JOHANNA GICZI AND KATALIN SZŐKE*

Official Statistics and Big Data

Intersections. EEJSP
4(1): 159-182.
DOI: 10.17356/ieejsp.v4i1.408
http://intersections.tk.mta.hu

* [giczi@tatk.elte.hu] (ELTE TáTK, Budapest); (Researcher, Central Bank of Hungary)

Abstract

More than five years ago Eurostat started a project with the aim to 'tame' sources of Big Data in a way that they can be incoporated into official statistical systems. In order to solve the problems a statistician might be faced with during the official statistical application of Big Data, first of all, we give an overview of traditional data collection, and then point to the differences one has to face when dealing with Big Data. We introduce common sources of data (traditional, administrative) and highlight the ways huge sets of data are different compared to them. Next, we discuss characteristics of Big Data versus traditional statistical methods based on the qualitative criteria of official statistics, and we also elaborate on the problems of analysing Big Data. Finally, we provide a list of use cases for Big Data in official statistical data collections.

Keywords: Big Data, official Statistics, big Data Paradigm, Big Data quality standards.

Opinions differ concerning how much data there is in the world. Some IT-specialists say that 2.5 exabytes (2.5 x 10^{18} B or 2.5 EB) of data are created every day (5 exabytes would be enough to store all words ever spoken by human beings), while the experts of IBM estimate that today the total amount of data in the world is doubled every two years – that is, today in 24 days the same amount of data is created as previously throughout our entire history. From this it is clear that thanks to the development of the information and communication technologies a huge amount of data is created, and for data scientists it is a waste not to exploit this huge amount. Big Data is special not only because of the great amount of data but also – mainly due to the spread of social media and mobile phone services – their changeable nature. Although today with development of technology it is increasingly possible to collect, process, store and organise this large amount of data, official statistics is still struggling to hammer out and apply methodologies that are based on Big Data. Almost five years ago, Eurostat – the main statistical organisation of the EU – started a project the aim of which is to 'tame' sources of Big Data in a way that they can be incorporated into official statistical systems.[1] Furthermore, by exploiting the opportunities provided by these large sets of data, the project also aims at producing statistical data faster and of a better quality than before in order to make data analyses more diverse and, in some cases, more detailed, and to draw more accurate conclusions and prepare more accurate forecasts, at the same time decreasing the burden on data providers. To achieve these goals, Big Data must become an integral part of official statistical data collection.

1. Definitions and the taxonomy of Big Data

The Oxford Dictionaries define Big Data as follows: 'Extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.'[2] Meanwhile in Wikipedia one finds: 'Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.'[3] Gartner, Inc. (2017) provides the following definition: 'Big Data are high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.' Based on the above, Big Data can be characterised by the following three concepts (the 3V-definition): volume, variety, velocity (Glasson et al., 2013) Besides the 3V-definition, the literature mentions other characteristics as well: from the point of view of statistics the velocity of data is of special importance. This term refers to how good the quality of the data is, and to what extent they reflect reality (DeVan, 2016). In the case of the traditional, statistical way of data collection high data quality

[1] This issue has special importance within the EU, as is also illustrated by the fact that one of the basic projects of the European Statistical System is the 'ESS. VIP Big Data' dealing with the development of Big Data for official statistical purposes.

[2] https://en.oxforddictionaries.com/definition/big_data

[3] https://en.wikipedia.org/wiki/Big_data

is one of the most important criteria, however at the same time it is also a great challenge.[4]

Big Data sources can be grouped in several ways (Glasson et al., 2013). First of all, as mentioned earlier, based on their generation, there are three types of data (Vale, 2013). These three types are also different in terms of the participants of the communication process: human-sourced information[5] is created through communication between human beings, process-mediated data[6] result from communication between human and machine, while machine generated data[7] are of course the result of communication between machines.

2. The Big Data Paradigm

From the foregoing, it is already quite clear that compared to traditional data collection procedures and methods, Big Data is of a completely different nature, and operates based on a different 'logic'. This problem includes questions of information technology and professional dilemmas. The former is not discussed here in detail, it is only touched upon in a couple of points.

[4] Concerning the definition of Big Data variability, visualisation, value (that is valuable, useful results from the data), validity and volatility (that is the period of validity) are also important (DeVan, 2016).

[5] Human-sourced information means the subjective records of human experiences. Earlier these records were stored in the form of books, works of art, then photographs, videos and audio storage devices, but today in almost all cases they are generated digitally (on personal computers or social media). Official statistics only have limited access to these, usually weakly structured, uncontrolled sets of data. Facebook comments, likes and posts, tweets, blogs, vlogs, personal documents, pictures and videos shared on photo or video sharing sites (Pinterest, Instagram, YouTube), searches on the Internet, text messages sent by mobile phones and emails belong in this category.

[6] Process-mediated/transaction data are data generated in the course of certain (mainly business) processes. These are well-structured, usually RDBMS-data (data from relational database management systems) or metadata. One type of these data come from databases of usually governmental institutions (e.g. public offices), electronic healthcare databases, medical records, databases of hospital visits, insurance records, data of banks and stock data, business data of enterprises (if legislation requires the storage of their records). Transaction data form another type of process-mediated data: they are generated through transactions between two entities. These are, for example, commercial transactions (e.g. online shopping), debit and credit card transactions, data from e-commerce (including transactions launched from mobile devices), etc.

[7] Machine generated data are typically referred when the phenomena behind the trendy expression IoT (the Internet of Things) is being discussed. Log files and data of fixed and moving sensors also belong in this category. This type of information could be defined as the data of billions of sensors measuring and recording the events of the physical worlds. As globally more and more sensors are introduced and activated, the amount of data created by sensors is growing, too. All in all this is the type of data the amount of which is growing the fastest. Sensor data are data of sensors on household devices, weather or air pollution sensors, satellite data or data of traffic monitoring systems/web-cameras. Also there are data generated by tracking devices such as route/tracking or geolocation data of mobile phones (e.g. GPS data). Data from computer systems (logs and web-server logs) are created in the course of the operation of computers, in a text format, about system events.

