Introduction – The Value Chain of Scanner and Web Scraped Data

Jens Mehrhoff*

Abstract – With the advent of scanner and web scraped data, “big data” sources are increasingly finding their way into official statistics. This second part of the special issue on “Big Data and Statistics” is devoted to developments in the use of these data for consumer price indices. To what extent are big data different to more traditional data sources such as the collection of prices in the field, and how do they change the process of producing consumer price indices? The four papers in this special issue address these questions by means of the experiences gained in the statistical offices of France, Sweden and the Netherlands. This introduction puts them into perspective vis-à-vis the value chain of scanner and web scraped data and looks at some further issues for research in this field.

JEL Classification: C43, C55, C82, E31
Keywords: consumer price indices, big data, scanner data, web scraped data

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution’s or Insee’s views.

* *Deutsche Bundesbank (jens.mehrhoff@bundesbank.de)*

Received on 21 July 2019

To cite this article : Mehrhoff, J. (2019). Introduction – The Value Chain of Scanner and Web Scraped Data. *Economie et Statistique / Economics and Statistics*, 509, 5–11.
https://doi.org/10.24187/ecostat.2019.509.1980

Setting the Scene

Consumer price indices are the gauge used to assess price stability, which makes them the single most important measure for central banks's monetary policy-making. With the advent of scanner and web scraped data, "big data" sources are increasingly finding their way into consumer price indices internationally. This second part of the special issue on "Big Data and Statistics" is devoted to developments in the use of scanner and web scraped data for consumer price indices.
The underlying questions of the four papers in this special issue are to what extent Big Data are different, or similar, to more traditional data sources such as the collection of prices in the field, and how they change the process of producing consumer price indices. While both approaches share the obvious same target – measuring the average rate of change in consumer prices – how this number is derived differs in multiple ways. First and foremost, scanner and web scraped data give access to a much broader continuum of products than classical sampling allows. The supposedly better coverage of goods and services comes at a cost, though: churn due to new and disappearing products, i.e. a dynamic product universe. Moreover, quantities sold (with scanner data) or at least a popularity ranking (from websites) become available too, thus allowing the calculation of weighted indices rather than the need to rely on unweighted formulae. The cost here is chain drift, i.e. the index might show spurious trends over time.

In this introduction, we put the four papers into perspective vis-à-vis the value chain of scanner and web scraped data, considering three stylised phases: i) collecting data; ii) processing data; and iii) disseminating results. We conclude by looking at some further issues for research in this field.

Collecting Data

Thanks to the pioneers in using these new data sources, there are now best practices for collecting scanner and web scraped data. The Eurostat *Practical Guide for Processing Supermarket Scanner Data* (2017) lists recommendations, which generically apply also outside the realm of supermarket scanner data. In particular, building a relationship with the data owners appears to be key. Supermarket chains and online retailers were afraid that their data might be misused by their competitors; once mutual trust is established, these reservations can be eliminated.

In terms of scanner data an arrangement might take the form of a *quid pro quo*, i.e. the data providers get some kind of market benchmarks as well as data analyses in return for their figures. In no case are micro data or competitor information disseminated. For web scraped data, the owner of the website might be open to provide an application programming interface, better known as API, rather than block the sta-tistical office's IP address if they understand who uses their data for which purposes.

Another approach to the collection of data is the establishment of a legal framework that allows statistical offices access to such sources. Details on this will very much depend on institutional arrangements at the national level.

Independently of the desired or feasible level of aggregation in terms of time, outlets and regions, experimental data sets should be tested before establishing the data flows in production. On both ends, there are many technical issues to be resolved such as transmission format or data storage.

Processing Data

There have been several approaches to further break down the second phase, processing data. Though by and large similar they differ due to institutional arrangements such as the statistical office's current approach to consumer prices. Typical steps include but are not limited to the automatic classification of products, intermediate aggregation of "homogeneous" products, rule-based filtering of observations and the calculation of the final index.

In the same vein, **Marie Leclair and co-authors** review how a number of questions have been addressed in France in relation to price aggregation to produce indices, handling quality adjustments, classifying goods by homogeneous product variety and product relaunches and promotions.

Classification

The vast amount of products can no longer be classified to COICOP or breakdowns thereof manually but only automatically. The classification might come from the data owner, at least to some extent. Supermarkets, for example, have their own classification for scanner data which might be useful to this end. The same holds true for web shops, where the products might be presented in a structured way. However, should this information not be available or sufficiently detailed for the purpose, one has to rely on supervised machine learning techniques. Yet, this requires the construction of a small labelled data set in order to train the algorithm.

