Quality Control of Web-Scraped and Transaction Data (Scanner Data)

Ingolf Boettcher[1]

[1] *Statistics Austria, Vienna, Austria; ingolf.boettcher@statistik.gv.at*

Abstract

New data sources such as web-scraped data and business transaction data (e.g. online and scanner data from retailers for price statistics) have the potential to improve official statistics, both in terms of quality (more data) and efficiency (low data collection costs, lower response burden). However, when using new data sources statisticians have to review and eventually replace traditional data quality control procedures to comply with existing quality standards. The challenges to deal with are manifold: How to define and identify outliers in millions of transactions data sets? How to select representative data sets from the internet for official data production? How to validate integrity and completeness of web-scraped data? How to integrate new and diverging data set structures into established statistical production processes?
New kinds of skills ("data science") are required from statisticians to handle these issues, such as advanced knowledge of data manipulation and programming – but also the right amount of statistical creativity to transform new and ever-changing (big) data sources into high quality official statistics...

Keywords: web-scraping, scanner data, transaction data, online data collection

1 New Secondary (Price) Data Sources – Potential of Scanner Data and Web-Scraping

Statistics Austria deploys several price collection methods to compile consumer price indices. Price collection is organized centrally via email, fax, internet and telephone, and regionally via price collection in actual outlets. Two major developments make it necessary to modernize existing price collection methods:

Firstly, more and more consumer products are sold using flexible and/or individual pricing schemes (promotion prices, membership prices). Conventionally measured list prices are becoming less reliable. The actual prices paid are captured only in the retailer's transaction

data sets (e.g. scanner data from supermarket chains).
Secondly, the growing importance of online commerce: As regards e-commerce, conventionally measured list prices become less reliable as online pricing schemes are highly flexible. Website prices fluctuate significantly depending on increasingly complex price setting algorithms. Prices may depend on time / place (IP-Address) / identity (member vs. non-member), quantity, demand history, etc. (This is the case in particular for airfares and hotels, less so for food, clothing and electronics).

Both developments require official statistics to adapt data collection sources and methods in order to maintain high statistical quality standards. Despite all difficulties, they are an opportunity to improve the quality of official price statistics. This is because, in comparison to conventional data sources, transaction and online data potentially allow total coverage of the target universe (e.g. price paid for consumer products) along several dimensions: time (every day), products (all items) and markets (all stores). Also, data collection efficiency increases as transaction and online data allow a shift from manually collected prices to automated data collection processes. In addition the response burden for business is reduced.

Statistics Austria reacts on the growing importance of transaction and online data by conducting several pilot projects on the use of scanner data and web-scraping, supported by Eurostat.

The Austrian Scanner Data and Web-Scraping Projects

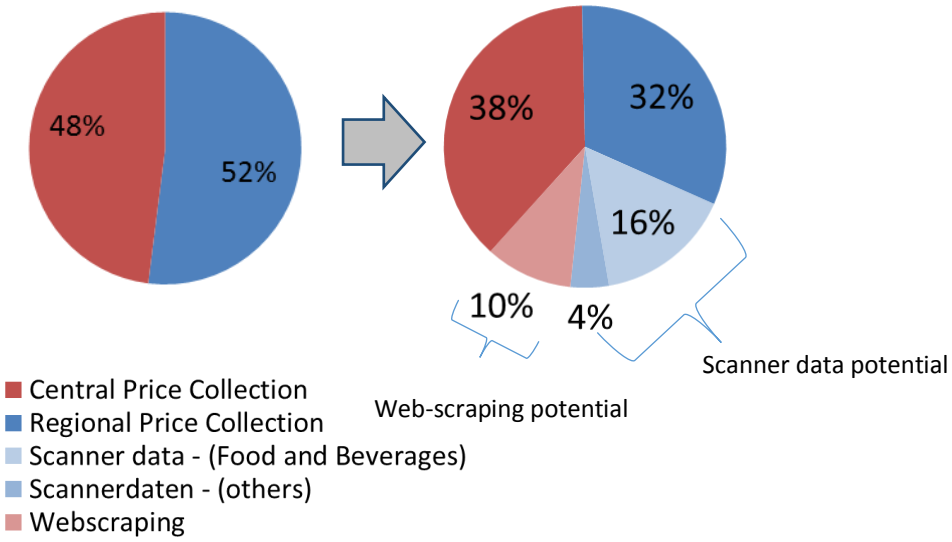
The Austrian scanner data project currently focuses on retail segments offering standardized consumer goods which are relatively easy to categorize and to process for CPI compilation. Food and beverages are the most important product groups for the scanner data project. Other possible retail segments are drugstores, medical and pharmaceutical products, non-durable household goods, small tools, accessories for cars and miscellaneous accessories

Mostly, these segments are currently covered by regional price collection in the 20 CPI regions. At the moment, durable goods (e.g. cars) and major appliances (e.g. washing machines) are not targeted by the scanner data project. These product groups often require

quality adjustments according to international standards and implemented by trained CPI staff at the central office.

