Statistical Commission
Forty-fifth session
4-7 March 2014
Item 3 (j) of the provisional agenda*
Items for discussion and decision: big data and modernization
of statistical systems

Big data and modernization of statistical systems

Report of the Secretary-General

Summary

  The present report, which was prepared in accordance with Economic and Social Council decision 2013/235, offers an assessment of the current use of big data for official statistics. The report provides an overview of recent activities of the official statistical community and presents the results of the global assessment on the use of big data for official statistics conducted by the Statistics Division in 2013.

  The report formulates a way forward through the creation of a working group at the global level, which builds on existing regional initiatives towards sharing with respect to methodological developments, best practices for strategic issues and training opportunities. The working group will also facilitate the international partnership for the use of big data in the transfer of technology to developing countries and in support of the post-2015 development agenda. The Statistical Commission is invited to express its views on the proposal made in the final section of the report.

_____
*E/CN.3/2014/1.

Please recycle

I. Introduction

1.   The theme of the Friday seminar on emerging issues in 2013 was "Big data for policy, development and official statistics".[1] Speakers from the private sector together with chief statisticians presented their views on the relevance of big data for policymaking and on the role that national statistical systems should assume with respect to exploiting the new sources of information. The afternoon session was chaired by the Chief Statistician of Australia. The conclusions drawn at this event were that big data constitute a source of information that cannot be ignored by official statisticians and that official statisticians must organize and take urgent action to exploit the possibilities and harness the challenges effectively.

2.   At the forty-fourth session of the Statistical Commission, as an immediate follow-up to this side event, the representative of Australia requested an assessment of the use of big data for official statistics, to be submitted to the Commission at its forty-fifth session. The present report responds to that request.

3.   As a result of the pervasiveness of the use of electronic devices and the all-around generation and availability of digital information, there has been a fundamental change in the nature of data, which are now generated continuously and in enormous quantities, and which we refer to as big data. They have highly distinct qualities that differentiate them from conventional-source data. The data from these innovative sources achieve a high level of distribution and are loosely structured, large in volume and often available in real time. Big data are data sources that can be described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision-making".

4.   In an era of declining responses to national household and business surveys, big data can provide policymakers with real-time evidence in areas such as prices, employment, economic output and development, and demographics.[2] Big data have the potential to produce more relevant and more timely statistics than traditional sources of official statistics, such as survey and administrative data sources. Most sources of big data reside in the private sector, and legislation designed to permit the use of big data for official statistical purposes has not yet been promulgated in most countries. Therefore, further broad-based dialogue is needed to consider not only legal, confidentiality and perception issues, but also the subject of the timely, trustworthy and credible use of big data. By incorporating big data sources into their production of official statistics, national, regional and international statistical organizations could be better positioned to obtain official statistics on the economy, society and the environment in terms of improved timeliness and cost-efficiency, and a lessened resource burden.

5.   Given the innovations in technology and the explosion in the quantity and diversity of real-time information, the Secretary-General of the United Nations launched, in 2009, the Global Pulse initiative to leverage innovations in digital data,

_____

[1] The seminar, a side event at the forty-fourth session of the Statistical Commission, was held on Friday, 22 February 2013. For details, see http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html.

[2] Organization for Economic Cooperation and Development (OECD), "Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by big data" (DSTI/ICCP(2012)9/FINAL), 2013.

rapid data collection and analysis to help decision makers gain a real-time understanding of how crises affect vulnerable populations. While stressing that big data are no panacea, the white paper produced by Global Pulse does state that "big data constitutes an historic opportunity to advance our common ability to support and protect human communities by understanding the information they increasingly produce in digital forms".[3]

6.   In the recommendations made to the Secretary-General by his High-level Panel of Eminent Persons on the Post-2015 Development Agenda in its report entitled "A new global partnership: eradicate poverty and transform economies through sustainable development", there is a call for a data revolution. Big data are recognized as constituting an important part of the data revolution, which could have the potential to contribute to improving some aspects of the quality of statistics, such as timeliness and completeness, without compromising the relevance, impartiality and methodological soundness of the statistics. The High-level Panel report states that

> Better data and statistics will help Governments track progress and make sure their decisions are evidence-based; they can also strengthen accountability. This is not just about Governments. International agencies, civil society organizations and the private sector should be involved. A true data revolution would draw on existing and new sources of data to fully integrate statistics into decision-making, promote open access to, and use of, data and ensure increased support for statistical systems (chap. 4, p. 24).

