Evaluating unit-value price indices in a dynamic item universe

Li-Chun Zhang[1,2,3], Ingvild Johansen[2], and Ragnhild Nygaard[2]

[1]*University of Southampton (email: L.Zhang@soton.ac.uk)*

[2]*Statistics Norway*

[3]*University of Oslo*

1 Introduction

Dynamic item universe poses a fundamental challenge to index numbers, which is made apparent given scanner data that cover an entire sub-universe of all consumption *items* by (item code, outlet). The early research was focused on supermarket data which consists largely of stable items. The attention has since gradually shifted towards the parts of consumption market that are characterised by high item churns, where the methodology initially introduced for supermarket data is no longer adequate. It is generally agreed that the available quantity and expenditure data should be incorporated, below any elementary aggregate (EA) that corresponds to the most detailed level of expenditure weights (Chessa, 2016; Dalén, 2017). Formulae that so far have received most attention include the Gini, Eltetö and Köves, Szulc (GEKS) index (Ivancic et al., 2011), the time product dummy (TPD) index (Aizcorbe et al., 2003; Krsinich, 2016) and the Geary-Khamis (GK)/Generalized unit value (GUV) index (Chessa, 2016; Von Auer , 2014). At the same time, one cannot but notice a lack of standard regarding the process, by which the particular index can be plausibly established in a given situation. To improve the situation, we outline and study in this paper two components to a systematic approach: a *Total Effect Framework (TEF)* and a set of *generic diagnostics*.

The TEF is defined along two dimensions. First, compiling the Consumer Price Index (CPI) in practice involves three necessary choices, which require broader considerations that are not dictated by the choice of ***index formula***. The three choices are (I) ***index base***, which can be moving every month or updated yearly, (II) ***reference universe***, depending on whether data other than the two comparison months are used, (III) ***homogeneous products (HPs)***, instead of applying the index formulae directly to the observed items. Thus, traditionally, the CPI can be described as a chained index where, over each 12 months cycle, a *bilateral index* is calculated with a *yearly fixed base* month and based on a set of *representative goods and services*. However, as it will be explained below, one needs to revisit these choices in order to fully capture the dynamic item universe of the relevant scanner data.

Second, the problems that complicate the above choices are caused by the dynamism of the item universe. A ***matched item*** between any two time points is an item that have the same (item code, outlet) on both occasions; an unmatched item is a ***dynamic item*** with respect to these two time points. We shall distinguish three types of dynamic items: a) ***replacement items***, such as relaunches or updates, b) ***regeneration items***, which are completely new or outgoing items, and c) ***(strongly) seasonal items*** that are present in the same month or months each year but not the other months.

The TEF is defined with respect to the choices (I) - (III) *and* the dynamic items (a) - (c), due to their inter-dependence: different dynamic items can have different effects on a given choice, and their effects on a given choice can vary with the other choices.

Chessa et al. (2017) list five "choice aspects" of an index. The first two aspects "product weighting" and "index formula" belong to what we refer to as the choice of index formula, which is a separate though dependent decision of the three necessary choices with broader implications above. As will be clarified in Section 2, the next two aspects "updating" and "window length" overlap largely the choices of index base and reference universe. Chessa et al. (2017) clearly favour multilateral index. The discussions in the sequel will show that several considerations are required for this choice. The last aspect "level of product differentiation" coincides with the choice of HP. We use the more cumbersome but unambiguous term "homogeneous product" (e.g. Chessa, 2016) instead of just "product", because the latter often designates simply "item" in the traditional literature. The proposed TEF provides a more structured coverage of the choice aspects, in that it allows one to differentiate the varying effects on these choices due to the different types of dynamic items over time.

Under the TEF, the different alternatives of a choice can be studied analytically wherever possible. Yet the complexity involved is often such that clear-cut conclusions cannot be reached a priori independent of the actual data. In the empirical studies reported in the literature, it is most common to find figures that compare alternative indices which are envisaged as possible production methods. However, it can be difficult to interpret the results, either because the indices involve different combinations of necessary choices in addition to the index formulae, or because some of the indices are not based on the 'best' choices for them. There is thus a need to develop more generic diagnostics as tools for empirical investigation, which can be applied across the commodity groups.

A generic diagnostic aims to isolate the choice aspect of concern, away from the other choices one would need to make in a production method, so that the likelihood of reaching a *partial* conclusion regarding that particular choice aspect is increased. To this end many indices employed in such diagnostics need not to be genuine candidates for production, but they are designed and introduced to generate useful empirical evidences for the particular choice aspect of concern. We develop a set of generic diagnostics (Table 11 in Section 7), pertaining to the effects of missing replacement, the formation and classification of HPs, a particular choice of index base or reference universe, etc. Connections to the relevant techniques in the recent literature will be reviewed and commented, although they are unlikely to be exhaustive due to limitations of our knowledge and effort in this respect. The generic diagnostics will be illustrated using scanner datasets mainly from the market of sport clothing and equipment which have high item churns. This work is part of an Eurostat grant agreement for the period 2018-2020. We hope that over time the set of generic diagnostics will be expanded and refined, based on the joint efforts of the whole research community, such that they can form a standard toolset of price index methodology in practice.

The paper is organized as follows. Section 2 to 5 provide a description of the TEF while Section 6 provides som results on index formula while Section 7 summarizes and makes some concluding remarks.

2 Total effect framework

In this Section we describe and clarify how we arrive at the elements along the two dimensions of TEF in more details. The effects of the dynamic items on the different choices, as well as the inter-dependences between the choices will be explored in the later Sections.

2.1 Three necessary choices

There are three necessary choices below the EA-level in a dynamic item universe.

(I) In practice any price index must be chained over time, such that it can be generically given as

$$P^t = P^b P^{b,t}$$

where P^b is the index for month b, which is not affected by the *short-term* $P^{b,t}$ from b to the month t, for as long as b is unchanged. For instance, in the current practice of CPI, b is December of a given year and t cycles through the next 12 months, before b is updated at the end of the following year and so on. The choice of b over time is referred to as the choice of index base. Three most obvious possibilities are summarised in Table 1.

(I.1) The base month b is updated once a year, denoted generically as month 0 in Table 1.

(I.2) The base month b is updated every month, denoted as month $t - 1$ in Table 1. Only one short-term index $P^{t-1,t}$ is calculated, before the base b is updated.

(I.3) In this case the base consists of all the 12 months of the previous year, denoted as 0,...,11 in Table 1. The whole 12-months fixed base is updated once a year. The short-term index $P^{0,t}$ between any two months 0 and t in the current year is an indirect index, calculated as the ratio between the chained long-term indices P^t and P^0.

Table 1: Index with moving base, yearly fixed base, or fixed 12-month base

Base	Update	Long-term P^t	Short-term $P^{0,t}$
(I.1) Fixed month 0	Yearly	$P^0 P^{0,t}$	$P^{0,t}$
(I.2) Moving month $t - 1$	Monthly	$P^{t-1}\ P^{t-1,t}$	$P^{0,1} \dots P^{t-1,t}$
(I.3) Fixed 12-months (0, ..., 11)	Yearly	$P^{t-12}\ P^{t-12,\mathrm{t}}$	$\frac{p^t}{p^0} = \frac{p^{t-12} p^{t-12,t}}{p^{0-12}\ p^{0-12\ ,0}}$

The distinctions (I.1) - (I.3) arise due to the dynamic item universe. Otherwise, provided a fixed item universe throughout time, the choices would no longer matter as long as one uses a transitive index.

(II) Given the index base b, we refer to the pair of item universes of b and t as the ***comparison universe***, for which an index $P^{b,t}$ is needed to yield P^t. The index $P^{b,t}$ can be calculated either using only the data from b and t, in which case it is a ***bilateral*** index, or it can make use of additional data from other time points, in which case it is a ***multilateral*** index. This requires a choice of the ***reference universe***, denoted by $R(b,t)$, consisting of the item universes of all the data used to compute $P^{b,t}$. For example, given index base (I.1) b = 0, $P^{0,t}$ is calculated as a bilateral direct index in the traditional CPI. But it is also possible to calculate $P^{0,t}$ as a multilateral index given (I.1). Reference universe is therefore a separate choice to index base.

(III) Formation and classification of HP is implicit in the traditional approach of CPI based on representative items. Each representative item can be regarded as an HP, and the unselected items are implicitly grouped with one or another of them. This enables one to treat the item universe as static over a given period of time. Given the scanner data that has full coverage of a CPI sub-universe, defining HPs among all the available items can be motivated theoretically, in order to accommodate price-related substitution involving the replacement items and to reduce the potential bias due to missing replacement. It is also desirable practically, in order to alleviate the resource that otherwise will be needed to identify item-by-item replacements. It follows that intermediate HP-aggregation of items should be introduced below the EAs. Indeed, it is attractive if the changing item universes can be 'matched' via the HPs over time, although this may not always be possible in practice. Finally, HP formation and classification is another separate choice, which is not dictated by the choice of a particular index formula.

2.2 Dynamic items

Given a comparison universe, there are three kinds of dynamic items in addition to the matched items, which are the seasonal, replacement and regeneration items.

Seasonal items It is standard to distinguish between the *strongly* and *weakly* seasonal items (IWGPS, 2004, Ch. 22). A strongly seasonal item is not available throughout a year; a weakly seasonal item may be available throughout a year, but has regular seasonal fluctuations in price or quantity. In the context of traditional CPI, IWGPS (2004) concludes that there "is, as yet, no consensus on what is the best practice in this area". There are European regulations on the treatment of seasonal items through the Harmonized index of Consumer Prices (HICP), but the regulations were not drawn up with scanner data in mind.

- ***Weakly seasonal items*** These are matched items with relatively large fluctuations of price and/or quantity. For instance, the fluctuations can be due to sales, which may or may not occur on a regular basis. As long as the sales data should be used as-is, they do not call for a special treatment, provided appropriate item identification, HP classification and index formula. A sales item that has a changing item code is classified as a replacement item instead. Another example of weakly seasonal item is one that is little transacted, so that it enters into the index only from time to time, when there exists a threshold value of transaction or quantity in practice. Such an item can be handled by an index formula that appropriately incorporates the quantity data.

- ***Strongly seasonal items*** In conception a *regular* seasonal item is present in the two months that are exactly one year apart from each other, but not necessarily otherwise. Examples are many food, clothing items, heating oil, etc. *Irregular* items arise for an individual consumer on an ad hoc even one-off basis, such as wedding expenditure, gifts, etc. Insofar as the consumers are considered in groups, there is no reason to treat the irregular strongly seasonal items differently than the matched items.

From now on we shall simply refer to regular strongly seasonal items as seasonal items. Seasonal items affect all the three necessary index choices. When it comes to index base, year-on-year monthly indices with base (I.3) are the most appropriate treatment. However, presumably one would then need to deal with an increased amount of replacement and regeneration items, and the short-time indices would need to be calculated indirectly as shown in Table 1. So the choice remains unsettled a priori. Since a multilateral index needs to deal with an increased amount of seasonal items than a bilateral index, the seasonal items affect also the choice of reference universe. Finally, the creation of HPs for matched items throughout a year requires clearly different considerations than for the seasonal items.

Replacement items These do not pose extra issues to an acceptable index formula, provided they can be appropriately identified, referred to as the *replaced* and *replacing* items, respectively. The creation of HP would transform the problem of missing replacement to potential HP misclassification. One may distinguish two types of replacement items.

- ***Relaunch*** (e.g. Chessa, 2013, 2016). These are the same (or essentially the same) items in terms of utility, but with different codes after repackaging. A relaunch replacement item is usually associated with an increasing price, since it defies commercial common sense that one would repackage an item before down-pricing it. A potential complication arises when an item code is changed when it is put on sale, without being repackaged physically. It is then formally a relaunch item but with a different price movement than a usual relaunch item that is associated with an increasing price.

- ***Update*** These are different items subjected to the substitution effect, despite they may have somewhat different utilities.

Insofar as the pair of replacing and replaced items are comparable in their characteristics, the hedonic method seems suitable in theory. In practice, however, the approach may be infeasible due to lack of metadata. One can aim at correct HP-matching, where the pair of items are identified as the same HP. The potential problems caused by these items are therefore HP misclassification or missing replacement without creating HPs. Given correct HP-matching of the replacement items, they do not pose further difficulties to the choice of index base or reference universe, since the sub-universes of all the replacement (and matched) items are then matched via the HPs and static over time.

Regeneration items These are ***new*** or ***out-going*** items, which are not replacement items from a substitution point of view. An example is microwave ovens first time they were introduced in the market. It is unclear what is the best theory and practice regarding the regeneration items. The hedonic method may not be able to correctly evaluate the prices of the new characteristics. It is possible to create new EAs and the associated expenditure weights in response to the new items. However, frequent EA restructuring or reweighting can be difficult in practice, nor has its theoretical basis been substantiated. The creation of HPs offers a practical treatment, despite it may be impossible to assign the correct HP of a new item if it is not subjected to substitution. For example, one might include the newly introduced iPhone X in the HP called 'top-end iPhones', so that it can be brought into the index immediately, although the resulting HP unit-value price might not appropriately capture the substitution effects with the existing items in that HP. Insofar as correct HP classification of regeneration items may be impossible in reality, they could also affect the choice of index base and reference universe to a greater or smaller extent.

3 HP formation and classification

In the analytic exposition and empirical illustrations below we use the Lehr or modified GK (MGK) index (Lamboray, 2017; Zhang et al., 2017), which is given by

$$P^{0,t} = V^{0,t} / Q^{0,t} \text{ where } Q^{0,t} = \frac{\sum_{i \in U_t} p_i q_i^t}{\sum_{i \in U_0} p_i q_i^o} \text{ and } p_i = \frac{\sum_{r \in R_i} p_i^r q_i^r}{\sum_{r \in R_i} q_i^r}$$

where $V^{0,t}$ is the expenditure ratio between the two months, and $Q^{0,t}$ can be considered as a reference-price (or intrinsic-worth) quantity index, with summation either over the relevant units – items or HPs, and R_i contains all the time points at which the unit i is present in the data. There are many other ways of setting p_i, all of which can be referred to as the GUV index (Dalén, 2001; De Haan, 2002; Von Auer, 2014). The use of this index in this paper is because it serves our present purposes more conveniently than the GK, TPD or GEKS index; it is not our intension to promote it as a generally superior index formula.

