Statistical Journal of the IAOS 39 (2023) 171–177
DOI 10.3233/SJI-220092
IOS Press

Overview of the use of big data for official statistics in Latin America and the Caribbean

Andrea Diniz da Silva*, Beatriz Menezes Marques de Oliveira, Ísis Gonçalves Peixoto and Lidiane Braga Sales de Souza
National School of Statistical Sciences, ENCE/IBGE, Rio de Janeiro, Brazil

Abstract. In 2020 and 2021, the challenges related to the decline in the financing of statistical production and the cooperation of respondents was exacerbated by the COVID-19 pandemic. This scenario led national statistical offices (NSOs) to accelerate consideration of alternative data sources to complement or even replace traditional survey data. In this context, the use of big data to produce statistics has become promising. The use of big data for statistics is already in practice in many parts of the Global North and has also been spreading rapidly in the South. Part of the success of this trend is due to the support of the United Nations Committee of Experts on Big Data and Data Science for Official Statistics (UNCEBD), in particular its four Regional Hubs for Big Data. To learn the extent of the use of big data for official statistics in Latin America and the Caribbean, the United Nations Regional Hub for Big Data in Brazil conducted a study of the practices of NSOs in the region. A very promising scenario was found regarding the use of big data from satellite imagery, web scraping and other big data sources, for applications such as the production of price statistics, land use and cover patterns and migration.

Keywords: Big data, official statistics, experimental statistics, Latin America and the Caribbean

1. Introduction

In 1994, the United Nations published Principle 5, according to which "data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents" as part of the Fundamental Principles of Official Statistics (FPOS) [1]. Despite not mentioning big data, it is intuitive that nowadays Principle 5 includes this (not so) new type of data, called big data, 25 years after the United Nations made it explicit when including big data as alternative data in its National Quality Assurance Frameworks Manual for Official Statistics [2].

From principles to implementation guidance, the need has increased for national statistical offices (NSOs) to include all types of data to leverage statistical production, which means big data as well. It is a predictable consequence of the pressure on NSOs to improve efficiency meaning collection of more statistics at lower cost, not only in terms of budget resources, but also reducing the burden on respondents

Since 2020, the challenges faced in the previous decades from the declining financing of statistical production and cooperation of respondents have been exacerbated by the COVID-19 pandemic. As a consequence, there is even more pressure on NSOs to consider alternative data sources to complement or replace traditional surveys. In this context, alternative data sources such as administrative records and big data have gained even more importance, becoming essential for the improvement of statistical production.

The use of big data to produce statistics is already in practice in many countries, such as Australia, Canada, China, France, Germany, Switzerland and Sweden among many others [3]. Part of the success of this trend is the support of the United Nations through the Committee of Experts on Big Data and Data Science for Official Statistics (UNCEBD), which was created at

*Corresponding author: Andrea Diniz da Silva, National School of Statistical Sciences, ENCE/IBGE, Rio de Janeiro, Brazil. E-mail: andrea.silva@ibge.gov.br.

the 45th session of the Statistical Commission to investigate the benefits and challenges of big data and its potential for monitoring the Sustainable Development Goals. Since 2014, UNCEBD has been growing, and now it already counts on the Global Platform, nine Task Teams and four Regional Hubs for Big Data in different parts of the world, such as Brazil, China, Rwanda and United Arab Emirates [4].

To learn of the extent of the use of big data for official statistics in Latin America and the Caribbean (LAC), the United Nations Regional Hub for Big Data in Brazil conducted a twofold study on the use of big data by the NSOs in the region. First, web scraping of the websites of the NSO was conducted and then consultation was carried out from November 2021 to April 2022. This paper provides an overview of the use of big data for official and experimental statistics in the region, the types and sources of big data in use or targeted the topics being considered and the challenges pointed out by the NSOs to include big data in their current collection efforts

2. Big data for statistics

There is an immense amount of data generated by today's digital life, which is found in the form of big data from varied sources. People's "digital footprints" leave a trail of data that can be analyzed and used to produce statistics. Digital interactions, such as online surveys, online purchases, posts or even the digital content consumed by people, generate a large volume of data. This emerges in a context of growing demand for data to support more informed decisions associated with the difficulties in carrying out surveys in the traditional way. The growing demand is motivated to some extent by the development of information technologies and state-of-the-art big data processing methods, the internet of things (IoT), the most important source of voluminous amounts of data, specifically "self-quantified, multimedia and social media data", which make it possible to obtain real-time insights with the power to map, describe, monitor, predict and prescribe the behavior and characteristics of a population, for public and private purposes [5,6]. On the other hand, difficulties in performing surveys are associated with the rising costs and increasing nonresponse rates. Therefore, attempts to produce statistics using big data combined with administrative data and sample surveys have become more necessary and frequent [7].