7.   In a broader context, the case has been made that big data have the potential to change many aspects of society. The TechAmerica Foundation Big Data stated in a recent report that "Big data has the potential to transform government and society itself".[4] Hidden within the immense volume, variety and velocity of data that are produced today are new information, facts, relationships, indicators and pointers, that either could not be practically discovered in the past, or simply did not exist before.[5] This new information, effectively captured, managed, and analysed, has the power to enhance profoundly the effectiveness of government".[6]

8.   Section II of the present report gives an overview of activities relevant to big data and modernization of statistical systems. Section III provides a summary of the sources of and challenges presented by big data. Section IV presents the results of the global assessment on the use of big data for official statistics, which was conducted by the Statistics Division from July to October 2013. Section V offers some case studies of the use of big data. The report concludes with a proposal on the way forward, which is in line with initiatives of partner organizations and which emphasizes the issues pertaining to the use of big data for official statistics by developing countries and for the post-2015 development agenda.

---

[3] Global Pulse, "Big data for development: challenges and opportunities", May 2012.

[4] Opening sentence of the executive summary of the TechAmerica Foundation Big Data Commission report entitled "Demystifying big data: a practical guide to transforming the business of government" (Washington, D.C., October 2012).

[5] Ibid., second sentence.

[6] Ibid., third sentence.

II. Overview of recent activities relevant to big data

9.   Governments in an increasing number of countries have acknowledged the importance of big data and established communities of practice and working groups to study its utilization and possible impact. The statistical community is gradually recognizing that there could be a paradigm shift in the making. The National Institute of Statistics of Italy acknowledges that "in order to exploit the (potentially enormous) treasure shining through the big data mountain, national statistical institutes will have to climb that mountain".[7] Statistics Netherlands admits that the official statistics community is only scratching the surface when it comes to exploring the opportunities offered by big data.[8]

10.   The Economic Commission for Europe (ECE), the Statistical Office of the European Commission (Eurostat), the Organization for Economic Cooperation and Development (OECD) and the Economic and Social Commission for Asia and the Pacific (ESCAP) organized a Meeting on the Management of Statistical Information Systems that was held simultaneously in Paris and Bangkok from 23 to 25 April 2013.[9] The key conclusion of the meeting was that statistical organizations have to combine efforts to face common issues in relation to using big data and work towards common solutions, such as producing statistics with a short life expectancy, taking a multidisciplinary approach to big data, and agreeing on a common classification for the different types of big data.

11.   The fifty-ninth International Statistical Institute (ISI) World Statistics Congress, held in China, Hong Kong Special Administrative Region from 25 to 30 August 2013, devoted much attention to the topic of big data, with one invited-paper session on the theme of big data; one special-topic session on the potential of Internet, big data and organic data for official statistics; and a satellite event on big data and statistical computing. At the special-topic session, John Dunne, of the Central Statistics Office of Ireland, presented a paper entitled "Big data coming soon ... to a national statistical institute near you",[10] which described some big data sources, such as mobile telephone roaming, electricity consumption or electronic payments, and presented ways in which these large volumes of timely data can be processed through cloud computing by third-party providers. Big data will also be a major topic in the scientific programme of the sixtieth World Statistical Congress to be held from 26 to 31 July 2015 in Rio de Janeiro, Brazil.[11] ISI congresses provide an opportunity for cooperation and exchange of knowledge among official, academic and business statisticians.

12.   Eurostat is investigating the potential use of big data for official statistics in areas such as price statistics (using Internet price data) and information and communications technologies (ICT) usage statistics. Moreover, the annual meeting of the Directors General of the European national statistical institutes in

_____

[7] Monica Scannapieco, Antonino Virgillito and Diego Zardetto, "Placing big data in official statistics: a big challenge?", 21 December 2012.

[8] Piet Daas and Mark van der Loo, "Big data (and official statistics)", working paper prepared for the Meeting on the Management of Statistical Information Systems (MSIS 2013), Paris and Bangkok, 23-25 April 2013.

[9] See http://www1.unece.org/stat/platform/display/msis/MSIS+2013.

[10] Available from http://www.statistics.gov.hk/wsc/STS018-P3-S.pdf.

[11] See http://www.isi2015.ibge.gov.br.

September 2013 in The Hague devoted one session to the topic of big data.[12] This resulted in the Scheveningen Memorandum on big data and official statistics (adopted by the European Statistical System Committee on 27 September 2013), which encourages members of the European Statistical System to develop a big data strategy, share experiences and collaborate at the level of the European Statistical System and beyond. An action plan and road map should be adopted by mid-2014, and integrated into the Eurostat work programme.

