NOVA IMS Information Management School

# MEGI

Master's degree Program in

**Statistics and Information Management**

# A Bibliometric Analysis on the use of Big Data Sources in the Compilation of Official Statistics

Zukiswa Mdingi

Master Thesis

presented as partial requirement for obtaining the Master's degree Program in Statistics and Information Management

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

# A Bibliometric Analysis on the use of Big Data Sources in the Compilation of Official Statistics

By

Zukiswa Mdingi

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management: with a specialization in Data Analytics

## Supervised by

Professor Bruno Damasio (PhD), Universidade Nova de Lisboa

May, 2024

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.
[Basel, 28 May 2024]

# ABSTRACT

The goal of any NSI is its ability to produce accurate, timely, reliable, and trusted official statistics. Historically, NSIs have relied solely on data collected from traditional data sources such as surveys and administrative data to compile official statistics. The dawn of big data is one that is expected to have a huge impact on organisations such as National Statistical Institutions (NSIs). The purpose of this study is to conduct a quantitative science study on the evolution of topics in the field of big data sources in the compilation of official statistics. The study will try to identify the evolution of topics in the research of big data sources in the compilation of official statistics, to identify which country(s) have made the most contributions to this research field, and to uncover emerging and relevant topics related to this field. It was concluded that, big data sources such as scanner data, mobile phone data, sensor data, and transactional data have been used institutions such as Statistics Netherlands, Bank Estonia, Italian National Institute of Statistics, US Census Bureau, and the Spanish National Institute of Statistics. Scanner data did seem to be the one big data source that has been used by a few institutions for the compilation of the Consumer Price Index (CPI). The study identified some of the key topics relating to big data and official statistics. Big data, official statistics, nowcasting, machine learning are some of the common keywords that have been covered in this research field. When analysing the bi-gram (two-words) of the most commonly found abstract words, emerging and relevant topics include web scraping, official statistics, data sources, and data governance.

**Key words:**

Official statistics; Big data; Bibliometrics; National Statistical Institutions (NSIs); Big data sources

**Sustainable Development Goals (SGD):**



# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| NSI | National Statistical Institutions |
| CPI | Consumer Price Index |
| GWG | Global Working Group |
| SDGs | Sustainable Development Goals |
| WoS | Web of Science |
| UN | United Nations |
| UNSD | United Nations Statistics Division |
| UNECE | United Nations Economic Commission for Europe |
| Eurostat | European Commission |

# CHAPTER ONE

## Introduction and background of the study

Official statistics is a quality label that indicates that statistics are developed, disseminated, and produced according to international standards as set out in the Fundamental Principles of Official Statistics (UNECE, 2022). Pieter (2022) describes the Fundamental Principles of Official Statistics as the backbone of official statistics, these principles play an essential role in maintaining the high quality of official statistics. The principles provide guidelines to official statistician who produce official statistics, from data collection all the way to how the statistics are to be communicated to its users. According to the United Nations (2014) for statistics to be classified as official statistics, output statistics need to achieve a certain quality level that indicates that the statistics are in accordance with the international statistical standards and agreements.

The dawn of big data is one that is expected to have a huge impact on organisations such as National Statistical Institutions (NSIs) whose business is mainly on data compilation and dissemination. With NSIs being responsible for compiling official statistics, this new influx of data is going to require NSIs to find a way to handle the situations that will eventually arise due to big data as it will have implications for society as a whole. The availability of big data sources has allowed some private organisations to publish their own versions of statistics which could mimic official statistics, this could lead to problems if society is not be able to differentiate between privately produced statistics and official statistics. Therefore, in trying to be relevant, NSIs are trying to find ways in which they can include these big data sources into their official statistics. However, NSIs will need to ensure that in trying to keep up with the demands of society they continue to publish quality official statistics in order to remain trusted providers of official statistics (Struijs et al., 2014). The United Nations (2023) cautions that some of these new data sources might not meet the statistical requirements as stipulated in the Fundamental Principles of Official Statistics or in the NSI's Quality Framework dimensions thus making them incompatible to use in the compilation of official statistics.

The goal of any NSI is its ability to produce accurate, timely, reliable, and trusted official statistics. Historically, NSIs have relied solely on data collected from traditional data sources such as surveys and administrative data to compile official statistics. Although these sources have produced reliable official statistics over the years, the advancements in technology have created the potential for NSIs to change their old data collection and dissemination practices (Ricciato et al., 2020; Serena et al, 2021). In recent years, NSIs have begun to explore and experiment with data that is becoming available through these new digital channels by trying to ascertain possibilities of introducing such data sources as inputs into official statistics. Although big data sources might provide advantages such as reducing respondent burden, lower costs, improved timeliness, granular breakdown, and higher frequency not all these new big data sources will be compatible for official statistics.