2.1 Issues of information technology

It is easy to see that in order to be able to manage Big Data, to collect, store, prepare and process the continuously flowing information certain prerequisites must be fulfilled.

1) It is necessary to have an increased computing capacity, which can be managed with MPP-solutions (massive parallel processing).

2) Also, tools of data redistribution and parallel processing are indispensable (knowledge and usability of MapReduce, Hadoop, Hortonworks Data Platform, RStudio, etc.).[8]

3) Furthermore, it is very important to use and know data reduction software technologies, which are not exclusively based on SQL (structured query language). At the same time for statisticians it is always a question what type of IT-skills are necessary for the application of Big Data.

2.2 Professional issues

In order to clear up problems which a statistician might be faced with during the official statistical application of Big Data, first of all we shall give an overview of traditional data collection, and also the differences one has to face when dealing with Big Data. After that we introduce the types of data sources and we shall point out in what way huge sets of data are different compared to them. Finally we shall discuss the characteristics of Big Data versus traditional statistical methods based on the qualitative criteria of official statistics, and also we shall elaborate on the problems of analysing Big Data.

2.3 Data collection and Big Data

Traditional approaches: the top-down paradigm. In accordance with the general practices of official statistics, before any data collection takes place, one must specify what kind of information they wish to collect, and for this hypotheses must be set up. After that one needs to go through the following steps: (1) design the data collection, (2) collect the data, (3) preparation of the data, (4) data analysis, (5) obtain information from the database leading to rejection or confirmation of the hypotheses.

In the top-down paradigm the researcher focuses on specifying the goal(s) of analysis during the design of the data collection. Planning, which is a key element of traditional data collection, involves the following: (1) creating the variables and definitions, conceptualisation and operationalisation, (2) choosing the statistical population (this can be a complete population or it can also be based on sampling), (3) application of lists or registries in order to reach the population, (4) the preparation of typologies and questionnaires.

[8] The programme of the SETI Institute is an excellent example of parallel processing, where with the help of volunteers, signals arriving from space are analysed in the search of traces of extra-terrestrial intelligence (or at least some patterns). The volunteers run the software on their own computers thereby multiplying the speed of the analysis.

For achieving the analytical goals, one must obtain specific information and set up (a) specific hypothesis/hypotheses. After that a model must be set up. At the end of the process one can provide some descriptive statistics, an estimate or a forecast. The Big Data approach: the bottom-up paradigm In the case of the Big Data paradigm one needs to follow a completely different logical pattern. Here data collection does not have to be designed in the traditional way (this is because the data are already given, or rather: they are present everywhere), and the usual order is turned upside down. In this case, instead of planning, one must start by (1) collecting the data, and then (2) prepare these data. The next step is (3) data mining (this usually means looking for correlations) and (4) customising the algorithms (by choosing scalable algorithms in the first place and avoiding aggregation). The final step is the discovery of new knowledge and the validation of the results (using heuristic [pattern search] technologies for forecasts/estimates).

In the case of this latter approach emphasis is laid on discovering accessible data in order to find valuable information which has not been extracted from the data yet. Obviously this methodology mainly offers solutions for problems researched by data scientists,[9] who are interested in questions like 'What is happening?' instead of 'Why?' and 'How?'. Due to the aforementioned special characteristics of Big Data, it is quite difficult to integrate it into the framework of traditional statistics.

2.4 Comparison of the types of data sources

After discussing the data collection paradigms, let us take a look at based on what factors we can define primary and secondary statistical data sources. Primary data sources are data collections where questionnaires are used (e.g. censuses), irrespectively of whether data collection is complete or is based on sampling. Secondary data sources are Big Data-type data sources, consisting of data collected from administrative sources.[10] Chart 1 summarises the main aspects of how different types of data sources are described.

Chart 1: The characteristics of data sources

Characteristic	**Primary statistical data source**	**Secondary statistical data source**	
		Administrative data source	**Big Data-type data source**
The data are designed for (a) statistical purpose(s)	yes	no	no
The concepts, definitions and typologies are clear and well-known	yes	often	rarely
The target population is well-defined	yes	often	no
Metadata are provided	yes	often	no
The data are structured	yes	yes	rarely

[9] This is a relatively new profession, which requires mathematical and statistical knowledge, programming skills, as well as experience and reliable expert knowledge in the given field.

[10] Istat (the Italian National Institute of Statistics) refers to these as tertiary data sources.

Characteristic	Primary statistical data source	Secondary statistical data source	
		Administrative data source	Big Data-type data source
The data apply to the statistical population	yes	usually	no
A preliminary processing is necessary for 'extracting' the statistical data	no	no	yes
Relevant/interesting data are directly accessible	yes	often	no
Extra variables are directly accessible	yes	often	no
The data completely cover the statistical population to be examined	yes (census) no (survey)	often	not yet
The data are representative or can be rendered representative or certain analyses	yes	often	no

Source: Istat ESTP, 2016.

It turns out from the chart that – considering their adequateness in terms of definition or typology, or (in some cases) also their structuredness – Big Data-type data sources do not live up to the requirements set up for traditional data collection methods completely. If we want Big Data to fulfil the quality requirements of official statistics, we need to improve the definability of the statistical population and the specificity of the target population (problems of 'data coverage'). If we manage to improve on the above criteria, that is, to measure their quality and to incorporate appropriate methods in the statistical data generation process, then we will also come closer to solving the problem of representativeness. The definition of metadata and extra variables also depends on how we manage to solve the issues mentioned above. For this it is necessary to increase computing capacities, to re-discuss ethical dilemmas and dilemmas of data protection, to set up a new set of rules and to discuss these issues at an international level.

2.5 Big Data and quality standards

In official statistics data and institutions must fulfil a number of quality standards. Out of these standards we shall now take a look at those that might be relevant when applying Big Data.

In official statistics representativeness is of key importance. Well-chosen representative samples used in traditional sampling describe the population quite well. In the case of Big Data the data are already given, however, according to the statistical definition they are not complete at all. Concerning the complete target population the problem of lack of coverage or over-coverage might occur, which in turn leads to distortion. Therefore Big Data sources can be seen as non-representative databases, for which reference data, necessary for the examination of validity, are essential. The selectivity/representativeness index is also important.[11] This shows how data from the Big Data source are different from the actual population. By setting up the criteria of 'ignorability' it is possible to address distortions related to coverage, sampling, measurement and respondents (for more information, see: Couper, 2013).