In this respect, the most common tools, methods and data sources that are being investigated for their potential to produce official statistics are web scraping, medical record systems, smart meters, genetic profile data, automatic identification systems (AIS), commercially compiled data, financial data, transaction data, mobile phone data, vessel and flight data, scanner data, sensors, pictures and videos, satellites, the global positioning system (GPS), geographic information systems (GIS) and machine learning [7].

In recent years, statistical agencies had registered more than a hundred projects based on big data sources, such as satellite imagery, remote sensors and smart meters among others, in the Big Data Project Inventory. Despite the apparent infinity of these data sources, their use is not straightforward and often lacks representativeness and stability. Regardless of these challenges, national statistical offices are using different big data sources to produce official statistics [8].

National statistical offices (NSOs) and international organizations (IOs) are using big data sources to compute sustainable development goal indicators to monitor the Agenda 2030. Latin American and Caribbean countries are generating indicators through these alternative data sources. Some examples are Honduras, which collects and combines data from satellite imagery and administrative records to map poverty and malnutrition [9]; Colombia, which is using satellite imagery to calculate indicators that require data on land surface such as green areas or streets [10]; and Brazil, where the Brazilian Institute of Geography and Statistics (IBGE) is carrying out an experimental project that proposes an alternative method of collecting price information using web scraping [11].

However, the use of big data generally has different aims between developed and developing countries. While in developed countries big data seem indeed to be a good information base to create reliable proxies of social indicators and to complete official statistics analyses, in developing countries the use of big data may be a viable alternative to traditional surveys [12].

3. The study of the use of big data

To investigate the use of big data for producing statistics, we conducted web scraping on the web pages of the national statistical offices (NSOs). This was an active search performed at the 34 NSO websites available in the M49 list of the United Nations [13]. This activity was carried out systematically in four steps: 1) visit the sites of the main NSOs to classify the address as active or inactive; 2) look for search engines and classify them

as available or not available; 3) for those with search engines, perform the search and classify the results as returned searched content, returned other content, or returned empty; 4) for those returning content, collect and classify the content found

The web scraping of the NSO webpages allowed identifyng only four countries in a universe of 22 with active pages and an available search engine, that use big data for statistics: Brazil, Chile, Colombia and Mexico. This means that these countries have made public at their websites the use of web scraping, satellite imagery and social networks to produce price and poverty indexes; obtain geographic information about the national territory in digital elevation models; as well as to conduct studies of subjective wellbeing.

Following the exploratory phase, the International Consultation on the use of big data for statistical production in Latin America and the Caribbean was sent in November 2021, by the United Nations Economic Commission for Latin America and the Caribbean (ECLAC), on behalf of the UN Regional Hub for Big Data in Brazil. A first reminder was sent two weeks after the first call and a second one month after. The questionnaire was available in three languages – Spanish Portuguese and English – the main official languages spoken in the region. It was organized into two blocks. The first one contained questions to identify the NSO and the respondent, and the second addressed the use of big data in statistical production. There were 13 questions (12 closed and one open-ended), not all mandatory.

The identification questions were all mandatory and included country, name of the NSO, name of the respondent, and e-mail for contact on the topic. As for the subject matter there were four main questions: "Q1. Does the Institute of Statistics use sources of big data in the production of official statistics?"; "Q2. Does the Institute of Statistics use sources of big data in the production of experimental statistics?"; "Q3. Is the Institute of Statistics conducting studies or tests of the use of big data sources in the production of statistics?"; and "Q4. Is the Institute of Statistics considering conducting studies or tests for the use of big data sources in the production of statistics?". For all of them the options were "Yes" or "No". Questions Q1 to Q3 were mandatory and Q4 was available only for the NSOs that answered "No" to Q1 Q2 and Q3.

For the NSOs that answered "Yes", additional mandatory questions about the type or source of big data and the topics covered were presented. Options were designed based on [9]. For type or source of big data the options were: web scraping, scanners, mobile phones/CDRs, social media, satellite imagery, smart meters, credit cards, road sensors, health records, ship identification, criminal records as well as "other". For the topics, the options were prices, population/migration, transport/mobility, geographical/spatial, labor market, agriculture/land use, tourism, health/disease, energy/environment, crime/corruption, poverty/inequality, disaster risk reduction as well as "other". A "select all that apply" (SATA) approach was used.

