MARS: A method for linking barcodes and stratifying products for price index calculation

**Keywords:** Transaction data, CPI, product homogeneity, GTIN, churn rate.

# Introduction

The increased availability of electronic transaction data 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 sales based weights at the most detailed product level and new products can be directly included in index calculations.

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 a weekly level and are specified by the barcode or Global Trade Item Number (GTIN) of each individual item. 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.

The problem of identifying “suitable” levels of product stratification has to be resolved before applying index methods to calculate price movements from period to period. The key question is what is found to be a suitable level of product stratification and how this notion could be formalised and operationalised. In addition, the size of electronic data sets calls for a method that enables statistical agencies to automate the stratification process to a high degree.

Section 2 contains a summary of a method that has recently been developed at Statistics Netherlands. Section 3 shows some results for product categories with high rates of churn. Section 4 summarises the findings and identifies directions for further research.

# Methods

## Sketch of the approach

As was stated in the introduction, the GTIN level is the most detailed and the most homogeneous product unit in terms of quality characteristics. On the other hand, this level may be too detailed in the sense that price changes associated with item relaunches will be missed. The method proposed in this paper finds a product level by balancing homogeneity and the degree by which products can be followed over time, which is referred to as “degree of product match” in this paper. The latter can be considered as a complementary measure of the rate of churn (the higher/lower the churn rate, the lower/higher the degree of product match).

The method described in Section 2.2 proposes measures for product homogeneity and for the degree of product match. The two measures are then combined into one measure, which yields a score for each partitioning of GTINs into items groups (called “products”). GTINs can be grouped according to a set of common characteristics. Transaction data sets may contain different attributes (or variables) and the question is which ones to select in order to form items groups. Different selections of attributes may lead to different items groups, that is, to different ways in which items can be partitioned. The method in Section 2.2 computes a score for each partition. The partition with the highest score is eventually selected.

Some examples are given in order to illustrate the idea behind the method. According to the method described in Section 2.2, the partition with each GTIN denoting a unique product will maximise the homogeneity measure. The GTIN level will be the ideal partition when there are no new items. When new items do enter an assortment in the course of a year, the method searches an item grouping that does not make substantial concessions to homogeneity while boosting the degree of product match.

## The method MARS

The stratification method operates on a time window, with a length that has to be set by the user (say 13 months, a calendar year plus December of the previous year, a common choice in the CPI). For each GTIN partition, the method calculates the following measures for homogeneity and degree of product match in every month of the time window:

* R squared of monthly product prices as a measure for product homogeneity;
* The share of the products from the first month of the window in the total sales of subsequent months (degree of product match).

The degree of product match operates as an adjustment factor on R squared, which explains the abbreviation MARS chosen for the method (Match Adjusted R Squared). Different choices can be made for the homogeneity and degree of product match measures, but this paper focuses on one set of choices for MARS. Suggestions and comparisons for other choices can be found in a more extended paper [1].

The version of MARS considered in this paper can be formalised as follows. Let $K$ denote a specific partition of items/GTINs and let $k$ denote an element of $K$, that is, a group of items with common characteristics, which will be referred to as a “product”. Let $\overbar{p}\_{t}^{k}$ denote the price of product $k\in K$ in month $t$ and let the quantities sold be denoted by $q\_{t}^{k}$. The product price is obtained by dividing the expenditures by quantities, both summed over the items belonging to $k$. Item prices and quantities are denoted as $p\_{i,t}$ and $q\_{i,t}$ and the set of items sold in month $t$ by $G\_{t}$. The weighted average price over all items in the set $G\_{t}$ is denoted by $\overbar{p}\_{t}$. R squared values $R\_{t}^{K}$ are computed each month $t$ for every partition $K$ as follows:

|  |  |
| --- | --- |
| $$R\_{t}^{K}=\frac{\sum\_{k\in K}^{}q\_{t}^{k}\left(\overbar{p}\_{t}^{k}-\overbar{p}\_{t}\right)^{2}}{\sum\_{i\in G\_{t}}^{}q\_{i,t}\left(p\_{i,t}-\overbar{p}\_{t}\right)^{2}}.$$ | (1) |

