ES/PR
John Johansson

PM till nämnden för KPI
Sammanträde nr 7
2019-10-16

Scanner data in the Swedish CPI, where are we now?

For information

This paper is a brief summary of the work that has been done during 2018 and 2019 in the scanner data field for the Swedish CPI.

Content

Background

Scanner data in the Swedish CPI (and HICP) has been used in production for over a decade. By scanner data we mean census electronic transaction data that contains information on turnover and quantities, divided into some sort of a product key. For most of the consumer goods, scanner data is divided into GTIN-codes (also known as EAN-codes) or PLU-codes. As for services, the products are defined as a combination of characteristics.

The first scanner data was received from a couple of supermarket chains, followed by scanner data on pharmaceutical product, alcoholic beverages and real estate agent fees. In the last two years new sources of scanner data have been obtained, such as package holidays, clothing and train tickets. The methodological principals of the use of scanner data have almost been the same since it got implemented at start. Price indices based on scanner data are calculated in the standard Swedish practice for elementary aggregates in the Swedish CPI: static bilateral Jevons indices with a PPS sample of products. The sample of products is chain specific, but is symmetrically matched against specific stores. Regionally differences in product assortment are therefore not taken into account in the product sample or in the product weights.

As new techniques of handling data have developed and new scanner data has been acquired, Statistics Sweden (SCB) is now facing new opportunities and challenges that could affect current principals on treating scanner data. SCB has been working the last two years (2018-2019) with a grants from Eurostat to try to implement new scanner data, as well as new techniques to handle existing scanner data. In this memo, we will summarize the work that has been done for this grants project. The first chapter will briefly present all the new scanner data that SCB has obtained, tested and implemented. In the following chapters, we will present some more specific insights from our work with scanner data. Most of the findings in this memo have already been presented to the CPI-board in earlier reports.

Purpose

The purpose of this memo is to briefly summarize the work that has been done during the last two years (2018-2019) to collect, test and implement scanner data in the Swedish CPI and HICP. The CPI-board is encouraged to discuss the general path that SCB should take in further testing and investigating of scanner data.

New scanner data

SCB has obtained numerous new scanner data the last two years. Some of them have been implemented in the Swedish CPI (as well in HICP), while some data are not yet implemented in production. Table 1 gives a description of the types of scanner data that have been obtained during the last two years. For fresh food (meat, fish, fruit and vegetables), product identification is often a PLU code. PLU codes work similar to GTIN codes, but do not necessarily have to be as homogenous as GTIN codes. Also, they are not necessarily generic codes, they can be shop specific codes. "Sales report based" scanner data contains information on turnover and quantities into specific groups of specifications. Register based scanner data is similar to sales report based, where the difference is that registers shows every transaction separately.

Table 1: Scanner data obtained by SCB in 2018-2019

Name	Type of scanner data	Implemented
Fresh food	PLU based	2018
Two additional supermarket chains	GTIN and PLU based	2019
Train tickets	Sales report based	2019
Package holidays	Sales report based	2019
Dental Care	Register based	2018
Petrol	Sales report based	Not implemented
Clothing	GTIN based	Not implemented

Two additional supermarket chains were implemented in 2019, and are treated with the same principals and in the same system as the other super market chains. No tests have been done on the petrol data. Calculating price indices on petrol scanner data should not be as challenging as for other types of scanner data, both product definition and product churn is expected not to be challenging.

Train tickets (Bubuioc and Tongur, 2019), package holidays (Johansson, 2019), dental care (Johansson, Ståhl and Tongur, 2018) and clothing (Bubuioc and

Tongur, 2019) have already been discussed in the CPI board, and will not be discussed further in this memo.

Fresh fruit and vegetables

When we were testing the scanner data for fresh fruit and vegetables, we detected high product churn for some of the products. A standard static approach wouldn’t then be feasible, since it requires many manual replacements in order to keep the basket none shrinking.

The new method for fruit and vegetables, where the most sold product is selected, was implemented in 2018. To find the most sold product, the retailer classification is used along with text mining. About 50 different retailer codes containing fruit or vegetables are manually distributed and mapped into one of the 27 product groups[1]. The retailer classification is stable when using this level of classification. However, the retailer classification does not perfectly match the product group classification. To get all items correctly classified we use text mining. Pearl expression is used to analyze and score the text strings to decide the right product group for the product. The text mining process needs to be updated continuously mainly due to changes in unit e.g. from kilo to piece. Lastly, the most sold item within each product group is selected.

