Experiences and best practice in developing new production processes for statistical products: Price statistics – the case of Slovenia

Mojca MAČEK KENK

Senior Adviser, Head of Section, Statistical Office of the Republic of Slovenia, Slovenia

Abstract. The use of scanner data appears as a promising new data source for HICP compilation. Transaction data obtained from retail chains contain data on turnover and quantities per item code and for their use we developed a new production process and improved the quality of the Slovenian HICP. After determining the list of the most important and largest retail chains, we put quite some effort into negotiating with retailers. We managed to obtain data on a regular basis from all of them. In 2018 we introduced a new methodology for computing inflation, where the price movements of food, beverages and tobacco are shown exclusively from the scanner data. Instead of 6,500 prices, around 70,000 prices are now used for these products with the new data source. We also use web-scraped data for calculating indices for computer equipment.

1. Introduction

In order to constantly adapt to the changes that are taking place in the field of statistics and to modernize and optimize statistical processes and improve quality, the Statistical Office of the Republic of Slovenia (SURS) wanted to modernize as much as possible the data collection methods for compiling the harmonised index of consumer prices (HICP). The aim was to explore new technical and methodological solutions for data collection and data compilation by using different methods and data sources, which is also one of the key areas of the ESS 2020 vision. Our activities focused on modernizing the consumer price collection and production process by using different tools and data sources and by analysing and exploring different new ways of automated data collection. Most of the work was done within two Eurostat's grants. Due to its importance, the work at SURS was organized within the framework of the project and was subject to project management.

The paper presents our experiences exploring the possibilities to regularly obtain scanner data from biggest retailers and find practical solutions for using them for HICP compilation. The last part of the paper is dedicated to the analysis of the use of web-scraped data for the HICP purposes.

2. Obtaining scanner data

The aim of modernising price statistics is to ensure that price collection methods remain appropriate in a world of increasingly dynamic markets for consumer goods, dynamic pricing and ingenious ways of providing discounts. Scanner data are a new quality data source. They are generated by point-of-sale terminals in shops and provide information at the level of the Global Trade Item Number (GTIN). Scanner data are aggregation of the turnover and quantity of individual transactions per GTIN for a given period and location and provide information on what the product is.

In the preparation phase, SURS focused on reviewing, examining and analysing existing documents and practices of MSs already using scanner data as a new data source for HICP compilation. This was also a very good starting point for further steps.

First of all, a detailed market research was done to determine the list of the most important and biggest retailers (retail chains) relevant for the Slovenian consumer consumption pattern. The selected retailers cover 75% of the market share of sales for food and non-alcoholic beverages and almost 70% of the market share of sales for alcoholic beverages and tobacco.

We prepared special letters inviting retailers to the meeting with the aim to present in detail why scanner data are so important for SURS. Due to the sensitive nature of the scanner data (detailed information at item code level - detailed item description, turnover and quantities sold) and treatment of these data as a business secret, it was expected that quite some effort will be necessary to convince retailers to participate in this project. For that reason, it was decided to include the top management in the negotiation process with retailers. At all the meetings, which were held at the level of directors, the mission of statistics, its activity and importance, and the data protection policy were presented as well as the current method of work in price statistics and the reasons for modernising our processes in times of modern information technology. Particularly pointed out was that such data, which are at the most detailed level and are a business secret for the retailers, will be particularly protected. We also took care of the legal basis. Compulsory reporting of scanner data was added to the Annual Programmes of Statistical Surveys, which are the legal basis for conducting statistical surveys in the current year.

The negotiation process lasted more than three years, since there were a number of meetings at expert level with each retailer after approval at the highest level, where we negotiated about the structure of the data and coordinated the sending procedures. We found out that negotiations are a very time consuming and on-going process with no time limitations. Some retailers were very sceptical at the beginning. They doubted whether we will manage to obtain the data also from other retailers. In the end, all retailers agreed to cooperate. The reason for the success was our comprehensive and strategic approach. For each retailer a special contract was prepared in which SURS committed in several articles to protect the data. Contracts contain information about detailed structure of the data, period of data transmission, format of the file and the method of data exchange (we developed a system for secure scanner data transmission).

