Use of Big Data in Compilation of SDG Indicators in the Arab Region

Challenges and Opportunities

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity Dignified Life

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

VISION

ESCWA, an innovative catalyst for a stable, just and flourishing Arab region

MISSION

Committed to the 2030 Agenda, ESCWA's passionate team produces innovative knowledge, fosters regional consensus and delivers transformational policy advice. Together, we work for a sustainable future for all.

E/ESCWA/CL4.SIT/2021/TP.2

Economic and Social Commission for Western Asia

Use of Big Data in Compilation of SDG Indicators in the Arab Region

Challenges and Opportunities

United Nations
Beirut

Author: Maria Simona Andreano and Giovanni Savio.

United Nations publication issued by ESCWA, United Nations House, Riad El Solh Square,
P.O. Box: 11-8575, Beirut, Lebanon.

Website: www.unescwa.org

Acknowledgements

This report was prepared by Maria Simona Andreano (University of Mercatorum in Rome) and Giovanni Savio, under the coordination of Ismail Lubbad, Statistician, ESCWA Statistics Division. The authors would like to express their gratitude to Haidar Fraihat, Senior Advisor of Technology at ESCWA, and Fouad Murad, Senior Programme Manager at ESCWA, and participants of the Regional Workshop on the Integration of Big Data and Geospatial Information for the Compilation of SDG Indicators in Arab Countries, 13-15 October 2020, for their comments on the presentation therein made.

The Authors also thank Zeina Sinno, Nada Moudallal and Mohamad Hossary, from the ESCWA Statistics Division, for the support provided.

Executive summary

Following the adoption by the United Nations General Assembly in September 2015 of the Sustainable Development Goals, National Statistics Offices are now compelled to undertake a data revolution, which implies a titanic challenge to their mandate: populate the SDGs matrices in all their economic, social and environmental dimensions.

The need for greater coverage, timeliness and disaggregation of the indicators to fit SDG data requirements implies imposing a very strong effort – especially for less statistically developed countries worldwide – to reconsider the ways statistics organizations work and plan for the future.

Recently, the COVID-19 crisis has added a further challenge, as the pandemic is disrupting operations routinely undertaken to collect and compile basic statistics, jeopardizing or delaying planned censuses, surveys and other data programmes.

This technical report analyses how non-conventional sources of information, particularly those offered by the Big Data revolution, can benefit official statisticians in filling in existing gaps and provide valuable insights on sustainable development monitoring and reporting.

This report provides details on Big Data definitions, reviews the ongoing and innovation projects using Big Data sources for official statistics, discusses the *pros* and *cons* of using this new breed of statistical data sources, analyzes concrete examples of using remote sensing information to get insights on prominent statistical indicators such as poverty rates, economic activity and prices, and provides examples of how to use "information from the above" to obtain information on effects of the recent pandemic.

The report concludes that Big Data has great potential in helping the collection of data and information on a number of focus areas, including mobility, transport, tourism, prices, corruption and crime, energy consumption, population density, land use, well-being, and the labour market. Their use, although posing ethical, legal, technical and reputational challenges for statisticians, can improve timeliness and accuracy, reach greater granularity, increase disaggregation capabilities, and fill in official data gaps for many SDG indicators by 2030.

Key messages

- *Following the adoption of the Sustainable Development Goals (SDGs), National Statistics Offices worldwide are requested to undertake a data revolution, which implies a titanic challenge to their mandate, namely to populate the SDGs matrices in all their economic, social and environmental dimensions.*

- *Recently, the COVID-19 crisis has added a further challenge, as the pandemic is disrupting operations routinely undertaken to collect and compile basic statistics, jeopardizing or delaying planned censuses, surveys and other data programmes.*

- *Non-conventional sources of information, particularly those offered by the Big Data revolution, can help official statisticians in filling in existing gaps and provide valuable insights on sustainable development monitoring and reporting.*

- *This report provides details on Big Data definitions, reviews the ongoing and innovation projects using Big Data sources for official statistics, and discusses the pros and cons of their use.*

- *The report concludes that Big Data have great potential in helping the collection of data on a number of phenomena. Their use, although posing ethical, legal, technical and reputational challenges, can improve timeliness and accuracy, allow for greater granularity, and fill in official data gaps for a number of economic, social and environmental SDG indicators.*

Contents

List of tables

List of figures

Introduction

Following the adoption by the United Nations General Assembly in September 2015 of the Sustainable Development Goals (SDGs), National Statistics Offices (NSOs) and Systems (NSSs) worldwide, as well as statistics offices of international organizations (IOs), were asked to undertake a data revolution. Indeed, statistics professionals should extend both the scope and disaggregation of the data traditionally produced, and measure new economic, social and environmental phenomena by 2030, leaving no one behind.

To populate the SDGs matrices in all their dimensions, NSOs and NSSs are facing a titanic challenge. Not only coverage of the indicators poses serious problems, but also disaggregation in all its forms (temporal, geographical, by sex, income, age, etc.) implies a strong data requirement, which forces statistics organizations to reconsider how they work and their future plans.

Recently, the COVID-19 crisis has added a further challenge, as the pandemic is disrupting operations that NSOs and NSSs routinely undertake to collect basic statistics, jeopardizing or delaying planned censuses, surveys and other data programmes.

Because of these enormous challenges, there is a growing consensus among stakeholders that Big Data, in whatever form they take, might strengthen the value of traditional data sources and statistics in monitoring sustainable well-being and facilitate the transformative agenda that organizations should implement in the forthcoming years.

However, Big Data have been on top of the attention of official statistics during the last five to ten years, and considerations about their true and effective value added for official statistics have been raised by many, with some reluctance to their use by many national and international organizations. Based on those views, Big Data should have now passed the phase of inflated expectations and reached a descending curve towards oblivion.

This technical report aims at scoping the concept of Big Data in statistics and identifying the *pros* and *cons*, advantages and disadvantages, of using Big Data for official statistics. It also reviews the main points of the discussion, trying to figure out what role Big Data might play in the near future of statistics, with a particular focus on the Arab region. The report is taking material from a presentation delivered during a virtual meeting organized by the United Nations Economic and Social Commission for Western Asia (ESCWA) titled "Regional Workshop on the Integration of Big Data and Geospatial Information for the Compilation of SDG Indicators in Arab Countries", that took place from 13 to15 October 2020.

After the executive summary and this introduction, the report discusses the actual challenges faced by statistics organizations in the compilation of the SDG indicators and analyzes the definitions of Big Data in section 2. Section 3 contains a review of actual projects underway in Big Data for official statistics, while sections 4 and 5 cover examples of use of remote sensing Big Data sources for mapping

poverty indicators, and a brief overview of sensor sources with concrete SDG applications. A discussion on challenges and opportunities for Big Data in SDG compilation is included in section 6. A summary of key points is contained in the final section of the report.

1. Definitions of Big Data

Big Data are becoming the by-product of the increasing digitalization of our modern day life, a phenomenon that is likely to endure for years to come. There is no uniformly accepted definition of Big Data. Nonetheless, it is important to provide a definition by taking into consideration the whole ecosystem that produces and uses them. Indeed, Big Data are not simply "lots of data" and, despite the name, size is not the only defining feature, as it should be accompanied by reference to their other features, the so-called "Vs".

The definition of Big Data provided by the TechAmerica Foundation, although rather general, is the one we favour. It states that: "Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information".

Big Data are generally characterized by the so-called four Vs, namely Velocity, Volume, Veracity and Variety, as shown in the infographics of the IBM Big Data & Analytics Hub (figure 1).

