MARS: A method for defining products and linking barcodes of item relaunches

Antonio G. Chessa[1]

Abstract

The occurrence of product relaunches at the barcode (GTIN) level is a well-known problem, which has to be resolved before price indices can be calculated. GTINs of outgoing items have to be linked to the GTINs of reintroduced items. Both items are usually of the same quality, but may have different barcodes and prices. The resulting price changes have to be captured when calculating price indices.

The growing popularity of transaction and web scraped data calls for a method that automates barcode linking to a high degree. This paper presents a method that groups GTINs into 'products' by balancing two measures: one measure quantifies the 'homogeneity' of GTINs within products, while the second measure expresses the degree to which products can be 'matched' each month with respect to a fixed comparison period. The two measures have opposite effects in a nested stratification scheme: tighter defined products are more homogeneous but have the same or worse product match over time compared to broader defined products.

A method (MARS) is proposed, which combines explained variance in product prices with product match over time. MARS can be used to evaluate and rank different partitions of GTINs, such that the partition or stratification scheme is selected with the highest combined value of R squared and product match.

MARS has been applied to a broad range of product types. Individual GTINs are suited as products for food and beverages, but not for product types with higher rates of churn, such as clothing, pharmacy products and electronics. In these cases, products are defined as combinations of characteristics: GTINs with the same characteristics are grouped into the same product. MARS not only solves and automates the product definition problem, but is also useful as a data monitoring tool.

Keywords: CPI, transaction data, GTIN, relaunch, product homogeneity, stratification.

1 Introduction

The increased availability of electronic transaction data sets for the consumer price index (CPI) offers possibilities to national statistical institutes (NSIs) to enhance the quality of index numbers. More refined methods can be applied that deal with the dynamics of consumption patterns in a more appropriate way than traditional fixed-basket methods. For instance, multilateral methods can be used to specify monthly weights based on actual sales at product level and new products can be directly included in index calculations (de Haan and van der Grient, 2011; Krsinich, 2014; Chessa, 2016; Chessa et al., 2017; ABS, 2017; Diewert and Fox, 2017; Van Loon and Roels, 2018).

Electronic transaction or scanner data sets contain expenditures and quantities sold of items purchased by consumers at physical or online sales points of a retail chain. The sales data are often aggregated by retailers to weekly sales and are specified by Global Trade Item Number

[1] CPI department, Statistics Netherlands. Email: ag.chessa@cbs.nl. The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

(GTIN, barcode) of each individual item.[2] Transaction data sets also contain characteristics, such as brand and package volume, of the items sold. While traditional price collection methods typically record prices of several tens of products in shops, electronic transaction data sets may contain several tens of thousands of items at the GTIN level for a single retail chain.

GTINs represent the most detailed product level in electronic transaction data sets. Each item has a unique barcode. In principle, this means that NSIs are given a set of tightly defined products. The ratio of monthly expenditure and quantity sold yields a transaction price, which can be followed for each product/GTIN from month to month. However, items may be removed from the market and reintroduced with a modified packaging, for instance, in order to fit within a retailer's new product line. Quality characteristics of such 'relaunch' items may remain the same, but the barcodes may change after reintroduction and also the prices compared with the prices under the previous GTINs. The barcodes of the old and new, reintroduced items have to be linked in order to capture price changes under such relaunches.

Typical market segments that are characterised by relaunches are pharmacy products, clothing and electronics. Rates of item churn may reach such high levels that each year new product lines are introduced that replace the former ones. The GTIN level is not appropriate as product level in such situations. GTINs of relaunch items have to be linked, which means that broader product concepts are needed.

Linking new to old GTINs can be handled manually for small samples of items. However, this becomes infeasible when NSIs aim at processing all GTINs each month, or at least those GTINs that account for a high percentage of total expenditure. To this date, a method for linking GTINs of relaunch items or, in more general terms, for defining products, that is both broadly applicable and efficient does not seem to exist. Recent studies from different NSIs have shown a need for such a method (Bilius et al., 2018; Hov and Johannesen, 2018; Keating and Murtagh, 2018). To find a generic and efficient method is the objective of the present paper.

Section 2 shows several examples of product types with different dynamics of GTINs that enter and leave an assortment. This section gives a first, rough impression of the possible impact of different choices with regard to product definition on a price index. The central element of this paper, the method MARS for defining products, is described in Section 3. MARS has been applied to different types of products, with different rates of churn: COICOP 01 items, clothing, pharmacy products and electronics. Some results are shown in Section 4.

Information about item characteristics in transaction data sets is often sparse. Whether the metadata are sufficient is an important but difficult question to answer. Special attention to this topic will be given in Section 5. A key aspect is obviously how MARS could be applied in production. A global outline of its application and points of attention are presented in Section 6. Section 7 concludes and identifies points of further research.

2 Assortment dynamics

As was mentioned in the previous section, certain types of product are affected more by item relaunches than other product types. This section gives several examples with different rates of churn. Combining GTINs based on common characteristics is one possible way of linking GTINs of relaunch items. This section also gives a first impression of the impact of linking versus not linking on a price index, which serves to highlight the importance of the problem of product definition.

[2] The term 'item' is equivalent to GTIN in this paper.

The focus in this paper is primarily on transaction data. Four product types from data sets of four different Dutch retail chains are considered: milk, cheese and eggs of a supermarket chain, infant garments of a department store chain, hair care of a pharmacy chain and televisions of an electronics retailer. About four years of data are used for the first two product types and three years of data for the other two product types.

The dynamics of products leaving and entering an assortment over time can be measured in different ways. Chessa et al. (2017) quantify the percentages of existing, leaving and entering products in each month with respect to the preceding month for different types of product. A similar measure is used in this section, which is modified on two points:

- The comparison or base month is fixed, and is taken to be the first month of a 13-month time window (December of the previous year);[3]
- The share of 'existing products' on the total number of products sold in a month is taken as a measure of assortment dynamics. Existing products are products that are sold both in the base month and in the current month.

These two choices can be translated into the following formal notation. Quantities of an item i sold in month t are denoted by $q_{i,t}$ and G_t is the set of items sold in month t. The comparison or base month is denoted as month 0. Let $G_{0,t}$ be the set of items/GTINs that are sold both in the base month and in (current) month t. The measure of dynamics proposed in this paper does not merely count numbers of products, but quantifies the numbers sold. This choice expresses the extent of churn more appropriately. For instance, a high number of new products with low sales is not necessarily problematic, in the sense that linking old and new GTINs hardly affects a price index in such situations because of their low expenditure shares.

The proposed measure of dynamics is defined as follows at GTIN level:

$$\frac{\sum_{i \in G_{0,t}} q_{i,t}}{\sum_{i \in G_t} q_{i,t}}. \qquad (1)$$

The numerator is equal to the number of items sold in month t that were also sold in the base month, and the denominator is equal to the total number of items sold in month t. It is easy to see that this measure is equal to 1 when there are no new items in month t, while it decreases when the sales quantities of new items increase. High values of the ratio therefore mean that the existing items prevail in the sales; in other words, the items sold in month t match well with the items sold in the base month. For this reason, expression (1) will be referred to as the 'degree of product match' in month t.

Other choices could be made for different aspects, such as a different base period and to include disappearing products as well. Admittedly, a better choice for the base period would be to take the whole previous year instead of a single month. Products may leave temporarily. A longer period would therefore be recommendable for seasonal products. But for non-seasonal items we do not expect significant differences, as was also noted in Chessa (2018, pp. 23-25).

