Multilateral indices and the relaunch problem: product clustering and alternative solutions

Jacco Daalmans, Statistics Netherlands[1],

Paper for the 2022 meeting of the Ottawa group

1. Introduction

Scanner data generated from electronic paying devices, like cashing desks, are a rich data source on transactions. It contains integral information on paid prices, quantities and product characteristics for all items of some business in a certain period of time. More and more countries use scanner data for (segments of) their Consumer Price indices (CPIs), or consider to do so. The widespread use of scanner data poses certain methodological demands on index methods. It is generally agreed that multilateral index methods are most appropriate. These methods simultaneously compute index values for all periods of an estimation window. Multilateral methods can flexibly deal with a sheer bulk of data. The sets of products are allowed to vary from period to period. New products can be incorporated at the moment of introduction and obsolete products can be removed when they disappear from the market. Furthermore, expenditure-based weights can be used to take economic significance into account. Multilateral methods are transitive, so that a choice of a base period can be avoided.

Most well-known multilateral methods belong to the class of 'matched-model' methods. These are methods that rely on the product identifiers. The prices of exactly the same items are compared across time. Thus, matched-model methods obey the well-known standard of comparing 'like with like'. Although matched-item methods have many desirable properties, these methods also have their drawbacks. An important complication is that part of the price changes is 'missed'. In particular, these are the changes that emerge when a certain product ceases to exist and buyers switch to an alternative product. A typical example is a relaunch: a product that is replaced by an (almost) identical one that usually has a larger price. This often occurs when there is a (minor) change of external appearance, for instance in the packing of the product. As characterized by Dalén (2017), relaunches are replacements of products for which the change in price is larger than the change of quality. Relaunches frequently occur

[1] j.daalmans@cbs.nl

for some consumption segments, like clothing and retail, but are less common for other segments, like restaurants. A further property of matched-model indices is that they are usually not affected by items that have been sold in one period only.

The above-mentioned complications have serious consequences. Chessa (2016) showed several examples in which ignorance of relaunches lead to a large bias. A well-known way of dealing with the 'relaunch problem' is to group similar items together into product clusters. This means that items with the same product characteristics are treated as one. A price index is calculated from the product clusters, rather than from the underlying items. If some product *A* has been replaced by a similar product *B*, the price change induced by the product change is internalized in the change of the cluster price. Product clustering is however only appropriate for sufficiently homogenous items; i.e. items that can replace each other. It does not make sense, for instance, to combine items with a different content into one cluster, like a 1 liter and a 200 ml pack of milk. The error that is obtained in this way is known as unit value bias and has been extensively studied in the literature, see e.g. Diewert and Von der Lippe (2016) and Silver (2010). To cite Triplett (2006): "Unit values are always suspect and provide no solution to the problem of computing basic components". In practice, however, some degree of heterogeneity cannot easily be avoided. Triplett (2006) states for instance that 'homogeneous commodities are rare'. Therefore, some degree of unit value bias is likely to exist in many practical applications. So we have that product clustering can solve the relaunch problem, but also creates another problem in the form of unit value bias. As concluded by Dalén (2017), the current literature lacks insights into the effects of product clustering. One question one can pose for instance is whether grouping items in not very homogeneous groups is better than no grouping at all.

One alternative for product clustering is imputation. This means that estimated prices are added to the data for products that have not been sold. The classical statistical meaning of imputation is to fill in unobserved prices. In traditional data collection, missing values might result if data collectors are unable to find a specified product in a shop, that is actually being sold. In today's scanner data, prices of unsold items are not observed. The aim of imputation is however broader than estimating missing prices. The purpose here is to negate the unfavorable effects of new and disappearing items on a matched-model index. With this aim, imputation can imply that a price is estimated for transactions that might not even have been taken place. Different imputation methods can be used, on which a nontrivial choice has to be made. The current literature contains imputation methods for a selection of indexes.

Another alternative for product clustering is product matching (or replacement). Traditionally, price are collected manually for a fixed basket of products. Whenever one product disappears from the market a replacement product is selected with more or less the same characteristics. A similar approach can be applied to scanner data. For instance, van

Loon (2019) describes a pragmatic, semi-automated method in which listings of new and disappeared products are analyzed using text mining analysis and manual verification. If the price collector determines that there is a relaunch, the new and old product code are linked and if necessary a quality adjustment is carried out. Manual intervention can however be practically impossible for large scanner data sets. Leclair *et al.* (2019) state that "The volume of data to be processed precludes human expert input to the choice of replacement products. An automated decision-making process should therefore be developed".

The main aim of this paper is to compare the effectiveness of clustering, imputation and product matching for their ability of solving the relaunch problem. It is organized as follows. Section 2 introduces three multilateral index methods, that will be considered in this paper: GEKS-Tornqvist, Geary Khamis and Time Product Dummy (TPD). Section 3 gives theoretical relations between product clustering and imputation and also presents a new imputation method for Geary Khamis and TPD. Section 4 introduces an automatable product matching method for scanner data. Section 5 presents the empirical results from a simulation study on the effectiveness of clustering, imputation and matching. Section 6 concludes.

2 Methods

Our focus will be on three multilateral index methods, that are well-known from the literature: GEKS-Törnqvist, Geary Khamis and a Time Product Dummy (TPD) method. Before we explain these methods, some notation needs to be introduced. The price of an item *i* at period *t* will be denoted by p_i^t and the quantity by q_i^t. The prices and quantities for the base period are referred to as p_i^0 and q_i^0.

2.1 GEKS-Törnqvist

A bilateral Törnqvist index $P_T^{0,t}$, that compares the prices of a base period 0 and a comparison period *t* is given by

$$P_T^{0,t} = \prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{(s_i^0 + s_i^t)/2}, \tag{2.1}$$

where s_i^0 and s_i^t are the expenditure shares for both periods, i.e. $s_i^t = p_i^t q_i^t / \sum_{i \in U^t} p_i^t q_i^t$. Further, U_M^{0t} is the set of items (the universe) which have been sold in both periods 0 and t and U^t is the set of products that have been sold in t.

In a multi-period setting, bilateral Törnqvist indexes can be computed for each pair of periods. However, the resulting indexes are not transitive. It is not necessarily true that: $P^{i,j} P^{j,k} = P^{i,k}$. This means that the choice of the base period matters. To resolve this issue a multilateral index can be used. A so-called GEKS method can be applied to create a transitive, multilateral

index from a set of bilateral indexes. A GEKS index is the geometric mean of the bilateral price indexes,

$$P_{GEKS}^{0,t} = \prod_{l=0}^{T} \left(P_T^{0,l} P_T^{l,t}\right)^{1/(T+1)}, \tag{2.2}$$

where *T* refers to the last period of a time series.

The resulting index, after applying GEKS to Törnqvist is known as GEKS-Törnqvist or CCDI (Caves, Christensen and Diewert, 1982; Inklaar and Diewert, 2016). The CCDI approach was followed by De Haan and Van der Grient (2011) and has been implemented by the National Statistical Institutes of Australia (ABS, 2017), Belgium (Van Loon, 2020), Luxembourg (Radjabov and Ferring, 2021) and Norway (Johansen and Nygaard, 2021)

2.2 Geary Khamis

A second method in this paper is the one by Geary Khamis (GK). The GK-index is given by

$$P_{GK}^{0,t} = \frac{\left(\sum_{i \in U^t} p_i^t q_i^t\right) / \left(\sum_{i \in U^0} p_i^0 q_i^0\right)}{\left(\sum_{i \in U^t} v_i^{GK} q_i^t\right) / \left(\sum_{i \in U^0} v_i^{GK} q_i^0\right)}, \tag{2.3}$$

where,

$$v_i^{GK} = \sum_{z=0}^{T} \varphi_i^z \frac{p_i^z}{P_{GK}^{0,z}} \tag{2.4}$$

Here, U^t is the population of items (universe) for period *t* and φ_i^z is a quantity share $\varphi_i^z = q_i^z / \sum_{t=0}^{T} q_i^t$.

The Geary Khamis price index in (2.3) can be explained as a ratio of a turnover index and a quantity index, where the so called transformation coefficients v_i^{GK} are used to weight the quantities. The coefficients v_i^{GK} are meant to say something about relative quality: the ratio of v_i^{GK} and v_j^{GK} indicate how many quantities of item *j* can be considered equivalent to one quantity of item *i* (in terms of quality). As shown in (2.4), the $v_i^{GK\prime}s$ are computed as a weighted mean of deflated prices.