Some authors also refer to other Vs that are relevant in order to fully describe Big Data. Amongst those characteristics are: their Value (information and insights that Big Data provide), Viability (quick and cost-effective assessment of a particular variable's relevance), Variability (due to changing definitions, irregularities in the data, existence of multitude of data dimensions resulting from multiple disparate data types and sources), and Visualization (the way of presenting the data in a manner that is readable and accessible). Figure 2 below tries to depict a hypothetical situation for a generic Big Data source in terms of an evaluation of the various V's characterizing the source.

The meanings of the four traditional Vs are summarized in table 1, which is drawn from the TechAmerica Foundation (2012). As correctly pointed out by Manske, Sangokoya, Pestre and Letouzé (2016), Big Data refers not only to data, but also to the whole ecosystem that produces and uses them. This gives rise to the three C's definition of Big Data, which are characterized by the union of Big Data Crumbs (new kind of passively generated data), Capacity (as the technical and human capacity to yield insights from this data), and Community (new actors from the private sector and the research community, for example).

Volatility refers to the "changing technology or business environments in which big data are produced, which could lead to invalid analyses and results, as well as to fragility in big data as a data source" (Hammer and others, 2017, p. 8). At first glance, the additional Vs may seem odd as they are not per se defining characteristics of the data or intrinsic to it. Nevertheless, volatility and veracity are extremely important additions for understanding the contribution that big data might make to compiling statistics and SDG indicators. A 6-V definition that includes "value", where value means that something useful is derived from the data, offers a superior definition, as it introduces the notion of cost-benefit, that is, the costs of investing in Big Data must be carefully weighed up against

what they might deliver in practical terms (figure 1). Like volatility and veracity, value is not an intrinsic characteristic, but as above, including this dimension is nevertheless useful.

Figure 1. Properties of Big Data – The 4 Vs (IBM Big Data & Analytics Hub)

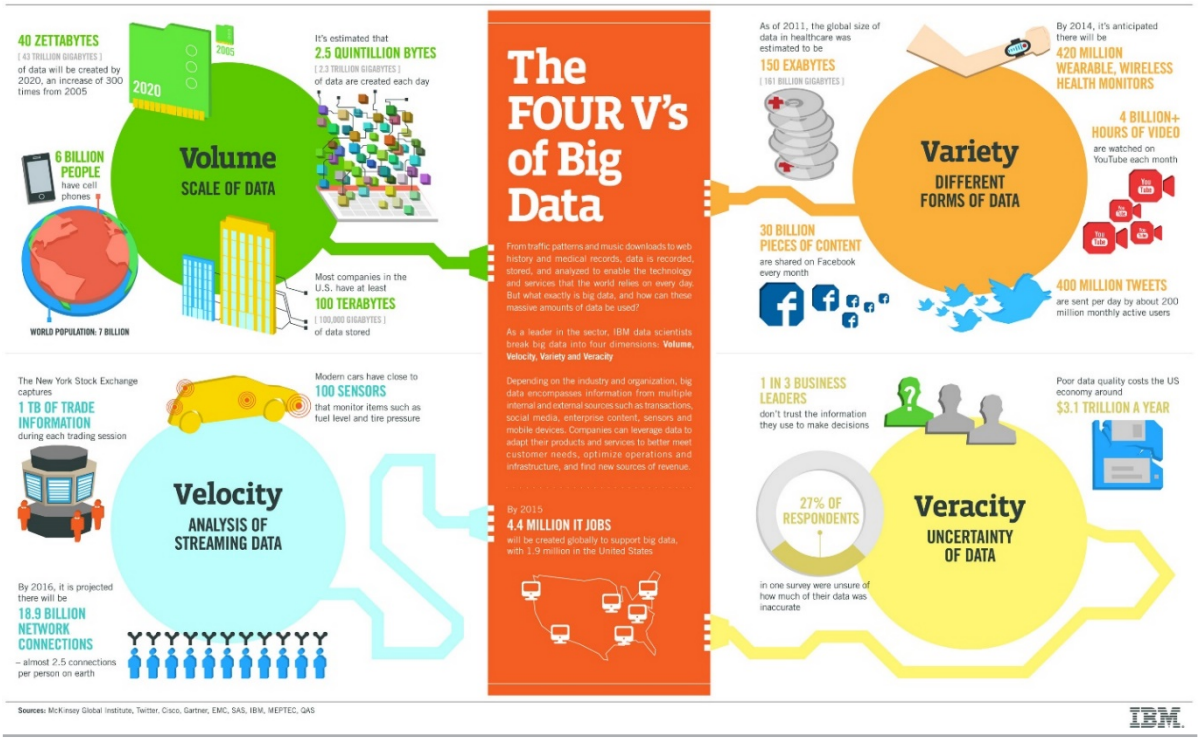


Figure 2. Radar chart with (theoretical) evaluation of various Vs of a Big Data source, i.e. low velocity, very high veracity and volume, and medium volatility and visualization

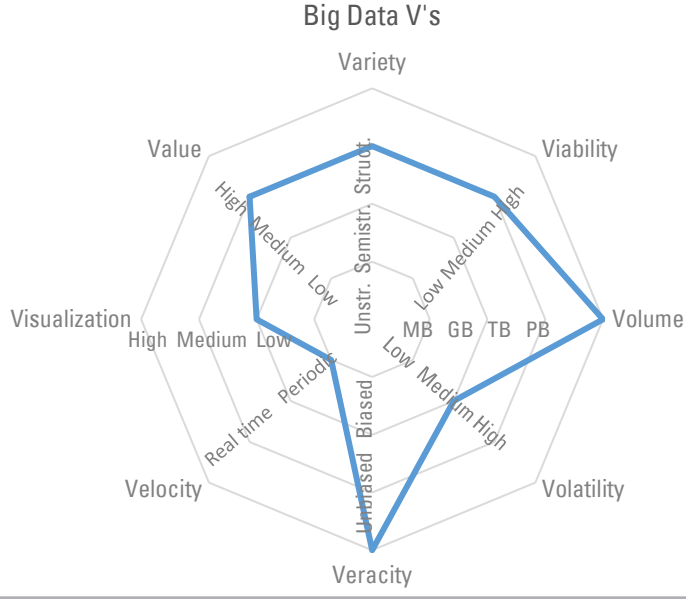


Table 1. Characteristics of Big Data

Characteristic	Description	Attribute	Driver
Volume	The sheer amount of data generated or data intensity that must be ingested, analyzed, and managed to make decisions based on complete data analysis	According to IDC's Digital Universe Study, the world's "digital universe" is in the process of generating 1.8 Zettabytes of information – with continuing exponential growth – projecting to 35 Zettabytes in 2020.	Increase in data sources, higher resolution sensors.
Velocity	How fast data is being produced and changed and the speed with which data must be received, understood and processed	• Accessibility: Information when, where and how the user wants it, at the point of impact; • Applicable: Relevant, valuable information for an enterprise at a torrential pace becomes a real-time phenomenon; • Time value: real-time analysis yields improved data-driven decisions.	• Increase in data sources; • Improved thru-put connectivity; • Enhanced computing power of data generating devices.
Variety	The rise of information coming from new sources, both inside and outside the walls of the enterprise or organization, creates integration, management, governance, and architectural pressures on IT	• Structured – 15 per cent of data today is structured, row, columns; • Unstructured – 85 per cent is unstructured or human generated information; • Semi-structured – The combination of structured and unstructured data is becoming paramount; • Complexity – where data sources are moving and residing.	• Mobile; • Social Media; • Videos; • Chat; • Genomics; • Sensors.
Veracity	The quality and provenance of received data	The quality of Big Data may be good, bad, or undefined due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations.	Data-based decisions require traceability and justification.

Source: TechAmerica Foundation (2012).

