CONFERENCE OF EUROPEAN STATISTICIANS WP.2-3

Workshop on Statistical Data Collection 'Visions on Future Surveying' 2 September 2016

3-5 October 2016, The Hague, Netherlands

Challenges and solutions to the use of internet data in the Dutch cpi

Robert Griffioen, Olav ten Bosch and Els Hoogteijling (Statisitcs Netherlands)

r.griffioen@cbs.nl, o.tenbosch@cbs.nl and emj.hoogteijling@cbs.nl

Abstract and Paper

For the computation of the Consumer Price Index (CPI) and the Harmonised Index of Consumer Prices (HICP) in the Netherlands, prices of consumer goods and services are acquired by using both conventional as well as innovated methods of collecting. CAPI and CAWI are considered as conventional and the use of scanner data and web scrapers on the Internet are modern methods. Namely the use of web scrapers has had an impact on the production processes within the CPI department.

This paper reports on the introduction of web scrapers for the collection of internet prices and metadata for the CPI and HICP computation. The use of internet data in the Dutch CPI in the current and near future will briefly be explained. Their core values and the important challenges that come with them will also be addressed, i.e. organisational, technological and legal issues. The concrete strategic solutions that have been taken in the CPI department will be discussed as well. We will conclude with a discussion of the possibilities and challenges for the application of web scrapers in other fields of Official Statistics.

Workshop on Statistical Data Collection (2016)
<The Hague, The Netherlands, 3-5 October>

Topic (i): Scanner data and big data

CHALLENGES AND SOLUTIONS TO THE USE OF INTERNET DATA IN THE DUTCH CPI

Invited Paper

Prepared by Robert Griffioen, Olav ten Bosch and Els Hoogteijling[1], Statistics Netherlands, the Netherlands

Abstract

For the computation of the Consumer Price Index (CPI) and the Harmonised Index of Consumer Prices (HICP) in the Netherlands, prices of consumer goods and services are acquired by using both traditional as well as innovative methods of data collection. CAPI and CAWI are considered as traditional methods, whereas scanner data and internet web scrapers are innovative methods. The introduction of innovative methods brings along new challenges that ask for new strategic solutions.

This paper reports on the introduction of web scrapers for the collection of internet prices and metadata for the computation of the CPI and HICP. The use of internet data in the Dutch CPI in the current and near future will be briefly explained. The core values of internet data and the challenges that come with them will be addressed. We will discuss organisational, technological and legal issues and explain the concrete solutions that have been taken in the CPI department. We conclude with a discussion on the opportunities and challenges for the application of web scrapers in other fields of official statistics.

[1] We would like to thank Leo Peeters for his editorial and linguistic contribution to this paper.

I. Introduction

1. The consumer price index (CPI) measures changes in the price levels of consumer goods and services purchased by households compared to a determined reference year. The CPI is computed by measuring the prices of a sample of representative items of which prices are collected periodically; in the Netherlands monthly. The representative items are grouped into different categories with weights that reflect their shares in total consumer expenditures.. European regulations (HICP) and national Dutch regulations (CPI) determine the categories of products that are used for computation within the context of the CPI and HICP.

2. Traditionally, CAPI is an important way of collecting data for the CPI. Trained interviewers visit the same shops each month to collect the prices of a list of predefined articles. Using scanner data is one of the methods used to modernize price collection. Another is collecting prices from the internet automatically using web scrapers. Apart from improving both the quality and the efficiency of price collection, it reduces the burden on our data suppliers, the retailers.

3. Internet data collection by web scraping is not new in official statistics. Research examples within Statistics Netherlands are for example gathering flight information from airline companies' websites (Hoekstra, ten Bosch and Harteveld, 2012), house prices from Dutch property websites (ten Bosch and Windmeijer, 2014), social media information for consumer confidence (Daas and Puts, 2014) and Google trends for health statistics (Reep and Buelens, 2013). A well-known example of the use of internet for CPI purposes is the Billion Prices Project (BPP) of MIT. In this project the websites of different retailers are scraped in order to make a price index for different countries, in particular some countries in Latin America (Cavallo, 2012).

4. The internet is irrefutably a Big Data source. Big Data is often characterised by the three v's for volume, variety and velocity. It is enormous in volume and the data grow continuously. It is semi-structured, because internet standards allow a large degree of freedom to design and implement a web page. Internet data can be collected with more detail and with a much higher velocity than traditional data collection and are available much earlier for statistical production. For instance, one could compute an hourly price index for airline tickets.