To compute the index, the equations (2.3) and (2.4) need to be solved simultaneously, because of an interdependency. The expression for $P_{GK}^{0,t}$ depends on v_i^{GK} and vice versa the expression for v_i^{GK} depends on $P_{GK}^{0,t}$. The GK-price index can be easily obtained by applying an iterative procedure, see e.g. Chessa (2016). An alternative, but practically more difficult way, consists of solving one algebraic equation. The reader is referred to Diewert (1999) and Diewert and Fox (2017) for more details. The use of transformation coefficients, or implicit

prices, is typical for a so-called Quality Adjusted Unit Value index (Dalén 1998 and 2001), or Generalized unit value index (Auer, 2014), a class of methods to which GK (and many others) belong to.
A variant of the GK method, the so called QU-method, is currently used to construct several components of the Dutch Consumer Price index. This approach combines Geary Khamis with product clustering as a solution to the relaunch problem.

2.3 Time Product Dummy (TPD)

A third index method in this paper is the Time Product Dummy (TPD), see e.g. Krsinich (2016), which can be considered the counterpart of the Country Product Dummy (CPD) model for cross-country comparisons.

This index is obtained by performing a log-linear regression model on the pooled data of all time periods 0,…,*T*. The regression model is formulated as follows:

$$\ln(p_i^t) = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{U-1} \gamma_i H_i \ + \epsilon_{it}, \tag{2.5}$$

The logarithm of the price, $\ln(p_i^t)$, is explained from a constant term, a time dummy D_i^t and a product dummy H_i . The time dummy variable D_i^t has the value one if the observation pertains to period *t*, ($t = 1, \dots, T$) and zero otherwise. Similarly, H_i is one if the observation relates to item *i* ($i = 1, \dots, U-1$) and zero otherwise. These product dummy variables are also known as 'fixed effects'. For a proper identification of the model, the dummies for an arbitrary item (*U*) and an arbitrary period (0) are excluded ($\delta^0 = 0$ and $\gamma_U = 0$). It is assumed that the errors ϵ_{it} are independently distributed with zero mean. Following Diewert's (2005) proposal, the regression equation is estimated by Weighted Least Squares (WLS), where the expenditure shares s_i^t are used as weights. The TPD index is obtained by exponentiating the estimated coefficients $\hat{\delta}^t$ for the time dummies:

$$P_{TPD}^{0,t} = \exp(\hat{\delta}^t). \tag{2.6}$$

As demonstrated by Rao (2005) the weighted TPD index can alternatively be written as follows:

$$P_{TPD}^{0,t} = \frac{\prod_{i \in U^t} \left(p_i^t / v_i^{TPD}\right)^{s_i^t}}{\prod_{i \in U^0} \left(p_i^0 / v_i^{TPD}\right)^{s_i^0}}, \tag{2.7}$$

where

$$v_i^{TPD} = \prod_{z=0}^{T} \left(p_i^z / P_{TPD}^{0,z}\right)^{w_i^z} \tag{2.8}$$

and $w_i^z = s_i^z / \sum_{x=0}^{T} s_i^x$.

The expressions (2.7) and (2.8) closely resemble the expressions (2.3) and (2.4) for Geary Khamis. Expression (2.7) computes a price index from the product prices and the transformation coefficients v_i^{TPD} and (2.8) derives the transformation coefficients from deflated prices . The equations (2.7) and (2.8) can be applied iteratively to arrive at the price index. There are however also slight differences in the computations between Geary Khamis and TPD: different weights are used and different means. Geary Khamis uses arithmetic means, whereas geometric means are used for TPD.

2.3 Relaunches

Now, let's see what happens when a relaunch occurs. Suppose that at some point of time, an item *i* is replaced by *i**.

Although the products *i* and *i** are essentially the same, the GEKS-Törnqvist does not capture the price change at the relaunch, because it relies on matched-models.

Geary Khamis uses the prices of all sold items, but also ignores that *i* and *i** are essentially the same. Consequently, different transformation coefficients (v_i^{GK} and v_{i*}^{GK}) are computed for both items. A large difference between both coefficients, might produce a discontinuity in the price index. Such a large break might be directly caused by a relaunch, since relaunches are often accompanied by a (large) price increase.

For TPD the ignorance of a relaunch translates into different product dummies for *i* and *i** with different coefficients in the regression model. Although both products closely resemble each other, the estimated coefficients can substantially differ. Thus, spurious index jumps might arise at the relaunch.

The above-stated is illustrated in the example below. Table 1 shows all prices. All quantities are the same, let's say ten. Here, Item 1 undergoes a relaunch. Its identifier changes from 1a to 1b, but the product remains the same.

Table 1. Example of a relaunch

	Prices			
Period	1	2	3	4
Item 1a	8	8	-	-
Item 1b	-	-	15	18
Item 2	10	10	10	10
Index GEKS- Törnqvist	1.00	1.00	0.97	1.03
Index GK	1.00	1.00	0.95	1.06
Index TPD	1.00	1.00	0.95	1.06

All indexes show a decrease at Period 3. This is counterintuitive, since there is no single item with a price decrease. Item 1b is introduced at a higher price than the last price for Item 1a. The price for Item 2 is constant.

The decrease of the GEKS-Törnqvist index at Period 3 can be explained from the 'index pair': $P_T^{1,4}P_T^{4,3} = 1.00 * 0.89 = 0.89$. All other pairs starting at 1 and ending at 3, i.e. $P_T^{1,1}P_T^{1,3}$, $P_T^{1,2}P_T^{2,3}$ and $P_T^{1,3}P_T^{3,3}$ are 1. The price increase due to the relaunch is not incorporated in any price pair, because of the different product identifiers.

The decrease of the GK index at period 3 can be explained from the different transformation coefficients. These are 8.26 and 15.98 for Items 1a and 1b. The shift towards the (supposedly) 'higher-quality item' 1b largely increases the quantity index, i.e. the denominator in (2.3). As this increase exceeds turnover growth, this pushes the price index down.

The decrease of the TPD index at period 3 can be explained from the estimated regression coefficients for the product dummies. The exponentiated coefficients are 1.00 and 2.05, meaning that product 1b is higher valued as product 1a. The shift towards 1b diminishes the price index.

3. Product clustering and imputation

Two well-known solutions for the relaunch problem for matched model methods are product clustering and imputation. Subsection 3.1 and 3.2 describe these techniques in general. Subsection 3.3 presents currently available imputation methods for GEKS-Törnqvist. Some new results on similarities between product clustering and imputation are presented in Subsection 3.4. Finally, Subsection 3.5 gives imputation alternatives for product clustering.

3.1 Product Clustering

Product clustering means that the most detailed products are clustered together into less detailed groups. Usually, "unit value indexes" are advocated for deriving price indexes at the product level, e.g. The CPI Manual (ILO *et al.* 2004a, Chapter 20), the 2008 System of National Accounts (SNA) and Balk (2005). A unit value price index computes a product price as a weighted mean of the item prices, where quantities are used as weights. The product quantity is simply the sum of the item quantities. In formula,

$$p_h^t = \frac{\sum_{i \in h} p_i^t q_i^t}{\sum_{i \in h} q_i^t} \tag{3.1}$$

and

$$q_h^t = \sum_{i \in h} q_i^t. \tag{3.2}$$

The subscript *h* will be used throughout this paper for the cluster. The summation $i \in h$ means 'taken over all items *i* that contribute to cluster *h*'. Mathematically, a cluster forms a partition of the items, meaning that all items contribute to exactly one cluster and each cluster contains at least one item. Advantages of product clustering are ease and broad applicability: it can be applied to many price index methods, at least to all methods from Section 2. A disadvantage is that unit value index is only appropriate for homogeneous items. As already mentioned in the introduction, severe unit value bias can result if product clustering is applied on heterogeneous items.

The impact of product clustering can be reduced in the *hybrid* approach in Chessa (2016). This means that 'stable' items – the items that exist on the market for each period - are not clustered. Only the other items, those that enter or leave the market, are combined into clusters. Here, clusters are defined afterwards, when all data have been collected.

3.2 Imputation

Imputation usually means that (possibly) unobserved prices are replaced by estimates. In the literature it is mainly used for methods that compute an index as (weighted) mean of item price changes, like Törnqvist. Where product clustering can be performed in one way – by unit value prices – imputation can be done using a range of methods. On the one hand, this offers more flexibility, on the other hand a nontrivial decision on the imputation method has to be made. Imputation and product clustering have in common that it might lead to error; where product clustering raises unit value bias; imputation gives estimation error. The artificial estimates added to the data and the error entailed by estimation have been reasons for reluctance at statistical offices to adopt imputation. (see e.g. Triplett, 2006). Imputation might give problems with the interpretation of the results, especially for seasonal items. This can happen for instance if a nonzero price is imputed for an out-of-season period. In that case, a non-sold can be found to have the largest impact on the price change. More about this can be found in the discussion in Section 6.

3.3 Existing imputation methods for Törnqvist

This subsection describes the available imputation methods for a Törnqvist index. Subsection 3.2.1 explains different strategies for choosing the prices to impute. Subsection 3.2.2. deals with estimation methods for the imputed prices.

