Statistical Journal of the IAOS 32 (2016) 43–52
DOI 10.3233/SJI-160965
IOS Press

Big Data: What can official statistics expect?

Peter Hackl
Institute for Statistics and Mathematics, Vienna University of Economics and Business, Building D4, Level 4, Welthandelsplatz 1, 1020 Vienna, Austria
Löfflerweg 9, 7100 Neusiedl/See, Austria
Tel.: +43 676 9379337; E-mail: peter.g.hackl@gmail.com

Abstract. New data sources are becoming available, in particular as a consequence of the enormously growing use of electronic media. These new data sources, often called the Big Data, open new opportunities for official statistics: Statistical products may be delivered in shorter time, with more detailed breakdowns, at lesser costs, and with reduced response burden. The paper gives a short overview of Big Data pilots and projects conducted in official statistics at national and international levels. The experiences from these projects in using the new data sources indicate that there is no uniform methodological approach for using the new data in the various statistical domains nor to take advantage of the new opportunities. Official statistics cannot expect that Big Data will substitute actual data sources like data from surveys and administration, but they may play a role as supplements for existing data in the production of certain statistics. A number of challenges are involved in using Big Data in official statistics: Solutions for methodological and technological issues are needed, a quality framework must be developed, and the staff has to get acquainted with new skills. Further issues concern the statistical environment such as legislative requirements, costs of sourcing of Big Data, and privacy, and have only indirectly to do with the statistical production. Finding solutions for these issues and developing standards that will be accepted internationally may require substantial efforts and take some time.

Keywords: Big Data, official statistics, international initiatives, Big Data pilots, challenges

1. Data needs in official statistics

The traditional data sources in official statistics are sample surveys, censuses, and administrative bodies.

Sample surveys have a sound methodological basis that has been developed over many decades. A strength of sample surveys is the control over the data collection which allows inferences on the corresponding population; the NSIs know to deal adequately with quality issues like non-response and survey errors. Problems of increasing size are the high cost of surveys and the growing resistance against the response burden.

Censuses have the strength that they allow results for small geographic areas and population sub-groups. They are simple in terms of statistical methodology but at enormous financial costs.

Administrative bodies are the owners of data for their specific purposes. Strengths are that the data of such bodies contain information on a full population of well-defined units, and that these data are continuously updated. Examples of data from administrative sources are tax data, social insurance data, credit card data, and counts of births, deaths, etc. For an NSI, access to administrative data, and to the actual updates of the data, requires establishing a working partnership with the owner of the data. In using the data, the NSI has to be aware of quality issues, among them the issue of representativity; quality issues are specific for each dataset.

Most administrative data that are used by NSIs are owned by public authorities like ministries. But many other bodies in business and other areas are owner of data which may also be useful for NSIs. A good example is the retail business: The big retail chains nowadays use scanners in the billing process of retail sales. Output of this process is a dataset covering each individual transaction. Such a dataset can serve as the basis for accounting, for warehousing, for sales forecasts and other analyses, etc. It can equally well serve as price information of the traded commodities for calculating price indices like the consumer price index (CPI).

Scanner data are an example for data that can serve as alternative to traditional data for the production of official statistics. Such alternatives are highly welcomed by the NSIs. The NSIs hope to reduce the response burden on enterprises and households as well as their costs for data collections and to improve the quality of the price statistics. Of course, the NSIs face challenges when using scanner data: The investment costs have to be financed, partnerships with data providers have to be established, and methodological issues, e.g., the treatment of rebates, have to be solved.

The amount of data that are stored and electronically available is rapidly increasing. The main reason for this trend is the immense development of the information technologies and the growing use of these technologies in all phases of industrial production, of commercial and public services, and of our private lives. Byproducts of these processes are amounts of data in such quantities and also complexity that traditional modes of managing and processing the data are not suitable or efficient. In this situation, the notion Big Data appeared and started to be subject to wide discussions about potentials of these masses of data. Although the notion Big Data is summarizing diverse data situations and fuzzy in the many proposed definitions, it is subject of a huge interest; a query for Big Data in an internet search machine results in 750 millions of results.

Official statistics, always looking for feasible alternatives to data collection, realized that Big Data might provide opportunities for improving the data collection situation in their offices. Various activities and initiatives were started by the NSIs of various countries and also at the international level.

The paper gives an account of the development of this interest up to the present time. Based on this account, a critical view is taken on the potentials of Big Data for the production of official statistics. The paper avoids the notion Big Data as much as possible, specifying instead concretely the data of interest in each context.