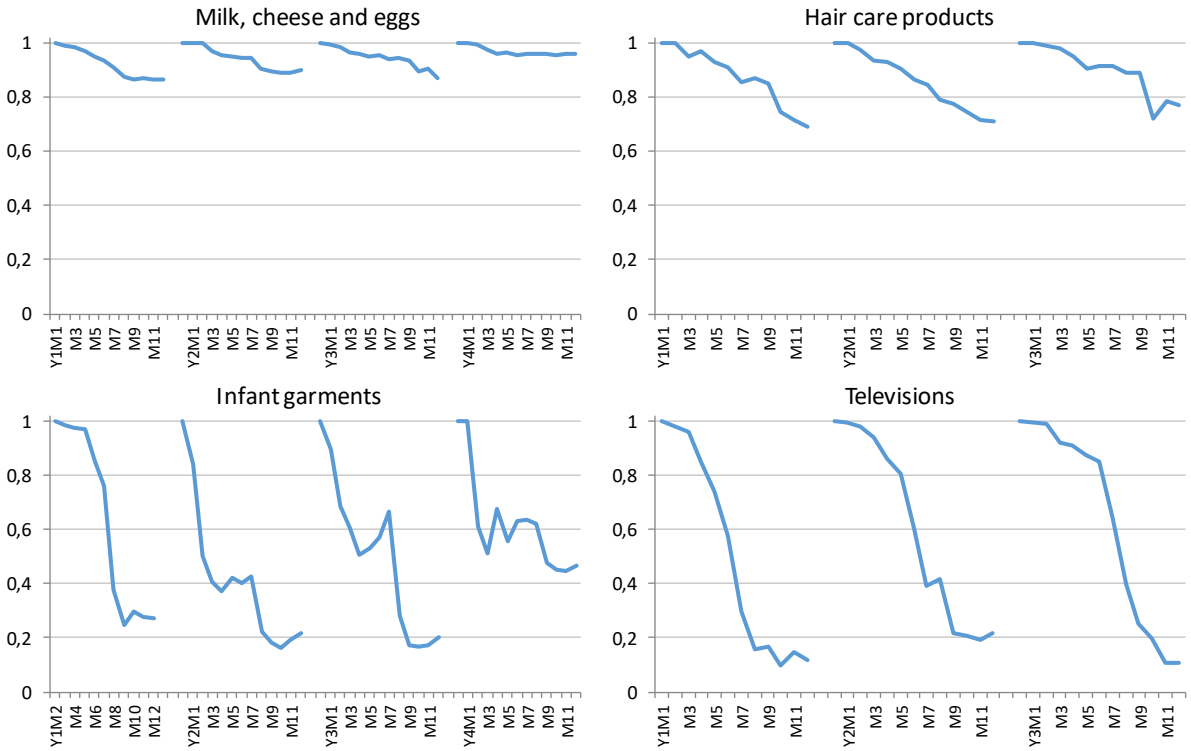
In the above definition of product match, adding the quantities sold for disappearing items that were still sold in the base month to the denominator of (1) would not influence the results when comparing different stratification schemes, since the denominator would be the same in

[3] For time windows shorter than 13 months (e.g. because December is not the first month in the data), the base month is obviously the first month for which data are available.

every scheme. In other definitions of product match, such as the version based on numbers of products in Chessa (2018), disappearing items have a very small effect.

Examples of product match are shown in Figure 1 for the four types of product mentioned above. The graphs clearly illustrate how strong product match can vary across different types of product. Rates of churn are relatively low for milk, cheese and eggs. Most items that were sold in the base month are still sold at the end of a year, as the existing items dominate the sales. The shares of existing items in the sales quantities for hair care drop to about 70 per cent at the end of each year, so that new items account for about 30 per cent. In this case, we are less confident of choosing GTINs as products. Relaunches are known to occur in this market segment (Chessa, 2013).

Figure 1. Degree of product match for four product categories.



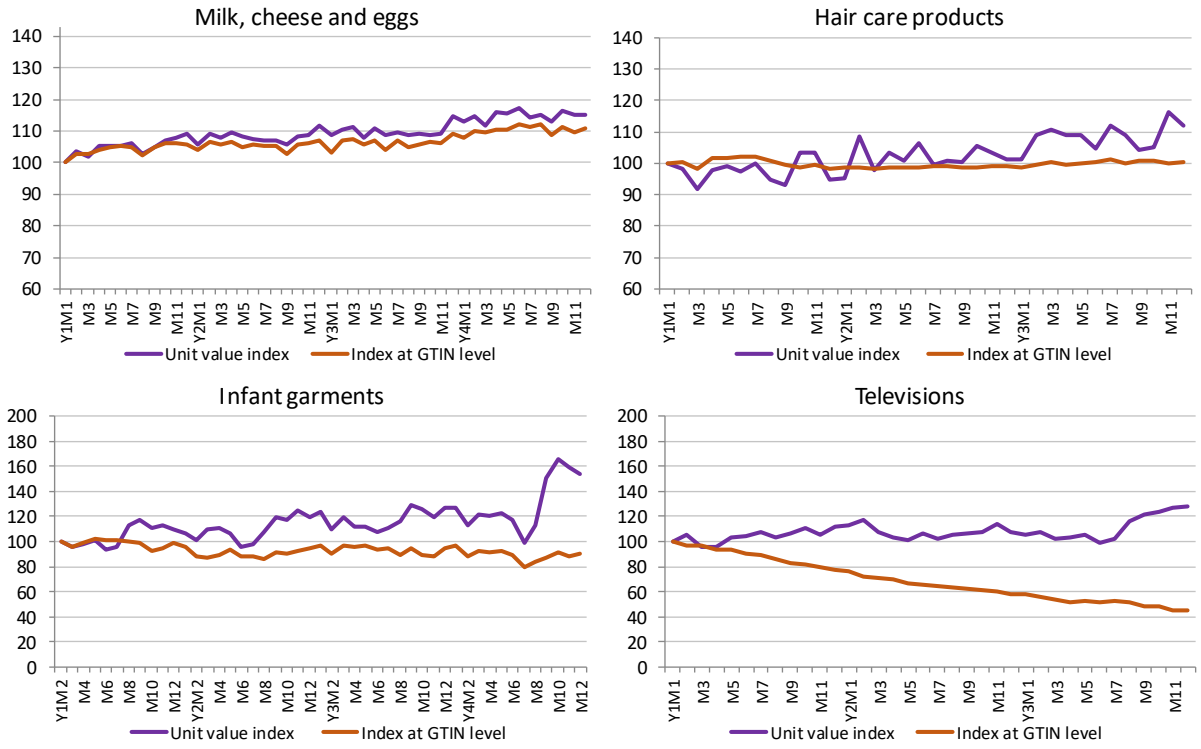
The other two product types, televisions and infant garments, show extremely low product match values at the end of each year. Item turnover is very high in the course of a year. Almost entirely new product lines are introduced each year, which practically replace the previous one. Infant garments are influenced by fashion trends, which may offer an explanation for the high turnover rates and the rapidly decreasing product match.

Traditional bilateral matched model approaches are hard to use at GTIN level under such circumstances, because of the poor continuity of GTINs over time. This also holds for more sophisticated methods like multilateral methods, since these methods are not able to identify price changes either when relaunches occur. A separate method is developed for handling this problem, which is the purpose of MARS.

Figure 2 shows price indices when each GTIN is taken as a separate product. These indices are calculated with the 'QU method'; more specifically, the multilateral Geary-Khamis method

applied to the time domain (Chessa, 2016).[4] Making no distinction among GTINs represents the other extreme of the stratification spectrum. In that case, all GTINs would be considered of the same quality. Expenditures and sales quantities are summed over all GTINs within a product category. The ratio yields a weighted average price, known as 'unit value' (ILO et al., 2004). The unit value indices are shown as well in the figure below.

Figure 2. Price indices at GTIN level and unit value indices for the four product categories (1st month = 100).



The graphs show large differences between the two indices, especially for televisions. Existing models usually decrease in price after being introduced. New models are often more expensive than the preceding models. The index at GTIN level does not consider any of these higher prices as price changes with respect to older models. This explains why this index decreases. Higher prices of new models are seen as price increases from a unit value perspective, which explains the behaviour of the unit value index. New products may have higher prices because they are relaunches, but also because they differ in terms of quality. Shifts in buying behaviour towards more expensive, higher quality products are also considered as price increases by the unit value. Similar explanations for the differences can be given for the other three product groups.

Although the indices represent two extreme cases of product stratification, the differences nevertheless make clear that product definition may have a big impact on a price index. We are therefore dealing with an influential choice aspect, which requires a solid and efficient method in order to decide and justify eventual choices.

[4] The term 'QU method' was introduced in previous work (Chessa, 2016), which in fact denotes a family of 'quality adjusted' or 'generalised' unit value methods (Auer, 2014). The term 'QU method' has become synonymous with the Geary-Khamis method in CPI applications over time.

3 Product definition with MARS

3.1 Preliminary remarks and terminology

We start this section by introducing some terminology. The term *item* was already introduced at the beginning of this paper, which is used here interchangeably with GTIN. Different items, that is, with different barcodes but not necessarily with regard to quality, may have to be linked when relaunches occur. The more generic term *product* is introduced to denote a set of one or more GTINs, which share certain quality characteristics. A product can therefore also be viewed as a combination of characteristics. The latter term is used here as a specific 'value' of the more generic term *attribute* or *variable*. For example, screen size is an attribute of televisions, and 42 inch is a specific characteristic.

Items are subdivided into products, and the set of all products forms a *partition* of GTINs. The term *stratification* is used in this paper as well for partition, although 'partition' is the formal mathematical term. In a partition, each item is assigned to exactly one product, such that products are pairwise disjoint (different products do not have items in common). GTINs may also be chosen as different products, as was illustrated in the previous section, so the set of GTINs is one of the possible partitions. In the examples with unit values there is only one product, which contains all GTINs of a product category.