3.2.1. Imputation strategies

The first step for an imputation method consists of determination which values must be imputed. De Haan and Krsinich (2014b) discuss a Single and Double imputation method for Törnqvist indexes (SI and DI, respectively). These indexes are given by

$$P_{SI}^{0,t} = \prod_{i \in U_M^{ot}} \left(\frac{p_i^t}{p_i^0}\right)^{(s_i^0+s_i^t)/2} \prod_{i \in U_D^{ot}} \left(\frac{\hat{p}_i^t}{p_i^0}\right)^{s_i^0/2} \prod_{i \in U_N^{ot}} \left(\frac{p_i^t}{\hat{p}_i^0}\right)^{s_i^t/2} \tag{3.3}$$

and

$$P_{DI}^{0,t} = \prod_{i \in U_M^{ot}} \left(\frac{p_i^t}{p_i^0}\right)^{(s_i^0+s_i^t)/2} \prod_{i \in U_D^{ot}} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0}\right)^{s_i^0/2} \prod_{i \in U_N^{ot}} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0}\right)^{s_i^t/2} \tag{3.4}$$

where, U_M^{ot}, U_D^{ot} and U_N^{ot} refer to the sets (or universes) of 'mutual existing', 'disappearing' and 'new' items. Mutual existing are items that have been purchased in periods 0 and t. Disappearing items have been in period 0. New items are the ones that have come into existence in t. The hats in $\hat{p}_i^t$ and $\hat{p}_i^0$ mean that a price has been imputed.

The single imputation method compares the prices of new and disappearing items with an estimated price for the period without observation. A main advantage of the SI-method is that it exploits all observed prices. As mentioned by de Haan and Daalmans (2019): "This so-called single imputation method is a natural choice as it restricts imputations to the missing prices and leaves unaffected all the observed prices, both for unmatched and matched items". The double imputation method, as proposed by De Haan (2004a) and Hill and Melser (2008), imputes for new and disappearing items the two prices of the base and comparison periods. The motivation for this is that there might be bias in the imputations. It is hoped that the base and comparison period are affected in the same way, so that possible bias (partly) cancels out. Replacement of observed prices by estimates, can however be considered a drawback of the method. Single and Double imputation are not the only possibilities for a Törnqvist index. Silver and Heravi (2007) discussed a so-called full imputation method that imputes all observed and nonobserved prices. Their application was particularly intended for products whose characteristics might change between two successive periods, like houses.

3.2.2. Imputation methods

We now arrive at the question how the imputed values $\hat{p}_i^t$ can be determined for a direct Törnqvist index. Several methods are available. Three alternatives are discussed below. The first two are existing variants of the hedonic regression method, the third alternative is a novel one and uses unit values.

1. Regression imputation - unilateral

This method is based on the following hedonic regression model (see e.g. De Haan and Krsinich, 2014b)

$$\ln(p_i^t) = \alpha^t + \sum_{k=1}^{K} \beta_k^t z_{ik} + \epsilon_{it}. \tag{3.5}$$

This model explains the logarithmic of the price from a constant term and *K* product characteristics. Here, z_{ik} denotes the quantity of the *k*-th characteristic (*k*=1,…,*K*) for item *i*, which is assumed time invariant and ϵ_{it} is a zero-mean error term. The model in (3.5) is estimated from the observed prices for period *t*, using weighted least squares (WLS) with expenditure shares s_i^t as weights. The imputations are given by $\hat{p}_i^t = \exp(\hat{\alpha})\exp(\hat{\delta}^t)\exp(\sum_{k=1}^{K}\hat{\beta}_k z_{ik})$, where $\hat{\alpha}^t$ and $\hat{\beta}^t$ are estimated coefficients.

2. Regression imputation - bilateral

De Haan (2004b) proposed the following 'bilateral' model

$$\ln(p_i^t) = \alpha + \delta^1 D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \epsilon_{it}, \tag{3.6}$$

that should be estimated from the pooled data of periods 0 and *t*. Here, D_i^t is an indicator that takes the value one for period *t* and zero for period 0. This model is estimated by WLS, where the weights are given by average expenditure shares, i.e. $(s_i^0 + s_i^t)/2$ for matched items and $(s_i^0/2)$ or $(s_i^t/2)$ for unmatched ones. The imputations are given by $\hat{p}_i^t = \exp(\hat{\alpha})\exp(\hat{\delta}^t)\exp(\sum_{k=1}^{K}\hat{\beta}_k z_{ik})$ and $\hat{p}_i^0 = \exp(\hat{\alpha})\exp(\sum_{k=1}^{K}\hat{\beta}_k z_{ik})$. The bilateral regression method has the potential of being more efficient than the model in (3.5), since more data are used for the estimation of the regression coefficients. On the other hand, the one-period regression method can be considered more appropriate if the relation between the price and the product characteristics rapidly changes.

3. Unit value imputation

Unit value imputation means that a unit value price is used for imputation

$$\hat{p}_i^t = p_h^t = \frac{\sum_{i\in h} p_i^t q_i^t}{\sum_{i\in h} q_i^t}. \tag{3.7}$$

where *h* is the cluster that contains *i*. To the best knowledge of the author, unit value imputation has not been proposed before, despite of the easiness of the approach. A further reason for considering this method is a similarity with clustering, which will be explained in

Subsection 3.3. Usually, regression includes the main effects of the auxiliary variables only. The unit value method in (3.7) on the other hand uses stratification cells that are demarcated by a full interaction of the stratification variables. If interaction effects were actually important in the determination of a price, unit value imputation can be expected to perform better. On the other hand, the regression approach seems to be more appropriate for problems with many available auxiliary variables. The unit value approach might require a choice between auxiliary variables to prevent an insufficient number of observations for proper estimation of the model and the same applies to product clustering.

In addition to three methods above, several other imputation methods exist. For instance, some authors in the literature argue that the fact that an item has not been sold means that the prices should be relatively high. This is the main idea behind the so-called 'Hicksian reservation prices' (Hicks, 1940), which should be estimated from an econometric model. We will however not elaborate on this in the remainder of this paper. Also often mentioned is the *carry-forward* method which simply imputes the last observed price. This method can be extended by a correction for inflation, e.g. Diewert (2018, ch. 5). For brevity, we do not consider these methods in the remainder of this paper.

3.3 Relation between product clustering and imputation

Imputation and product clustering are often presented as two distinct methods, each with their own properties. This subsection shows however that there is a close relation between both. It can be shown that the results from product clustering can also be achieved by some specific form of imputation, at least for all matched-model methods from Section 2. Product clustering replaces a group of items by one 'cluster' product, where the price for the cluster is set to an average (unit value) price of the underlying items. Alternatively, one can keep the original items and replace each item's price by the unit value price of the cluster. As shown below, both approaches can produce the same results. The relations between both methods are paraphrased below. The reader is referred to Appendix A for clarification and proofs.

- *Bilateral Törnqvist*: Product clustering is equivalent to a fully imputed price index, in which all observed and non-observed prices are imputed by unit value prices of their cluster.
- *Geary Khamis* and *TPD*: Product clustering is identical to imputation of all observed and unobserved prices and quantities. Prices are imputed by a unit value price ($\hat{p}_i^t = p_h^t$), where *h* is the cluster that contains i. Quantities are imputed by the total quantity share of the cluster quantity q_h^t. That is $\hat{q}_i^t = q_h^t(Q_i/Q_h)$, where $Q_i = \sum_t q_i^t$ and $Q_h = \sum_t q_h^t$.

Hence, in order to obtain the same results as for product clustering, prices need to be imputed for Törnqvist, while prices and quantities need to be imputed for TPD and Geary Khamis. Unit value price are imputed for all items that belong to a cluster. Quantities are imputed such that the relative quantity shares remain constant over time for all items within a cluster.

Clustering means that a set of products is replaced by a single 'cluster' product. The previous relations mean that imposing a single cluster product is equivalent to the use of several products, each with the same cluster price and for Geary Khamis and TPD also the same quantity shares. The relation between imputation and product clustering has similarities with the consistency in aggregation property. This property is met if the values of an index calculated in two stages, i.e. by first calculating separate indices for sub-components and then aggregating them, coincides with the value of an index computed in one stage. As explained above, prices and quantities can be imputed such that that this property is met.

The foregoing sheds further light on the properties of product clustering. It shows that this method is rather rigorous. For instance, for Geary Khamis and TPD, product clustering is equivalent to replacing prices and quantities: not only unobserved data are replaced by estimates, but also the observed ones. As it is counterintuitive to override observed values, the following subsection presents a new method that preserves the observed prices.

3.4 Imputation alternatives for product clustering

Last subsection showed that product clustering produces results that can also be established by a rather rigorous imputation method. This subsection exploits the versatility of imputation methods to find alternatives for product clustering. The idea is to develop a new imputation method that mimics product clustering, while keeping the distortion to the observed values to a minimum. The new method is supposed to obey the following conditions:

- For a fixed population, in which all items are sold each period, results are the same as for the original index without correction;
- Imputed prices and quantities for unsold products are the same as the ones implied by product clustering;
- For each product cluster, the unit value prices are the same as for the uncorrected (GK or TPD) index;
- For each product cluster, total cluster quantities are the same as for the uncorrected index.