2. Alternative data sources

As mentioned above, scanner data are an example of data that can serve as alternative to traditional data collection for the production of official statistics. Table 1 shows various types of alternative data and their potential use for the production of official statistics. Various NSIs have started to experiment with the mentioned data in order to gain experience and to understand better the potentials, challenges and problems in using these data.

2.1. Scanner data

A number of initiatives are related to the use of scanner data in official statistics. In an early study, the Norwegian NSI used scanner data to compute a subindex of the CPI for food and non-alcoholic beverages [1]. At about the same time, scanner data were also used by the NSIs of The Netherlands and Switzerland for the production of price indices. The EU has established a project "Scanner Data" within the EES task force "Multi-purpose consumer price statistics" [2]. The NSIs of 17 European countries are working on the use of scanner data for the production of CPIs; 10 of them experiment with scanner data, among them the NSIs of The Netherlands, Italy, Sweden, Norway, and Switzerland. Workshops on the use of scanner data have annually been organized starting in 2011. Guidelines on obtaining and using scanner data are developed by Eurostat. Similar projects are in progress in various other countries, among them China and South Africa.

Several problems need to be tackled when using scanner data for the production of official statistics. A major concern is the dependence of the NSI on the data owners; a sustainable working relation between the partners is crucial. Other issues are the monitoring of data quality and the discrepancies between EANs, the International Article Numbers used in the bar-code of the retail business, and the COICOP-codes used in official statistics. To be mentioned are also the enormous efforts and financial resources needed in the development of a sound methodology.

2.2. Mobile phone data

Mobile phone data are information on calls and transmissions of text (SMS); the telecom providers document for each communication the time and the location of the involved phone mast. These mobile phone data are of interest for measuring tourism flows in tourism statistics and for population, migration and mobility statistics.

Eurostat instigated a feasibility study on the use of mobile positioning data for tourism. The study was conducted between 2012 and 2014 with participants from Estonia, Finland, France, and Germany including mobile network operators. The reports of the study cover technical, financial, and legal aspects, as well as methodological and quality issues [3].

National projects on the use of mobile phone data are reported by the NSIs of Italy, the UK, Slovenia, and New Zealand.

Table 1
Alternative data sources and their potential Use in official statistics

Type of data	Areas of potential use in official statistics
Scanner data	price statistics, economic statistics
Mobile phone call/text times and positions	Tourism statistics, population and migration statistics
Traffic sensor data	Transport statistics
Smart energy meter data	Population statistics, housing statistics
Satellite images, remote sensor data	Agriculture, forestry, fishery, environment statistics
Social media data, internet data	Labour statistics, population and migration statistics, income and consumption statistics, price statistics, health statistics

2.3. Other alternative data sources

Road traffic sensors produce a variety of different data, such as data from traffic loops and from traffic webcams, and transaction data from toll payment systems. Statistics Finland has gained experience with using traffic sensor data for transport statistics; it also has developed models for commuting times of individuals. The Dutch CBS used traffic sensor data for producing transport statistics and traffic statistics.

Smart energy meter data have been investigated by the British ONS for statistics on mobility and migration.

Important areas are satellite images and remote sensing data. The Australian ABS has experience in using satellite images for agriculture statistics and environment statistics. Similar projects have been conducted by other NSIs, e.g., the Canadian StatCan and SCAD from Abu Dhabi.

Interesting sources of alternative data are the social media. Projects conducted by the Australian ABS and by the INEGI of Mexico investigated the use of such data in the production in various statistical domains like health statistics, statistics of income and consumption, labour statistics, population and migration statistics, and tourism statistics. The data which are generated in the social media can be blogs or comments posted in Facebook, Twitter, or another social media site. Similar information is available in emails and other text messages.

2.4. Internet data

The internet is a global system of interconnected computer networks; it consists of millions of private, public, academic, business, and government networks of local to global scope, linked by electronic, wireless, and optical networking technologies. The internet is a huge repository of information provided by private individuals, units from private business, and public institutions. Data from private individuals are generated by using the social media, sending emails and other text messages, conducting internet searches. Business data are provided by e-commerce companies like Amazon and Geizhals, giving access on their sites to prices for books, CDs, electronics, photo equipments, toys, etc. Similarly, agencies like Booking or Opodo provides price information for flights, hotels, and car rental.

Like other alternative data, internet data, in particular price data, have the potential to be used by NSIs in the production of official statistics.