In official statistics, the comparability of data is another key factor. Due to the fact that in different countries statisticians work with differing concepts (e.g. the definitions of household, family or unemployment are not identical), sometimes it is difficult to render different specific statistics comparable in terms of time or space. When it comes to the comparability of Big Data and data from traditional collections, we have to face similar challenges:

- Differences between definitions. In official statistics, thanks to the harmonisation of official statistical services and EU and national standards and rules, the concept to be measured is defined very precisely. However, we have to keep in mind that the concept behind a variable set up based on a Big Data-type source is usually not identical to the concept used in official statistics, therefore our primary task is to harmonise the different concept structures.

- The concept of population. In official statistics a (statistical) population is a set of items to be examined. These items can be characterised by providing their features. In the case of Big Data-type sources, the accessible population is usually not identical to the population to be characterised. Therefore, we need to establish methodologies for producing the latter based on the former. For example, in the case of a mobile service provider if the given population consists of the users of the service, but the population to be examined is somewhat different, then the two concepts are not identical, and distortion will occur (if the population we wish to examine consists of the inhabitants of Hungary, then there will be persons whom it will be impossible to observe through this way [e.g. those <children or elderly people> who do not have a mobile subscription], and also, there might also be people who have several accounts. This way the population to be examined will be distorted).

- The concept of the statistical unit. Based on the collected data official statistics defines, analyses and gives information about different units or groups of units. However, when applying Big Data, we must check if every piece of information necessary for the management of different statistical units is available. This is important because the relevance and statistical units of a Big Data source are different from those of traditional statistics, so in order to produce the necessary information further methods and models are needed. A good example of this is when in a Big Data-type data source the statistical units are the cell phone subscriptions and not the people. In this case, as the statisticians would like to make statements about the behaviour and habits of the people, it might pose a problem that some people have more subscriptions, while some do not have any.

[11] Selectivity/representativeness is one of the most important dimensions of concern. A non-representative database may be useful for certain purposes but not suitable for others. The question is whether there are reference data that can be used to validate the validity.

2.6 Methodological issues: the pros and cons of using Big Data

Chart 2 summarises and complements the foregoing by introducing those experiences which are for and against the usage of Big Data in official statistics.

Chart 2: Arguments for and against the Big Data Methodology

Pros	**Cons/Challenges**
No sample	No sample – representativeness
Real time	Coverage (overcoverage/lack of coverage)→distortion
Real behaviour, not based on self-declaration	Measurement of the quality of input and output data
Less burden for the respondent	The method of using the data source, loss of potential validating data source
Can be combined with other databases	Comparability (with current statistics)
Discovery of new knowledge	IT infrastructure, support
Expenses (in the long run)	Expenses (in the short term)
	Accessibility of the data
	Data protection
	Stability

Source: Our own compilation.

In Chart 2 the lack of samples is listed among the pros, as well as the cons of Big Data. On the one hand it is good, as the possibility of sampling errors can be excluded. However, on the other hand, as pointed out earlier, from the point of view of representativeness this is a problem. As in this case, we know very little about the statistical population, and therefore the identity of the sample units is not clear either. Without a deep knowledge about the statistical sample it cannot be guaranteed that the researcher can produce statistics which apply to the whole target population; this means that one of the main advantages of qualitative research is lost here.

The issue of stability is also among the challenges: this is also a problem for traditional data collection (e.g. a high level of non-response rate can destabilise a survey). Big Data is a flow of data which can change easily and fast; therefore, it might happen that a homepage ceases to exist, someone deletes the data-collecting application from their smartphone or they deny access to their device, etc.

At the same time, it is a clear advantage of Big Data that it is real-time. These types of data might even be accessed real-time and they can be analysed much faster than data collected in the traditional way. However, the collection of real-time data can also be blocked, and as a result this advantage can be lost. This is because for official statistical institutions it is a problem that Big Data are possessed by other institutions, organisations or persons, therefore in some cases it is costly to access them, and access to real-time, individual and personal data can be difficult due to ethical or data protection issues. In the majority of cases these data are highly unstructured, and in almost all of the cases there are also 'useless' pieces of data among them. Therefore, if this 'noise' cannot be excluded from the database, the quick usability of the data might be hindered.

Another advantage of Big Data is that in contrast to self-declared data, they show people's real behaviour. This makes it possible to avoid some significant sampling errors of traditional data collection (such as non-response, respondent error, distortion and effect of the interviewer).[12]

Today – as mentioned earlier – it is an important goal of traditional statistics to reduce the burden on respondents. Therefore, if the responses to the questions of surveys are partly or completely also available from (an)other data source(s) (such as Big Data or administrative data sources) or if they can be deduced from them, then the respondent is not saddled with the questionnaire and expenses are reduced, too.

It also supports the usage of Big Data that these types of databases can easily be combined with other databases. According to the present state of knowledge, Big Data sources can be used in official statistics in a complementary or validating function, with the help of data fusion procedures

2.7 Problems of analysis

Although we have already shown several Big Data-related problems, here we would like to present some more problems which one might be faced with during the analysis of data.

In the case of Big Data traditional procedures of data analysis do not work. Firstly, during the analysis of huge amounts of data one will be faced with the limits of complexity and computing capacity (e.g. matrix inversion, the principle of least squares estimators, GLM maximum likelihood vs. Newton–Raphson algorithm). Most traditional algorithms are difficult to parallelise, so it is very cumbersome to have several processors work on the details at the same time (e.g. Hadoop cannot cope with such a task). However, due to the size of the data set computing capacities cannot be increased in any other way. Secondly traditional statistical procedures are very sensitive to data errors and extreme values, so it is obligatory to perform checks and data cleansing. However, a significant part of Big Data is 'noisy' and unstructured, what is more, the size of the data set is so huge, that it is impossible to edit, add new input or manage outliers in a simple way. The management of duplicates is also a difficulty. In statistical offices, the traditional data management procedures, the well-structured monitoring systems and data cleansing algorithms can guarantee that the databases generated from surveys do not contain any duplicates. In the case of Big Data, different types of procedures must be created for this.

