UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE (ECE) CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting of the Group of Experts on Consumer Price Indices (2016)
<Geneva, 2-4 May>

Session 2: Big Data

ON THE USE OF INTERNET DATA FOR THE DUTCH CPI

Invited Paper

Prepared by Robert Griffioen and Olav ten Bosch, Statistics Netherlands, the Netherlands

Abstract

In the Dutch CPI there are currently two ways in which Internet data is used as a data source for statistics. Firstly, we use a so-called “robot tool” that extracts specific targets from several websites. Secondly, we use web scrapers that collect bulk data, 1,000 to 100,000 records, from about 10 specific websites. In this paper, we briefly explain both types of internet data collection and the importance of the use of Internet data in general. We describe how we CPI department currently deals with the processing of the bulk or big data and describe the plans to scale up the use of the bulk data even further. We explain some of the difficulties in production and the challenges we expect with a scale up as is planned. Finally, we discuss possible solutions to current and future challenges.

1. Introduction

The internet is irrefutable a Big Data source. Big Data is often characterised by the three v's of volume, variety and velocity. It is very large in volume and growing continuously. It is semi-structured, because internet standards allow for a large degree of freedom to design and implement a web page. Internet data can be collected with a much higher velocity than data collected from shops and are therefore much faster available for statistical production. One could make for instance an hourly price index.

Internet has already been used as a source for Official Statistics. Examples at Statistics Netherlands are the use of flight information from websites of airlines (Hoekstra, Bosch and Harteveld, 2012), house prices from Dutch property websites (Bosch and Windmeijer, 2014), social media information for consumer confidence (Daas and Puts, 2014) and Google trends for health statistics (Reep en Buelens, 2013). One of the early projects in the domain of prices is the Billion prices project (BPP) of MIT that use the websites of different retailers to make a price index of different countries, in particular some Latin American countries (Cavallo, 2012).

The BPP started its data collection in 2007, which is almost simultaneously when Statistics Netherlands began developing ideas of how to use the Internet for statistics (Daas, Roos en Puts, 2008).The first concrete experiment started in 2009 with the collection of data of airline tickets (Hoekstra, Bosch and Harteveld, 2012). The use of the Internet as a source for the CPI has started at Statistics Netherlands in 2012 (Bakker, Griffioen en Willenborg, 2012). It is unclear to what extent the BPP follows international concepts, regulations and formats. As a consequence there could be differences in demands to the sample, representativeness of the web scrapers, the collected data and meta data, and the price index method (the calculation of the partial indices and their aggregation). For instance, the BBP uses one large retailer as representative for the clothing market (BPP, 2013), while the Dutch CPI plans to use as many as 20 to 25 retailers to cover the clothing market as fully as possible.

In this paper, we will discuss in much more detail the use of internet data in the Dutch CPI. The aim of this paper is to give insight in how the Dutch CPI is currently and in the near future collecting and processing Internet data. In a previous paper we discussed the technique of web scraping and gave detailed information on what we collect from a retailer's website (Griffioen, Haan and Willenborg, 2014). In this paper the emphasis is on what the Dutch CPI does with the data. For instance, we will address issues like how Internet data are processed to a price index and how data are validated during production. These and other issues will be addressed for the current situation of the Dutch CPI in Section 3. In Section 4, we will address the same issues for the future situation. First, in Section 2, we will start with an introduction of web scraping. In Section 5, we will conclude with a discussion of the relation between the future situation of the Dutch CPI and Big Data. The same issues as were discussed in the previous sections will be addressed, but now from a Big Data perspective.

2. Internet data as a source for the Dutch CPI: Why, how and what?

There are many reasons to study the Internet as a data source for statistics. We mention a few of them. A more elaborate version can be found in ten Bosch and Windmeijer (2014). First of all, data collection via the Internet reduces the response burden of outlets. Secondly, it increases efficiency of a Statistical Office by reducing the cost: data collection via shops is more expensive. Efficiency within a Statistical Office further increases by the automatic processing of the data.