2.5. Global pulse initiative

Since 2009, this initiative of the UN Secretary-General, Ban Ki-moon, promotes the discovery, development and adoption of Big Data innovations for sustainable development and humanitarian actions. A typical project is “Estimating Migration Flows”, a study exploring whether internet search data could be analyzed in order to estimate migration flows and produce a proxy for migration statistics. The project demonstrates, like other Global Pulse projects, how alternative data can be used for estimating indicators which typically are in the portfolio of official statistics. See [5] for more on Global Pulse case studies.

2.6. Alternative data: Some issues

It should be obvious that the use of alternative data in the production of official statistics requires coping with a number of issues.

Partnership with data owners: To get access to the data, an agreement between the NSI and the owner of the data is usually necessary.

- Most NSIs have good relations with public authorities but little or no experience in negotiating with owners from the private sector.
- Using alternative data makes the NSI dependent of the owner of the data. This is only feasible if the relation with the owner of the data is sustainable.

The use of alternative data requires solving methodological issues.

- Whereas in sample surveys the data are collected in a mode which guarantees that the sample is representative for the target population, this is not necessarily true for the data from the sources mentioned in Table 1.
- Quality criteria need to be designed or adapted for alternative data and statistical products that are based on alternative data.

New tools and skills are needed to handle alternative data.

- In particular for data from the internet, IT-tools are needed for handling the large data amounts. New tools have been developed in recent years.
- The staffs of the NSIs have to learn to use the new tools. The new profile of a "data scientist", i.e., an expert with skills in statistics, data engineering, high performance computing, data warehousing, et al., will be more common for staff members in the future.

Other points are legal issues, e.g., the personal data protection, and the considerable investments necessary to adapt an NSI to the use of alternative data. A comprehensive and detailed discussion of opportunities and challenges of big data is given by [4].

3. Historical background

The use of scanner data by NSIs as alternative to traditional data sources has been mentioned in the previous Section of the paper. Early projects with scanner and also other types of alternative data were conducted by the Australian ABS, the Dutch CBS, the Italian ISTAT, INEGI in Mexico, and others.

Since about 2010, NSIs and international organizations in official statistics became interested in the notion "Big Data". Various initiatives started at the UN level and also on the regional level such as in the EU.

The Global Pulse Initiative of UN Secretary-General Ban Ki-moon from 2009 was speaking explicitly of "Big Data innovations", aiming at promoting the use of internet and other alternative data which are available in regions without highly developed statistical infrastructure. The Global Pulse website reports 20 successful case studies on the use of Big Data and analytics in projects of sustainable development [5].

In October 2012, the Seminar of the High Level Group (HLG) on Streamlining Statistical Production and Services in St. Petersburg stated the need for "a document explaining the issues surrounding the use of Big Data in the official statistics community". In June 2013, the report "What does 'Big data' mean for official statistics?" was given by a Task Team of the HLG for the Modernisation of Statistical Production and Services to the Conference of European Statisticians [6]. In 2014, the HLG Big Data Project was started. The project, coordinated by the UNECE Secretariat, aims to analyse the major strategic questions posed by the emergence of Big Data [7]. A practical element of the project is a web-accessible environment for the storage and analysis of large-scale datasets, called the Big Data Sandbox, used for collaboration on practical projects across participating institutions [8]. The Big Data Inventory, also established within the HLG Big Data Project, is a website where any institution can give a report about its projects in standardized form [9].

In September 2013, the DGINS conference, the annual meeting of the heads of the European NSIs, adopted the Scheveningen Memorandum [10] resulting in the Big Data Action Plan and Roadmap 1.0 for the European Statistical System that was adopted in mid-2014 [2]. This action plan aims at the integration of Big Data sources into the production of European and national statistics. For the period 2016–2020, partnership models, the necessary IT architecture, and skills shall be developed, and pilot projects conducted. From 2020 onwards, the use of Big Data shall be fully integrated into official statistics. Eurostat has been instigating three major pilots, e.g., exploring the use of mobile phone data in tourism and population statistics [3].

In February 2013, a seminar on "Big Data for Policy, Development and Official Statistics" was held within the frame programme to the 44th Statistical Commission of the UN. Chief statisticians from India, Australia, The Netherlands, et al. as well as representatives from SAS, Google, and Amazon gave presentations of their views. In May 2014, the UN Global Working Group (GWG) on Big Data for Official Statistics was established. Aims of this GWG are to complement regional achievements, the provision of a strategic vision, the direction and coordination of a global programme, and the promotion of practical use of Big Data. Group members are six developed countries: Australia, Denmark, Italy, Mexico, The Netherlands, and USA; six developing countries: Bangladesh, China, Colombia, Morocco, Philippines, and Tanzania; and seven international organizations, among them the UNSD, UNECE, OECD, and the World Bank. The inaugurating conference "Big Data for Official Statistics" took place in Beijing in October 2014. The con-

ference adopted the Terms of Reference of the UN GWG and established eight task teams, five on general issues: advocacy and communication; Big Data and SDG indicators; access and partnerships; training, skills, capacity building; and cross-cutting issues (like a quality framework); and three task teams related to mobile phone data, satellite imagery, and social media data. The 2nd Global Conference on Big Data for Official Statistics took place in October 2015 in Abu Dhabi.