3. Processing scanner data

Retailers send us the data weekly, at the most detailed level, for all their stores and for the entire range of products they offer. At the beginning they had a lot of work preparing appropriate data structure and setting up all procedures for automatic data transmission. The structure and characteristics of the scanner data are not the same across retailers. Despite the wishes to adjust the structure of the data sets to our requests, some retailers did not agree. Changes and adjustments would cause them additional burden both in terms of human and financial resources. For that reason, almost for each retailer, in the end, we had to find a solution at SURS for arranging the data in such a way that we can use them in automated procedures.

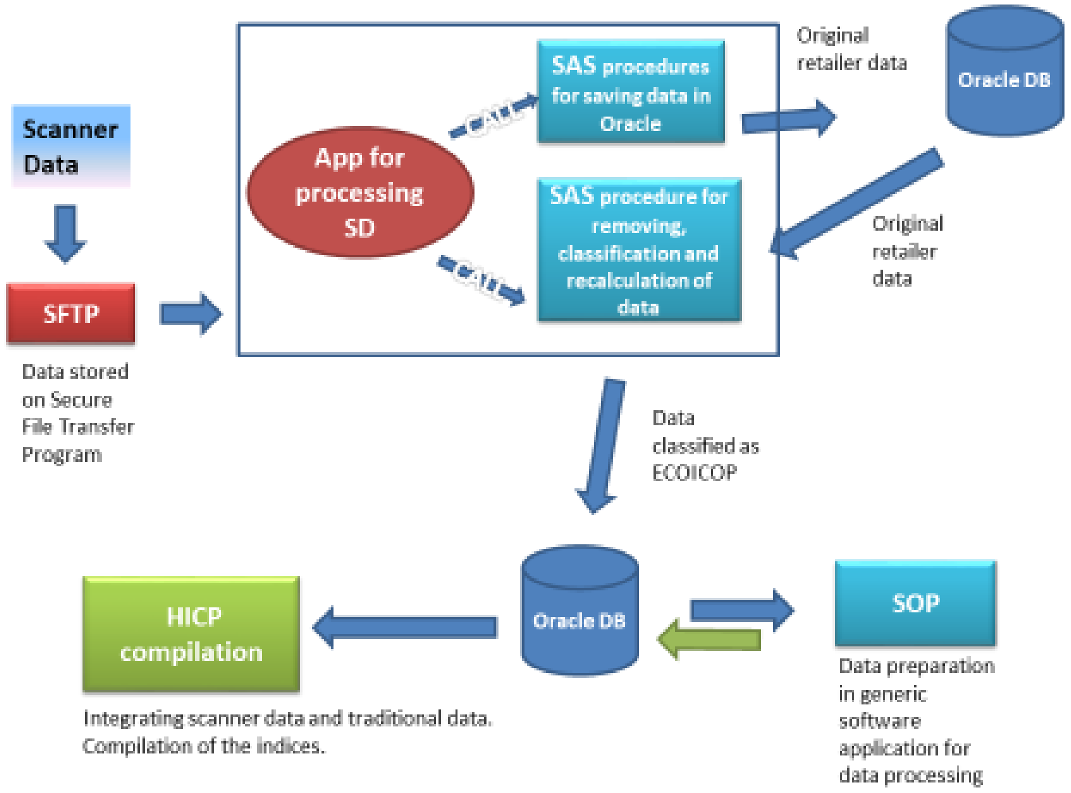
First, scanner data of each retailer were included into regular processing according to the old methodology, the system of representative products. This was a simulation of the traditional price collection. We promised retailers that we will include the received scanner data as soon as possible in our regular processes, and at the same time stop collecting data in their stores so that we will no longer burden them. We received the data gradually and gradually included them in the HICP compilation. Since the receiving of the first data and the data of the last retailer lasted more than three years, we could fulfil the promises we made to them only with this approach. After receiving data from the retailer, the main task at SURS was to find the technical solution for mapping the GTIN code to an elementary aggregate in HICP. At the beginning we had a lot of manual work to prepare links between our selected products at this retailer and its GTIN codes. The search was manual and we helped ourselves with detailed product descriptions. To be able to match individual products between time periods, the GTIN code was used as the product identifier. We had to prepare certain adjustments to the software and then procedures were running automatically every month.

