			7.91	7.88	7.61	7.89	8.00	7.72	7.81	7.34	7.87	7.73	7.89
GTIN-12			4.95	4.92	4.81	4.96	5.05	5.08	5.04	5.03	5.00	4.88	4.98
GTIN-13	8.40	7.67	5.00	4.00	4.00	4.00	2.00	2.00	1.00		1.00	1.00	1.00
GTIN-14	10.04	10.06	8.00	5.00	4.00	4.00	1.00	1.00	1.00	1.00		2.00	1.00
GTIN-15	6.91	6.11	5.25	5.50	5.33	4.00	3.00	1.33	3.00	2.00	2.00	1.00	1.00
GTIN-16	5.00	3.50	3.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00	2.00	3.00	1.00
GTIN-17	14.00	12.40	8.00	9.33	7.33	11.00	5.25	5.67	3.00	8.00	3.00	2.00	1.00
GTIN-18			4.91	4.92	4.81	4.96	4.92	4.92	4.93	5.03	4.92	4.91	4.91
GTIN-19	7.95	6.60	4.00	6.00	3.00	2.00	2.00	2.00	2.00				1.00
GTIN-20							1.00	1.00					
GTIN-21	9.89	9.50	8.00	8.00	8.00	4.50	4.00	4.00	2.00	2.00	2.00	2.00	2.00
GTIN-22	6.85	6.00	4.67	5.00	4.00	3.50	3.50	3.00	3.00	2.00	1.00	1.00	1.00
GTIN-23	13.43	11.20	10.67	9.50	9.00	5.00	3.50	3.50	3.00	2.00	1.00		
GTIN-24	7.90	6.79	5.75	5.33	4.50	4.50	4.00	4.00	4.00	3.00	2.00	1.00	1.00
GTIN-25	6.85	6.21	5.00	5.33	6.00	5.00	5.00	4.00	2.00	1.00	1.00	1.00	
GTIN-26	9.80	9.68	8.33	9.25	8.50	6.75	5.40	5.33	6.50	7.00	7.00	8.00	4.00
GTIN-27	14.20	11.00	7.67	9.33	7.50	7.00	5.33	5.50	6.00	4.00	5.00	2.00	1.00
GTIN-28	9.81	9.76	7.33	8.00	5.25	5.67	3.50	5.00	8.00	5.00	5.00	3.00	2.00
GTIN-29	5.84	4.83	5.30	5.55	5.38	4.17	3.17	2.50	2.00			1.00	1.00
GTIN-30	8.84	8.09	6.90	6.40	7.33	4.75	5.50	2.67	2.50	4.00	2.00	5.00	3.00

ProdID	q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}
GTIN-01	103	120	37	19	9	6	6	3	2	1	1	1	1
GTIN-02			81	221	137	163	116	90	141	251	405	487	337
GTIN-03			17	14	8	9	5	3	9	18	27	43	18
GTIN-04	108	15	2	1	1	1	1	1	1				1
GTIN-05			42	75	99	93	23	141	79	51	67	82	50
GTIN-06				34	23	26	20	18	28	73	134	153	97
GTIN-07			23	43	30	31	23	17	30	36	31	37	23
GTIN-08	5	4	1	1	1	1	1	1				1	
GTIN-09			41	62	58	57	45	28	45	64	55	66	39
GTIN-10	74	41	8	6	5	2	2	2	1	1	1	1	1
GTIN-11			64	65	62	53	37	25	31	50	52	73	37
GTIN-12			39	66	91	81	20	130	68	40	64	73	46
GTIN-13	5	3	1	1	1	1	1	1	1		1	1	1
GTIN-14	49	16	2	2	1	1	1	1	1	1		1	1
GTIN-15	99	66	8	4	3	5	3	3	1	1	1	1	1
GTIN-16	2	2	2	1	1	1	1	1	1	1	1	1	1
GTIN-17	7	5	5	3	3	1	4	3	1	1	1	1	1
GTIN-18			32	53	67	96	38	73	43	32	48	54	34
GTIN-19	22	10	2	1	1	1	1	1	1				1
GTIN-20							1	1					
GTIN-21	27	14	6	2	1	2	1	1	1	1	1	1	1
GTIN-22	61	26	3	3	2	2	2	1	1	1	1	1	1
GTIN-23	7	5	3	2	2	2	4	2	1	1	1		
GTIN-24	39	19	4	3	2	2	2	1	1	1	1	1	1
GTIN-25	47	47	5	3	1	1	1	1	1	1	1	1	
GTIN-26	30	34	18	8	4	4	5	3	2	1	1	1	1
GTIN-27	5	5	3	3	2	2	3	2	1	1	1	1	1
GTIN-28	47	33	6	4	4	3	2	1	1	1	1	1	1
GTIN-29	50	42	30	22	8	6	6	2	1			1	1
GTIN-30	82	56	10	5	3	4	2	3	2	1	1	1	1