13.   Big data were also part of the discussions at the ECE Seminar on Statistical Data Collection held in Geneva from 25 to 27 September 2013.[13] At the Seminar, Peter Struijs and Piet Daas of Statistics Netherlands presented a working paper (No. 31) entitled "Big data, big impact?", which stated that it is obvious that big data will have a big impact on the statistical community and on official statistics. The specifics of this impact will become clear only gradually, but some features are already visible or foreseeable. National statistical systems will be subject to more competition from actors outside their community. They will need to adjust their means of generating statistics and find a new balance, making use of the new possibilities offered by big data. This may require a paradigm shift from a survey-oriented to a more secondary data-focussed orientation, where model-based approaches are the norm. With the entry into the big data era, statistical offices have to incorporate data scientists in their workforce as a norm rather than as an exception, and the organizational culture will change accordingly. Through an early recognition of the opportunities and challenges presented by big data, the official statistical community will be well placed to leapfrog its role in information provision from these new sources.

14.   At its second meeting held on 21 and 22 October 2013, the 2013/2014 Bureau of the Conference of European Statisticians, the governing body of ECE in statistics, reviewed in depth the topic of big data.[14] The main recommendations emanating from the review were that the key priority areas of big data should be specified and tackled as a collaborative activity by the international statistical community, and that a mechanism for sharing information on knowledge and experience of using big data should be established. The Bureau also approved a project proposal[15] on the role of big data, with the objectives of:

       (a)   Identifying the main possibilities offered by big data and providing guidance for statistical organizations, and developing a coordinated response to the main strategic and methodological issues that big data pose for the official statistics industry;

       (b)   Demonstrating the feasibility of efficient production of both novel products and "mainstream" official statistics using big data sources, and the possibility of replicating these approaches across different national contexts;

       (c)   Facilitating the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using big data sources.

_____

[12] See http://www.cbs-events.nl/dgins2013/.

[13] See http://www.unece.org/stats/documents/2013.09.coll.html.

[14] See ECE/CES/BUR/2013/OCT/2.

[15] This project is part of the initiative on modernization of statistical production under the High-level Group for the Modernization of Statistical Production and Services coordinated by ECE.

15. Big data and the modernization of statistical systems were also in the agenda of some recent events in Asia. The Statistics Division and the National Bureau of Statistics of China, with the support of ESCAP, organized an international seminar on modernizing official statistics from 24 to 26 October 2013 in Tianjin, China. This seminar focused on modernization of statistical organizations and included several presentations on big data. The Eleventh Management Seminar for the Heads of National Statistical Offices in Asia and the Pacific,[16] which was organized by the Statistical Institute for Asia and the Pacific (SIAP) of ESCAP on 21 and 22 November 2013 in Chiba, Japan, focused on the data revolution initiative and the potential role of big data in the post-2015 development agenda.

III. Big data: sources and challenges

16. An informal task team, established by the High-level Group for the Modernization of Statistical Production and Services and coordinated by ECE, summarized the sources of big data and the challenges related to the use of big data in a report entitled "What does big data mean for official statistics" (10 March 2013). The report was presented at the sixty-first plenary session of the Conference of European Statisticians held from 10 to 12 June 2013. The findings of this report are discussed below. The breakdowns proposed by the ECE task team, which were provided in the report, have been used in the questionnaire for the global survey conducted by the Statistics Division in the third quarter of 2013 (see sect. IV).

17. Nowadays, data are automatically and continuously generated in digital form in many different ways. These different data sources are potentially of interest for use in official statistics to measure certain social, environmental, financial or economic trends with greater accuracy and timeliness. In general, big data sources can be classified as follows:

- Sources arising from the administration of a programme, be it governmental or not, e.g., electronic medical records, hospital visits, insurance records, bank records and food banks

- Commercial or transactional sources arising from the transaction between two entities, e.g., credit card transactions and online transactions (including from mobile devices)

- Sensor network sources, e.g., satellite imaging, road sensors and climate sensors

- Tracking device sources, e.g., tracking data from mobile telephones and the Global Positioning System (GPS)

- Behavioural data sources, e.g., online searches (about a product, a service or any other type of information) and online page views

- Opinion data sources, e.g., comments on social media

18. Administrative data are one of the main data sources for the production of official statistics by national statistical systems. Traditionally, these data are received in a highly structured manner from public administrations, and then processed, stored, managed and used by the statistical institutes. Administrative data are currently not a big data source, but could become "big" when the velocity and

_____

[16] See http://www.unsiap.or.jp/training/6_Leading_MQS/6_ms11.php.

volume increase — for instance, when statistical institutes start using administrative data more extensively by collecting them in real time, or on a daily or weekly basis, instead of once a year or once a month, which is usually the case.