The rise of big data sources has provided some opportunities for NSIs however before these opportunities can materialise, NSIs will be required to address a few new challenges introduced by these big data sources. Some of the challenges that NSIs are faced with relate to legal, technological, financial, and methodological issues. Big data sources are known to be unstructured, volatile, and selective therefore, the biggest challenge that official statisticians are faced with concerns methodology. This is because data sources are meant to be observational data and were not designed

for data analysis. Due to this, traditional statistical methods cannot be applied. In order to address these challenges NSIs need to adopt new strategies that will allow them to integrate big data sources with traditional data sources in attempts to produce more accurate and reliable results. In addition to producing accurate and reliable official statistics, NSIs also need to ensure they guarantee the quality of the statistics (Alrousan et al., 2023; Braaksma and Zeelenberg, 2020; Struijs et al., 2014). Although the new big data sources might mean more data, Mehrhoff (2018) warns that some data sources may have limited accuracy and thus including them in the compilation of official statistics could be a threat to the integrity of published statistics.

The purpose of this study is to conduct a quantitative science study on the evolution of topics in the field of big data sources in the compilation of official statistics. This study will endeavor to analyse a sizable amount of research articles on the topic of using big data sources to compile official statistics, to study the evolution of topics in this field and also identify if there are any emerging topics currently developing in this field.

This study intends to investigate and analyse research articles published on the impact of big data on NSIs in the compilation of official statistics. The study will try to identify the evolution of topics in the research of big data sources in the compilation of official statistics, to identify which country(s) have made the most contributions to this research field, and to uncover emerging and relevant topics related to this field.

The objectives of this study are to (i) examine the contributions made to the research by authors and countries; (ii) use bibliometric analysis methods to study the citation, conceptual, and social structures of the research field on the use of big data sources in the compilation of official statistics; (iii) identify future research directions according to the developments of relevant topics.

The study will consist of five chapters, chapter one will give the introduction and background of the study. Chapter two will present the literature review. Following chapter two will provide a breakdown of the methodology used which will be chapter three. Chapter four will provide the analysis and discussion of results. The final chapter will be presenting the conclusion, limitations, and future work.

# CHAPTER TWO

# Literature review

## 2.1 Exploring the use of big data in official statistics.

This chapter will be on the literature review of the study. This chapter will include sections that unpack the definition and requirements of official statistics according to international statistical standards set for official statistics, discuss how statistical institutions have used big data in the compilation of their official statistics, discuss the opportunities and challenges that are associated with using big data in official statistics, and finally the last section will discuss the possibilities of the future of official statistics.

Over the past decade the world has evolved and has significantly advanced technologically thus creating a more technological and modern society. The evolutions seen leading to this modern society have resulted in new digital data sources and technologies. Within this newly modern society, National Statistical Institutions (NSIs) have endeavoured to examine possible benefits and opportunities that they could yield for the field of official statistics.

For official statistics to be efficient for its economy, they need to be reliable but most importantly be produced in a timely manner to ensure relevancy. World events that had significant impacts not only on domestic but global economies such as the Great Financial Crisis or more recently the Covid-19 pandemic showed the importance of having reliable official statistics available in a timely manner (ESCAP, 2022). If NSIs were to go the route of incorporating big data sources in the compilation of their official statistics, official statisticians would be required to produce official statistics in accordance with the fundamental principles of official statistics.

Big data is a word that is largely affiliated with data that is becoming available due to the digital world, this data is associated with data that comes in large volumes and high frequency (Daas et.al, 2015). UNECE (2013) adds that big data is characterized by the 3Vs, volume, velocity, and variety. Big data has the potential to create opportunities for official statistics such as better accuracy, better detail, more frequent and timely statistics. However, this data could also lead to challenges such as uncontrolled changes in the source leading to continuity and comparability issues, data may not be a direct phenomenon of the required statistics of interest and the data could be selective and volatile (Braaksma and Zeelenberg, 2020).