HP code	Attributes	ProdID
HP1	Cotton; Long; 2	GTIN-30
HP2	Cotton; Long; 2	GTIN-03
HP3	Cotton; Short; 1	GTIN-01
HP3	Cotton; Short; 1	GTIN-04
HP3	Cotton; Short; 1	GTIN-05
HP3	Cotton; Short; 1	GTIN-09
HP3	Cotton; Short; 1	GTIN-11
HP3	Cotton; Short; 1	GTIN-12
HP3	Cotton; Short; 1	GTIN-15
HP3	Cotton; Short; 1	GTIN-16
HP3	Cotton; Short; 1	GTIN-17
HP3	Cotton; Short; 1	GTIN-18
HP3	Cotton; Short; 1	GTIN-19
HP3	Cotton; Short; 1	GTIN-20
HP3	Cotton; Short; 1	GTIN-22
HP3	Cotton; Short; 1	GTIN-23
HP3	Cotton; Short; 1	GTIN-24
HP3	Cotton; Short; 1	GTIN-25
HP3	Cotton; Short; 1	GTIN-29
HP4	Cotton; Short; 2	GTIN-07
HP4	Cotton; Short; 2	GTIN-10
HP4	Cotton; Short; 2	GTIN-14
HP4	Cotton; Short; 2	GTIN-21
HP4	Cotton; Short; 2	GTIN-26
HP4	Cotton; Short; 2	GTIN-28
HP5	Cotton; Short; 3	GTIN-27
HP6	Organic; Short; 1	GTIN-02
HP6	Organic; Short; 1	GTIN-06
HP6	Organic; Short; 1	GTIN-08
HP6	Organic; Short; 1	GTIN-13

Annex D: Data example 3

ProdID	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
Model 1a	100.00	98.00	96.04	94.12	92.24	90.39	88.58					
Model 1b						102.00	99.96	97.96	96.00	94.08	92.20	
Model 1c											104.00	102.00
Model 2a		100.00	98.00	96.04	94.12	92.24	90.39	88.58	86.81			
Model 2b								102.00	99.96	97.96	96.00	94.08
Model 3c											104.00	102.00
Model 3a	100.00	99.00	98.01	97.03	96.06	95.10	94.15	93.21	92.27	91.35	90.44	89.53
Model 3b											104.00	102.00

ProdID	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}
Model 1a	53	60	65	70	75	50	40					
Model 1b						20	40	60	80	85	75	
Model 1c											20	40
Model 2a		30	35	40	45	45	45	40	25			
Model 2b								20	30	40	40	40
Model 3c											30	30
Model 3a	10	10	10	10	10	10	10	10	10	10	10	10
Model 3b											20	20

References

[1] Aizcorbe A. (2022), Outlet Substitution Bias Estimates for Ride Sharing and Taxi Rides in New York City, BEA Working Paper Series, WP2022-1.

[2] Chessa A. (2016), A new methodology for processing scanner data in the Dutch CPI, EURONA 1/2016.

[3] Chessa A. (2019), MARS: A method for defining products and linking barcodes of item relaunches, paper presented at the 16th meeting of the Ottawa group, Rio de Janeiro, Brazil.

[4] Dalèn J. (2017), Unit values in scanner data some operational issues, paper presented at the 15th meeting of the Ottawa Group, Eltville, Germany.

[5] de Haan J. (2001), Generalized Fisher Price Indexes and the Use of Scanner Data in the CPI, Paper presented at the 6th Meeting of the Ottawa group, Canberra, Australia.

[6] de Haan J., Hendriks R. and Scholz M (2021), Price Measurement Using Scanner Data: Time–Product Dummy Versus Time Dummy Hedonic Indexes, Review of Income and Wealth, Series 67, Number 2.

[7] Diewert E.W. and von der Lippe P. (2010), Notes on Unit Value Index Bias, Discussion Paper 10-08, Department of Economics, University of British Columbia.

[8] Diewert E.W., Fox K., and de Haan J. (2016), A newly identified source of potential CPI bias: Weekly versus monthly unit value price indexes, Economics Letters, vol. 141, issue C, 169-172.

[9] European Central Bank (2021), Inflation measurement and its assessment in the ECBs monetary policy strategy review, ECB Occasional Paper Series. Available at `hiips://wwwecbeuropa.eu/pub/pdf/scpops/ecb.op265 ~a3fb0b611d.en.pdf`

[10] Eurostat (2021), Recommendations on the bridged overlap method. Available at `hiips://ec.europa.eu/eurostat/documents/272892/7048317/Recommendation-bridged-overlap-June-2021.pdf/`

[11] Eurostat (2022), Guide on the use of multilateral methods in the HICP (Version 2022). Available at `hiips://ec.europa.eu/eurostat/product?code=KS-GQ-21-020`

[12] ILO, IMF, OECD, Eurostat, UNECE and World Bank (2020), Consumer Price Index Manual Concepts and methods.

[13] Ivancic L. and Fox K. (2013), Understanding price variation across stores and supermarket chains: some implications for CPI aggregation methods, Review of Income and Wealth, Series 59, Number 4.

[14] Konny C. G., Williams B.K., and Friedman D. M. (2019), Big Data in the U.S. Consumer Price Index: Experiences & Plans

[15] Melser D. and Webster M. (2021), Multilateral methods, substitution bias and chain drift: some empirical comparisons, Review of Income and Wealth, Series 67, Number 3.

[16] Von Auer L. (2017), Processing scanner data by an augmented GUV index, EURONA 1/2017.

[17] Von Auer L. (2019), The Nature of Chain Drift, paper presented at the 16th meeting of the Ottawa group, Rio de Janeiro, Brazil.

[18] Von Auer L. (2014), The Generalized Unit Value Index Family, Review of Income and Wealth, Series 60, Number 4.

[19] Williams B. and Sager E. (2019), A New Vehicles Transaction Price Index: Offsetting the Effects of Price Discrimination and Product Cycle Bias with a Year-Over-Year Index, BLS Working Paper 514