Statistics Netherlands

Division PIM
Department of Methodology

The Hague
The Netherlands

Collecting clothing data from the Internet

Robert Griffioen, Jan de Haan and Leon Willenborg

Remarks:
The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. We are grateful to our colleague Karlijn Bakker with whom we collaborated in the web scraping project.

Project number: -
BPA number: -
Datum: *24 April 2014*

Collecting clothing data from the Internet

Summary: At Statistics Netherlands the question was raised whether or not data observed on the Internet, in particular on clothing, could be used for CPI purposes. Such data might replace data that is currently being observed by price collectors visiting the shops. They typically collect prices for a relatively small number of items. To answer the question raised, an 'Internet robot' was developed and set to work to collect clothing data from a particular web shop on a daily basis. This web scraping has been carried out during the past two years. The paper reports on our findings and discusses a number of issues we encountered.

Keywords: CPI, data collection, web shops, benefits, risks

1. Introduction

Over the past few years, Statistics Netherlands has been experimenting with the collection of prices from the Internet through web scraping or the use of Internet robots, as it is also referred to. Online prices could perhaps replace part of the prices observed by price collectors for the compilation of the Consumer Price Index (CPI).[1] Online prices could also replace data that is already being collected from the Internet in a less efficient manner. Apart from efficiency considerations, web scraping offers the possibility to monitor prices as frequently as desired, allowing the estimation of high-frequency price indexes.[2]

Importantly, data on quantities purchased cannot be observed via the Internet. The lack of quantity data is problematic for the construction of price indexes, but the problem is not new to statistical agencies. Weighting information at the item level is generally lacking (unless scanner data is available), and so the agencies are forced to construct unweighted indexes. For most products, the sample of narrowly defined items, or product specifications, is kept fixed, at least for some time, and the index is based on matched items to compare 'like with like'. When new items are introduced into the sample to replace disappearing items, quality-adjustments should be carried out in order to measure pure price change.

[1] Hoekstra, ten Bosch and Harteveld (2012) describe some first experiences with the use of web scraping software, which is part of a broader project at Statistics Netherlands on 'big data' (Daas et al., 2011).

[2] In the Billion Prices Project, a research initiative at MIT that uses online data to study high-frequency price dynamics and inflation, daily price indexes have been calculated for several countries around the world, including the Netherlands. For an example on data for Argentina, see Cavallo (2012). The indexes are currently compiled by PriceStats, a private company; see www.PriceStats.com. De Haan and Hendriks (2013) looked at different methods to construct high-frequency price indexes from online data, including the so-called time-product dummy method.

Item samples in the CPI have traditionally been relatively small, particularly to keep things manageable and control costs. A large part of the costs associated with the compilation a CPI stems from price collection at the stores. If web scraping turns out to be successful, the costs could be reduced substantially, even when observing all items displayed on the website instead of taking small samples. The costs could be reduced further if it were possible to develop an automated computer system without manual interventions to estimate price index numbers.

This paper reports on a pilot project for clothing. Given its weight in the CPI and the large number of price quotes that can be observed, web scraping has the potential of increasing efficiency, provided that the quality of online data and the resulting index numbers are sufficient. We decided to start off with investigating the website of a single retailer. Unlike online-only stores, this retailer also has many physical stores across the Netherlands.

Clothing is considered one of the ‘hard to measure’ goods in the CPI. For different approaches to the treatment of clothing in a CPI, see *Consumer Price Index Manual: Theory and Practice* (ILO et al., 2004). It is well known that matched-model price indexes for clothing will be severely downward biased since the prices of individual apparel items (as identified by item numbers or with web scraping possibly by web IDs) typically decrease over time.[3] An example using U.S. scanner data on women’s tops is given by Greenlees and McClelland (2010).

Greenlees and McClelland show that the downward bias in a matched-model index for women’s tops can be eliminated through the use of hedonic regression. However, collecting item characteristics on the Internet and processing them, either manually or by developing an automatic procedure, will be quite laborious as characteristics are embedded in product descriptions. Also, for many clothing items it may not be possible to observe all relevant characteristics, and in some cases it might actually be unclear what the relevant characteristics are.

Given these difficulties, the aim of our paper is not to compare different approaches to estimating clothing price indexes or to propose some preferred method. Although we do present tentative index numbers, we focus mainly on the data collection part and describe our experiences with collecting data from the Internet via web scraping. An important lesson is that the way the ‘robot’ is designed can have an effect on the data collected.

The paper is structured as follows. Section 2 discusses the web scraping process and the information we have collected. Section 3 presents evidence on the dynamics of the items observed on the website of the retailer under study. Section 4 describes the classification for clothing we developed and presents price indexes based on average prices at the lowest level of the classification. The use of websites as a data source is

[3] A problem with fashion goods such as clothing is that fashion itself can be regarded as a quality-determining feature, and part of the price decline could be attributed to deterioration in quality. But measuring these fashion effects is virtually impossible. Seasonality in itself is obviously another problem.

not without problems, and Section 5 discusses both methodological issues and issues related specifically to web scraping. Section 6 concludes the paper by summarizing the potential advantages and disadvantages of web scraping data for CPI purposes and by discussing our findings.