One of the things we have learned with web scraping is that it is good to distinguish between two quite different types of use: (i) collection of data from internet sites with many items, e.g. prices and characteristics of TV sets in electronic web shops, and (ii) collection of data from internet many data sources with only few items, e.g. the price of a cinema ticket from a cinema website. The first case can be approached with advanced internet robots that run without user interaction weekly or even daily to collect larger portions of observations at each run. For the latter category it would become too expensive to write dedicated robots for every single website and quite a different approach was chosen.

For the second case we developed what we call "the internet robot tool". With this tool you can mark a price on a website and its meta data and save the navigation path from the home page to the page with that particular product information. Once marked, every time the data collector collects its data, the tool indicates whether the price, its meta data or the path to the information has changed. The Dutch CPI uses this tool in cases where only few prices per site have to be observed, such as the price of a cinema ticket. In that case

twenty to thirty prices from different websites are collected. Since the price and relevant meta data does not change very often and because the collection process is a few clicks away much time can be saved.

For the first case, the collection of bulk data, the research department of Statistics Netherlands developed a so called “robot framework”. This framework uses the fact that websites of clothing retailers (and possibly other webshops) have similarities in structure, such as menu’s, pagination, product overview pages and detail pages and product metadata such as size, color, type and category. The robot framework uses these similarities to ease the configuration of scraping. But websites also have differences, for example the navigation from main page to the product pages. To handle the differences between the websites one needs to configure the robot framework using a configuration file . To give an idea, a parameter in this configuration file is the main page of the retailer’s website. Also the position of the price and its meta data is configured this way.

At the moment we collect at least the following data from each retailer’s website using this robot framework:

- **ID**: unique internet/web address of the item.
- **Type**: information on whether the item belongs to the women’s department, men’s department or children’s department. There is no overlap between the three departments.
- **Name**: name of the item as displayed on the website. It is not necessarily a unique identifier.
- **Short description**: description used for classifying the items.[1]
- **Price**: ‘offer price’ of the item as displayed on the website.

3. Current processing of internet data

3.1 Numbers and Facts

At the moment the CPI of Statistics Netherlands uses the Internet data of three retailers for production. In this case each retailer is ‘scraped’ by an Internet robot. The production of one retailer started at the first of January 2015. The other two have been taken in production on the first of January of 2016 . Furthermore, there are 16 internet robots collecting data from 19 other retailers. Their data are waiting for analysis and validation before using them for production. We will tell more about this in the next section 4.1.

All internet robots so far scrape clothing retailers are used for clothing purposes. There are two main reasons for this. The first one is a cost reason: Statistics Netherlands wants to reduce the data collection in shops as much as possible, because this is very costly and a large part of data collections in shops was for clothing purposes. One should keep in mind that data collection from the internet is not for free, for instance only think of the maintenance of the internet robots in the context of the dynamic web. However, compared to data collection in shops it is still cheaper. The second reason is a methodological reason, clothing is a good that rarely needs quality corrections, making it easier to automate the process.

3.2 The method: from prices of the internet to price indices

A lowest Coicop (LCoicop) index $lCI_{g,y,m}$, for instance the men’s clothing index, of a retailer *r* for month *m* of year *y* is calculated according to the following method.

$$lCI^r_{G,y,m} = \sum_{g \in G} w^r_{g,m} I^r_{g,y,m} \; and \;\; I^r_{g,y,m} = \frac{P^r_{g,y,m}}{P^r_{g,y-1,12}} \tag{1}$$

where $\sum_{g \in G} w^c_{g,m} = 1$. It is a direct method, with December (month 12) of the previous year *y-1* as a base month, where a $lCI_{g,y,m}$ index is a weighted sum of elementary indices $I_{g,y,m}$ for each product type or group *g* of all groups *G* (Willenborg and Bakker, 2013). The weight of each group *g* for month *m* is represented by $w_{g,m}$.

