UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION

Inclusive and Sustainable Industrial Development Working Paper Series
WP 11 | 2017

BIG DATA – ITS RELEVANCE AND IMPACT ON INDUSTRIAL STATISTICS

DEPARTMENT OF POLICY, RESEARCH AND STATISTICS

WORKING PAPER 11/2017

Big Data – Its relevance and impact on industrial statistics

Shyam Upadhyaya
UNIDO

Petra Kynclova
UNIDO

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION

Vienna, 2017

Acknowledgment

The paper was prepared in the process of exploring alternate data sources of industrial statistics in the context of new role of UNIDO in monitoring global indicators of sustainable development goals. The paper was presented earlier internally to Division of Industrial Policy, External Relations and Field Representation and externally to the Twenty Sixth session of the Committee for Coordination of Statistical Activities held on 1-2 October 2015 in Bangkok. Authors express sincere appreciation for comments and feedbacks received during those presentations. Constructive comments on index numbers of industrial production provided by Katarina Peitl are greatly appreciated. The paper has been edited and compiled by Niki Rodousakis.

Table of Contents

List of Figures

List of Abbreviations

BPP	Billion Prices Project
CPI	Consumer Price Index
CPS	Cyber-Physical Systems
DDI	Data-Driven Innovation
ECB	European Central Bank
EU	European Union
GDP	Gross Domestic Product
GMCI	Global Manufacturing Competitiveness Index
GWG	Global Working Group
HLG	High-level Group for Partnership, Coordination and Capacity-Building for statistics for the 2030 Agenda for Sustainable Development
HLG-MOS	High-Level Group for the Modernisation of Official Statistics
ICHEC	Irish Centre for High-End Computing
ICT	Information and Communications Technology
IT	Information Technology
IIP	Index numbers of Industrial Production
IoT	Internet of Things
MDGs	Millennium Development Goals
NGO	Non-governmental Organization
NSO	National Statistical Office
OECD	Organization of Economic Cooperation and Development
PMI™	Purchasing Managers' Index™
R&D	Research and Development
SDGs	Sustainable Development Goals
SPF	Survey of Professional Forecasters
UN	United Nations
UNDESA	United Nations Department of Economic and Social Affairs
UNECE	United Nations Economic Commission for Europe
UNIDO	United Nations Industrial Development Organization
UNSD	United Nations Statistics Division

1. Introduction

Official statistics are the main source of highly reliable information about society. They are mostly compiled on the basis of internationally recommended methodological standards to provide a complete and meaningful real-world analysis. Yet society is rapidly changing and new technologies emerge continuously. We are witnessing an expansion of new types and sources of data that are more widely available. Larger volumes of data have triggered higher demand for data storage facilities with new technological sophistication. We are living in the era of Big Data, in which information about the world around us is collected and disseminated every single moment.

These changes have created opportunities for massive exchange of information. At the same time, Big Data poses a challenge to society as a whole. There is concern about the confidentiality of personal data and re-use of these data by companies for commercial purposes. The amount of information shared by users on social networks is rising without sufficient attention being paid to the privacy of citizens. There is also concern about the quality of data. Internet generated data do not necessarily heed traditional statistical concepts such as scope and the coverage of data, representativeness and level of significance. Data are not collected using specially designated statistical tools but are instead generated by internet use for normal day-to-day activities. Statisticians are thus faced with the question what precisely this change implies for the exclusive role of national statistical offices (NSOs) as the main custodians of official statistics.

Before the era of Big Data, only few alternative data sources were available, and official statistics represented the single most valuable source of data for researchers and policymakers. Emerging data sources represent a new challenge in the field of statistics but it might also be necessary to establish an additional system of legal and methodological regulations for their use. The United Nations has developed its own principles with reference to Big Data (United Nations 2013) and professionalism in terms of data collection, processing, storage and presentation together with a respect for confidentiality of the data provided by respondents.

Big Data may represent an available low cost source of data to complement traditional surveys, especially with regard to short-term predictions. Using alternative sources may inspire researchers to develop new methodologies for official statistics and to promote an incremental and system-wide integration of statistical and IT systems together with NSOs and international agencies. Moreover, to use Big Data in practice requires the engagement of “data scientists”, i.e.

researchers with an analytical mind and programming skills, to extract the desired information from the large volume of external data and to collaborate with traditionally trained statisticians.

The role of international organizations in official statistics consists of formulating international statistical standards, introducing best practice methodologies and providing technical assistance to NSOs in order to enhance global monitoring. Hence, an emerging volume of alternative data sources is of relevance for international organizations as well. In September 2015, the United Nations General Assembly adopted the 2030 Agenda for Sustainable Development (United Nations 2015). *Distinguishing Big Data risks and opportunities, the UN Data Revolution – A World That Counts* concept was introduced to monitor the progress of the sustainable development goals (SDGs) and to produce better coverage and high quality data (United Nations 2014). Data revolution encompasses the technology of data collection as well as production and dissemination by (combining new data sources and technologies with traditional data.

This paper intends to provide an overview of possible ways to implement Big Data sources into UNIDO's industrial statistics programme. It does not intend to describe the extent of usage of big data and does not make any value judgment due to lack of sufficient information. Some of the main challenges for the future work necessary to build the groundwork to foster Big Data as a valuable and reliable component of official statistics are addressed in this paper.

2. What is Big Data?

The modern world is rapidly changing with the help of emerging technologies which are producing an avalanche of data. New tools and applications are being used for data analysis, data capturing, visualization, storage and sharing. Some data are produced by sensors of the increasing number of electronic devices being used by individuals or are generated on the web, sometimes without people even noticing (information about our activities is often being collected without us being aware of it). Big Data has become an issue of particular interest for many researchers and private companies trying to define and capture it.

The term "Big Data" or "Big Data Analytics" is defined by Gartner[1] as follows:

> *"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.*"

[1] http://www.gartner.com/it-glossary/big-data/

In other words, Big Data represents data sets collected from various sources and produced in enormous amounts and high frequency. These data are so large and/or complex that traditional data processing tools and methods cannot adequately process them. Big Data is generally characterized as data of increasing volume, velocity and variety; known also as the 3 Vs (Laney 2001).

- **Volume** refers to the exponentially increasing magnitude of data. The size of Big Data is expressed in terabytes and petabytes that are driven by emerging new data collection tools, such as social media, mobile applications or sensors. The amount of this collected information requires improved storage systems and data management technologies.

- **Variety** refers to Big Data's heterogeneous structure and the complexity of its formats. Big Data can appear in a structured pattern, such as databases, as well as in semi-structured or unstructured documents, images, video, emails, coordinates, etc. Analytical tools are being continuously improved and they are now able to deal with very heterogeneously structured data. Such data do, however, consume more time and storage.

- **Velocity** refers to the speed at which Big Data are created, spread, stored, analysed and visualized. The data are virtually produced in real-time, e.g. internet logs or location tracking by mobile phones. This higher frequency of data creates pressure to increase the current speed of data processing.

The concept of the 3 Vs has frequently been expanded by additional dimensions to elaborate the definition of Big Data.

- **Variability** and **Complexity** were introduced by SAS[2]. Variability refers to inconsistent data flows across time, emphasizing the fact that the majority of Big Data appears in periodic peaks. The complexity attribute is related to multiple data sources which makes it difficult to link data across systems. The essential challenge for SAS is to connect and match data from various sources.

- **Veracity** was listed by IBM as the fourth V of Big Data[3] and entails trusting the data's accuracy. The data might be biased, contain noise and other abnormalities, or even be out of date. The challenge consists in deciding whether the stored data are relevant for a meaningful data analysis.

[2] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[3] https://www-01.ibm.com/software/data/bigdata/

- **Value** was introduced by Oracle[4] to complete the list of Vs. Big Data can contain significant economic value. The challenge consists in identifying good information, in other words, determining whether such value can be obtained by extracting and analysing Big Data.

Several other Vs, such as **Validity, Variability,** and **Visualization**, have been proposed by other sources, but mostly overlap with the already mentioned dimensions of Big Data.

Big Data can also be presented as a wide ecosystem covering not only flows of data, but also all related institutions which produce and use them (Data-Pop Alliance 2015). The produced data are often largely unstructured, i.e. they do not match any predefined data models or any database structure.

The rise of Big Data has had a significant impact on businesses which have now started investing in data innovations to explore and better understand the market and customer behaviour. Big Data do not only entail advantages for the company, they are also used to develop new and improve products based on customers' needs (Marr, 4 Ways Big Data Will Change Every Business 2015). New data sources and developing technologies represent incentives for the fourth industrial revolution. The German government introduced a high-tech strategic project called "Industrie 4.0" promoting the computerization of manufacturing[5]. The trend of automation in manufacturing supported by Big Data technologies, including cyber-physical systems (CPS)[6], the Internet of Things and cloud computing are the main focus of the project (Marr 2016).

The *Industrie 4.0* has introduced another Big Data Analytics concept used in manufacturing based on 6 Cs (Lee, et al. 2013):

- Connection (sensor and networks)
- Cloud (computing and data on demand)
- Cyber (model & memory)
- Content/context (meaning and correlation)

[4] http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf
[5] http://www.plattform-i40.de/I40/Navigation/EN/Industrie40/WhatIsIndustrie40/what-is-industrie40.html
[6] Wikipedia defines a CPS as "a mechanism controlled or monitored by computer-based algorithms, tightly integrated with the internet and its users. In cyber-physical systems, physical and software components are deeply intertwined, each operating on different spatial and temporal scales, exhibiting multiple and distinct behavioural modalities, and interacting with each other in a myriad of ways that change with context. Examples of CPS include smart grid, autonomous automobile systems, medical monitoring, process control systems, robotics systems, and automatic pilot avionics."

- Community (sharing & collaboration)
- Customization (personalization and value).