GTINs should provide a suitable level of stratification for product categories with high degrees of product match. GTINs also represent the most detailed level of product homogeneity in transaction (scanner) data sets. GTINs can therefore be considered as a serious stratification candidate for milk, cheese and eggs in Section 2. We will return to this when the results of the method MARS are presented in Section 4. How to select a suitable level of stratification for product categories with low degrees of product match at GTIN level is less obvious. Broader defined products will increase product match, but homogeneity may be affected.

There are many ways of partitioning GTINs into products. Available attributes of GTINs can be used for this purpose. Different selections and combinations of attributes give rise to different partitions. Each product in a partition contains GTINs with the same characteristics. For example, the attributes brand, screen size and screen type yield one partition of the set of televisions. The characteristics 'Samsung', 'between 51 and 59 inch' and 'Ultra HD' define a specific product of this partition. The GTINs within each product are then considered to be of the same or comparable quality. An attribute may be selected or not. This means that a set of GTINs can be partitioned into 2^n ways for n attributes, to which the partition with GTINs as distinct products can be added (thus yielding $2^n + 1$ partitions).

Attributes provide an efficient way of partitioning a set of GTINs. However, merely using attributes only generates a subset of the entire set of possible partitions. GTINs can be combined also without making use of attributes, although enumerating all possible combinations will be impossible in practice, given the large number of GTINs in transaction data sets.[5] We will first focus on partitioning by either GTINs or attributes (Section 4.1). Hybrid stratification schemes, which combine 'stable' GTINs as separate products with broader defined products that link new to exiting GTINs, are studied in Section 4.2. An important feature is that the number of attributes in transaction data is often limited, which raises the question whether the available set is sufficient and how this could be verified. Section 5 makes several suggestions with regard to this difficult problem.

[5] The total number of partitions of a set with n elements (e.g. GTINs) is equal to the *Bell number*. For example, a set with 3 GTINs has 5 possible partitions, and a set with 6 GTINs can already be partitioned in 203 different ways. For more details, see https://en.wikipedia.org/wiki/Partition_of_a_set .

3.2 Formalisation of MARS

From the introduction to this section it may be clear that numerous ways of partitioning a set of GTINs exist, each of which may have a different impact on product match and homogeneity. The aim is to find a method that balances these two properties in an optimal way in some sense. Measures of product match and homogeneity will be set up in order to operationalise this idea. The two measures are eventually combined, which allows to evaluate and rank GTIN partitions.

First, some notation is introduced in addition to the notation already used in Section 2. We denote a partition by K and use k to indicate an element of a partition, which we have called a 'product'. In theory, different partitions K_t can be defined each month t, for example by changing the set of attributes. However, such dynamic cases are very complex and are probably not well-understood yet with regard to price index calculation. This study therefore deals with situations where products, once defined, are kept fixed for some period. A time window of 13 months would be in line with CPI convention, as product definitions are typically reviewed and possibly revised at the end of each year. We can therefore drop the time dimension from the notation for partitions.

We denote the degree of product match for a partition K in month t with respect to the base month by μ_t^K and the degree of product homogeneity by R_t^K. It is useful to think about desirable properties for these two measures. As this study is a novel field in the processing of large electronic data sets and index calculation, a first attempt towards defining properties is given below.

Property 1. For two partitions K and K' such that K' is a refinement of K, so every element of K' is a subset of an element of K, the degree of product match of the refinement K' cannot be larger than the product match of K, that is: $\mu_t^{K'} \leq \mu_t^K$.

Property 2. For two partitions K and K' such that K' is a refinement of K, the refinement K' is at least as homogeneous as K. In formal terms, we have $R_t^{K'} \geq R_t^K$ for all t.

It is reasonable to expect that broader defined products will increase product match, or at least stay the same, while the opposite holds for homogeneity. This is in fact what the two properties say. Measures that satisfy both properties are defined below.

Product match

Expression (1) applies to GTINs as separate products, so a generalisation is needed. We introduce $K_{0,t} \subseteq K$ for the set of products that are sold both in base month 0 and a second month t, with $t \geq 0$. In practical applications, month 0 will usually be December of the previous year and t a month in a 13-month window that runs until December of the present year. Let q_t^k denote the number of items sold for product k in month t. The degree of product match of partition K in month t is defined as follows:

$$\mu_t^K = \frac{\sum_{k \in K_{0,t}} q_t^k}{\sum_{i \in G_t} q_{i,t}}. \tag{2}$$

It is easily verified that this measure satisfies Property 1. Note that $0 \leq \mu_t^K \leq 1$ for all K.

Product homogeneity

By a set of homogeneous products we usually intend that the products are of the same quality. Finding a measure for the homogeneity or heterogeneity of a set of products boils down to finding a method that expresses their quality differences. Hedonics is an approach that comes to mind when reflecting about this complex problem. Although this class of methods has been broadly studied, it is certainly not without limitations (Chessa et al., 2017).

Index methods usually express differences among products in terms of prices. This paper also takes item prices to set up a measure of product homogeneity. Alternative approaches have not yet been studied. We introduce the following notation for prices. Let the price of item i in month t be denoted by $p_{i,t}$. For a product $k \in K$, let G_t^k denote the set of items in product k. Let $\bar{p}_t^k$ denote the unit value for product k in month t, that is:

$$\bar{p}_t^k = \frac{\sum_{i \in G_t^k} p_{i,t} q_{i,t}}{\sum_{i \in G_t^k} q_{i,t}}, \tag{3}$$

where the denominator is equal to the previously introduced quantity q_t^k. The unit value over all items in month t is denoted by $\bar{p}_t$:

$$\bar{p}_t = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t}}{\sum_{i \in G_t} q_{i,t}}. \tag{4}$$

MARS uses the proportion of explained variance in product prices, relative to the total variance in item prices, as a measure of product homogeneity. The contribution of each product or item is weighted by the quantities sold. This yields the following weighted R squared measure:

$$R_t^K = \frac{\sum_{k \in K} q_t^k \left(\bar{p}_t^k - \bar{p}_t\right)^2}{\sum_{i \in G_t} q_{i,t} \left(p_{i,t} - \bar{p}_t\right)^2}. \tag{5}$$

More precisely, this is in fact a measure of heterogeneity between products. The complementary measure is the price variance of GTINs within products. We want this measure to be as low as possible, and, as a consequence, the explained variance as high as possible. Higher values of R_t^K thus denote better homogeneity.

Note that $R_t^K = 0$ when all items are combined into one product and $R_t^K = 1$ when each item is a separate product. Expression (5) satisfies Property 2. Together with the previously mentioned properties, this implies that $0 \leq R_t^K \leq 1$ for all K.

An alternative homogeneity measure could be defined by using coefficients of variation of the products in a partition. These statistics are commonly used in price statistics, not only in the CPI for data analyses, but also in PPPs. However, it can be shown that coefficients of variation do not satisfy Property 2 in general.[6]

MARS: Combining product match and homogeneity

The method MARS aims at evaluating and ranking item partitions. To this end, the measures of homogeneity and product match will be combined. Some guidance on suitable functions could be provided by considering properties of rank orderings of partitions.

[6] A counterexample with three items and quantity weighted coefficients of variation is sufficient to show this.

Transaction data sets, but also other data sources like web scraped data or traditionally collected data, are usually incomplete. For example, the available product variables are a subset of attributes that characterise items, and data sets are usually delivered by retailers in some aggregate form. These considerations motivate the following property.

Property 3. For any two sets of partitions $\mathcal{K}'$ and $\mathcal{K}$, with $\mathcal{K}' \subset \mathcal{K}$, the ordering of partitions on $\mathcal{K}'$ is equal to the partial ordering of the same partitions obtained on the larger set $\mathcal{K}$.

This property tells us that the ranking of partitions should not be affected by the fact that we are usually dealing with subsets of data in practice. And if we would possess all imaginable data, the original ordering of partitions should stay the same. Property 3 has direct implications for the form of the combined measure of homogeneity and product match. For instance, an arithmetic mean of (2) and (5) does not satisfy Property 3. A multiplicative form does satisfy this property, which is the choice made for the method MARS. In this paper, R squared and degree of product match are combined as follows in every month t:

$$M_t^K = R_t^K \mu_t^K. \tag{6}$$

A multiplicative function also has the characteristic that partitions with either low values for R squared or degree of product match will be suppressed.

It will be clear that $0 \leq M_t^K \leq 1$, since expressions (2) and (5) also have this property. Expression (6) allows us to evaluate and rank item partitions, such that the partition with the highest value of M_t^K is preferred. MARS yields values in every month t, so, in theory, the ranking of partitions may differ from month to month. The values of MARS in different months have to be combined in some way in order to produce one ranking. Different methods can be thought of, which will be described and compared in Section 4.