Each elementary index $I_{g,y,m}$ represents a clothing type like a dress or jeans trousers. The statistical group is broadly defined such that new appearing articles in a year can be incorporated in the calculation. All clothing types together form a clothing classification (Willenborg & van Hooff, 2013). Willenborg and van Hooff derived this classification from the clothing descriptions present on the internet of the first clothing retailer we took in production. However, with a few exceptions this clothing classification also fitted the other two retailers we took

[1] Note that price indexes for clothing are only published at the level of departments, which is equivalent to the three-digit COICOP level.

in production this year. Another reason why we came up with this classification is that at this level of aggregation the CPI possesses weight information.

To classify goods from the internet we made a system of classification rules. An example of such a rule for jeans trousers is “jeans and not jackets”. Thus the process of classification is automated. However, the maintenance of classification rules is not. They are manually maintained over the year. For instance, a CPI production employee is looking for the pattern of declining number of items of a certain product group and increasing number of items in another group (this is a misclassification) or a pattern of increasing number of items that could not be classified. Depending on the pattern the employee adapts the classification rule. Note that maintenance thus consists of a detection and correction process.

The price of each group $P_{g,y,m}$ is the arithmetic mean of all prices that are collected of the retailer during a month, so $P_{g,y,m}$ is:

$$P^r_{g,y,m} = \frac{1}{n^r_{g,y,m}} \sum_{d \in m} \sum_{a \in g} p^r_{a,y,m,d} \qquad (2)$$

Here (lowercase) *p* is a price of a product found on the website of the retailer *r* with a unique webID that fits the description of product group *g*. The number *n* is the number of prices found on the website during month m: that is the sum of days *d* of the month of the sum of all daily article prices. For practical reasons we deduplicate the daily data, so in effect only one price per article per day is used in the calculation.

The weights $w_{g,m}$ of equation (1) are based upon revenue or sales volume from other sources (than internet). These contain the revenue of all groups *g*, but the grouping from the external source can be different from the internal clothing classification. Thus some handwork is needed to tune the external classification to the internal one.

One of the motivations to use the above unit value method was the wish to make fully use of the available data. But we will not address the choice of method further in this paper. Once an unit value or group method is used, one of the most determining quality factors of the index is the homogeneity of the groups. This is for example also true for the method of the Billion Prices Project, though they are using a chain index with geometric mean (Cavallo, 2012). The group’s average is not representative if it contains non-homogeneous items. To guarantee the group’s homogeneity as much as possible we have chosen retailers that have one or a few brands of which items belong to same market segment. Furthermore, suspicious items can be excluded by the above described classification rules.

To give an idea about the indices produced with this method we made two figures to compare the data collected in stores and the Internet data. The data differs and the method differs. The data of shop collection includes also the retailers of the Internet data, but this is their shop data collection. Figure 1 is the data of the men’s clothing. Figure 2 is the data of the women’s clothing. In both figures the red line represents the data collected in stores “Stores Clothing General”: all items of all retailers of men, women and children of the shop data collection together. The green one is of Internet Retailer 1, the retailer that we have for one year in production. The blue one is of Internet retailer 2, one of retailers that we took this year in production.

Figure 1. Comparison Stores versus Internet: Men’s clothing

Figure 2. Comparison Stores versus Internet: Ladies clothing

At first sight it seems that the price indices produced with internet data are inferior, because for instance there seasonal pattern is flatter than the price index produced with data collected from stores. This is true if you assume that the latter represents the truth or is closest to reality. However, our clothing market expert often receives questions from retailers why the prices index still has this traditional strong seasonal pattern, because according to them it should be more flat. The reason for this is that they have more than two times a year a new collection. Furthermore, they keep their collections small to be certain to sell their goods. As a consequence a big sale with many low prices is unnecessary. Given this information, it is seems that a price index based on Internet data is at least of similar quality as an index based on data collection of outlets.