While using scanner data only as a substitute for traditionally collected data in a regular production, we were intensively studying and developing procedures for more comprehensive usage of data from the retailers' databases. We followed the Eurostat's guidelines for processing supermarket scanner data to a great extent, because they base on current best practices in order to ensure harmonization of consumer price indices.

We prepared a detailed plan for changing the methodology. Because the system of representatives was abandoned, it was necessary to change both the calculation of indices (introduction of a new index type) and the weighting of data at the lowest level. We prepared a system of data control and criteria for data capture for processing. We defined rules for the production of new weights and rules and procedures for the new calculation of indices at the level of elementary aggregates. All these changes also required a change in the production process. We developed a new software solution that enables central management of received scanner data. It supports the process of storing data in a database, sorting, linking, removing and converting data according to the new methodology. For each retailer, catalogues were prepared where their data were linked to the European Classification of Individual Consumption by Purpose (ECOICOP) at the most thorough 6th level. We prepared test calculations based on various indices and a variety of analyses, with which we tested both the new methodology and the procedures themselves.

A lot of work was done by adapting all existing procedures so that we could integrate all the innovations into an automated data processing process and, in parallel with the regular process, prepared the calculation of indices according to the new methodology. The new monthly production process is presented in Picture 1.

Picture 1: New monthly production process


Scanner Data
SFTP
Data stored on Secure File Transfer Program
App for processing SD
CALL
CALL
SAS procedures for saving data in Oracle
SAS procedure for removing, classification and recalculation of data
Original retailer data
Oracle DB
Original retailer data
Data classified as ECOICOP
Oracle DB
HICP compilation
Integrating scanner data and traditional data. Compilation of the indices.
SOP
Data preparation in generic software application for data processing

In 2018 we introduced a new methodology in our HICP. For the first two divisions of the ECOICOP "01 Food and non-alcoholic beverages" and "02 Alcoholic beverages and tobacco" scanner data are the only source, while in other divisions scanner data are integrated with traditionally collected prices.

For users, we prepared detailed information on changes and also the comparison of monthly indices calculated according to the new methodology with published indices for 2017. The 12-month average showed no difference at the all-item index between the test and the published indices. The variability of monthly indices with the new methodology at lower levels of ECOICOP is slightly higher than in the old methodology, particularly in groups in which goods are frequently discounted, which reflects the actual situation on the market.

4. Web scraping

Within the project, we also analysed the use of web scraping. At the beginning we identified the target product groups and studied the possibility of using web scraping at selected websites. We selected websites with data on profit rents and second hand motor cars, since we monthly collect a large amount of data from these sites and data collection lasts several days.