Product clustering method does not satisfy the first property, which is a cause for avoidable unit value bias. The first two properties imply that the results of the new method can be expected to lie somewhere in between "product clustering" and the "no correction scenario".

If almost all items were imputed, the result would be similar to product clustering. If no items were imputed, the results would be the same as for the original method (without correction).

It is easily verified that a Single Imputation Method combined with unit value imputation satisfies all above-stated properties for GEKS-Törnqvist.

Below, a novel imputation method is proposed for Geary-Khamis and TPD that also satisfied the mentioned properties. This new method imputes the prices and quantities as follows:

$$\hat{p}_i^t = \begin{cases} p_i^t & \text{if } p_i^t \text{ is available} \\ p_h^t & \text{otherwise} \end{cases} \tag{3.8}$$

As before p_h^t is the unit value price. So, original prices are used, if available, otherwise a unit value price is imputed. Further, we have

$$\hat{q}_i^t = \begin{cases} q_i^t \left(\sum_{i \in h} I_q(i,t)\, Q_i \,/ Q_h \right) & \text{if } q_i^t \text{ is available} \\ q_h^t (Q_i / Q_h) & \text{otherwise} \end{cases} \tag{3.9}$$

where $I_q(i,t)$ is an indicator function, which is one if q_i^t is available and zero otherwise. If all item prices within a cluster are available, we get $\hat{q}_i^t = q_i^t$, so that the observed quantities are used for the index compilation. If some items within a cluster have not been purchased, all observed quantities are reduced by the same factor and a nonzero quantity is imputed for the unobserved products. It is easily verified that the imputation method given by (3.8)-(3.9) satisfies the aforementioned properties.

Example

Consider the four-period example in Table 2 below. Three different items can be distinguished. Item 1 is subject to a relaunch. It is called Item 1a before and Item 1b after. Table 3 shows the Geary Khamis and TPD indexes for different scenarios. In the first scenario no correction has been made for the relaunch. The second and third show the results after product clustering and imputation. The bottom line shows results based on the "true" indexes, in which items 1a and 1b are considered the same. Here, the 'imputation' index closely approximates the 'true' index, while product clustering leads to much deviation from the original result. Doing no-correction is even better than product clustering. The problem is that clustering produces severe unit value bias. For instance, all period 4 prices are lower than the period 3 prices. The 'product clustering' still yields an increased price index, that can be explained from a shift of market share from the relatively cheap product 4 to the more expensive product 3.

Table 2. Available prices and quantities

	Prices				Quantities			
	1	2	3	4	1	2	3	4
Item 1a	12	11	-	-	10	5	-	-
Item 1b	-	-	16	15	-	-	10	5
Item 2	20	19	19	15	5	15	10	30
Item 3	8	7	7	5	20	20	20	10

Table 3. Results for Geary-Khamis and TPD

	Geary-Khamis				TPD			
Period	1	2	3	4	1	2	3	4
No correction	100.00	90.73	88.36	71.84	100.00	90.91	88.93	72.36
Clustering	100.00	110.53	112.83	117.69	100.00	110.53	112.83	117.69
Imputation	100.00	95.79	100.98	82.11	100.00	94.97	100.30	81.83
True (no relaunch)	100.00	96.27	103.22	80.89	100.00	95.65	102.86	80.77

4. Product matching

Product replacement means that disappeared products are matched with new products. Whenever an item leaves the market, a new item is chosen to replace that item and both items are considered the same. The main challenge is to find appropriate replacements. Ideally, a replacement item has similar characteristics as the items they replace. Text mining of the product descriptions can be applied to find the replacements, combined with expert judgement. Such a semi-automated system might require a lot of efforts, though. The current section proposes a very easy to apply replacement method. It basically assumes that all items within a stratum are perfect substitutes. This might not be fully realistic for many applications, but a similar assumption is also made in product clustering.

The proposed method matches new and disappeared products, but does not apply any correction to temporarily unsold products. All disappeared and new products within a stratum are matched, provided that at least one new and one disappeared product is available. Otherwise, no matching is done. If the number of new (disappeared) items exceeds the number of disappearing (new) items, disappeared (new) items are replicated until the point that the number of new and disappearing items are the same. Replication means that an item is created with the same prices as the replicated item. The quantity is equally divided by the number of replications, thus leaving the total quantity constant. For the purpose of replicability, the matching of disappeared and new items is not done randomly, but based on average turnover share. The new item with the highest average turnover share is matched with the disappearing item with the highest average turnover share. The same occurs for the items with the second highest turnover share and so on. The choice of average turnover share has been made arbitrarily, several other alternatives are possible as well. For instance, for

Geary Khamis implicit prices can be considered. A formal description of the algorithm applied in this paper is stated in Appendix B.

5. Empirical evaluation

This section presents an empirical study that compares product clustering, imputation and matching. The main aim of the study is to find out whether these methods adequately deal with relaunches. We use data sets for three product categories: TV's, chocolate and potatoe products.

The TV data set is the same as in De Haan and Daalmans (2019). It contains 17 months of scanner data on 313 TVs sold by a Dutch retail chain; where online sales are excluded. Items are identified by European Article Number, the European version of GTIN, and item prices are calculated as unit values across all the stores belonging to this retail chain. Twelve product characteristics are available (including screen size, processor type, and brand). The chocolate and potato data sets contain 12 month data for respectively 11,004 and 3,374 products. These data sets have been actually used for the CPI production. The EAN is used as a product identifier. Only one product characteristic is available, i.e. "product quantity".

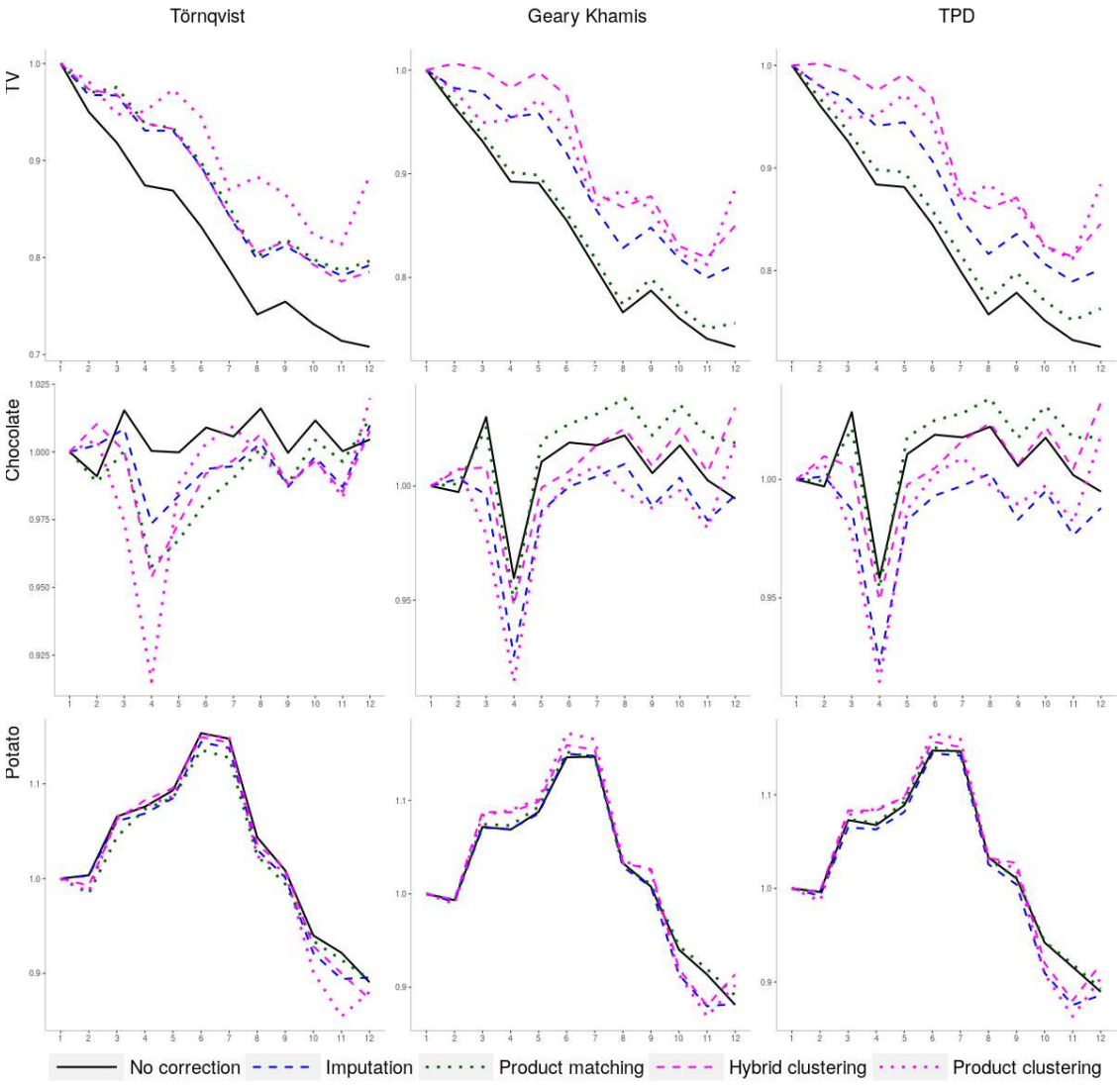
Subsection 5.1 empirically compares product clustering, imputation and matching. The true composition of the relaunches is not known for these data. This means that we are able to compare the impact of the corrections but unable to conclude which method works best. Subsequently, in Subsection 5.2, we simulate relaunches in the data ourselves. Because we are in full control of all breaks in the data, we can check whether the correction methods produce satisfactory results.