3.3 Current practice of the monthly CPI validation

The validation of the monthly CPI indices that are produced with the internet data is carried out in four different stages. The first validation stage is just after the internet robot collected the data from a website. The data collection automatically triggers a process that produces indicators like file size, number of price observations, number of collected categories, number of warnings, errors et cetera. These indicators are automatically sent to the members of the robot team for inspection. If something went wrong the robot is manually restarted. For some of the robots there is an automatic restart if some indicators pass a threshold.

The second stage at which the data is validated is just after the data enters the CPI production process. Some additional indicators such as the totals after deduplication are produced and these indicators are plotted for a time period of a year or more. Such charts make it easy to spot anomalies spotted by eye. This process is carried out by the CPI team that does the first checks of the 'big' scanner and internet data. If approved, they load the data into the CPI production systems.

The third stage at which the data is validated is when the data has been classified and elementary indices have been produced. This stage is executed by what CPI at Statistics Netherlands calls the consumer analyst: a market expert for different types of goods like for instance clothing. The consumer analyst investigates patterns at the group level as for instance the increased or decreased number of items in a group such as described in the previous Section 3.3.

The fourth and last stage of the validation process at the last stage of production in which the complete CPI is produced. This is performed by what the CPI of Statistics Netherlands calls the Calculation and Validation role. This person checks all data top down for suspicious values. For the data collected in stores these people have an analysis tool that can 'down drill' from first and second digit Indices to the micro data of a data collector.

4. The CPI at 2018

4.1 Numbers and facts

The CPI of Statistics Netherlands is executing an innovation program until 2018 to renew its data collection. One of its targets is to take the 18 internet robots mentioned in Section 3.1 in production. To fully stop the data collection in clothing stores, or at least reduce it as much as possible, even more internet robots may be necessary. One of the research programs this year is to find out whether and how we can cover the full clothing market with internet data collection.

It is unlikely that the Dutch CPI will go for internet data of other product types like books or electronics, because scanner data are the preferred data source. Another more likely reason for collecting data from the Internet is to obtain meta-data or product information additionally to scanner data.

4.2 The method

There are two challenges in the future that will affect the current method described in Section 3.2. First, as was mentioned before, the classification process is automatic, but the maintenance of the classification rules is not. For a limited number of three retailers this is still feasible, but, the maintenance of the classification rules for 18 or more retailers could become laborious. Our experience until now is that during a year descriptions change on purpose for example for marketing reasons and because of mistakes like typo's, or other reasons.

Therefore, the research department of Statistics Netherlands is investigating ways to automatize the maintenance of the classification rules, for example using machine learning algorithms. Not only can this be done using advanced text mining techniques, research has started to use the pictures of clothing products as well, as these may contain useful additional information. Although there is lot of research going on in this area, the expectation for the future is that there will always be some manual work left.

The second methodological challenge is the fact that not all data is homogeneous. At this moment we possess Internet data of retailers of whom the different clothing product types are not so homogeneous, partly because of the many different brands in their collection. To solve this problem it might be necessary to adapt the method that we discussed in Section 3.2. Another possibility is to apply other methods like methods of the ILO manual (ILO et al., 2004) or a new method like the QU-method (Chessa, 2016). The application of QU-method or other scanner data methods is not straightforward, for instance because we miss weight information at the lowest item level.

If we apply more sophisticated methods, like QU or hedonic methods, it is possible that much more additional information has to be extracted: more product characteristics. A drawback or greater challenge is that we also need more sophisticated automatic extraction and classification techniques, because it is impossible to do this manually. The greater challenge resides in the fact that often this additional product information is present in unstructured text parts.

Our future choice of a price index method for Internet data will depend on the quality of the index and the amount of work it will cost to maintain it in production. The latter is dependent on the successfulness of research, in particular the research involving automatic classification.