As explained before, the creation of HP is theoretically motivated for the replacement items but not necessarily the seasonal and regeneration items. Ideally, to motivate the HPs in a given situation, one would like to focus on the potential effect of HP formation and misclassification against the effects of missing unmatched replacement items without the HPs. However, in reality one may be unable to identify and exclude the seasonal and regeneration items in the data. Below we shall first explore the effects of seasonal and regeneration items analytically. Afterwards, we shall consider the effects and propose diagnostics of missing replacement items, HP misclassification and HP formation, respectively.

3.1 HP for seasonal and regeneration items

Denote by U a matched item universe. Let k be a new (or seasonal) item in month t, which is not a replacement of any item in U_0. Without HP classification, let $P^{0,t}_{(k)}$ and $P^{0,t}$ and be the GUV index calculated without and with item k, respectively, i.e.

$$P^{0,t}_{(k)} = \frac{V^{0,t}_{(k)}}{Q^{0,t}_{(k)}} \qquad V^{0,t}_{(k)} = \frac{\sum_{i\in U} p_i^t q_i^t}{\sum_{i\in U} p_i^0 q_i^0} = \frac{A}{B} \qquad Q^{0,t}_{(k)} = \frac{\sum_{i\in U} p_i q_i^t}{\sum_{i\in U} p_i q_i^0} = \frac{S}{T}$$

$$P^{0,t} = \frac{V^{0,t}}{Q^{0,t}} \qquad V^{0,t} = \frac{\sum_{i\in U} p_i^t q_i^t + p_k^t q_k^t}{\sum_{i\in U} p_i^0 q_i^0} = \frac{A+d}{B} \qquad Q^{0,t} = \frac{\sum_{i\in U} p_i q_i^t + p_k^t q_k^t}{\sum_{i\in U} p_i q_i^0} = \frac{S+d}{T}$$

$$\Rightarrow \quad \begin{cases} P^{0,t}_{(k)} \geq P^{0,t} & \Leftrightarrow \quad \frac{A}{S} \geq \frac{A+d}{S+d} \quad \Leftrightarrow \quad A \geq S \\ P^{0,t}_{(k)} \leq P^{0,t} & \Leftrightarrow \quad \frac{A}{S} \leq \frac{A+d}{S+d} \quad \Leftrightarrow \quad A \leq S \end{cases}$$

It can be seen that if there is a general upward trend of the prices, i.e. $A \geq S$, incorporating item k will move the index downwards; whereas the opposite is the case, provided a general downward trend, i.e. $A \leq S$. Thus, without HPs, including regeneration (and seasonal) items into the index may cause a systematic *regeneration (or seasonal) effect*.

Next, suppose HP classification pertains to all the items, denoted by C = $\{c_1,...,c_M\}$. Suppose the item k above is classified as $k \in c_i$, i.e. the ith HP. We have

$$P^{0,t}_{(k)} = \frac{V^{0,t}_{(k)}}{Q^{0,t}_{(k)}} \qquad V^{0,t}_{(k)} = \frac{\sum_{j=1}^M p_j^t q_j^t}{\sum_{j=1}^M p_j^0 q_j^0} \qquad Q^{0,t}_{(k)} = \frac{\sum_{j\neq i} p_j q_j^t + p_i q_i^t}{\sum_{j\neq i} p_j q_j^0 + p_i q_i^0}$$

$$P^{0,t} = \frac{V^{0,t}}{Q^{0,t}} \qquad V^{0,t} = \frac{\sum_{j=1}^M p_j^t q_j^t + p_k^t q_k^t}{\sum_{j=1}^M p_j^0 q_j^0} \qquad Q^{0,t} = \frac{(\sum_{j\neq i} p_j q_j^t + p_i' q_i^t) + p_i' q_k^t}{\sum_{j\neq i} p_j q_j^0 + p_i' q_i^0}$$

where p_i' is calculated for c_i with the item k and p_i that without k. It is now possible for $P^{0,t}$ to be either above or below $P^{0,t}_{(k)}$. In other words, classifying the regeneration and seasonal items to the HPs can perturb the direction of any systematic effects otherwise, despite this treatment of these two types of dynamic items may not be ideal theoretically speaking.

3.2 Effects of missing replacement

Missing replacement is practically unavoidable without the HPs. Let $U_{0\cup t} = U_0 \cup U_t$ contain all the items in the comparison universe at months 0 and t, and $U_{0t} = U_0 \cap U_t$ only the matched items. Let an unmatched item j, for $j \in U_t \setminus U_{0t}$, be the replacement of an item i, for $i \in U_0 \setminus U_{0t}$. Let $\tilde{p}_i = \tilde{p}_j$ be their reference price, provided they are identified as such. Otherwise, in the case of missing replacement, one may either compute a matched-item index without $\{i,j\}$, or a distinct-item index with $\{i,j\}$ where the two are treated as distinct unmatched items.

Matched-item index The GUV index for the matched items in U_{0t} is based on

$$V_M^{0,t} = \frac{\sum_{k\in U_{0t}} p_k^t q_k^t}{\sum_{k\in U_{0t}} p_k^0 q_k^0} \equiv \frac{A}{B} \qquad \text{and} \qquad Q_M^{0,t} = \frac{\sum_{k\in U_{0t}} p_k q_k^t}{\sum_{k\in U_{0t}} p_k q_k^0} \equiv \frac{C}{D}$$

where p_k is the reference price of $k \in U_{0t}$. The index for $U_{0t} \cup \{i,j\}$ is based on

$$V^{0,t} = \frac{A + p_j^t q_j^t}{B + p_i^0 q_i^0} \qquad \text{and} \qquad Q^{0,t} = \frac{C + \tilde{p}_j q_j^t}{D + \tilde{p}_i q_i^0}$$

As discussed earlier, unmatched replacement items (i,j) may arise from relaunch or update. In either case, one tends to have $p_j^t > p_i^0$. Provided the price increase associated with relaunch or update is higher than that among the matched items. i.e.

$$p_i^0 < p_j^t \qquad \text{and} \quad \frac{p_j^t}{\tilde{p}_j} > \frac{A}{C} \quad \text{and} \quad \frac{\tilde{p}_i}{p_i^0} > \frac{D}{B}$$

we have

$$\Big(\frac{V^{0,t}}{Q^{0,t}}\Big) / \Big(\frac{V_M^{0,t}}{Q_M^{0,t}}\Big) = \Big(\big(\frac{A + p_j^t q_j^t}{C + \tilde{p}_j q_j^t}\big) / \big(\frac{A}{C}\big)\Big)\Big(\big(\frac{D + \tilde{p}_i q_i^0}{B + p_i^0 q_i^0}\big) / \big(\frac{D}{B}\big)\Big) > 1$$

Sine the argument can be extended to multiple missing replacements $(i_1,j_1),\dots,(i_K,j_K)$, missing replacement tends to induce a negative bias of the matched-item index $V_M^{0,t} / Q_M^{0,t}$.

Distinct-item index Let p_i and p_j be the respective reference prices, when the two replacement items are treated as distinct items. Let $P_A^{0,t} = V^{0,t}/Q_A^{0,t}$ be the resulting index, since $V^{0,t}$ given above is unaffected whether the two are match or not, where

$$Q_A^{0,t} = \frac{\sum_{g\in U_t; g\neq j} p_g q_g^t + p_j q_j^t}{\sum_{k\in U_0; k\neq i} p_k q_k^t + p_i q_i^0} = \frac{C + p_j q_j^t}{D + p_i q_i^0}$$

Let $P^{0,t} = V^{0,t}/Q^{0,t}$ be the index provided the two are correctly match, as given above. Without losing generality, suppose

$$\min(p_i, p_j) \leq \tilde{p}_i = \tilde{p}_j \leq \max(p_i, p_j)$$

such that

$$\Big(\frac{V^{0,t}}{Q^{0,t}}\Big)/\Big(\frac{V^{0,t}}{Q_A^{0,t}}\Big) = \Big(\frac{C + p_j q_j^t}{C + \tilde{p}_j q_j^t}\Big)\Big(\frac{D + \tilde{p}_i q_i^0}{D + p_i q_i^0}\Big) \quad \Rightarrow \quad \begin{cases} P^{0,t} \geq P_A^{0,t} & \text{if} \quad p_i \leq p_j \\ P^{0,t} \leq P_A^{0,t} & \text{if} \quad p_i \geq p_j \end{cases}$$

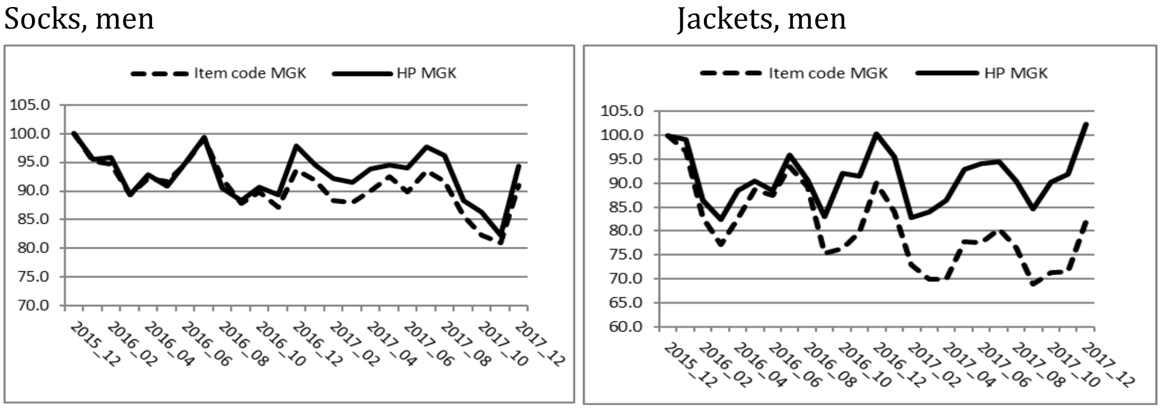
Since the argument can be extended to multiple missing replacements, missing replacement leads to a downward bias as the prices increase, and an upward bias as the prices decrease.

Missing replacement diagnostics Table 2 summarises the set-ups of this diagnostic. Under set-up (A), the matched-item index $P_M^{0,t}$ and the distinct-item $P_A^{0,t}$ provide two diagnostics of the missing replacement effects. In situations where the replacement items are associated with an overall upwards price trend, both of the indices would have a negative bias, yielding a kind of lower bound of any index that better handles the replacement items. Ideally one would like to exclude from $P_A^{0,t}$ the seasonal and regeneration items in $U_t \backslash U_{0t}$ and $U_0 \backslash U_{0t}$. However, to the extent the seasonal items and regeneration items are associated with a relatively higher price, including them in $P_A^{0,t}$ may move it downwards in the same direction as the missing replacements. The two indices $P_M^{0,t}$ and $P_A^{0,t}$ should be compared to the benchmark index $P^{0,t}$, which however requires extra cost associated with explicit item-matching of replacement items. The other choices involving index base and reference universe should be kept the same to isolate the effects of missing replacement. This implies to use only any index that pertains to the items in $U_0 \cup U_t$, such as a bilateral index or a multilateral GK/MGK index. Thus, under a more practical set-up (B), one may use as the benchmark an index based on suitable HPs pertaining to all the items, denoted by $P_{A,HP}^{0,t}$. To limit the potential confounding HP misclassification effects, the HPs should not be too coarse in this diagnostic. One may compare $P_{A,HP}^{0,t}$ to an index of the matched-item universe, which can be $P_M^{0,t}$ or $P_{M,HP}^{0,t}$ based on the same HPs of matched items.

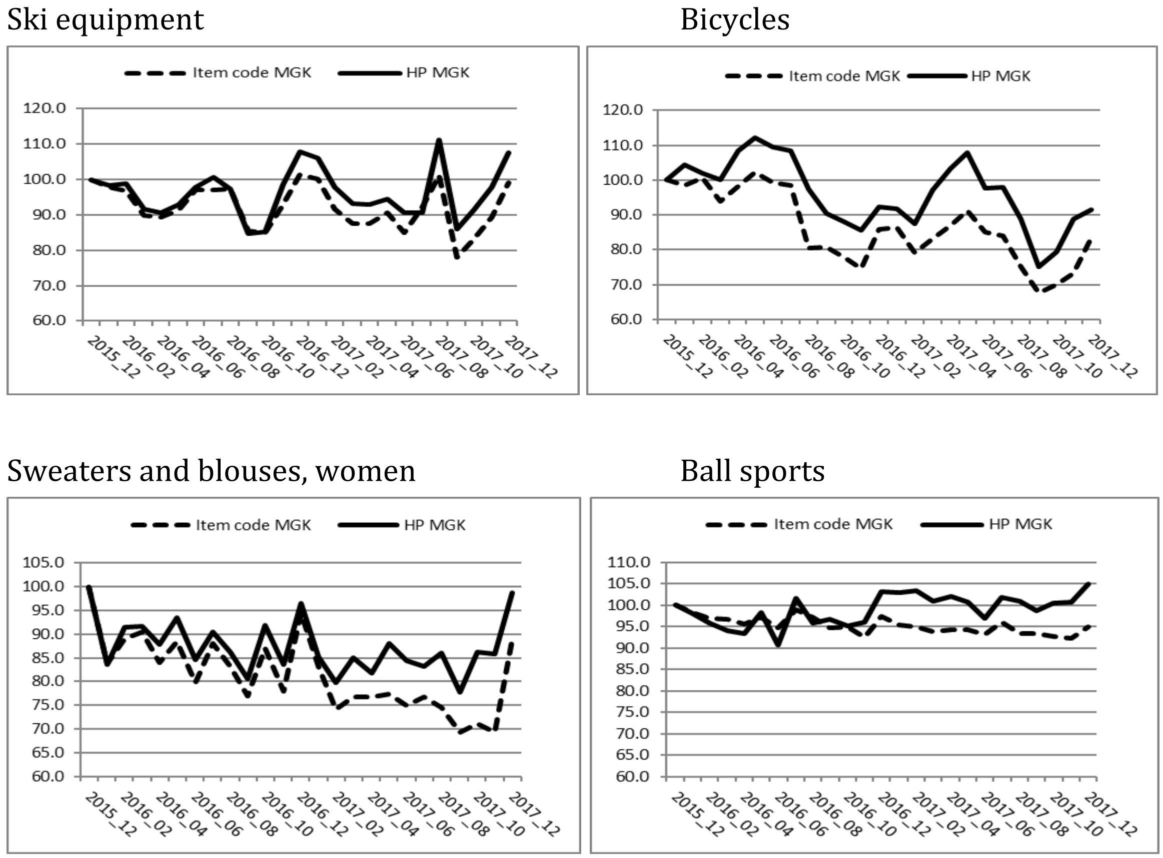
Table 2: Set-up of generic diagnostic for missing replacement effects

Set-up	Base	$R(0,t)$	HP	Formula = GUV	Comment
A	(I.1)	$U_0 \cap U_t$	No	$P_M^{0,t}$	Matched-item
	(I.1)	$U_0 \cup U_t$	No	$P_A^{0,t}$	Distinct-item
	(I.1)	$U_0 \cup U_t$	No	$P^{0,t}$	Explicit item-matching needed
B	(I.1)	$U_0 \cap U_t$	No/Yes	$P_M^{0,t}$ *or* $P_{M,HP}^{0,t}$	HP not too coarse
	(I.1)	$U_0 \cup U_t$	Yes	$P_{A,HP}^{0,t}$	More practical than $P^{0,t}$

Application Clothing and consumer electronics are examples of high-churn items, where it is necessary to study the effects of missing replacements. The application in this case uses the set-up (B) in Table 2, $P_M^{0,t}$vs. $P_{A,HP}^{0,t}$, based on data from one major sport equipment chain in Norway in 2016 and 2017. The definition of the HPs will be explored in more details in Section 3.3 and 3.4. As expected, the diagnostic (Figure 1) reveals a systematic downward bias in $P_M^{0,t}$, despite the effects may be small for certain commodity groups like men's socks. It is seen that the HP-based index is not more volatile that the item-based index.