The purpose of integrating cyber-physical systems and cloud computing is to achieve more effective and resource-saving processes in manufacturing.

3. Types of Big Data

Big Data represent a complicated structure of various datasets combining structured and unstructured features. Two different classifications of Big Data exist in official statistics, classifications by type (UNECE 2017) and classifications by size (Dornik and Hendry 2015).

3.1. Classification by type

Big Data are derived from various sources. The internationally agreed classification system was developed by UNECE's Task Team on Big Data in June 2013; the classification system is summarized in Appendix I[7]. The first classification encompasses data that is derived from **social networks** that provide human-generated data reflecting various human experiences. This information is largely unstructured and often unregulated. Such data come from sources such as Facebook, Twitter, blogs or YouTube, stored in a digital form on personal computers as well as on social networks and are characterized by high frequency. However, such data are hardly ever representative, nor are they based on a common set of definitions.

The second classification covers **traditional business systems** including both administrative data types, data collected by public agencies, medical care, social security or tax records, and data produced by businesses, such as data on customers, sales, costs, profits, assets and liability. Such process-mediated data are highly structured and usually include reference tables, relationships and metadata to set the context. The challenge of processing this type of data relates to gaining access to proprietary and confidential data, the representativeness of the given sample, and inconsistencies in methodologies and definitions across different accounting systems. Such administrative and business data have been the most important source of Big Data for industrial statistics for many years and are likely to remain the most important input for such data in the future.

Yet another category is defined as the **Internet of Things** (IoT) established by machine-generated data such as various sensor data (weather, pollution or traffic), mobile phone location records and satellites. This type of data derives from the significant growth of sensors and

[7] http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

machines measuring and recording events and conditions in the physical world. Many businesses specifically build their know-how on processing sensor-recorded data. The data are well-structured; however, they appear in high volumes and sizes which require proper storage, high-level computer processing techniques and well-trained IT specialists. The challenge this type of data poses is its limited use for general purposes. The most fruitful uses are in the travel, tourism, traffic, migration and environmental industries.

3.2. Classification by size

Big Data can also be classified based on their numerical features, such as the size of data sets. Three types of Big Data shapes based on size can be distinguished (Dornik and Hendry 2015):

- Tall
- Fat
- Huge.

We can define the dimension of a data set, the number of variables *D*, and the number of observations *N*. **"Tall"** data set is represented by $N >> D$, that is, it does not entail all too many variables, but a high number of observations, namely a cross-section of observations of many individuals at a single point in time, for example, financial transactions or search queries.

"Fat" data sets contain a high number of variables, but not so many observations, $D >> N$. This type of data is also described as high-dimensional data, i.e. the dimension may in fact be larger than the sample size. It is thus represented by a time series of observations on many variables. In practice, high-dimensional data are increasing exponentially, as are numerous statistical and data science methods. The most frequently used techniques are machine learning methods, dimension reduction techniques or Bayesian methods.

Big Data sets are **"Huge"** when they entail many variables and many observations, i.e. a very large number of *D* and *N*. This type of data is also known in econometrics as multidimensional panel data.

4. UN initiatives in data revolution

With the penetration of new technologies in information systems, interest in alternative data sources has increased rapidly. The data revolution has the potential to transform society. The UN Secretary General's Independent Expert Advisory Group on the Data Revolution for Sustainable Development launched a report, *A World That Counts: Mobilising the Data*

Revolution for Sustainable Development, which describes the data revolution and highlights its potential global challenges, opportunities and risks.

New technology has played a crucial role in the expansion of data volume worldwide. First, internet use has created a new data source on human activities. Second, data from traditional sources have become more easily accessible. The results of various surveys are being placed online to provide open data for the public. Data are becoming more easily available and more detailed due to various types of disaggregation. Moreover, they are disseminated much more frequently than ever before.

These developments have convinced policymakers that the data revolution represents an opportunity to improve the data that are essential for sustainable development and for decision-making to that end. The UN report defines data revolution as: "An explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the "internet of things", and from other sources, such as qualitative data, citizen-generated data and perceptions data". Additionally, it describes the data revolution as "a growing demand for data from all parts of society"[8].

It has, however, also been emphasized that the data revolution poses many new challenges and risks that must be addressed and prevented in order to ensure that the fundamental elements of human rights are not violated: privacy, respect for minorities or data sovereignty. The majority of new data is collected in a passive manner, namely as "digital footprints" on social networks or by sensor-enabled objects while individuals might not even notice the volume of data they are leaving behind. The many constraints to access to information include language, poverty, lack of education, lack of technology and infrastructure, remoteness and discrimination. The gap between the countries that have access to data and those that do not further widens the inequality between them. Industrialized economies can take advantage of the new data sources more than developing countries can, where the public sector accounts for most of the information due to the weaker private sector. The governments of developing countries will have to establish and enforce legal frameworks to make use of Big Data and to simultaneously ensure data privacy and security. NSOs can play key role in defining the relevance of Big Data by maintaining data quality and independence and fostering a balance between public and private interests. In addition, promoting data collaboration between public institutions and the private sector enables government institutions to be better prepared to adapt new data sources.

[8] United Nations, A World That Counts: Mobilizing the Data Revolution for Sustainable Development, 2014, 6.

Effective adaptation of new technologies in the field of data production and dissemination support the achievement of sustainable development. Moreover, a revamped statistical system that is capable of using Big Data can exploit the advantages of Big Data to monitor the progress made towards achieving sustainable development goals.

The objective to improve the process of data monitoring and accountability was initiated with the introduction of the Millennium Development Goals (MDGs) followed by an urgent call for implementing new data sources to monitor achievement of the Sustainable Development Goals (SDGs). To ensure that the data revolution advances sustainable development, concrete principles and standards[9] need to be developed and specified in guidelines. The following recommendations on key principles for the data revolution were introduced in *A World That Counts* (United Nations 2014). The main goals address increasing data quality, ensuring data transparency and safeguarding human rights protection.

1. **Data quality and integrity**

   To ensure high data quality and prevent results that may be misleading, new standards for official statistics need to be adopted following the UN Fundamental Principles of Official Statistics and the work of independent third parties.

2. **Data disaggregation**

   The commitment to "leave no one behind" has been an essential feature of all discussions related to the 2030 agenda and the SDGs. Therefore, data should be collected for various variables and disaggregated to ensure better coverage and cross-section comparability. The disaggregation should be relevant to the given programme, policy or other issue under consideration, with respect for individual privacy. .

3. **Data timeliness**

   Data quality also depends on data timeliness. The standards of data collection and dissemination should be developed to safeguard consistent time series of high quality and timely data from national statistical offices, international agencies, as well as from private Big Data sources. The data cycle should correspond to the decision cycle.[10]

[9] United Nations, A World That Counts: Mobilizing the Data Revolution for Sustainable Development, 2014.
[10] United Nations, A World That Counts: Mobilizing the Data Revolution for Sustainable Development, 2014.

4. Data transparency and openness

All publicly funded data sets or data related to public interests should be open to the public, also including data produced by the private sector. A legal framework for data use should be introduced as well, which states clear rules for commercial and non-commercial applications. Data should be provided in a machine-readable, standard format to be easily processed by any computer software. An analysis of the data should be accompanied by corresponding findings and include a description of the methodology and tools used.

5. Data usability and curation

Every single citizen should have access to data which must be easy to understand for the majority of people. Thus, all data should be made available on a user-friendly platform to deliver the data to a broader public, particularly to non-technical users, and to enable all data users to provide feedback and improve the current status of all provided data.

6. Data protection and privacy

As more data is becoming available, particularly from alternative or even unknown data sources, new threats and risks need to be addressed to safeguard human rights and to protect privacy. Legal frameworks must be developed to regulate data practices such as data mining, use and re-use of data for other purposes, data transfer and dissemination. Moreover, citizens' rights and data producers need to be protected from government or non-government demands and attacks by hackers. The establishment of such a robust legal system should also enhance the rights to freedom of expression of individuals who correctly provide, collect and analyse data, in order to protect them from recrimination.

7. Data governance and independence

National statistical offices play a key role in the production of national data and are responsible for independent monitoring, especially with regard to monitoring SDG indicators. Their autonomous role needs to be strengthened to ensure their independence.

8. **Data resources and capacity**

   The capability of national statistical offices to produce high quality official statistics requires huge investments in national capacity building. The increasing influence of Big Data requires investments in capacity for data science. The data revolution for sustainable and inclusive development is demanding increasing resources and support in developing countries from international agencies to produce transparent data that is compiled in line with all agreed standards.

9. **Data rights**

   All legal or regulatory mechanisms and programmes related to the data revolution for sustainable development should respect the fundamental rights of citizens, including the right to identity, the right to privacy, freedom of expression and equality.

The strengthening of national capacities is a fundamental driver of the data revolution. Sizeable and continuous investments in technology and innovation are essential for increasing literacy, access to information and the use of ICT. Improving data quality, coverage and timeliness for SDG monitoring should become a key component of the data revolution by identifying available and missing data. Together with the existing MDG monitoring scheme, new methodologies and analytical tools need to be developed to upgrade the availability of measured indicators. Moreover, other stakeholders such as the private sector, NGOs, the media or academia should be invited to participate in this global effort. New stakeholders could introduce an innovative approach to the data revolution to better forecast long-term trends and identify critical research gaps. Global cooperation and coordination should be managed under the governance and leadership of the United Nations to mobilize initiatives, coordinate actions and share experiences, particularly in the case of SDG monitoring.

The United Nations Statistical Commission took a leading role at its 45th session in 2014 by establishing the Global Working Group (GWG) on Big Data for Official Statistics. The objectives of the global working group cover vision, direction and coordination of a global programme on Big Data in official statistics, including investigating potential benefits and challenges of its practical use for the measurements of indicators of the 2030 Agenda for Sustainable Development (Statistical Commission 2016).