For fruit and vegetables the main concern both before and after implementation is whether the price is given for a kilo or for a piece. The main variables to check are unit and number of sold items. Unit should be in kilo, and number of sold items should be in decimals. By comparing these variables over time we know which product groups that needs more monitoring during the monthly production process. For example we have learned that cucumber, grapes, and lemons needs to be thoroughly checked every month.

Fish

In 2018, SCB presented a micro study on fresh fish on the scanner data workshop in Oslo. The CPI basket of fish was shrinking in 2018 as a consequence of problems finding replacement for disappearing items, and we therefore found it necessary to analyze scanner data for fish. The sample selection for fish was prior to 2019 performed as standard sample selection for scanner data in the Swedish CPI. The standard way is to cross match a PPS

[1] Product groups are defined by SCB and are groups within COICOP 4

sample of products with a sample of stores. The same sample of products is then measured in every store.

The study showed that the majority of the fish sold, had shop specific codes. An overall sample of products didn't represent the supply of fish of the selected shops. It also showed that a monthly chaining approach didn't have higher matching proportions then a fixed December base approach. This could be explained by the fact that products come and go and have a seasonal pattern with different kinds of fish in the summer. A monthly chaining approach also showed tendencies of a downward chaining bias.

As a result of the study SCB changed sampling method for fish for 2019 to a cut off of products per store, instead of a PPS sample from the total supply of fish for every supermarket chain. Replacements in 2019 are therefore made for each store separately instead of as before, generic replacements for products in all stores at the same time. The matching proportions for fish have increased with approximately 20 percentage points for 2019 compared to 2018.

Dumping filter on abnormal prices

Scanner data for groceries captures all sales per item and store. Extreme low prices can be included from promotions, sales and various forms of special offers such as gifts or goods with short date, where the latter two types of prices should not be included in the CPI. A study was presented to the CPI board in 2018 (Bilius, Bubuioc and Tongur) with a proposal on a decision rule to automatically filter extreme prices. Three different kind of approaches were proposed, all based on filter prices that goes under a benchmark of a percentage deviation of a mean price. The automatic filters gave small effects on aggregated CIOCOP 1 (exclusive fresh food) level, and were sooner implemented in the Swedish CPI. Manually reviewed price distortion was able to be replaced by an automatic process.

Testing of dynamic and multilateral index methods

SCB has conducted several micro studies on comparing the currently used fixed basket with alternative index methods applicable to scanner data. In one of these studies, the task of using more data was approached. In the study by

Bilius and Tongur (2017), the obstacle of data coding was mitigated in a frugal and easy way by applying a “reversal” of the coded items in the existing CPI sample for daily necessities, COICOP 01. This was achieved in two steps.

First, the given CPI sample of some 800 items per retailer was taken and matched, through GTIN with scanner data during the two studied years (2015 and 2016) to identify corresponding retailer code. Second, having both retailer categories and the corresponding COICOP codes from the sample, much more items could be “linked” to COICOP through this “blindfold” approach.

The outcome was that the uncontrolled matching in the sense presented here rendered total COICOP 01 index based on monthly chaining that was very similar to the actual fixed basket version. In both cases, unweighted Jevons was applied.

SCB presented a comparative micro study on static, dynamic bilateral and multilateral methods for Coffee (Ståhl, Tongur, and Bilius) in Geneva 2018. A general conclusion from that study was that the compilation of homogenous groups was crucial for resulting indices, as well as big sales are affecting indices with non fixed weights.

Product group classification of scanner data

While increasing the use of transaction data in the Swedish CPI, for many product areas, a coherent and efficient strategy for the classification of products to product groups is needed. SCB has identified the need to explore various possibilities for automatic classification of retailer codes to product groups of required detail, using e.g. text mining, machine learning, or direct mappings of retailer classified product groups to CPI product groups.

Our preliminary findings below come from an ongoing study, with focus so far on supermarket scanner data. As to the possibility of creating and maintaining direct mappings of retailer classification to product groups, we tentatively conclude that:

- Some retailer classified groups might be difficult to classify consistently to one unique product group, and vice versa, some product groups might be difficult to construct from a specified list of retailer classification.
- In general, retailer classifications are sufficiently detailed for the required product group classification, with only a minor number of problematic groups requiring special treatment.

- In general, retailer classifications are stable over time, with only a few changes per year and retailer.
- Securing access to retailer classifications of sufficient detail and quality is important when forming agreements with data providers.