It is also a problem that the majority of Big Data-based analyses rely on the examination of correlations. However, this method also involves the possibility of false correlations (see ecological misconceptions); and if correlations are not clear, then this leads to the 'death of the cause' (Scannapieco, Virgillito and Zardetto, 2013).

If we use Big Data for statistical purposes, thereby substituting data based on statistical data collection, then the aforementioned problems will be more nuanced, as Big Data will also have to be subjected to the same procedures (data preparation,

[12] For the sake of completeness here it must be mentioned that Big Data might involve certain distortions which are unknown and therefore cannot be corrected.

micro-validation, outlier management and aggregation) as data used in traditional data collection. However, complexity and capacity still pose a problem.

The currently used methods of official statistics (designed sampling methods built upon models, regression, general linear models, etc.), which are successful or unsuccessful depending on specific characteristics of traditional baseline data are capable of managing and analysing data of high quality, but (compared to Big Data) in very small quantities.

Based on the foregoing it might seem that the currently used methods of analysis have nothing to do with Big Data. What could then be the solution? Literature agrees that in order to be able to manage Big Data, we need a radical paradigm shift in statistical methodology:

- We need to use robust procedures – even if to a certain extent they erode accuracy: At the same time the criteria of accuracy and quality have to be laid down in every case: The level of accuracy can only decline if at the same time other qualitative components are improved:
- The methodology for analysing Big Data have to rest on approximating and not exact techniques, which can cope with the noisy target functions.[13]
- We need a paradigm shift. We need to accept that Big Data allows different types of analyses (Scannapieco, Virgillito and Zardetto, 2013).

However, these compromises are also affected by the fact that even if we want to develop, produce or publish official statistics based on Big Data only partially, then we have to make the latter live up to the same requirements as official statistics. As a result, the moment Big Data becomes part of official statistics, it will (partly) lose its Big Data-nature.

3. The utilisation of Big Data in official statistics: international experiences

The Italian and Dutch statistical offices are leading in developing and using methods based on Big Data. Here we shall introduce some projects the results of which are already being used successfully by official statistical service providers.

3.1 The analysis of social media by official statistics

In the Netherlands ca. 70 per cent of the population use one or more social media sites (Daas and van der Loo, 2013), out of which Facebook and Twitter are the most popular. Research studies have analysed tweets posted on the Dutch Twitter – the forum where the most public Dutch language content is available – with the aim of

[13] Very often the quadratic function (also in the case of Big Data) is determined by measurement data containing measurement errors (noise) to a greater or lesser extent. The target function can be estimated directly from the raw measurement data (historic estimate) or in a way that a distribution function is run on the raw data first (parametric estimate).

finding out more about the correlation between their content and 'general mood'.[14] The tweets showed a strong correlation between the general mood of the public and the economic situation and consumer confidence.[15] The correlation with the situation of the economy was so strong that the researchers examined it on a weekly and monthly basis, (As a criticism of the results, like several other studies, such as Pléh and Unoka, 2016, we would like to point out that social media posts do not necessarily reflect the real opinion of the individual, they rather show a conformity with an expected norm, which presents the given person in a more positive light compared to reality.[16])

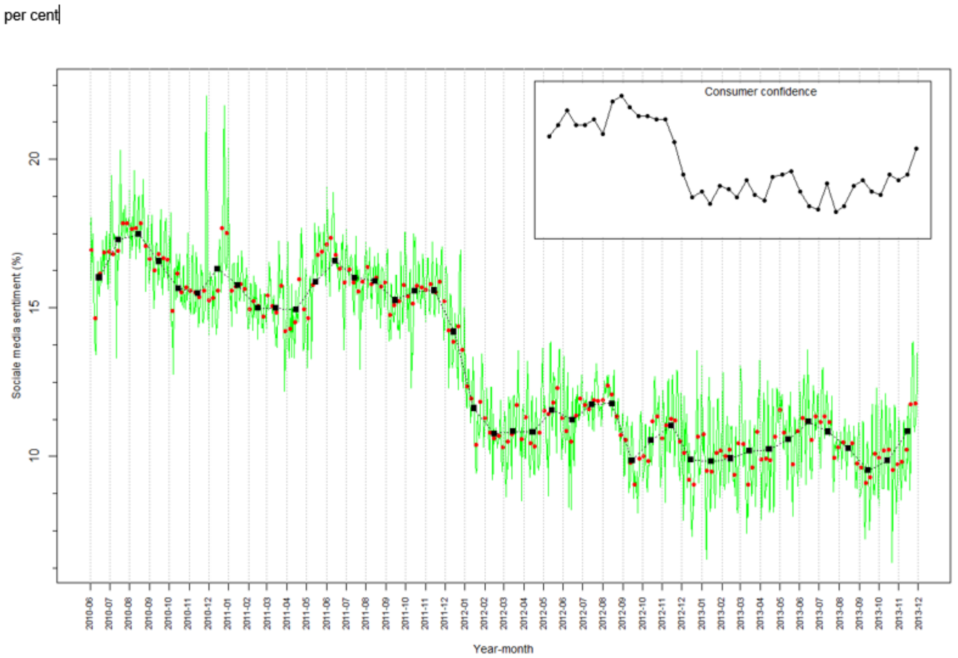

Figure 1. Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. In the insert the development of consumer confidence is shown for the identical period.
Source: Daas and Puts, 2014.

[14] From the analyses, it also turns out that almost 50 per cent of dialogues are pointless 'babbling'; the topic of the remaining discussions was about free time (10 per cent), work (7 per cent), TV and radio (5 per cent) and politics (3 per cent).

[15] In sociology, the association between particular and generalised trust is a debated issue. While the former brings economic development, the latter inhibits it (Fukuyama, 1995; Knack and Keefer, 1997; Raiser et al., 2001). In official statistics, the level of trust is dealt with by well-being studies.

[16] The social media messages sent in December were much more positive than those sent in the periods before and after.