5. In the Netherlands the CPI team uses two different methods to collect data from the internet: (i) data from web sites with many similar items, e.g. prices and characteristics of garments from clothing web shops, and (ii) data from websites with only a few items, e.g. the prices of cinema tickets from cinema websites. For the first type of data we use advanced internet robots that run weekly or even daily without user interaction to collect a large number of observations with each run. We call these web scrapers 'bulk scrapers'. In Statistics Netherlands the bulk scrapers are only used to measure the prices of garments. In 2016 three web scrapers are currently used to collect data for the monthly production of the CPI. The web scraper used for the first retailer was taken into production in January 2015. Using scrapers for data from two other retailers went into production in January 2016. Additionally, sixteen bulk scrapers are collecting data from other retailers. Their data are ready for analyses followed by validation before being adopted for production.

6. For the second type of data we developed a so called 'internet robot tool'. With this tool you can mark a price on a website and its meta data and save the navigation path to the location on the internet. Once marked, the tool indicates whether the price, its meta data or the path to the information has changed since the previous visit to the website. The internet robot tool was developed because it would be too expensive to build dedicated robots for every single website and use only a limited number of prices. The internet robot tool is used to collect data for a variety of products like for instance cinema tickets, driving lessons and pizza delivery services.

7. In this paper, more than four years of experience in using internet data for the Dutch CPI will be discussed. In Section II we summarize the core values of using internet data. In Section III we discuss the challenges that come along with the new data. We did not aim to be exhaustive, but we highlighted the issues that can be interesting for other fields of official statistics and in other national statistical institutes. In Section IV we will address solutions for the challenges mentioned in Section III, focusing on organisational solutions.

Finally, Section V is a discussion on how the values, challenges and solutions for the use of internet data in the CPI, could be applied to other areas of official statistics.

II. The core value of using internet data in the CPI

8. There are many reasons to consider internet as a data source for the CPI and HICP. Many of the prices collected by visiting shops, telephone surveys, or using electronic questionnaires can nowadays be found on the internet. Furthermore, they can be collected efficiently by internet robots. The huge amount of price and product information available on the internet, and the increasing speed of processing make it possible to develop new detailed and fast price indicators.

9. Internet is not just another communication channel, in many situations the internet is becoming the main, or even the only, communication channel. Consulting the internet for shopping, travel arrangements, hotel and restaurant reservations, car maintenance scheduling, job vacancies, second-hand goods, is increasing in the Netherlands. As internet is becoming a natural part of our daily lives, official statistics cannot afford to lag behind. Otherwise, official statistics might become biased towards traditional channels and the bias may increase every year.

10. Another important reason to use internet data is to reduce costs. Traditional data collection methods are very labour intensive as many interviewers have to be hired to collect data in the shops. The costs of the data collection are mostly determined by their wages. The wage of a single web scraper expert is higher than the wage of a single interviewer, but we need much more interviewers, making a strong business case for internet data collection. The most important reason for introducing web scrapers by Statistics Netherlands to collect prices for clothing is to reduce price collecting in clothing outlets by 50 per cent.

11. For the retailers the reduction of costs comes from diminishing the response burden. The employees in shops no longer have to provide price collectors with information on articles and prices. Instead the servers of the retailers website are burdened by the web scrapers of Statistics Netherlands. However, this burden can be minimized by building intelligent robots.

12. We can be confident that the data on retailers websites are of good quality. Websites are publicly accessible and made to attract a large group of potential consumers. Both the structure of a clothing website, which is a kind of a classification of goods, and the descriptions of the articles must have a high standard level of quality and look familiar to the consumer. If consumers cannot quickly find what they want, they will choose to visit a competitors web shop. Furthermore, the retailer has strong financial incentives to put the actual prices on the website. If the prices are lower than intended the company loses money. If they are higher, the company might lose clients that choose to buy their products from another web shop.

13. Studying internet data sources for official statistics allows national statistical institutes to observe, and in the end to understand, price patterns better than they can with traditional data collection. For example, traditionally the price of an airline ticket is collected manually several times a month. Using internet robots, the price may be observed more frequently: daily, hourly, or even every 10 minutes. Only by collecting data very frequently for a certain period, can we learn more about their volatility and thus decide what would be an optimal collection strategy. It looks like we can suddenly see in the dark, where before we could only touch and guess.