Several NSIs have started initiatives to explore whether Big Data provide opportunities to deliver a more efficient and effective statistical service. A good example is the Australian ABS. A Big Data Strategy was developed aiming at establishing an integrated multifaceted capability for systematically exploiting the potential value of Big Data for official statistics. The ABS Big Data Flagship Project is intended to coordinate research and development efforts that will build a sound methodological foundation for the mainstream use of Big Data in statistical production and analysis [11]. Among the other NSIs to be mentioned in this context are the Dutch CBS and the Italian ISTAT.

In September 2014, UNSD and UNECE conducted a survey on Big Data projects in statistical organizations [12]. From 78 NSIs and 28 international organizations who were invited, 32 NSIs and three international organizations responded. From them 37% worked already and 43% were planning to work with Big Data; 57 Big Data projects were reported.

4. Big Data: Potentials and challenges

According to Wikipedia, Big Data is "a blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications". This very general description allows for many different types of data; cf. Table 1. It is clear that it is not possible to give a general assessment of the potentials of such an enormous scope of data to be used in the production of official statistics. The potential of using any of these data sets in official statistics depends on the statistical product for which data are needed as input, and whether any inherent biases or measurement errors of an eligible data source make it unsuitable for being used properly or at all.

The expectations of official statistics in the use of Big Data are very high. Looking at the collection of data, the reduction to the response burden is an obvious goal and will go hand in hand with lower costs of the statistical production. A main advantage of most types of Big Data is their high actuality; the timeliness of statistical products may be improved. The availability of huge data volumes may make more detailed breakdowns of statistics possible as well as improved accuracy. Finally, the wide range of Big Data may allow the production of new statistical indicators.

On the other hand, a number of challenges are involved in using Big Data in official statistics. Challenges which are mentioned in the report of the Task Team of the HLG to the CES [6] are the following. In order to deal with the data eligible for the production of a certain statistics, the NSI must be competent in the methodological and technological issues related to these data: Suitable statistical and IT methods must be available, staff must be prepared, and policies and directives about the management and the protection of the data must be in force. Moreover, legislative requirements for getting access and using the data must be fulfilled, and the costs of sourcing the data must in balance with the benefits. Priority areas mentioned by the HLG Big Data report are partnership (with data owners), methodology and technology, skills, quality, and privacy, the latter encompassing managing the public trust and acceptance of (private) data re-use and its link to other sources.

The following discussion is restricted to aspects of the statistical production and is not dealing with issues of the statistical environment like legislative requirements, costs of sourcing of the data, or privacy. Most of the requirements on competencies and skills of the staff, the technology and the statistical methodology cannot be stated in general terms but are very specific for each type of data and the purpose for which the data are used. This will be illustrated in the following on the basis of four cases:

- Mobile positioning data for tourism flow statistics
- Web scraping data for tourism accommodation statistics
- Scanner data for price statistics
- Satellite images and sensing data for agricultural statistics

In all these cases, projects have been conducted, some of them resulting in routine application by NSIs.

4.1. Mobile positioning data for tourism flow statistics

The use of mobile positioning data for the production of tourism statistics has been extensively investi-

gated in the feasibility study instigated in 2012 by Eurostat [3]. Other projects on the use of mobile positioning data for tourism flow statistics were conducted by the Dutch CBS in cooperation with Vodafone, and by the NSI of Estonia.

An interesting project was conducted by the Dutch CBS where log data were registered by an app installed on the mobile phones of a group of respondents. The data were not specifically collected for tourism statistics but more generally for mobility statistics and ICT use statistics. However, Heerschap et al. [13] discuss the potential of the data collection approach for tourism statistics.

4.1.1. Mobile positioning data as complement for survey and administrative data

Each time when a mobile device is used for making a call, sending an SMS, or having a data session, one record of the mobile positioning data is generated, the Call Detail Record (CDR). The CDR contains the ID of the mobile device, the date and time of the contact, the kind of communication (call, SMS, data), the receiver of the call (call, SMS), and the location of the phone mast by which the communication is transmitted.