5. Big Data in official statistics

Big Data has proven to have a huge impact in the private sector. It has fostered the development of new tools for analysing a large amount of structured and unstructured data, to establish a new research field known as data science, and to adjust data storage capabilities. The Big Data industry has grown tremendously with the support of the private sector. However, the impact of Big Data on official statistics, i.e. on the institutions whose core business consists of producing and analysing statistical data, has yet to be assessed. In this section, we describe the role of national statistical offices (NSOs) and many other international or non-governmental agencies that play a key role in providing essential data on society. NSOs and international or non-governmental agencies undertake efforts to compile data in compliance with the internationally recommended methodological standards. To date, official statistics account for the majority of data products and represent a reliable data source for scientific research, education and evidence-based policymaking.

5.1. Role of international organizations

The role of international organizations in official statistics consists of formulating international statistical standards, introducing best practice methodologies and providing technical assistance to NSOs to assist them in monitoring sustainable development. Hence, an emerging volume of alternative data sources is also of relevance for international organizations.

At a high-level Seminar on Streamlining Statistical Production and Services in 2012, the participants requested "a document explaining the issues surrounding the use of Big Data in the official statistics community"[11]. They called for this document to have a strategic focus, aimed at the heads and senior managers of statistical organizations (UNECE 2013).

The first step towards using the power of Big Data to support the production of official statistics were made in 2014, when the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) launched a project to create a "Sandbox", a web-based collaborative environment, hosted in Ireland by ICHEC (Irish Centre for High-End Computing). The main objective of this project is for experts from national and international statistical organizations to come together to share their ideas and experiences on the use of Big Data for official statistics. The group has also formulated types of Big Data and created a Big Data inventory[12] to centralize pilot projects of national and international organizations related to this topic.

[11] UNECE, What does Big Data mean for official statistics?, 2013.
[12] http://www1.unece.org/stat/platform/display/BDI/UNECE+Big+Data+Inventory+Home

At the 45th session of the Statistical Commission, the Global Working Group (GWG) on Big Data for Official Statistics was created to explore the potential benefits and challenges of Big Data with a special focus on monitoring and reporting the sustainable development goals[13]. The global working group created eight task teams on advocacy and communication, Big Data and Sustainable Development Goals, access and partnerships, training, skills and capacity building, cross-cutting issues, mobile phone data, satellite imagery data and social media data.

The global working group conducted a global survey on Big Data in official statistics approved by the Statistical Commission. The primary goal of the survey was to examine the practical experiences of national statistical offices with Big Data, and to identify all of their concerns in order to recommend future steps for the implementation of Big Data in the modernization of statistical production. The survey was conducted from June to August 2015 containing questions on the management of Big Data, advocacy, communication, link to the Sustainable Development Goals, access, privacy, confidentiality, skills, training and concrete experience with the use of Big Data (Statistical Commission 2016).

About a half of the national statistical offices reported at least one Big Data-related project, mostly on price statistics (scanner data), transport and labour statistics (web-scraping data), tourism and population (mobile phone data) and agriculture (satellite imagery data). Among the main reasons for using Big Data are: they can be collected faster and are produced in a timelier fashion, and they reduce the respondents' burden due to the modernization of the statistical production process. For developing countries, Big Data sources also carry a significant benefit in terms of monitoring the Sustainable Development Goals. The survey of NSOs indicated that both developing and developed countries continue to rely on traditional statistical methods in the processing and analysis of Big Data sources. Moreover, national statistical offices expressed a need for skilled methodologists on Big Data issues, recommended the development of a quality framework, estimation methods and data access issues. Based on the global survey and panel discussions on Big Data in official statistics, the three following priorities were formulated: 1) Big Data methodology and estimation, 2) training and capacity building, and 3) data access and partnerships.

The OECD has presented the data revolution from another perspective, namely as having the potential to enhance resource efficiency and productivity, economic competitiveness and social well-being as it begins transforming all industries in the economy, including low-tech industries and manufacturing (OECD 2015). The OECD has introduced Data-Driven Innovation (DDI) as

[13] http://unstats.un.org/bigdata/

a significant improvement of existing or development of new products, processes, organizational methods and markets emerging from this phenomenon (OECD 2015). DDI presents the data as the new R&D for 21st century innovation systems. However, the report does not discuss the potential for incorporating Big Data into official statistics, and Big Data sources are still only considered complementary information to official statistics production (Reimsbach-Kounatze 2015).

Growing interest in the power of Big Data to promote development has also compelled the World Bank to present a pilot report exploring the potential of Big Data for development in Central American countries (The World Bank 2014). The report does not consider Big Data to be a panacea and refers to the increasing probability of biased data, spurious correlations and thus misleading conclusions and interpretations. Several case studies demonstrate the challenges associated with Big Data and considerations that need to be taken into account to promote a more in-depth discussion of these issues.

In addition to the above mentioned initiatives, the United Nations Secretary-General launched the United Nations Global Pulse in 2009, as a flagship innovation initiative on Big Data[14]. Global Pulse aims to link together experts from UN agencies, governments, academia and the private sector to explore and develop Big Data innovation programmes for sustainable development and humanitarian action. The programmes focus particularly on issues such as monitoring, evaluation and privacy protection, covering areas such as food security, agriculture, employment, infectious disease, urbanization, disaster response and others. They aim to find and access relevant sources of Big Data and learn from existing expertise in data science to develop appropriate and functional data analytics tools.

Increasing demand for collaboration between various professional groups resulted in the first United Nations World Data Forum hosted in South Africa in January 2017[15]. The forum was organized with the guidance of the UN Statistical Commission and the support of the UN Statistics Division of the UN Department of Economic and Social Affairs (DESA) and the High-level Group for Partnership, Coordination and Capacity-Building for Statistics for the 2030 Agenda for Sustainable Development (HLG). The urgent call to modernize the data production process to achieve SDGs was recognized and followed by the launch of the Global

[14] http://www.unglobalpulse.org/
[15] http://undataforum.org

Action Plan for Sustainable Development Data[16] to evaluate, build and strengthen NSO capacity (adopted by the UN Statistical Commission at its 48th Session in March 2017).

5.2. Role of national statistical offices

National statistical offices (NSOs) are the leading statistical agencies in a country. They collect, compile and publish national data using the country's statistical programme. They are committed to protecting the confidentiality of all collected information. The increasing amount and variety of data probably represent the biggest challenge since national statistical offices were first established. NSOs remain responsible for providing the foundation for good decision making by governments, businesses, households and community members, in consideration of ethical and legal commitments towards individuals linked to the data. The key focus of Big Data should be observable trends or patterns, not personal data about individuals. How Big Data sources are incorporated into official statistics may differ from country to country and depends on data availability, technological capacity and the level of data science skills of NSO staff members.

What NSOs should prioritize is access to Big Data sources and successful collaborations with public and private organizations. NSOs should cooperate rather than compete with the private sector to be able to more effectively incorporate Big Data in official statistics. However, NSOs should remain impartial and independent in order to continue building and retaining public trust.

The most promising collaboration is that between NSOs and researchers from academia. Such collaboration could bring new insights into the undiscovered potentials of Big Data sources for traditional official statistics. Research on the development of new methods, algorithms, systems and even software tools requires fruitful and constructive feedback from experienced official statisticians to uphold the key principles of official statistics.

UN Global Pulse in collaboration with UNDP has provided step-by-step guidance for development practitioners to leverage new sources of data. The report introduces data innovation and advocates the use of new or non-traditional data sources and methods to gain a more nuanced understanding of development challenges (UNDP, UN Global Pulse 2016). One of the key statements of the report is that a development practitioner does not need to necessarily be a data scientist to launch a data innovation project.

[16] http://undataforum.org/WorldDataForum/wp-content/uploads/2017/01/Cape-Town-Action-Plan-For-Data-Jan2017.pdf

5.3. Evaluating and using Big Data in official statistics

Big Data is a low-cost data source, especially for short-term forecasts. The opportunity to use Big Data as an alternative source should be an incentive for researchers to collaborate with official statisticians in order to develop new methodologies combining both Big Data and official statistics, and to promote an incremental and system-wide integration of statistical and IT systems together with NSOs and international agencies. Moreover, to use Big Data in practice requires the engagement of data scientists and researchers with an analytical mind and programming skills, to extract the desired information from the large volume of external data and to collaborate with traditionally trained statisticians.

There are three possible ways to use Big Data for the production of official statistics (Economic Commission for Europe 2014):

a) Fully replace statistical sources based on common definitions and classifications;

b) Partially replace statistical sources and supplement the information by means of record linkage, matching or other procedures;

c) Provide completely new statistical data that may complement already available statistical information.

Since no updated statistical system of recommendations and methods for Big Data has been established, businesses producing statistics opt for the third method c). They produce their own statistical data, while not having to cope with problems of harmonizing different structures of data based on statistical classifications, linking data and other corresponding procedures. The first two methods may reduce the costs related to conducting a survey, but this might be expensive in terms of time and other resources such as well-trained data scientists, improvement of technological capacity, etc. in the long term.

The most widely discussed question in terms of Big Data use in official statistics is whether it represents an opportunity or poses a threat to official statistics. Adopting Big Data in official statistics entails numerous challenges which are linked to particular risks and opportunities. The challenges are of a legislative, privacy, financial, management, methodological and technological nature (UNECE 2013).

5.3.1. Legislative and privacy challenges

Every single moment, individuals around the globe leave various digital footprints, sometimes without even noticing it. Such data may represent very useful information for further analysis.

Many people have started voicing their concerns about data privacy and the re-use of data by companies for commercial purposes. Others do not mind as long as the provided services are free of charge. It is unclear who the actual owner of the data is – the person who is the subject of the given information, the person or organization collecting the data, the person compiling or analysing the data, or society itself. Moreover, the development of new social phenomena evolves at a much faster speed than legislation on the issue. It is therefore particularly challenging to create a legal framework that covers ownership, stewardship, copyright or any further re-use of the data. Even if the data can be used in a legal way, it does not necessarily ensure careful and proper treatment of data. This ethical aspect is the most sensitive issue as regards Big Data use.