Initially, all products need to be classified. In addition to information from the data owner, typically product codes (such as GTINs), descriptions (i.e. text) and other metadata (e.g. size) are available. A major challenge in this respect is feature engineering. In most cases, product descriptions are not natural text but use specific vocabularies and rely on different kinds of shorthand. Product codes, in general, follow some kind of a structure. Also, every month new products will appear and need to be classified as well. Already classified products should not be re-classified in this exercise. Nonetheless, the quality of the classification over time should be assessed. A further complication is the identification of re-launches, e.g. when the very same product is sold in a different packaging but gets a new product code.

Product Aggregation

A first step in the calculation of elementary indices is the definition of the so-called homogenous product. Due to product churn and the sheer amount of observations, the classical fixed basket approach would only be viable if a small but fixed sample was drawn from the data. With the approach of using most of the data gathered, a trade-off between product homogeneity and product continuity arises. In this case, the problem is elevated by re-launches, whose identification is not at all straightforward.

The dilemma here is that it is *per definitionem* impossible to come up with an optimal solution. It is advisable to test different scenarios for the product definition and investigate a homogeneity measure and a continuity measure independently as well as their development over time rather than a single summary statistic. In particular high churn and seasonal products need special attention; for consumer electronics, say, hedonic quality adjustment might still be the best option. Eventually, product continuity must not be bought at the expense of (unit value) bias.

As an example of implementation, **Can Tongur** discusses the issue of preserving the fixed basket approach, despite the introduction of scanner data in Sweden, and why the traditional manual item replacement strategy, with quality and quantity adjustments, is still a relevant method to ensure comparability.

Filtering

If a fixed sample is drawn from the data, the problems associated with scanner and web scraped data are similar to the situation of traditional price collection and include imputation and quality adjustment. If the intention is to use most of the available information, on the other hand, some rules are necessary to pre-process the raw data. Filters usually remove product codes that are not representative over time, observations considered to be suspect and potentially products with low sales or that are likely to be dumped.

Product codes that are not representative include product groups out of the scope (e.g. clothing for supermarkets) and generic codes used by the data owner in a non-stable manner. Suspect observations refer to both outliers, e.g. unusually low or erroneous prices, and influential products, e.g. extreme expenditure shares or high leverage. Low sales filter introduce a coarse weighting, leaving only the relevant products in the index, thus mimicking a weighted formula. Dump filters try to minimise the downward effect of disappearing products in clearance sales.

Index Calculation

After the data set has potentially been further edited, e.g. imputations for missing prices, the final index can be calculated. Choices include a fixed basket with a bilateral formula and multilateral approaches in a dynamic product universe. In no case should weighted indices be chain-linked at a high frequency such as monthly. These have shown to be subject to severe drift.

If a bilateral approach is chosen, we are again very much in the same situation as with traditional price collection. The major difference now is that, if scanner data are used, weights from the current period and formulae such as Fisher or Tornqvist can be employed. On the contrary, if a multilateral approach is chosen, several decisions have to be taken: which particular multilateral approach should be implemented using how many months as the estimation window and how should the disseminated time series be extended in real time without revisions? There is no consensus on the “right” answers here and it might be more straightforward to search for robust methods – those that produce reliable estimates even for challenging product groups – rather than some economic or statistical justifications.

Though now used in intertemporal comparisons, multilateral approaches originally come from the literature on international purchasing power parity comparisons. While these approaches are, hence, obviously not tailored to the problem at hand, they do the trick and ensure freedom of chain drift, which is considered a *conditio sine qua non*. Plenty of methods have been suggested for interspatial comparisons but the following three emerged to be preferred in the time domain (in no particular order): time-product dummy (TPD), Geary-Khamis (GK) and Gini-Eltetö-Köves-Szulc (GEKS). The TPD method derives the price index from a log-linear regression framework, the GK method does it through the solution of a harmonic eigenvalue problem, and the GEKS method transitivises bilateral indices through geometric averaging.

While all of three aforementioned approaches satisfy circularity, that is the chain-linked index defined as the product of the short-term indices is equal to the direct index, when data for the next month are added the entire time series would be subject to revisions. When using any of these methods this is, unfortunately, unavoidable. To circumvent the problem of revisions, the estimation window is shifted forward while keeping its length fixed and the new index is spliced onto an already disseminated figure. Typically, the estimation windows should cover no less than 13 months and the splicing is performed onto the previous month (movement splice), the same month in the previous year (window splice) or something similar.