CDRs can be used to derive indicators of both domestic tourism flows and inbound flows. The positioning data allow estimating visiting frequencies of destinations but also durations of stay. Inbound flows are based on roaming data which obviously are only available if the SIM card of the mobile is from the country of usual residence of the visitor. CDRs are an excellent source of data for estimating indicators of same-day tourism flows. CDRs are also useful for estimating the number of overnight stays of the visitors, covering not only stays in registered accommodations but also, e.g., in private houses of a relative or friend.

Some methodological issues need to be taken into account when using the CDRs for estimating tourism flows.

- Classification: For estimating tourism flows it must be possible to recognize whether a CDR is generated within a touristic activity, and if yes, the type of activity. The identification of the type of touristic activity on the basis of CDRs might be erroneous and cause biases, e.g., over-estimation of the number of same-day trips.
- Representativity: The counts of CDRs reflect the habits in using mobile phones during travelling; over- and under-coverage of the number of visitors are possible and need to be investigated. Determinants of the use habits and counts of CDRs may be the costs of roaming service, the design of network of phone masts, and others.
- Combination of information: Tourism statistics are usually reported with breakdowns according to dimensions like the purpose of the trip, the usual environment of the visitor, the means of transport, and the type of accommodation. The CDRs need to be combined with corresponding data from other sources in order to allow standard reports. Such sources can be surveys among visitors and businesses and data from administrative bodies like immigration service.
- Quality issues: Assessing the quality of the CDRs and in particular of the statistical products which are obtained from the CDRs in combination with data from other sources requires experience with this type of data. Accuracy, comparability, relevance, and other dimensions of quality of the tourism flow statistics are to be reported as part of the corresponding metadata.

It is not the competencies and skills of the staff which are needed to deal with CDRs in the production of tourism flow statistics that are the challenge. Lessons have to be learned about the classification issues and about the possibilities to combine CDRs with data from other sources. Experience from pilots together with focused research will allow assessing the representativity and the quality of the resulting statistics.

A main conclusion of the Eurostat feasibility study was that mobile positioning data may complement in the future currently used methods of tourism statistics.

It should be mentioned that CDRs were similarly used in projects on migration statistics, population statistics, and passengers transport statistics.

4.1.2. Log data on the use of mobile phones

The pilot conducted by the Dutch CBS in 2012 to 2013 was based on data that were generated by an app that was installed on the phone or mobile device of a group of respondents; these log data combined with data provided by each of the respondents and data from other sources can be used to estimate indicators of tourism statistics. The approach that was developed in this project can be used for different purposes and was tested in two pilots, with focus not specifically on tourism statistics but on mobility statistics and ICT use statistics.

Basis to the approach is the close cooperation with the respondents. The respondent has to agree that an app is installed on his or her phone or mobile device.

In addition, the respondent has to provide background data such as age, sex, income, region, and composition of the family/group. These background data allow controlling the sample and weighting of the data in the estimation process. The app generates the main body of data, measuring every five minutes the location by GPS and thus allowing tracking the movements of a person through time. Moreover, specific questions – so-called pop-up questions – are triggered by the app on the basis of, e.g., a change in location. Information triggered by the app may contain the purpose of the journey, the mode of transport, the paid price, the type of accommodation, restaurant visits, satisfaction, activities, etc.

Based on the data generated by the app and in combination with data from other sources, tourism flow statistics can be estimated with detailed breakdowns. An issue of this approach is the representativity of results which is determined by the population of interest and the sample of cooperating respondents. In general, it is not easy to find persons who are willing to act as respondent. The data that were registered by the app can be expected to have a better quality than the same data collected through a survey.

Challenges are – besides the organization of the respondents and representativity – the costs of the data collection, the technology of the data handling, privacy concerns, and of course the quality of the resulting statistics.

4.2. Web scraping data in tourism statistics

Hotels, motels, youth hostels, bed and breakfasts, campsites, and other types of tourism accommodation use websites as a means to inform and attract potential guests. An NSI can collect the information of these websites in order to fill and update a tourism accommodation register and to derive tourism accommodation statistics. This activity is called web scraping. Tools for web scraping are so-called robots or web crawlers, software which crawls the internet gathering the desired information. The web crawler can be specialized to gather information from a specific website such as Tripadvisor.com, Booking.com, or other specific tourism websites where tourism accommodation providers post their information. Alternatively, scraping can address the internet in general or parts of the internet, looking for web pages which contain keywords related to tourism.