Socks, men

Jackets, men


Ski equipment

Bicycles

Sweaters and blouses, women

Ball sports

Figure 1: Missing replacement diagnostic, setup (B)

3.3 HP classification

Let C = {c_1,...,c_M} be the HPs in the comparison universe (U_0,U_t). Each HP may consist of more than one item. Suppose every item in $U_{0\cup t}$, matched or not, must belong to one or another HP, denoted by $i \in c_k$ for some $1 \leq k \leq M$. HP misclassification effects may be studied conditional on the classification of items in U_0, in which case HP misclassification is the case if an item $j \in U_t \setminus U_{0t}$ is classified as $j \in c_l$ when it should be $j \in c_k$, *given* how the items in U_0 are classified. Whether the items in U_0 are correctly classified will be considered as a question of HP formation in Section 3.4. While such a decomposition is not free from potential problems, it does provide a practical means to disentangle the dual problem of HP formation and classification.

Dalén (2017) lists product, seller/geography and time as the relevant dimensions for achieving homogeneity. At this stage there are no completely generic algorithms for HP classification based on available metadata. However, since unit-value prices are needed for each HP, a minimum requirement is that the items within an HP must either have the same quantity unit or the same *kind* of quantity units for which a unit-value price across these items makes sense. Next, in addition to outlet/retailer, ***brand blocking*** seems a useful rule in practice, by which an HP is limited to items of the same brand. For example, an iPhone and a Samsung phone would then never be classified as the same HP. Brand blocking has been incorporated in many empirical studies (e.g. Chessa, 2016). Conceptually speaking, brand blocking may lead to fragmentation of an HP, when substitution actually takes place across the items of different brands. Even then, however, in theory it is still possible for an appropriate index formula to capture the substitution effects between the 'fragmented' HPs, by appropriate use of the quantity data.

It may happen that the available metadata are not rich enough to allow one to arrive at sufficiently HPs. In such situations it seems worth considering to use price directly as an additional criterion for HP classification, at least conditional on brand blocking. Brand blocking increases the plausibility of using price as a proxy quality measure, since it is unintuitive for a producer to price its own products contrarily. However, since using price for HP classification is not without potential drawbacks, there is a need for more careful analysis and suitable diagnostics.

Nearest price cluster Let p_k be the reference price of the items in c_k, where the HPs are arranged such that $p_k \geq p_g$ if $k \geq g$ where c_k and c_g are the kth and gth HP, respectively. By the method of *nearest price cluster (NPC)*, an unmatched item $j \in U_t \setminus U_{0t}$ is classified to c_k, provided

$$|p_k - p^t_j| < |p_l - p^t_j| \text{ for any } l \neq k \text{ and } 1 \leq l \leq M$$

Misclassification is the case if one should have classified $j \in c_g$, where $g \neq k$. Let p'_k be the updated reference price of c_k after including item j, and p'_g be that of c_g had j been included in c_g. We have then two likely situations

$$\begin{cases} p_g > p'_g > p'_k > p_k & if\ k < g \\ p_g < p'_g < p'_k < p_k & if\ k > g \end{cases}$$

Let $Q^{0,t}_g$ be the reference-price quantity index with correct classification $j \in c_g$, and $Q^{0,t}_k$ that with misclassification $j \in c_k$. We have

$$\frac{Q_g^{0,t}}{Q_k^{0,t}} = \Big(\frac{\sum\limits_{l\neq g,k} p_l q_l^t + p_g' q_g^t + p_g' q_j^t + p_k q_k^t}{\sum\limits_{l\neq g,k} p_l q_l^t + p_g q_g^t + p_k' q_j^t + p_k' q_k^t}\Big)\Big(\frac{\sum\limits_{l\neq g,k} p_l q_l^0 + p_g q_g^0 + p_k' q_k^0}{\sum\limits_{l\neq g,k} p_l q_l^0 + p_g' q_g^0 + p_k q_k^0}\Big)$$

where $q_j^0 = 0$ since $j \notin U_0$. In the special case of $q_i^t = q_i^0$ for any other item $i \neq j$, the ratio is less than 1, so that misclassification would cause a positive bias of the resulting GUV index. Similarly in the opposite case of $g > k$, which would cause a negative bias of the resulting GUV index. However, the result is inclusive generally, so it seems that misclassification may cause bias in either direction, as suggested by Dalén (2017).

HP misclassification sensitivity diagnostic We propose a generic diagnostic for the sensitivity due to HP misclassification. The idea is simple. If NPC classification causes bias in either direction, one can possibly induce a bias in one direction if one systematically classify an unmatched item in one direction of price. By the method of *lower nearest price cluster (LNPC)*, an unmatched item $j \in U_t \setminus U_{0t}$ is classified to c_k, provided

$$0 < p_j^t - p_k < p_j^t - p_l \; for\ any\ 1 \leq l < k.$$

Similarly one can possibly induce a bias in the opposite direction if one systematically classify an unmatched item in the opposite direction. By the method of *upper nearest price cluster (UNPC)*, an unmatched item $j \in U_t \setminus U_{0t}$ is classified to c_k, provided

$$0 < p_k - p_j^t < p_l - p_j^t \; for\ any\ k < l \leq M.$$

Finally, for the risk of disregarding price in HP classification, consider the method of *random price cluster (RPC)*, where an unmatched item $j \in U_t \setminus U_{0t}$ is classified to c_k, for k that is randomly selected from $\{1,...,M\}$.

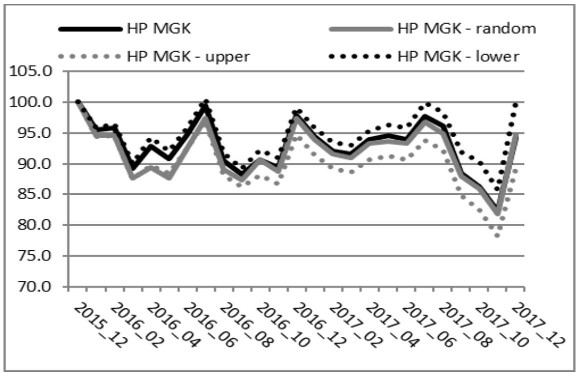
Table 3: Set-up of generic diagnostic for HP misclassification sensitivity

Base	$R(0,t)$	HP	Formula	Comment
(I.1)	$U_0 \cup U_t$	NPC	GUV	Blocking by outlet/retail and brand
(I.1)	$U_0 \cup U_t$	LNPC	GUV	Bound of bias in one direction
(I.1)	$U_0 \cup U_t$	UNPC	GUV	Bound of bias in opposite direction
(I.1)	$U_0 \cup U_t$	RPC	GUV	Risk indicator for disregarding price

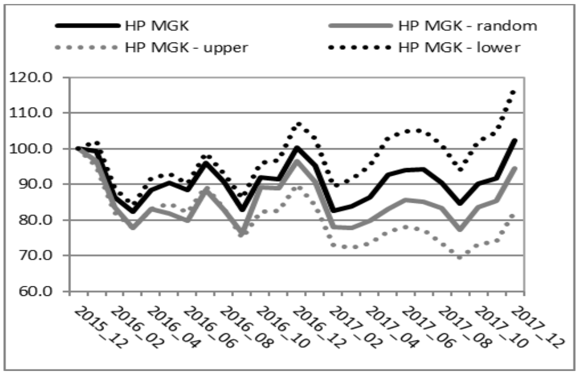
Table 3 summarises the set-up for this generic diagnostic, where the index base and formula are subject to one's choice. Provided blocking by outlet/retailer and brand in addition to other relevant metadata, one may expect HP misclassification by LNPC and UNPC to indicate the practical bounds of the bias due to misclassification by NPC. The NPC classification would seem sensitive if the resulting index is close to one of the bounds. In addition, the RPC classification provides an indication of the potential risk of bias due to misclassification if one disregards the price data altogether.

Application As discussed in Section 3.4, use of price for HP classification is needed in the sport clothing and equipment data of this study, since the HPs remain too heterogenous only based on the available metadata. We apply the generic diagnostic above to explore the sensitivity of NPC in addition to classification by available metadata, which includes the chains' own classification and blocking by outlet/retailer and brand. The results for six commodity groups are shown in Figure 2. As intended the methods of LNPC and UNPC provide, respectively, the upper and lower bounds of misclassification bias. The index using the method of NPC is kept at a distance to either bound, with a stable trend overtime and without exhibiting additional volatility. None of these observations suggests a high sensitivity of the NPC method in this case.

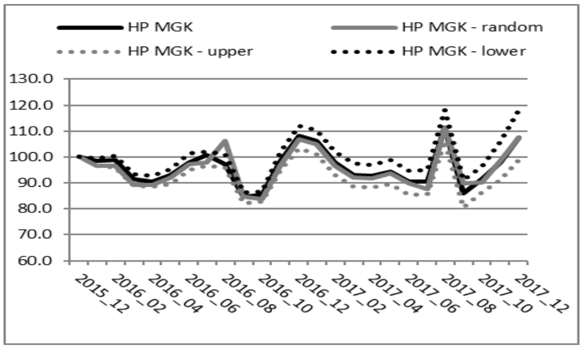
Socks, men



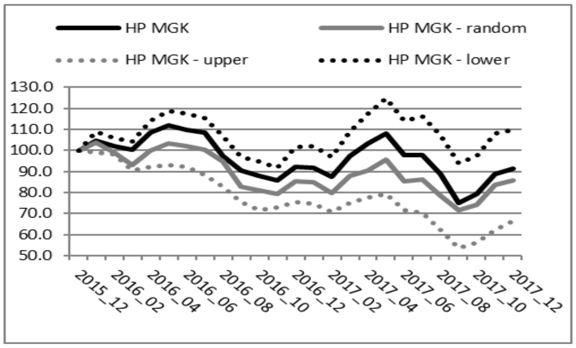
Jackets, men

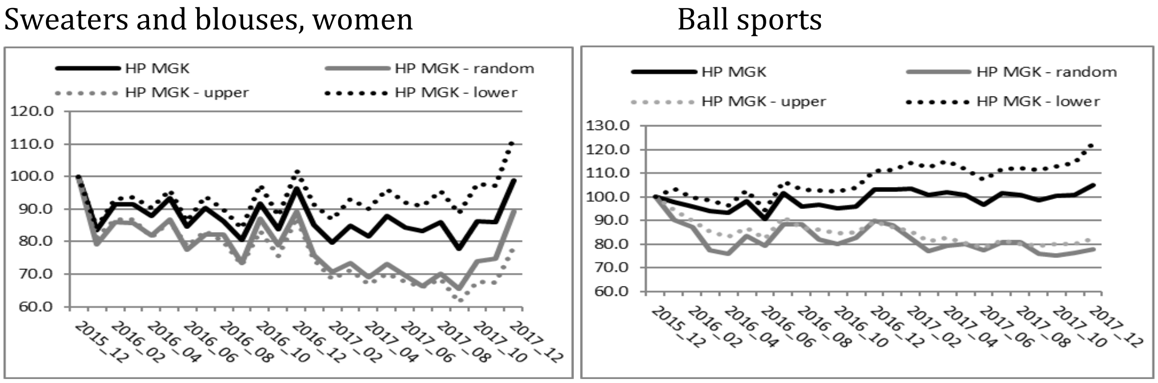


Ski equipment



Bicycles




Sweaters and blouses, women

Ball sports

Figure 2: HP misclassification sensitivity diagnostic

The index using the method of RPC is quite close to that using NPC in two of the commodity groups. Where the two methods differ, the index by RPC is always lower, even approaching the lower bound in two cases. To understand this result, one may examine the distribution of prices in a given commodity group. Figure 3 shows the price distribution for men's jackets, outdoor anoraks in October 2017, which shows that it is clearly skewed towards the lower end of price. Most of the existing items belong therefore to an HP closer to that end of price. It follows that, when the unmatched items are randomly classified into the existing HPs, relatively more items will incorrectly end up in the high-end HPs. This pulls the reference prices of these HPs downwards, generating a similar effect as by the method of UNPC, which biases the reference-price quantity index $Q^{0,t}$ upwards and the corresponding index $V^{0,t}/Q^{0,t}$ downwards.

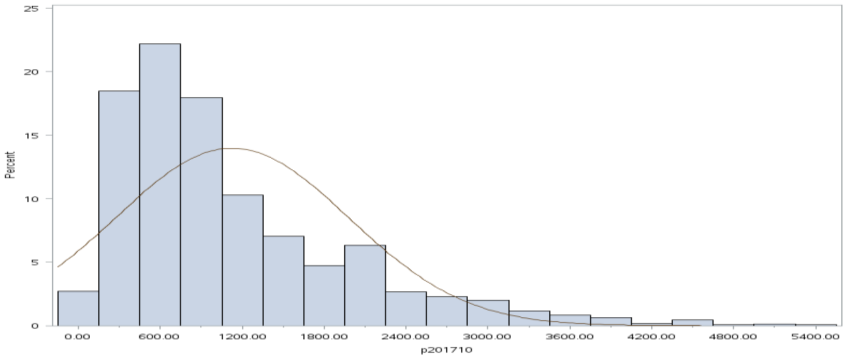


Figure 3: Distribution of price for men's jackets in October 2017

In summary, the proposed generic diagnostic helps one to identify the potential risk of bias, if one disregards the price data when the HPs only based on the available metadata are too heterogenous. Moreover, the artificial methods of LNPC and UNPC allows one to check whether NPC classification (or a plausible alternative) may be sensitive in reality.