14. The introduction of the internet robot tool is another example of the use of internet in the CPI. This tool is used to collect a few prices on many websites. It compares the prices and meta data of the current month and the previous month. If these do not change, the data is saved automatically. In other words, if the information on the website is constant over time, no additional manual work is needed. Before, the data collector had to open all the websites and check the information. We have concluded that working with the internet robot tool for these specific websites cuts time for price collecting by 85 per cent compared to the prior situation.

III. Technical, organisational and legal challenges of using internet data in the CPI

15. Data collection from the internet and its subsequent processing for producing statistics requires new skills and reorganizing work processes after switching from the traditional ways of data collection. This applies to both the CPI teams consumption analysts who select the items for the CPI "basket", and the data collectors. The analysts using internet as a data source must have a comprehensive understanding of consumer and retailer behaviour on the internet and find statistical methods to measure these. In addition they need to know and to understand the technical aspects of the internet in order to scrape it. They need to understand how a HTML page is structured, what an URL is and what JavaScript is.

16. Automatic classification of articles is one of the most important challenges if we want to process internet data in an efficient way. There is simply too much data to classify these data manually. In CPI production we do still work with rule based systems: the design and maintenance of the rules is done manually. Thus for each new retailer a new set of rules has been designed. Furthermore, rules have to be maintained periodically because of changes in clothing assortments. In order to reduce this repeating manual work we need smarter rule sets, rules that apply to more retailers, and probably artificial intelligence systems like neural networks. The CPI department and the Methodology department within Statistics Netherlands work closely together on the development of methods for automatic classification.

17. The IT-experts who build the web scrapers must both have a general knowledge of the internet and its structure, and experiences with rather new IT-tools and programming languages. They must have programming skills in for instance Java, JavaScript and Python and experience with frameworks built with these programming languages. These are new tools not only for the statisticians, but also for the IT-experts in the national statistical institutes. Though it should not be a big effort to acquire the new skills for the IT-experts, it will still be a time consuming process.

18. Even if the building of the robots is done completely by the IT experts, we have concluded that it is necessary to have sufficient knowledge of programming languages for processing and analysing purposes within the CPI-team itself. Most statisticians master traditional tools for data analysis like Excel and Access. However, the use of big data make the processing and analysis more complex. IT-tools are required that facilitate better software engineering principles. We need reliable performance when data size is scaled up. The CPI-team has chosen Python and MS T-SQL as main IT-tools to process the data and carry out the analyses.

19. Next to the acquisition of new knowledge and skills both by IT-experts and CPI statisticians, we have to adapt work processes in order to process the new data sources in an efficient way. We give two examples of the organisational challenges that we encountered in the production process of the CPI-team. To introduce these challenges we begin with a history of internet data within the CPI team.

20. The first web scrapers were developed and maintained by researchers of the Research department. In the first years there was a limited number of web scrapers and these were used for research only and not for statistical production. The researchers had acquired sufficient knowledge of the CPI, because they used the collected data also for their own research. The positive side effect that this had was that robot failures, often due to website changes, were quickly fixed. After a few years the research project went into production, with many web scrapers running in a production cycle and the original researchers left the project. Thus the question arose: 'Who is responsible for web scraper development and maintenance for production?' The IT-department and Research department had the knowledge, but are traditionally not part of the daily production process. The Data Collection department and the CPI team are experienced with data collection for production, but they had insufficient resources and knowledge.

21. A related problem is the validation of the internet data. Initially, in the research phase, researchers from the Research Department validated the data at all processing stages from the data collection to computing price indices. Later on, when they were only responsible for the data collection stage, they only checked for obvious errors: a suspicious file size or missing variables. More subtle cases such as changes in product names or product descriptions were overlooked. These kind of incidents were detected by the CPI department later in the process. Sometimes even too late, when the CPI Department were ready to use the web scrapers data for

production. An obvious problem inherent to using internet data, is that you cannot go back in time to recollect data if something in the process appears to have gone wrong. So you have to be aware if the data is inadequate as soon as possible, in order to collect the missing data in the next day or the next week even though these data correspond to a later period.