19.   The use of big data in official statistics presents many challenges, which fall primarily into the following categories:

- Legislative, i.e., with respect to access to and use of data

- Privacy, i.e., managing public trust and acceptance of reuse of data and their link to other sources

- Financial, i.e., potential costs of sourcing data versus benefits

- Management, e.g., policies and directives about the management and protection of the data

- Methodological, i.e., data quality and suitability of statistical methods

- Technological, i.e., issues related to information technologies.

IV.   Results from the global survey on the use of big data

20.   Taking the descriptions given in the report of the ECE task team, the Statistics Division developed a questionnaire on the use of big data for official statistics. The questionnaire had three main parts: sources, challenges and areas of use. The questions on sources and challenges encompassed all the points mentioned above, with some elaboration on management aspects, which were indicated as "management issues, such as adhering to new policies and regulations, and developing human resources with the necessary set of skills and expertise". This global assessment was undertaken with the objective of providing information on national priorities, developments and experiences regarding the current or planned use of big data for official statistics.

21.   The questionnaire was sent to the statistical administrations of over 200 economies in the period from July to September 2013: it was sent out in English in July 2013, Spanish in August and French in September. It could be completed either online or in PDF format. By 2 November 2013, 107 responses had been received. The full set of results are made available to the Statistical Commission as a background paper. A summary of the results is given below.

A.   Sources of big data

22.   The exact phrasing of the question regarding the sources of big data was: "Please indicate which of the following big data sources will likely be used in the next 12 months by your office or other agencies that are part of the national statistical system of your country." If the answer was Yes, an explanation on the specifics of the data source was requested. While over 50 per cent of the economies replied positively about the use of administrative sources as big data, the figure was about 25 per cent for each of the other (five) data sources. Several countries raised the issue whether extensive administrative records are big data or not. More generally, administrative data are a cornerstone of the production of many statistical

products, but whether they should be considered within the same context as big data remains a point of discussion.

23.   Overall, many countries replied positively with respect to the use of one or more of the big data sources. However, further analysis revealed that some of these positive answers were not compatible with the definition of big data, as used in the reference report of the ECE task team. To lend more power to the conclusions of this survey on big data, the responses were scrutinized and the positive replies narrowed down to a selected group entailing what were deemed "relevant practices". The positive answers that did not pass the bar were all those without an explanation, with a very limited explanation or with an irrelevant explanation. In consequence, instead of the initial figures of 25-50 per cent for positive responses, the adjusted figures were almost 20 per cent for administrative data sources, about 10 per cent for transactional sources, sensor networks and tracking devices, and less than 5 per cent for the behavioural and opinion sources. Detailed results will be made available on the website of the Statistics Division, with separate columns for all countries with positive answers and those countries with "relevant practices".

24.   Among the most-cited sources of administrative data are data from tax and value added tax (VAT) offices. Other sources are person registers and business registers. Transactional sources cited were supermarket scanner data, telephone communication, credit card transactions and toll collection data. In terms of sensor or satellite sources, several countries stated that they used satellite imaging for survey design. Satellite imaging was also stated as a source for studying crop production statistics. Road sensors were used for both transportation statistics (the traffic index in the Netherlands) and environmental statistics (measuring air quality in the United Kingdom of Great Britain and Northern Ireland).

25.   Tracking or GPS sources were used by a few countries. For example, tracking of mobile telephones was used to improve tourism statistics in Estonia, which is given as a case study in the next section. In Canada, GPS data from trucks, rail, ship and air transport are used to supplement transport statistics. While behavioural data were not used by many countries, some countries had plans to make use of this source. For example, in Italy, Internet queries will possibly be used (a) as auxiliary information to reduce costs associated with surveys, (b) for forecasting purposes and (c) to increase geographical detail of estimates (especially for labour-market indicators).

26.   Similarly, very few countries make use of opinion or social media sources for big data. The Netherlands indicated some research efforts based on publicly available social media messages (see the case study in sect. IV). In Singapore, research is being conducted on how this source of big data can be used for early event detection in the case of food safety issues.

