BIG DATA: POTENTIAL, CHALLENGES, AND IMPLICATIONS IN OFFICIAL STATISTICS

Ogerta Elezaj[1], **Dhimitri Tole**[2]

Abstract: The data explosion called "data deluge", is already starting to transform public institutions redefining their way of producing statistics in response to Big Data. The use of Big Data is considered as an innovation in the production of official statistics facing a range of opportunities, challenges and risks. This "data deluge" requires a number of challenges to be addressed in various domains: technological, legal, methodological, and statistical. Even though big data is changing the paradigm of producing statistics in many public organizations, an open debate still exists involving both IT specialists and statisticians of national statistical institutions.

In this paper we will provide an overview regarding the concepts of Big Data as a data source in production of official statistics by government institutions with the main focus on providing a synoptic overview of opportunities, challenges and risks. Following this, in the next section we will analyse a case study related to the potential use of mobile positing data, and how this data could be used to produce national statistical indicators in the country. This study serves as an example to identify some critical issues on challenges and risks, draw conclusions and give recommendations on the proper ways to shift to Big Data paradigm usage in the government sector in Albania.

JEL Classification Numbers: C55; **DOI:** http://dx.doi.org/10.12955/cbup.v6.1139

Keywords: big data, official statistics, tourism statistics, mobile position data

Introduction

Presently huge amounts of data are produced, with most of it being unstructured. Based on the estimates about the quantity of data produced, organizations are facing technical and organizational challenges to extract useful insight from them. Streams of data flows come from different sources such as web data, online social networks, smart grid and sensor data and time and location data.

The term Big Data refers to structured or unstructured data which cannot be processed and stored using tradition data warehouse solutions while applying classic rules of data and information processing. Big data is considered to be one of the key drivers in information management for many organizations requiring a different paradigm for data processing in a real-time economy where rapid access and process of complex data matters more than even. The two main factors of importance of Big Data are the usage of open source technologies for storing and processing data and the large volume of data spanning many tens to hundreds of tera and petabytes.

Different definitions regarding Big Data can be found in literature but the most often used in industry is provided by Gartner (Beyer & Laney, 2012) "Big data is 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“. It is clear that Big Data does not only refer to the size of the data, but data scientists describe it as having different dimensions known as 5V: scalable data (Volume), different type of data (Variety), generation rate (Velocity), biases and noise in data (Veracity), added-value (Value).

Different examples of the usage of Big Data appear in many areas such as health care, manufacturing, media and entertainment, the Internet of Things, and governments. The main goal from the collection of Big Data by these organizations is to get insights from the massive data sets available. Big data brings an innovative direction to data analytics, shifting it from descriptive statistics to predictive and prescriptive analytics.

As far as National Statistical Institutes are concerned, the question is how this innovative approach can be implemented in a business statistical production so that it adds value to the process itself. Big data should be considered an important data source for official statistics which allow organizations to reveal figures at low costs (Glasson et al., 2013). How this can be achieved in practice is a topic of interests for many National Statistical Institutes. Overall, there is an urgent need for greater conceptual clarity, technical specificity, political breadth and strategic foresight to be taken into account by producers of official statistics.

[1] Faculty of Economy, University of Tirana, Albania, ogertaelezaj@feut.edu.al
[2] Faculty of Economy, University of Tirana, Albania, dhimiter.tole@unitir.edu.al

In this paper we provide an overview of the concepts on Big Data under the official statistics perspective related to ongoing change management about the collection and production of statistics.. In the second section we identify the challenges and opportunities of introducing Big Data in the production of statistics. In the third section, we will provide some real examples of Big Data in the production of statistics in different countries. In section four, we will describe a case study related to the use of data of scrape prices taken from the internet, and how this data could be used within price statistics.

Finally, to conclude some recommendations will be provided regarding the steps that Albanian providers of official statistics should take in order to add value to statistical information products through rethinking of traditional statistical methods.

The Potential of Big Data

Using Big Data can benefit social-economic statistics and ultimate policymaking because of the possibility to produce new indicators, support a timelier forecasting of existing indicators and use innovative data sources in the production of official statistics. Bases on these new data sources, new indicators can be produced seeking patterns and correlations to give an indication whether a phenomenon is happening or not. For example, Google web search can be used to predict stock market liquidity (Arouri et al., 2014) or the IMF newscasts GDP using Google trends data. More specifically, to monitor and predict financial development the IMF launched a project called SWIFT where the financial transaction data coming from banks all over the world was processed to assess network connections and cross-border transactions. More than 25 million Swift messages are sent daily, and different indicators can be built to capture global and regional financial flows and this data can then be used to for early GDP estimation or to support banks and the government in anti-money-laundering processes. In other countries Big Data are considered an alternative data source to produce statistics with a lower cost than when conducting survey or administrative procedures (UNECE, 2013). For example, Estonia is using mobile data for balance of payment statistics to replace border surveys. Mobile positioning data can be used to estimate the inbound travels by non-residents into the network of resident mobile operators and outbound travels through roaming activities. The main benefit of this new approach is a reduction in costs however, it also results in an improvement in data quality and accuracy as the statistics produced are real-time statistics. This can be an opportunity to be taken into account for countries where tourism is an important sector of the economy.