There is a growing literature on how long the estimation window should be, which proves particularly challenging for strongly seasonal items exhibiting trends, and how exactly the extension should be performed. Since chain-linking of consumer prices indices is today the standard, at least the latter question might be answered by looking at the way the overall index is calculated. Evidence points to that some kind of anchoring mitigates path dependency of the index; the classical chain-linking approaches reflect this by either referring to the average of the previous year (annual overlap) or the last quarter/month of the previous year (one-quarter/month overlap).

A final word in this respect is due. While it is already regressive to invent yet another approach which comes closer to the fully transitive benchmark index with one or the other data set, there is a severe complication with that benchmark particularly when products are seasonally unavailable. Extending the time window has a contrary effect: the index loses what is known as “characteristicity”. What does that mean? The relative differences in price levels of the products are accounted for implicitly by multilateral methods. This adjustment is an average over the estimation window. However, should products within the elementary aggregate show differing trends, that time average is just wrong (it is not “stationary”). For strongly seasonal items expressly this can lead to obscure index numbers in the benchmark and different estimation windows can lead to hugely divergent time series.

An illustration of index calculation is found in the article by **Antonio G. Chessa & Robert Griffioen**; more precisely they investigate whether web-scraping of online prices of consumer goods is a feasible alternative to scanner data given the lack of transaction data.

Disseminating Results

Most likely, statistical offices will not disseminate very detailed information, above all not if it would allow identification of a data owner. Thus, the elementary indices are aggregated from that level, and potentially even a regional breakdown, to COICOP using weights from business statistics, for example. But this also means that data users, more often than not, get just the same level of detail from the publication using scanner and web scraped data as they get from traditional price collection. In this sense, statistical offices might be using big data sources but they are still disseminating "small statistics".

Furthermore, indices from scanner and web scraped data have shown to be more volatile than traditional indices. While the traditional price collection of matched models shows little to no noise in the price developments, the new methods introduce a lot of noise in the time series. This is all the more true for weighted indices and using scanner data. Basically, and despite the estimation window, multilateral methods perform cross-section averaging only. An area for further research is whether time averaging can help in dampening the noise and amplifying the signal component.

A notable exception in the level of detail disseminated is **Isabelle Léonard and co-authors** who calculate indices that measure differences in consumer price levels between different areas of metropolitan France, focusing specifically on food products sold in supermarkets.

Wrapping Things Up

Recent developments now allow the standardisation of implementing scanner and web scraped data across different statistical offices. As regards scanner data, the Dominick's Finer Foods data set is publically available from the University of Chicago Booth School of Business to build capacity.[1] Several workshops have been developed in using different tools for web-scraping that only need adaptation to the specific case at hand.[2] For the calculation of indices, a beta version of an R package is available that enables the use of the most common methods.[3]

The update of the 2004 Consumer Price Index Manual will include a research agenda, which of course includes scanner data and web-scraping. It does not put into question the very approach from the standpoint of economic theory, though – which is also due to its intention of being more practically applicable. So-called cost of living indices recognise that quantities consumed depend on prices. They do not, on the other hand, recognise that consumers stock-pile a product when it goes on sale, thus violating a basic assumption, i.e. purchases of goods made during a period coincide with the consumption of these purchased goods within the period. But cross-production substitution is dwarfed by intertemporal substitution and, as a consequence, static estimation may provide misleading results.

Finally, scanner and web scraped data represent an admittedly "big" but biased non-probabilistic sample – not the population. There are transactions that are in the scope but are not recorded electronically, not available to the statistical office, deleted in the filtering step, cannot be matched or linked, and so forth. After all, not more data are better, better data are better. Scanner and web scraped data can be very

[1] *https://github.com/eurostat/dff*
[2] *https://unstats.un.org/bigdata/taskteams/scannerdata/workshops/Presentation_webscraping_Bogota_Statistics%20Belgium.pdf*
[3] *https://cran.r-project.org/package=IndexNumR*

precise but at the same time may have limited accuracy. The danger lies in blindly trusting that these new data sources must give us better answers; in fact, big data are not capturing all transactions, just some, and we might not even know which ones are missing. That is why the combination of more traditional data with big data is the ticket to reducing coverage bias. □