The legislative challenges are closely linked with challenges addressing privacy concerns. Privacy is generally defined as the right of individuals to control or influence what information related to them may be disclosed (UNECE 2013). There are two approaches to privacy related to Big Data: access to private data and its subsequent use, and how to ensure data privacy and confidentiality by avoiding a misuse of data.

Access to data differs from country to country; some countries provide open access to government and non-government organizations, others only provide data from public authorities. The public perception of NSOs is also of high relevance. It is crucial for NSOs to stay in line with public opinion, because credibility and public trust are extremely important. In addition, a diversion from official data sources could result in loss of transparency, which may have an impact on democracy.

5.3.1. Management challenges

The biggest management and financial challenge relates to human resources. Although Big Data are easy to obtain, their analysis can be very time consuming and requires specially trained staff, known as data scientists. Data science, sometimes also called data-driven science, is an interdisciplinary field aiming to mine the relevant information from data in various forms, either in a structured or unstructured way. It can be described as a parallel continuation of already established analytical data tools such as statistics, data mining, predictive analytics, machine learning and many others, depending on the objective of the analysis.

The most important aspect in this regard is the approach taken by the trained data scientists. They are expected to have an open and analytical mind set, together with high level IT skills to be able to find, manage, extract, analyse, visualize and interpret Big Data sources. Such trained staff members equipped with high level skills will naturally have an influence on costs.

The era of rapidly emerging data sources demands willingness of all participating organizations to collaborate to obtain more precise and complex information about our society. Several parties are considered essential for sharing knowledge and experiences. They include Big Data providers who have full access to data. Establishing a proper relationship with them could be complicated by many factors, such as data privacy or ownership. Many commercial partners exist as well, such as Google and Facebook, which already have a large amount of knowledge about Big Data use and who could become very useful advisors to NSOs by sharing their experiences. Although such partners could provide extensive knowledge on IT services such as storage, cloud expertise, security, etc., collaboration may imply large financial injections (Struijs, Braaksma and Daas 2014).

The most promising collaboration is that with academia. This partnership has a history and has been very beneficial for both parties. Academia has become a useful partner for Big Data methodological and technical questions. It also plays an essential role in training future data scientists by creating new study programmes adapted to Big Data challenges. The collaboration between NSOs and research institutions ought to benefit from public funding, such as research grants.

5.3.2. Financial challenges

A frequently heard argument in favour of using Big Data in official statistics is that it is a low cost source of data. However, the data are often owned by a person or institution that might only share the data under certain conditions, principally to sell them. The possibility of further public dissemination of the data might represent a problem as well.

Even if the data are freely available, they should be analysed to retrieve the specifically required information. This is the stage where traditional statistics should be replaced by data science techniques which are suitable for analysing Big Data. Nevertheless, the analysis requires proper data pre-processing and management. Hiring additional staff members or training existing ones may entail a significant increase in costs.

Another essential financial feature is investment in technical capacity which is necessary to make Big Data analysis possible. The majority of data science techniques are time consuming and require adequate technical equipment to handle the volume and variety of Big Data. Moreover, these new technological systems should safeguard data security and compliance with data protection, which is closely related to the legislative and privacy challenges mentioned above.

5.3.3. Methodological challenges

Methodology is another huge challenge for official statistics in dealing with Big Data sources. Official statistics is built on the sampling theory: statisticians identify a target population, develop a survey to cover this population, draw a sample, collect the data and process them (UNECE 2013). Traditional surveys ensure appropriate representativeness of the sample before continuing with the data collection process. Big Data, by contrast, is a continuous flow of data which is not bounded. The unstructured nature of many Big Data sources makes it very difficult to extract relevant statistical information. Moreover, the subpopulations Big Data sources represent are not typical target populations for official statistics. For instance, people who use social media networks do not necessarily reflect the actual structure of society. Big Data tend to be more selective of an investigated target population than representative. The appropriate representativeness of the sample can be determined by comparing the characteristics of the population covered with those of the target population. This step might be more difficult in cases in which no suitable characteristic features are available to conduct such a comparison. One example is social media where people often register with a username and other information, such as age or sex, but it is not clear whether these data are correct (Daas, Puts, Bulenes, & Hurk, 2015).

Although social media are the biggest potential source of Big Data, they need to be treated very prudently. The resulting information may be biased not only in terms of representation of the population, but also because of intentional posting of false and misleading messages. This phenomenon, known as “flaming”, “hating” or “trolling” has been spreading quickly on social networks as forms of provocation (McCosker 2014).

The mentioned methodological considerations imply that Big Data requires a completely different approach because of where the data is derived from and the fact that the information these data hold need to be retrieved by specially designed methods. Using trained data scientists and implementing statistical learning and data mining techniques is therefore highly recommended for obtaining relevant statistical results from these data. Emerging Big Data sources also require the development of new tools for data visualization.

Another very important factor in official statistics is the reference period, i.e. the time period the values refer to. In many Big Data cases, it would be very difficult to determine an exact reference period.

The primary aim of official statistics is to provide data in the form of time series analyses, investigating trends and possible future developments for the purpose of continuity and to avoid

data gaps. However, many Big Data sources consist of fluctuating information which requires special attention during data processing.

The information provided by official statistics has a very unique value, since it is based on international statistical standards and thus allows international comparisons of the relevant information among countries. International statistical standards cover definitions, classifications and other related standards. Unstructured Big Data derived from various sources can thus not really compete.

The key challenge is undoubtedly the quality of data provided by organizations. Institutions responsible for the production of official statistics are considered trustworthy and reliable. Implementing Big Data sources in their core competencies should not negatively affect these positive attributes.

5.3.4. Technological challenges

Technological challenges generally refer to the level of knowledge, experience, skills and all related costs created by investment in technological innovation. Technological development has never been as prevalent as today. The requirements on computing power or storage facilities are becoming more demanding, which is also the case for Big Data analysis.

Large investments in equipment are necessary for data processing to be carried out within an acceptable time period and to store relevant data until it is needed. The development of a proper IT environment does not resolve all problems, since highly skilled data scientists are crucial for processing large data sets and producing relevant outputs.

Storage could become a significant factor with regard to long-term statistics, because keeping a large amount of data in-house may be very expensive.

The main difference between traditional official statistics and Big Data sources is evident in the information they provide to society. Whereas Big Data can be used to generate a steady flow of information about **what** is happening or **how**, traditional research focuses on understanding a problem in more depth. Official statistics seek to explain **why** we observe certain trends or deviations thereof.

With regard to all of the above mentioned challenges, Big Data is obviously no panacea for official statistics. It can be considered potentially relevant complementary information to standard data. The biggest advantage of Big Data is that statistics can be compiled very quickly without long waiting periods to collect the corresponding data. Big Data can help provide first

flash estimates to policy- and decision makers who can use this most recent information to make decisions on which measures to take (Baldacci, et al. 2016).

6. Big Data in industrial statistics

Several research projects have demonstrated the possibilities Big Data offer to official statistics (UNECE 2016). However, most of the research activities involve satellite images for geospatial use of statistics or mobile phone data to improve tourism and migration statistics. The use of Big Data in an economic sense has not yet been thoroughly investigated and is mostly represented by internet sales data on prices or aggregated private data due to proprietary and privacy concerns by the owners of Big Data. Thus applications of alternative data sources to industrial statistics, in particular to manufacturing-related activities, remain unexplored.

To date, UNIDO has carried out one study on the relevance of Big Data for industrial statistics. However, no case study has been conducted, which incorporates alternative data sources to adjust UNIDO estimates. There are many practical aspects that have to be taken account in order to do so (Landefeld 2015). Official statistics is based on internationally agreed methodological standards to improve international comparability. This harmonized system of internationally comparable data for industrial statistics was developed based on the International Recommendations for Industrial Statistics (United Nations 2008) as the main guideline for international agencies and national statistical offices. Various sources of Big Data might be incompatible with these methodological standards. There is thus a call for action for cooperation between data scientists and traditionally trained statisticians. Data scientists specialized in data mining and processing have to learn and adhere to international standards when extracting and compiling data for official statistics, particularly those related to manufacturing activities. However, this will require significant investments in terms of time and resources in professional training as well as in building a professional IT infrastructure and capacity.

It is without debate that alternative data sources are not replacing the role of official statistics, but they represent a new supplementary approach for data collection and adjustment. The NSOs might benefit from an upturn of Big Data in the sense of complementing traditional surveys and filling possible data gaps. Nevertheless, data gaps in official statistics could emerge from the data gaps obtained from Big Data. Similar concerns can also be raised about the coverage and representativeness of the sample or differences in reference year. For instance, low coverage is typical in least developed countries with a lack of statistical capacity building. The question remains whether some Big Data sources on industrial statistics even exist in these countries to provide reasonable estimates.

Although Big Data represent a low cost complementary source for data, incorporating them into the system of NSOs would require huge investments to train data scientists so they can describe potential inconsistencies and deficiencies and in developing partnerships and data protocols with businesses and administrators that own such Big Data.

Guidelines and recommendations on Big Data use for macroeconomic nowcasting has been presented in Baldacci et al., 2016 to improve short-term estimates of macroeconomic indicators. The advantage of Big Data is the frequency and amount of data, which entails a higher number of nowcast updates, either weekly or even daily. Thus, policy- and decision makers would be able to immediately revise their actions based on the new updated estimates.

Baldacci et al., 2016 present a step-by-step approach on how to identify and choose Big Data sources and pre-treatment and econometric modelling of such data until a comparative evaluation of the results can be made to decide whether and to what extent the use of Big Data is appropriate or not.