3.4 HP formation

The main challenge of establishing the HPs is to achieve the appropriate balance between capturing replacement items without increasing the unit value bias, which might occur if the HPs are too coarsely defined and not sufficiently homogeneous. Formally, the discussion in Section 3.3 presumes given HP formation C = $\{c_1,...,c_M\}$. The question is how to arrive at C in the first place. As always, suppose that the available metadata has been used to form the initial ***item groups***, including outlet/retailer and brand blocking. The question remains whether and how to arrive at the final HPs, if there remains too much heterogeneity among the items *within* these groups. There are three obvious possibilities.

- By size, e.g. roughly equal number of items for each HP without overlapping price ranges between any two HPs.

- By turnover, e.g. roughly equal amount of turnover for each HP without overlapping price ranges between two HPs.

- By analysis of variance (ANOVA), e.g. minimum within-HP and maximum between-HP variance of prices. Non-overlapping price ranges between two HPs is ensured by ANOVA.

ANOVA is a standard statistical technique for decomposing the total variability in a dataset, where a relatively lower sum of within-group variances indicates greater group homogeneity. Chessa (2018) propose to use *match adjusted R-squared (MARS)* as the basis for forming the HPs, where the adjusted R-square is an ANOVA measure. In addition, the mismatch rate of the HPs over time is incorporated as a penalty, to avoid exceedingly fine HPs favoured by ANOVA on its own, when the number of groups is allowed to be as high as possible. Provided the HPs are only updated periodically, say, once a year, the match rate can be raised to maximum, if every unmatched item later on in the same period is classified into one of the existing HPs, e.g. using the method of NPC. Thus, in practice the MARS may not differ much from the ANOVA approach, if the classification of HPs in a given month t is decomposed into HP formation of existing (or yearly anticipated) items in the base month b and classification of unmatched items in month t.

One can easily envisage the effect of forming equal-size HPs in a plot like Figure 3: relatively more groups will be created where the price is more densely distributed, despite this is not effective from the perspective of ANOVA, if two neighbour HPs have very close prices anyway. Thus, as long as the number of HPs is allowed to be greater than that by ANOVA, the two methods can give similar results, provided the index formula can appropriately account for a fragmented ANOVA-HP by two or more equal-size HPs.

The method of equal-turnover HPs is analogous to the method of forming sampling strata in business surveys, in order to deal with the skewed distribution of the survey outcomes that are typical there. However, since price index is a different target than population totals that are typically the interest of business surveys, it is unclear whether one should simply adopt the method as such, especially provided the corresponding quantity and turnover can be taken care of by the index formula directly.

Below we outline two generic diagnostics for HP formation, one for the differences at a single time point, and the other for the effects of a chosen HP formation over time.

HP heterogeneity diagnostic Given the metadata item groups as usual, one may explore how the different methods behave in relation to each other, as the number of final HPs vary. Table 4 summarises the set-up.

Table 4: Set-up of generic diagnostic for HP heterogeneity

Base	$R(0,t)$	HP	Formula	Comment (for all)
(I.1)	$U_0 \cup U_t$	Size	GUV	Blocking by outlet/retail and brand...
(I.1)	$U_0 \cup U_t$	Turnover	GUV	... varying the number of HPs...
(I.1)	$U_0 \cup U_t$	ANOVA	GUV	... missing replacement as number increases

The HP formation according to size and turnover are both based on all the items between month 0 and t. The ANOVA approach however, is applied to month 0, and new items are classified into existing HPs each month.

On the one hand, provided initial heterogeneous group and that the index formula can handle the fragmentation of an HP into several nested HPs, the index should change as the number of HPs starts to increase from 1 but it should not be volatile as the no. of HPs increases further. On the other hand, as the number of HPs increases, the index will eventually approach a distinct-item index (discussed in Section 3.2), which is subjected to the effects of missing replacement.

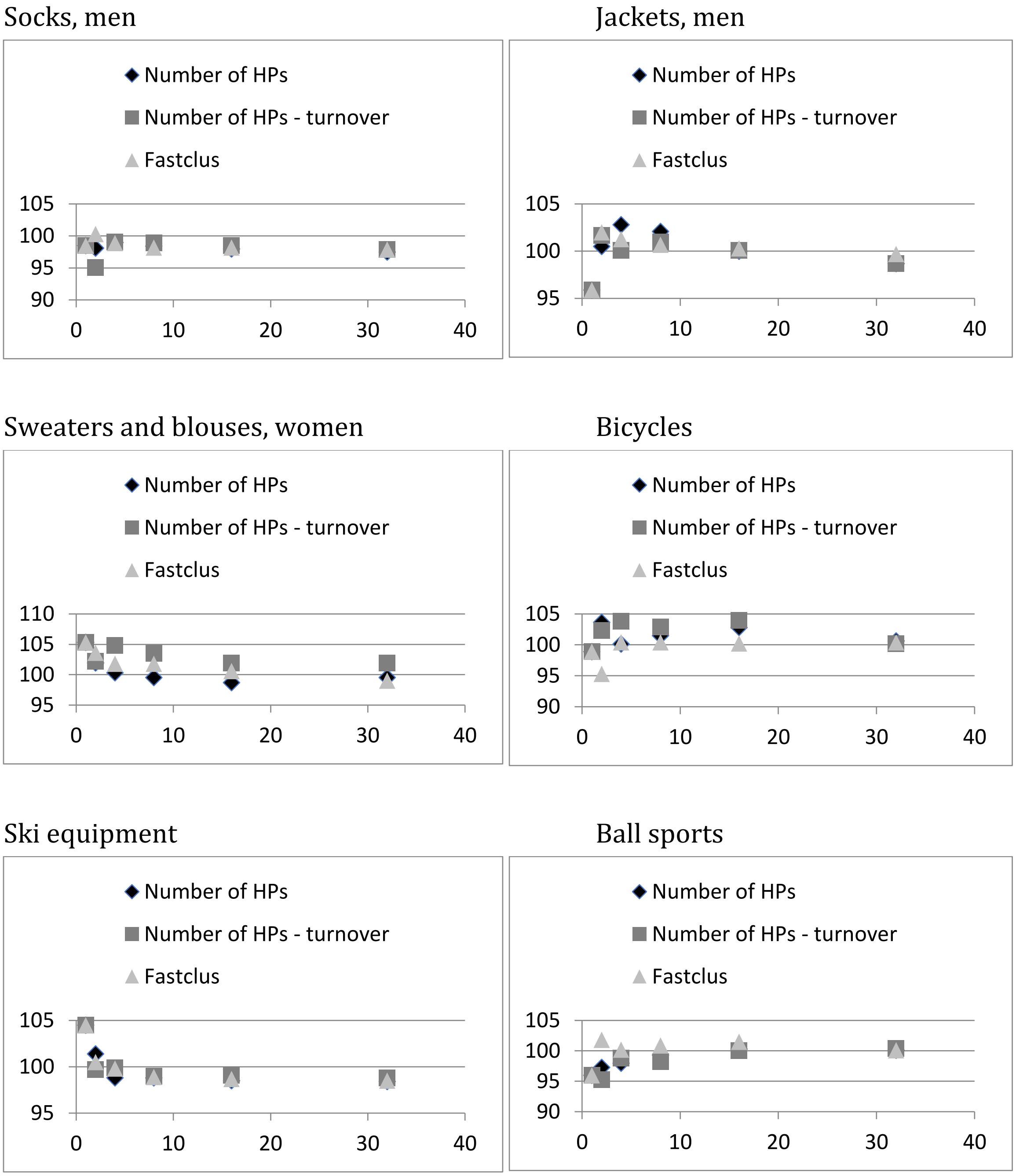


Figure 4: HP heterogeneity diagnostic

Application The results of applying the diagnostic of Table 4 are shown in Figure 4. The number of HPs increases from 1 to 2, 4, 8, 16 and 32. The ANOVA approach is implemented using the SAS procedure FASTCLUS. It can be seen that the indices are volatile initially, as the number of HPs starts to increase from 1, but become less so once the number reaches approximately 4. As the number increases further, the indices become more similar to each other in terms of general direction. As mentioned above, if we continue to increase the number of HPs, all the indices will eventually converge to the distinct-item index, which is subjected to the effects of missing replacement. In most of the commodity groups analyzed, this would lead to a negative bias, as discussed in Section 3.2. It seems that a sensible choice of the number HPs may be the point at which the index stops being volatile and starts to exhibit a steady trend, the latter of which may be an indication that the effects of missing replacement starts to take hold of the index. Between the three methods, the ANOVA-HPs and size-HPs are generally closer to each other at this point and onwards. Taken altogether, the diagnostic suggests e.g. forming approximately 4 ANOVA-HPs for these data may be a sensible choice.

HP formation diagnostic A generic diagnostic of a chosen HP formation over time can be based on the following intuition: a suitable HP formation for a dynamic item universe should also perform reasonably in the special case of fixed-item universe, where explicit HP formation is not absolutely necessary. We proceed as follows.

- Form the HPs using *all* the items in U_0, according to the chosen method below the item groups; note the initial item groups based only on available metadata.
- For month t, identify the matched item universe $U_{0t} = U_0 \cap U_t$; calculate the matched-item index $P_M^{0,t}$, the matched-HP index $P_{M,HP}^{0,t}$ and the matched-group index $P_{M,G}^{0,t}$.

Notice that the only difference between $P_M^{0,t}$ and $P_{M,HP}^{0,t}$ is the use of HP in the latter but not the former. Since both are calculated for the same matched item universe, there are no misclassification errors involved, and the difference between the two is entirely due to the HP-formation. Compared to $P_{M,HP}^{0,t}$, the index $P_{M,G}^{0,t}$ is based on a 'rougher' HP formation, which only uses the available metadata but not any method that treats the remaining heterogeneity within each item group. The set-up is summarised in Table 5.

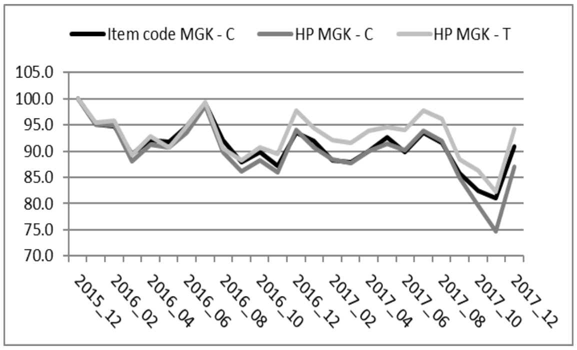
Table 5: Set-up of generic diagnostic for HP formation

Base	$R(0,t)$	HP	Formula	Comment
(I.1)	$U_0 \cap U_t$	No	$P_M^{0,t}$	Optional: superlative index formula
(I.1)	$U_0 \cap U_t$	Item group	$P_{M,G}^{0,t}$	HP = item group by metadata only
(I.1)	$U_0 \cap U_t$	ANOVA-HP	$P_{M,HP}^{0,t}$	Optional: other method below item group

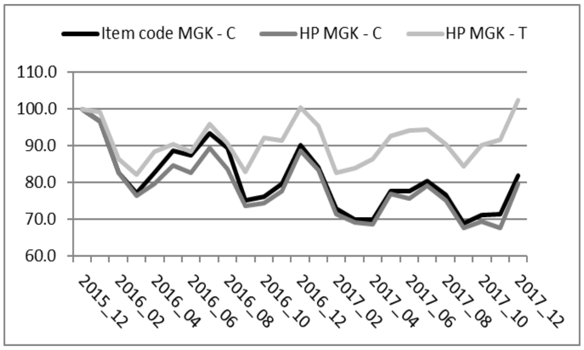
The diagnostic above is based on the matched item universe U_{0t} between 0 and t. This may conceivably run into difficulties in markets with extremely high churn rates, as the distance between 0 and t increases, so that the indices can only be computed based on a small fraction of the available data. In such cases, one may consider using a moving index base (I.2) in the set-up, and apply the diagnostic to the month-to-month matched item universe, where the HPs are initially formed using all the items in the (moving) base month.