22. Forming a Development and Operations (DevOps) team can be useful in overcoming these organisational challenges. The concept of a DevOps-team stems from the IT-world and is meant to improve communication and to share knowledge between software development, IT-operations and the final customer. Statistics Netherlands formed a DevOps team for the development and maintenance of the web scrapers. The team consists of employees from the CPI-team and the IT-department. The CPI-employees pass on their knowledge of price indices to the IT-employees. They inform which retailers hold the biggest market shares and are therefore more important than others, and which data on a website is relevant for the CPI. The IT employees train the CPI employees the basic skills to repair and maintain the web scrapers. The members of the DevOps team are collectively responsible for keeping the robots running in production. IT experts and CPI team members work together on the basic maintenance of the robots, though IT-experts are responsible for the more complex web scraping problems.

23. To tackle the problem of data validation one person within the team is specialised in monitoring the data. She is familiar with the web scraper data quality indicators and the indicators of the first stages of CPI-production. We have implemented a number of indicators to detect the changes on a website. The number of indicators will enhance in the coming years as new types of problems appear. The data analyses and subsequently designing new indicators will always be reactive: you are always one step behind the problem.

24. A last challenge is the legal issue. The Dutch statistical legislation does not (yet) make it compulsory to grant access to websites for statistical purposes. Consequently, the owner can block the web scrapers access to the website. This is however not very likely to happen, as websites are visited by many other web scrapers as well and web servers are built to handle this. We have often experienced that a sudden change of the website would be more likely to occur. These changes can lead to missing or inadequate data as mentioned earlier.

25. For the CPI we have investigated ways to prevent blocking the web sites for our web scrapers and how to anticipate website changes. We are working on professional relationship management with retailers and organisations that supply the data; including the internet retailers. When we contact them we inquire on their policies regarding their web sites and request them to inform us on any (large) website changes.

IV. More challenges and critical success factors

26. Technical, organisational en legal solutions are not sufficient to make a successful transformation from traditional data collection to the use of web scrapers. We also need qualities like a shared vision, a common urgency, enthusiasm and courage. Furthermore, flexible human resource solutions can help to make the transition in a relatively short period.

27. If urgency and enthusiasm are shared between management and statisticians, the transition to modernizing data collection can be done more easily. It was very useful to discuss the opportunities and challenges of internet data already in an early stage. The new opportunities and challenges of internet data collection were obvious to all of our CPI statisticians.

28. The CPI statisticians are aware of the fact that it is impossible to maintain traditional data collection without loss of quantity in a period of shrinking budgets, let alone to expand it. Internet data collection makes it possible to enhance the amount of data collection (and improve the quality of the CPI), without high additional costs. It gives them new challenges to solve, but also new opportunities to improve statistics. We notice that our CPI statisticians keep coming with new ideas for internet data collection by web scrapers. We also notice that they are very enthusiastic working with the new data and learning new skills to master them.

29. Another important aspect is courage. Doing research and building pilot systems for web scraping is fun. However, taking the web scrapers into production is much more challenging. We made the challenge controllable by taking small steps. In 2015 the first robot was taken into production for the CPI. In 2016 two more robots followed. We have now reached the point that many more robots can succeed. They will completely replace the traditional price collection for clothing in shops by 2018. Of course we will make the necessary preparations and plans and adapt the processes in time to keep the transition under control, but in the end we'll 'just do it'.

30. We have taken several human resource (HR) measures in order to acquire the necessary knowledge and IT competences in a relatively short period. The first one is hiring new specialists prepared with the needed knowledge and IT skills, or who are capable of acquiring them within a short period. Due to shrinking budgets, statistical institutes are limited in hiring new personnel, so we invited skilled and motivated people already actively working within Statistics Netherlands to apply for the web scraping and data processing vacancies.

31. Additionally, extra training of current personnel to keep their knowledge and skills up to date is very important. We trained our CPI statisticians in programming languages like SQL and Python. A small expert group followed external courses to enhance their knowledge. Afterwards they shared their knowledge and experience and trained the beginners. Many of our CPI statisticians also took part in a pilot within the Economic Division of Statistics Netherlands to follow online courses from a professional online training centre. The training centre organizes feedback from teachers and the students can obtain certificates.

32. Next, we extended our cooperation with other departments within Statistics Netherlands and with external partners. The CPI team works closely together with experts from the Research department on the modernisation of data collection methods and new methods for computing price indices. We also work with universities and colleges, inter alia, by offering internships to their students. Students can acquire work experience, while they can offer modern knowledge and IT-skills. Statistics Netherlands employed some interns to collaborate with new graduates in developing the use of machine learning for automatic classification.