B.   Challenges in the use of big data

27.   The exact question for this part was "Are [issues] a major challenge for the national statistical system in your country?", with No (not a challenge), No opinion (has not been discussed) or Yes (is a challenge) as possible answers. As regards all six challenges, yes, (this is a challenge) was the reply of most countries, followed by No opinion (indicating that those issues had not yet been thoroughly discussed) and by a small number of Nos (this is not a challenge). Overall, methodological,

information technology and management challenges were indicated most often, followed closely by legislative and privacy challenges.

28.   Respondents agreed that one of the major challenges will be resolving methodological issues, and suggested that more research would be needed in this area. Data quality, representativeness, volatility and dimensionality all present methodological difficulties for the exploitation of big data sources. Information technology issues were also indicated as a major challenge in many countries because the existing processing and analysis systems do not scale for big data. The nature of big data (especially the volume but also the velocity and variety) leads to major challenges for the development of necessary information technology tools and information technology architecture, in particular in the developing world. Management issues such as adhering to new policies and regulations, and developing human resources with the necessary set of skills and expertise were viewed as being a major challenge in most countries.

29.   The systematic use of non-governmental data sources poses a legislative challenge in the context of the big data opportunity for Governments. While some countries already have legislation in place to ensure access to both governmental and non-governmental sources, the lack of legislation guaranteeing access to non-governmental sources was cited as a challenge in other countries. Privacy issues are also a major concern in many countries as regards big data use. The protection of personal data in this area is very important; hence, it is necessary to strike an optimum balance between the protection of personal data and their statistical use.

30.   Financial issues were also deemed challenging by a number of countries, in particular the costs of sourcing data, which were perceived as a major challenge in the use of big data. Also, the financial resources needed to acquire the latest hardware and software for handling big data may be inadequate or lacking. In addition, it seems that private companies are increasingly recognizing the value of the data they hold and are therefore increasing their prices.

31.   Other major challenges related to the use of big data by national statistical systems were suggested. In particular, the issue of how to integrate big data into statistical systems was raised by many respondents. How can this new source of data be fitted into an existing statistical programme? Most agree that big data will not replace official statistics, but, rather, will act as a complement thereto, for example, by providing very rapidly available statistics and beta indicators for key statistics. As big data constitute a relatively new concept in the field of official statistics, there is a need to cultivate enthusiasm among official statisticians.

C.   Areas of use and potential areas of use

32.   In the third and last part of the global survey, the respondents were asked about the areas of use (or research into use) of big data in the next 12 months. The areas of enquiry included "Demographic and social statistics", "Vital and civil registration statistics", "Economic and financial statistics", "Price statistics", "Transportation statistics", "Environmental statistics" and "Other domains of official statistics". The possible answers were No or Yes, where a Yes needed to be explained. The areas that were most often indicated as area of use were demographic and social statistics and economic and financial statistics. However, as in the case of the use of sources of big data, the positive answers in this part of the survey needed

to be re-evaluated by identifying "relevant practices". After this re-evaluation, it turned out that the relevant practices of use of big data were in the areas of price statistics and economic and financial statistics. In particular, a number of countries cited the usage of scanner data and/or Web-scraping techniques for the creation of frequently updated price indices that complemented the standard consumer price index. Overall, the proportion of relevant practices for areas of use was just over 10 per cent for price statistics, for economic and financial statistics and for demographic and social statistics, and about 5 per cent for each of the other areas of use.

33.   In the area of demographic and social statistics, there were a few mentions of how mobile telephone data could be used for travel behaviour and tourism, social media sentiment for consumer confidence and online vacancy data for employment, and of the potential use of Google trends. Other examples included cooperation with mobile telephone operators to calculate geo-referenced commuting statistics. In the area of economic and financial statistics, research into the use of satellite imaging and ground sensor data for estimating crop yields and the use of mobile device location tracking data for service population mobility statistics were mentioned. A number of countries explained how they used big data in the area of transportation statistics. Detailed data from highway tolls and road sensors are being used to prepare traffic statistics.

V.   Big data and the modernization of statistical systems

34.   In order to confront the challenges mentioned in the replies to the global survey, statistical systems will need to be modernized. It was suggested that more research is needed to overcome the methodological difficulties impeding the exploitation of big data sources. Owing to the nature of big data, issues of representativeness and population coverage may be problematic. Further, the variety and potential temporality of the data (e.g., how long will people keep using Facebook?) pose problems from the official statistics' point of view, given that comparability, continuity and coherence of statistics are among the main quality dimensions. Changes in methodology may include the more frequent use of modelling and may require closer cooperation with academia. More research and experimental studies need to be conducted to explore the potential applications of big data in official statistics, and research may need to become a part of the statistical production process.