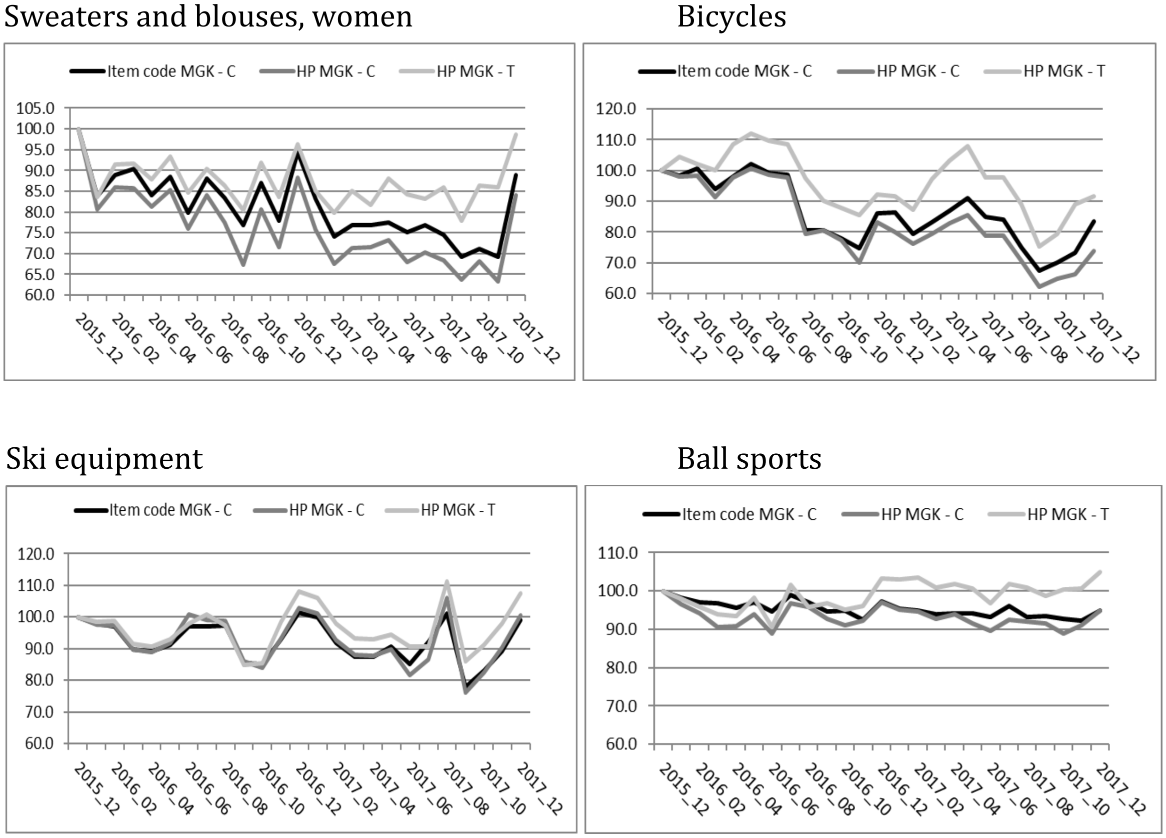
Application For ease of elaboration, the results of applying the diagnostic of Table 5 are presented in two figures. All the indices are calculated as the GUV index. In Figure 5, the index $P_M^{0,t}$ is compared to $P_{M,HP}^{0,t}$, referred to as item code MGK-C and HP MGK-C, respectively. The two indices are quite close to each other in most commodity groups, both in terms of short-term movements and the development over time. The HP-based index is only slightly more volatile. Figure 5 shows the index $P_{A,HP}^{0,t}$ given in Section 3.2, referred to as HP MGK-T. The difference between $P_M^{0,t}$ and $P_{A,HP}^{0,t}$ is a diagnostic of the effects missing replacement, which is seen to dominate the potential imperfection of HP formation, i.e. the difference between $P_M^{0,t}$ and $P_{M,HP}^{0,t}$, in all the commodity groups. It follows that the benefits of adopting the HP formation are likely to outweigh the issues caused by imperfect HP formation for these data.

Socks, men



Jackets, men





Sweaters and blouses, women

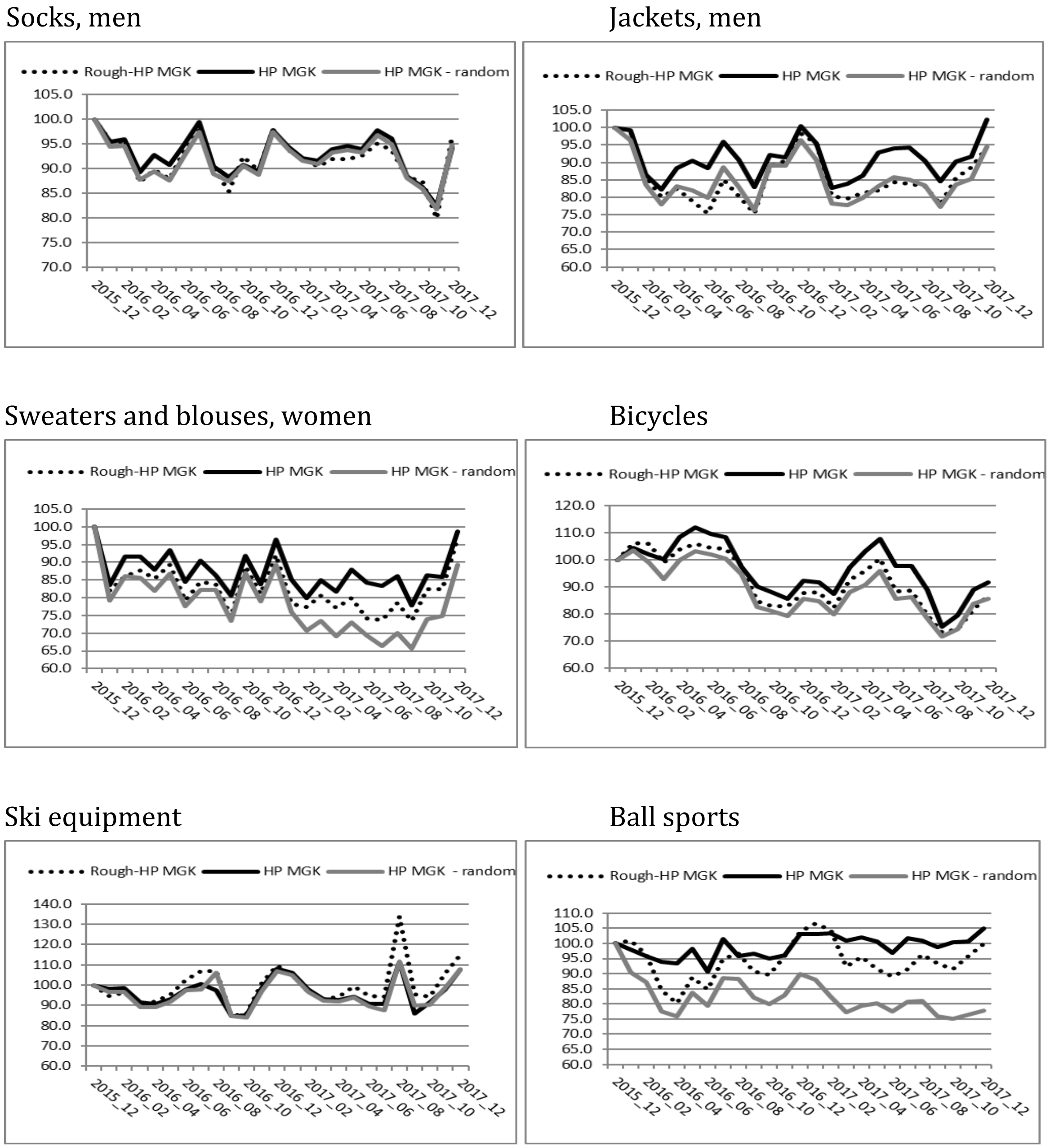
Bicycles

Ski equipment

Ball sports

Figure 5: HP formation diagnostic, plot 1

Next, In Figure 6, the index $P_{M,G}^{0,t}$ is compared to $P_{M,HP}^{0,t}$, indicated by rough-HP MGK and HP MGK, respectively. The index by RPC classification in Section 3.3 is included in addition. The rough-HP index $P_{M,G}^{0,t}$ is somewhat more volatile than $P_{M,HP}^{0,t}$. Indeed, the former is closer to the index based on RPC classification in all the commodity groups. Since RPC classification is clearly an inappropriate treatment of the heterogeneity below the item groups, disregarding the heterogeneity as by $P_{M,G}^{0,t}$ seems hardly acceptable for these data.


Socks, men

Jackets, men

Sweaters and blouses, women

Bicycles

Ski equipment

Ball sports

Figure 6: HP formation diagnostic, plot 2

4 Reference universe

In the case of bilateral index given the comparison months, say $(0,t)$, the reference universe $U_{0\cup t} = U_0 \cup U_t$ is naturally the *only* choice for *all* candidate index formulae, as long as one does not wish to reduce the index to the matched item universe only. The choice is certainly less obvious when it comes to multilateral index. For instance, while the GK index is often applied using an expanding window in the reported empirical studies, a 13-months rolling window is typically the choice for the GEKS index. Insofar as a multilateral index formula does not dictate the choice of reference universe, an extra decision is needed. In this Section we consider diagnostics for the choice of multilateral reference universe.

Notice that the need to choose the multilateral reference universe does not occur when using multilateral index for international (or spatial) price comparison, where the only natural choice would be the reference universe containing *all* the countries concerned. This serves as a reminder of the extra complications involved in adapting the methodology of spatial multilateral index for temporal price comparison (Zhang et al., 2017).

Problems with dynamic items For the comparison months $(0,t)$, where $t > 1$, all the difficulties for the choice of reference universe are caused by the different types of unmatched dynamic items in $U_{0\cup t} \setminus U_{0t}$.

- *Replacement items* By involving the intermediate months between 0 and t, a multilateral index can be based on multiple bilateral comparisons, each of which has a larger amount of matched items than U_{0t}. Still, adopting multilateral reference universe is unlikely to be more important than the use of HPs for dealing with replacement items.

- *Seasonal items* As discussed earlier, year-on-year monthly price comparison is the natural way for dealing with seasonal items. This requires only a bilateral reference universe. The most plausible multilateral reference universe should contain a full 13-months cycle, where the total amount of seasonal items (compared to the base month 0) must be larger over *all* the months than just in the bilateral case.

- *Regeneration items* As discussed before, the use of HPs can perturb the systematic biasing effects of regeneration items. Given fixed index base (I.1), a new entry item/HP in month t will keep causing the 'same' problem for a bilateral index in all the subsequent months, until the next time the base month is updated. For a multilateral index though, the problem can possibly be reduced from month $t + 1$ on, through the choice of index formula, provided the item/HP is matched between t and a later month. Similarly, given index base (I.3), a new item/HP in month t causes the 'same' problem to any bilateral index for $(t-12,t)$, $(t-11,t+1)$, ..., until it becomes matched for $(t,t+12)$ and onwards. But the problem can possibly be reduced using an appropriate multilateral index from $t + 1$ on, based on the reference universe of the previous 13 months.

- *Spurious items* A multilateral index must also deal with the items that are *spurious* to (U_0,U_t), which are the items that neither appear in month 0 nor t but only in the other months of the reference universe. A spurious item can be a seasonal item absent in months 0 and t, or it can be one with very short life span. A spurious item can have different effects depending on the index formula. For instance, it does not affect the GK index, but it does affect the GEKS index.

Expanding window diagnostics (EWD) Given the index base (I.1), the multilateral short-term index $P^{0,t}$ is calculated using the reference universe $U_0 \cup \cdots \cup U_t$, which expands as t increases and is referred to as the *fixed base monthly expanding (FBME)* window. Chessa (2017) compares an FBME index to the same index calculated based on a large fixed window of several years, and finds the differences to be small empirically for the HP-based GK indices using different lengths of the time window and updating methods, based on data from a department store and a supermarket chain. Chessa et al. (2017) further demonstrate that these effects are small against those due to the other "choice aspects" mentioned earlier in Section 1. Van Loon and Roels (2018) compare different window splicing options including the FBME window for indices based on SKU (stock keeping unit) codes, using the Belgium food market scanner data over a large fixed window. The different choices again seem to show only small effects for a given index formula.

Of course, these findings may not hold in other markets with higher churn rates and greater price fluctuations. Below we propose two generic diagnostics for the expanding window, focusing on the sensitivity of the corresponding multilateral reference universe.

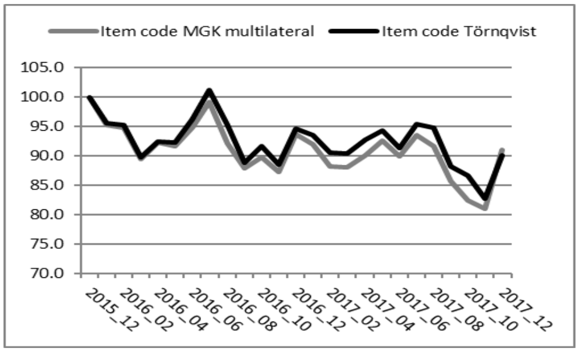
EWD-1 The basic idea is the following. In the case of fixed item universe, denoted by $U_M = U_0 \cap \cdots \cap U_T$ that is constructed for the whole study period consisting of T + 1 months, the choice of multilateral reference universe should be inconsequential provided the use of a transitive index. Indeed, it is possible to calculate a bilateral superlative index, denoted by $P_{sup}^{0,t}$, which can serve as the benchmark ideal, while avoiding the HPs and their confounding effects. The extent to which the direct index calculated using the FBME window differs from the benchmark index would thus provide a diagnostic of the sensitivity of the particular choice of reference universe. Table 6 summarises the set-up, where $P^{0,t}$ denotes a multilateral index of choice. Notice that for each $t = 2,...,T$, the index is multilateral using the FBME window, since it uses all the data associated with U_M between month 0 and t, despite the item universe U_M itself is held fixed over time.

Table 6: Set-up of generic diagnostic for expanding window (EWD-1)

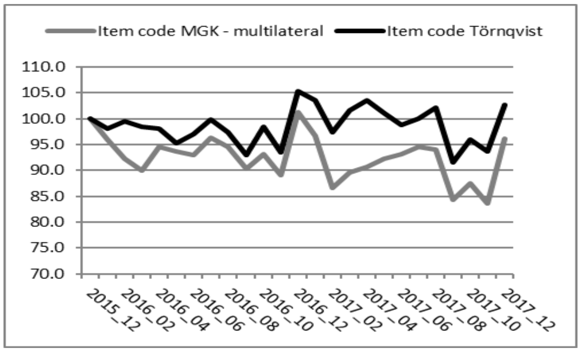
Base	$R(0,t)$	HP	Formula	Comment
(I.1)	$U_0 \cap \cdots \cap U_T$	No	$P^{0,t}$	Multilateral index, fixed item universe 0 to T
(I.1)	$U_0 \cap \cdots \cap U_T$	No	$P_{sup}^{0,t}$	A bilateral superlative index of choice

Application Figure 7 shows the results of applying EWD-1 to our data, where the Törnqvist index is used as the bilateral superlative index $P_{sup}^{0,t}$ in Table 6, against the multilateral GUV index using the FBME window. The two indices show little difference in volatility. Their long-term trends agree quite well with each other. The multilateral index seems to lie mostly below the Törnqvist index, indicating possibly a small negative bias over the study period. But the effect seems small compared to some of the effects associated with the choice of HPs earlier.