6.1. A step-by-step approach to evaluate Big Data

The step-by-step approach introduced by Baldacci et al., 2016 describes the use of Big Data sources for economic nowcasting to adjust early estimates of the main macroeconomic indicators. The key question is deciding whether to use Big Data or not, i.e. the value added of incorporating Big Data to obtain official early macroeconomic estimates. For instance, nowcasting macroeconomic indicators, such as GDP growth, may be based on Big Data, which in turn is based on coincident and leading indicators of economic activity (Carriero, Clark and Marcellino 2015).

It must be highlighted that the use of Big Data sources for nowcasting economic indicators is still based on only a few pilot empirical studies and has not yet been tested. A step-by-step manual developed by Baldacci, et al., 2016 describes the seven steps for assessing the value of using Big Data for economic nowcasting.

1) Step 1: Big Data usefulness within a nowcasting exercise

Firstly, it is important to note that the information provided by Big Data should be considered as being potentially relevant complementary information to standard data. The use of Big Data is only recommended if it can solve existing problems or improve the quality and timeliness of nowcasting faster than traditionally used coincident and leading indicators. Moreover, Big Data-based indicators should not be affected by any uncertainity or even spurious correlations with the target variable.

2) Step 2: Big Data search

The official classification of Big Data, defined by UNECE in 2013, has already been discussed in Chapter 3. All three classification types can be potentially considered relevant for economic nowcasting. Nevertheless, it is very difficult to create general guidelines for the use of Big Data because they need to be chosen very carefully with respect to the target nowcasting variable. In other words, Big Data sources require special treatment based on the individual characteristics of the given economic indicator.

3) Step 3: Assesment of Big Data accessibility and quality

One major concern is data avilability as most of the data are provided by private institutions. In addition, data availability of Big Data may be quite unstable. The data currently provided by a given institution may not be available in, for instance, five years from now. The continuity and reliability of data provision may thus not be guaranteed to ensure the methodological stability of estimates. For example, Google Trends could stop providing its services or could start charging for the data. It is also possible for new data platforms and types to arise in the near future and that the established methodology is redefined.

Determining whether Big Data are appropriate for nowcasting applications, one relevant factor is the overall number of temporal observations in the frequency of the target economic indicator in months or quarters (Baldacci, et al. 2016). Moreover, the available and used Big Data must include well-structured documentation which is able to provide relevant metadata to assure the quality of nowcasting.

The quality evaluation of Big Data sources should also entail a proper assessment of the presence of a bias. The first type of bias is related to the sample population which does not necessarily reflect the true structure of the target society. For example, Google Trends covers only population with unlimited access to internet and Google-based search tools. Another source of bias is non-response to online surveys. If the structure of respondents is biased, then the prediction might also be biased and the data needs to be corrected for the bias.

The nowcasting of macroeconomic indicators is generally very sensitive to various economic shocks which may often be unpredictable. The same source of instability may also apply to Big Data, since the size and quality of data changes rapidly over time in comparison with standard data collection.

4) **Step 4: Big Data preparation**

Proper data preparation is one of the most essential factors of Big Data processing in nowcasting exercises. The format of Big Data is typically unstructured and their processing should be supported by a given prior data set to choose an appropriate transformation.

Although it would be convenient to build a Big Data IT environment with incorporated procedures for an automatic conversion of data into structured data sets, no unique tranformation exists that can be used as a panacea for all Big Data sources suitable for economic nowcasting.

The data has to be cleaned or filtered by removing deterministic patterns, such as outliers, calendar effects or missing observations. If the target variable is seasonally adjusted, the seasonal component should be removed as well from the Big Data-based variables.

In short, Big Data tranformation and preparation is a very time and resource demanding process.

5) **Step 5: Designing a Big Data modelling strategy**

After all Big Data has been pre-treated, it can be analysed. This stage requires appropriate Big Data econometrics. Many statistical and data science methods have been developed to capture the specific information being sought from the Big Data sources. The techniques can be categorized into five main approaches: 1) machine learning methods, 2) heuristic optimization, 3) dimensionality reduction techniques, 4) shrinkage estimators and Bayesian methods, and 5) nowcast pooling (Baldacci, et al. 2016).

6) **Step 6: Results evaluation of Big Data based nowcasting**

The impact of Big Data-based results should not be overestimated, therefore, a critical and comprehensive evaluation must be carried out. A proper evaluation should include both standard models and models based on Big Data. Models involving Big Data should be preferred when they significantly improve the reliability, accuracy and timeliness of the target variable in nowcasting. Moreover, special attention needs to be paid to several factors. The size of Big Data and the number of models are larger, which increases the risk of false positives and the instability associated with Big Data. The biggest problem of Big Data sources analysis is the risk of misinterpreting correlations and causes, which was the case with Google flu trends (Arthur 2014). Researchers are aware of the so-called “Big Data hubris” and warn that too much weight is given to analyses which are inherently not relevant and should be carefully re-evaluated.

7) Step 7: Implementation of Big Data-based nowcasting

The seventh and final step entails implementation of Big Data-based nowcasting and the dissemination strategy. Decisions have to be made to ensure the timeliness and reliability of nowcasting supported by clear metadata, references and methodological papers in order to facilitate access to the information by potential users.

6.2. Examples of Big Data in industrial statistics

The goal of official statistics is to produce relevant, objective and accurate data to keep users well informed and foster good policy and decision-making. To ensure all public expectations are met, data collection, processing and dissemination may cause a modest delay in data release with respect to the reference period. Many private companies have realized that they do not need to wait until the official statistics are released and have decided to compile their own indicators. These indicators are computed exclusively by private companies to provide an overview of current business conditions in a given economy.

Plenty of economic indicators are produced by the private sector to track business conditions such as:

- The Billion Prices Project (BPP),
- The Markit PMI™ (Purchasing Managers' Index™),
- Global Manufacturing Competitiveness Index (GMCI).

The main concern with regard to privately compiled indicators is whether they are relevant and useful in terms of complementing official estimates on economic indicators, how the data can be accessed by the public and how much they cost.

The Billion Prices Project

An interesting example of Big Data use to compile economic statistics is the Billion Prices Project (BPP)[17] produced by the Massachusetts Institute of Technology (Cavallo and Rigobon 2016). The BBP index represents an alternative approach to measuring inflation based on aggregated price information from thousands of online retailers. This indicator which is compiled online does not replace the traditional consumer price index (CPI), but may definitely

[17] http://bpp.mit.edu/

be a useful tool for short-term estimates or to fill gaps of inflation data time series (Cavallo and Bertolotto 2016).

The sample of retailers is chosen to be representative of retail transactions, thus focuses exclusively on large multichannel retailers (online-only retailers are disregarded), and categories of goods are selected with respect to the traditional structure of the consumer price index basket. The available data consist of 25 countries and provide information on categories covering at least 70 percent of the weights in consumer price index baskets (Cavallo and Rigobon 2016).

The main concern remains when asking whether online prices differ from offline prices. A direct comparison was conducted by the Billion Prices Project, which found that on average, 70 per cent of the price levels were identical in the offline and online samples (Cavallo and Rigobon 2016). The biggest differences were observed in retailers selling electronics or apparel, and the lowest in drugstores and office supply retailers. The conducted research also suggests that online prices may predict forthcoming changes in offline prices.

Figure 1 illustrates a comparison between the official consumer price index (CPI) and the index based on online prices in the United States. Figure 2 compares monthly (month-to-month) growth rates of both indices between 2011 and 2015 (The Billion Prices Project 2017). The lagged pattern between online and offline prices is clearly visible. The CPI's trend is slightly delayed behind the corresponding online prices. The same pattern is also visible in monthly growth rates, which supports the idea that online prices can be used for short-term predictions of the CPI.