5.1 Effects of the correction methods on the original data

This subsection compares clustering, hybrid clustering, imputation and product matching for three data sets (TV, chocolate, and potatoes) and three multilateral methods (GEKS-Törnqvist, TPD and Geary Khamis). For brevity, one imputation method is considered per index method. For GEKS-Törnqvist, this is the regression based Single Imputation method, as we will see later that this method gives the most accurate results for the data under consideration. For Geary Khamis and TPD, this is the new method from Subsection 3.4. The same stratification is used for all methods. For the TV data set screen size is used. Although a finer stratification would be better for TV's, it will become clear later why this choice has been made.

The product strata for chocolate and potato are derived from 'product quantity'. The total number of product strata amounts to 4, 16 and 10 for TV's, chocolates and potatoes. All 12 available product characteristics in the TV data are used as auxiliary variables in the imputation method. For the other two data sets no other auxiliary information is used besides the stratification variable. Figure 1 illustrates the results of the correction methods on three data sets.

Figure 1. Results of different correction methods on three data sets



In general, differences in results between the correction methods can be large. For TV's each correction method leads to an upward shift of the price index, which is to be expected, as relaunches are known to be accompanied by price increases. For chocolates the effect is a bit

ambiguous. For potatoes the effect is relatively small, implying that the price indexes are hardly affected by relaunches.

Product clustering has a relatively large impact for TV's and the chocolates. It does not only shift the price index, but also affects the shape of the graphs, i.e. the short term movements, thus suggesting the presence of unit value bias. Imputation has a lower impact than clustering for most cases, especially for TV's. Product matching has a relatively large impact on GEKS-Törnqvist, but a lower influence on Geary Khamis and TPD. The first follows from an increase of the number of matched 'items' in each pair of periods. The latter can be explained because it does not alter the data set much: it neither adds nor adjusts prices and quantities.

5.2 Simulated relaunches

We now arrive at the main question of this section: whether imputation, product clustering and matching adequately solve the relaunch problem. To answer this, relaunches were simulated in a fixed population, or static universe, context. Static universes for the three data sets have been artificially created by selecting products that have been sold in all periods under consideration. In the television data we focus on 12 months and all 54 out of 336 GTIN's that have been sold in each of these 12 months. Similarly, 2,661 out of 11,004 GTIN's have been selected for chocolates and 284 out of 3,374 for potatoes. The total turnover shares of the selected items over all periods are 55%, 57% and 71% for TV's, chocolates and potatoes. Four scenarios are considered. In the first, relaunches are added randomly. In a second scenario, the probability of a relaunch depends on price change. In the third scenario, relaunches are accompanied by extraordinary, artificially added price increases. The fourth scenario deals with relaunches in a single month. The scenarios are described in more detail below.

Scenario 1: Random relaunches. A relaunch occurs in 20 percent of the cases. This means that a product is replaced by a different product with the same characteristics. Relaunches have been randomly selected among all periods and products.

Scenario 2: Selective relaunches. A relaunch randomly occurs in 20 percent of the cases in which a price has been observed in two adjacent periods, but only those periods are considered for which the price change is above the median value of all price changes for all products.

Scenario 3: Relaunches with additional price increase. Same as scenario 2, but additionally, each relaunch goes along with a simulated, permanent price increase by 20%.

Scenario 4: Relaunches occur at one time period for randomly selected 75% of all items. All relaunches take place at month 4 and lead to a permanent 20% price increase.

Scenarios 1 and 2 compare the results of a correction method – imputation, clustering and matching – with those of the original data, without simulated relaunches. The closer the approximation, the better. Scenarios 3 and 4 artificially add a price increase after each relaunch. The benchmark series do not include the relaunches, but do contain the added price increases. Thus, it is assumed that these price changes have actually happened. It is checked again whether a correction method is capable to correct the improper treatment of relaunches.

All scenarios have been applied 100 times. This number has been pragmatically chosen and not evaluated. The following evaluation criteria have been used:

- Median absolute difference: median difference of a corrected index and the benchmark (original) index, i.e. median $\{\text{abs}(P_{scenario}^{0,t} - P_{benchmark}^{0,t})\}$, computed over all combinations of 12 periods and 100 replications;
- 90th percentile of the absolute difference: same as above, but the median is replaced by the 90th percentile.

The first criterion says something on bias. The difference between the second and the first criteria tells something about the dispersion of these results.

We consider a 'base' scenario, in which no correction method has been applied and we compare it with product clustering, imputation and matching. As before, one imputation method is considered per index method. For GEKS-Törnqvist, this is the Single Imputation regression method. For Geary Khamis and TPD, this is the new method from Subsection 3.4.

The need for a correction method becomes especially pertinent for Scenarios 3 and 4, where relaunches are accompanied by extraordinary price changes. On the other hand, if relaunches occur more randomly, then correction methods can do more harm than good, as shown in the results in Tables 4 and 5 for Scenario 1 on the TV and chocolate data. Here, "no correction" appears to be the best option. For the potatoes this is opposite: all correction methods work better than no correction. An explanation lies in the fact that the stratifications cells are more homogeneous, which eases the correction possibilities. For the Scenarios 2,3 and 4 imputation turns out to be the best method for GEKS-Törnqvist on TV data and product matching for most other cases. The imputation method benefits from the auxiliary information that is available for TV's, additional to the variables used for stratification. Such additional information is not used for the other data sets and it is not exploited by other correction methods either.

Table 4. Results for four scenarios of simulated relaunches, median abs. difference

	---------TV's---------				--------Chocolates--------				--------Potatoes--------			
	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4
GEKS-Törnqvist												
No correction	1.18	2.29	18.70	11.14	0.37	1.64	24.83	14.77	1.96	3.32	29.53	13.33
Imputation	0.65	0.64	1.69	1.24	0.47	0.74	1.60	1.83	1.78	1.42	0.99	0.83
Matching	1.38	1.09	1.71	2.10	0.46	0.39	0.46	0.20	0.48	0.34	0.76	0.95
Hybrid clust.	4.53	4.31	6.57	4.18	0.76	0.68	1.21	0.74	1.21	1.16	1.33	1.10
Clustering	4.56	4.56	7.42	5.63	0.78	0.78	1.43	1.01	1.27	1.27	1.56	1.44
Geary Khamis												
No correction	1.32	1.59	18.75	7.39	0.16	1.67	24.98	11.90	1.40	3.46	30.28	10.42
Imputation	4.74	4.76	7.32	3.81	0.94	0.63	1.03	0.34	1.31	1.05	1.28	2.06
Matching	1.48	1.32	2.40	1.44	0.58	0.52	1.13	4.09	0.53	0.53	0.77	1.58
Hybrid clust.	5.23	5.04	8.07	4.81	1.02	0.90	1.36	0.85	1.28	1.31	1.51	1.31
Clustering	5.52	5.52	9.01	6.66	1.07	1.07	1.74	1.25	1.37	1.37	1.60	1.54
TPD												
No correction	0.74	2.42	18.98	10.54	0.20	2.41	25.11	14.61	1.22	5.09	31.26	13.40
Imputation	4.83	4.80	6.76	5.11	1.47	1.30	1.41	3.04	1.15	0.91	1.40	1.28
Matching	2.33	1.80	2.73	3.62	0.98	0.82	0.57	0.54	0.51	0.32	0.71	1.05
Hybrid clust.	5.47	5.23	7.56	4.76	1.00	0.87	0.89	0.81	1.26	1.25	1.78	1.20
Clustering	5.63	5.63	8.52	6.70	1.03	1.03	1.06	1.04	1.37	1.37	2.02	1.50

The best method for each data set and index is underlined

Table 5. Results for four scenarios of simulated relaunches. 90th percentile, abs. difference