2. Data collection via web scraping

2.1 Introduction

For constructing a price index, data on both prices and quantities (or expenditures) is needed. The latter data is needed to aggregate the price relatives of individual items, or items, i.e. to weight the various price relatives. Compared with scanner data, an obvious drawback of web scraping is that weighting information cannot be observed on the Internet. What is also needed to construct price indexes is information on the characteristics of the individual items to classify the items into product categories and adjust for compositional change.

On the website of the retailer studied, referred to as “S”, a variety of characteristics is available about the items offered for sale. In particular, there are short product descriptions in alphanumerical form that can be used for matching and classification purposes. The website also displays pictures of all of the items. While photos are a rich source of information which, depending on the method chosen to compile price indexes, can potentially be very useful, it will be difficult if not impossible to extract characteristics information from photos in a fully automated way. The information on prices observed on the website of “S” is numerical, hence immediately readable and ready for use.

In what follows we will describe what data has actually been collected and how this was done.

2.2 Web scraping strategy

To understand how web scraping works, one has to know a little bit about hypertext mark-up language (HTML). An HTML page that a web browser shows is essentially a *tree of nodes*. Different nodes can have a different meaning and can have different content. Moreover, as there is no formal agreement on when to use what kind of tree structure, different HTML pages typically have different tree structures. In this sense the Internet is unstructured.

Web scraping for CPI purposes requires two types of data: *i*) data needed to compile price indexes, i.e. prices and characteristics, and *ii*) data needed to navigate through the website, i.e. to jump to the data of interest. A certain HTML page can contain both types of information. To extract the relevant information from the HTML page one can ‘query’[4] the tree to look at parts of it, to look at particular nodes, and to retrieve specific information from a node. To navigate through a website, the robot looks for URLs in web pages. For example, on the home page of the website of “S”,

[4] We ‘query’ an HTML page with xpath and regular expressions.

the robot searches for URLs that will direct it to the women's, men's and children's clothing departments.

There are usually many ways to navigate through a website in order to collect the required information. It is worthwhile to keep the robot lean and to try and minimize navigation, i.e. to visit as few web pages and use as little information as possible, for three reasons. First, this will make web scraping software robust against website changes. That is, it reduces the probability that the collected information is affected by changes made to the website. Second, it will keep the robot simple and thus make software administration easier. Third, it will reduce respondent burden: the less web pages are visited, the less web server requests are made.

Our strategy therefore was to collect data solely from the "view all" page of each department of the online store. This kept navigation to a minimum, and we expected that most, if not all, items would be observed in this manner.[5] However, at a later stage we decided to visit more web pages with the purpose of collecting additional item characteristics.

In section 2.3 we will describe what kind of information we have been collecting for each item. Note that the website is scraped on a daily basis. A daily frequency was chosen for research purposes. Statistics Netherlands has no intention of increasing the publication frequency for price indexes.

2.3 Scraped data

The following information has been extracted from the website of "S".

- **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.[6]
- **Price**: 'offer price' of the item as displayed on the website.

An important issue is whether item IDs are unique identifiers and stable over time (i.e. available during some time period). From information received from the retailer and looking at three months of data, we concluded that they are. We removed a few records that had identical item IDs but different item types or names, after correcting for language differences; sometimes the English rather than the Dutch name for the department was mentioned, which resulted in a lot of suspicious records. We also removed identical records pertaining to the same day.

[5] Our expectations were based on past experiences with housing market websites (ten Bosch and Windmeijer, 2014).

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

At a later stage of the project we decided to collect additional information on the items embedded in *long item descriptions,* including fabric used, care label, print, waist, etc., and also on item size and colour. This information might prove useful for choosing a proper index number method. For example, it could be useful to refine the classification (see Section 4.1) or to estimate hedonic models. The information contained in the long descriptions is semi-structured and in alphanumerical form, and to be able to use it we have to parse the descriptions and extract the details.

2.4 Some experiences

We began collecting data on a daily basis from the website of “S” on 31 January 2012. At present (by the end of April 2014), the data is still being gathered, so we now have daily data for more than two years. During this period we had to fix the robot about ten times because of website changes. In three instances, when a number of URLs required for navigation had changed, it took us up to two days to repair the robot. In the remaining cases the changes to the website were minor. For example, in some cases the data collected had different nodes than before.

It is interesting to note that one of the three major changes was not easily detected. The changes made to the website had as an unfortunate side effect that the first page of each department was scraped multiple times, apparently because the (incorrectly specified) URLs of other pages led to this first page. Even though the amount of data collected was monitored to check for malfunctioning of the robot, the error was not revealed. Later on, when we had a closer look at the data, we detected the error and removed the duplicates. This data collection error is clearly visible in the results, as will become clear in Section 3.