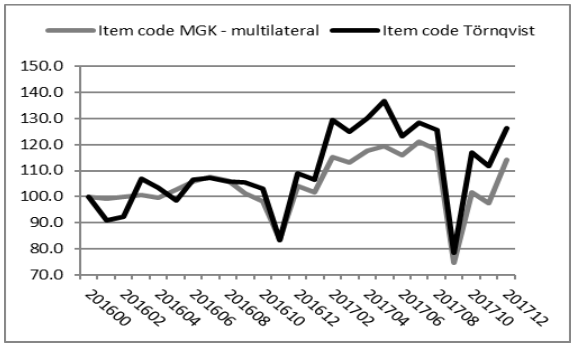
Socks, men



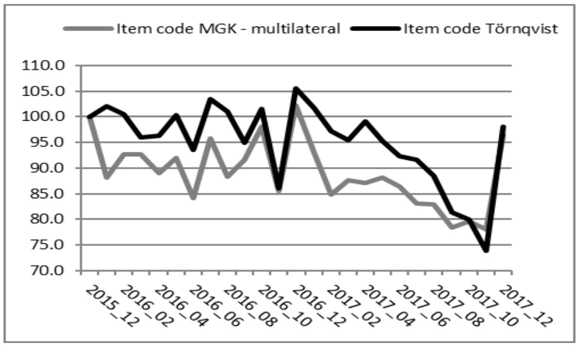
Trousers, men



Fishing equipement



Sweaters, men



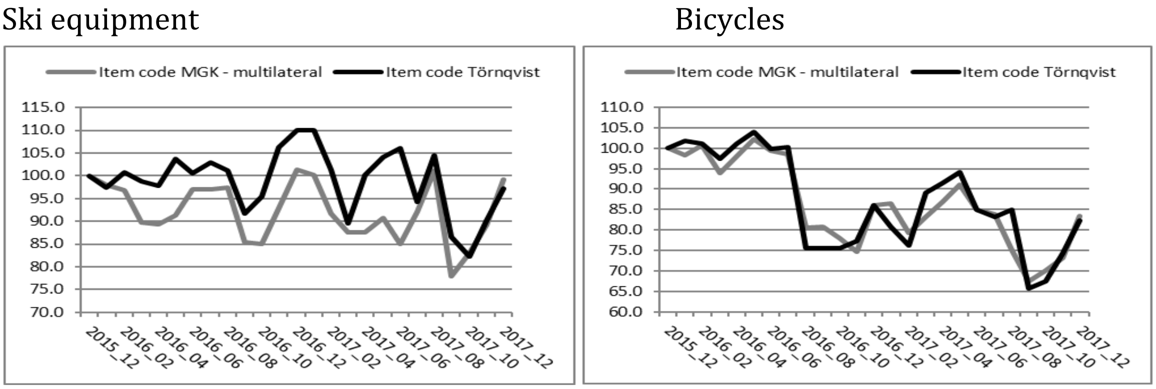


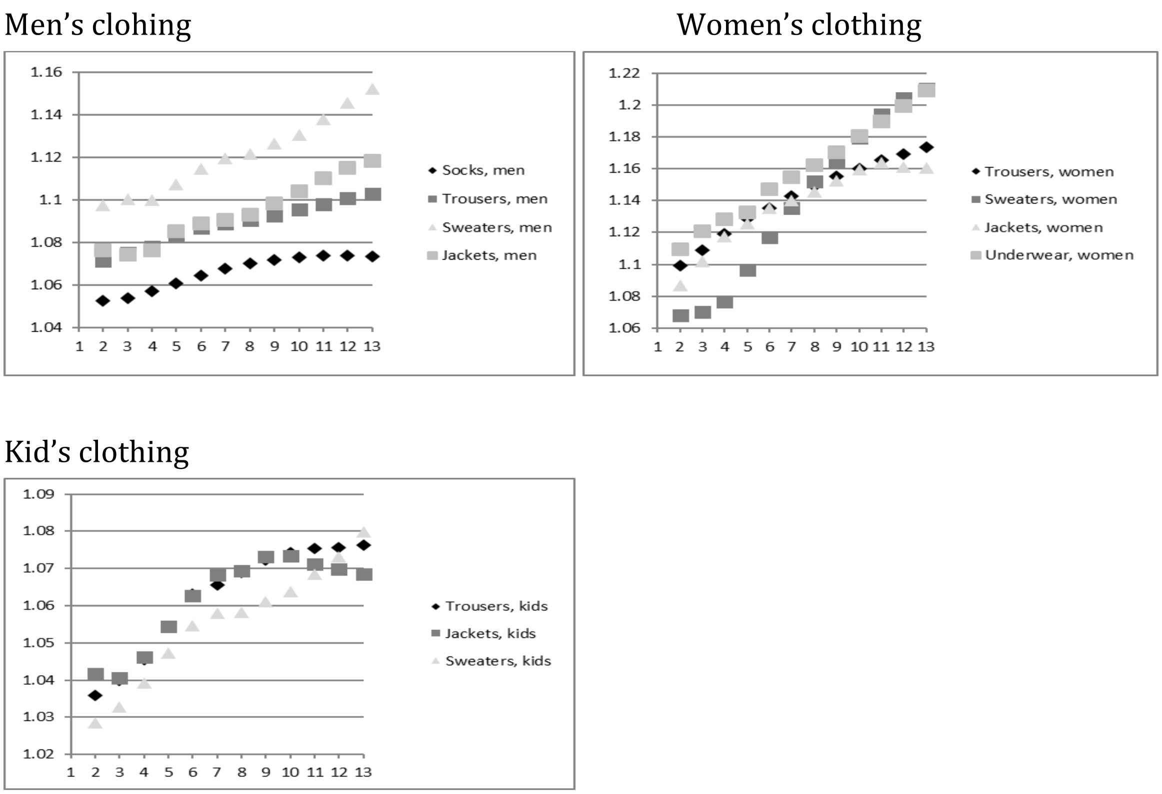
Figure 7: Expanding window diagnostic EWD-1

EWD-2 A straightforward generic diagnostic for the sensitivity of the FBME window for a multilateral index formula is simply to calculate it *repeatedly* for the *same* comparison months (0,t), but using different possible FBME windows. Table 7 summarises the set-up when the window is expanding backwards for the same index $P^{0,t}$. It is equally possible to expand the window forwards, by adding months t + 1, ..., 12 one at a time. Moreover, one might wish to apply the diagnostic using HPs instead, in order to check the sensitivity *given* the choice of HPs. In any case, the more the index $P^{0,t}$ varies over the different reference universes, the more sensitive is the choice.

Table 7: Set-up of generic diagnostic for expanding window (EWD-2)

Base	$R(0,t)$	HP	Formula	Comment
(I.1)	$U_0 \cup \cdots \cup U_t$	No	$P^{0,t}$	Minimum multilateral choice
(I.1)	$U_{-1} \cup U_0 \cup \cdots \cup U_t$	No	$P^{0,t}$	One additional time point
	⋮			⋮
(I.1)	$U_{t-12} \cup \cdots \cup U_t$	No	$P^{0,t}$	Full 13-months window backwards

Application Figure 8 shows the GEKS index from November to December 2017, on applying the EWD-2 to our data. The length of the expanding window is given on the X-axis. The sensitivity of the GEKS index varies across the commodity groups: the indices for kid's clothing vary least, while those for women's clothing vary most. The variation is the largest for women's sweater. In all the cases, the GEKS increases with the length of window. Since the bilateral GEKS index reduces to a matched-item index, the bias may be largely due to the effects of missing items.


Men's clohing

Women's clothing

Kid's clothing

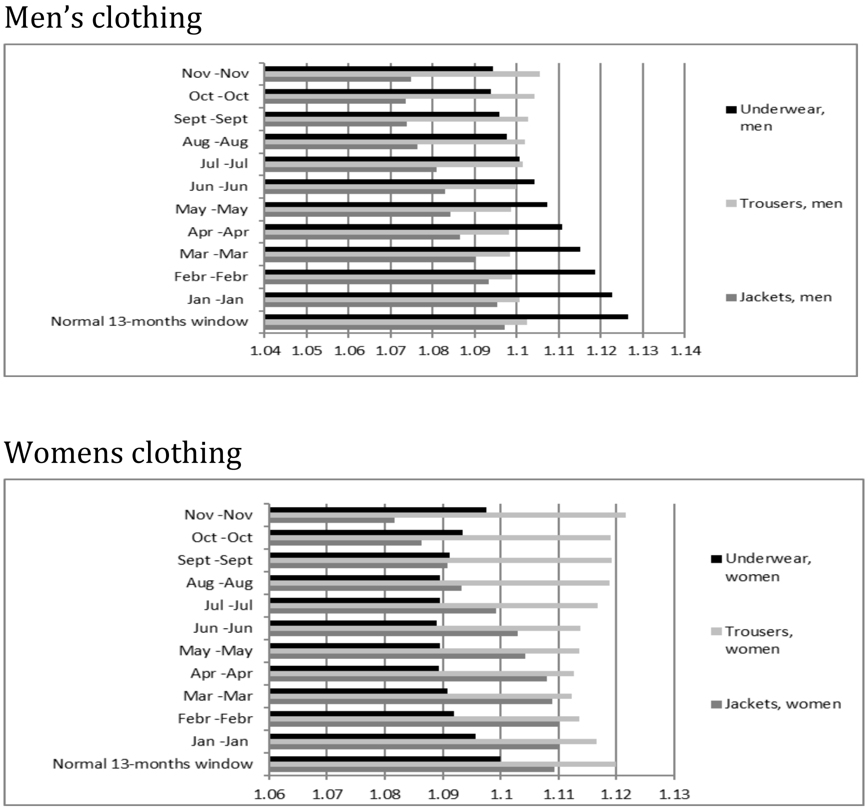
Figure 8: Expanding window diagnostic EWD-2

Fixed-length window diagnostic A multilateral reference universe is a *moving base fixed-length (MBFL)* window, if it consists of a fixed number of months and is applied with a moving base over time. Table 8 specifies a simple diagnostic. Given the comparison months $(0,t)$, calculate a chosen multilateral index $P^{0,t}$ *repeatedly*, using all the possible full 13-months windows $(t-12,...,0,...,t)$, $(t-11,...,t,t+1)$, ..., $(0,...,t,...,12)$. The greater the index varies with the different MBFL windows, the more sensitive is the choice.

Table 8: Set-up of generic diagnostic for fixed length window

Base	$R(0,t)$	HP	Formula	Comment
(I.1)	$U_{t-12} \cup \cdots \cup U_t$	No	$P^{0,t}$	Usual choice in practice
(I.1)	$U_{t-11} \cup \cdots \cup U_{t+1}$	No	$P^{0,t}$	Whole window moved one month forward
	$\vdots$			$\vdots$
(I.1)	$U_0 \cup \cdots \cup U_{12}$	No	$P^{0,t}$	The last possible 13-months window

Application The diagnostic of Table 8 is applied to the GEKS index from November to December 2016, based on all possible 13-months windows. The results are given in Figure 9. The stability of the MBFL window varies across different commodity groups: for a group like trousers for men the index changes little across the different reference universes, but the variation is noticeable for many other groups, probably due to different churn rates.


Men’s clothing

Womens clothing

Figure 9: Sensitivity of MBFL window diagnostic

5 Index base

Provided the HPs are able to capture all or nearly all the replacement items, the choice of index base (defined in table 1) is most affected by the seasonal and regeneration items.

- For a fixed base index (I.1), one can use an annual basket that includes the seasonal items from the other months with imputed prices in base month 0. However, imputation of non-existing price (or quantity) is not easy theoretically, and it complicates the interpretation of the index. It requires these items to be explicitly identified, which can be resource-demanding or impossible for scanner data. Meanwhile, since some seasonal items are absent in the index for months (0,12), their effects do not persist over time. As noted before, a regeneration item in month t could have a lingering effect for the subsequent months, when the base month is fixed for a whole year. But the effect may differ depending on whether the reference universe is multilateral or bilateral, as well as on the choice of index formula.

- The inclusion of seasonal or regeneration items in a month-to-month chained index with moving base month (I.2) is often found to lead to bias, e.g. if the typical relatively high first-appearance prices cannot be appropriately accounted for. Since generally month-to-month chaining cannot satisfy the identity and fixed-basket tests (Zhang et al., 2017), the risk of chain drifting cannot be avoided in a dynamic universe.

- While choosing the 12-month fixed base (I.3) and year-on-year monthly price comparison is the best option for seasonal items, one would presumably have to deal with an increased amount of regeneration items, although the effects can possibly be mitigated by a suitable choice of reference universe and index formula. Meanwhile, the choice requires a series of 12-month initial indices, before year-on-year monthly indices can be calculated. Let $P^0,...,P^{11}$ be the indices for the first 12 months. Since e.g. P^5/P^2 would be present in any future indirect index from March to June in the same year, whatever the initiation errors of $P^0,...,P^{11}$ may have, they will affect all future short-term indices.

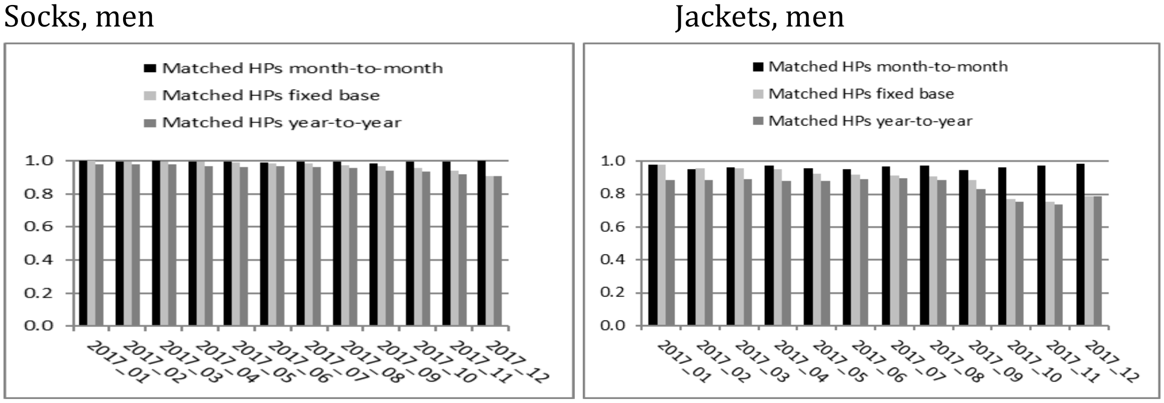
It seems important to check the actual dynamics in a given market. Chessa et al. (2017) used three statistics, which they refer to as “flows”, which are the items that are sold in one month and the next, “outflow”, which are the items sold in one month but not the next, and “inflows”, which are the items sold in one month but not the previous. In other words, this provides a generic diagnostic of the month-to-month dynamics.