The use of web scraping data for creating or updating a database of tourism accommodations has been investigated in a pilot study conducted by the Dutch CBS [13]. The Italian ISTAT is using web scraping data for updating the farm register, completing the register with information on agritourism farms [14].

4.2.1. Web scraping data for tourism accommodation statistics

Tourism accommodation data that can be downloaded from related websites are: the name and the address, also characteristics like the number of rooms, prices, available facilities, chamber of commerce registration number, guest reviews, in some cases job vacancies.

A challenge in using web scraping data is first of all the processing of the data, cleaning them and extracting the relevant information:

- A consistent classification of the various types of tourism accommodations is crucial; the information provided on the website of an accommodation unit might be insufficient for a consistent classification of the unit.
- Names and addresses of the accommodation units need standardisation.
- A de-duplication procedure is needed to recognize a unit that is found on more than one website, sometimes under slightly modified spelling of name and address.
- Prices need to be clearly defined and comparable.
- The tourism accommodation register needs to contain the number of rooms but also the number of beds; the latter often is not mentioned in the touristic websites.

Another challenge is the representativity of the accommodation units found in the internet for the whole population of tourism accommodations. Assessment of the quality of both the web scraping data and the tourism accommodation statistical is another rather complex issue.

The pilot study conducted by the Dutch CBS in 2012 and 2013 came to the conclusion that information from the internet is currently not sufficient as the sole source of information to compile the population of units for the tourism accommodation statistics.

4.2.2. Other uses of web scraping data

Web scraping was also used in a project of the Italian ISTAT for collecting price data on consumer electronics and airfares for the production of consumer price indices [15]. Within Eurostat's programme of work, a project was started in 2013 on the use and analysis of internet prices; the intention was the development of a web scraping software that assists the Consumer

Price Index (CPI) specialists in the automated collection of prices. Related to these projects is the Massachusetts Institute of Technology (MIT) Billion Prices Project [16], an academic initiative with focus on research on high-frequency price dynamics and inflation measurement: Price indices are based on daily automated internet collection of prices. The BPP approach is used by PriceStats [17] to provide measures of inflation in 22 economies worldwide on a daily basis as well as other related products.

Within the Global Pulse Initiative, the project "Nowcasting food prices in Indonesia" was conducted [5]. This project investigates how people's self-reporting of commodity prices through Twitter can be used to provide real-time price indicators.

Technologies based on web scraping have been used in other projects such as for statistics of job vacancies [18] and for statistics of ICT usage by enterprises [19].

4.3. Scanner data for price statistics

In Section 2 the use of scanner data, i.e., price data recorded by point-of-sales in supermarkets, for estimating price indices is mentioned. Early attempts to use scanner data go back ten and more years. Within the EES task force "Multi-purpose consumer price statistics" [2], the NSIs of 17 European countries participated in the EU project "Scanner Data", experimenting with the use of scanner data for the production of CPIs. A task team of the Big Data Sandbox was experimenting on the computation of price indexes using the various tools that are available in the Sandbox [8].

NSIs use codes of the Classification of Individual Consumption according to Purpose or COICOP-codes as the standard for measuring CPIs. As scanner data are a by-product of transactions in the retail business, scanner data are available only for a few COICOP divisions like "Food and non-alcoholic beverages" and "Alcoholic beverages, tobacco and narcotics", items which have a weight of less than 15% in the market basket of the Austrian CPI, to give an example [20]. Obviously, scanner data can substitute only the corresponding minor parts of the currently used price collections which are represented in the CPIs. Nevertheless, the use of scanner data instead of price collections in the retail shops can improve both efficiency of the statistical production and the quality of the CPIs. Regional breakdowns may become possible as well as new products.

Several challenges in using scanner data need to be tackled. The major concern is representativity:

- The transactions recorded in supermarkets which use scanners may differ from the transactions in other retail shops. Such systematic deviations need to be taken into account in order to avoid a bias in the CPI.
- The big retail chains use rebates extensively as a marketing tool. The effects of rebates on the CPIs need to be investigated.

The retail industry uses International Article Numbers or EAN-codes in order to characterize the items. The EAN-codes are not harmonized with COICOP-codes, another potential source for biased CPIs.

Like in other cases of using alternative or Big Data, experiences from pilot studies will be needed to assess the quality of the scanner data and in particular of the statistical products which are obtained on the basis of these data, often in combination with data from other sources like price collections.

4.4. Satellite images and sensing data

For the production of agricultural statistics such as indicators on land use and crop yields, the data are collected in annual or sub-annual surveys and multi-annual agricultural censuses. An alternative data source is satellite data. A pioneer in using satellite data is the Australian ABS. The ABS has developed the methodology for predicting crop yields based on satellite data [21]. Other countries which have conducted projects on the use of satellite data for agricultural statistics are Mexico, Colombia, China, and Abu Dhabi. The Colombian NSI, DANE, has been using satellite images in conducting the agricultural census [22].