The above experience led us to build in standard checks during data collection rather than checking for errors afterwards. In case there are suspicious indicator values or many duplicate records, or if the number of items found in one of the departments is zero, the data collection process will restart automatically. This also saves us from manually restarting the data collection process when the connection fails, which has happened occasionally.

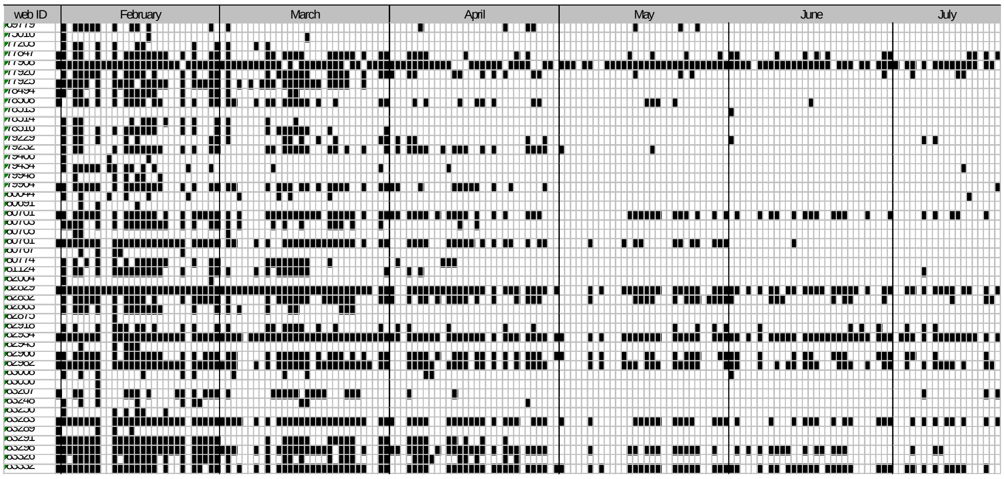
3. Results on items observed

3.1 Dynamics of the set of items

In this section we will describe the size and composition of the collection of items offered daily on the website of “S” and the dynamics of the collection. Having an idea of the dynamics involved can be important when designing an appropriate price index number method. In particular, we wanted to know the number of items that are new (‘births’), disappear (‘deaths’), change from ‘regular’ to ‘sales’, or continue to be available on the website without changing to ‘sales’ items.

It turned out that many items are temporarily unobserved. Table 1 shows an example of what happened between February and July 2012 for a selection of items. Only a few items were observed almost continuously, as indicated by the black rectangles, but most items were observed during a relatively small number of days.

Table 1. Selection of items observed daily by the robot


web ID
February
March
April
May
June
July

It may be that items were temporarily out of stock. It may also be that the website is dynamic in that a ‘disappeared’ item can still be purchased but has been temporarily replaced by a similar item, perhaps of a different colour. The latter reminds us that the way in which the robot navigates through the website can affect the set of items observed. Moreover, web scraping was done during a very short time period early in the morning of each day. So it is not necessarily true that through web scraping we observe the entire population of items that can be purchased.

Figure 1 shows the total number of clothing items on the website of “S” from 31 January 2012 to 28 February 2014, as observed daily by the robot, after cleaning the data (i.e., after removing suspicious values and duplicates).

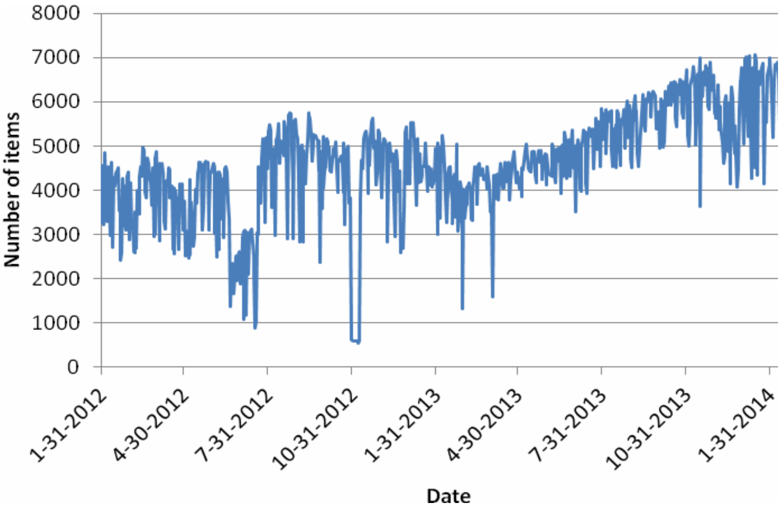


Figure 1. Daily number of observed items (after data cleaning)

The number of items observed during this period roughly increased from 4,000 to 6,000. The volatility is significant, with day-to-day changes sometimes being more than 60%. A few strong drops in the number of items were due to changes in the website, and which have led to erroneous data collection. The biggest drop is the one in mid-June 2012, which lasted for quite some time.