Big Data types are classified according to a definition that is mostly based on data sources, (United Nations Economic and Social Council, 2013), as follows:

- Data and information 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., camera data, satellite imaging, road sensors and climate sensors, such as those pertaining to Remote Sensing data sources;
- 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;
- Geographic information system (GIS) data and information of various sources and types.

Administrative data, traditionally organized in a structured way by public administrations (sometimes referred to by statisticians as administrative records), should not be classified as Big Data, but could become such if the velocity and volume characteristics would increase, as it seems appropriate in the future.

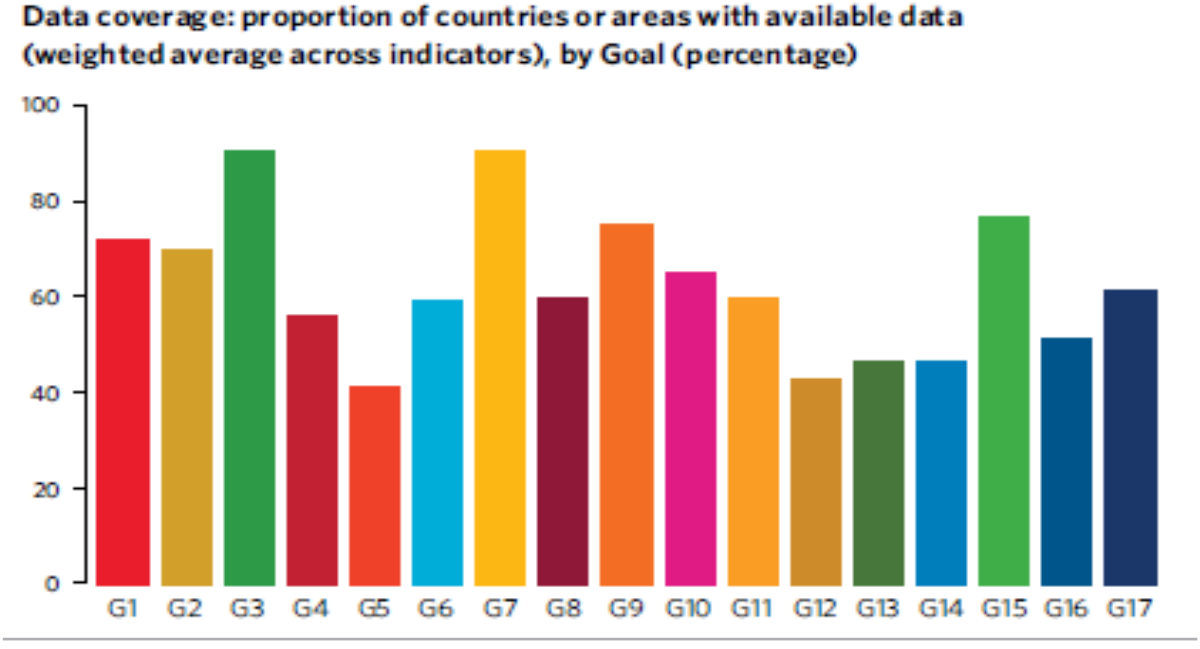
2. Project under Way in Big Data

Although good progress has been made in increasing the availability of internationally comparable data for the SDGs, important data gaps still persist in terms of geographic coverage, timeliness and the level of disaggregation.

An analysis of the indicators recognized in the Global SDG Indicators Database, available at https://unstats.un.org/sdgs/indicators/database, reveals that for four out of the 17 SDGs, less than half of 194 countries or areas have internationally comparable data (figure 3). This is particularly worrying for Goal 5 (on gender equality), as well as Goals 12, 13 and 14 (on responsible consumption and production, climate action, and life below water, respectively). More importantly, countries with available data have only few observations over time, making it difficult for policy makers to monitor progress and identify dynamics.

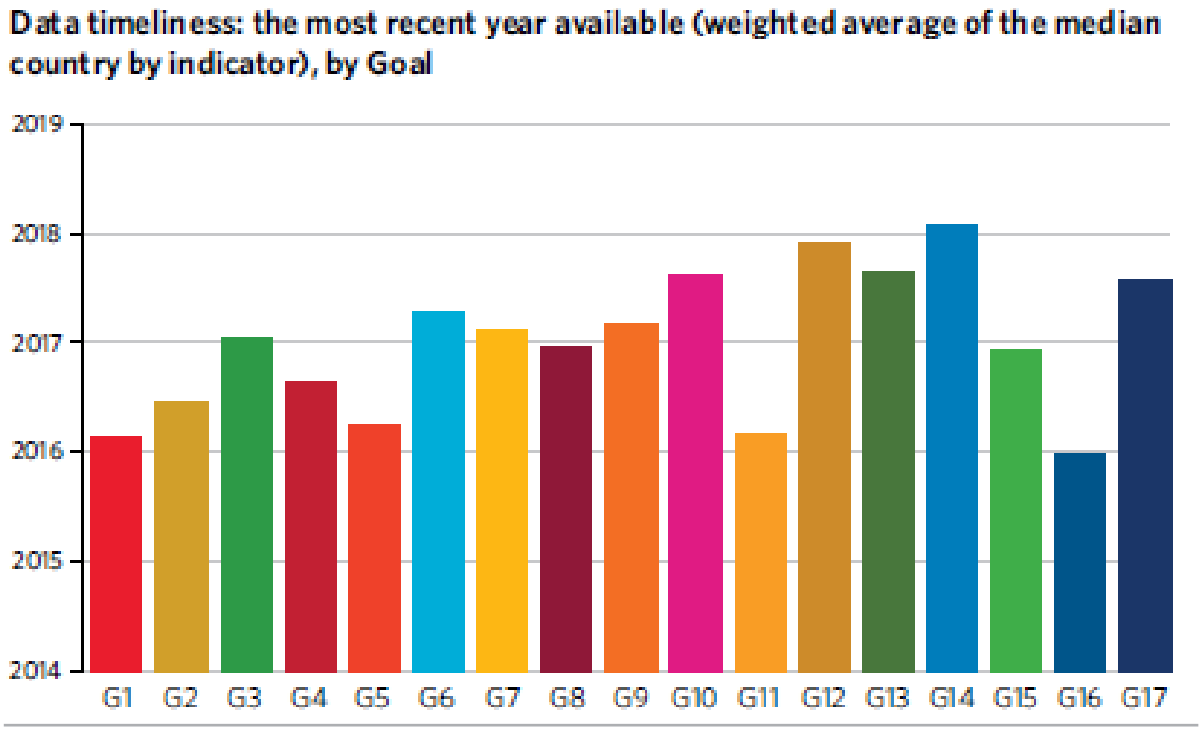
A number of SDGs are available with a long time lag (figure 3). In at least 50 per cent of countries or areas in the database, the latest data available for poverty indicators (Goal 1) are for 2016 or earlier. A similar situation is found for indicators on gender equality (Goal 5), sustainable cities (Goal 11), and peace, justice, and strong institutions (Goal 16). This means that, persisting this way, information on the 2030 SDGs will be available only four years after the target is going to be analyzed, namely in 2034.

Figure 3. Data coverage by SDG



Source: The Sustainable Development Goals Report 2020. Available at https://unstats.un.org/sdgs/report/2020/.