Figure 1: Comparison of CPI and online price index in the United States

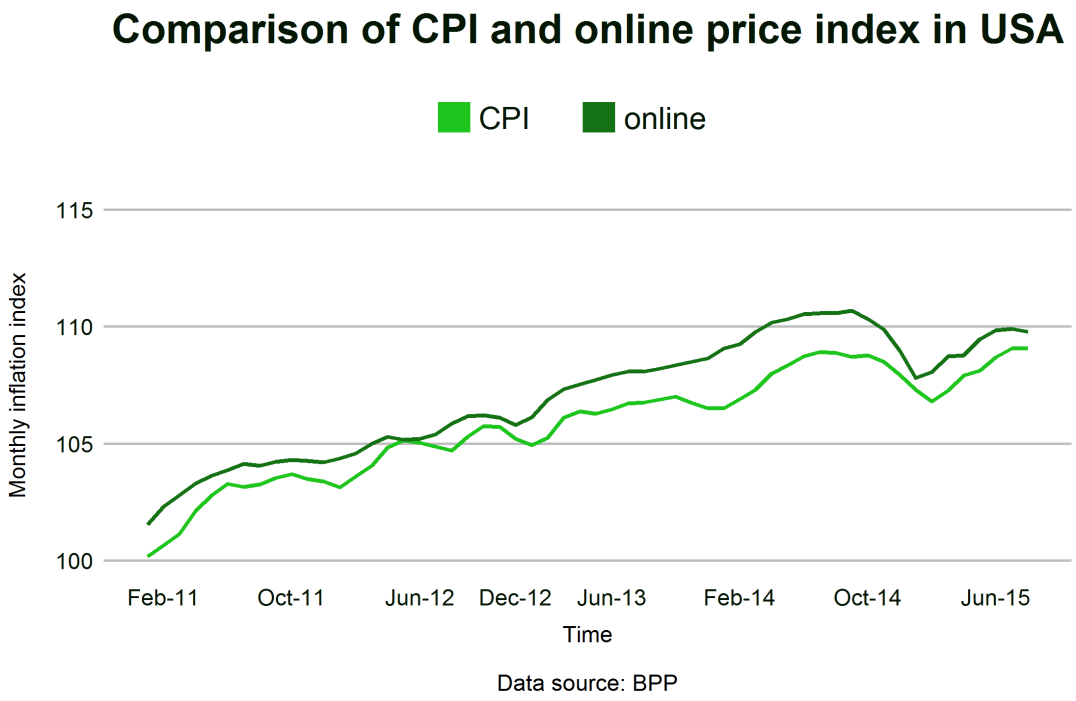
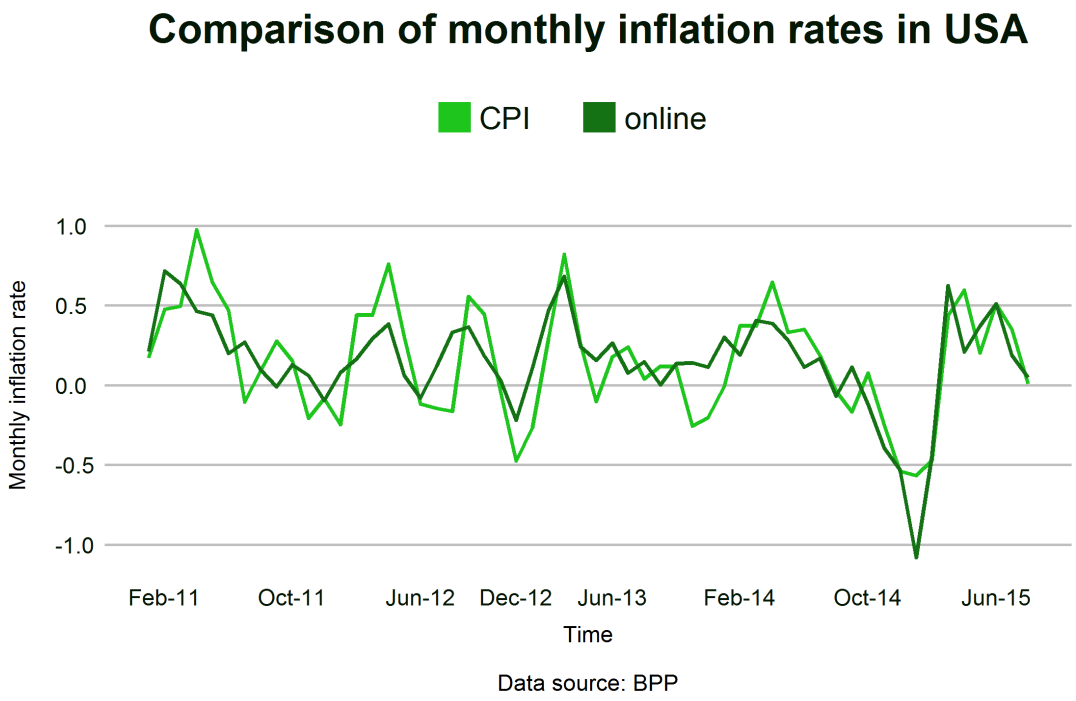


Figure 2: Comparison of monthly (month-to-month) inflation growth rates in United States



The Markit PMI™ (Purchasing Managers' Index™)

One of UNIDO Statistics' responsibilities is the collection and quarterly publication of index numbers of industrial production (IIP) (UNIDO 2011). IIP measures the growth of the volume of industrial production free from price fluctuations and represents a reliable approximation of value added growth in the short term. For the purposes of IIP production, the Purchasing Managers' Index™ (PMI™) may help determine potential trends in cases where data are not available or have to be estimated. The PMI™ series published by Markit are the economic indicators used by central banks and financial markets to forecast official economic data[18]. They are produced monthly, cover more than 30 countries and survey over 20,000 companies. The questionnaire investigates development in manufacturing, construction, services and the entire economy. The survey respondents are purchasing managers who are meticulously selected to represent the true structure of the respective industry in the economy based on international standards and the appropriate weighting system. Managers are asked to determine whether their business conditions have either improved, deteriorated or remained the same compared to the previous month. The biggest advantage of this indicator is its fast compilation since it does not require any actual numbers from companies.

UNIDO has been publishing quarterly reports on world manufacturing since 2011 by collecting IIP for more than 70 countries worldwide. The Manufacturing PMI™ is released monthly on the first working day of the month. IHS Markit compares Manufacturing PMI™ to corresponding official data (for example, see Markit Eurozone Manufacturing PMI™) (IHS Markit 2017). Manufacturing PMI indices may represent a possible means for estimating quarterly IIP or replacing missing official figures.

Global Manufacturing Competitiveness Index (GMCI)

The Global Manufacturing Competitiveness Index (GMCI) is an index published by the Deloitte Touche Tohmatsu Limited (DTTL) Global Consumer & Industrial Products Industry Group and the Council on Competitiveness (Deloitte 2016). The GMCI is an output of the CEO survey which investigates how manufacturing CEOs view competitiveness around the world. In 2016, the survey counted 540 valid responses. It is divided into three sections:

- Business confidence and current environment – respondents share opinions on the global economic environment at the country and industry level;

[18] https://www.markit.com/product/pmi

- Manufacturing competitiveness – respondents rate the relative importance of components that drive the competitiveness of a country's manufacturing sector to rank the manufacturing competitiveness of 40 countries (now and a five-year perspective);
- Demographics – investigating a company's demographic conditions.

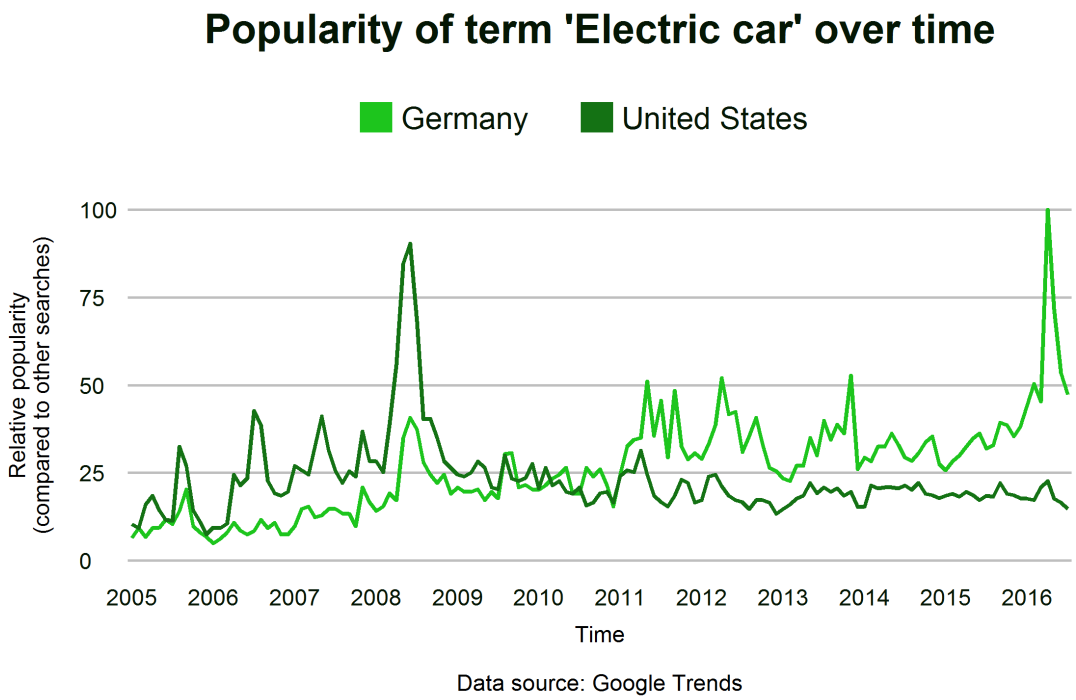
UNIDO also compiles a composite index of industrial performance through its Competitive Industrial Performance (CIP) index. Industrial competitiveness is defined as the capacity of countries to increase their presence in international and domestic markets whilst simultaneously developing industrial sectors and activities with higher value added and technological content[19].

Google Trends

Another alternative data source to the above mentioned indices produced by private companies is Google Trends, launched by Google to allow users to explore various interests through search query data. The use of Google Trends has been central in studies using Big Data in economics (Shawn & Stridsberg, 2015, Combes & Bortoli, 2016). For instance, trends derived from the search can be used to nowcast the sales figures for commodities. Google Trends might indicate demand for certain manufacturing goods since they represent users' intentions (the data correspond to the relative popularity of a search term). In other words, Google Trends provides information on how often a term was searched relative to the total number of global searches. Based on conducted case studies, it is still not clear how effective Google Trends is in predicting developments and continues to be a hotly discussed issue.

[19] For detailed methodology and trends in UNIDO's CIP index, see UNIDO (2017), Competitive Industrial Performance Report 2016. Volume I.

Figure 3: Relative popularity of term "Electric car" over time in Germany and United States[20]



A simple example of Google Trends' use of data is depicted in Figure 3. It demonstrates the relative popularity of the term "electric car" in Germany and the United States in recent years, i.e. how popular the product was compared to other searches. The recent announcement that Germany will subsidize electric car purchases to give the sluggish growth in the industry a jolt considerably increased online interest in electric cars (causing a noticeable peak in the graph in 2016). California electric automaker Tesla Motors began development on the Tesla Roadster in 2004, which was first delivered to customers in 2008. This event is visible in the graph as the peak in mid-2008 in both Germany and the US. Growth arrest in the US after 2009 did not result in a loss of interest in electric cars. It only meant that the popularity of searches on electric cars decreased compared to other searches, although the total number of searches for that term might have stayed the same. In other words, other topics dominated people's interests.

[20] https://www.google.com/trends/explore?cat=47&date=2005-01-01%202016-08-31,2005-01-01%202016-08-31&geo=DE,US&q=%2Fm%2F03nlf2w,%2Fm%2F03nlf2w

6.3. How can Big Data be used in official statistics?

The previous chapter provided various examples of how Big Data sources can be used in economics, which seem to be adequate or often even recommended to adjust official statistics data production.