35.   The existing information technology infrastructure and capabilities need to be enhanced so as to be able to support the processing of high-volume and high-velocity data sets from various sources. Also, standardization of methods and tools among national statistical offices should be considered. There is a need to acquire the latest technology (software, hardware and people ware) to keep abreast of the technological needs in the collection, processing and use of big data. Offices may have to consider cloud computing instead of trying to move large amounts of data to their own servers. Alternatively, the custodian of the big data source may be requested to perform the initial computations and provide more aggregated and better-structured data to the statistical office. Sufficient resources will be needed to build the required information technology infrastructure to source for relevant data frequently, link and analyse them and present them in a suitable format for users.

36.   Both the development and the retention of staff with the necessary skills in advanced data analysis and computational informatics will be a major challenge. The new big data sources require changes in the organization of data collection and data processing. Most national statistical offices do not have staff with the appropriate skills, and would therefore require capacity-building and intensive training in the area of human resources in order to be able to exploit these data sources. The current expertise, based on the classic statistical know-how, is not sufficient to ensure adequate treatment of this new typology of data, and it is therefore necessary to provide for the acquisition and training of adequate new professional types (e.g., data scientists).

37.   Additional legal steps are needed to enable the production of official statistics using big data. The current legislative framework for statistics in many countries does not cover access to and use of big data, both within government and from private sector. Thus, it will be particularly difficult to gain access to the big data collected and kept by other parties. Furthermore, a privacy framework is needed that sets the ground rules for how big data sets can be combined, protected, shared, exposed, analysed and retained. This would address the significant issue of public trust in the appropriate use by government of the personal data of individuals. It is important to maintain public trust: individuals must be sure that their personal information will be well protected — not disclosed or misused. For example, in the area of mobile telephone location data, which are one of the potential data sources of big data, even if identification is suppressed, people will still be highly concerned about the transfer of such information from the mobile telecommunication providers to other parties for any purpose and about the sharing of that information. Similarly, mobile device providers need guarantees that privacy rights will not be violated when they turn over their data to the Government.

38.   The public's reaction to the use of big data for official statistics is an essential determinant for their future use. Public trust and confidence in the use of personal data must be earned (not damaged) by the government, and even more so by the private sector. The government sector will need to invest considerable efforts in establishing an authorizing framework that sets out the rules for access to and use of automatically generated digital data with a personal content and that provides transparency, in respect of the roles of the data custodians and the relevant government offices. Only within such an environment can privacy and confidentiality of personal information be assured and can public trust be established in the use of big data in official statistics.

VI.   Case studies

A.   Australia: use of satellite images for agriculture statistics

39.   Satellite images are being used in the area of land-cover and in a research programme where agricultural land-use features are employed to estimate particular types of crops. The spatially generated features are utilized as input into statistical models created to perform calculations in the compilation of agriculture statistics. Potentially, they can also be applied to the compilation of fisheries and forestry as well as environment statistics. The raw data are accessed and processed on a cloud server of the supplier's system outside the statistical organization's computing

environment. The image analysis algorithms for extracting spatial land-use features are developed in joint research with university collaborators. This involves development of algorithms, spatial and time-series models, geo-coding systems and statistical estimation processes.

B. Brazil: high-frequency monitoring data for water accounts

40.   By an agreement signed in 2012 by the National Institute of Geography and Statistics (statistics authority), the National Water Agency (the water resources federal authority) and the Secretariat of Water Resources and Urban Environment of the Ministry of Environment (for water resources policy coordination), a committee was created that holds the mandate to develop water accounts in Brazil. The National Water Agency monitors the quantity and quality of water resources through the national hydro-meteorological network, composed of almost 22,000 monitoring stations, which collect high-frequency data regarding water flows, rainfall, water quality, evaporation and sediments. In other words, the National Water Agency collects a high volume of data on water every day, processes these data and makes them freely accessible on the institutional website of the Agency[17] or from the National Water Resources Information System.[18] The Information System constitutes a comprehensive system of collection, treatment, storage and provision of data and information related to water resources management. The National Institute combines the data provided through the Information System with data obtained from its regular collection of economic and environmental statistics to compile the water accounts of Brazil.

C. Bhutan: mobile communication technology for the consumer price index

41.   In Bhutan, the collection of consumer price data is supported and facilitated by modern communications technology. The consumer price index is produced on a monthly basis, based on data that are collected from the various districts by local statistical officers using an android mobile telephone application. The officers collect data for all items in the consumer price index basket from the relevant shops in their respective districts and submit the data automatically to the server in the head office of the National Statistics Bureau in Thimphu. The head office combines the data from the districts and produces the national-level consumer price index and inflation figures at a monthly level.