	---------TV's---------				--------Chocolates--------				--------Potatoes--------			
	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4
GEKS-Törnqvist												
No correction	3.27	5.17	37.82	14.69	1.00	3.32	45.94	15.78	5.50	7.73	43.23	19.40
Imputation	1.55	1.48	2.67	1.82	1.28	1.36	4.51	2.35	3.01	2.31	2.54	1.65
Matching	3.40	2.62	3.94	3.86	1.19	1.01	1.44	0.57	1.76	1.35	2.41	1.89
Hybrid clust.	8.41	8.19	10.83	7.89	2.04	1.95	2.62	1.97	2.37	2.34	2.98	2.55
Clustering	8.47	8.47	11.55	10.38	2.08	2.08	2.86	2.51	2.47	2.47	3.13	3.04
Geary Khamis												
No correction	3.04	3.15	37.49	10.29	0.52	2.93	45.33	12.81	3.80	7.88	43.04	15.50
Imputation	7.67	7.62	11.02	6.61	1.92	1.71	2.30	1.49	3.01	2.65	3.58	3.59
Matching	3.75	3.58	6.53	4.66	1.65	1.60	2.60	4.95	1.81	1.68	2.65	3.49
Hybrid clust.	7.93	7.75	11.99	8.10	2.14	2.08	2.84	2.10	3.18	3.16	4.29	3.62
Clustering	7.90	7.90	12.67	10.00	2.26	2.26	3.27	2.69	3.14	3.14	4.59	4.08
TPD												
No correction	1.95	4.65	39.37	14.09	0.56	4.43	46.81	15.74	3.38	9.60	45.09	18.02
Imputation	7.91	7.85	10.26	7.68	2.24	2.05	2.51	3.80	2.81	2.34	3.70	2.51
Matching	5.03	4.16	6.38	6.33	1.69	1.47	1.54	1.23	1.83	1.45	2.62	2.65
Hybrid clust.	8.19	7.96	11.21	7.98	2.07	1.99	2.40	2.04	3.12	3.11	4.65	3.57
Clustering	8.12	8.12	11.97	9.99	2.10	2.10	2.65	2.60	3.00	3.00	5.21	4.10

The best method for each data set and index is underlined

The results further show that standard clustering and hybrid clustering work bad for TV's. Unit value bias is a serious problem for these data and the chosen stratification therein. For the other data sets, product clustering produces better results, indicating that unit value bias is much less of a problem. But, also for these other data sets, other correction methods perform better, especially 'product matching'.

Although the good results for product matching are very clear, these might be too optimistic because relaunches have been artificially created in the test data by replacing one product by another. Since this is operation is exactly adverse to product matching, it is already known beforehand that product matching might work. Reality can be more diffuse. Usually, it is unknown whether a data set contains products that can be properly combined in a one-to-one way. To reduce this advantage, broadly defined product strata have been chosen, which reduces the fraction of properly matched products. The fraction of items that have been matched to a different item in the original data sets amounts to 75%-90% for TV's, 93-94% for chocolates and 60%-66% for potatoes. These percentages differ by scenario. The results for the chocolate data are especially promising for other real-life applications. On the one hand, only a small fraction of products have been matched exactly the same as in the original data. On the other hand, the results for this data set are still good.

So far, only one imputation method has been considered for each index method. Below, different imputation methods for GEKS-Törnqvist from Section 3 are compared. These methods include unilateral and bilateral regression as well as unit value imputation, combined with a single or double imputation approach. Tables 6 and 7 show that a unilateral regression method for single imputation works best for most applications. This is why this method was included in the previous evaluation.

Table 6. Results for four scenarios of simulated relaunches, median abs. difference

	---------**TV's**---------				--------**Chocolates**--------				---------**Potatoes**--------			
	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4
Single imp.												
Regr. unilat.	0.65	<u>0.64</u>	<u>1.69</u>	<u>1.24</u>	0.47	0.74	<u>1.60</u>	<u>1.83</u>	1.78	1.42	<u>0.99</u>	0.83
Regr. bilat.	<u>0.44</u>	0.88	4.38	2.82	0.35	<u>0.68</u>	3.21	2.97	1.53	0.91	1.66	0.74
Unit value	3.03	3.06	6.01	4.11	0.68	0.90	2.46	2.05	1.20	<u>0.76</u>	0.99	<u>0.46</u>
Double imp.												
Regr. unilat.	0.64	0.74	2.44	1.64	0.44	0.99	4.13	3.57	1.78	0.95	2.31	0.86
Regr. bilat.	0.49	1.26	7.38	4.63	<u>0.25</u>	0.71	7.83	5.86	1.35	0.85	7.29	3.35
Unit value	2.87	3.44	8.72	5.00	0.48	0.91	5.81	3.88	<u>1.05</u>	0.68	4.26	1.53

The best method is underlined. Regr = regression; Unilat = unilateral, bilat = bilateral

Table 7. Results for four scenarios of simulated relaunches, 90th percentile abs. difference

	---------TV's---------				--------Chocolates--------				--------Potatoes--------			
	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4	Sc.1	Sc.2	Sc.3	Sc.4
Single imp.												
Regr. unilat.	1.55	1.48	2.67	1.82	1.28	1.36	4.51	2.35	3.01	2.31	2.54	1.65
Regr. bilat.	1.15	1.53	8.23	3.67	1.04	1.17	7.71	3.52	2.86	1.88	5.04	1.57
Unit value	6.09	5.74	8.61	6.21	1.53	1.45	5.36	2.50	2.18	1.66	3.17	1.07
Double imp.												
Regr. unilat.	1.55	1.58	4.00	2.41	1.26	1.55	9.90	4.13	3.48	2.31	6.84	2.11
Regr. bilat.	1.32	2.20	14.44	5.92	0.79	1.45	16.65	6.44	3.28	2.32	14.35	5.51
Unit value	5.98	6.02	11.25	7.09	1.25	1.46	12.03	4.34	2.19	1.84	10.45	3.50

The best method is underlined. Regr = regression; Unilat = unilateral, bilat = bilateral

6. Discussion

Matched-model indexes are often used for today's index compilation in official statistics. It is well known that these methods do not get along well with relaunches, i.e. products that leave the market and return with slightly difference appearances. A way to cope with relaunches is to combine similar items into product clusters. This means that a price index is computed from a product cluster rather than from their underlying items. Product clustering relies on average "unit value" prices. The literature discourages unit value prices for heterogeneous items. At the same time, it has been noted that some degree of inhomogeneity cannot easily be avoided in statistical practice. Thus, some kind of contradiction is involved with the awareness of unit value bias on the one hand and the application of product clustering on the other hand. This contradiction can be partly explained from the fact that some degree of clustering might be practically inevitable in the construction of data sets for index compilation. For instance, it can be necessary to cluster daily data into weekly or monthly data (aggregation in time), or to cluster individual stores into a chain store level. If the first level of aggregation is done on narrowly defined products, unit value bias should not be too much of an issue. The risks of product clustering are the highest for the second step in which price indices are compiled from a set of essentially different products.

This paper has empirically compared several correction methods. In the analysis relaunches have been artificially created in three data sets and it has been verified which method best mitigates the negative effects of relaunches. This application has demonstrated that product clustering might lead to serious unit value bias, depending on the data and the homogeneity of the product clusters.

A first alternative for product clustering is so-called hybrid clustering. This approach only applies clustering to the selection of items that have not been purchased at the entire estimation window. Although less drastic than 'full' clustering, unit value bias might still occur

due to the replacement of items with actually observed prices by product clusters with unit value prices. The empirical results in this paper show that hybrid clustering slightly reduces the bias from 'full' product clustering.

Another alternative for product clustering is imputation. Imputation means that possibly unobserved prices are replaced by estimates. Different estimation techniques give rise to different methods. Product clustering and imputation are often presented as entirely different approaches. It has been shown in Section 3 that for three index methods, GEKS-Törnqvist, Geary Khamis and TPD, an imputation method can be constructed that give exactly the same results as product clustering. The unit value bias that might arise for product clustering corresponds to a rather rigorous substitution of prices in the imputation approach. Not only unobserved prices are replaced by estimates, but also observed ones, which does not make sense intuitively. Subsection 3.4 has proposed alternative imputation methods that behave similarly as clustering, although with much less distortion to the data. Specifically, observed prices are not adjusted.

In the empirical evaluation, imputation mostly performs a better correction for simulated relaunches, especially for GEKS-Törnqvist. Despite this good performance, the imputation approach raises serious practical concerns. A transaction data set provides the complete picture of all sales of a shop. The requirement for a statistical method to add information to integral data seems superfluous. Besides this intuitive argument, the interpretation of the added data also poses problems, particularly for seasonal items. Imputation might mean an addition of a nonzero price for an out of season product, e.g. for Easter eggs in September. The imputed price should be interpreted as a price of a comparable replacement product, but this is difficult to explain. Statistical institutes compute so called impact and contributions to determine which products affect a price index the most. After imputation, it might happen that a non-sold product is found to be a main driver of the price change between two periods, which would also be very difficult to explain. On the other hand, in product clustering all detailed product information gets lost; the data cannot be analyzed anymore below the cluster level.