Figure 4. Data timeliness by SDG



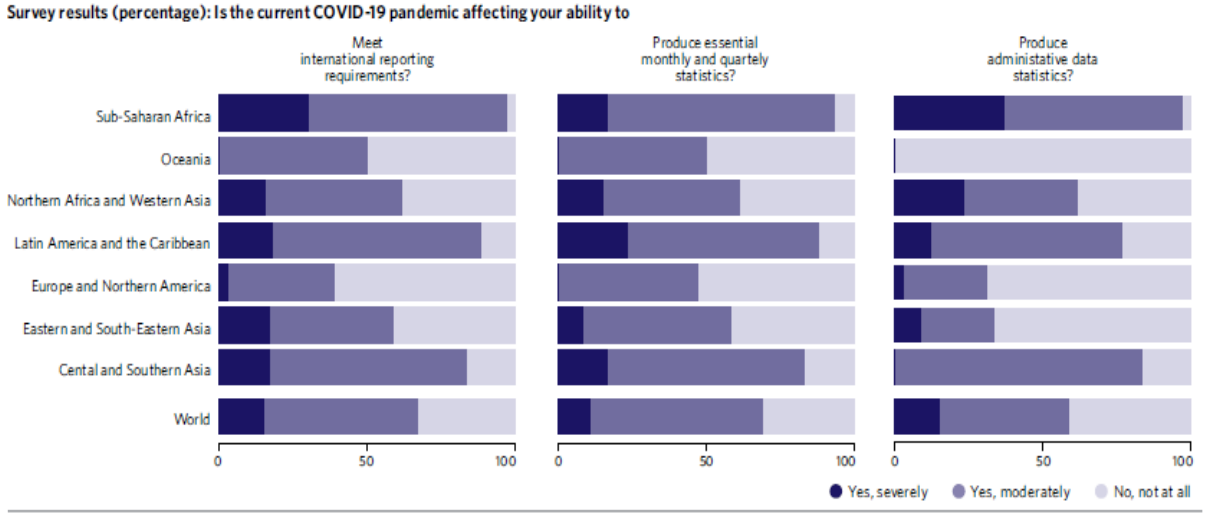
Source: The Sustainable Development Goals Report 2020. Available at https://unstats.un.org/sdgs/report/2020/.

As governments attempt to contain the spread of COVID-19, regular data collection operations are being disrupted. This jeopardizes the ability of many NSOs to deliver official monthly and quarterly statistics, as well as the data necessary to monitor progress on the SDGs (figure 4).

A recent survey conducted by the United Nations and the World Bank (with responses from 122 countries) shows that the pandemic has affected the operations of a number of NSOs: 65 per cent of headquarters are partially or fully closed, 90 per cent have instructed staff to work from home, and 96 per cent have partially or fully stopped face-to-face data collection. In Northern Africa and Western Asia, 62 per cent of countries surveyed indicated that the production of regular statistics, including monthly and quarterly data, as well as business registers, was affected: the same percentage indicated that they were having difficulty meeting international data reporting requirements.

According to survey results, nine out of ten NSOs in low- and lower-middle-income countries have seen funding cuts and are struggling to maintain normal operations during the pandemic.

Figure 5. COVID-19 Impact on Statistics Activities



Source: The Sustainable Development Goals Report 2020. Available at https://unstats.un.org/sdgs/report/2020/.

Role of big data

Facing this situation, a quite obvious question arises: is there any possible role that Big Data can play to assist and facilitate NSOs and NSSs worldwide in their efforts to provide users with a more complete and timely picture of SDG developments, particularly in the ESCWA region?

Let us consider now a very broad overview of exemplary sources that could potentially be utilized in compiling SDG indicators, obtained in great part from the UN Global Pulse infographics.

- Goal 1: Spending patterns on mobile phone services used as proxy indicators of income levels; remote sensing of night lights and population data;
- Goal 2: Crowdsourcing or tracking of food prices listed online;
- Goal 3: Mapping the movement of mobile phone users;
- Goal 4: Citizens reporting on reasons for students drop out;
- Goal 5: Financial transactions indicating spending patterns and responses to shocks;
- Goal 6: Sensors connected to water pumps tracking access to clean water;
- Goal 7: Smart meters allowing companies to modify levels of electricity, water and gas distributed in order to reduce waste and ensure proper supply in peak periods;
- Goal 8: Patterns in global postal traffic, night lights from remote sensing can provide indications on economic growth, remittances and trade;
- Goal 9: Data from GPS devices used for traffic control and improvement of public transportation;
- Goal 10: Speech-to-text analytics on local radio can reveal discrimination;
- Goal 11: Satellite remote sensing can reveal encroachment on lands, forests and parks;
- Goal 12: Online search patterns and e-commerce transactions can provide information on transition to energy efficient products;
- Goal 13: Combination of satellite images, crowd-sourced witness accounts and open data can help track deforestation;

- Goal 14: Maritime vessels tracking data, obtained through remote sensing, can reveal illegal and unregulated traffics and fishing activities;
- Goal 15: Social media data can help disaster management with real-time information on victims' location, as well as effects and strengths of forest fires;
- Goal 16: Sentiment analysis of social media might reveal public opinion on governance, public services or human rights;
- Goal 17: Partnerships to enable combinations of statistics, mobile and internet data can provide better and real-time of today's connected real world.

Further insights on ongoing international projects using Big Data sources for SDG monitoring can be found in the United Nations Big Data Project Inventory portal. (examined on 9 October, 2020). Table 2 shows the reported projects by Big Data source and thematic area.

Although projects are sometimes speculative and aspirational, with some of them foreseeing the use of more sources, one can easily get an idea of the most important data sources in use, and their focus areas.

Big Data sources, such as web scraping, scanner data and mobile phone, almost monopolize the sources of information used in the inventory. The first and the third ones are also the main sources used in the ESCWA region, based on the responses to the survey questionnaire delivered by the organization to its member countries before the Regional Workshop on the Integration of Big Data and Geospatial Information for the Compilation of SDG Indicators in Arab Countries. As we speak, technology, information and communications technologies (ICT) and digitization offer new sources of big data of various Vs attributions.

Technology blinds, converges, and mergers produce even more volatile sources of data that can be utilized by NSOs, NSSs and the international community, including the SDGs.

Other main sources used in the Arab region include social media, which, in our inventory, are being used in six ongoing projects. Coming to the project area, the most scrutinized are prices, labour markets, mobility and tourism, whilst prices and environment/energy are the most analyzed in the ESCWA member countries.

Looking more closely at the outcomes of the questionnaire for the Arab countries, particularly at the component of Big Data, to which an entire section is dedicated, survey results indicate that procedures for verifying the potential of Big Data sources for statistical purposes, including the development of SDG indicators, exist in only five countries (36 per cent).

These countries have identified the main sources of Big Data used for the following: Internet, mobile phone, satellite or aerial photos, social media, road sensors, public transport use, and others such as credit cards, electricity consumption meter/electricity meter.

Statistical agencies in six countries (55 per cent) are involved in Big Data processing projects using various technologies and tools such as data visualization tools and software, GIS and Hadoop Clusters.

Only three of those countries (23 per cent) use Big Data sources to measure SDG indicators, namely the proportion of individuals who own a mobile phone by gender, the percentage of the population covered by the mobile phone network, by technology, and the proportion of individuals who use the Internet.

Big Data projects are being held in eight countries, in partnership with international organizations and technology partners.

The survey results show that only three (23 per cent) of the NSOs have developed new methods of calculating estimates or a methodological framework specifically related to the use of Big Data sources.

Challenges and obstacles that these agencies face, which prevent the use of Big Data sources in the production of official statistics, including SDG indicators, include an inadequate legal framework, limited access to data sets, and human resources not adequately skilled to access and manage Big Data sources (57 per cent), high costs of accessing that data (50 per cent), lack of technological tools, statisticians' view of Big Data (43 per cent), and difficulty in applying methods and methodologies (36 per cent).

As per the last UN Global Pulse Annual Report, there are some ongoing innovation projects worth mentioning, aimed at improving SDG understanding with the use of Big Data. It should be noted, however, that those projects are mainly aimed at analyzing realities in Asian and African countries (table 3). Brief descriptions of those projects are outlined below.