4.3 Future practice of the monthly CPI validation

In Section 3.3 we discussed the current situation of how the internet data is validated in the monthly production process. We distinguished four different stages at which data are validated. In the current situation (2016) the first and second stage of validation is carried out by a so called development and operations team (DevOps team). This team consists of researchers, IT people and CPI employees. The idea is that the number of researchers involved in the team will decrease as the work is changing from very experimental to proven technology. Right now the advantage of such a team is that the knowledge and experience is multi-disciplinary. In addition, the CPI production department knows exactly the status of the collection process so that they can anticipate on changes that may affect the CPI production as soon as possible.

The DevOps team is still under construction, but the first results are promising. The methods of data validation will be further improved over time. Below we present some of the ideas for further improvements.

The first idea is that data validation has to be as automated as possible. For instance, for all known errors we want to make automatic detection methods. If possible, these errors also have to be automatically corrected. One can think of one batch with automatic detection and correction methods for all different retailers. So if one finds a new error in one particular data set, the detection and correction method will be put in the generic batch such that it will automatically be applied to all other data sets. What has to be considered is the tradeoff between genericity and performance of the batch.

The second idea is not to wait with the production until the third week, but produce it after the first week. Traditionally, the Dutch CPI uses data of three weeks to produce the price index. In this case, with data collected from stores only after two weeks a sufficient amount of data was received to make a sensible index. However, because we are working with large data sets now there is a sufficient amount of data to make an index after a week. This will give the analyst the opportunity to see in what direction the index is heading.

The third idea is that manually human detection should be top down. This does not necessarily have to start at the highest aggregation level, but could start at a lower aggregation level, for example the starting point of analysis with clothing could the general clothing level. That is why we call this kind of analysis the meso-

analysis. Manual inspection or an outlier should only occur when its value falls outside some boundary value. In other words, only investigate in case errors are visible at some higher aggregated level. This meso-analysis has to be supported by a system that can detect and mark these suspicious values. Furthermore, the system has to facilitate 'down drilling': a system that can zoom in from a high aggregation level to lower levels of aggregation, preferable down to the lowest micro data level.

The fourth idea is that the automated system should not only detect and correct the errors, but also generate process information about it. So it can later be determined what type of error occurred, when it occurred and how many times it occurred. This will give us insight into to data and with it we could for instance find the correct values for the above mentioned boundary values to mark suspicious values at a given level of aggregation. Since information on the web is dynamic this process information will give us insight about these dynamics.

The fifth and final idea is that some errors that cannot be automatically corrected and are also not manually corrected during the monthly validation process because their impact is too low, are periodically repaired. The reason for this is that these errors can have a significant effect on the price index in the long run, but to wait for that moment can generate a lot of work and stress just before the monthly production. Examples of these errors are classification errors like t-shirts of which some are spelled as 'tshirt'. First only a few t-shirts are misspelled, however it could be that in the end all are misspelled. Of course this is a known error that we are able to automatically fix, but there will always be unexpected classification errors.

5. Discussion: The relation of the current and future situation of the Dutch CPI and Big Data

In this paper we described the way the Dutch CPI department collects and uses Internet data for production purposes. We described the current and near future situation. In this chapter we describe the future situation of the Dutch CPI with respect to Big Data principles such as volume, velocity, variety and selectivity. The latter is an issue that according to Buelens et al. (2014) has to be addressed when using Big Data.

Volume is an important determiner of how to deal with Internet data. Currently, we already have a highly automated process to handle the large volume of price data. However, the maintenance of classification rules is still handwork. In the future situation, where we further grow into 20 Internet robots and approximately 500.000 records a day, we have to do this differently. Therefore, the production, the rule maintenance and the analysis has to be as automated as possible (Section 4.3). Furthermore, to not lose grip on the data, top-down analysis methods have to be implemented. Manual data inspection should only occur in case some indicators on the meso level indicate some serious errors, otherwise the manual work will soon outgrow us. This is only possible when the process is fully automated, and advanced methods such as machine learning replace the manual work. In addition, we need a system to perform these meso analyses. Another prerequisite to handle the large volumes of data is to produce process statistics (Section 4.3). This could be used for further automation or as additional variables that are used during meso analysis.