Another method considered in this paper is product matching. Each disappearing product is matched to a new product and both are considered the same. In contrast to the previous methods, it neither adjusts nor adds any data. Product matching had the best results in our empirical evaluation. Complications with the interpretation of results may emerge; although these seem less severe than those of imputation. It is unclear for instance which name should be given to a matched product: the name of the first product or the name of the second product. This problem seems however less relevant when applied to narrowly defined strata. Another disadvantage of matching is that a matching method relies on arbitrary choices, e.g.

which item is matched to another item? The matching method in this paper has been developed on ad-hoc basis. It can be further developed in the future.

A limitation of this study is the limited amount of test data. The empirical results can be further extended in the future. The simulation framework as introduced in the current paper can be reused for future evaluations.

As a final conclusion, the popularly applied clustering method has the advantages of easiness, broad applicability and good interpretability. This method can easily lead to unit value bias, that can remain hidden to a practitioner who is unaware of this problem. To avoid unit value bias, it is crucial that the method is applied to homogeneous product strata. The so called MARS method can be applied to define these strata, see Chessa (2021). To find out whether unit value bias is present in a given data set and product stratification, a price index after product clustering can be compared with the index without clustering. To filter out the effects of relaunches, a comparison should be preferably conducted on the fixed population of products that have been sold each period. As demonstrated in this paper, imputation and especially product matching reduce the error arising from relaunches. Although these methods have been less well established and the interpretability might be an issue, the empirical results in this paper demonstrate that these methods reduce estimation error.

References

Auer, L. von (2014), The Generalized Unit Value Index, *Review of Income and Wealth*, 60, 843-861.

Australian Bureau of Statistics (ABS) (2017), *An Implementation Plan to Maximise the Use of Transactions Data in the CPI*, Information Paper 6401.0.60.004, ABS, Canberra, Australia.

Balk, B.M. (2005), Price Indexes for Elementary Aggregates: The Sampling Approach, *Journal of Official Statistics*, 21, 4, 675–699.

Caves, D.W., L.R. Christensen and W.E. Diewert (1982), The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity, *Econometrica* 50, 1393-1414.

Chessa, A. (2016), A new methodology for processing scanner data in the Dutch CPI, *EURONA*, 1, 49-69.

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

Dalén, J. (1998), *On the Statistical Objective of a Laspeyres' Price Index*, Paper presented at the fourth meeting of the Ottawa Group, 22-24 April 1998, Washington D.C., USA.

Dalén, J. (2001), *Statistical Targets for Price Indexes in Dynamic Universes*, Paper presented at the sixth meeting of the Ottawa Group, 2-6 April 2001, Canberra, Australia.

Dalén, J. (2017), *Unit Values and Aggregation in Scanner Data – Towards a Best Practice*, Paper presented at the 15th meeting of the Ottawa Group, 10-12 May 2017, Altville am Rhein, Germany.

Diewert, W.E. (1999), Axiomatic and Economic Approaches to International Comparisons. In A. Heston and R.E. Lipsey (eds.), *International and Interarea Comparisons of Income, Output and Prices , Studies in Income and Wealth*, Vol. 61, pp.13-87. Chicago: University of Chicago Press.

Diewert, W. E. (2005), Weighted Country Product Dummy Variable Regressions and Index Number Formulae, *Review of Income and Wealth*, 51, 561–570.

Diewert, W. and P. Lippe (2016), Notes on Unit Value Index Bias. *Jahrbücher für Nationalökonomie und Statistik*, 230, 690-708.

Diewert, W.E., and K.J. Fox (2017), *Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data*, Discussion paper *17-02*, Vancouver School of Economics, The University of British Columbia, Vancouver, Canada.

Diewert, W.E. (2018), *Scanner Data, Elementary Price Indexes and the Chain Drift Problem*, Discussion Paper 18-06, Vancouver School of Economics, University of British Columbia, Vancouver, Canada

de Haan, J. (2004a), Direct and Indirect Time Dummy Approaches to Hedonic Price Measurement, *Journal of Economic and Social Measurement*, 29, 427-443.

de Haan, J. (2004b), *The Time Dummy Index as a Special Case of the Imputation Törnqvist Index*, paper presented at The Eighth Meeting of the International Working Group on Price Indices (the Ottawa Group), Helsinki, Finland.

de Haan, J. and H.A. van der Grient (2011), Eliminating Chain Drift in Price Indexes Based on Scanner Data, *Journal of Econometrics*, 161, 36-46.

de Haan, J. and F. Krsinich (2014a), *Time Dummy Hedonic and Quality-Adjusted Unit Value Indexes: Do They Really Differ?*, Paper presented at the Society for Economic Measurement Conference, 18-20 August 2014, Chicago, U.S

de Haan, J. and F. Krsinich (2014b), Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes, *Journal of Business & Economic Statistics*, 32, 341-358.

de Haan, J. and J. Daalmans (2019), *Scanner Data in the CPI: the imputation CCDI Index revisited*, working paper, Statistics Netherlands, The Hague.

Hicks, J.R. (1940), The Valuation of the Social Income, *Economica* 7, 105-140.

Hill, R. and D. Melser (2008), Hedonic Imputation and the Price Index Problem: An Application to Housing, *Economic Inquiry* 46, 593-609.

Inklaar, R. and W.E. Diewert (2016), Measuring Industry Productivity and Cross Country Convergence, *Journal of Econometrics* 192, 426-433.

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, International Labour Office.

Johansen, I. and R. Nygaard (2021), *Ny beregningsmetode for prisindeks for matvarer og alkoholfrie drikkevarer i KPI*. Statistics Norway.

Krsinich, F. (2016), The FEWS index: fixed effects with a window splice. *Journal of Official Statistics*, 32, 375-404.

Leclair, M., I. Léonard, G. Rateau, P. Sillard, G. Varlet and P. Vernédal (2019), Scanner Data: Advances in Methodology and New Challenges for Computing Consumer Price Indices. *Economie et Statistique / Economics and Statistics*, 509, 13-29.

Radjabov B., M. Ferring (2021), *The Implementation of a Multilateral Price Index Method for Scanner Data in the Luxembourg CPI*. STATEC Working Paper N° 121.

Rao, D. S. P. (2005). On the Equivalence of Weighted Country-Product-Dummy (CPD) Method and the Rao-System for Multilateral Price Comparisons. *Review of Income and Wealth*, 51, 571–580.

Silver, M. and S. Heravi, S. (2007), The Difference Between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes. *Journal of Business & Economic Statistics*, 25, 239–246.

Silver M. (2010), The wrongs and rights of unit value indices. *Review of Income and Wealth*, 56, 206-223.

Triplett, J.E. (1986), The economic interpretation of hedonic methods. *Survey of Current Business*, 86, 36-40.

Triplett, J.E. (2006), *Handbook on hedonic indexes and quality adjustments in price indexes: special application to information technology products*, OECD, Paris.

Van Loon K. (2019), *Redefining what products are in the context of scanner data and web scraping, experiences from Belgium.* Paper presented at the 16th meeting of the Ottawa Group, 08-10 May 2019, Rio de Janiero, Brasil.

Van Loon, K. (2020*). Scanner data and web scraping in the Belgian CPI*. Presentation available here: https://www.nationalacademies.org/event/10-07-2020/docs/D9B1BADE81B7AD61F557A33D673AA79C7809506872D8

Appendix A. Proofs for the claims in Section 3

This appendix gives the proofs for the claims on the equivalence between imputation and product clustering.

Törnqvist

Lemma 1. Product clustering on a direct, bilateral Törnqvist index is equivalent to a fully imputed Törnqvist index, in which each mutual existing, new and disappearing item is imputed with a unit value price.

Proof:

Törnqvist, applied on product clusters *h*, can be written as

$$P_T^{0,t} = \prod_h \left(\frac{p_h^t}{p_h^0}\right)^{(s_h^0+s_h^t)/2}, \quad \text{(A.1)}$$

where p_h^t and p_h^0 are unit value prices for periods *t* and 0. Let U_h^{ot} denote the set of mutual exisisting, new and disappearing items in stratification cell *h*.

Expression (A.1) can be stated as

$$P_T^{0,t} = \prod_h \left(\frac{p_h^t}{p_h^0}\right)^{\sum_{[i\,\epsilon\, U_h^{ot}]}\,(s_i^0+s_i^t)/2}, \quad \text{(A.2)}$$

which is equivalent to

$$P_T^{0,t} = \prod_h \prod_{i\in U_h^{ot}} \left(\frac{p_h^t}{p_h^0}\right)^{(s_i^0+s_i^t)/2}. \quad \text{(A.3)}$$

Expression (A.3) is a Törnqvist index, fully imputed with unit value prices. □

Geary Khamis

Lemma 2. Product clustering in Geary Khamis is equivalent to:

- *imputing each price p_i^t by the unit value price $\hat{p}_i^t = p_h^t$*
- *imputing each quantity q_i^t by $\hat{q}_i^t = q_h^t\,(Q_i/Q_h)$*

where Q_i and Q_h are the total quantity for item i and cluster h aggregated over time.