In one of such projects, UN Global Pulse worked with the United Nations Office of Internal Oversight Services on analyzing Twitter data to understand the effectiveness of social media campaigns, and whether they could be improved.

United Nations agencies working in Lesotho rolled out another initiative to better understand people's perceptions of the country's social, economic and environmental gains to help inform further work on the SDGs. The initiative used collective intelligence to model new forms of data collection (through perception surveys and a social media analysis) in order to build feedback mechanisms that could improve decision making and citizen reporting.

Another initiative developed an interactive visualization and analysis dashboard that uses data from the latest census of agriculture to identify smallholder farmers and to subsequently generate insights that can inform policies relevant to small and medium enterprises in the agriculture sector.

Table 2. Big Data Project Inventory

BD source	No. of projects	Project topic	No. of projects
Web scraping	24	Prices	33
Scanner	21	Labour market	11
Mobile phone/CDR	17	Mobility	10
Satellite imagery	8	Tourism	10
Social media	6	Environment/Energy	9
Road sensor	5	Transportation	8
Smart meter	5	Geo-spatial	5
Health records	3	IS/ICT	4
Credit cards	3	Agriculture	3
Ship identification	2	Vital stats./Civil registration	2

Table 3. Innovation Projects, UN Global Pulse Annual Report 2019

SDG	Theme	Data source	Project topic
All	Online perception and effective communication	• Twitter data, social media; • Perception surveys, social media.	• Gauging the Effectiveness of Social Media Advocacy Campaigns; • Supporting Country Teams to Understand Online Perceptions of the SDGs.
2	Food security	Agriculture census micro-data.	Mapping Smallholder Farmers in Indonesia to Inform Policies.
12,16,17	Societal impact of AI	Social media (UN speeches).	Understanding the Risks in AI-Generated Texts.
1,2,3,8,11,16	Disaster risk-reduction	• Mobile phone; • Mobility, population, economics, conflicts and others.	• Understanding Population Movement After the 2018 Central Sulawesi Disasters; • Using AI to Model Displacement in Somalia.
5,10	Women's safety and rights	• Social media (radio data); • Mobility data, interviews.	• End of Violence Against Women and Girls in Uganda; • After Dark: Encouraging Safe Transit for Women Travelling at Night.
3	Health and well-being	Social media (radio data).	Use of radio broadcasts to augment early detection of health risks.

Another project aggregates data on potential causes of displacement from internal and external sources, including information on conflict events and fatalities, wages and commodity prices, climate-related anomalies, and historical displacement flows. Different machine learning models for making predictions are tested, and results are presented in a dashboard that illustrates historical arrivals, alongside predictions for the three top-performing models. In a second iteration, the dashboard is extended through standardization of the collection of input data, extension of the prediction horizon form, and build-up of a second, internal dashboard that allows users to compare predictions across all models, explore the input data, and assess the performance of the algorithms.

In a gender-focused project, a radio content analysis tool is used and a real-time gender perceptions dashboard is created to unearth discussions and topics regarding sexual and gender-based violence across a specific country. Automated speech recognition tool uses artificial intelligence technology to pull recordings of radio conversations around particular themes and transcribe them. Once identified, the relevant snippets of information – in this case regarding violence against women and girls – are analyzed and visualized using the dashboard to allow authorities to access more timely and granular information.

In order to explore the utility of mining radio broadcasts for early signals of health risks, another project analyzed data from radio broadcasts using a speech-to-text technology developed to automatically transcribe radio talk into text. The study investigated key health-related signals based on specific keywords. Preliminary results showed potential correlations between health topics people discussed on the radio and recorded health metrics on the ground.

Those are just few examples on how Big Data might contribute to a better compilation of sustainable development indicators in a specific field and country or local area, thus providing policymakers with valuable insights that help them orient their interventions.

3. Remote Sensing and Goals 1 and 8

In this and the next sections, we focus on two main sources of Big Data that seem to be the most prominent in helping statisticians monitor and report on SDG indicators, namely remote sensing and sensors data. The main advantage of using remote sensing information is that, in general, remote sensing data are available for free and involve a wide spectrum of relevant applications. Sensors are generally mounted on satellites, which detect the energy reflected from forests, water, grass, soil, paved roads and built-up areas (figure 6).

Particularly, there is a growing literature using information on night lights (as, for example, that shown in figure 7) to proxy many important phenomena, starting from finding strong correlations between the intensity of night lights and a number of human-induced changes: i.e. in the economic area, Gross Domestic Product (GDP) and prices; in the social area, poverty rates; in the demographic area, population and migration flows; in the environmental area, emissions, pollution and land degradation; and in other areas, wars, smuggling, informal activities, tourism and urbanization.

Figure 6. How remote sensing concretely works

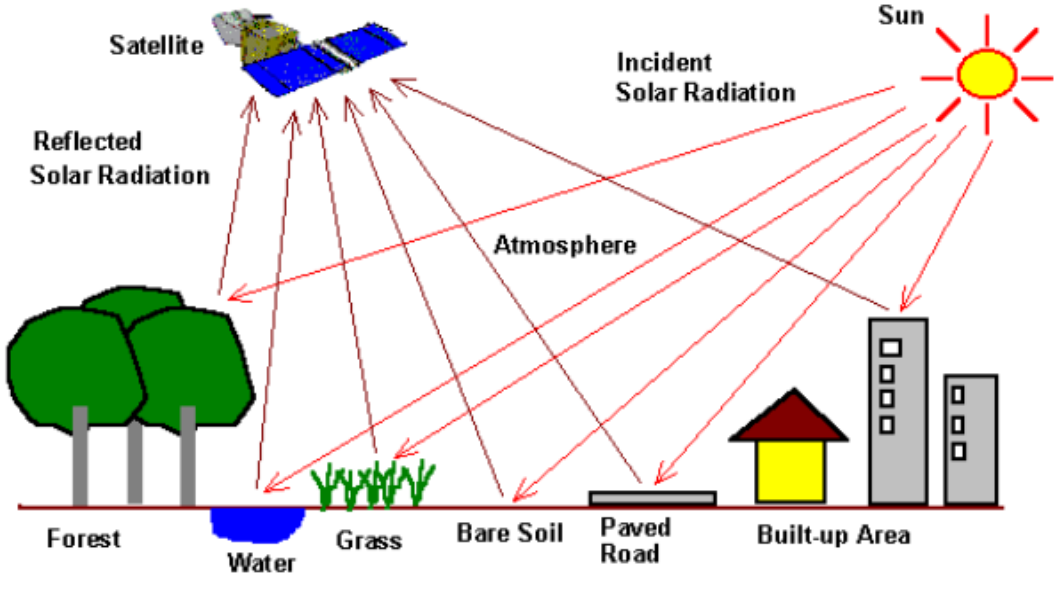


Figure 7. Lights during night as captured by a NASA satellite

Concerning Goal 1, on poverty, in a seminal paper Elvidge and others (2009) use LandScan population annual data and DMSP-OLS data of NASA (lights during night), both at one square kilometre resolution, to derive a Poverty Index given by $PI = \frac{Pop}{NL}$, in order to obtain a calibration between PI and official poverty rates drawn from the World Development Indicators of the World Bank. The estimated coefficients are then applied to the cells data of population and night-time lights to obtain maps of poverty at a finer geographical level.

Data availability on poverty at the national level is dramatically scarce worldwide, and the information at the disaggregated level by sex, geographic area, sector and income level is practically unavailable. Therefore, the motto of "leaving no one behind" appears as a mere daydream if it is not accompanied by specific actions to improve data quality and availability.