Every day new items are shown on the website. We identify the 'birth' of an item as the first observation by the robot. Initially, when web scraping was started, a newly observed item could have been shown on the website before so that it was not really 'born' on that day. But this problem only arises in the early stage of the observation period and then gradually vanishes. Figure 2 displays the daily number of 'births' during the same period as in Figure 1, except that the first few days have now been skipped.

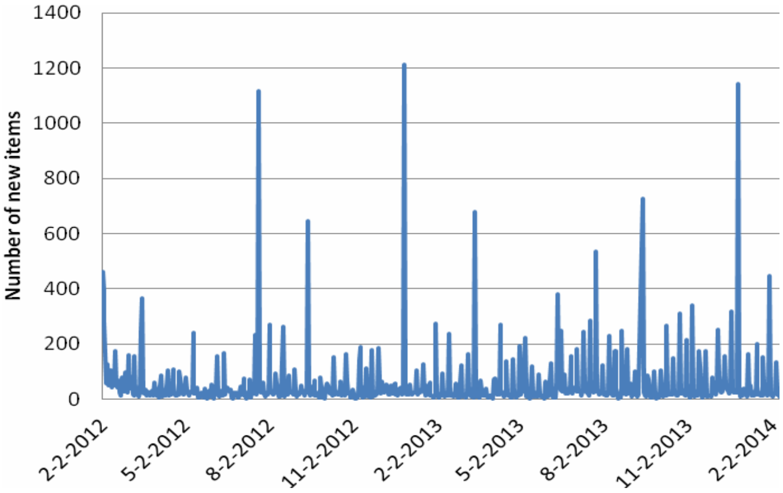


Figure 2. Daily number of 'births'

New items are constantly being introduced. There are several significant spikes in Figure 2, the largest ones on 21 July 2012, 27 December 2012 and 27 December 2014. The dates of most spikes are just after the period that many items went from the regular collection to the sales collection (See Section 3.2). This suggests that the existing collection was replaced for the greater part by a new collection. However, as we will see in Section 3.2, a fraction of the new items was immediately sold as 'sales items', without first being part of the regular collection.

An item 'dies' when it is permanently removed from the collection. Since items are often temporarily unobserved, it is not very useful to try and determine the 'death' of an item on a daily basis.

3.2 Regular and sales items

The website distinguishes between a *regular collection*, existing of items offered at 'regular' prices, and a *sales collection*, largely existing of items offered at reduced prices. Figure 3 shows the number of items that have changed from the regular to the sales collection. As we expected, the peaks during 11 June – 20 July 2012, 3 – 27 December 2012, 11 June – 15 July 2013 and 9 December 2013 – 7 January 2014 coincide with sale periods in the physical shops of "S". There are several other peaks in Figure 3, notably on 10 April 1012, 8 October 2012, 8 April 2013 and 7 October 2013, which are difficult to explain.

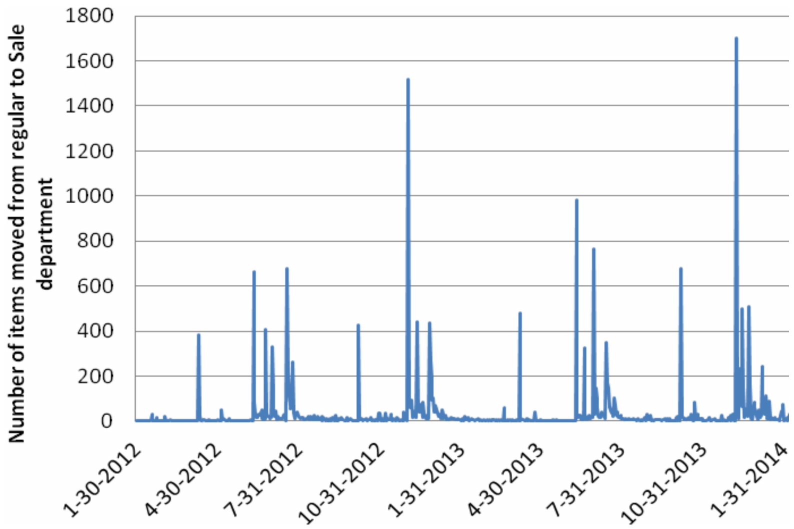


Figure 3. Daily number of 'regular items' that become 'sales items'

Figures 4 and 5 show the number of regular and sales items, respectively, during the two-year period for each of the three departments distinguished on the website of "S" (women's, men's and children's clothing). Note that the women's department is by far the largest in terms of the observed number of items.