Dash: daily aggregates
Round dot: weekly aggregated

3.2 The application of data generated by sensors in official statistics

In the Netherlands traffic sensors collect data about a road network of more than 6,000 km, process almost 24 thousand pieces of data a minute, and publish them within 75 seconds, as a result making it possible for drivers to avoid traffic jams and increasing road safety. By analysing these detailed data, official statistics can make important traffic estimates (e.g. it is possible to give estimates about the number of personal cars moving into and out of the country, broken down by vehicle types, as well as the amount of freight traffic, broken down by nationality of the vehicle) and complement their own data collections. Currently it is a problem of the project that with traditional statistical tools it is possible to process a day's amount of sensor data, however, in order to process an amount of three months, one needs to use Big Data-type tools.[17] In the future, this problem can be solved by increasing computing capacities.

Satellites also provide sensor data. By using them, it is possible to check the characteristics of land use frequently and accurately. The Austrian statistical office's land use study is based on satellite images, and is a very good example for this (Tam and Clarke, 2015). This study provides useful data for the whole of official statistics, as well as environmental statistics. Satellite images are used for identifying what land is used for also in Australia. In the framework of the research programme the images are analysed according to land use characteristics in order to estimate the ratio of crops. Land use characteristics are identified with the help of an algorithm (Daas and van der Loo, 2013).

3.3 The application of data generated by mobile devices in official statistics

In the framework of a consortium project the Bank of Estonia, the University of Tartu and Positium LBS have established a solution for using data generated by mobile devices for the purposes of official statistics. Positium LBS (which was founded exclusively for this purpose) collects and processes anonymous data of mobile service providers, thereby providing reliable information on border crossings (into and out of Estonia). With the help of a PDM software (Product Data Management software) – which is partly in operation within the system of the service providers and is controlled by them, and is partly controlled by Positium LBS as data mediator – business secrets and personal data are protected, as the respondents receive a randomly chosen nickname/code, and it is impossible to identify the telephone number/owner. Data collection is based on active and passive tracking. In the case of active tracking the mobile devices are located and tracked using the MPS (Mobile Positioning System), while in the case of smart phones this is done with the help of GPS. In the case of passive tracking (which is mainly used for inside business or marketing purposes) the data are automatically stored at the mobile service providers (either in a memory or a log file).

[17] In one minute ca. 460,000, in one hour ca. 27 million, in a day ca. 600 million and in a year ca. 240 trillion pieces of data are generated.

The cooperation of these three institutions is hampered by the fact that mobile service providers are sales driven, and they are interested in keeping their customers, whose secrets they would like to keep in order to retain their credibility. Therefore, the partners have to face several professional, methodological and legal challenges in order to collect the positioning data.

After the collection of the data Positium LBS concludes a quality assessment, during which characteristic errors must be filtered out and corrected. After that the company performs spatial interpolation by using a special geographical information module. The collection of statistics by analysing the usage of mobile devices in time and space shows several methodological characteristics. The use of mobile devices is equally widespread in developed and developing countries, irrespectively of income, age or other social characteristics (but of course depending on network coverage and density). Thanks to this data can be collected easily and from a broad range of users. The method is cost-effective, as the results are recorded automatically, and as the respondents do not have to be contacted directly (like in traditional data collection), this reduces the costs even further (Daas and van der Loo, 2013).

4. The possible use of Big Data in the Central Statistical Office

As we have shown earlier – some of the European statistical offices are already using Big Data-based methods or combining Big Data and traditional techniques. Now we shall introduce those areas where the Hungarian Central Statistical Office could successfully use the results of the above-mentioned projects – either by complementing data collection with Big Data sources or by validating the data collected so far. We shall list the possibilities of use according to the types of Big Data sources: transactions generated by the Internet, sensors or processes (see: Appendix F1).

4.1 The possibilities of using data generated by mobile communication in Hungarian official statistics

The Hungarian Central Statistical Office collects data on domestic and international migration on a yearly basis, in the framework of the OSAP programme (National Statistical Data Collection Programme). The data sources are the census, the micro-census and the LUSZ programme (about the population's travelling habits). The first two are held every ten years, while the latter is held annually. However, if we used positioning data, we could get information on the mobility of the population more often, even if these data would not be complete due to problems of network coverage mentioned earlier.

Data collected from mobile phones could also be used in tourism statistics collected more frequently than a year or in creating an estimating process, which is more accurate than today's methods. Therefore, they could be used especially well in monitoring border crossings. This would be useful as with the establishment of the Schengen area and the cessation of border control we have less information on the number and nationality of border crossers than earlier. For these problems – similarly to the Estonian project shown earlier – access to mobile phone information could provide a solution. However, due to Hungarian data protection legislation access to these data is highly problematic, and (currently) it is also very costly, although this area holds a lot of opportunities (and as pointed out earlier, Estonian migration research already uses passive positioning data in practice).

Currently there is already a methodological research study in the pipeline at the Hungarian Central Statistical Office dealing with the usage of mobile phone data, in the framework of a Eurostat grant, with an aim of using these mobile data instead of the data collected for our time balance survey collected in the traditional way. In order to achieve this, we would like to create a mobile app, with the help of which we will be able to compare the GPS data of smartphones with the answers of persons who fill in the time balance survey.

4.2 The possibilities of using sensor data in Hungarian official statistics

In surveying travelling and mobility we could use sensor-based Big Data. As we have shown earlier, the Netherlands has a road sensor network, and Hungary has more and more such sources (e.g. the video footage of the National Toll Payment Services Plc. or the National Police Headquarters). Based on this, by using the appropriate coding and data protection techniques not only the number of border crossings, but also migration, commuting and tourist habits could be observed.

In terms of transportation and vehicle statistics, the data are often produced by collecting data from administrative data sources. By using sensors, this could be done more often and faster and the data could be analysed based on other criteria, too (area, nationality, type). Sensor data could be used not only in official statistics but also in other areas like city planning or the transformation of transportation.

