On the Operational Definition of Homogeneous

Products in Transaction Data

**Keywords:** Consumer Price Index, Scanner data, Transaction data, Homogeneous products, Unit value bias.

# Introduction

Statistics Sweden has been using transaction data in its monthly production of the Consumer Price Index since 2012. Although different possibilities for utilizing the new data source were discussed before the introduction (e.g. [1,2]), the decision was taken to mimic the previous setup used for manual price collection, however with an increased sample size and in some cases also a more detailed product definition. Since then, more transaction data of various forms have been introduced, and are continuously being introduced, into the Swedish CPI. As a result, it has become increasingly evident that methodologies need to evolve in order to make the best use of this new type of data.

One of the more urgent matters is the need for a unified treatment of the concept of homogeneous products. The impact that a particular price change in transaction data has on an index is often highly dependent on how the distinction between homogeneous and heterogeneous products has been made in practice. This observation applies not only to the type of fixed base indices used currently by Statistics Sweden, but also to other types of indices. The method used for partitioning transaction data into homogeneous products should therefore, in our opinion, be considered one of the most important features when discussing comparability between countries. In this paper we will report on preliminary results from work currently being done at Statistics Sweden in this area.

# homogeneous and heterogeneous products

In theory, the concept of homogeneous versus heterogeneous products is simple. Balk ([3], p.173) gives a definition in the form of an “operational question”:

*“Does it make (economic) sense to add up the quantities […] of the elements? If the answer to this question is ‘yes’, then the elementary aggregate is called homogeneous. The elementary aggregate can then be considered as a single commodity.”*

If products are homogeneous, a unit value is the proper representation of the associated price. As is well known, however, a unit value should not be used to compute the average price of heterogeneous products, since this would lead to unit value bias in the resulting index. (Von Auer, [4], refer to this type of bias as *assignment bias*, highlighting the fact that the bias comes from the inadequate assignment of products to groups which are not truly homogeneous.) A perhaps less known type of bias is the opposite one; by treating homogeneous products as if they were heterogeneous, a type of bias similar to that referred to by Von Auer [4] as *assortment bias*, can occur. Real price changes will then not be shown properly in the index. Silver [5] provides some intuition for this type of bias in terms of the cost of living, by citing the following example:

*“If […] the prices of goods A and B were 10 and 12, respectively, in both the reference and current periods, but there was a shift in quantities from 6 for both A and B in the reference period to 8 for A and 4 for B in the current period, a […] price index number formula for heterogeneous goods would give an answer of unity […]. However, the correct answer for homogeneous goods would be a […] fall of 3 per cent. […] The good, and cost of living with regard to this good, is, on average, now cheaper.”*

# Operational definitions of homogeneity in transaction data

Although the concepts of homogeneity and heterogeneity seem simple in theory, the practical choice of an operational definition is not always trivial. In many cases, practical matters such as data availability, set boundaries for the level of detail that is possible. With the increased access to highly detailed transaction data, however, possibilities for choosing between different operational definitions of homogeneity have increased.

When it comes to scanner data for daily necessities, the discussion has mostly been focused on so called “relaunches”; i.e. products that remain essentially the same but change for example package size between months. If homogeneity is defined on the most detailed level, i.e. with respect to GTIN or PLU codes, a price change introduced at the same time as a relaunch will not show up as a change in the price index, unless manual adjustments are made. Similar problems occur in for example clothing, although in this case the products often differ to a greater extent; Dalén [6] therefore refer to them as “near relaunches”.

In practical terms, the selection of an operational definition for what is to be considered a homogeneous product within a particular product group, is thus a question of trading two potential risks of bias against each other, trying to find the right balance. As a tool to assist this balancing act, Chessa [7] suggests a score function, referred to as the MARS measure (Match Adjusted R Squared). The idea behind MARS is to weight two factors against each other; the degree of homogeneity in prices within a suggested definition of homogeneous products, and the amount of matching over time for this same definition. A summary score is produced that can be used to compare different operational definitions (different “product scenarios”) with each other.