Let $K\_{0,t}$ denote the set of products in partition $K$ that are sold both in month $t$ and the base month 0 (i.e., the first month of the time window). The degree of product match $μ\_{t}^{K}$ of partition $K$ in month $t$ is defined as the share of the quantities of base month products in the total quantities sold in month $t$:

|  |  |
| --- | --- |
| $$μ\_{t}^{K}=\frac{\sum\_{k\in K\_{0,t}}^{}q\_{t}^{k}}{\sum\_{i\in G\_{t}}^{}q\_{i,t}}.$$ | (2) |

MARS combines the two measures by taking the product of (1) and (2):

|  |  |
| --- | --- |
| $$M\_{t}^{K}=R\_{t}^{K}μ\_{t}^{K}.$$ | (3) |

This measure is evaluated for every item partition that one wants to consider. The three measures (1)-(3) take values between 0 and 1. Taking GTINs as products implies that $R\_{t}^{K}=1$ in every month $t$. If there are no new items in month $t$, that is, items that were not sold in month 0, then $μ\_{t}^{K}=1$. The entrance of new items will decrease the degree of product match if new items end up forming new products under an item partition.

# Results

The method MARS has been applied to different types of products, ranging from grocery to pharmacy products, electronics and clothing. Grocery items are quite stable, so that their prices can be followed over a fairly long period of time. MARS shows that the GTIN level is the most suitable level of product stratification in most of these cases. It is interesting to look at the results of MARS for product types with high rates of churn and relaunches. Figure 1 shows some results for televisions and hair care products.



Figure . Results of MARS (formula (3)) for three GTIN partitions for televisions and hair care products in three years.

Figure 1 clearly shows that the rate of churn for televisions is extremely high: MARS reaches very low values as a consequence of the same, very low values for degree of product match (R squared is equal to 1 for GTINs as products). The share of new models reaches values between 80 and 90 per cent in the last months of each year.

When applying MARS, the attention should focus on its values in the last months of the time window, when the high rates of churn become most apparent. Figure 1 makes clear that the GTIN level is not suitable as product stratification level. GTINs should rather be grouped, and doing this by combining televisions with the same screen size, screen type and whether or not 3D is supported, gives the best result in the above example.

Similar findings are obtained for hair care. New items appear on the market at lower rates than for televisions, but their sales shares still reach fairly high levels. Relaunches are a typical feature in this market segment. Figure 1 shows that also in this case the GTIN level is not appropriate. Defining item groups by brand and package volume gives better results according to MARS, which implies that old and new GTINs will be linked and that price changes associated with relaunches will be captured. The addition of a third attribute hardly improves the results.

# Conclusions

Until recently, decisions about defining products have been made by following a pragmatic approach. GTINs are found to be suitable products for stable assortments, while dynamic assortments are handled by selecting a set of item attributes, such that items/GTINs with the same characteristics are combined into the same group. How to select attributes was still an open question.

MARS offers a method for comparing different product stratification choices and ranks different item partitions according to scores that combine measures of homogeneity and degree of product match over time. MARS can contribute significantly to automating the product stratification process to a high degree, which is highly desirable when processing large electronic data sets. MARS may therefore become a valuable tool for supporting consumer analysts in their decisions about product definition, selection of attributes, etc.

The first results obtained with MARS look promising: GTINs are inappropriate as products in cases with frequent relaunches and high rates of churn, while the GTIN level results as a safe option for stable assortments like grocery items. MARS also tends to select attributes that are found to be relevant by consumer analysts, such as package volume, and screen size and screen type for televisions.

NSIs are often faced with limited metadata in transaction data sets. A major challenge therefore is to find out whether such data could be used for product stratification and index calculation purposes. One route could be to supplement metadata with web scraped data. Alternatively, price segments could be defined and explored as an additional variable. Statistics Netherlands and other NSIs are currently investigating both possibilities.

# References

1. A.G. Chessa, Product definition and index calculation with MARS-QU: Applications to consumer electronics, Report Statistics Netherlands (2018), 44 pages.