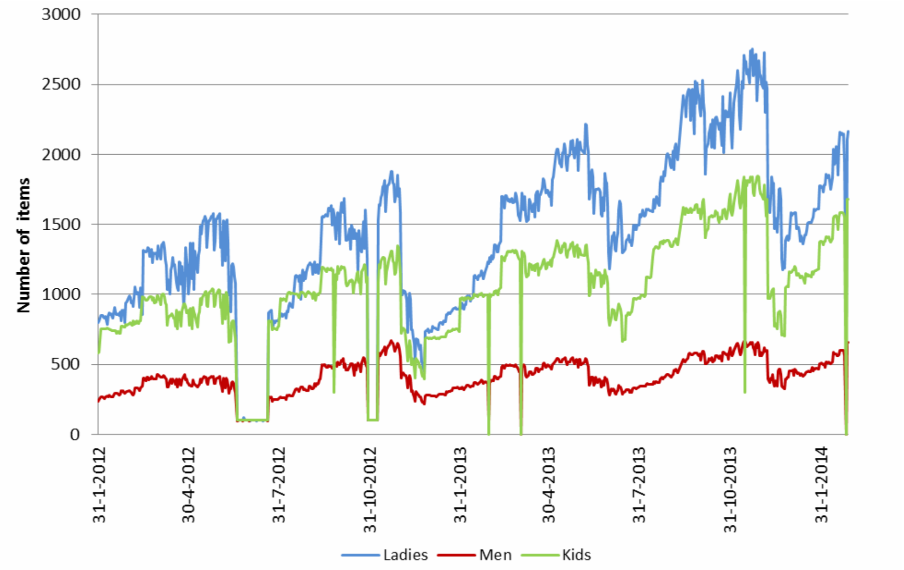


Figure 4. Daily number of 'regular items'

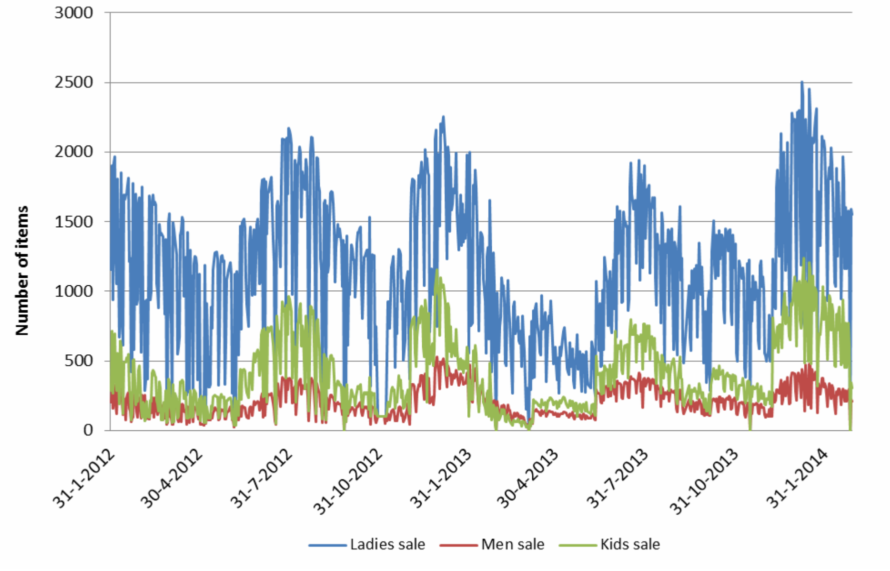


Figure 5. Daily number of 'sales items'

The number of items in each sales sub-department fluctuates much more than that in the corresponding regular sub-department. A clear seasonal pattern is visible for all sub-departments. The sales periods that can be identified in Figure 5 correspond to the periods in Figure 3 when many items moved from regular to sales. The peaks for women's sales in Figure 5 are on 31 July 2012, 6 January 2013, 25 July 2013 and 8 July 2014.

3.3 Summary of findings

Below, we summarize the main findings about the dynamics of the items observed on the website of retailer "S", which may affect the choice of price index number method.

- The robot currently does not observe the entire collection on a daily basis. We will return to this issue in Section 5.1.
- The number of items observed fluctuates substantially from day to day. This is due to the fact that the total population of items is only partially observed on a daily basis.
- As expected, the numbers of regular and sales items exhibit a clear seasonal pattern. The sales periods as identified on the website correspond to the sales periods observed in the physical shops.

In Section 4 we will present tentative index number calculations based on taking ratios of average prices at the lowest level of the clothing classification used. This classification has been developed during our web-scraping project and will therefore be discussed in some detail.

4. Classification and tentative price index numbers

4.1 A classification for clothing

As part of our effort to investigate the usefulness of online data for CPI purposes, a *classification system* for clothing, referred to as Klcl, has been developed. Short item descriptions found on the website of “S” have been used as source material. Yet, Klcl does not contain any descriptions that are specific for “S” and should thus apply to any web store. In particular, the classification is independent of the categorization used by web stores themselves.[7]

Klcl distinguishes four (rather than three, as on the website of “S”) departments, i.e. women’s, men’s, children’s and babies’ clothing. A further breakdown is based on a number of dimensions, including underclothing versus outerwear, part(s) of the body covered, function (clothing to be worn every day at work or privately, to be worn at home or during the night, party and formal dresses, etc.). The categories are the same for each of the four departments, resulting in a matrix-like structure. Some cells will obviously be empty; there are no suits for babies, for instance.