Fig. 1. Response to the International Consultation on the use of big data for statistics in Latin America and the Caribbean.

Additionally, all NSOs, regardless the previous answers, were given an opportunity to express their thoughts on the use of big data for official and experimental statistics in an open-ended question. This item had the following description: "Use this space to talk about the limitations of the use of big data sources in the statistical production in the country, about the needs of the Institute in this area and about other issues that you find relevant."

3.1. Response rate

Until May 2022, the response rate reached 70% in Latin America (14 of 20 countries) and less than 1% in the Caribbean (2 of 28 countries). The 16 NSOs that reported their situation regarding the use of big data were from Argentina, Belize, Brazil, Chile, Costa Rica, Colombia, Cuba, Ecuador, Mexico, Panama, Paraguay, Peru, Dominican Republic, Suriname, Uruguay and Venezuela (Fig. 1).

4. Use of big data by National Statistical Offices in LAC

Seven countries answered positively to the question "*Does the Statistical Office use sources of big data in*

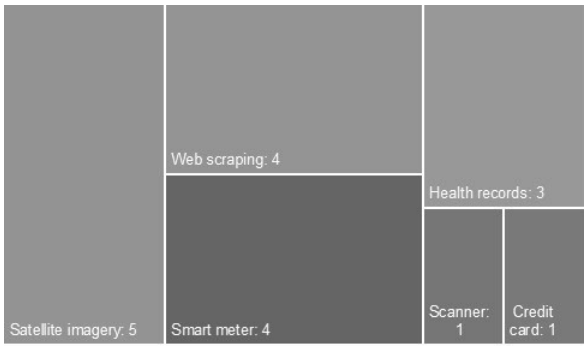

Fig. 2. Big data used for official statistics.

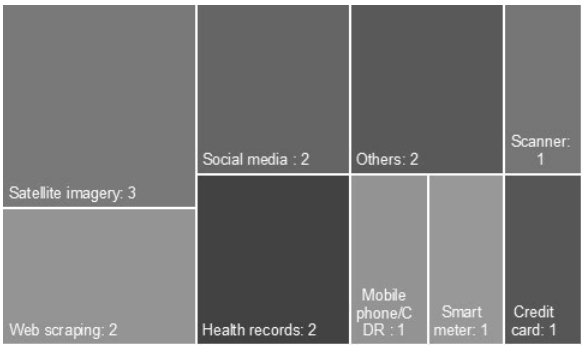

Fig. 4. Big data used for experimental statistics.

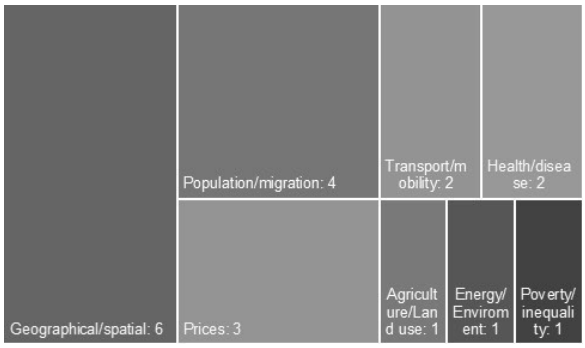

Fig. 3. Topics using big data in official statistics.

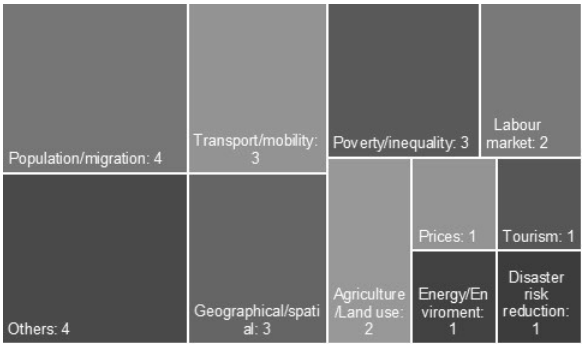

Fig. 5. Topics of big data in experimental statistics.

the production of OFICIAL STATISTICS?". They were Argentina, Brazil, Chile, Colombia, Costa Rica, Peru and Uruguay. Except for Costa Rica, which is in Central America, all of them are in South America. These countries accounted for 44% of the total respondents, which means that almost half of the countries that responded were already using big data for official statistics. This shows a very positive scenario in the region considering that the use of big data is a very recent practice in NSOs around the world and that the methods and tools are still in development.