One approach to overcome this is to use a price index to deflate prices and then combine the deflated prices of each product over all months. However, we prefer to stick to the approach proposed in this paper, since the addition of a price index would make the method more complex, computationally more intensive and also dependent on index method. The method presented in this paper can be combined with any index method, which is a big advantage since different index methods are normally used in the CPI for different forms of price collection.

A separate remark is made for the partition where all items are combined into one product (unit value case). The multiplicative form of measure (6) implies that $M_t^K = 0$ for all t, which means that the single product partition will always be rejected. If this is found to be a limitation, then a simple remedy could be to increase both the numerator and the denominator of (5) by some constant, say 1. This yields a monotonic transformation of R squared, which therefore still satisfies Property 2, takes values in (0, 1] and preserves the value 1 for the GTIN level. Also Property 3 is still satisfied. However, in practice we do not expect that we have to use modified measures.

Example

We illustrate MARS with an example with a small number of items and only one month of data. Consider three GTINs, say A, B and C. The prices and quantities of the GTINs in some month are given in Table 1. The 'status' of each GTIN is also given: GTIN A was sold in the base month and is still sold but is about to leave, GTIN B is new and GTIN C is a regularly sold item. GTIN B could be seen as a relaunch of GTIN A.

Table 1. Prices, quantities and status of the three GTINs.

GTIN	Price	Quantity	Status
A	2.00	1	Exiting
B	4.00	20	New
C	2.00	40	Sold in both months, not exiting

We could also specify attributes in order to construct partitions. But, for simplicity, attributes and characteristics are excluded from this example since the number of GTINs is very small. Three GTINs can be partitioned in five ways:

- One partition with each GTIN as a separate product, which we denote as A-B-C;
- Three partitions with two products each, which we denote as AB-C, A-BC and AC-B, where, for example, AB-C means that GTINs A and B are combined into one product (AB), while GTIN C is a separate product;
- One partition with one product (ABC), which combines all GTINs.

The results are shown in Table 2. A clear preference emerges for partition AB-C, in which exiting GTIN A is linked to new GTIN B, while 'persisting' GTIN C is treated as a separate product. This partition maximises the degree of product match, while making a minor concession in terms of homogeneity. We can thus say that MARS has picked up the relaunch. Product match is also maximised by linking new GTIN B to C, but this partition (A-BC) highly affects homogeneity.

Table 2. Results of MARS for the five partitions.

Partition	R squared	Product match	Combined
A-B-C	1	0.672	0.672
AB-C	0.929	1	0.929
A-BC	0.008	1	0.008
AC-B	1	0.672	0.672
ABC	0	1	0

4 Results for different product types

4.1 Partitioning by GTINs or attributes

This section presents the results of applying MARS to the data of the four product categories that were introduced in Section 2. The method is applied to each of the three or four years of data. The product variables that are available in the four transaction data sets are used to set up partitions, which are evaluated and ranked with MARS, including the partition in which each GTIN is a separate product. The available product variables are shown in Table 3.

Several remarks can be made about the variables. GTIN classifiers are subdivisions of GTINs as used by retailers. This information is provided to Statistics Netherlands and is used to facilitate the mapping from GTINs to COICOP in the CPI. However, the most detailed classifiers could also be used as additional attributes, as is done in this study. For example, the broadest of the two classifiers of milk, cheese and eggs contains seven classes. The second classifier is a

further refinement. For example, 'dairy beverages' (first classifier) is subdivided into seven classes (e.g. milk, buttermilk). The classifiers for hair care make a distinction between conditioners and shampoo, and the most detailed classifier mainly by hair type (normal, dry, damaged, anti-dandruff).

Table 3. Product variables in the four transaction data sets.

Product category	**Variables/attributes**
Milk, cheese, eggs	Brand, package volume, 2 GTIN classifiers
Infant garments	Type of garment, volume (#items), fabric, sleeve length, colour, fit, size
Hair care	Brand, package volume, 2 GTIN classifiers
Televisions	Brand, screen size, screen type

Apart from package volume, the other attributes are categorical variables. This also holds for screen size (televisions), which is expressed as a range (e.g. from 28 to 32 inch). Screen type in fact means display technology (e.g. OLED, Ultra HD). Because of the level of detail used by the retailer to specify colours for clothing, we decided to compress colours to three classes (white, black and coloured). The other attributes were used as specified by the retailers. The example with colour shows that different choices can also be made with regard to how the specified characteristics (i.e. the different colours) are used. Using them as specified or compressing the range of colours further increases the number of partitions. This illustrates again how complex the problem of product definition is from a combinatorial perspective.

Partitions have been set up by using the product variables in Table 3. These partitions, together with the partition in which every GTIN represents a separate product, are evaluated and ranked with MARS. The method MARS yields a score in each month for every partition. The monthly scores are combined into a single score, which is eventually used to rank the partitions. In this section, the MARS scores of the last three months are taken to compute an average score for each partition. The idea behind this choice is that the effects of churn become more apparent towards the end of a year, as the products that are sold in the base month will typically dominate sales in the first months of a year. The results are shown in Table 4. An obvious question is whether the rankings of the partitions will often change when the three-month period is extended. We will come back to this later in this section.

Table 4. Partitions selected by MARS for the four product categories in each year.

Product category	**Year 1**	**Year 2**	**Year 3**	**Year 4**
Milk, cheese, eggs	GTINs	GTINs	GTINs	GTINs
Infant garments	Type, volume, sleeve length	Type, volume, colour, fit	All attributes	All attributes, except fabric
Hair care	All attributes	Brand, volume	Brand, volume	
Televisions	Screen size, screen type	Screen size, screen type, brand	Screen size, screen type, brand	

These results invite us to make a number of remarks:

- The partition 'GTINs as products' is only chosen for milk, cheese and eggs, and emerges as the best partition in each year;
- The results tell us that products for clothing, hair care and televisions should be defined by sets of attributes. The three product categories are characterised by moderately to rapidly decreasing degrees of product match at GTIN level (high rates of churn);
- For milk, cheese and eggs, hair care and televisions, the selection of attributes is quite stable over the years. The results for infant garments show more variability.

These findings probably summarise what results could be expected beforehand in a broad sense. GTINs are an appropriate choice for milk, cheese and eggs, and also for other COICOP 01 groups (not shown here) but not for the other product categories because of the higher rates of churn.

GTINs are not selected as products for hair care, infant garments and televisions. The degrees of product match at GTIN level are very low for infant garments and televisions in the second half of each year (Figure 2). New GTINs may thus have to be linked to GTINs that leave the stores. The results show that a small set of attributes is sufficient in most cases. However, the number of attributes for infant garments selected in the third and fourth year is considerably larger than in the first two years. Apparently, the degree of product match increases significantly for tighter defined products in the third and fourth year.

The variability of the set of selected attributes from year to year raises an important question on how MARS could be used in CPI production. In practice, decisions about product definition are made for the next year. The results in this study apply to the situation where the data are known in advance. It is therefore of big practical interest to know to what extent the price indices in this section change when the best partition is used to compute an index with the data of the next year. We will return to this question in Section 6.

Figure 3 shows the monthly MARS scores for various GTIN partitions. Several remarks can be made based on these graphs. First, the GTIN scheme has the highest MARS scores in the first months of each year, which confirms the idea expressed above to shift the focus towards the second half of a year with regard to evaluating and ranking partitions. As was noted previously, this is most apparent for hair care, infant garments and televisions because of the high rates of churn.

Second, the MARS scores in the second half of each year lead to the same ordering of the partitions in almost all months in the graphs of Figure 3. This means that extending the three-month period up to six months in order to calculate an average MARS score for every partition would hardly change the results. The stability of the behaviour of MARS is obviously a pleasant feature.

Third, the partition that results by selecting all attributes for milk, cheese and eggs shows a particular behaviour in the second and fourth year. It reaches much lower MARS scores than the other two partitions in the second half of these two years. Product match turns out to be much lower than at GTIN level. This should not happen, since the GTIN level is a refinement of any partition based on attributes. This means that Property 1 in Section 3.2 is violated. The rapid decreases in product match for the 'all attributes' partition are caused by changes in the names of some GTIN classifiers. Products with changed names are considered as new products, which in reality are existing products with new classifier descriptions. This case shows the practical usefulness of Property 1 and that MARS can also be used as a data monitoring tool.

The price indices that correspond with the highest ranked partitions in each year (Table 4) are compared with the index for GTINs as products and the unit value index. The indices are shown in Figure 4. The index for milk, cheese and eggs at GTIN level is the same as the index for the best partition. The indices for the best partitions for the other product categories show large

Figure 3. MARS scores for different partitions for the four product categories.