Apart from data availability, which reflects the fact that not all countries around the world conduct household surveys – the main statistics source used for poverty estimations – global estimates of poverty show quite well-known problems. These include: high data collection and processing costs; lack of timeliness in data; availability of data; different timing and frequency of data collection; uncertainty in the survey cycle at the country level; lack of inter-comparability of surveys among countries; and different impacts of measurement errors at the national level that, in turn, might depend on a number of critical issues.

Figure 8. Examples of estimation of poverty gaps at $5.50 per day for Latin American and Caribbean countries, using night lights

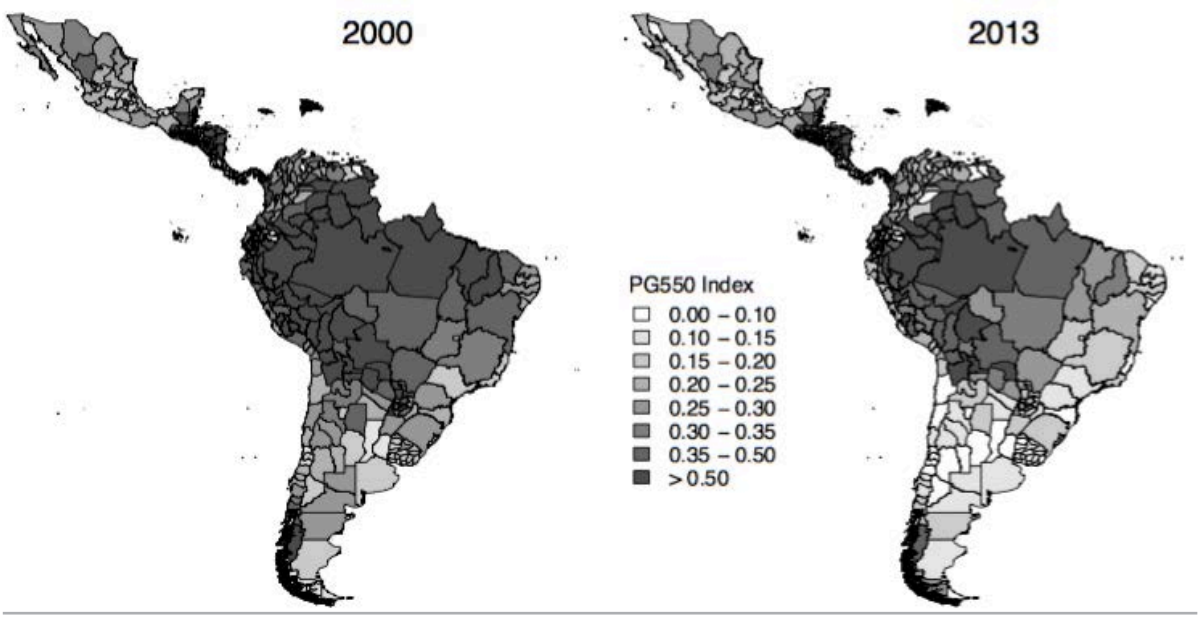
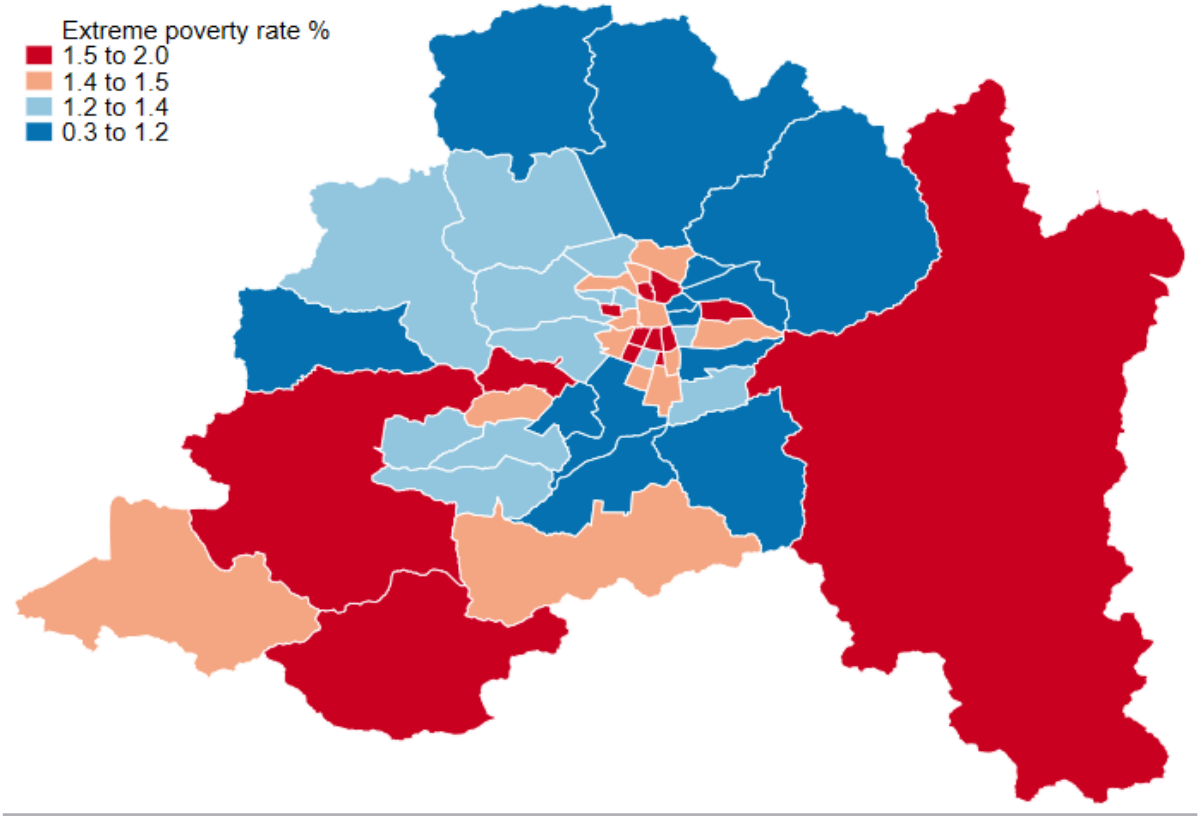


Figure 9. Examples of estimation of extreme poverty rate using night lights for the municipalities of Santiago, Chile, in 2015



When household surveys are available, the use of micro-data to produce reliable estimates at fine-scale spatial level through small area estimation techniques might represent an extraordinary undertaking due to the likely low reliability of the available information. A most challenging statistics issue is the lack of a homogeneous time series of subnational statistics on poverty, which should be of paramount importance in directing aid, policy interventions and sustainable development resources, especially when "territory matters".