Table 9: Set-up of generic diagnostic for flow dynamics

Base	$R(b,t)$	HP	Share = flow/(flow + outflow + inflow)
(I.1)	$U_0 \cup U_t$	No/Yes	By count or expenditure
(I.2)	$U_{t-1} \cup U_t$	No/Yes	By count or expenditure
(I.3)	$U_{t-12} \cup U_t$	No/Yes	By count or expenditure

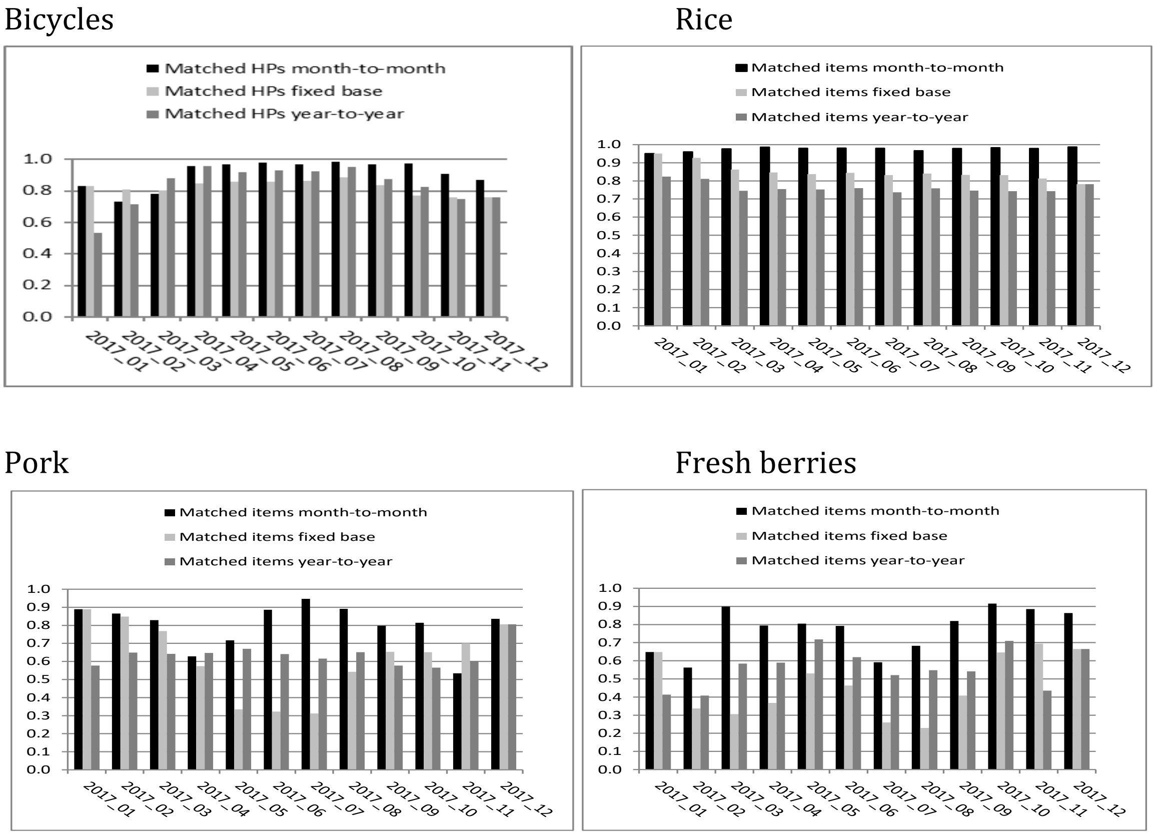
We extend the diagnostic in Table 9. The unit of flow can be item or HP. One may calculate a share of the flow either based on the counts or expenditure of the chosen unit. When all the replacement items are captured by the flow of HPs, the difference of the share to unity in the case of index base (I.2) measures the monthly total of seasonal and regeneration items. Notice that there is no difference between the shares by (I.1) and (I.2) in the first month of a yearly cycle. Since the problems concerning seasonal items are absent in the case of index base (I.3), the difference between the shares by (I.1) and (I.3) measures the total amount of seasonal items between months 0 and t and the regeneration items between t−12 and 0. The difference between them disappears at the end of a yearly cycle, when the difference to unity measures the total amount of regeneration items over a year.

Application Figure 10 shows the dynamics of HPs in some commodity groups both for sport clothing and equipment as well as for food. In most groups the shares are similar for index bases (I.1) and (I.3). For groups with prominent seasonal patterns, there can be large differences between the two as for instance for bicycles, pork and fresh berries. For fresh berries the two differ in most months indicating a greater sensitivity of the choice (I.1) than that of the choice (I.3). Thus, applying the *same* choice of index base to *all* the commodity groups has its risks.



Socks, men

Jackets, men



Bicycles

Rice

Pork

Fresh berries

Figure 10: Flow dynamics diagnostic

Sensitivity of fixed base month diagnostic The most resource-demanding task for diagnosing empirically the effects of seasonal and regeneration items is to explicitly identify these items in the vast amount of scanner data. Insofar as this may be impossible in many applications, one would need less rigorous diagnostics which can still be helpful. Table 10 provides two generic diagnostics for base (I.1) and (I.3), respectively. In both cases, the use of HPs is needed to remove the effects of replacement items as much as possible.

In the diagnostic for index base (I.1) one calculates *repeatedly* the index for a given month *t*, using 12 different base months. It is preferable to use an index that pertains only to the items in these two months, i.e. unaffected by the spurious unmatched HPs in any other month. The idea is to explore the sensitivity of base month chosen for (I.1). For instance, one may compare P^t given b = December to the average of 12 P^ts given different base months. Or, one may compare P^t given b = December to P^t given b = July to see what happens if the base month is in summer instead of winter.

Given index base (I.3), the index $P^{t-12,t}$ is mainly subjected to the effects of regeneration items, since seasonal items are absent between $t-12$ and t, provided the replacement items are captured by the HPs. To check this, one may compare a bilateral index that is much affected by the unmatched HPs and a multilateral index that is maybe less affected by the unmatched HPs. To the extent the multilateral index is able to mitigate the effects of regeneration items, the difference to the bilateral index would provide an indication of the effect of regeneration items. Moreover, if the differences are large, then some effect may potentially remain even if one chooses the multilateral index that is more resilient to such effects.

Table 10: Set-up of generic diagnostic for index base

Base	$R(b,t)$	HP	Formula	Comment
(I.1) base 0	$b = t - 1$: $U_{t-1} \cup U_t$	Yes	GUV	$P^t = P^{0,t-1}P^{t-1,t}$
(I.1) base 0	$b = t - 2$: $U_{t-2} \cup U_t$	Yes	GUV	$P^t = P^{0,t-2}P^{t-2,t}$
⋮		⋮		
(I.1) base 0	$b = t - 12$: $U_{t-12} \cup U_t$	Yes	GUV	$P^t = P^{0,t-12}P^{t-12,t}$
Base	$R(t - 12,t)$	HP	Formula	Comment
(I.3)	$U_{t-12} \cup \cdots \cup U_t$	Yes	GEKS	Multilateral reference universe
(I.3)	$U_{t-12} \cup U_t$	Yes	GEKS	Affected by regeneration items

Application The results from applying the diagnostic for index base (I.1) is given in Figure 11. The index using December as the base month is compared to the average of the same index using all 12 possible base months. The two are quite close to each other in most cases, where the particular choice of December is not sensitive. There are nevertheless exceptions, such as the summer time indices of ski equipment, where the two differ quite much. The result is in line with that of the flow diagnostics above: the choice of fixed base month (I.1) can be sensitive for commodity groups dominated by seasonal items.

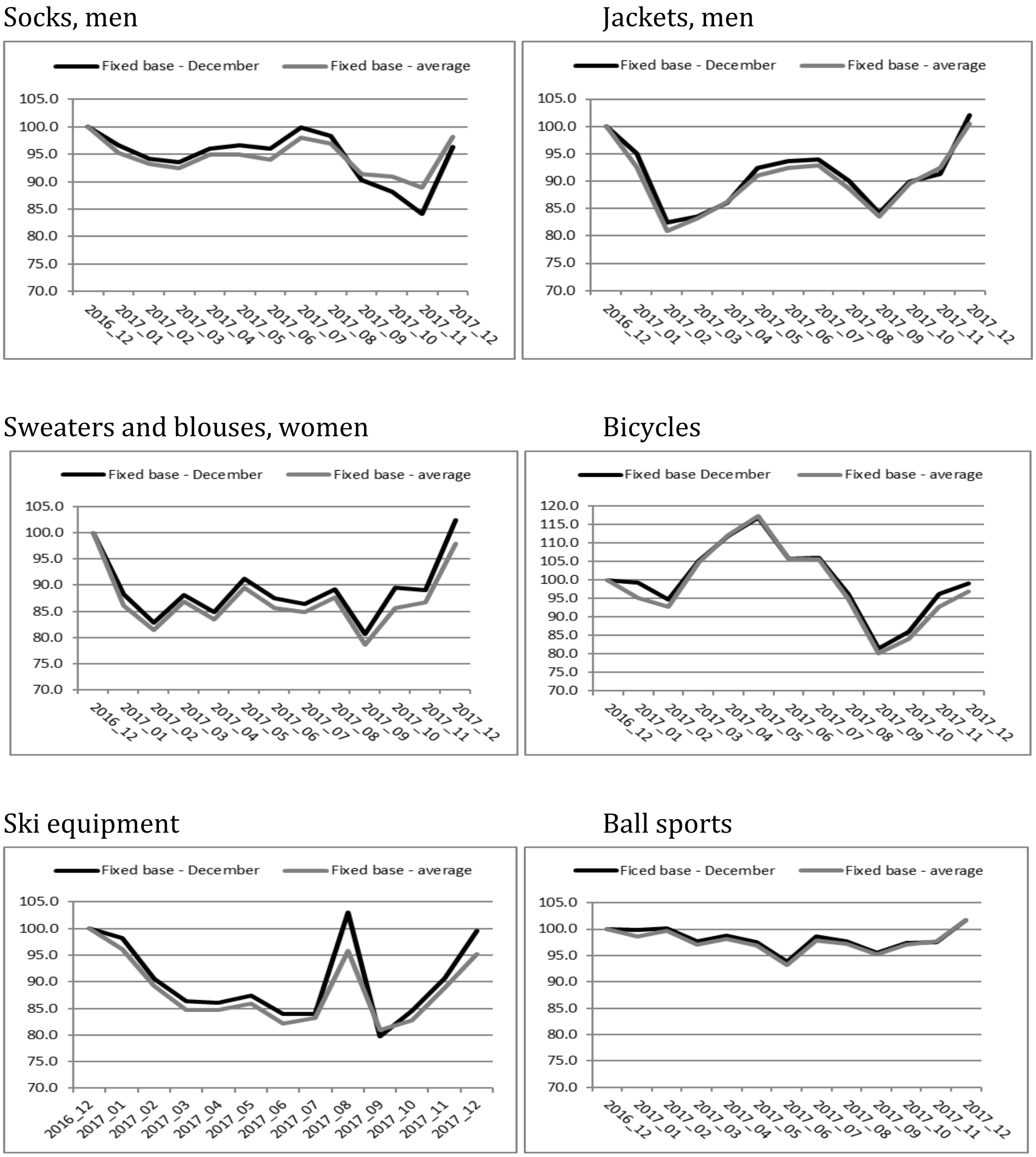
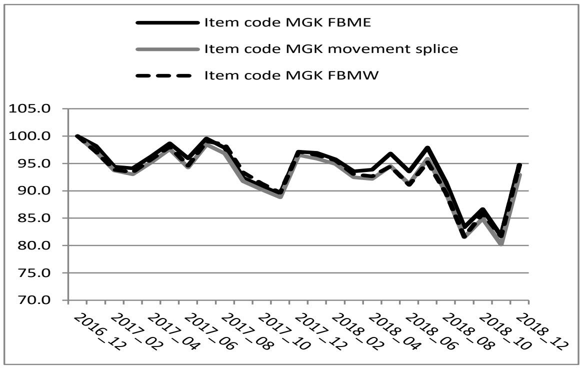


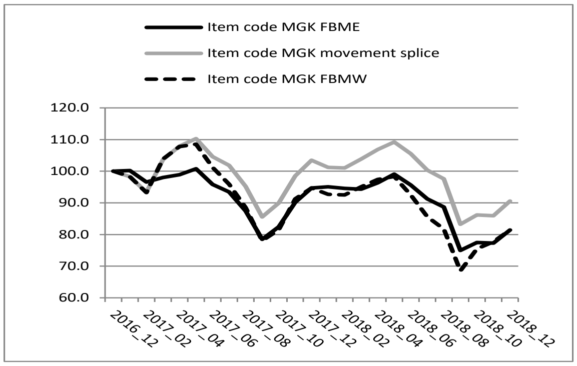
Figure 11: Sensitivity of fixed base (I.1) diagnostic

Splicing So far we have analyzed choices concerning HPs, reference universe and index base. In the comparisons below indices combining the last two choices are analyzed. Figure 12 shows MGK based on item codes for some commodity groups covering food as well as sport clothing and equipment. The indices are calculated as MGK using different time windows and splicing techniques. MGK with a FBME uses all data between December and current month *t*. The window expands as *t* increases. At the end of each yearly cycle it coincides with the 13-months multilateral reference universe. MGK with *movement splice* is based on a 13-months window, where the index from last month *t-1* is spliced onto the existing time series. The MGK with *fixed base moving window* (FBMW) however, is a combination of the other two as it is based on a 13-months window but at the same time uses December as fixed base, and where the last month of the window is always compared to December (Lamboray, 2017, Van Loon and Roels, 2018). The FBMW equals the movement splice the first month of the yearly cycle, and equals the FBME at the end of the yearly cycle. Except for some short-term deviation, many of the commodity groups show rather similar price development. However, for bicycles and jackets for men for instance, there seem to be a more persistent effect, which might be related to how the different time windows, in combination with the splicing techniques, are able to capture seasonal items.