Smart meters are capable of storing the data collected from the environment (temperature, air pressure or CO2 levels), and these data could serve the work of energy and environmental statistics. This type of problem is rooted in the fact that although the sensors are already in operation and make measurements every second, data are forwarded less frequently (in the majority of cases the Central Statistical Office receives monthly, quarterly or yearly data). However, if these data could be connected with other types of data, one could get a real-time view of a city's operation, and this way energy consumption and traffic could be measured based on different dimensions, too.[18]

4.3 The possibilities of using the web crawling method in Hungarian official statistics

Currently a significant part of the data needed for calculating the price index is collected by the colleagues of the Central Statistical Office in some appointed shops, and a smaller part of these data comes from on-line surfaces. However, with the help of the web crawling method, where with a special software, data can be scraped from web surfaces in a structured format, the data of real estate agencies could be

[18] The Senseable City Lab project of the Massachusetts Institute of Technology monitors the events of the city in real time (energy consumption, traffic) with the help of sensors.

downloaded and the changes of the real estate market could be estimated. The office could obtain information on the most components of the index of consumer prices the same way.

Furthermore, by web crawling the homepages of companies the office could collect information on infocommunication tools used by businesses. Currently this type of information is collected through questionnaires.[19] Based on the experiences of the Italian National Institute of Statistics data collected through web crawling can complement traditional data collections quite well (Barcaroli et al., 2014).

Using this method, it is possible to gain information on job vacancies from job sites (in which city/at what type of company what kind of jobs are needed). These data can be used for the statistical estimation of the number of job vacancies.

4.4 The possibilities of using process generated data in Hungarian official statistics

In official data collections, the Central Statistical Office's household budget data collection serves as a basis for examining the consumption patterns of Hungarian households. The respondents in the sample must write a consumption diary throughout the year, where they are asked to make an itemized list about the quantity and price of the purchased products. This is a significant burden on the shoulders of the respondents, however, without such a data collection we would not have any information on the consumption characteristics and expenditures of the households. However, process generated data (e.g. data resulting from bank card payments or sales data of the shops, where we can see when, where, for how much and how many items were purchased) would significantly improve the quality and accuracy of statistical data on consumption.

The cash machine data of the National Tax and Customs Administration are also process generated data. Based on their quantity and frequency, these data can be considered as Big Data, and they could complement the data collections of the Central Statistical Office.

4.5 Monitoring of the daily air ticket prices in Hungary and the calculation of prices, using the web crawling method

The monitoring of air ticket prices is one of the small components in calculating the index of consumer prices[20] in Hungary. Currently for this purpose data are collected manually, but this way only a limited amount of data can be collected, and it is not possible to closely follow the fast changes of prices. The monitoring of air tickets with a Big Data-based method offers a solution for these problems. A current project at the Central Statistical Office uses the web crawling method to automatically collect data through Google. The frequency of 'data crawling' can be set depending on

[19] OSAP 1840: Az információs és kommunikációs technológiák állományának minőségi és mennyiségi adatai.

[20] The index of consumer prices is an indicator showing the average changes of the products and services purchased by the population (the households), that is the changes of the level of consumer prices.

the purpose of the research: price changes can be checked several times a day or even every hour or every minute.

With this project the Central Statistical Office intends to 'reproduce' the current, manual data collection method. The web crawling method not only accelerates but also simplifies the process, that is, besides improving quality this reform of data collection improves efficacy in itself (this method does not require any human resources, and the released labour force can be used for other types of analysing and development tasks). This way it is possible to gain information quickly and in a cost effective way (that is for free) and the daily changes of the prices can be compared.

It is a disadvantage of the method that currently the colleagues at the office are not monitoring all of the web pages selling air tickets.

5. Methodology

The project monitors the prices of tickets to four destinations from Budapest to Rome, Berlin, London and Paris. The travelling period was set as days 10-11 of the month, plus/minus two days. The price data are collected in the five-month period before the reference month (e.g. for a journey in July the data are collected from February to June, on a daily basis), and the price index is calculated from the average of these prices.

With the traditional method, the office can collect one price a month for each of the destinations, however, with the web crawling method data are collected on a daily basis, therefore the possibilities of processing can be extended, too. It is possible to calculate the average of the daily minimum prices, the standard deviation of prices or also a daily price index can be calculated. The reference period can be changed, too: either the same period last year or any of the months from the previous year.

6. Results

In order to calculate the monthly price indices, the Central Statistical Office collects data throughout the five months preceding the reference month. During the pilot project we only managed to collect a two months' data, so the price index calculated from these data cannot be analysed with the office's methodology, but the characteristics of the change patterns can be analysed already (however, in the case of a continuous web crawling the results could already be compared). Figure 3 shows the price index calculated from the daily price index, the monthly average price index calculated from the daily price indices and the price index calculated from the data of the 21st day of each month. This way the daily and monthly price indices can be compared.

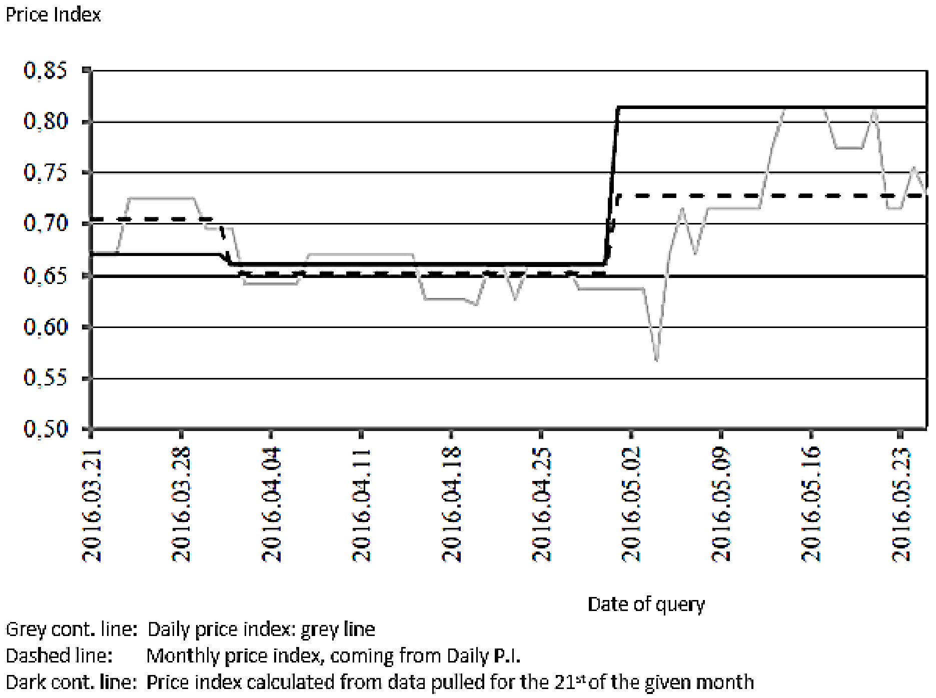

Figure 3 The price index of flights from Budapest to Paris, 21 March 2016 – 26 May 2016 (reference period: December 2016).
Comment: The figure shows the ticket reservations for flights to Paris in July 2016.
Source: Our own figure.