It might be useful to further refine the current structure of Klcl. In practice, however, we are limited by the metadata available on the website and the format in which it is available. Some formats are easier to use than others. The short item descriptions on the website of “S”, and probably on other retailers’ websites as well, can be read and processed directly. As mentioned in Section 2.3, additional item information on the fabric used, details on the fabrication, care label, etc. is present in long descriptions on the website. This information might prove useful to refine the classification in the future, but so far we have only used the short descriptions to classify the items.

A more detailed classification will increase the homogeneity of the item categories. However, a too detailed classification will be unstable over time: existing categories will disappear and new categories will appear. Put differently, it is not possible to develop a stable classification with completely homogenous item categories. But this does not mean that collecting additional item information using the long item descriptions will not be useful. Additional information may be required to control for changes in the composition of the item categories, or more generally, to control for quality change using hedonic regression or otherwise.

Due to the bulk of data collected by the Internet robot, the classification of items had to be done automatically. Our *automatic coding* method yields a high percentage of correct matches, about 98%. This has been achieved by using the classification that the retailer uses to structure the website. The key advantage of this approach is that it can provide a unique meaning to a term which is otherwise ambiguous. ‘Jeans’ is

[7] However, when automatically coding items, use of a web shop’s classification may be very beneficial, it has turned out in our investigation. It helps create a narrow context in which a piece of clothing is unique, which it is not without this context. This allows clothing items to be coded automatically in the majority of cases, and as it turned out the quality of the results was extremely good.

an example of this. It depends on the department (women's clothing, men's clothing, children's clothing) what code is to be associated with such an item.

In other automatic coding applications, e.g. on jobs, the input is often riddled with spelling and grammatical errors, abbreviations, special terminology, etc. The success of our automatic coding was due to a large extent to the good quality of the short item descriptions found on the website of "S", which use a specialized vocabulary with a well-defined delineation.

4.2 Monthly price indexes for item groups

We calculated tentative price indexes at the lowest level of the current classification, for example for women's tops, men's jeans, girl's dresses, men's jackets, etc. Figure 6 shows the frequency distribution of the 12,000 prices observed for men's jackets during February 2012 to June 2013.

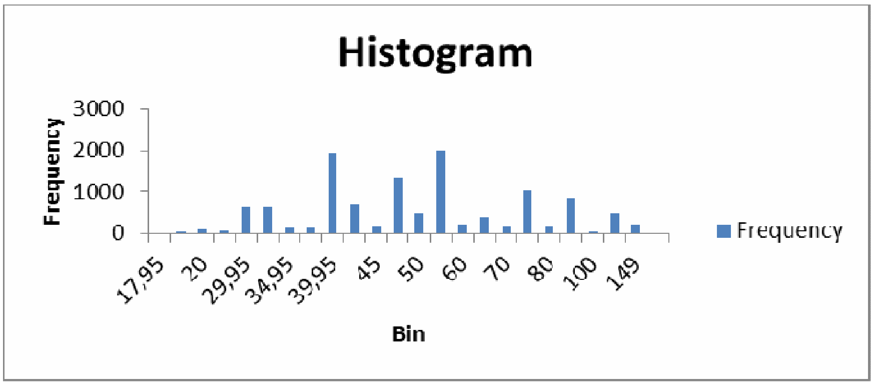


Figure 6. Frequency distribution of prices of men's jackets

Like most other categories at the lowest or elementary level of the classification, this particular category cannot be deemed homogeneous. It consists of woollen jackets, velvet jackets, cotton jackets, tweed jackets, linen jackets, cord jackets, jackets made from patterned material, blazers, gilets, etc. The *heterogeneity* results in significant price dispersion, with prices ranging roughly from 20 to 150 euros.

Quantities purchased cannot be observed online, and so the elementary price indexes are necessarily unweighted. We calculated monthly price indexes simply as ratios of average prices. Since prices are observed on a daily basis, items that are observed more frequently within a month will have a bigger weight in monthly average prices than other items. Obviously, the measured changes in average prices do not reflect pure price change because they are affected by compositional changes due to exits and entries of items (partly as a result of the robot not observing all items each day). This leads to highly volatile elementary indexes.

Using fixed weights from an external source, the (unweighted) elementary indexes have been aggregated to form price indexes for each department and store "S" as a whole. Figure 7 shows the indexes for the three departments, covering slightly more than a year. The indexes exhibit a seasonal pattern with average prices being lowest in January and July.