Challenges of Big Data

The opportunities and challenges of using Big Data in producing official statistics are matter for open debate because of the complexity of data management and the nature of the products of official statistics. An overview is given bellow outlining the main challenges.

- Big Data management

Challenges for Big Data management include data processing challenges, managing data across organizations and limitations in the technologies and expertise required to manage large volumes of data. Traditional statistical analyses are based on target populations, structured data and sampling theory. Since Big Data are mostly unstructured and in different formats such as video, image, and textual data, techniques of information retrieval such as data mining, data reduction and artificial intelligence should be considered for use by official statisticians to cope with these challenges (Quick & Choo, 2014).

Retrieving, storing, processing and transferring the huge volume of data is a challenge. Technological innovations like high performance computing, storage facilities and high bandwidth data channels may partially solve these issues. Because the confidentially and security of the collected data is a top priority in public organizations, using cheap cloud-computing solutions is not the best option to be utilized by them. Working groups in big data projects should be composed of members with multidisciplinary skills from different professional backgrounds. Statistical agencies, banks and public agencies should train their staff.

- Data Access, ethical and privacy concerns

In the majority of cases, the owners of Big Data are outside of the control of national or international institutions. The data are held by private sector organizations and accessing and using them requires legal agreements to be arranged with data owners where confidentiality and data privacy should be well specified. As Big Data contain personal data, institutions should invest in IT infrastructure and to adopt security techniques such as cryptography, data anonymization and statistical disclosure control (Moreno et al., 2016). Furthermore, there is a conflict between demands for privacy and principles of official statistics to ensure openness and transparency in disseminating these statistics. (Henninger, 2013).

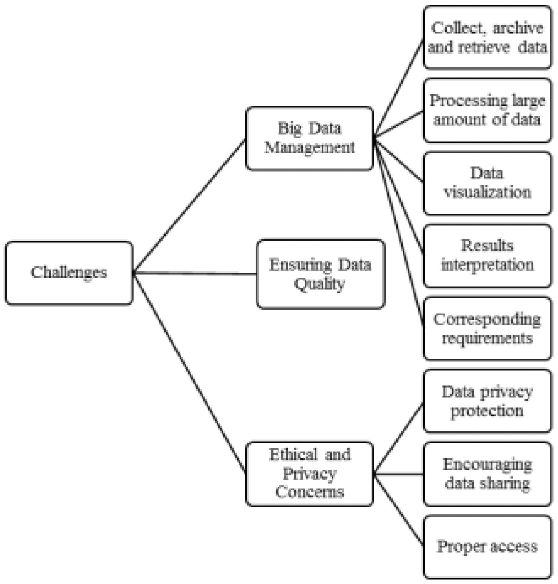
- Data quality

The quality of indicators produced using Big Data should be assessed to insure that the new indicators produced meet the minimum data quality standards for real fiscal, monetary, financial or other statistics. Most of the data coming from Big Data is unstructured and proper data manipulation techniques should be applied to transform them into time series data. The processing means cleaning, outlier detection and treatment, imputation of missing variables which should be done in line with tradition statistical methods to ensure data quality. Still, there is a lot to be done regarding methodology of Big Data processing to adhere to a data quality standard.

Even though the existing methodology issues, Big Data can still uncover insights in data by alerting that a social-economic phenomena is happening. For example the government can draw conclusions about different social-economic trends or opinions by conducting sentiment analyses on social network data.

These challenges are summarized in Figure 1.

Figure 1: Challenges of Big Data applications



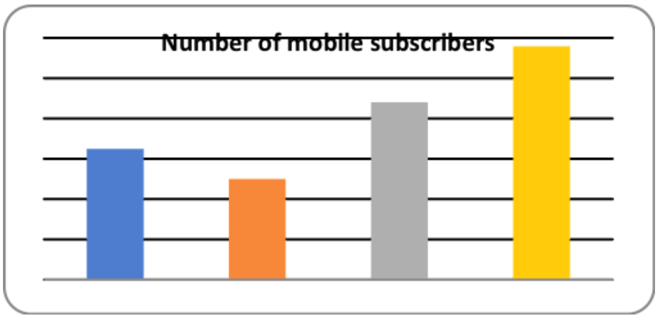
Source: Authors

Potential application of Big Data in Albania

Albania, like Europe and the world is transforming its economy and society through the use of the internet and other digital technologies. The government is promoting digital initiatives to create new opportunities for citizens and businesses. Recently, a government data center was established following the international standards. The data center has implemented a modern IT infrastructure and is accessible to public organizations. Also, a unique public service portal is implemented and the majority of e-services to citizens and businesses are offered via this portal. In the “Digital Agenda of Albania 2015-2020” Cross-Cutting Strategy, 2018) no vision, policies and strategic priorities exist for Big Data. Also, based on some interviews with employees on “IT Directories” in different ministries, no initiatives using Big Data are taken. However, in the private sector there are some limited efforts in using Big Data especially in the banking sector. The Albanian government should launch projects to investigate and explore practical aspects of Big Data by implementing pilots. The main goal of these projects should be to create a Big Data action plan and start identifying potential Big Data sources, skills and competences needed, methodological changes, the benefits of using them, and the definition and implementation of the required IT infrastructures.