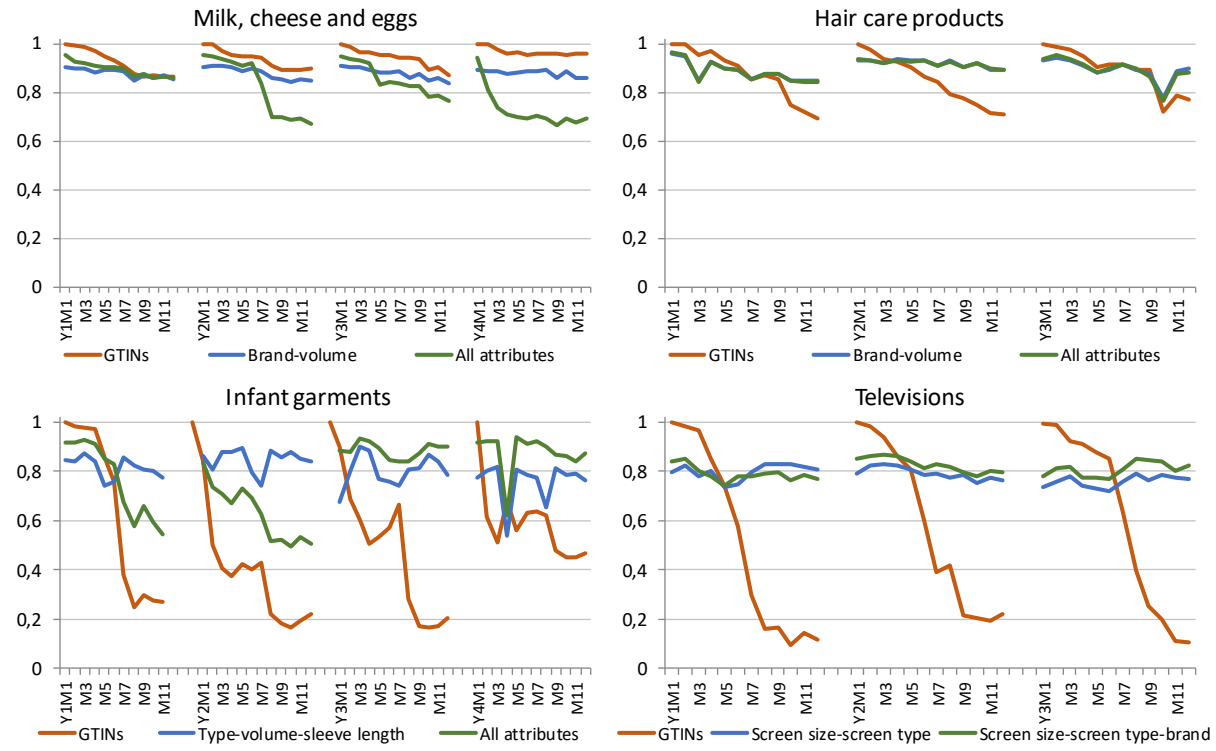
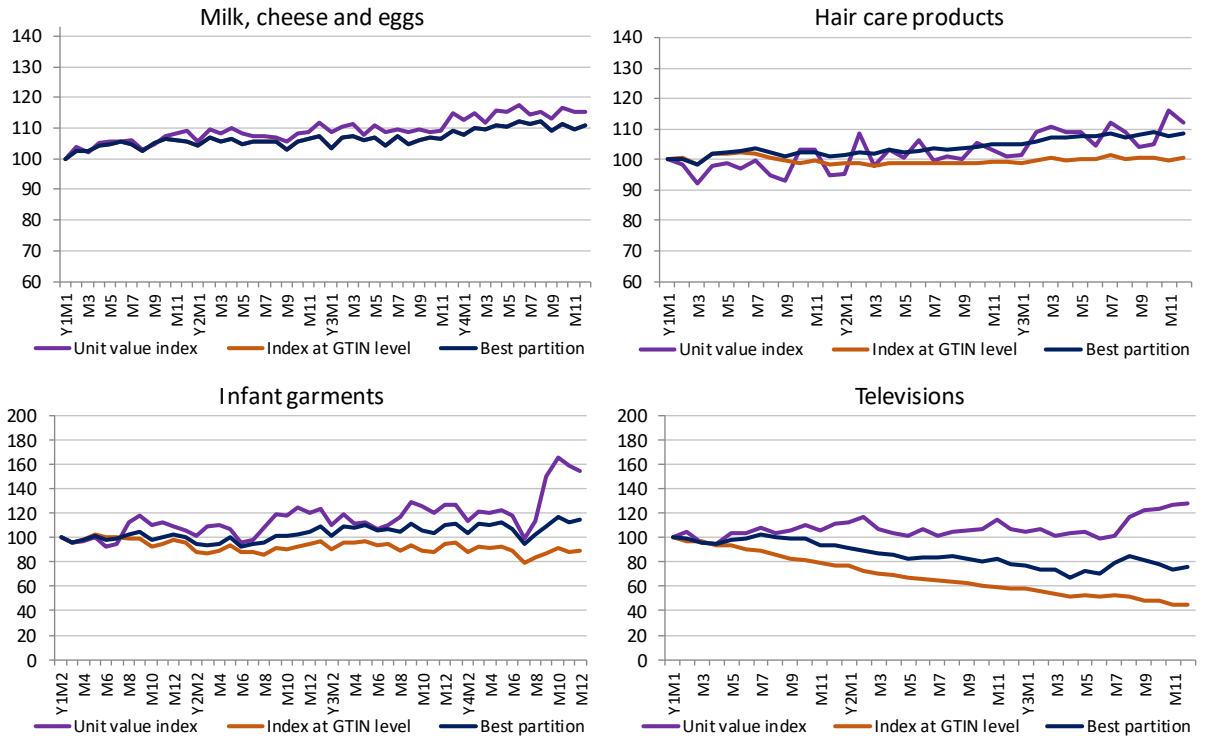


Figure 4. Price indices for the yearly best partitions, compared with the indices at GTIN level and the unit value indices. The first two indices are calculated with the QU method (Geary-Khamis).



differences with the indices at GTIN level and the unit value index. The partitions are based on attributes, which make it possible to pick up price differences between new and exiting GTINs

with the same characteristics. This results in higher indices compared with the indices at GTIN level for hair care, infant garments and televisions.

The indices for the best partitions for infant garments and televisions also differ substantially from the unit value indices. The differences between these two indices for hair care are smaller. Although products are defined as combinations of characteristics, this does not mean that we should expect the corresponding indices to behave like a unit value index. Products may be tightly defined in such partitions, as is indicated by the figures in Table 5. In relation to this, note that MARS allows new and disappearing products to occur also at broader product levels than GTIN.

Table 5. Average number of GTINs per product.

Product category	Year 1	Year 2	Year 3	Year 4
Milk, cheese, eggs	1	1	1	1
Infant garments	20.3	24.6	4.2	4.6
Hair care	3.0	5.4	6.0	
Televisions	21.7	6.6	7.1	

In some cases, the average number of GTINs per product is quite large. This may be caused by the relatively small number of attributes in transaction data sets. On the other hand, it cannot be excluded that broader groups are sufficiently homogeneous. The data for infant garments contain seven attributes. The partition with four attributes reaches high MARS scores and clearly dominates the partition based on all available attributes in the first two years (Figure 3). Intuitively, this could be interpreted as a sign that the clothing data contain enough attributes. But whether attributes are sufficient or not is a hard problem to deal with from a methodological point of view. We will return to this question in Section 5.

4.2 Hybrid stratification schemes

The method MARS is not restricted to stratification by either GTIN or attributes, but takes a broad perspective on partitioning a set of items. This section gives an illustration of this idea. Some results of the previous section could be criticised by hardliners, as the composition of products may still found to be heterogeneous. Such objections could be met by considering 'hybrid' partitions, in which 'stable' GTINs are taken as separate products, while broader products are defined to link new to exiting ('unstable') GTINs by means of common characteristics. This approach is described below and is compared with the results of the previous section. Hybrid stratification schemes for the four product categories are set up as follows:

- A GTIN is defined here as 'stable' if: (1) It is sold in the base period, and (2) its expenditure does not decline by more than a predefined threshold during a year. In this study, this threshold is set at 75 per cent. The GTINs that satisfy these two conditions are defined as separate products;
- GTINs that enter after the base period and existing GTINs that decline by more than 75 per cent are combined in broader products. These GTINs are grouped by using the attributes in Table 4 for hair care, infant garments and televisions. For milk, cheese and eggs this is done by brand and volume (the GTIN classifiers are excluded because of naming issues).