Altogether about 20% of the CPI basket of goods may be covered using scanner data.
Figure 1 below depicts the potential of scanner data (and in addition of web-scraped data).

Figure 1 –Share of central and regional CPI price collection (% weight), scanner data and web-scraping potential



The Austrian web-scraping project focuses on products and services for which price data is currently manually collected on the internet. At the moment, the most important target segments for web-scraping are transportation (e.g. flight tickets, train tickets, holiday package tours), technical equipment, clothing and shoes and hotels.

With the exception of 'clothing and shoes', these segments are mainly covered by central price collection. Momentarily, altogether about 10% of the CPI basket of goods may be covered using web-scraping.

In addition, results of the Austrian Household Budget Survey 2015 are going to provide detailed information on the market share of internet purchases in different product and service

segments. Accordingly, the Austrian CPI price collection will adjust the sample of online shops and their CPI share to correctly represent the growing importance of ecommerce.

2 Challenges of transaction and web-scraped data for quality control

New data sources such as transaction and web-scraped data are subject to existing quality control standards. However, there is a lack of applicable quality control procedures and guidelines regarding data from large and vast secondary data sources (transaction data, web-scraped data, social media data, internet of things data, etc.). Compilers of official statistics will find it hard to apply existing data quality frameworks to large data source. In fact, official statistics quality frameworks have for a very long time focused on *primary statistical data* sources (e.g ESS quality report 2014). In these quality frameworks the *output* requirements of official statistics are thoroughly described. In the last years, updated quality frameworks focus more and more on the integration of *secondary non-statistical data* (data that has been collected for the purpose of compiling official statistics). In particular the integration of *administrational secondary non-statistical data* for official statistics has been in the spotlight of quality frameworks (see UNECE 2011). Recently, works have started to provide guidelines and frameworks for the quality of *input data* for official statistics users to guide statistician when compiling official statistics from Big Data sources (Eurostat 2015; Struijs P. and Daas P. 2014).

When using large secondary data sources such as scanner data and web-scraped data, compilers of official statistics might struggle to align their processes to existing quality guidelines and frameworks. Table 1 and 2 below depict several novel challenges encountered by the Austrian scanner data and web-scraping pilot project. The tables contain the most relevant quality criteria for input data and only depict problems and measurement methods that are unique to transaction and web-scraped data – while quality problems known when dealing with conventional primary or secondary data sources are not treated.

Table 1 – Novel quality problems and measurement methods with transaction data

Input data quality criteria	**Transaction data /scanner data**	
	Novel quality problem (for consumer price statistics)	**Measurement Method**
Relevance	Data may contain transactions that are out of scope. -e.g. expenditures of business (out of scope for consumer price indices)	Information by data providers; otherwise unresolved
Accuracy	Volume and variety of data sets are too large to identify and clean erroneous/ untrustworthy/ inconsistent data sets with conventional methods.	Extent in % of erroneous / inconsistent data is monitered and excluded
Timeliness/Punctuality Accessibility	- (no new kind of quality problem)	-divergence from formal data delivery agreement between data provider and NSI
Completeness	Volume and variety of data sets are too large to identify missing values with conventional methods. (Scanner data: natural attrition of Unique idenfiers is extremely high)	Number and level of target values are measured against historical values from previous deliveries
Clarity / interpretability	(no new kind of quality problem)	Information by data providers about: -format and definition of variables -data transformation checks performed before delivery (e.g. aggregation)

Challenges of the use of scanner data

Table 1 shows that the use of transaction / scanner data poses several quality challenges that need to be addressed. Replacing traditional price collection with scanner data leads to a high dependency of Statistical Offices on the providing retailers. Therefore, the quality provisions of the delivered transaction data should preferably be laid down in a formal agreement. Scanner data is a secondary data source and may include data types, classifications,

characteristics and elements that are hard to integrate with the existing CPI production system. Processing scanner data can be difficult as each retailer usually deploys different database structures, data types and product classifications. Extensive data cleaning and index compilation procedures need to be developed for each scanner data provider. In particular, there are two main tasks to achieve when processing scanner data after receiving it from retailers and before CPI compilation: matching individual articles between time periods and assigning/mapping GTINs to a CPI elementary aggregate (EA) and COICOP (sub-)class. Also, size and structure of the data files might require investments in IT infrastructure. Finally, scanner data raise methodological issues that need to be addressed to ensure that existing rules establishing comparable CPIs in the EU are not violated.