A potential use of Big Data using Mobile Positioning data to produce tourism statistics will be explained. Based on the official data reported by the government (AKEP, 2018), Albania has a total number of 5,558,492 mobile subscribers for the year of 2017 and as is shown in Figure 2. There is a positive trend and using them as a data source will be a great potential to extract patterns and mine data.

Figure 2 Number of mobile subscribers in Albania



Source AKEP (https://www.akep.al/en/statistika)

In the data warehouses of mobile data providers, Call Detail Records, represent the network activities as stored information regarding inbound roaming data, domestic data and outbound roaming data. The data are identified by user identifiers, time when the call was made, and the location based on the antenna ID. This data has a geographical dimension which can be used during data analyses. It also contains demographic data such as age, gender and nationality of users. To access this data there are legislative restrictions as mobile phone data is sensitive and protected by law. The government should sign a memorandum of understanding which clearly specifies the access and outlines the way by which the data is processed to protect the users’ privacy. Hashing algorithms should be applied to the data before data processing, to avoid direct or indirect identification of users. This would be a part of a detailed data security policy. The system that processes the data has to be hosted in the national data center, which is a secure government cloud environment. The mobile companies in Albania are not connected to this center, so solution for the transfer data should be developed.

The processed data can be used to estimate trip duration, inbound and outbound international travels. Analyzing the patterns of connected SIM cards, we can estimate the number of inbound travels from non-residents in the resident mobile operators and the number of outbound travels from roaming activity. Base on some predefined algorithms which consider short and long visits and adjust for transit travel (harbors, airports, transit roads) and look at permanent workers from other countries and border noise (ship traffic and random switching) predication can be made and statistical indicators can be produced. The data source of information for official tourism statistics is the Ministry of Interior Affairs, General Directorate Policy, which is an administrative source. The National Institute of Statistics in Albania does not conduct a frequent survey for tourism, because of restriction in financial resources.

The main benefits for the Albanian government will be an enhancement in data quality providing accurate data with a more detailed breakdown by region and demographic characteristics. Supplementary and new indicators in previously unavailable magnitude can be produced. Also, the data can be available in real time giving the possibility to produce monthly and quarterly estimates.

Conclusion

Most of the developed countries in Europe and in world, are trying to adopt Big Data technology in important areas such as healthcare, crime, agriculture and financial segments because of data deluge. Using this technology, countries face a lot of challenges which should be addressed including but not limited to technical, legal and methodological issues. In Albania, public institutions have taken no initial actions to use Big Data for good planning and decision making. For Albania, putting Big Data into an official statistics environment requires a radical paradigm shift in the legal framework and in data processing methodology. Regarding the IT infrastructure, we can conclude that the national data center has the appropriate infrastructure to start some pilot projects such as the one explained in this paper, mobile position for producing tourism indicators. The opportunities and challenges of using Big Data in producing official statistics should be a matter of an open debate because of the complexity of data management and the nature of the products of official statistics. Numerous applications of Big Data are already taken by other international organizations and Albanian institutions have to build good collaboration with them to get support in their respective area of expertise. Statistical agency in Albania should further step up its involvement in Big Data projects.

References

AKEP (2018, March). Treguesit statistikorë të tregut të komunikimeve elektronike Viti 2018, 1. Retrieved from https://www.akep.al/images/stories/AKEP/publikime/raporte/Raporti%20Vjetor%202017%20AKEP.pdf

Arouri, M., Aouadi, A., Foulquier, P., & Teulon, F. (2013). Can Information Demand Help to Predict Stock Market Liquidity? Google it! Working Paper: IPAG Business School, France, 024, 1-30.

Beyer, M.A. & Laney, D., (2012). The Importance of “Big Data”: A Definition. Gartner Publications, pp.1–9.

Cross-Cutting Strategy “DIGITAL AGENDA OF ALBANIA 2015-2020” (2018). Retrieved from http://akshi.gov.al/wp-content/uploads/2018/03/Digital_Agenda_Strategy_2015_-_2020.pdf

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. Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622 Accessed: 01.03.2018.

Henninger, M. (2013). The value and challenges of public sector information. Cosmopolitan Civil Societies: An Interdisciplinary Journal, 5(3), 75. doi:10.5130/ccs.v5i3.3429

Moreno, J., Serrano, M., & Fernández-Medina, E. (2016). Main Issues in Big Data Security. Future Internet, 8(3), 44.

UNECE(2013) “ What does ‘big data’ mean for official statistics?”, 10 March 2013. Retrieved from https://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf

Quick, D., & Choo, K. R. (2014). Impacts of increasing volume of digital forensic data: A survey and future research challenges. Digital Investigation, 11(4), 273-294.