Satellite sensing data are measurements of the amount of light reflected by the agricultural objects such as fields planted by certain crops. A major challenge in using satellite sensing data is the interpretation of the satellite images and in particular the classification of the content. Classification of land use implies to distinguish between agriculture, forest, grassland, mixed use, non-agricultural use, and other uses. For land in agricultural use, the type and amount of crops that is grown and other data need to be identified. The translation of satellite sensing signals into crop production statistics requires statistical modelling and may raise interpretability issues. Tam [21] has suggested an approach to combine satellite sensing data with data provided by farmers in agricultural surveys at the unit record level for predicting crop yields.

Satellite images can be used in the production of statistics on land use, e.g., by identifying land parcels. Interimage is an open-source knowledge-based framework for automatic image interpretation; DANE has been using Interimage for object extraction in conducting the agricultural census.

Depending on the region for which the satellite data are used, data might be missing, e.g., due to cloud covers. Typically, satellite sensing data are available once every fortnight; this means that crop yield statistics can be produced with a much higher frequency on the basis of satellite data than on the basis of traditional data collections.

Like for other types of Big Data, assessing the various quality dimensions of alternative data and the resulting statistical products like accuracy, relevance, consistency, interpretability and timeliness needs to gain experience in using the satellite data.

4.5. Issues in using Big data: A summary

Issues in using Big Data in the statistical production that need special attention are specific for the type of data sources and for the use that is intended for the data.

For all applications of Big Data, the representativity of the available data for the population of interest is of major concern. The coverage of Big Data populations may deviate from the target populations, resulting in over- and under-coverage and in consequence in biased statistics. Data collected from the internet are tied to internet users, a population which may substantially deviate from the target population. For similar reasons, data related to users of mobile phones and to scanner data may lead to biased statistics. The combination of Big Data with data from surveys and registers may allow for compensating deficiencies in covering the target population. Careful analyses will help to understand the needs and find individual solutions for each application of Big Data.

The quality of Big Data and of the statistics that are produced on the basis of these data is another major issue for the NSIs. Like representativity issues, quality issues are specific for each dataset and application. A quality framework and quality criteria need to be designed or adapted for the various types of Big Data and statistical products that are based on these data. This should allow assessing consequences of using these data for the relevance of the statistical products, their accuracy, comparability, and other dimensions of quality. Metadata reports need to be adapted correspondingly. Analyses will be needed to assess the quality of Big Data and of the statistical products which are obtained on the basis of these data.

Classification problems are a serious issue, in particular for using internet data. This was illustrated above by the classification of tourism flows on the basis of CDRs. Another example is the classification of land use and grown crops on the basis of satellite sensing and images.

As mentioned above, further issues like legislative requirements, costs of sourcing of Big Data or privacy have to do with the statistical environment and only indirectly with the statistical production.

5. Conclusions

For a number of years, the notion Big Data has been finding great interest from many sides. This can be measured by the number of conferences, workshops, and other events, and the enormous amount of publications in journals, proceedings, and even books. Within official statistics, projects like the HLG Big Data Project, the ESS BIGD Project, or the Global Pulse Initiative as well as national initiatives like the ABS Big Data Flagship Project have been established, and quite substantial investments have been made to find out and clarify the potentials of using Big Data in various statistical domains.

Among statisticians, Big Data have raised expectations of various kinds: Reduction of the response burden and lower costs of the statistical production; improved timeliness of statistical products and more detailed breakdowns; higher accuracy; creation of new statistical products.

Various projects and pilot studies have been conducted and resulted in insights about and experiences with various types of data which are subsumed under Big Data. It quickly became obvious that no common methodological approach can be found for using the various types of Big Data in the production of official statistics. The challenges statisticians are confronted with are specific for each application and type of data. A uniform concept for dealing with Big Data is not visible.

Clearly, the easy availability of all these Big Data and the potential of many of these data for the production of official statistics will have the consequence that NSIs have and will find ways to properly use these data sources. The NSIs need to make preparations:

- Sound solutions for methodological issues have to be developed like suitable statistical methods for handling the various types of Big Data and for matching them with data from surveys and administrative sources.
- New skills of the staff will be required with respect to statistical methods and to IT tools for handling these data.
- A quality framework is required which contains standards for the assessment of quality dimensions of Big Data and of statistical products derived from them. Quality dimensions that need special attention in the context of Big Data are relevance, interpretability, and comparability.
- Further preparations for the use of Big Data concern the statistical environment, in particular legislation, partnerships, budget, privacy.