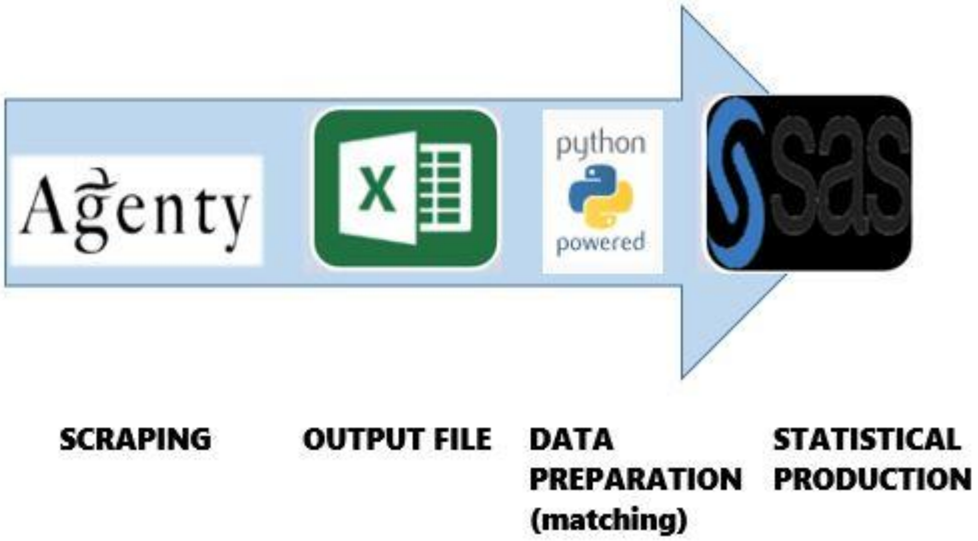
We contacted the owners of selected websites and asked them to allow web scraping. Unfortunately, they did not. After negotiating we managed to sign an agreement with the owners of the website with data on profit rents for sending us monthly data directly from their database. The owners of the website with data on second-hand motor cars requested payment for such a service, so we continue with traditional data collection. We found out that web scraping is not allowed by owners whose websites are their products and the source of their income since they invest a lot of effort in collecting and processing data.

Therefore, we decided for web scraping of retailers' websites and we select retailers for computer equipment. First we checked the general conditions of individual webs. We found that one web retailer prohibits any automated data queries. We asked for permission to automatically scrap the data but our request was denied. Two web retailers prohibit further use of their data without their permission and the others have no such statements. After that we prepared a list of technical criteria for all computer equipment included in the basket for HICP compilation.

All websites for testing were well structured, so we focused on testing interactive, freely available online applications that enable web scraping of structured data. Their greatest applicable value is their simplicity and user guidelines available to users within the applications. Before choosing a test tool for scraping data, we checked the operation and usability of the three tools. We checked the IROBOT and KIMONOLABS tools, but the most useful was the Import.IO tool, which was also used by other countries. We also decided to use it, as the tool was free and at the same time the optimal choice for our needs.

After a while the Import.IO tool became payable and SURS decided to purchase a Data Scraping Studio license (renamed Agenty), since free tools could no longer be used. The new tool proved to be more reliable, provides more functionality as well as automation and more detailed definition of variables in HTML code (with CSS selectors). We managed to prepare all the procedures for including the scraped data for computer equipment according to the system of representative products into HICP compilation in January 2018. The process of processing data obtained with web scraping is presented in Picture 2.

Picture 2: The process of processing data obtained with web scraping



5. Conclusion and future plans

Using scanner data in statistical processes greatly increased the dependence on retailers, so we will maintain good relations with them to ensure that data continue to be supplied and issues are resolved in a timely fashion. For each retailer we have prepared and will also regularly provide them with specific analyses of the movements of their prices in comparison with the trends in the prices of all retailers at the ECOICOP level. This is one of the ways to thank them for their efforts and could help them in their business decisions. We will also try to obtain scanner data from other retailers, especially from specialized stores (such as drugstores).

A great challenge for us was the amount of data that had to be processed and for which the appropriate IT solutions had to be developed. We put a lot of effort in developing of our own software solution, which is constantly being upgraded in order to optimize processes as much as possible.

We will continue to work on improving the methodology and, in particular, data processing operations. We will also actively participate in the development of new methods and finding new solutions for processing scanner data at European level, because cooperation between countries greatly contributes to the spread of experiences and best practices. Very important for us were study visits we made at statistical offices of Norway, Sweden and the Netherlands. They presented their systems, experiences and the practice of using scanner data and answered our specific questions, which has been of great help to us.

Within our new project we will continue our work on web scraping. The first objective of the project is to upgrade web scraping for computer equipment. We will try to develop our own custom-made tools and also prepare a new methodology for more comprehensive use of web scraped data. Our second objective is extension of web scraping and analysing of manually and automatically collected data for airline tickets and holiday packages.

6. References

[1] Eurostat (2017), Practical Guide for Processing Supermarket Scanner data

[2] SURS (2018), Changes in computing inflation in 2018