Detection of relaunches

In order to keep our fixed basket non shrinking, SCB allows replacement in the basket to be more widely defined then what a relaunch can be defined. A relaunch is defined as a product that has changed packaging and GTIN-code, but remains the same ingredients and packaging size. In the Swedish CPI we allow replacements to have minor changes in packaging size as well as minor changes in ingredients. SCB names the production procedure to make replacements in the scanner data basket for "basket analysis", and is a manual procedure with support from SAS-programs, where an official on SCB decides which product that is most suitable as a replacement (and therefore a semi-automatic procedure).

As a continuation of the project on automatic classification of GTIN-codes to product groups, SCB has improved the detection of relaunches in the basket analysis. SCB has included the retailer category codes in the basket analysis, in order to refine the proposed replacements. The retailer category codes is saved as a key to every single product that has been selected in the base month. When a product is showing signs on going out from the market, a SAS-program suggest relaunches based on the retailer codes and text string similarities (using pearl expression).

SCB has also obtained a "relaunch code" from one of the suppliers of scanner data for daily necessities. SCB believes that the relaunch code is too specific, and can only detect strict relaunches. For the majority of the replacements done in 2019, the relaunch code has not been able to detect any suggestion for a replacement. In those few times when the relaunch code has suggested a replacement, further basket analysis haven't been necessary.

Conclusions

Based on our micro studies on e.g. fish, coffee and clothing, SCB sees no argument to implement a bilateral dynamic method for scanner data. Our view is supported by the draft of the new ILO CPI manual on scanner data (page 366):

> *"**10.61** Also, when using a census of varieties, not a sample, a weighted index number formula should be used. Again, variety turnover poses a significant problem. To maximize the number of matches in the data, chaining at high frequency will be needed. This can lead to a significant drift in the index. Multilateral price index number methods, which are drift-free by construction, are currently the most suitable method to handle a census of items and varieties from scanner data. Scanner data offers many opportunities for new research and develomemoents."*

Chaining should be considered in a multilateral method in order to be drift free. SCB is participating in Eurostat on drafting a recommendation on multilateral methods, and will follow the research done on new index methods. Until there is a consensus of which multilateral index method to use, SCB will continue improving the quality and efficiency within our current index formula.

With our improvements on the basket analysis and with the automatic price filter, we have been able to efficient our production. Our project on mapping retailer codes to our product groups has given us the opportunity to test new methods in a bigger scale, as well a more efficient sample selection procedure. We believe that we need to work more with our IT-infrastructure for scanner data, in order to improve our production and testing of scanner data. We need to be more efficient overall, in order to e.g. increase our sample of products.

Furthermore, except that we are going to test scanner data on clothing and petrol, we want to estimate the effect that our basket analysis has in comparison to the sampling error. The tradeoff of the sampling error making replacements of a sample of products versus the bias of making no replacements for a census of products would be a valuable insights for SCB. Maybe it is possible to have the best of both worlds, by calculating scanner data on a census of products (with no sampling error) and only make sufficient replacements for relaunches (to reduce the bias) that we are able to automatically detect?

References

Bilius, Å., Bubuioc, R., Tongur, C (2018). ”Hantering av extrema priser i kassaregisterdata”. Memorandum to the Swedish CPI board, meeting 4 (in Swedish).

Bilius, Å., Tongur, C (2017). “Empirisk jämförelse av fast varukorg visavi månadsmatchning för dagligvaror”. Memorandum to the Swedish CPI board, meeting 3 (in Swedish).

Bubuioc, R., Tongur, C (2019). ”Järnvägsresor i KPI - ändringar från 2019”. Memorandum to the Swedish CPI board, meeting 6 (in Swedish).

Bubuioc, R., Tongur, C (2019). ” Preliminary findings in scanner data on clothing”. Memorandum to the Swedish CPI board, meeting 7.

IWGPS (2019), “Consumer Price Index Manual, draft version”. Available online: https://www.imf.org/en/Data/Statistics/cpi-manual (downloaded 2019-09-27).

Johansson, J (2019). ”Effekter av ny insamlingsmetod för flygcharter”. Memorandum to the Swedish CPI board, meeting 6 (in Swedish).

Johansson, J., Ståhl, O., Tongur, C (2018). ”Prisvariabel i mätningen av tandvård”. Memorandum to the Swedish CPI board, meeting 4 (in Swedish).