In this case the Big Data-based solution offers higher quality and accuracy, as well as faster results. The project was launched only recently. In the next step the Central Statistical Office would like to analyse the data over a longer period of time and it would like to compare the results with those data that were obtained manually.

7. Summary

The following questions are discussed in numerous fora dealing with statistics: 'Why is Big Data so interesting for official statistics?' and 'Why is not it enough to use the traditional data collection methods?' Based on the foregoing these questions can be answered with the following arguments:

Financing pressures. The 2007 economic crisis forced the private as well as the public sphere to find more cost effective solutions for financing their activities. Traditional data collections might be costly, therefore official statistics must find other, alternative data sources. Big Data is an alternative data source with the help of which it is possible to exploit the potential of administrative data, hopefully without any expenses spent on data collection. The use of Big Data-type sources is a reaction to social and market changes, but it can also actively transform the way people work with official statistics. These changes might affect mainly the following areas:

Improving the quality of traditional data collections. Those working with the traditional method must overcome several challenges. Through Big Data it is possible to obtain additional information which help us to create a better sample frame which can be managed more easily, to further develop our sampling methods, to create more accurate calibrating, estimating and imputation methods, to validate data from other sources (e.g. from traditional data collections or data adoptions), to reduce non-response rates or describe their pattern (with questionnaires certain social groups tend to be less accessible or are not accessible at all), and also to enrich out toolbox used for data analysis.

By using Big Data, we could *reduce the burden for respondents*. This aspect speaks for itself. Every respondent – private individuals and corporations alike – welcomes shorter questionnaires and forms, as this way they are able to spend less time on providing data for official statistics. If data are accessible from different sources as well, it will be unnecessary to ask data providers.

By actively using Big Data, it is possible to gain new information and new types of data, which were not accessible through traditional methods, and also new connections can be discovered, which could not have been discovered without these huge amounts of data. Also, this way innovative tools and methodologies can be created, which can later become milestones in official statistical procedures. With regard to short-term goals, we believe that by using Big Data it is possible to develop new welfare indicators; to link general economic, agricultural and environmental statistics from multiple aspects; to supplement data collections on household consumption and earnings by developing new measurement techniques; to measure consumer confidence and to understand consumer behaviour better.

The points listed here are only a fraction of how Big Data may be put to use. However, we still have many questions, both technical and professional. We are sure that statistical data collection is facing a paradigm shift that radically changes the status of official statistics.

Answering the following questions can help in setting our heading: 'What is our purpose – reproduction or creating a new calculation method?'; 'What to do if a statistician does not have the right IT tools and IT expertise to handle Big Data?'; 'In order to handle the new approach effectively, should statisticians become more knowledgeable in their respective fields, or rather develop their IT skills (such as programming language knowledge)?'; 'Can Big Data be built into the current data generation process?'; 'Will statistical results grow more reliable and accurate by the use of Big Data?'; 'Speed vs. Accuracy, what is the role of official statistics? Which of the two is more important?'

In our opinion, a balance between the latter two factors must be found, since the aim is to ensure not only quick results but also a methodological guarantee.

In our view, Big Data, like the online surveys that have become an integral part of the data capture method, will also find its place in official statistics without making conventional data collection procedures superfluous.

References

Barcaroli, G., Nurra, A., Scarnò, M. and Summa, D. (2014) *Use of Web Crawling and TextMining Techniques in the Istat Survey on 'Information and Communication Technology in Enterprises'.* Istat. Available at: http://www.q2014.at/fileadmin/user_upload/Iad_in_ICT_survey_PAPER.pdf Accessed: 01.20.2017.

Couper, M. P. (2013) Is the Sky Falling? New Technology, Changing Media and the Future of Surveys. *Survey Research Methods*, 7(3): 145–156. DOI: http://dx.doi.org/10.18148/ srm/2013.v7i3.5751

Daas, P. J. H. and Puts, M. J. H. (2014) *Social Media Sentiment and Consumer Confidence.* Statistics Paper Series No. 5. European Central Bank. Frankfurt.

Daas, P. J. H. and van Der Loo, M. (2013) *Big Data and Official Statistics.* United Nations Economic Commission for Europe, Eurostat, Organisation for Economic Cooperation and Development, United Nations Economic and Social Commission for Asia and the Pacific. Discussion paper. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/ Topic_4_Daas.pdf Accessed: 01.20.2017.

Devan, A. (2016) *The 7 V's of Big Data.* Impact Radius blog. Available at: https://www.impactradius.com/ blog/7-vs-big-data/ Accessed: 01.20.2017.

Fukuyama, F. (1995) *Trust.* The Free Press. New York.

Gartner, INC. (2017) *Big Data.* IT Glossary. Available at: http://www.gartner.com/it-glossary/big-data/

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M. and Khan, A. (2013) *What Does 'Big Data' Mean for Official Statistics?.* Paper prepared for the High-Level Group for the Modernization of Statistical Production and Services. 10 March. Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622 Accessed: 01.20.2017.

Groves, R. M. (2011) Three Eras of Survey Research. *Public Opinion Quarterly,* 75(5): 861-871. DOI: https://doi.org/10.1093/poq/nfr057

Infodiagram.com (2014) *Visualizing Big Data concepts: Strong and Loose Relation Diagram.* Available at: https://blog.infodiagram.com/2014/04/visualizing-big-data-concepts-strong.html Accessed: 01.20.2017.