Satellite imagery and web scraping were indicated as among the top three most used big data sources, which can be explained by the fact that they provide free and open data. Smart meters was also among the top three sources, although this source is private in most of the countries (Fig. 2).

Big data offers an immensity of possibilities and this diversity can be seen in the number of topics already being produced in Latin America (Fig. 3). Six of seven countries that use big data for official statistics are producing geographical and spatial information, which is consistent with the fact that satellite imagery is among the top three most used big data sources. Migration is one of the topics that can benefit from social network data and scraped prices are increasingly used in CPI. Considering that, it is also possible to see consistency between the indicators produced and the big data presented in the previous section.

The use of big data in the production of experimental statistics is less frequent compared to the use in official statistics. For experimental statistics the use of big data is practiced in 31% of the NSOs, mostly located in South America (Argentina, Brazil, Colombia, Mexico, and Uruguay).

It is worth to pointing out that all these countries gave an affirmative response both regarding the use of big data in experimental statistics and official statistics, except for Mexico. This could be related to the difficulty of distinguishing between experimental and official statistics.

According to the Office for National Statistics of United Kingdom (ONS) [14] and the Brazilian Institute of Geography and Statistics (IBGE) [15], experimental statistics are statistics that are in the testing phase and not yet fully developed. For the National Administrative Department of Statistics (DANE) [16], experimental statistics arise from projects that have innovative as-

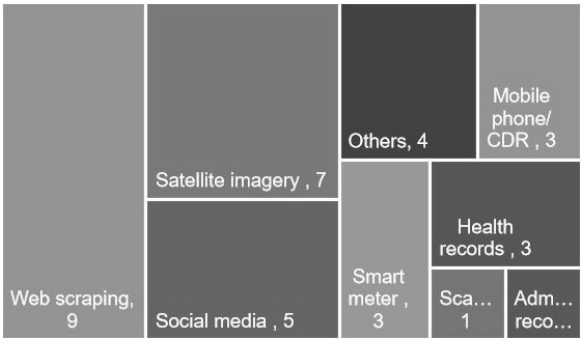

Fig. 6. Big data used in studies or tests.

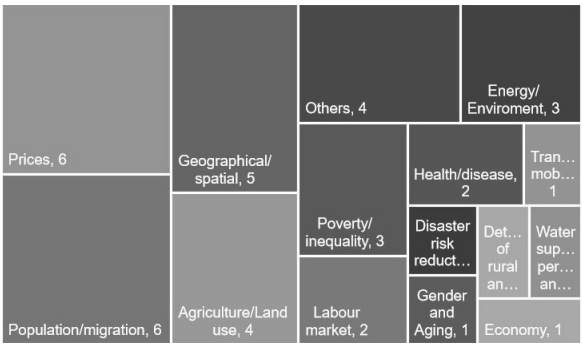

Fig. 7. Topics with big data in studies or tests.

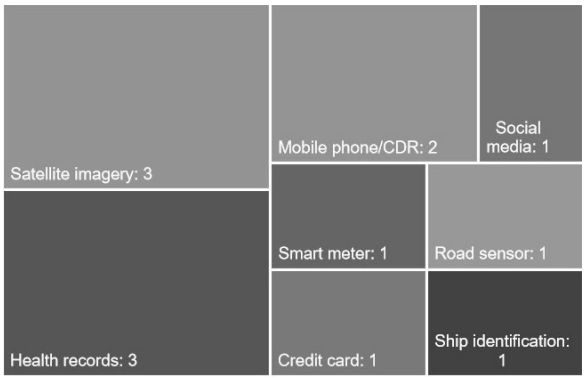

Fig. 8. Big data considered for future use.

Fig. 9. Topics with big data considered for future use.

pects, whether using new methods or new data sources. And since they still can be improved, they did not reach the quality, reliability and stability of data to be included in the inventory of regular statistics. However, not all NSOs have a clear definition about this concept, which may lead to confusion.

Regarding the responses, the most used source of big data for producing experimental statistics was satellite imagery, followed by social media, web scraping and health records (Fig. 4). It is noteworthy that the field "others" received two responses which were all related to the use of administrative records. The continuation of satellite imagery among the most used types of big data also for experimental statistics can be explained by same reason indicated before, i. e., it may be linked to the availability of cost-free data. When considering web scraping and social media together, another mostly cost-fee data source emerges: the Internet.

As for the topic population/migration is the main one produced using big data for experimental statistics, while transport/mobility, geographical/spatial and poverty/inequality together are the second most common (Fig. 5) Under "Others" economic activity appears twice.