Table 2 - Novel quality problems and measurement methods with web-scraped data

Input data quality criteria	**Web-scraped data**	
	Novel quality problem (for consumer price statistics)	**Measurement Method**
Relevance	Representatives of online data (are products offered really sold and by whom?)	Information by data providers; otherwise unresolved
Accuracy	Website content may be IP-specific (a user who frequently checks a website or a web-scraper might lead to different price displays than first-time users)	Comparison of automatically and manually collected data
Timeliness/Punctuality	the amount of data makes it difficult to jugde data quality within a reasonable amount of time	-quantative instead of qualitive processing of data
Accessibility	Websites might identify web-scrapers and block them	unresolved
Completeness	Websites change frequently Releveant variables and URLs might not be identified and scraped	Number and level of target values are measured against historical values from previous data collection activities
clarity / interpretability	- (no new kind of quality problem)	

Challenges of the use of web-scraped data

The Austrian web-scraping project is faced with frequently changings websites. This requires the re-programming of the respective web-scrapers. It can be expected that price index staff can spot changes to websites more easily and will immediately resolve malfunctioning web-scrapers.

Development of automatic price collection quality assurance processes

Price statistics staff uses the web-scraping software and create automation scripts to continuously download price data from eligible online retailers. This step includes checking the compatibility of the specific extraction methods applied on the selected data-sources (online retailers). Quantitative as well as imitating approaches are considered. The Quantitative approach aims at continuously harvesting all the available price data from selected websites. The imitative approach collects automatically the data according to criteria, which are currently already applied in the manual price collection. The extracted data is analyzed and cleaned for price index compilation.

Part of the quality assurance is the comparison of automatically collected price data with manually collected prices. Predefined research routines and consistency checks will be deployed. It would be beneficial to deploy another web-scraper software whose results could be automatically compared with the results of the first web-scraper. The irregular maintenance work needed to run the web-scraper has to be assessed and quantified. Maintenance is required to assure quality when website architecture is changed. There is evidence that the resources needed to perform the irregular maintenance work depends on the individual website and heavily affects the total work load. Thus, a critical cost effectiveness analysis is needed when applying automatic price collection methods.

3 Quality Control of new secondary data sources - Need for "data science"

References

New kinds of skills ("data science") are required from statisticians to build up quality assurance processes that comply with existing quality standards on data output. Advanced knowledge of data manipulation and programming is necessary to succeed in this task. Statisticians will have to invest into training of staff within their unit and improve and integrate cooperation with colleagues from other departments able to handle large data sources (e.g. IT , data collection departments). All in all, the right amount of statistical creativity is necessary to transform new and ever-changing (big) data sources into high quality official statistics. Large secondary data sources require individual data cleaning and editing processes. Measurement of big data input data quality will make more flexible measurement methods and quality benchmarks necessary. To facilitate these challenges, the statistical community should continue the work on guidance and quality frameworks for integrating large new secondary data sources into official statistics. This is especially important, as NSIs usually are facing financial constraints. There might be a danger of only using the advantages of new secondary data sources (low collection costs, high coverage) and to neglect the disadvantages (need to develop new quality measurement processes). Large new secondary data sources have the potential to improve official statistics but also to cause faulty and biased outcomes as the amount of quality related decisions by statisticians increase.

References

ESS quality report.(2014), [Online] Available: http://ec.europa.eu/eurostat/web/quality/quality-reporting

Eurostat (2015). HICP Recommendation on Obtaining Scanner Data (Draft August 2015)

UNECE (2011). Using Administrative and Secondary Sources for Official Statistics. [Online] Available: http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf

Struijs P. and Daas P. (2014). Quality Approaches to Big Data in Official Statistics. Paper presented at the Q2014. [Online] Available: http://www.q2014.at/papers-presentations.html