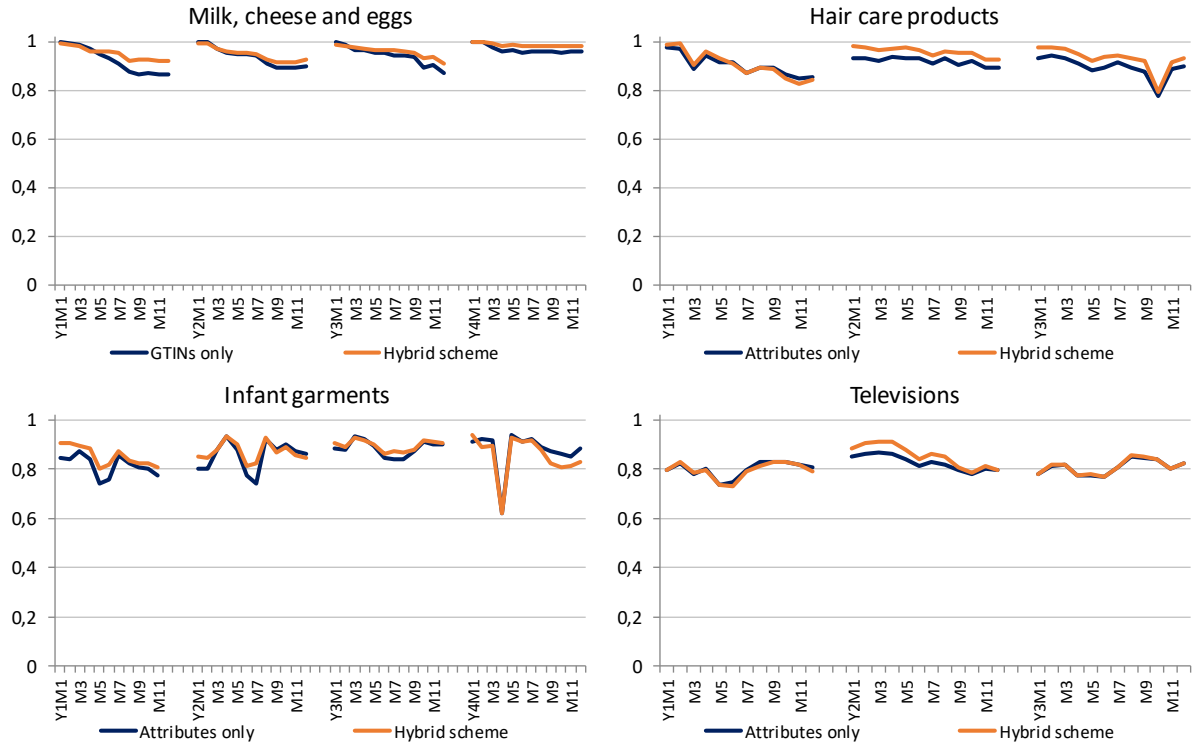
The table below shows that product definitions become considerably tighter than in the three cases where all GTINs are partitioned by using attributes.

Table 6. Average number of GTINs per product for the hybrid schemes.

Product category	Year 1	Year 2	Year 3	Year 4
Milk, cheese, eggs	1.3	1.2	1.2	1.2
Infant garments	4.1	8.0	3.1	1.9
Hair care	1.3	1.6	2.4	
Televisions	10.4	4.8	5.7	

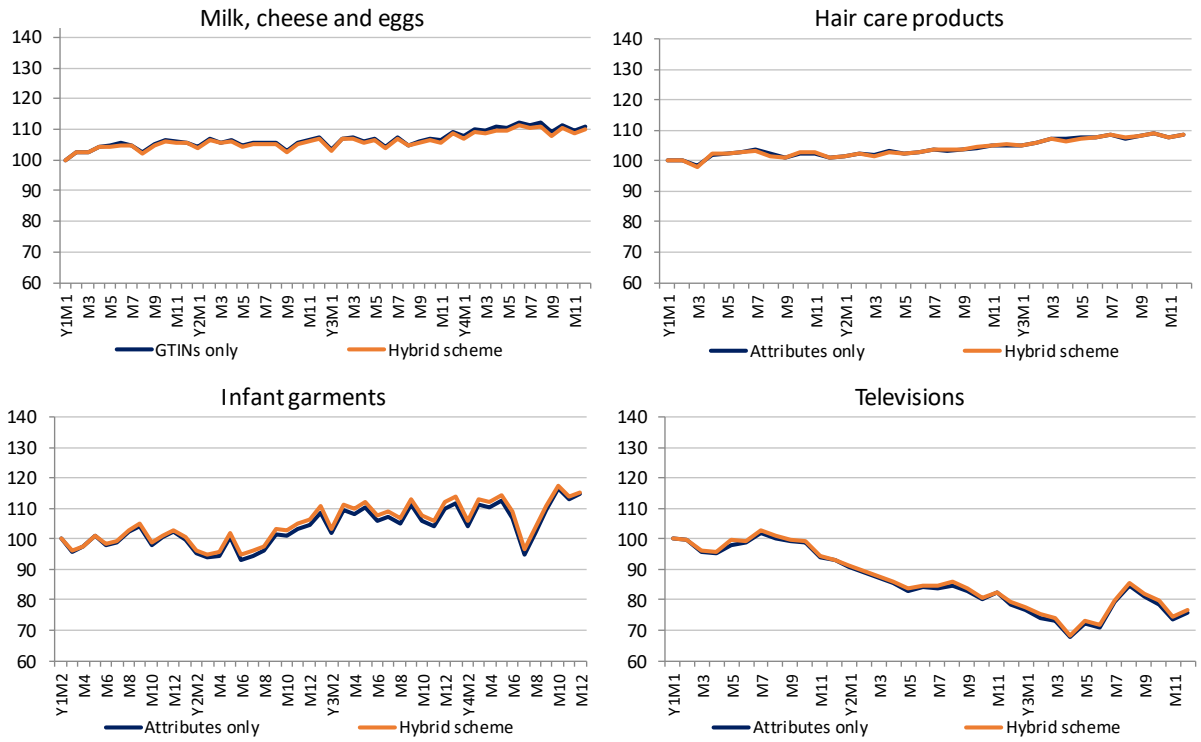
The MARS scores for the hybrid partitions are compared with the 'best' partitions obtained in the previous section with either GTINs or attributes. The results are shown in Figure 5. The hybrid stratification schemes yield higher MARS scores in most years, in particular for milk, cheese and eggs, and for hair care. Hardly any improvement is obtained for infant garments and for televisions, which can be explained by the low degrees of product match at GTIN level.

Figure 5. MARS scores for the hybrid stratification schemes, compared with the best partitions of Figure 3.



The price indices for the two sets of partitions are compared in Figure 6. Differences are very small, which shows that the previously obtained partitions yield robust price indices, at least, given the available information in the transaction data sets. The hybrid schemes may take more time to work with in a production environment, since the stability of GTINs requires frequent monitoring. As a consequence of this, product definitions may also have to be changed more often during a year than when working with either GTINs or attributes. Nevertheless, it seems worth including hybrid schemes as well in practice, as they give additional insight and are also useful as a sensitivity analysis.

Figure 6. Price indices for the hybrid schemes, compared with the indices for the best partitions of Figure 4.



5 Sufficiency of metadata

Transaction data sets usually contain limited product information. An important question is whether the available information is sufficient to define products. The impression that could be obtained from the results in sections 4.1 and 4.2 is that the information about characteristics may be sufficient: products are tightly defined in most cases and some of the available attributes are superfluous in several years.

However, the question remains whether the results in the two previous sections would change if the transaction data sets contained more attributes. Statistical offices have invested a lot in web scraping in recent years, which, in theory, is an excellent way of supplementing transaction data with scraped information about characteristics.[7] Consumer electronics is a typical example of a product category with detailed specifications. Televisions therefore provide an interesting case to verify whether additional attributes would influence the results. Moreover, it is the category with the highest average number of GTINs per product after the additional analysis in Section 4.2 (year 1, Table 6).

Before receiving transaction data, Statistics Netherlands built a scraper for collecting price and product information of different categories from the website of the same electronics retailer included in this study. Five attributes are added from the scraped data: whether or not TVs come with 3D, smart tv or a curved screen, resolution and number of processor cores. MARS was applied to the extended data set. Only attributes are used to form partitions (hybrid schemes are not considered). The highest ranked partitions are shown in Table 7.

[7] This holds subject to certain conditions, such as the continuous availability of an appropriate linking key (e.g. GTIN or retailer's own item code).

