Towards a Roadmap for Efficient Use of Electronic Transaction Data in the Swedish CPI

Peter Nilsson[1], Olivia Ståhl[2]

Abstract: Statistics Sweden has recently started investigating the possibilities of implementing electronic transaction data on clothing into the CPI. As this would constitute an important next step in our use of transaction data, we believe that now is a good time to start reviewing already existing methods and processes to make sure that principles are fully consistent between subgroups and that the best possible use is made of data. This paper discusses our ideas for creating a practical internal guideline, or "roadmap", for the future use and implementation of electronic transaction data in the Swedish CPI. The roadmap should specify all the required steps necessary when incorporating new transaction data into the CPI, as well as when making quality improvements in sub-indices already based on transaction data. The principles of the guideline should be clearly stated and practical, and of course in line with international recommendations. The work on this internal guideline have started in the beginning of 2019, and the process will be continuously discussed in the national board of experts tied to the Consumer Price Index in Sweden.

Keywords: Consumer price index, Scanner data, Electronic Transaction data, Continuous quality improvement.

1. INTRODUCTION

Starting with scanner data on daily necessities, Statistics Sweden has over the last couple of years incorporated transaction data into many different product areas in the monthly production of the Consumer Price Index, as well as the European Harmonised Index of Consumer Prices (HICP). Transaction data now constitutes approximately 27% av the Swedish CPI basket and covers the whole of COICOP 01 and 02, as well as a parts of COICOP 04, 05, 06, 07, 09 and 12 (cf. figure 1). This can be contrasted with the situation less than 10 years ago; Our main approach to price measurement was then the manual collection of shelf prices, in outlets or over the telephone, a method that now constitutes only about 30% of the basket.

In the beginning of this transition towards more transaction prices, the experience at Statistics Sweden was often that each new data source was associated with a whole new set of conceptual problems and practical implementation issues; Each data set in this sense required its own type of analysis and implementation strategy. With time, most of the issues have however proven to be quite similar between product areas, and we now believe that it is possible to "streamline" the process of analysing new transaction data sets as well as that of implementing these new data sources into our production environments, to a much larger extent than we previously thought. We have thus started working towards an internal "roadmap for transaction data implementation", with the purpose of making sure that consistency between different sub-indices is upheld and important details not missed out as we in the next coming years transition towards even more transaction data and towards making better use of the data that we already have.

[1] Head of Price Unit, Statistics Sweden. Contact information: peter.nilsson@scb.se.
[2] Economist/Statistician, Statistics Sweden. Contact information: olivia.stahl@scb.se.

In this paper we will shortly describe the production situation at Statistics Sweden today when it comes to transaction data, and the reasons for why we see the need in the future for a more systematic treatment of different product areas and thus for the "roadmap" described earlier. We also put forward preliminary ideas on the content of such a roadmap as well as some thoughts about the process for constructing it.

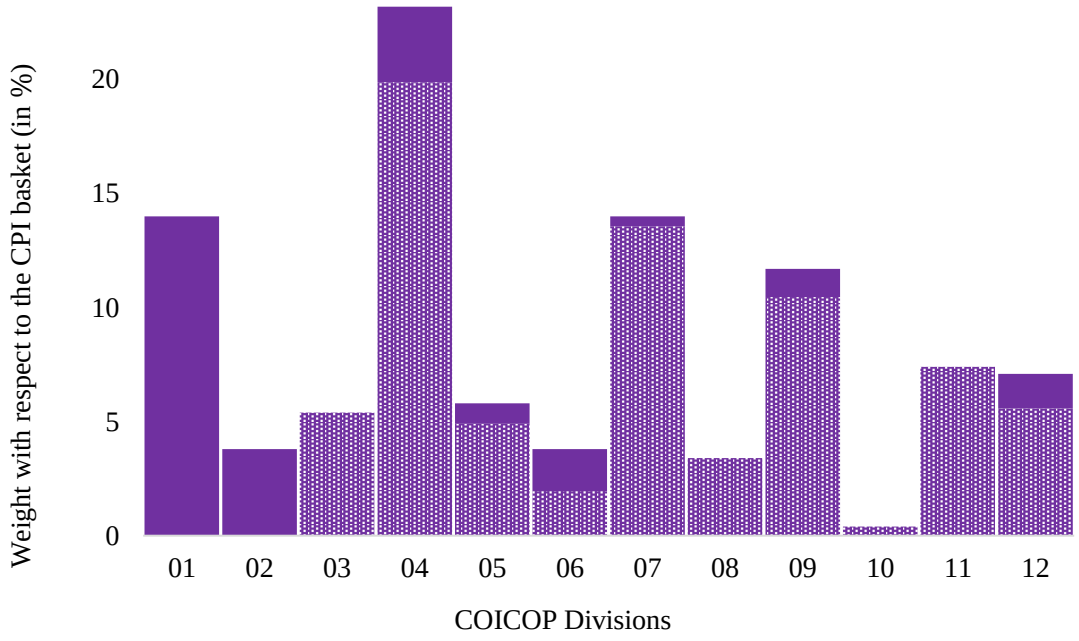

Figure 1: Transaction data coverage (dark purple) in different COICOP divisions of the 2019 CPI basket.