Official statistics cannot expect that Big Data will substitute actual data sources like data from surveys and administration. Experiences from pilots show that scanner data from retail trade, mobile positioning data, some sorts of web scraping data, and satellite images and sensing data can be used as supplements for existing data in the production of certain statistics. Implementing the use of such data will allow taking advantage of their high actuality and might reduce the response burden on some respondents. However, finding solutions for dealing with the various issues mentioned above and developing standards that will be accepted internationally may require substantial efforts and take some time. Exertions of the international statistical bodies will be crucial for ensuring that these and similar activities are well coordinated and the outputs efficiently communicated at the strategic level.

Acknowledgement

I am grateful for helpful comments from a referee.

References

[1] J. Rodriguez and F. Haraldsen, The Use of Scanner Data in the Norwegian CPI: The 'New' Index for Food and Non-Alcoholic Beverages, *Economic Survey* **4** (2006), 21–28.

[2] Eurostat. ESS Big Data Action Plan and Roadmap 1.0. 2014. Available from: www.cros-portal.eu/content/ess-big-data-action-plan-and-roadmap-10.

[3] Eurostat. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Consolidated Report. ISBN 978-92-79-39762-2. Luxembourg: Publications Office of the European Union, 2014.

[4] R. Kitchin, The opportunities, challenges and risks of big data for official statistics, *Statistical Journal of the IAOS* **31** (2015), 471–481; DOI 10.3233/SJI-150906.

[5] UNSD. UN Secretary-General's Global Pulse Initiative. 2015. Available from: www.unglobalpulse.org/big-data-development-case-studies.

[6] UNECE. What Does "Big Data" Mean for Official Statistics? 2013. Available from: www1.unece.org/stat/platform/display/hlgbas.

[7] S. Vale, International collaboration to understand the relevance of Big Data for official statistics, *Statistical Journal of the IAOS* **31** (2015), 159–163; DOI 10.3233/SJI-150889.

[8] UNECE. How big is Big Data? Exploring the role of Big Data in Official Statistics. 2014. Available from: www1.unece.org/stat/platform/display/bigdata/How+big+is+Big+Data.

[9] UNECE. The Big Data Inventory. 2014. Available from: www1.unece.org/stat/platform/display/BDI/UNECE+Big+Data+Inventory+Home.

[10] Scheveningen Memorandum. 2013. Available from: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version_0.pdf.

[11] S.M. Tam and F. Clarke, Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics, *International Statistical Review* **83** (2015), 436–448.

[12] UNSD. Big Data Project Survey. UNSD/UNECE joint survey – Fall 2014. 2015. Available from: unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf.

[13] N. Heerschap, O. Shirley, A. Priem and M. Offermans, Innovation of tourism statistics through the use of new big data sources, *The Hague: Statistics Netherlands*, 2014.

[14] G. Barcaroli, D. Fusco, P. Giordano, M. Greco, V. Moretti, P. Righi and M. Scarnò, Istat Farm Register: Data Collection by Using Web Scraping for Agritourism Farms. Unpublished data, 2015.

[15] G. Giannini, R. Lo Conte, S. Mosca, F. Polidoro and F. Rossetti, Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation, *Statistical Journal of the IAOS* **31** (2015), 165–176.

[16] BPP. The Billion Prices Project @ MIT. 2015. Available from: http://bpp.mit.edu/.

[17] Price Stats. 2015. Available from: http://www.pricestats.com/.

[18] A. Virgillito, Experiment report: Job Vacancies. 2014. Available from: http://www1.unece.org/stat/platform/display/bigdata/Experiment+report%3A++Job+Vacancies.

[19] D. Summa, Experiment report: Web Scraping. 2014. Available from: http://www1.unece.org/stat/platform/display/bigdata/Experiment+report%3A++Web+Scraping.

[20] Statistik Austria. Warenkorb und Gewichtung des HVPI 2015. Available from: http://www.statistik.at/web_de/statistiken/wirtschaft/preise/verbraucherpreisindex_vpi_hvpi/warenkorb_und_gewichtung/index.html.

[21] S.M. Tam, A Statistical Framework for Analysing Big Data, *The Survey Statistician* **72** (2015), 36–51.

[22] S.Y. Rodriguez Figueroa and S.L. Moreno Mayorga, Big data for the National Agricultural Census, Colombia 2014. Unpublished data, 2015.