Proof:

The Geary Khamis index can be obtained by iteratively solving the equations (2.3) and (2.4). We demonstrate that these iterations evolve similarly for product clustering and imputation. Let us denote v_i^{imp} the implicit price of item *i*, after imputation. Similarly, v_h^{clu} is the implicit price for product cluster *h.*

We first show that, given a price index $P_{GK}^{0,t}$, equation (2.4) implies that $v_i^{imp} = v_h^{clu}$ for all $i \in h$. In other words: imputation and product clustering give the same implicit prices.

$$v_i^{imp} = \sum_{z=0}^{T} \varphi_i^z \frac{\hat{p}_i^z}{P_{GK}^{0,z}} = \sum_{z=0}^{T} \left(\frac{\hat{q}_i^z}{\sum_z \hat{q}_i^z} \right) \frac{p_h^z}{P_{GK}^{0,z}} = \sum_{z=0}^{T} \left(\frac{q_h^z \, (Q_i/Q_h)}{\sum_z q_h^z \, (Q_i/Q_h)} \right) \frac{p_h^z}{P_{GK}^{0,z}}$$

$$= \sum_{z=0}^{T} \left(\frac{q_h^z}{\sum_z q_h^z} \right) \frac{p_h^z}{P_{GK}^{0,z}} = v_h^{clu} \qquad \text{(A.4)}$$

The second thing to show is that equation (2.3) gives rise to the same price indexes, i.e. $P_{imp}^{0,t}$= $P_{clu}^{0,t}$, if the implicit prices are the same, i.e. if $v_i^{imp} = v_h^{clu}$ for all $i \in h$.

Imputation gives the following price index

$$P_{imp}^{0,t} = \frac{(\sum_{i \in U^t} \hat{p}_i^t \hat{q}_i^t)/(\sum_{i \in U^0} \hat{p}_i^0 \hat{q}_i^0)}{(\sum_{i \in U^t} v_i^{imp} \hat{q}_i^t)/(\sum_{i \in U^0} v_i^p \hat{q}_i^0)}. \qquad \text{(A.5)}$$

Let's denote the set of product clusters at time *t* by U_h^t. Then, (A.5) can be re-expressed as

$$\frac{\left(\sum_{h \in U_h^t} \sum_{i \in h} p_h^t q_h^t \, (Q_i/Q_h)\right) / \left(\sum_{h \in U_h^0} \sum_{i \in h} p_h^0 q_h^0 \, (Q_i/Q_h)\right)}{\left(\sum_{h \in U_h^t} \sum_{i \in h} v_h^{clu} q_h^t \, (Q_i/Q_h)\right) / \left(\sum_{h \in U_h^0} \sum_{i \in h} v_h^{clu} q_h^0 \, (Q_i/Q_h)\right)} =$$

$$\frac{\left(\sum_{h \in U_h^t} p_h^t q_h^t\right) / \left(\sum_{h \in U_h^0} p_h^0 q_h^0\right)}{\left(\sum_{h \in U_h^t} v_h^{clu} q_h^0\right) / \left(\sum_{h \in U_h^0} v_h^{clu} q_h^0\right)} = P_{clus}^{0,t} \qquad \text{(A.6)}$$

Above it has been shown we get the same indexes for imputation and product clustering after applying (2.3) and (2.4). The same also holds true if we repeatedly apply these equations. This proofs the equivalence of imputation and product clustering. □

TPD

Lemma 3. Product clustering in TPD is equivalent to:

- *imputing each price* p_i^t *by the unit value price* $\hat{p}_i^t = p_h^t$
- *imputing each quantity* q_i^t *by* $\hat{q}_i^t = q_h^t \, (Q_i/Q_h)$

where Q_i *and* Q_h *are the total quantity for item i and cluster h, aggregated over time.*

Proof:

The TPD index can be obtained by iteratively solving the equations (2.7) and (2.8). We demonstrate that these iterations evolve similarly for product clustering and imputation. The proof is similar to the one for Geary Khamis (Lemma 2).

First, we show that, given a price index $P^{0,t}$, equations (2.8) implies that $v_i^{imp} = v_h^{clus}$ for all $i \in h$. In other words: imputation and product clustering give the same implicit prices.

Applying (2.8) to the imputed data gives

$$v_i^{imp} = \prod_{z=0}^{T} \left(\hat{p}_i^z / P_{TPD}^{0,z}\right)^{\widehat{w}_i^z} = \prod_{z=0}^{T} \left(p_h^z / P_{TPD}^{0,z}\right)^{w_h^z} = v_h^{clus}. \tag{A.7}$$

Hence, imputation and product clustering lead to the same implicit prices. In the last equality we use that $\widehat{w}_i^z = \widehat{w}_h^z$. This follows from

$$\begin{aligned} &\widehat{w}_i^z = \hat{p}_i^z \hat{q}_i^z / \sum_z \hat{p}_i^z \hat{q}_i^z = (p_h^z q_h^z\, Q_i/Q_h) / \sum_z (p_h^z q_h^z\, Q_i/Q_h) = \\ &(p_h^z q_h^z) / \sum_z (p_h^z q_h^z) = w_h^z \end{aligned} \tag{A.8}$$

The second thing to show is that if we have the same implicit prices, $v_i^{imp} = v_h^{clus}$ for all $i \in h$, then imputation and product clustering lead to the same price index $P^{0,t}$.

Applying (2.7) on the imputed data gives

$$\begin{aligned} P^{0,t} &= \frac{\prod_{i \in U(t)} \left(\hat{p}_i^t / v_i^{imp}\right)^{s_i^t}}{\prod_{i \in U(0)} \left(\hat{p}_i^0 / v_i^{imp}\right)^{s_i^0}} = \frac{\prod_{h \in U(t)} \prod_{i \in h} \left(p_h^t / v_h^{clus}\right)^{s_i^t}}{\prod_{h \in U(0)} \prod_{i \in h} \left(p_h^0 / v_h^{clus}\right)^{s_i^0}} \\ &= \frac{\prod_{h \in U(t)} \left(p_h^t / v_h^{clus}\right)^{s_h^t}}{\prod_{h \in U(0)} \left(p_h^0 / v_h^{clus}\right)^{s_h^{0,}}}. \end{aligned} \tag{A.9}$$

The last expression is the price index for product clustering. This shows that imputation and clustering result in the same price index. □

Appendix B. Product matching algorithm

The product matching algorithm in Section 4 can be more formally summarized as follows:

Consider a period *t,* with *t > 0*. Let $N_{h,t}$ and $D_{h,t}$ denote the set of new and disappearing items for stratum *h*. New items are the ones that have been sold in *t* for the first time. Disappearing items have been sold in *t*–1 for the last time. The sizes of these sets are $|N_{h,t}|$ and $|D_{h,t}|$.

- If $|N_{h,t}| = 0$ or $|D_{h,t}| = 0$, do not do any matching for stratum *h*.
- If $|N_{h,t}| > |D_{h,t}| > 0$:

- Sort the items in $D_{h,t}$ decreasingly on the average turnover share over the periods 1,…, *t*-1
- Compute $r_{h,t} = (|N_{h,t}| \bmod |D_{h,t}|)$, i.e. the remainder when $|N_{h,t}|$ is divided by $|D_{h,t}|$.
- Replicate the first $r_{h,t}$ items of $D_{h,t}$ $\lfloor |N_{h,t}| / |D_{h,t}| \rfloor + 1$ times. Replicate all other items $\lfloor |N_{h,t}| / |D_{h,t}| \rfloor$ times. The prices of all replications are set equal to the prices of the original items. The quantities of all items are divided by the number of replications of that item. After this step the numbers of new and disappearing items are equal.
- Compute the average turnover share for all new and disappearing items over all periods in which these have been sold. Sort the new and the disappearing items decreasingly on average turnover share. Match the new and disappeared items one on one, based on the order in the sorted lists.

- If $|D_{h,t}| \geq |N_{h,t}| > 0$. The other way around:

- Sort the items in $N_{h,t}$ decreasingly on the average turnover share over the periods *t*,…,*T*
- Compute $r_{h,t} = (|D_{h,t}| \bmod |N_{h,t}|)$, i.e. the remainder when $|D_{h,t}|$ is divided by $|N_{h,t}|$.
- Replicate the first $r_{h,t}$ items of $N_{h,t}$ $\lfloor |D_{h,t}| / |N_{h,t}| \rfloor + 1$ times. Replicate all other items $\lfloor |D_{h,t}| / |N_{h,t}| \rfloor$ times. The prices of all replications are set equal to the prices of the original items. The quantities of all items are divided by the number of replications of that item. After this step the numbers of new and disappearing items are equal.
- Compute the average turnover share for all new and disappearing products. This average share is computed over all periods in which an item has been sold. Sort the new and the disappearing items decreasingly on average turnover share. Match the new and disappeared items one on one, based on the order in the sorted lists.