2. BACKGROUND: DIVERGING CURRENT PRACTICES

Currently at Statistics Sweden, transaction data is not always treated in a consistent way between subgroups. This is mostly due to the rapid transition into transaction data during the last couple of years, where decisions have had to be made successively, not permitting a broader overview until recently.

As already mentioned, the first type of transaction data introduced into the Swedish CPI was supermarket scanner data. This data consists of weekly average prices for daily necessities such as foodstuff, non-alcoholic beverages, tobacco and hygiene products. Except for fresh fruit and vegetables, which are treated separately, indices for these products are based on a static sample of "product offers", where a product offer is in this case identified in the beginning of the year by a barcode and an outlet ID, and then replaced with other barcodes manually throughout the year if judged necessary.

For transaction data other than the supermarket scanner data, methods used today are somewhat diverging. For example, random sampling is used within some product areas whereas the full material is used in others. The amount of manual replacements made also differ quite a lot and there is a concern that the allocation of resources between product areas in this regard is not fully optimal.

Another area where practices are diverging today is in the methods used for stratifying data into so called "homogeneous products". Until now, this issue have not been well coordinated between product areas. Although for most areas the level of stratification used today is probably well balanced, there are examples of product groups where we suspect that specifications optimally

should be either somewhat tighter, or somewhat wider, than they currently are. For example, an analysis performed recently showed that the specifications previously used for dental services where too loose, thus resulting in a risk of unit value bias. (This result lead to a change in the specifications used for this product group as of 2019.) At the same time, there have been discussions in our CPI board about the specifications used for medicines. Here the worry is the opposite, i.e. that the current level of specification, which includes item-specific "barcodes", might be too tight; It has been argued that perhaps consumers care only about which active pharmaceutical ingredient they buy, in which case the current practice of Statistics Sweden wouldn't show the correct price development. Statistics Sweden has recently started to look into this issue in some detail (see Ståhl, 2019, for some preliminary results). Our main conclusion thus far is that in many cases it might well have a larger influence on the final index series than, for example, the choice of index formula or sampling design, a conclusion which is consistent with results of similar studies (e.g. recently Nygaard, 2019).

Another area which we feel requires an overview is that of micro and macro editing. The introduction of electronic transaction data has set new demands on the editing process, and might call for different approaches than the ones we are used to.

Further, the IT setup used for our electronic transaction data today is not fully optimal. In our discussions with data providers, the aim has always been to minimize response burden; We believe that this is the best approach to achieve high quality data. As a result, we now have quite a large divergence between product areas with respect to the in-house data structure and storage facilities. Organizing the data in a more systematic way will not only facilitate the editing possibilities and decrease the risk of processing errors, but also enable ad-hoc methodological studies to be performed more easily.

Finally, user communication issues have been topical at the Swedish price unit during recent years. In the process of deciding whether or not to transition to a new data source and associated methodology it is important to keep all quality dimensions in mind - users of the CPI usually do not care only about accuracy but also about comparability over time. We are currently working on a new policy on implementation issues, which will be discussed in our CPI board later this year. Increased consistency is desirable also here, as a way to keep the methods of the Consumer Price Index appearing transparent to our main users.

3. PRELIMINARY IDEAS FOR THE ROADMAP

Although it is certainly true that different types of products have different dynamics and might require different approaches to price index compilation, we believe that a higher degree of consistency between product groups is desirable when it comes to the use of electronic transaction data in the Swedish CPI. This does not mean that the same index number formula or sampling design must be used for all areas. What we aim at is a consistency with respect to the principles used and the analyses performed, not with respect to the methods finally chosen. (This is also in line with how the Swedish CPI is computed in general; About 70 different surveys form the basis of the current index, and methodologies are tailored to fit the product group in question taking e.g. resource allocation issues and underlying economic assumptions into account.)

Of course, we do not yet have all the answers needed to construct our "roadmap", and indeed some answers still wait to be found within the international price index research community. By starting to work more purposefully on these issues we, however, hope for the knowledge that we will acquire over the next coming years to be more easily translated into a broader understanding within the CPI team. The work should help us systematizing the quality improvements made within this area and increase the economies of scale as more and more transaction data need to be analysed and incorporated into the CPI. Finally, we believe that clear principles and distinct processes will be even more important in the future, as we expect the situation to change from consisting of a

couple of large transaction data providers into many small providers. Below, we list the main topics that we have discussed so far.