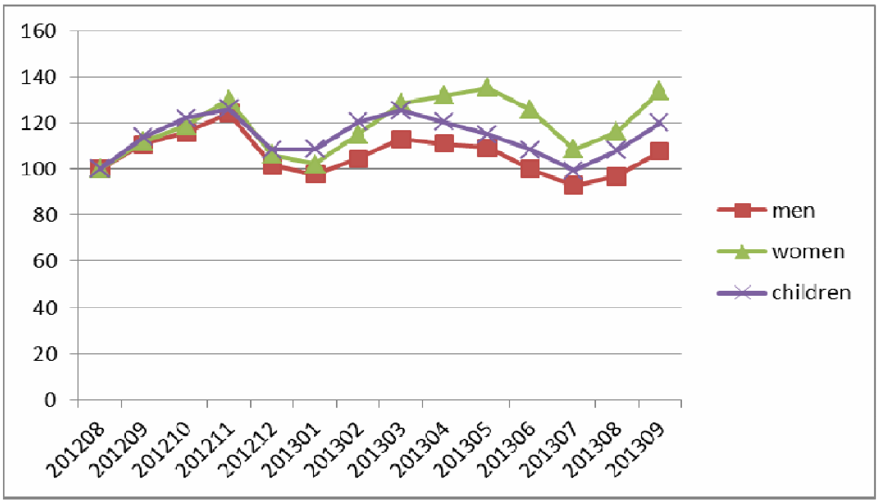


Figure 7. Tentative price indexes for three departments

We did simulations to study the impact of the frequency of observation. Alternative price indexes were calculated using the observations during the first three weeks of each calendar month, the first three Mondays of each month, and a single day per month, respectively. The results of the second simulation were very similar to the original results. Apparently there is quite some redundancy in the data. This does not necessarily mean that it is useful to drastically reduce the number of observations if such data would be used in CPI production. In case of problems with the robot it could in fact be reassuring to have some *redundancy* in the data.

5. Issues and risks

5.1 Methodological issues

An important issue of 'big data' for official statistics is representativeness; see e.g. Buelens et al. (2013) and Daas and Puts (2013). At the beginning of our project we expected to observe the complete item population of retailer "S". However, this is definitely not true on a daily basis. In particular, only a choice of colours of a certain item is observed each day. Whether the entire population is observed on a monthly basis has yet to be investigated. It is possible that some colours of an item are never shown on the "view all" page of the website. We will be able to check this at a later stage as we are now collecting additional details of the observed items (see Section 2.3), including their colour.

It may be that some items are never shown on the "view all" page. This does not seem very likely though. Hence, with the current web scraping strategy, we would expect to observe the entire population of items offered for sale on a monthly basis, although we might not observe every colour of each item.

A related question is whether the set of items offered for sale on the website is the relevant population for CPI purposes. Generally speaking it is, because purchases by households via the Internet are part of the *scope* of a CPI.[8] But there is a caveat. The web shop and the physical stores of "S" should be considered different types of outlet because the services provided are different. It could be that the collection in the physical stores is a only subset of the collection shown on websites. According to staff of "S", however, the collections are the same, except that newly introduced items can generally be purchased somewhat earlier via the website.

A major issue with web scraping is that, unlike with scanner data, quantities cannot be observed. Thus, weighting information is lacking, both for aggregating the prices of individual items to product category indexes and to aggregate across different product categories. However, this issue is not new to statistical agencies: indexes at the elementary aggregation level are usually unweighted and weighting information at upper levels typically stems from other sources. Another issue is how to deal with *regular and sales prices*. But this issue too is essentially the same as what statistical agencies have been facing with data collected in the physical stores.

What is different is that web scraping in its current form delivers daily price quotes. This raises the question as to whether we actually want to use this daily information and/or whether daily observation is needed. Usual practice is to observe prices only once per month or, alternatively, a few times per month. In the latter case, a simple unweighted average is taken to obtain a 'more representative' picture of the average price across the whole month. With online data, we could do something similar and average the daily price observations for each item, taking into account both regular and sales prices. This could be worthwhile if there is a high correlation between the frequency with which items are displayed on the "view all" page of the website (and therefore the frequency with which we observe the items) and their popularity in terms of quantities purchased. In that case we would probably end up with more accurate estimates of the desired *unit values* than by averaging a few price quotes. Unfortunately we are not able to investigate the issue since we do not have access to scanner data for retailer "S". Perhaps the retailer itself has a clue whether or not such a relationship exists.

As mentioned earlier, having additional detailed information on item characteristics is crucial. The current product categories at the lowest level of the classification are still quite heterogeneous. Additional information of items will be helpful to refine the classification and/or to adjust for compositional change within categories using hedonic regression. It should be noted, however, that the need for collecting detailed information of characteristics may depend on the methodology chosen. For instance, if manual intervention is chosen for selecting (a sample of) items to be priced and for making *quality adjustments*, looking at photos of the items might be just as good as collecting characteristics information. It might even be better because the required information may not be available on the website. Retailers put information on their

[8] For our national CPI, purchases by foreign households (as well as purchases by business) should be excluded.

website in order to increase sales, which is not necessarily the same information that is needed to properly estimate a CPI.