V. Conclusions and discussion

33. The use of internet web scrapers is a very good alternative for traditional price collection for the CPI. Though web scrapers are not (yet) applicable for all product categories, they are applicable for large parts of the consumer market. Internet is becoming an important sales channel and in some cases the main sales channel for consumer goods and services. The number of web sites that offer consumer products is growing every day. The prices on the websites are reliable. The costs of price collection using web scrapers are lower than the costs for traditional price collection. Finally, we can collect much more prices and therefore make a CPI of higher quality. Statistics Netherlands took the first web scraper into production for the CPI in 2015. We aim to end the traditional price collection for garments in 2018.

34. Can the core values of internet data collection that were mentioned in Section II also be applied to other official statistics? The value of internet as a growing communication channel is universal. Internet is an important medium and it's use will only grow in the coming years. As internet is becoming a natural part of our daily lives, official statistics cannot afford to lag behind. Not using internet as a data source in official statistics might not only cause a bias in data sources but also in statistical publications.

35. The advantage of cost reduction can of course only be determined by comparing the costs of the present statistical data collection to the costs and possibilities of collecting data from the internet. The business case for internet data will probably grow stronger over the years, as experience, knowledge and tools can also apply to other statistics. The quality of the data depends on the interest of the web owner. If they are for commercial purposes, as is the case of clothing retailers, the quality is probably high. Internet data will certainly lead to new opportunities, see the examples of the consumer confidence (Daas and Puts, 2014) and health statistics (Reep en Buelens, 2013) mentioned in the Introduction (Section 1). However, the quality of the final statistics may be hard to compare with the original statistics due to enormous methodological differences.

For internet data sources like Twitter (Daas and Puts, 2014) the increase in quality is not obvious, because of a biased selection of the sample space.

36. The challenge of acquiring up-to-data knowledge and modern IT-skills is relevant for all fields of official statistics. The impact of the organisational challenges of course depends on the current state of knowledge and the actual organisation of statistical production. In many statistical institutes, Research and IT-knowledge are formally separated from the statistical divisions. If web scrapers are introduced to collect statistical data, the statistical divisions will be challenged to keep the web scrapers running in production. The challenge of data validation also depends on the current organisation of statistical production. If one unit is responsible for the validation at all production stages, there will be no problem. If there are separate units for data collection and data validation, the data validation challenge must be solved.

37. The idea of a DevOps team, a team that breaks with the traditional organisational boundaries, is possibly an idea that can be applied elsewhere. One of the advantages of relationship management is, that it's sometimes easier to ask for information that already exists, than gathering it yourself using complex analyses. Probably, also not a new idea, but useful to repeat. For all kinds of solutions enthusiasm is a key word. If people are enthusiastic about the new data they will gladly accept training to master them and they will gladly accept a new role in a DevOps team or a relationship management team. Naturally, we still have challenges ahead of us, enthusiasm and courage will keep us going.

38. The legal challenge is universal. The websites are not owned by the statistical institutes, but by other parties.

39. The demand for employees with new knowledge and up-to-data skills is a challenge, but not new. Most solutions of acquiring new employees, training the current employees and collaboration with internal and external partners are probably not new either. The training with online courses probably is new.

References

ten Bosch, O. and Windmeijer, D. (2014), "On the Use of Internet Robots for Official Statistics", Paper presented at the Meeting on the Management of Statistical Information Systems (MSIS 2014), 14-16 April, Dublin, Ireland and Manila, Philippines.

Cavallo, R. (2012), "Online and Official Price Indexes: Measuring Argentina's Inflation", *Journal of Monetary Economics*, online version, 25 October 2012.

Daas, P.J.H. and Puts, M.J.H. (2014), "Social Media Sentiment and Consumer Confidence", European Central Bank Statistics Paper Series No. 5, Frankfurt, Germany.

Hoekstra, R., ten Bosch, O. and Harteveld, F. (2012), "Automated Data Collection from Web Sources for Official Statistics: First Experiences", *Statistical Journal of the IAOS* 28, 99-111.

Reep, C. en Buelens, B. (2013), "Mogelijkheden van Google zoekgedrag als verrijking van de Gezondheidsstatistieken, Internal report, Statistics Netherlands, Heerlen, The Netherlands.