TOPIC 1) OBTAINING A FIRST CONTACT WITH POTENTIAL DATA PROVIDER

During 2018, staff from Statistics Sweden's price unit have worked together with the communication division on a draft communication strategy for how to approach new potential data providers as well as on a list of priorities. From 2019 a reorganization within Statistics Sweden has resulted in a transfer of responsibilities for these issues from the price unit to the division for data collection, and work will continue in this new constellation.

TOPIC 2) REQUIREMENTS ON EXPERIMENTAL DATA SETS

A practical checklist for what type of information is needed to be able to perform preliminary analyses on experimental transaction datasets will be developed during the coming year. This checklist should not only be of help to price statisticians but also, and perhaps more importantly, to the staff working at the new data collection division who will be involved in these issues.

TOPIC 3) SYSTEMIZING CONTINUOUS DATA DELIVERY

As already mentioned, we believe that transaction data providers should have a large amount of freedom with respect to the form in which data are delivered; The role of Statistics Sweden is mainly to assist with competence on legal, technical and methodological issues. We thus need to have IT systems flexible enough to incorporate all of the different types of input data delivered and store them in a consistent way. Discussions with the IT department on this issue have started and a pre-study is planned for 2019.

TOPIC 4) METHODOLOGICAL ISSUES

Classification

During 2019, automatic classification methods will be evaluated. If evaluations prove to be successful, the methods will be incorporated into the product sampling of daily necessities for the year 2020. Specifications of the classification methods will in the future need to be part of the roadmap framework.

Stratification

The issue of finding appropriate operational definitions for homogeneous products in electronic transaction data will probably always will be associated with some degree of subjectivity, but we believe that practical methods such as the MARS score function of Chessa (2018) can be of great help. Work on this topic is ongoing and will continue into 2020.

Data cleaning

In the future, the methods used for data cleaning and outlier detection will also be made more consistent between product areas.

Editing

Editing processes used for electronic transaction data will be reviewed and hopefully result in practical guidelines. We are currently mainly discussing moving from top-town approaches to probabilistic ones, as well as shifting more focus to macro level comparisons.

Index formula

When it comes to the issue of index number formula, we aim at a methodology that takes the actual behaviour of consumers into account in a better way than our current methods do. Statistics Sweden's research on index number methods for transaction data have, however, only just started. Our aim is to adjust our internal guidelines continuously as we learn more about these issues from the international research community as well as from our own experiences. In the end, we hope to be able to set up a list of criteria's that should be fulfilled for a particular index number formula or strategy to be selected. We see this as a potential way to keep being transparent towards our users as methodologies become more complex.

Sampling design issues

As for all types of statistical surveys we aim at an "optimal" sample size, which could mean surveying the whole data material or excluding only extreme or highly unusual values. Such a decision must, however, always be preceded by some sort of analysis contrasting sampling variance with the risk of different types of systematic errors.

TOPIC 5) IMPLEMENTATION ISSUES

The draft policy that will be discussed in the CPI board later this year includes a list of concrete requirements such as e.g. the availability of parallel series of at least one full year before implementation can be discussed, and topics related to user communication strategies.

4. FINAL REMARKS

We believe that a highly important prerequisite for the successful implementation of more transaction data into the Swedish CPI is a close collaboration between staff at the price unit and other parts of Statistics Sweden such as e.g. the data collection division, the methodology department and the IT department. The topics suggested in this paper are only a start, and we expect to further develop this list in collaboration with the other divisions.

REFERENCES

Chessa, A. (2018). Product definition and index calculation with MARS-QU: Applications to consumer electronics. Report provided to participants of the Statistics Norway scanner data workshop held in Oslo in September 2018.

Nygaard, R. (2019). Exploring the use of scanner data in the Norwegian CPI for products with high churn. Abstract presented at the New Techniques and Technologies for Statistics (NTTS) conference held in Brussels in March 2019. (Retrieved April 29 2019 from https://coms.events/ntts2019/data/x_abstracts/x_abstract_37.docx.)

Sammar, M., Norberg, A. and Tongur, C. (2013). Issues on the use of scanner data in the CPI. Paper presented at the thirteenth Ottawa group meeting held in Copenhagen in May 2013.

Ståhl, O. (2019). On the Operational Definition of Homogeneous Products in Transaction Data. Abstract accepted for presentation at the New Techniques and Technologies for Statistics (NTTS) conference held in Brussels in March 2019. (Retrieved April 29 2019 from https://coms.events/ntts2019/data/x_abstracts/x_abstract_154.docx.)