Official statistics represent the only source of reference data that are produced with certain time lags. The estimated contribution of increasing data sources is to mitigate the time lags between data provision. Big Data may provide a useful source of different extrapolators such as commodity sales, price and registration data, airlines, hotels and motels, investment company data (FDI), banks, etc. National statistical offices and international agencies could thus derive their nowcasted values and estimates from the reference data, from official statistics, and extrapolators, for instance, Google trends or PMI™.

Data availability and its public access, which is mostly determined by data ownership, are the biggest obstacles. As already mentioned, UNIDO is responsible for publishing quarterly indices of industrial production (IIP) to report on world manufacturing production. Data on industrial production are collected from national statistical offices. In case of missing data, UNIDO conducts imputations or projections, where appropriate. These estimates are generally replaced as soon as the officially reported values become available in national statistical publications (UNIDO 2011). Surveys investigating business conditions would be valuable sources for adjusting missing data or nowcasting IIP. These surveys are cost effective and easy to compile, which also increases the frequency of data dissemination.

There are various other Big Data sources that could potentially be used to adjust official industrial statistics estimates. Important information can also be extracted from the Survey of Professional Forecasters (SPF) conducted, for instance, by the European Central Bank (ECB)[21] or in the US by the Federal Reserve Bank of Philadelphia[22]. Such surveys are conducted on a quarterly basis and usually forecast key economic variables based on experts' predictions. Other frequently cited surveys are the Blue Chip Economic Indicators and Blue Chip Financial Forecasts, available on a monthly basis[23].

One widely discussed and promoted index is the Purchasing Managers' Index™ (PMI™) which has already been introduced above. However, the Purchasing Managers' Index™ is owned and licensed by IHS Markit and the data are not available to the public, but exclusively to subscribers. For UNIDO to gain open access to PMI data, data sharing protocols and a

[21] http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html
[22] https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters
[23] https://lrus.wolterskluwer.com/product-family/blue-chip

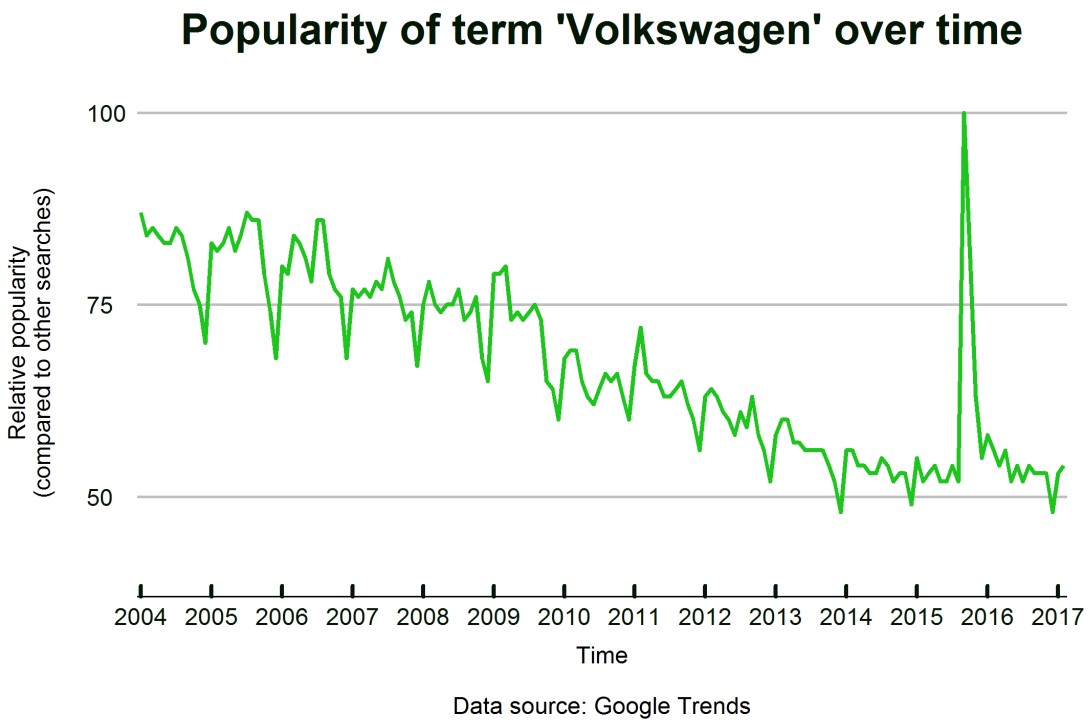
partnership would have to be established, which may represent additional costs. For this reason, no potential comparison of IIP and PMI™ has been made so far.

Google Trends is also frequently examined as a potential extrapolator, since the data are provided by Google free of charge. The biggest advantages of Google Trends are easy access, flexibility and long-term availability. People can search for a given term and its relative popularity over time worldwide or in a specific country. It is very important to correctly understand the term 'relative popularity', since a decreasing trend of index numbers does not necessarily mean a significant decline in the absolute number of searches. The Google Trends index is calculated by dividing each data point by the total number of searches in the region and time range scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics[24].

Figure 4 depicts the relative popularity of the term "Volkswagen" between 2004 and 2017 (monthly data) worldwide. Volkswagen is one of the world's largest automakers whose car production is increasing every year[25]. However, Google Trends shows a decreasing tendency in terms of its relative popularity. It is obvious that the popularity of the term "Volkswagen" compared to other searches does not accurately reflect the actual car production or sales. Moreover, we can see an extreme peak in September 2016. This effect was not caused by an extraordinary interest by people in buying Volkswagen cars, but was probably linked to people searching for more information on the Volkswagen emission scandal.

[24] https://support.google.com/trends
[25] http://www.forbes.com/sites/bertelschmitt/2017/01/30/its-official-volkswagen-worlds-largest-automaker-2016-or-maybe-toyota/#7871b2fe20ba

Figure 4: Relative popularity of term "Volkswagen" over time worldwide

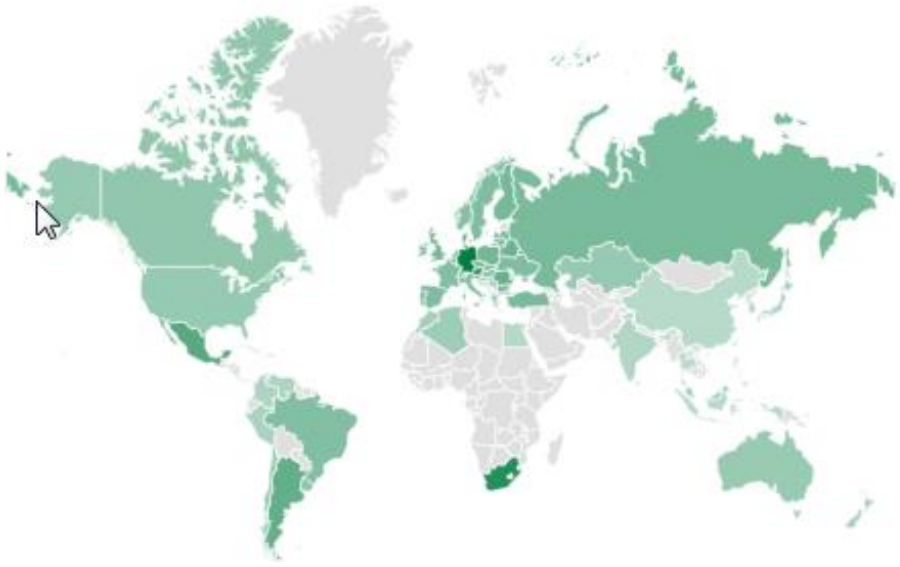


The second issue relating to using Google Trends for the purposes of official statistics is its coverage of the population. Figure 5 illustrates the relative popularity of the term "Volkswagen" by region. We observe that the term 'Volkswagen' was more frequently searched in Germany and South Africa than in other countries. Clearly, some regions are not covered at all, apparently due to limited internet access. For this reason, Google Trends does not provide a fully representative sample of the entire population, which official statistics, on the other hand, do.

The significance of a proper choice of extrapolators is demonstrated by an example of the index of industrial production (IIP) for motor vehicles, trailers and semi-trailers in the US. The IIP is generally derived from data obtained by the Federal Reserve, often referred to as "the Fed"[26]. We have compared the index numbers of industrial production with IIP constructed for the US' domestic auto production data retrieved from the Bureau of Economic Analysis[27]. The resulting IIP for ISIC Revision 4 Division 29, motor vehicles, trailers and semi-trailers, from November 2012 to December 2016, are presented in Figure 6. The graph shows analogous patterns in the case of peaks. Nevertheless, looking at the time series taking a longer perspective reveals contradictory trends. Whereas the official IIP constructed from Federal Reserve data demonstrates a decreasing tendency, the BEA car production numbers show an opposite trend.