In this study, we will compute the MARS score for selected product groups and analyse the results. We will also compare the resulting indices.

# some preliminary results on the impact of the operational definition

Figure 1 shows preliminary (experimental) results of fixed base indices computed based on two different homogeneous product definitions, for the product group ‘Fresh salmon’.[[1]](#footnote-1) The most detailed grouping consists if unique item codes (*GTIN*) sold in unique stores (*Store Id*). In the broader definition, fresh salmon of the same *brand* and *type,* sold in the same store, are considered to be the same product, i.e. fully homogeneous. The effect on index in this example is quite large and will be further investigated in the coming study.

In figure 2, a similar picture is shown, containing preliminary computations done for the product group ‘Package holidays’. For this product group, Statistics Sweden is currently receiving test transaction data from several service providers, some of which are provided on a highly detailed specification level, and some on a more aggregated level. In figure 2, the two different aggregation scenarios are compared for a subset of the data material.[[2]](#footnote-2) Again, the effect of using different operational definitions of homogeneity is notable and needs to be further analysed.

Finally, figure 3 compares two alternative index series for the product group ‘dental services’.[[3]](#footnote-3) In this case the evaluation concerns not the *product dimension*, which is highly stable over time, but the *service level* of different dental service providers. In the figure, indices based on a highly detailed definition of the service provider (a specific dentist, working in a specific clinic) is compared to an index computed on a more aggregated level, differentiating only between private and public dentists working in different geographical areas.

**Figure 1: Experimental price indices for fresh salmon for 2018, based on two different homogeneous product definitions.**

**Figure 2: Experimental price indices for package holidays for 2018, based on two different homogeneous product definitions.**

**Figure 3: Experimental price indices for dental services for 2017, based on two different homogeneous product definitions.**

# Conclusions

In the present paper only preliminary results have been included, since the work is still ongoing. At the conference, final results will presented as well as a discussion of the main conclusions drawn.

# References

1. A. Norberg, M. Sammar and C. Tongur, A study on scanner data in the Swedish consumer price index, Room document at the twelfth Ottawa group meeting held 4-6 May 2011.
2. M. Sammar, A. Norberg and C. Tongur, Issues on the use of scanner data in the CPI, Paper presented at the thirteenth Ottawa group meeting held 1-3 May 2013.
3. M. Silver, An Index Number Formula Problem: The Aggregation of Broadly Comparable Items, Journal of Official Statistics, Vol. 27, No. 4, 2011, pp. 553–567.
4. B. Balk, Price and Quantity Index Numbers – Models for Measuring Aggregate Change and Difference, 2008, Cambridge University Press.
5. L. von Auer, Processing scanner data by an augmented GUV index, Eurostat Review on National Accounts and Macroeconomic Indicators, 2017, pp. 73-91.
6. J. Dalén, Unit values and aggregation in scanner data – towards a best practice, Paper presented at the fifteenth Ottawa group meeting held 10-12 May 2017.
7. A. Chessa, Product definition and index calculation with MARS-QU: Applications to consumer electronics, 2018, Report provided to participants of the Statistics Norway scanner data workshop held in Oslo in September 2018.
1. Experimental indices shown in this plot are based on a small selection of the full scanner data material provided to Statistics Sweden on a monthly basis, as not to disclose any detailed information. They are also computed using a different index number formula than the one used in today’s production, which means that results are not comparable with published numbers. [↑](#footnote-ref-1)
2. Again, only a subset of the full material is used in the comparison as not to disclose any of the data providers. For this plot, additional noise have also been added to the data. This, however, does not distort the overall picture. Results are for these reasons not comparable to published numbers. [↑](#footnote-ref-2)
3. The data material used for the experimental indices includes certain individual subsidies, which is not the case for the data used in the official CPI today. Results are therefore not comparable to published index numbers without further adjustments. [↑](#footnote-ref-3)