The quality of the data observed on the website does not seem to be a big issue. We expect the prices information to be ‘correct’ because making errors can have major consequences for the retailer. Nevertheless, some data checking is always useful. In the future, new pricing strategies could arise. Prices shown on websites may become ‘more dynamic’ in that they change due to a ‘lowest price guarantee policy’. This is perhaps not a major issue. Things would become much more complicated if retailers started charging different prices for different customers, for example depending on their past purchases. To inform potential buyers, it is likely that prices will continue to be shown on websites, but these prices would then be like advisory or maximum prices.

5.2 Risks of data collection via web scraping

As was mentioned in Section 2, a potential risk is changes in the structure of the website that have an effect on the information the robot uses for navigation or data extraction. This did indeed occur several times during our project. It can also happen that the format of the data to be collected changes. An example is a change from text to picture (bitmap). Even if the robot would properly observe the picture, which is certainly possible, the changed format would make it difficult to extract the desired information.

There is also the risk that the retailer suddenly closes its website for the robot, for instance because web scraping adversely affects the website’s performance. A good working relationship with the retailer could prevent this from happening. Of course a website can also seize to exist, but this is likely to happen only when the retailer in question closes down.

6. Conclusions and discussion

We will summarize our experiences with the observation of prices and metadata via web scraping for CPI purposes. The main advantages are:

- price collection via web scraping is cheaper than price collection in physical stores;
- given the relatively low collection costs, there is an incentive to rely on ‘big data’ and circumvent small sample problems (e.g. high sampling variance);
- the quality of online data tends to be very good;
- some item characteristics can be easily observed.

The main disadvantages are:

- website changes can lead to data problems;
- the choice of web scraping strategy can affect the information collected and item representativeness;
- weighting information is unavailable;

- the available information on characteristics may be insufficient, depending on the need for quality adjustment.

There is little we can do about the last two disadvantages. The first disadvantage can be mitigated by establishing a good working relationship with the web store so as to be prepared for website changes. A good relationship could also reduce the need for reverse engineering to understand what is happening. The second disadvantage calls for a web scraping strategy that is tailored to the needs of the specific method chosen to estimate price indexes rather than a strategy aiming at obtaining as much data as possible. For example, if the traditional, small sample CPI methodology is adapted to online data, then the strategy will most likely differ from the strategy required for a ‘big data’ solution.

A potential advantage of web scraping we have not discussed so far is that it could be an effective way, and sometimes perhaps the only way, to collect information on clothing prices and characteristics. The information put on the website by retailers such as “S” comes from a transactional database, which is constantly being updated. The information collected via daily web scraping can be seen as snapshots from the transactional database. We use the daily snapshots to build a kind of data warehouse, allowing queries with a time dimension to be made. It is likely that many retailers do not keep historical data on prices and characteristics over a long period of time, in which case building our own data warehouse might be the only opportunity to obtain such data.

Scanner data could be an alternative data source. It has the advantage of providing both prices (unit values) and quantities sold, enabling us to calculate weighted price indexes at all levels of aggregation. A potential problem, particularly for clothing, is that a large proportion of goods that are ordered online is returned by the customers. The scanner data sets may or may not take returns into account. If ‘gross’ quantities are registered, then the resulting weights could be far off the mark. But even if ‘net’ quantities are registered, it may not be exactly clear how returns have been treated. Also, the use of ‘net’ scanner data in CPI production can lead to timeliness problems as customers return the goods (at no additional charges) after some time, up to a few weeks. Another potential problem with scanner data is that the information on item characteristics may be scarce. The scanner data Statistics Netherlands receives from chains or web stores contain short item descriptions, which are not necessarily the same as the short descriptions found on websites.

If the problem with returns can be resolved, we could try and enrich scanner data by merging them with online data, including long descriptions, provided that a common item identifier is available. This approach will not be possible for small independent shops that cannot provide us with scanner data. Moreover, we are unable to develop and maintain Internet robots for many small shops, and some physical stores have no website anyway. Visiting small shops to collect prices may therefore still be needed. Combining different collection methods, data sources and index number methods for clothing seems to be the natural way forward.

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.

Buelens, B., Daas, P., Burger, J., van den Brakel, J. and Puts, M. (2013), "Initial Insights into Representativeness in the Context of Big Data", Report PPM-2013-09-30-BBUS, Statistics Netherlands, The Hague, The Netherlands.

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

Daas, P., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O. and Ma, Y. (2011), "New Data Sources for Statistics: Experiences at Statistics Netherlands", Discussion Paper no. 201109, Statistics Netherlands, The Hague, The Netherlands.

Daas, P. and Puts, M. (2013), "Big Data as a Source of Statistical Information", Report PPM-2013-12-09-PDAS, Statistics Netherlands, The Hague, The Netherlands.

Greenlees, J. and McClelland, R. (2010), "Superlative and Regression-Based Consumer Price Indexes for Apparel Using U.S. Scanner Data", Paper presented at the Conference of the International Association for Research in Income and Wealth, 27 August 2010, St. Gallen, Switzerland.

de Haan, J. and Hendriks, R. (2013), "Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes", Paper presented at the Economic Measurement Group Workshop, 28-29 November 2013, Sydney, Australia.

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.