42.   Even though the current application in Bhutan does not involve huge amounts of data, the use of mobile devices and wireless transmission could be further developed and applied to price collections with a much greater quantity of data. The approach used in Bhutan, starting with the use of small-scale electronic data collection, could also be seen as a stepping stone for other developing countries. Automating of input of digit data and wireless transfer to a centralized server are important first steps towards tapping into the potential of big data sources.

_____

[17] See http://www2.ana.gov.br.

[18] Http://www.snirh.gov.br/telemetria/consultardadoshidrovisualizacaogeo/inicial.action.

D. Colombia: satellite images for agriculture and transport statistics

43. Colombia has been using satellite images as a source of big data in a number of areas. One example in this regard is a pilot test using satellite images to complement the work of the survey staff of the national agricultural census. The diagnostic phase of the test aims at determining the types of variables that can be captured through satellite imaging and that can support information gathered by the census operation. Another example entails use by the United Nations Office on Drugs and Crime and the Ministry of the Interior and Justice of satellite images to measure and monitor coca crops in Colombia through the integrated system of illicit crop monitoring.

44. The National Roads Institute of Colombia uses GPS data to improve traffic circulation and to serve as input for transport statistics. With this method, cars do not have to stop at toll booths; instead, an electronic tracking device installed in the vehicle is read when it enters the toll. Such dynamic tolls ensure a high level of comfort, safety and service. The tracking device also contains all the information concerning the vehicle, which complements that of the National Single Transit Register. Thus far, this new method has been tested in 10 toll stations in Colombia and has already enhanced control of traffic flows and has led to the strengthening of transport statistics.

E. Estonia: mobile positioning data for international travel statistics

45. The methodology for the compilation of international travel statistics, which uses passive mobile positioning, was developed jointly by the Central Bank of Estonia, the University of Tartu and Positium LBS. Positium is a company that collects and processes anonymous data from mobile operators with statistical models and extends these data to the general data set. This methodology gives a reliable overview as regards the physical persons crossing the Estonian border to travel abroad (outbound travel) or to enter Estonia (inbound travel). The mobile positioning data are anonymous. They are processed, analysed and presented by using a methodology that does not allow for any kind of identification of the phone number or phone owner.

46. The Central Bank uses international travel statistics to calculate the imports and exports of travel services for the current account of the balance of payments. In compliance with the balance-of-payments methodology, mobile positioning determines the residence of a passenger by the permanent-residence criterion, regardless of the resident's citizenship. The presumable residence of phone owners is determined by the use of phone roaming and local mobile networks. The number, length and nature of the visits of residents and non-residents of Estonia are determined by the location-based use patterns of mobile phones, depending on the country and region of use.

47. Statistics on outbound travel reflect the same-day and overnight visits of residents to foreign countries. Outbound travel by residents of Estonia may include visits to a number of countries. The country of destination is the farthest country from Estonia or the country (countries) where at least two consecutive days have been spent. The other countries, which are visited without overnight stay and which do not satisfy the criteria for destination country, are classified as transit countries.

The number and length of visits by residents and non-residents are calculated with statistical models and extended to the general data set, while also taking into account other official data sources (number of passengers in Tallinn port and Tallinn airport, crossings of frontiers on Estonia's eastern border, official accommodation statistics).

48. The anonymous mobile positioning data are used in compliance with Estonian and European Commission legislation on data protection and personal data protection (the Estonian personal data protection act; directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data; and directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector). The mobile positioning methodology has been developed and used in various fields in Estonia since 2006 in cooperation with the data protection inspectorate with regard to personal data protection issues.

F. Netherlands: social media as a potential data source for official statistics

49. An increasing number of people are active on social media. Here, people voluntarily share information, discuss topics of interest and contact family and friends. In the Netherlands, social media are used by a large proportion of the population; about 70 per cent have reported being active on one or more social media platforms. Since the responses to the questionnaires of Statistics Netherlands continue to decline, the potential usability of the messages created and shared voluntarily in social media as a data source for official statistics was investigated. For this purpose, publically available social media messages created on various social media platforms, such as Twitter and Facebook, as well as the public messages posted on news sites, web forums and blogs, were studied. The messages were obtained from a commercial company that routinely harvested all publically available messages written in Dutch on the Dutch-language part of the web.