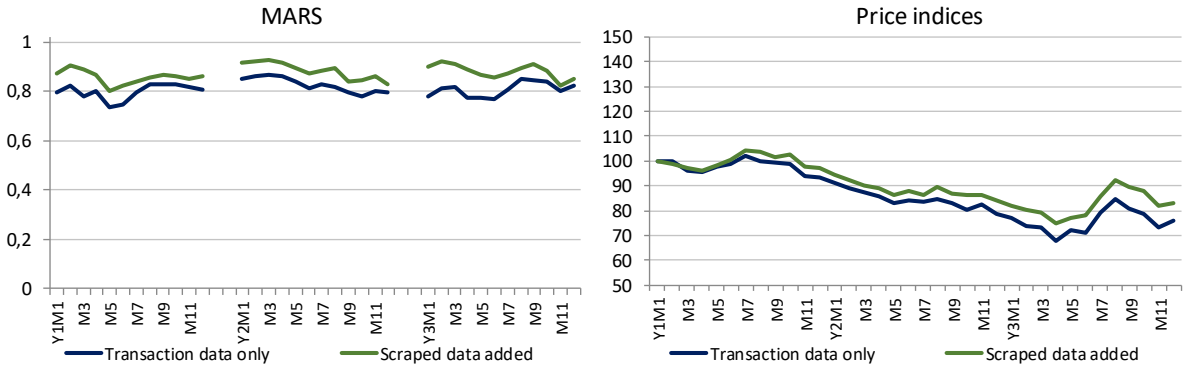
Table 7. Best GTIN partitions for televisions after adding attributes from web scraped data.

Period	Transaction data only	Scraped data added
Year 1	Screen size, screen type	Same + 3D
Year 2	Screen size, screen type, brand	Screen size, screen type, 3D, curved screen
Year 3	Screen size, screen type, brand	Same + 3D

The results show that the selection of attributes has changed each year. The attributes that were selected when only using transaction data still belong to the selection, apart from brand in year 2. The attribute 3D enters the set of attributes in each year after adding web scraped data, and whether or not televisions have a curved screen replaces brand in the second year, together with 3D. The numbers of attributes are still limited to three or four, but the question is to what extent the MARS scores and the price indices change. These results are shown in Figure 7.

Adding attributes from the scraped data clearly improves the previously obtained MARS scores (Figure 5) and also affects the indices that are exclusively based on transaction data. The attribute 3D is added in each year and can therefore be considered as the main driver behind the upward effect on the previously calculated price index for TVs. A closer inspection of the data shows that televisions with 3D were most popular in year 1. The expenditure share for televisions that are equipped with this feature gradually went down in subsequent years. Televisions with 3D are more expensive on average than TVs without 3D. Consequently, not distinguishing TVs by 'with 3D' and 'without 3D' has a downward effect on product prices, which explains why the index based on transaction data is lower than the one after adding scraped data. In fact, the latter adjusts for a shift towards lower quality products, resulting in an upward effect on the index.

Figure 7. MARS and QU indices for televisions, before and after adding web scraped attributes.



The differences between the two indices in Figure 7 are quite large: the differences between the year on year indices are 3.6 percentage points in year 2 and 5.2 in the third year. Scraped data are not always available, which raises the question whether other possibilities can be thought of to assess the sufficiency of attribute information in transaction data.

The addition of web scraped attributes leads to refinements of the partitions obtained with transaction data. If supplementary data sources are not available, then we might think of a number of alternative ways of obtaining refinements and quantifying their impact:

- CPI analysts could be asked to inspect products with the highest expenditure shares and suggest possible refinements;

- Visual inspection and judgement may take a lot of time. Automated alternatives are highly appreciated in an era with less resources and bigger data sets. An approach that may be worth investigating is to introduce an artificial or 'dummy' categorical attribute and define a range of 'values' (characteristics). The idea then is to draw one of these values uniformly at random for every GTIN. Different simulations can be done, each of which may lead to a refined partition that can be evaluated with MARS.

It is clear that the second suggestion generates refined partitions in an artificial way. But if none of the simulated partitions leads to improved values of MARS, it may be concluded that no other attributes are needed to define products. In the opposite case, the partitions that yield higher MARS scores could be inspected by analysts and checked for heterogeneity of refined products and the impact of these partitions on the index can be quantified. The resulting findings can be used to decide whether to retain the product refinements or not, and to decide on additional data collection for future index calculations.

We added a dummy attribute, first with two and next with three values, to the attributes in the transaction data of televisions. 100 simulations were run for both two and three values, and each set of simulated values was combined with the attributes selected by only making use of the transaction data (Table 4). A total of 200 refined partitions were thus evaluated with MARS. The simulations showed that the MARS scores were improved only in the first year, and only in the case with two values of the dummy attribute.

This exercise was repeated by selecting brand or not at random in each simulation. This strategy led to higher MARS scores also in the second and third year. These preliminary results seem to confirm that the attributes contained in the transaction data set of TVs are not sufficient. Different strategies can be thought of, which are worth exploring and developing further.

6 Using MARS in the CPI

Although the method MARS is a result of recent research, the results obtained until now have motivated Statistics Netherlands to start implementing MARS in its CPI in the course of this year. The following reasons support this decision:

- MARS can be applied to a broad range of product types with varying degrees of churn;
- MARS can be seen as a general approach to the problem of product definition: for example, products can be defined as separate GTINs, by a set of attributes or a combination of both;
- MARS does <u>not</u> require price indices, so it can be combined with any price index method. This is important for a number of reasons, which are explained in the next bullets;
- Different index methods are used in the CPI for different data sources, so that MARS does not have to be limited to transaction data;
- The problem of product definition and attribute selection can be automated with MARS to a high degree, which can lead to considerable time savings during monthly production;
- The degree of automation that can be achieved also allows to increase the amount of data processed for index computation in specific cases, such as fixed basket methods;
- MARS lends itself very well for visualisation purposes;
- MARS can also be used as a data monitoring tool, which is an important added feature with the growing use of large electronic data sets. The example with milk, cheese and eggs in Figure 3 illustrates this additional feature.

An important question is how MARS could be applied in a balanced way in production. Different factors should be considered, such as:

- The 'retail dimension'. Should MARS be applied to each retail chain separately or is it possible to combine the data of different retail chains for the same product category?
- The 'product dimension'. Also in this case the question is which level of aggregation would be feasible within resources and time constraints. The elementary aggregate level would probably be preferred from a conceptual point of view.
- The 'time dimension'. This refers to the frequency of maintenance of product definitions during a year.

Retail dimension

The question is whether applying MARS to the data of each retail chain would lead to a workable solution. If we take supermarkets, then according to the Dutch situation we would apply MARS in COICOP 01 a number of times equal to 12 retail chains times 57 COICOP-5 aggregates, that is, at the lowest COICOP level. This shows that it is hard to imagine an application of MARS to data of each retail chain separately. Analysts may want to inspect the results. Our proposal therefore is to apply MARS to the pooled data of all retail chains for the same product category. This does not imply calculating unit values of products by summing over the expenditures and sold quantities of all retailers. Retail chains are kept separate, so that unit values are calculated per product and per retail chain. Retail chain then acts like an additional attribute when applying MARS to data aggregated over retail chains. This choice will lead to the same product definitions for each retail chain for the same product category.

Product dimension

Also in this case, a downside of choosing a detailed level (elementary aggregate, COICOP-6) is that the number of applications of MARS could become large. COICOP-5 would be a better choice in this respect. This should be possible for product categories that share important attributes (e.g. brand, package volume, GTIN classifiers). Food, beverages and clothing could possibly be handled at COICOP-5. Other COICOPs may have to be treated at a lower level of aggregation if product categories are diverse, but this has to be judged from case to case.