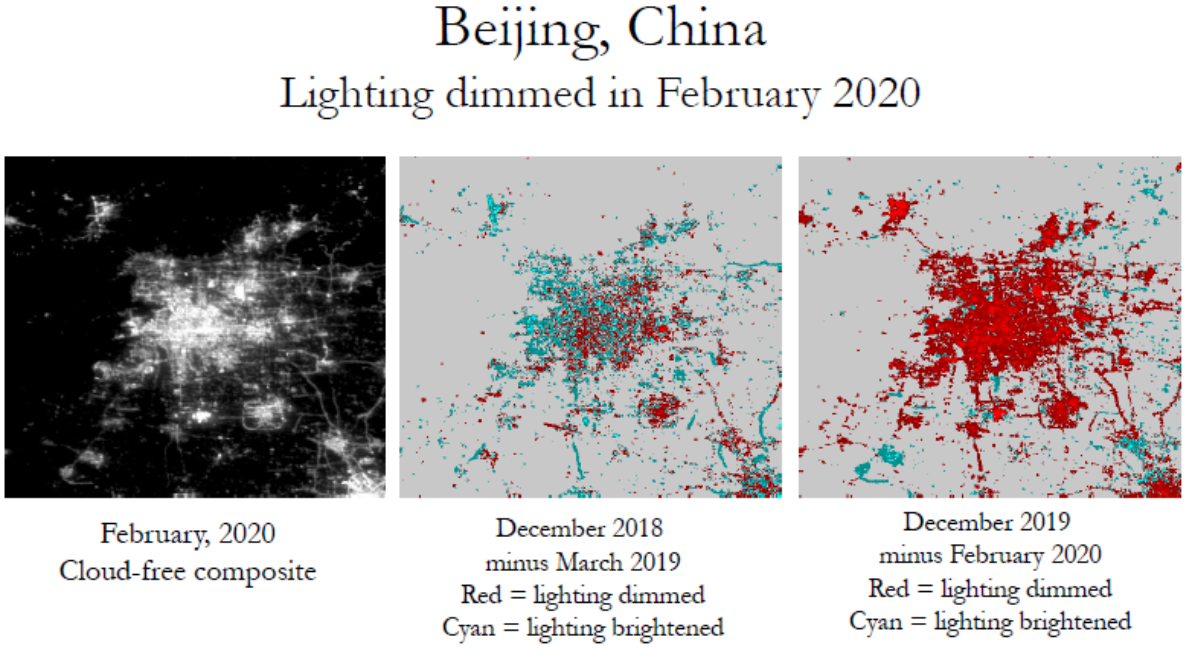
Following Elvidge and others (2009), Andreano and others (2020) and Cecchini and others (2020) use two different econometric frameworks – fractional panel and multinomial logit models – to obtain time series maps of estimates of poverty rates for SDG indicator 1.1.1 for Latin American and Caribbean countries using solely information on night lights and population, with a geographical and temporal detail not available in statistics officially released by NSOs of the region. Examples of the results of such estimates are shown in figure 8 and figure 9.

In another parallel application, using panel data models applied to night lights and population data, estimates of GDP and purchasing power parities are obtained in time series at fine geographical level for Eastern European and the Commonwealth of Independent States countries (Andreano and others, 2019).

Night lights might offer valuable insights and understanding of the socio-economic conditions and their changes in time of pandemic, such as for the recent COVID-19.

Figure 10 shows dimming and recovering night lights in Beijing, China, before, during and after the impact of the pandemic. It is clear that night lights might provide a clear overview of the impact of the events on lives and conditions in the affected area (Elvidge and others, 2020).

Figure 10. Night lights before, during and after the impact of the pandemic in Bejing, China

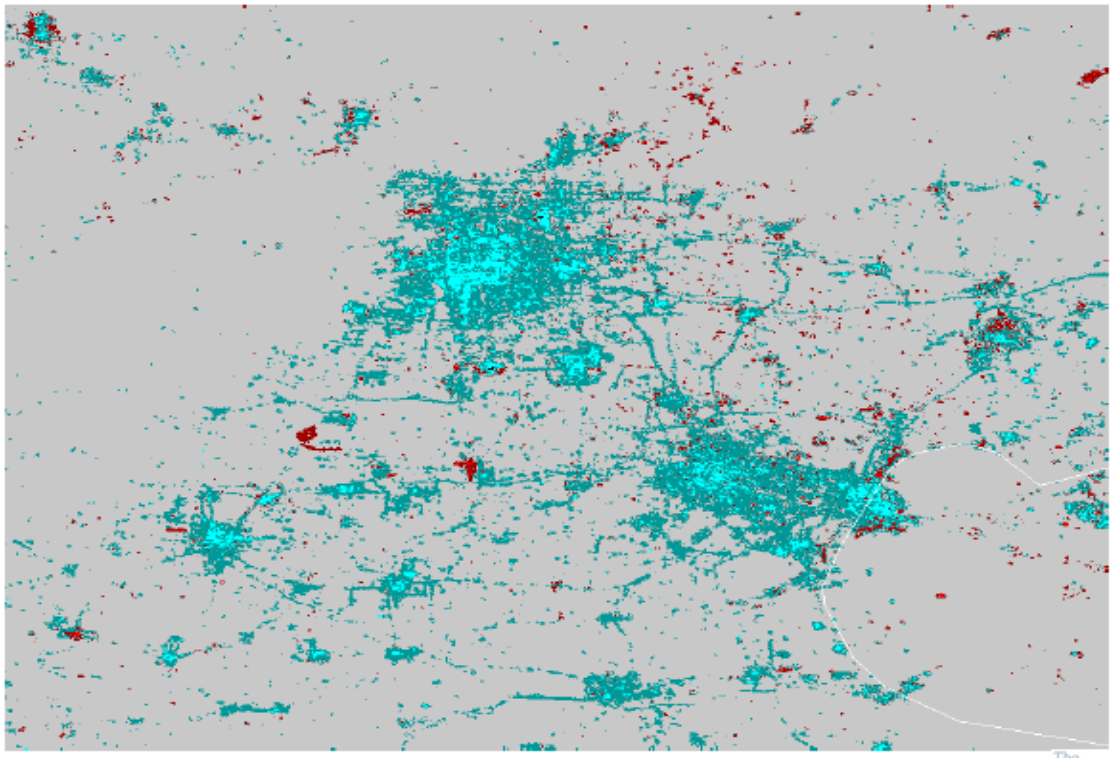


Beijing, China

February 2020 minus March 2020

Lighting has largely recovered!

Red = lighting dimmed. Cyan = lighting brightened.



4. Other Applications: Sensors

Sensors are becoming another important Big Data source for SDG monitoring, and numerous examples of concrete applications are nowadays available. Areas of main interest are agriculture, water, environment and nature, cities, transportation and homes. Here we will briefly mention two interesting cases aimed at specifically addressing data collection through sensors and estimation of specific SDG indicators.

Sensors in cities, transportations and homes are often described within the term "smart city", indicating actions aimed at improving today's city, both functionally and structurally, using ICT as an infrastructure. The basic component towards these improvements are sensors that are installed around the city to monitor urban flows in the widest sense and to detect any issues in the city life that need to be fixed or improved. This group comprises sensors at homes, which aim to improve the efficient use of resources as well as civil buildings safety.

An important example is the project "Improving transport planning through real time data analytics", carried out by the UN Global Pulse Lab Jakarta, Indonesia. Jakarta is well-known for its traffic jams, which, among other effects, prolong commutes and complicate the scheduling of public transport. The Smart City team within the Jakarta Government and Pulse Lab Jakarta are collaborating to explore real-time bus location data, service demand data, and real-time traffic information. The project aims to enhance transport planning and operational decision-making within the Jakarta Government through real-time data analytics, thus providing statistical evidence on SDG indicator 11.2.1 – proportion of population that has convenient access to public transport, by sex, age and persons with disabilities.

Sensors are also being increasingly used in rural areas, nature and water, including in developing countries. Sensors in remote areas can be seen as a step towards the goal of leaving no one behind as they provide data from areas for which information is lacking. This also includes disaster forecast as a priority area to enable early warning of, for example, flooding, landslides and avalanches.

An example, which deals specifically with SDG indicator 15.5.1 (forest area as a proportion of total land area), is the Wadi Drone Project implemented by students of New York University Abu Dhabi in the United Arab Emirates. The system leverages commercial drone technology and proprietary software for wildlife conservation and environmental protection. The Wadi Drone collects data in regions where deploying communications infrastructure would spoil the natural heritage or present a human risk to physically retrieving data. It does so by retrieving information from ground-based scientific measurement devices. In Wadi Wurayah National Park, the drone flies over mountains and through valleys to wirelessly download photographs taken by ground-based camera traps that automatically capture images of wildlife as they pass in front of their motion sensors. The Wadi Drone serves the conservation efforts of the Emirates Wildlife Society by both increasing the rate at which

photographic data of wildlife can be analyzed by experts and reducing the human risk associated with the current method of hiking to retrieve photos from remote camera traps. Wadi Drones further eliminate the need to employ a costly helicopter to reach camera traps during the summer months when high temperatures pose dangerous hiking conditions.