[26] https://www.federalreserve.gov
[27] http://www.bea.gov

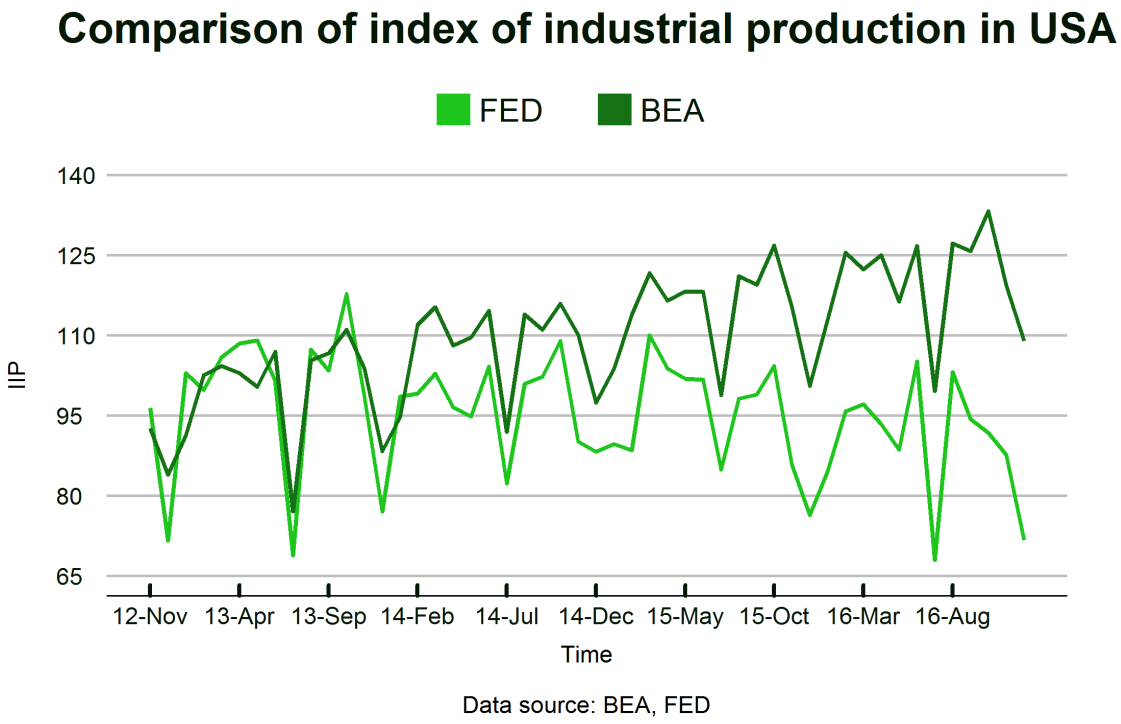
Figure 5: Relative popularity of term "Volkswagen" by region between 2004 and 2017



For this example, an alternative data source than the official statistics sources was chosen to calculate IIP. The resulting values demonstrate that even official data can change the trend of a given time series. This example emphasizes a very important factor related to Big Data sources and raises the question which data can be used as extrapolators. No universal checking mechanism exists to ensure the quality of estimates based on Big Data constructed extrapolators. Moreover, the unofficial data sources may vary over time, for example, some companies could end their data production and others might start producing data.

Big Data has definitely changed the world, and not only the world of data. Official statistics cannot ignore new trends and the fact that data are everywhere. However, as proven above, implementing Big Data sources in official statistics requires careful treatment and a frequent revision of methodology.

Figure 6: Comparison of index of industrial production (IIP) of motor vehicles, trailers and semi-trailers in the US constructed for two different sources of data



Conclusions

Big Data refers to the entire ecosystem related not only to flows of data (big data streams), but also encompasses all institutions that produce and use Big Data and the development of corresponding methods and tools.

Although no concrete steps for the implementation of Big Data in industrial statistics have been taken so far, several options in this direction are possible. Incorporating Big Data or other alternative data sources into industrial statistics may entail similar problems as described in this paper for the general use of Big Data in official statistics. Big Data are not a panacea for industrial statistics because they may only represent a source of data available for short-term estimates. Many applications would require an individual approach based on characteristic features of the target variable. Filling data gaps might be problematic since the missing values could appear in both data sources, in official ones and in alternative data sources.

To improve the quality and reliability of results derived from Big Data, a close cooperation should be established between national statistical offices and academia to develop and test new methods suitable for corresponding estimates. Collaboration with private companies can also

bring new insights and increase data availability, and should be taken into account whenever possible.

Currently, the main task of UNIDO Statistics is to develop a concise statistical methodology for the use of Big Data in industrial statistics with a special focus on creating partnerships and data protocols with owners of such data. UNIDO with its international mandate in industrial statistics should play a leading role in all activities relating to the design of standards for Big Data in this field. In any case, a transition from the official statistics based system to a new Big Data-driven system would require a huge investment in terms of time and resources in professional staff training and IT capacity building. Despite these challenges, Big Data will have an impact on the future of statistics. We need to embrace Big Data in a timely fashion and use it to our advantage.

7. Bibliography

Arthur, Charles. *Google Flu Trends is no longer good at predicting flu, scientists find.* The Guardian, 2014.

Baldacci, Emanuele, et al. *Big Data and Macroeconomic Nowcasting: from data access to modelling.* EUROSTAT, 2016.

Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino. "Real-Time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility." *Journal of the Royal Statistical Society Series A* 178, no. 4 (2015): 837–862.

Cavallo, A, and R Rigobon. "The Billion Prices Project: Using Online Prices for Measurement and Reserach." *Journal of Economic Perspectives* 30, no. 2 (2016): 151-178.

Cavallo, Alberto, and Manuel Bertolotto. "Filling the Gap in Argentina's Inflation Data." 2016.

Combes, Stéphaniea, and Clément Bortoli. "Nowcasting with Google Trends, the more is not always the better." *First International Conference on Advanced Research Methods and Analytics, CARMA2016.* Valencia, 2016.

Daas, Piet JH, Marco J. Puts, Bart Bulenes, and Paul AM van den Hurk. "Big Data as a Source for Official Statistics." *Journal of Official Statistics* 32, no. 2 (2015): 249-262.

Data-Pop Alliance. "Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America." 2015.

Deloitte. "Global Manufacturing Competitivness Index 2016." 2016.

Dornik, Jurgen A., and David F. Hendry. "Statistical model selection with "Big Data"." *Cogent Economics & Finance* 3, no. 1 (2015).

Economic Commission for Europe. "Big data- an opportunity or a threat to official statistics?" Paris, 2014.

EUROSTAT. "European Statistics Code of Practice." 2014.

Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International Journal of Information Management* 35, no. 2 (2015): 137-144.

IHS Markit. *Markit Eurozone Manufacturing PMI™.* IHS Markit, 2017.

Landefeld, Steven. *Strengthening UNIDO Statistical Capacity for Sustainable Development Goal Monitoring* . UNIDO, 2015.

Laney, Douglas. *3-D Data Management: Controlling Data Volume, Velocity and Variety.* META Group Reserach Note, 2001.

Lee, Jay, Edzel Lapira, Behrad Bagheri, and Hung-an Kao. “Recent advances and trends in predictive manufacturing systems in big data environment.” *Manufacturing Letters* 1, no. 1 (2013): 38-41.

Marr, Bernard. “4 Ways Big Data Will Change Every Business.” *The Forbes*, 8 September 2015.

—. “What Everyone Must Know About Industry 4.0.” *The Forbes*, 20 June 2016.

McCosker, Anthony. “Trolling as provocation: YouTube’s agonistic publics.” *Convergence: The International Journal of Research into New Media Technologies* 20, no. 2 (2014): 201-2017.

OECD. *Data-Driven Innovation: Big Data for Growth and Well-Being.* Paris: OECD Publishing, 2015.

Reimsbach-Kounatze, Christian. *The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies.* Paris: OECD Digital Economy Papers, 2015.

Shawn, Lim Kai Jie, and Douglas Stridsberg. “Feeling the Market's Pulse with Google Trends.” *International Federation of Technical Analysts' Journal*, 2015.

Statistical Commission. “Report on the Global Working Group on Big Data for Official Statistics.” 2016.

Struijs, Peter, Barteld Braaksma, and Piet JH Daas. “Official Statistics and Big Data.” *Big Data & Society*, 2014: 1-6.

The Billion Prices Project. *Online – Offline Price Comparison.* Harvard/MIT Dataverse, 2017.

The World Bank. *Central America: Big Data in Action for Development.* The World Bank, 2014.

UNDP, UN Global Pulse. “A Guide to Data Innovation for Development: From Idea to.” 2016.

UNECE. “Big Data Inventory.” 2016.

UNECE. *Classification of Types of Big Data.* 2017.

UNECE. “What does "Big Data"mean for official statistics?” 2013.

UNIDO. *Competitive Industrial Performance Report 2016. Volume I.* Vienna: United Nations Industrial Development Organization, 2013.

UNIDO. “Strengthening UNIDO Statistical Capacity for Sustainable Development Goal Monitoring.” 2015.

UNIDO. “World Manufacturing Production, Quarter 1 (including methodology) - 2011.” 2011.

United Nations. *A World That Counts: Mobilising The Data Revolution for Sustainable Development.* New York: United Nations, 2014.

United Nations. “Fundamental principles of official statistics.” 2013.

United Nations. “International Recommendations for Industrial Statistics (IRIS).” 2008.

United Nations. “Resolution adopted by the General Assembly on 25 September 2015.” 2015.

Appendix I - UNECE Big Data classification

Social networks (human-sourced information)			
		1100	Social networks: Facebook, Twitter, Tumblr, etc.
		1200	Blogs and comments
		1300	Personal documents
		1400	Pictures: Instagram, Flickr, Picasa, etc.
		1500	Videos: YouTube, etc.
		1600	Internet searches
		1700	Mobile data content: text messages
		1800	User-generated maps
		1900	E-Mail
Traditional business systems (process-mediated data)			
21	Data produced by public agencies		
		2110	Medical records
22	Data produced by businesses		
		2210	Commercial transactions
		2220	Banking/stock records
		2230	E-commerce
		2240	Credit cards
Internet of Things (machine-generated data)			
31	Data from sensors		
	311	Fixed sensors	
		3111	Home automation
		3112	Weather/pollution sensors

		3113	Traffic sensors/webcam
		3114	Scientific sensors
		3115	Security/surveillance videos/images
	312	Mobile sensors (tracking)	
		3121	Mobile phone location
		3122	Cars
		3123	Satellite images
32	Data from computer systems		
		3210	Logs
		3220	Web logs

UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION

Vienna International Centre · P.O. Box 300 9 · 1400 Vienna · Austria
Tel.: (+43-1) 26026-0 · E-mail: info@unido.org
www.unido.org