Eleven countries answered positively the question "Is the Institute of Statistics conducting studies or tests for the use of big data sources in the production of statistics?". They were Argentina, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Mexico, Peru and Uruguay. This means that answers come from all over Latin America and the Caribbean area, since Costa Rica is in Central America, Cuba and Dominican Republic are in the Caribbean, and the remaining countries are in South America. The NSOs conducting studies or tests to use big data for experimental or official statistics amounted to 68.75%. This result indicates that NSOs are committed to developing research on the use of this alternative approach.

The most indicated sources were web scraping; satellite imagery; social media and mobile phone/CDR, smart meters and health records (Fig. 6). Again web scraping and satellite imagery are among the most used sources in addition to social media.

Regarding the topics, the most mentioned were price index and population/migration, geographical/spatial and agriculture/land use (Fig. 7). Other topics also

frequently mentioned were poverty/inequality and energy/environment. These results are consistent with the most cited type of big data used for the studies. Population/migration can be obtained from social media, geographical/spatial data from satellite imagery and both from web scraping.

Only five NSOs declared that they were not using big data to conduct studies or tests, those from Belize, Panama, Paraguay, Suriname and Venezuela. For those, the question "Does the Institute of Statistics consider carrying out studies or tests for the use of big data sources in the production of statistics?" was posed. Panama, Paraguay and Venezuela answered positively, pointing to the use of satellite imagery and health records. Surprisingly, web scraping and social media were not mentioned (Fig. 8), breaking the pattern seen regarding the use of big data and studies on the use of big data.

For to those NSOs, big data could be used to produce statistics on demographics, geography, prices and health care, which are also common topics also among the NSOs that already produce official and experimental statistics using big data, among other topics (Fig. 9).

Challenges and limitations of the use of big data have been widely discussed in the literature [7,8,17–19], even though some of these works are not exclusively about big data, also including themes such as survey data and records. To obtain a view on how it works in Latin America and the Caribbean, an open-ended space was provided under the heading "Use this space to write about the limitations of the use of big data sources in the statistical production in your country, about the needs of the Statistical Office in this area and other issues that you may find relevant."

Lack of scientific and technological capacity, insufficient trained human resources, unstandardized internal structure of data, lack of legislation, poor quality, lack of cooperation policy, and bureaucratization to access big data were some of the most common concerns reported by the NSOs participating in the consultation.

Scientific and technological capacity and training are the biggest concerns in the region, being mentioned by Brazil, Chile, Colombia, Ecuador, Paraguay, Suriname and Venezuela in South America; Belize, Costa Rica and Panama in Central America; and Cuba and Dominican Republic in the Caribbean.

The concern with standardization and the quality of big data was expressed by Mexico, Suriname, Uruguay, Colombia and Dominican Republic, but it also may be an issue in other countries. With the exception of Suriname, the institutes of these countries are already carrying out studies or tests for the use of big data sources for the production of statistics.

5. Final remarks

The study showed that the use of big data is already a reality in Latin America. Of the 14 Latin American countries that responded the consultation, only two (Belize and Suriname), did not state an intention to use this type of alternative data. Unfortunately, the low response rate in the Caribbean did not allow an overview of what is happening in this part of the region. However, the two respondents, Cuba and Dominican Republic, are producing experimental statistics based on big data.

Among types of data sources investigated, satellite imagery web scraping stood out. This can be explained by the fact they are more associated with open data. Many free satellite images from authoritative sources are available, and the Internet is a virtually unlimited source of free data, although less reliable. In both cases, open source and free tools as well as tutorials are available. Other data sources, although private, such as mobile phone data, transaction data and many others, can be easily accessed with consent via the standard terms of use.

The topics produced using big data are linked to the type of data most used. As expected, prices and geographic information were among the most popular topics. Population/migration data were considered to be important. Despite not being among the topics in official statistics, they were mentioned in studies and plans. Perhaps because official statistics come from the traditional censuses and surveys, there is an opportunity of look at other alternatives.

Despite the very positive scenario the study highlighted the low visibility of the use of big data in the region, since the preliminary research carried out on the pages of the NSOs identified the use of big data in only four countries. Giving more visibility could help understanding the state of the art in the region and facilitate cooperation, especially among neighboring countries.

There are a number of immediate challenges, though not unique to big data, namely access, methods and tools for big data. These challenges are related to others such as legislation and confidentiality. Even if all these challenges ae overcome, others can emerge related to computing infrastructure and trained professionals, among others. There is much to be done, but advances in technology, methodology and cooperation are making great strides. In this way, the use of big data for official statistics will soon be the norm.