The velocity with which Internet data can be collected is much higher than the data collection of stores. We are not considering a daily statistic, but we plan to use the possibility of higher velocity for data validation purposes by making provisional weekly statistics. This idea is not only restricted to the Internet data, but could also be applied to scanner data.

The variety of Internet data expresses itself in the fact that we have two types of Internet robots (Section 2). If we restrict ourselves to the bulk scraping, the principles implemented in the robot framework shows that websites of clothing retailers have a lot in common. It also shows that in spite of these commonalities there are always differences between the sites. Furthermore, there are differences between the collected data of the different retailer websites. This calls for some variety in the production processes of the data from the different retailers. If we make use of the advanced prices index methods mentioned in Section 4.2, we have to make use of extended clothing descriptions. It is here where the highest variability resides between the retailer websites and is a great challenge to automatize this.

Selectivity in the words of Buelens et al (2014) is that Big data is generated in a different way than data generated by probability sampling theory. The Big Data is often only a selection of the population that the statistic aims to target and is therefore not representative. In our case we aim to target the population of all clothing. Traditionally and in the present case of Internet data there is no clothing sample frame. Clothing is just like other product types of the CPI selected by cut-off selection (Haan, Opperdoes and Schut, 1999). In case of clothing it is a bit more complicated. Clothing selection happens via retailer selection. In theory the retailers

with the highest revenues are chosen first and from these retailers clothing types with highest revenues are chosen. But in practice this is not completely true. We have this detailed information only for a few retailers and in some cases this information is outdated. However, if we compare the use of internet data with the traditional data collection, we see that the Internet case is not that bad. In fact, using Internet data is even better, because we now possess all data of a retailer instead of only a few items. So if we select the retailers that cover 80 percent of the market, we have a kind of register for the full 80 percent. Another implication of relevance for selectivity is that we use other external sources to determine the population and not only the Internet. Yet another important issue for selectivity is whether units can be identified. In both the traditional as well as the Internet case there are no clear units for the complete sample space, within a retailer the items can be clearly distinguished. Again, the Internet case could be more positive, because it may be possible to uniquely identify items over the different retailers via their pictures.

In conclusion, the future of the clothing part of the CPI is heavily influence by the ideas of Big Data, which is not a surprise since Internet data is Big Data. The quality of the 'Internet' indices are good compared to the indices produced with data collected from stores. Furthermore, much work has yet to be carried out to get all the new data into production. To get a grip on this data sets in production new algorithms and data analysis tools have to be developed. Most of the ideas of this paper also apply to scanner data. Since in the future scanner data and Internet data will be the main sources for the CPI, this paper sketches the future for a large part of the Dutch CPI.

6. 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.

Bakker, K., Griffioen, A.R. en Willenborg, L. (2012), "Gebruik Internetdata van een retailer voor CPI-berekening: Exploratie van mogelijkheden", Internal Report, May 2012, The Hague, The Netherlands.

Billion Prices Project (2013), "Global Retailers Data 2008 to 2013", http://bpp.mit.edu/datasets/.

Buelens, B., Daas, P., Burger, J., Puts M. and van den Brakel, J., (2014), "Selectivity of Big Data", Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

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

Chessa, A.G. (2016), "Processing scanner data in the Dutch CPI: A new methodology and first experiences",. Paper to be presented at the *Meeting of the Group of Experts on Consumer Price Indices*, 2-4 May 2016, Geneva, Switzerland.

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.

Griffioen, A.R., de Haan, J. and Willenborg, L. (2014), "Collecting clothing data from the Internet", Paper presented at Group of Experts on Consumer Price Indices, 26-28 May 2014, Geneva, Swiss.

Haan, de J., Opperdoes, E. and Schut, C. (1999), "Item selection in the Consumer Price Index: Cut-off versus probability sampling", Survey Methodology, June 1999, Vol. 25, No. 1, pp. 31-41, Statistics Canada, Catalogue No. 12-001.

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.

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004), *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO Publications.

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