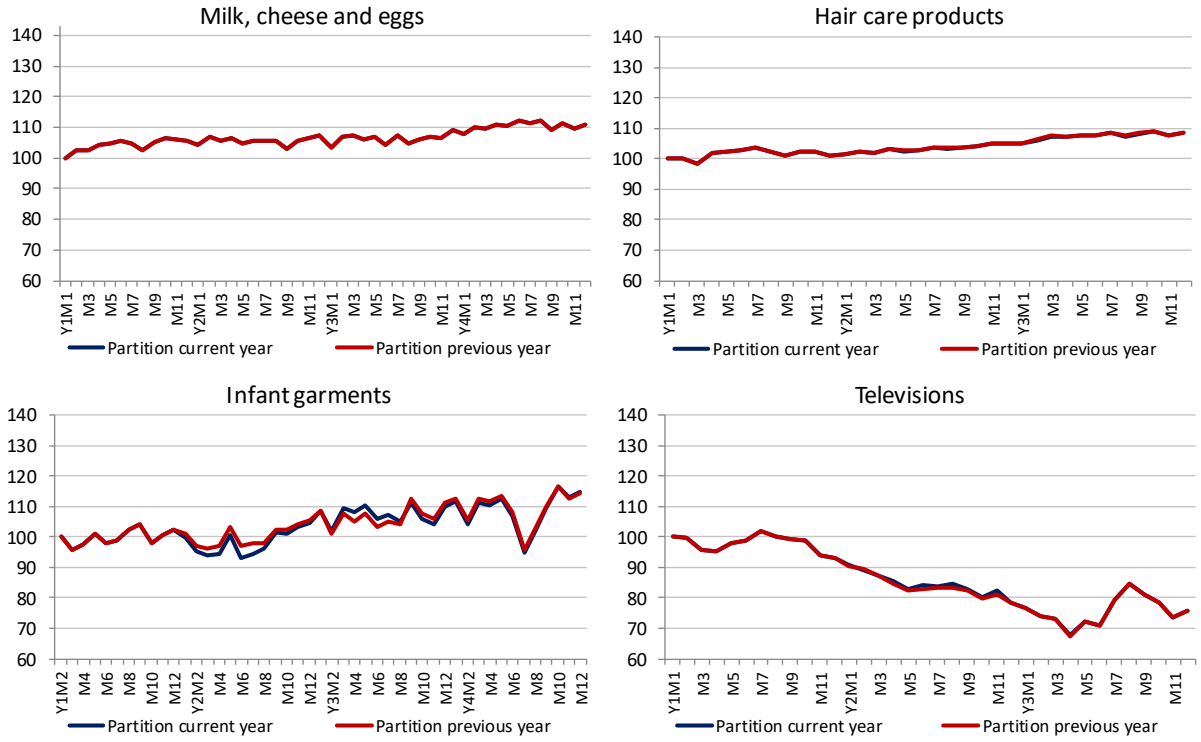
Time dimension

Another important question is how often product definitions should be checked and possibly revised during a year. Will once a year, at the end of a year, be sufficient? Or should more checks be carried out? Once a year is most compatible with current CPI routine and also saves time. A higher frequency has the merit of up to date monitoring and index adjustments. In the latter case the question is how that could be performed. The results in Section 4.1 showed that product definitions may change over time. Infant garments shows changes each year, while product definitions are more stable for the other categories. Product definitions were established by making use of the data of the same year. In practice, decisions have to be taken for the next year. An interesting question therefore is to what extent the indices will change if these are based on the product definitions established in the preceding years. The results are shown in Figure 8.

The price indices obtained with the product definitions of the preceding years are very accurate. The indices for televisions are based only on transaction data. The index would also be accurate in the case where scraped attributes are added. These results suggest that it may be sufficient to check and possibly revise product definitions at the end of a year. Some monitoring,

say after half a year, is nevertheless useful, in particular to detect new products with new attributes.

Figure 8. QU indices when using the partitions of the preceding year.



The aforementioned considerations and the results obtained in the previous sections are useful to set up a generic scheme for applying MARS in practice. Before giving this outline, it should be noted that MARS requires data to be in a readily usable format, with all variables specified in separate fields. Text mining of characteristics should therefore be carried out before applying MARS. The same holds for other pre-processing stages, like standardisation of units of measurement.

For every combination of retail chains and COICOP level (5 or 6), MARS can be applied by following the general procedure below:

1. Set a time window (12 months + base period, December previous year or whole year);
2. Generate partitions (by GTINs, attributes);
3. Calculate R squared, product match and MARS scores for every partition in the 12 months;
4. Rank the partitions by their average MARS scores over the last three months;
5. Analysts may want to inspect products with the largest numbers of GTINs in the highest ranked partition for further refinement. The 'dummy attribute' approach suggested in the previous section could also be used for this purpose.

Retail chain should be included as an additional attribute in step 2 when different retail chains are combined. Finally, it is desirable to work towards a user friendly, interactive tool. Users should have the possibility to preselect certain attributes when they are found to be important (package volume) or to redefine the range of values of an attribute (e.g. comprime a detailed range of colours to a smaller number of classes).

7 Final remarks

Price collection is traditionally carried out by following prices of representative items that satisfy product definitions. This approach is feasible for relatively small product samples. So, historically, the problem of product definition is not new. But it needs a new, more efficient approach when NSIs consider a move towards big electronic data sets and aim at processing a significant part of such data sets or even to their full extent.

Recent research at Statistics Netherlands has led to a method that fits this philosophy. To the author's knowledge, the method MARS is not only new in the field of product definition but also in the sense that it shifts the boundaries of processing big transaction data sets in the CPI, which was mainly concerned with index methods (beside classification methods for linking GTINs to COICOPs). MARS addresses product definition as a combinatorial optimisation problem which does not, and should not by nature, involve index theory. It is concerned with product homogeneity, and therefore also with product prices, but this is the point where the two fields meet and which MARS does not cross.

One of the merits of MARS therefore is that it can be combined with any index method. An extensive list of the merits of the method was given at the beginning of the previous section. This section focuses on identifying points of further research and concludes with the following list:

- MARS has been extensively applied to transaction data. Web scraping is a rapidly growing field, which motivates applying MARS also to web scraped data. This requires finding proxies for quantities sold (e.g. number of scraped product prices per month, see Chessa and Griffioen (2019)). First exercises with scraped data indicate that MARS also works well for this type of data;
- MARS has been applied so far by using all item prices in its R squared measure. An interesting question is whether discount prices should be included;
- The number of product attributes in transaction data sets is usually limited. But in cases where many attributes are available, enumerating all partitions may become problematic (say, when the number of attributes exceeds 10). For such situations, it may be interesting to consider heuristic methods like simulated annealing (Kirkpatrick et al., 1983; Granville et al., 1994);
- MARS could be classified as an unsupervised method. It may be worth studying supervised versions of the method, for instance, by setting up a 'training data set' in which exiting and new GTINs of relaunch items are linked. MARS combines homogeneity and product match measures by calculating in fact a geometric mean. A weighted geometric mean could be used in a supervised version of MARS, where the homogeneity and matching weights are tuned to the target training data.

Acknowledgements

This research was funded by a grant assigned by Eurostat, for which the author wants to express his gratitude. The author also wants to thank colleagues at Statistics Netherlands and of other NSIs for the many useful discussions and feedback.

References

ABS (2017). Making greater use of transactions data to compile the Consumer Price Index. Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, 10-12 May 2017, Eltville am Rhein, Germany.

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

Bilius, Å, Ståhl, O., and Tongur, C. (2018). Coverage bias and the effect of re-launches in scanner data: A coffee index. Poster presented at the *Meeting of the Group of Experts on Consumer Price Indices*, 7-9 May 2018, Geneva, Switzerland.

Chessa, A.G. (2013). Comparing scanner data and survey data for measuring price change of drugstore articles. Paper presented at the *Workshop on Scanner Data for HICP*, 26-27 September 2013, Lisbon, Portugal.

Chessa, A.G. (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, issue 1/2016, 49-69.

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

Chessa, A.G., and Griffioen, R. (2019). Comparing scanner data and web scraped data for consumer price indices. To appear in *Economie et Statistique/Economics and Statistics*. Special issue: *Big Data, Statistics and Economics*, September 2019.

Chessa, A.G., Verburg, J., and Willenborg, L. (2017). A comparison of price index methods for scanner data. Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, 10-12 May 2017, Eltville am Rhein, Germany.

Diewert, W.E., and Fox, K.J. (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.

Granville, V., Krivanek, M., and Rasson, J.-P. (1994). Simulated annealing: A proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16 (6), 652-656.

de Haan, J., and van der Grient, H.A. (2011). Eliminating chain drift in price indices based on scanner data. *Journal of Econometrics*, 161, 36-46.

Hov, K., and Johannessen, R. (2018). Using scanner data for sports equipment. Paper presented at the *Meeting of the Group of Experts on Consumer Price Indices*, 7-9 May 2018, Geneva, Switzerland.

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO Publications.

Keating, J., and Murtagh, M. (2018). Quality adjustment in the Irish CPI. Paper presented at the *Meeting of the Group of Experts on Consumer Price Indices*, 7-9 May 2018, Geneva, Switzerland.

Kirkpatrick, S., Gelatt Jr, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, 220 (4598), 671-680.

Krsinich, F. (2014). The FEWS Index: Fixed Effects with a Window Splice – Non-Revisable Quality-Adjusted Price Indices with No Characteristic Information. Paper presented at the *Meeting of the Group of Experts on Consumer Price Indices*, 26-28 May 2014, Geneva, Switzerland.

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*, 7-9 May 2018, Geneva, Switzerland.