5. Challenges and Opportunities

From the catalogue of Big Data projects, and the discussion in sections 4 and 5, it is clear that Big Data can help in a profound manner in collecting information on a number of focus areas, including mobility, transport, tourism, prices, corruption and crime, energy consumption, population density, land use, well-being, cities and the labour market.

Today, analysis of Big Data is quite familiar to the private sector, with consumer analysis, personalized services, predictive exercises and tools being developed and used for marketing, advertising, forecasting and management. Similar techniques could be adopted to gain real-time insights into people's well-being and to target aid interventions to vulnerable groups.

Big Data can be used in conjunction with, or as a replacement for, traditional data sources to improve and enhance existing statistics. There is a growing interest in combining different Big Data sources (i.e. mobile data and earth observation) in order to exploit as much as possible potentialities, synergies and complementarities offered by various and heterogeneous information systems.

Big Data may also provide solutions for data gaps in the developing world, where traditional approaches to data collection have so far failed. There is a growing consensus that Big Data, which represent real transactions, may in some cases be better than survey data. In this respect, it has also been stressed that Big Data may provide more honest data, with greater veracity than traditional survey data.

Advantages of Big Data highlighted by literature include cost savings, increased timeliness, burden reduction, possibility to go deeper in granularity, sometimes greater accuracy and international comparability, higher variety and new time-series of indicators made available to data producers.

Big Data may also offer opportunities to re-think the role of official statistics and reposition them in view of a wider and more complete data ecosystem.

Many Big Data sources are supranational or global in scope. This globalized aspect of Big Data offers exciting, although strategically sensitive, opportunities to re-thinking national production models. For example, while switching from a national to a collaborative international production model is fascinating, it still poses many challenges. As statistical legislation and data protection are often weak in many developing countries, focusing on Big Data before addressing these fundamental issues might be dangerous.

Furthermore, many Big Data are proprietary, which implies legal impediments, access costs and confidentiality issues.

Another aspect that deserves some attention is the concentration of the ownership of some Big Data sources, especially those obtainable from digital activities, in a few hands. Having the ownership in other hands than NSOs/NSSs/IOs implies a reputational risk for official organizations, and the primacy of these

organizations as unique providers of data might be easily challenged, with official entities consequently exposed to greater vulnerability.

Generally, costs and investment on staff (human capital formation) are important, and organizations must also face rapid technology changes, the ones with which Big Data grow.

Scarce representativeness, low accuracy and weak meta-data are also worth mentioning as *cons* when dealing with Big Data. Many Big Data sources do not fulfil all data quality requirements that an NSO would like to have.

For example, Google Trend data are often cited as an important data source to be used for forecasting/now-casting purposes. However, those sources are not created by statisticians, or for statistical purposes. They simply represent a self-selected (non-probabilistic) sample, with generating mechanisms often unknown. Therefore, there is no guarantee that the data are representative, unless they cover the full population of interest, as it is the case of satellite remote sensing data (Andreano and others, 2019).

Volatility and instability inherent in Big Data sources might also pose important issues, whereas NSOs privilege stability of disseminated data, and users are uncomfortable with data that are subject to large revisions. Part of these challenges are cited by ESCWA countries in their answers to the questionnaire discussed above. Countries have identified, the following issues, in order of importance: (a) inadequate legal framework, and limited access to data; (b) human resources not skilled enough, and high costs of accessing the data; and (c) lack of technological tools, and statisticians' perception of Big Data.

Another important issue related to Big Data is cybersecurity. Big Data come in many cases as open source. Quite often, they are not designed with security as a primary function, and this can cause issues when information stores are sensitive or confidential, such as customer information, credit card numbers or contact details. All those issues imply, amongst others, that agencies using Big Data should dedicate great attention to the use of third-party applications and securing privileged user access and service level agreements. Security instruments might include encryption, centralized key management, user access control, intrusion detection and prevention, as well as the establishment of physical security systems.

6. Conclusions

Perhaps we are only at the beginning of the Big Data era: the "data revolution" is in part in the "data evolution" that Big Data actually represent. Big Data, if "tackled" properly, might offer some tempting openings for NSOs/NSSs/IOs, particularly improved timeliness and accuracy, greater granularity and disaggregation capabilities, and ability to fill in official data gaps for many SDG indicators by 2030.

This can be relevant, especially for statistically developing countries, and during times of exogenous adverse constraints (such as COVID-19 and natural and human-induced disasters).

Big Data offer a wide spectrum of alternatives, covering a number of key areas for official statistics in the economic, social and environmental domains. However, their use poses challenges of different nature to be carefully considered: ethical, legal, technical and reputational.

Quality considerations should always be at the top of the agenda of institutions tasked for providing official data, and "quality" includes, perhaps at first place from a user's perspective, the "availability" dimension.

Facing the availability requirement, NSOs and NSSs should undertake an honest exam about advantages and disadvantages of accompanying traditional data sources (such as official surveys and censuses) with the information possibly gathered from Big Data sources, in order to supplement shortcomings and failures of official statistics with the information obtainable from innovative sources of evidence.

References

C. D. Elvidge, P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri and E. Bright, "A global poverty map derived from satellite data" (Computers & Geosciences, Elsevier, 2009).

J. E. Steele, P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem and L. Bengtsson, "Mapping poverty using mobile phone and satellite data" (Journal of the Royal Society Interface, 2017).

M. S. Andreano, R. Benedetti, F. Piersimoni, P. Postiglione and G. Savio, "Sampling and modelling issues using big data in now-casting", in New Statistical Developments in Data Science (eds. Verde R., Ferrari F., Petrucci A. and Racioppi F.), Springer Verlag, 2019.

M. S. Andreano, R. Benedetti, F. Piersimoni, P. Postiglione and G. Savio, "Mapping GDP and PPPs at sub-national level through earth observation in Eastern Europe and CIS Countries", Voprosy Statistiki, Rosstat, 2019.

M. S. Andreano, R. Benedetti, F. Piersimoni and G. Savio, "Mapping poverty indices for Latin American and the Caribbean countries through satellite remote sensing" (Social Indicators Research, Springer, 2020).

S. Cecchini, G. Savio and V. Tromben, "Shedding light on territorial poverty in Chile", submitted to Regional Science Policy and Practice, Wiley, 2020.

C. D. Elvidge, F.-C. Hsu, T. Ghosh, and M. Zhizhin, "World tour of COVID-19 impacts on nighttime lights", Earth Observation Group, Payne Institute for Public Policy Colorado School of Mines, 21 April2020. Available at https://payneinstitute.mines.edu/wp-content/uploads/sites/149/2020/04/World-Tour-of-COVID-19-Impacts-on-Nighttime-Lights-3.pdf.

TechAmerica Foundation, "Demystifying Big Data: A practical guide to transforming the business of government" (Technical Report, 2012).

United Nations Economic and Social Council, "Big Data and modernization of statistical systems – Report of the Secretary-General", Forty-fifth session of the UN Statistical Commission, NY, 4-7 March 2014, doc. E/CN.3/2014/11. December. New York, pp. 1-16, 2013.

J. Manske, D. Sangokoya, G. Pestre, and E. Letouzé, "Opportunities and requirements for leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America", White Paper Series Data-Pop Alliance, pp. 1-71, 2016.

2100218