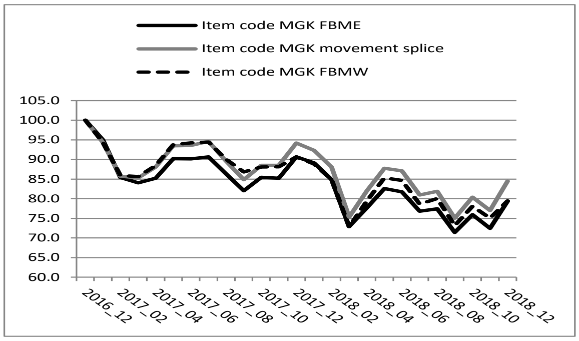
Socks, men



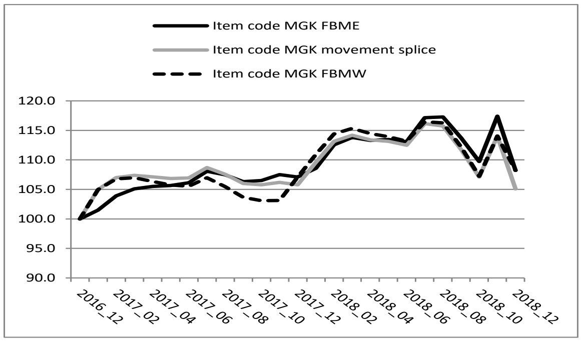
Bicycles



Jackets, men



Rice



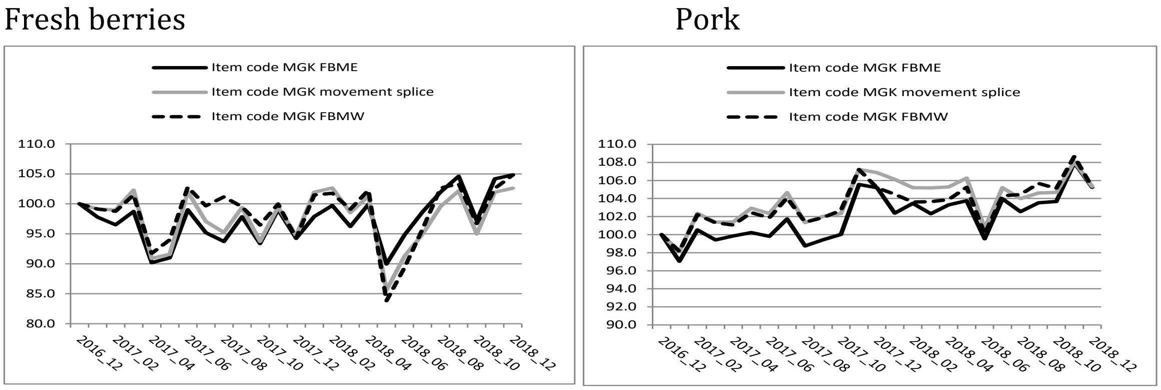


Figure 12: Different splicing techniques

6 Results on different index formulas

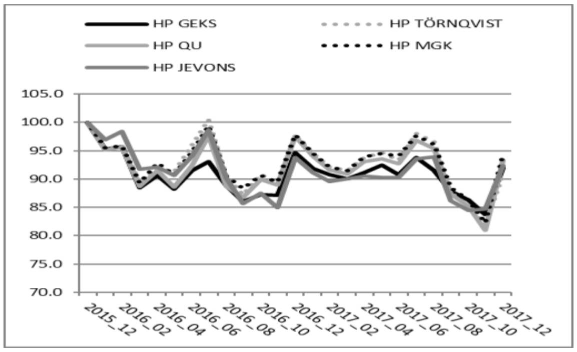
At this stage one does not have an ideal or ‘superlative’ index for a dynamic item universe. Provided replacement items can be handled by the creation of HPs, their effects on the index formula choice are much reduced. The best treatment of seasonal items may depend more critically on the combined choices of index base and reference universe.

One does not need to resort to multilateral index to avoid drift, since it can easily be achieved by a bilateral index. However, a bilateral index suffers a fundamental short-coming regarding the regeneration items: by definition it cannot be genuinely responsive, because there is simply no comparison data for an unmatched item (or HP) in $U_{0\cup t} \setminus U_{0t}$. Even when an index formula is formally responsive, so that it does not reduce to a matched-universe index for U_{0t}, its treatment of the new and out-going items (or HPs) can only be limited due to this lack data for price comparison. Given that there are data for price comparison, the question becomes one of index formula and the answer is still not obvious.

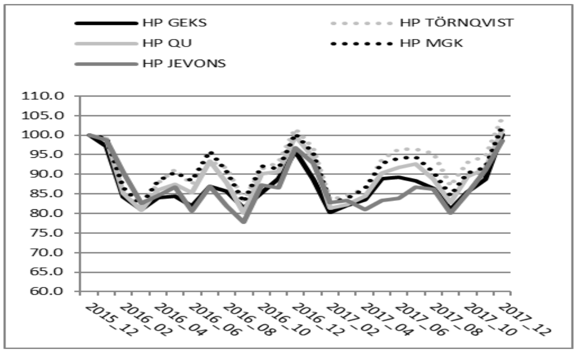
So far the different choices related to HPs, reference universe and index base have been studied. In this Section we compare different index formulas based on different combinations of these necessary choices. In Figure 13 all the index series are based on HPs, but differ according to index base and reference universe. The *multilateral indices;* GEKS (Rolling year GEKS) index is based on a 13-months window and movement splice while the GK/”Quality adjusted unit value index” (QAUV) or “QU-index” in short (Chessa, 2016) and the MGK indices are based on FBME window. The *bilateral* indices; Jevons index is based on a sample of the most sold items while splicing month-on month movements (the so-called “dynamic method”) while the Törnqvist index is calculated using a fixed December base.

The mix of the different methods illustrates how the build up of price indices is not simply a question of index formula, but a set of choices must be made. The results show that the deviations between the price indices are rather small. For some of the the commodity groups and especially the groups affected by seasonal pattern like ski equipment, we clearly see some short-term deviations, but the long-term trends seem to coincide to a larger extent. The index that seems to deviate the most seems to be the Jevons index that illustrates the need of using explicit weighting at elementary level. As shown earlier the HP formation is one of the most important issues to solve as the effects may be both systematic and large and the choice of HP formation seems to be as important as the choice of index formula itself when it comes to dynamic universe and items of high churn.

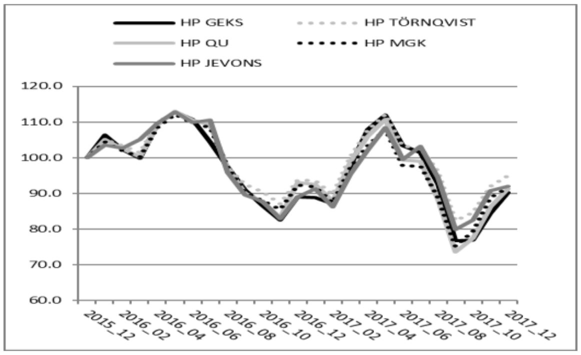
Socks, men



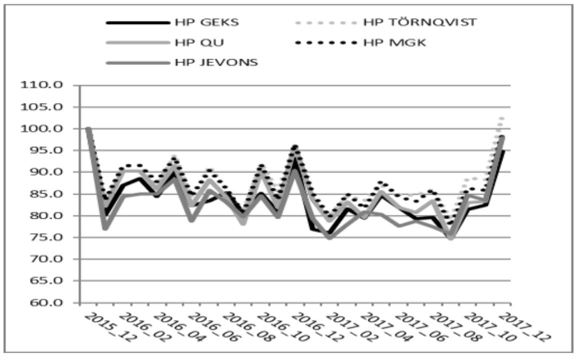
Jackets, men



Bicycles



Sweaters and blouses, women



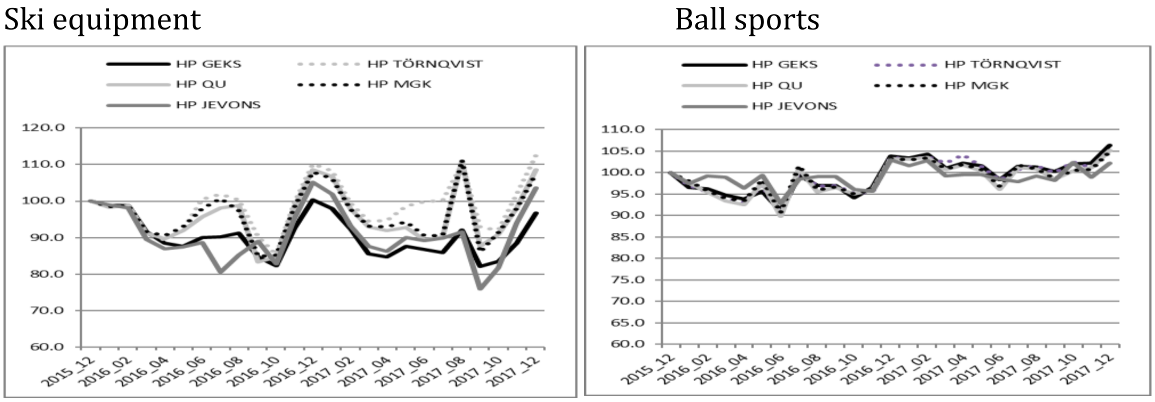


Figure 13: Different calculation methods

7 Summary and conclusions

Below we summarise our conclusions regarding each necessary choice, based on the general discussions and empirical results above. The generic diagnostics proposed in this paper are listed in Table 11.

Table 11: List of generic diagnostics

Choice	Effect	Feature / Comment
HP	Missing replacement Misclassification Heterogeneity Formation	Matched item universe vs. entire universe For given HPs; may use price within item groups Heterogenous item groups using only metadata Form HPs using all items; check only matched ones
$R(0,t)$	Expanding window Fixed-length window	EWD-1: using matched universe EWD-2: using entire universe $P^{0,t}$: same $(0,t)$, all possible MBFL windows
Base	Flow dynamics Sensitivity of base	Effects of season and regeneration if HP-based For index base (I.1) or (I.3)

Regarding creation of HPs

- For many markets, such as sport clothing, the effects of missing replacement can be large, in which cases the creation of HPs is necessary and important.
- Detectable heterogeneity can remain within the item groups that are formed only based on available metadata. Using price as a proxy quality measure for HP formation and classification below the item-group level can reduce the volatility as well as the bias of the index based on item groups.
- The number of HPs within each item group can be a key choice aspect in HP formation. In this respect the ANOVA approach seems the most intuitive option at present, and is expected to yield statistically most stable results.

Regarding multilateral vs. bilateral reference universe

- There are no clear arguments in favour of multilateral reference universe, with respect to the replacement items, the seasonal items and the spurious items to the given comparison universe. However, using multilateral instead of bilateral reference universe can make the index more responsive to regeneration items, provided suitable choice of the index formula.
- The 13-months fixed-length window is less ad hoc than the monthly expanding window. It more clearly reflects a necessary compromise towards the different dynamic items. The sensitivity of the choice can be explored empirically for a given market, using the MBFL window diagnostic both with and without the HPs.

Regarding index base

- Monthly moving base (I.2) can never avoid the risk of chain drifting in a dynamic universe. The potential effect depends on reference universe and calculation method.
- Single fixed base month (I.1) resembles the traditional CPI practice. It coincides with the 13-months multilateral reference universe at the end of each yearly cycle. The combined choice of (I.1) and FBME window is applicable from the beginning. It can be more problematic for COICOP groups dominated by seasonal items.
- Fixed 12-month base (I.3) suits best the measure of year-on-year price development. The short-term index for two months within the same year needs to be calculated indirectly. The index must be calculated differently in the first 12 months.

Our overall aim is to implement a generic solution which can be applied across different commodity groups and also to incorporate expenditure shares at most detailed level. No international consensus on calculation method has yet been reached, and we notice that National statistical institutes (NSIs) in similar situation may choose differently.

Based on a framework we have systematically gone through different choices, like formation of HPs, index base and reference universe. As the above summary illustrates, the choices are many and the anwers are not obvious. We regonize the importance of using HP in order to capture replacement items. We are also in favor of implementing a multilateral price index formula that better manages to capture regenerations items compared to their bilateral counterparts. Furthermore, using a fixed length 13-months window to capture the price development of seasonal items seems to be a good choice. For the time being, no final conclutions are made and the work will go on within the grant agreement.

References

[1] von Auer, L. (2014). The generalized unit value index family. *Review of Income and Wealth*, 60, 843-861.

[2] Chessa, A.G. (2018). *Product definition and index calculation with MARS-QU: Applications to consumer electronics.* Eurostat Grant Agreement project report.

[3] Chessa, A.G. (2017). *Comparisons of QU-GK indices for different lengths of the time window and updating methods.* Paper presented at an informal Meeting on multilateral methods, Luxembourg:14-15 March 2017.

[4] Chessa, A.G. (2016): *A new methodology for processing scanner data in the Dutch CPI.* EURONA, 1/2016, Eurostat, pp. 49-70.

[5] Chessa, A.G., Verburg, J. and Willenborg, L. (2017). *A comparison of price index methods for scanner data.* Paper presented at the fifteenth Ottawa Group meeting, Eltville, Germany.

[6] Dalén, J. (2017). *Unit values and aggregation in scanner data – towards a best practice*. Paper presented at the fifteenth Ottawa Group meeting, Eltville, Germany.

[7] Dalén, J. (2001). *Statistical targets for price indexes in dynamic universes*. Paper presented at the sixth Ottawa Group meeting, Canberra, Australia.

[8] Diewert, E. W. and Fox, K. J. (2017). *Substitution bias in multilateral methods for CPI construction using scanner data.* Paper presented at the fifteenth Ottawa Group meeting, Eltville, Germany.

[9] Geary, R. C. (1958). A note on comparisons of exchange rates and purchasing power between countries. *Journal of the Royal Statistical Society, Series A*, 121, 97-99.

[10] de Haan, J. (2002). Generalized Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI). *Journal of Official Statistics*, 18, 61-85.

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

[12] IWGPS (2004)*: Consumer Price Index Manual: Theory and Practice*. International Labour Organization.
http://www.ilo.org/public/english/bureau/stat/guides/cpi/CPI_Manual.html

[13] Ivancic, L., Fox, K. J. and Diewert, E. W. (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics*, 161, 24-35.

[14] Lamboray, C. (2017). *The Geary Khamis index and the Lehr index: how much do they differ?* Paper presented at the fifteenth Ottawa Group meeting, Eltville, Germany.

[15] Lehr, J. (1885). *Beitrüge zur Statistik der Preise insbesondere des Geldes und des Holzes*. F. D. Sauerländer Verlag, Frankfurt a. M.

[16] Van Loon, K and Roels, D. (2018). *Integrating big data in the Belgian CPI.* Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, Geneva, Switzerland: 7-9 May 2018.

[17] Zhang, L.-C., Johansen, I. and Nygaard, R. (2017). *Testing unit value data price indices.* Paper presented at the fifteenth Ottawa Group meeting, Eltville, Germany.