Acknowledgments

The authors are grateful to the anonymous reviewers for their constructive comments as well as to Katia Aragão Eusébio and Mariza Rayanne da Silva Pereira for the work carried out in the data collection phase. The second author thanks the Office to Coordinate Improvement of Higher Education Personnel (CAPES), and the third author thanks the Brazilian Institute of Geography and Statistics (IBGE) for the scholarship granted.

References

[1] United Nations. Fundamental Principles of Official Statistics. Resolution adopted by the General Assembly on 29 January 2014. [without reference to a Main Committee (A/68/L.36 and Add.1)]. Sixty-eighth session. Agenda item 9. 2014.

[2] United Nations. National Quality Assurance Frameworks Manual for Official Statistics. 2019. Including recommendations, the framework and implementation guidance. Department of Economic and Social Affairs. Statistics Division. Studies in Methods Series M No. 100.

[3] Halderen GV. SDGs and Big Data. 2021. How many countries are using big data sources for the SDG indicators? Are we in the midst of a big data revolution? Apresentação feita no Big Data for the SDGs – What Is The Way Forward. An Interactive Exchange of Views. Side Event of the 52nd Un Statistical Commission.

[4] UNBigData. Available from: https://unstats.un.org/bigdata/about/index.cshtml.

[5] Yaqoob I, Hashem IAT, Gani A, Mokhtar S, Ahmed E, Anuar NB, Vasilakos AV. Big data: From begin to future. International Journal of Information Management. Elsevier. 2016; 36(6): 1231-1247. doi: 10.1016/j.ijinfomgt.2016.07.009.

[6] Letouzé E, Jütting J. Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach. Data Pop Alliance White Paper Series. Available from: https://paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf.

[7] Japec L, Lyberg L. Big Data Initiatives in Official Statistics. In: Big Data Meets Survey Science. John Wiley & Sons, Ltd; 2020. pp. 273-302. doi: 10.1002/9781118976357.ch9.

[8] MacFeely S. The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. United Nations Conference on Trade and Development. 2019. doi: 10.1111/1758-5899.12595.

[9] Regional Knowledge Management Platform – 2030 Agenda in Latin America and the Caribbean. Available from: https://agenda2030lac.org/es/paises.

[10] DANE – Departamento Administrativo Nacional de Estadística. Estadísticas experimentales. Available from: https://www.dane.gov.co/index.php/estadisticas-por-tema/estadisticas-experimentales.

[11] Dantas TM, Silva LT. Coleta automática de dados para Índices de preço e ajustamento de qualidade utilizando web scraping. IBGE – Instituto Brasileiro de Geografia e Estatística. 2018. Avaliable from: https://eventos.ibge.gov.br/downloads/smi2018/apresentacoespdf/ST3%20-%20Tiago%20Mendes.pdf.

[12] di Bella E, Leporatti L, Maggino F. Big Data and Social Indicators: Actual Trends and New Perspectives. Soc Indic Res. 2018 Feb 1; 135(3): 869-78. doi: 10.1007/s11205-016-1495-y.

[13] UNSD – United Nations Statistical Division. National Statitical Offices Websites. Available from: https://unstats.un.org/unsd/methodology/m49/.

[14] ONS – Office for National Statistics. Methodology topics and statistical concepts: Guide to experimental statistics. Available from: https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/guidetoexperimentalstatistics.

[15] IBGE – National Brazilian Institute of Geography and Statistics. Experimental Statistics. Available from: https://www.ibge.gov.br/estatisticas/investigacoes-experimentais/estatisticas-experimentais.

[16] DANE – Departamento Administrativo Nacional de Estadística. Estadísticas experimentales. Available from: https://www.dane.gov.co/index.php/estadisticas-por-tema/estadisticas-experimentales.

[17] Braaksma B, Zeelenberg K, De Broe S. Big Data in Official Statistics. Big Data Meets Survey Science. 2020; 303-338. doi: 10.1002/9781118976357.ch10.

[18] Kitchin R. The Opportunities, Challenges and Risks of Big Data for Official Statistics. Statistical Journal of the International Association of Official Statistics. 2015 Jan 1; 471-481. doi: 10.2139/ssrn.2595075.

[19] Tam S-M, Clarke F. Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. International Statistical Review. 2015; 83(3): 436-448. doi: 10.1111/insr.12105.