50. Both the content and the sentiment of the messages were studied. Studies of the content of messages in Dutch on Twitter, the social media platform on which most publically available Dutch-language messages are created, revealed that nearly 50 per cent of those messages were composed of "pointless babble". The remainder predominantly discussed spare-time activities (10 per cent), work (7 per cent), media (television and radio) (5 per cent) and politics (3 per cent). Use of these more serious messages was hampered by the less serious "babble" messages. Determination of the sentiment in all messages created on all available platforms revealed a highly interesting potential use of these data for statistics. The sentiment in these messages was found to be highly correlated with Netherlands consumer confidence, in particular with sentiments regarding the economic situation. The latter relation was stable on a monthly and on a weekly basis. Daily figures, however, displayed highly volatile behaviour suggesting that it is possible to produce monthly and weekly sentiment indicators comparable with consumer confidence. The latter indicators can be produced on the first working day following the week studied, demonstrating the ability to deliver results quickly.

VII. Conclusions and the way forward

51. The statistical community has recognized the potential of big data for official statistics. Big data can help to better fulfil the mandate of providing timely and coherent statistics on the economy, society and the environment for decision-making, research and public debate. Moreover, with the progress of technology in geo-coding the units of the economy, society and environment and the increasing availability of information at the lowest level of geographical resolution, early progress is expected on the use of big data in developed and developing countries. Therefore, methodological developments, best practices for strategic issues and training opportunities need to be shared within the global statistical community across regional initiatives, if big data are to be taken advantage of, including in addressing issues pertaining to legislation, privacy, finance, management, methodology and technology. While addressing these big data issues, Governments and the corporate sector will be called upon to create an environment where the privacy and confidentiality of personal information for official statistical purposes can be assured, and public trust can be established in the use of big data for official statistics.

52. The global assessment of the use of big data for official statistics yielded some good examples, particularly from transactional, tracking, sensor and satellite data sources. Price statistics were compiled from scanner information, consumer expenditure from electronic cards transactions, travel and transport statistics from GPS tracking devices, well-being statistics from social media data and agriculture statistics from satellite imaging. The global survey also revealed many perceived challenges, especially with respect to methodological difficulties in the compilation of big data, information technology demands and privacy concerns.

53. It is evident that facing the challenges of big data implies modernizing the statistical system. For instance, the research arm of the statistical system will need to be strengthened, the information technology capabilities will need to be scaled up, human resources will have to be restructured to add data scientists, and partnerships will need to be forged with the private sector in the sharing of automatically generated digital information within an environment where privacy concerns are addressed and confidentiality rules obeyed.

54. Big data and the modernization of statistical systems introduce challenges and opportunities which are very similar for most countries. Experiences, practices and solutions could therefore be shared equally among national statistical institutes. Some initiatives in that direction have been started at the regional level. Within the context of the Conference of European Statisticians, an ECE project proposal was approved to develop the role of big data in the modernization of statistical production. Similar initiatives are under way through ESCAP in the Asia-Pacific region. Moreover, the launching of projects reflecting similar aspirations is being considered in other regions.

55. A global working group on big data is envisioned to complement the regional aspirations and initiatives within a global programme of work. This programme will allow for the sharing of practices among the regional commissions. Where the regional commissions could work on specific projects with a smaller number of countries at the regional level, the global working group would be sharing and managing issues so as to reach agreement on matters of methodology, best practices

for strategic issues, training and use of common platforms. As part of its work programme, the global working group should make the case that national statistical systems have a comparative advantage over other parties in their role of validating big data for policy use or combining big data sources with existing official statistics.

56.   The global working group could further address specific issues related to the role of big data for the post-2015 development agenda and should facilitate partnership with other United Nations and international organizations, civil society organizations and other members of the global partnership for the emerging data revolution initiative. One of the complaints regarding the Millennium Development Goals indicators was that their timeliness was poor. Big data are most often generated in real time and could potentially provide very timely indicators.

57.   To launch the global working group, an international conference will be organized in the second or third quarter of 2014 for the purpose of defining a work programme on common methodology, training, platforms and other issues related to the use of big data for official statistics.

VIII.   Points for discussion

58.   **The Statistical Commission is invited to express its views on:**

        (a)   **The need to further investigate the sources, challenges and areas of use of big data for official statistics at the global level, especially with respect to the circumstances of developing countries and the link to the post-2015 development agenda and the data revolution initiative;**

        (b)   **The creation of a global working group on the use of big data for official statistics whose activities would complement the work carried out by the regional commissions and manage the globally relevant issues.**