ISTAT ESTP (European Statistical Training Program) (2016) *Introduction to Big Data and Its Tools.* Course. 29 February – 2 March. Rome.

Knacks, S. and Keefer, P. (1997) Does Social Capital have an Economic Payoff? *Quarterly Journal of Economics,* 112(4): 1251-1288. DOI: https://doi.org/10.1162/003355300555475

Pléh, Cs. and Unoka, Zs. (2016) *Hány barátod is van? (English: How Many Friends Did You Have Again?).* Budapest: Oriold és társai Kft.

Raiser, M., Haerpfer, C., Nowotny, T. and Wallace, C. (2001) *Social Capital in Transition: A First Look at the Evidence*. EBRD Working paper. No. 61. London: EBRD.

Scannapieco, M., Virgillito, A. and Zardetto, D. (2013) *Placing Big Data in Official Statistics: A Big Challenge? Eurostat, Collaboration in Research and Methodology for Official Statistics*. Available at: https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_214.pdf Accessed: 01.20.2017.

Tam, S. M. and Clarke, F. (2015) *Big Data, Statistical Inference and Official Statistics*. Research Paper No. 1351.0.55.054. Canberra: Australian Bureau of Statistics. Available at: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/015937BADB90186B CA257E0B000 E428A/$File/1351055054_mar%202015.pdf Accessed: 01.20.2017.

UNECE Big Data Quality Task Team (2014) *A Suggested Framework for the Quality of Big Data*. December. Available at: http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20 Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2 Accessed: 01.20.2017.

Vale, S. (2013) *Classification of Types of Big Data*. UNECE Statistics Wikis. Available at: http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of +Big+Data Accessed: 01.20.2017.

Vukovich, G. (2015) Adatforradalom és hivatalos statisztika (English: Data Revolution and Official Statistics). Statisztikai Szemle, 93(8–9): 745-758.

Appendix

F1 Possible application for Big Data in the Official Statistics

Type of data	Statistics	HCSO– Name of data collection/Published data
	Mobile communication	
Mobile data	Tourism, Population statistics, Migration statistics	Census – (Publishing mobility data every ten years) OSAP 2290 – Data of the permanent population by settlement (broken down by sex and age) OSAP 2228 – Data of citizens participating in international migration OSAP 1943 – Tourism and other expenses of foreigners in Hungary OSAP 1114 – Basic data to the physical and non-physical workers' working time balance Time Use Statistics
	Internet	
Internet searches	Labour force statistics, Migration statistics	OSAP 2238 – Monthly labour report OSAP 2009 – Report on the number of jobs and job vacancies OSAP 1114 Basic data to the physical and non-physical workers' working time balance
E-commerce webpages	Price statistics	OSAP 1006 – Consumer price survey (Consumer Price Index) OSAP 1712 – Report on housing and building site market turnover (housing prices) OSAP 1007 – Price survey on industrial products and services OSAP 1831 Prices of construction activities Construction price statistics OSAP 2193 – Data services on foreign trade turnover outside the European Union (Foreign trade prices)

Type of data	Statistics	HCSO– Name of data collection/Published data
		OSAP 2130 – Quarterly survey on the output prices of business services
Enterprises webpages	Information society statistics	OSAP 1840 – Qualitative and quantitative data on the inventory of information and communication technologies
Enterprises webpages	Business register (GSZR)	Clarification of Business Register (GSZR) pontosítása
Job advertises' websites	Job vacancy statistics	OSAP 2009 – Report on the number of jobs and job vacancies
Websites for real estate ads	Price statistics (real estate market)	OSAP 1712 – Report on housing and building site market turnover OSAP 2418 – Construction cost estimation on standard dwelling types
Social media	Consumer satisfaction, GDP, Information society statistics	GDP calculation
Sensor data sources		
Traffic sensors	Transport statistics, Tourism	OSAP 1390 – Transport infrastructure: local roads and bridges OSAP 2297 – Fleet of road vehicles OSAP 1183 – Data on road passenger transport OSAP 1189 – OSAP 1654 – Road and fixed-line passenger transport performances Census – (Publishing mobility data every ten years) OSAP 2290 – Data of the permanent population by settlement (broken down by sex and age) OSAP 2228 – Data of citizens participating in international migration OSAP 1943 – Tourism and other expenses of foreigners in Hungary

Type of data	Statistics	HCSO– Name of data collection/Published data
		OSAP 1114 – Basic data to the physical and non-physical workers' working time balance
'Smart' measuring devices	Energy statistics	Energy and environment OSAP 1321 – Energy balance, industry OSAP 2221 – Energy balance of energy sector, energy commodities OSAP 1329 – Monthly energy balance report OSAP 1335 – Survey on energy use
Satellite imageries	Agriculture, land use, enviromental protection statistcs	Land use: OSAP 1082 – Land area and sown area, 1 June OSAP 2218 – June survey of private holdings OSAP1709 – Data on Protected Areas by National Law and 'Natura 2000' sites
Aircraft movements	Transport and air pollution statistics	Air transport OSAP 1725 – Traffic data of airports OSAP 1966 – Report on the traffic of airports OSAP 2160 – Inland waterway, air and pipeline transport performances OSAP 1066 – Air quality data
Process generational transactions		
Supermarket scanner and sales data	Price statistics, Household consumption statistics	Consumer Price Index OSAP 1006 – Consumer price survey OSAP 2153 – Household Budget and Living Condition Survey, monthly diary keeping OSAP 2154 – Household Budget and Living Condition Survey, annual interview OSAP 1045 – Report on the sales turnover of retail trade and catering OSAP 1646 – Report on the sales of retail trade and catering by commodity groups OSAP 2130 – Quarterly survey on the output prices of business services

Type of data	Statistics	HCSO– Name of data collection/Published data
Financial transactions	Household consumption statistics	OSAP 2153 – Household Budget and Living Condition Survey, monthly diary keeping OSAP 2154 – Household Budget and Living Condition Survey, annual interview
Volunteered Geografic Information (VGI, websites (OpenStreetMap, Wikimapia, Geowiki)	Land use	OSAP 1082 – Land area and sown area, 1 June OSAP 2218 – June survey of private holdings