## 2.2 Working groups on big data for official statistics

For many years NSIs used tradition data sources for the compilation of their official statistics. However, given the opportunities that could be leveraged off from using big data sources, NSIs have been exploring this avenue to assess the possibilities for official statistics. Following many discussions within the official statistics community, the United Nations (UN) reviewed the possibility of big data and official statistics. The UN noted that individual statistical institutions did not have enough resources to analyse and the tackle the issues that using big data in official statistics could have (United Nations, 2015b). To investigate the impact and viability of big data sources on official statistics, international statistical organisations have set up working groups and tasks forces to analyse this new phenomenon.

Some of these group include the HLG-MOS, the Global Working Group on Big Data for Official Statistics, and the UN Global Pulse.

In 2009 the United Nations Global Pulse was launched, this initiate was created to bring together experts from the UN, governments, academia, and the private sector to collaborate and develop big data projects. The initial programme focused on issues involving monitoring, evaluation, and privacy protection. The initiative aims to use access big data and to lean from the different experiences of its members in data science to develop their own functional data analytics tools (Global Pulse, 2023).

With the significant evolutions that have taken place over the last few decades, the Conference on European Statisticians established the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) which was tasked with advancing the modernisation of official statistics. The HLG-MOS operates as a collaborative platform where experts from different statistical organisations can come together and develop strategies and solutions as to how to advance the modernisation of official statistics (UNECE, 2023). The European Commission (2023) adds that the aim of HLG-MOS is to improve the efficiency of the statistical production process. In 2014 the HLG-MOS created a project "Sandbox" that was created to have a group of experts from national and international statistical organisations to share ideas and experiences on the use of big data in official statistics. According to the European Commission this project was created to be a collaborate platform where experts could come together to evaluate and test new tools, techniques, and sources which could be used in modern statistical analysis.

In 2014 at the 45$^{th}$ statistical conference, the UN Statistical Commission created the Global Working Group (GWG) on Big Data for Official Statistics. This group was created to identify opportunities and challenges that NSIs could potentially face if they were to incorporate big data in official statistics, linking specially to the reporting and monitoring of the sustainable development goals. In 2015 the working group conducted a survey on the use of big data in official statistics, the aim of this survey was to identify which NSIs have started using big data for their official statistics and what their experience was, to identify any concerns that NSIs had, and to get ideas on future steps for the implementation of big data in a modern statistical production. According to the results of the survey, the Statistical Commission (2016) noted that at least half of the NSIs had big data related projects. The survey showed that the main big data sources included scanner data for price statistics, web-scraping data for transport and labour statistics, mobile phone data for tourism and population statistics and satellite imagery data for agriculture statistics. This working group achieves its objectives through the eight working tasks teams that were established.

Eurostat (2022) also took a similar route and set up an experts' group that was tasked with "facilitating the use of new data sources for official statistics" by identifying new opportunities and coming up with a set of recommendations that would aid in the reuse of data collected by the private sector in official statistics. These working were a starting point for identifying if it would be possible to re-use this private data. This is because this new data that is being considered was collected for a single purpose and it is now being intended to be used for a different purpose therefore, additional issues could be found ranging from methodological to technical issues.

## 2.3 How NSIs have incorporated big data into their official statistics

Big data sources have evolved over the years and now includes a wide range of data (Ricciato et al, 2020). Big data sources that NSIs have experimented with include (United Nations, 2023):

Transactional data: point-of-sale processors or credit card data that is collected through an organisations electronic payment data where this type of data could provide more information on the retail sales industry. Transactional data is a desirable data source to NSIs as the richness and volume of the data provides more valuable insights thus allowing for more thorough analysis. Transactional data is available at daily or weekly frequencies, examples of traditional data include financial and logistics data. The challenge that has been experienced by NSIs with this data source is due the large and complex data structure that the data comes in. The large datasets required NSIs to upgrade their classical data processing methods as they were not able to process such large and complex data. NSIs have been able to overcome this hurdle, Statistics Canada developed a processing system for transactional data which has allowed them to ingest this large data source (Waye et al., 2018). More NSIs seem to have also been able to overcome the issue of complexities of transactional data as some NSIs are compiling some of the statistics using transactional data.

The US Census Bureau compiles the monthly retail trade statistics using the monthly retail trade survey however, the survey data is unable to provide information on localized and short-term shocks and has a reduced usefulness when it comes to real-time analysis due to the time lags. Trying to increase the frequency and granularity of the survey would prove to be infeasible as this would result in additional costs and increased respondent burden therefore, the Bureau had to find an alternative approach that would help resolve the shortcomings of the survey dataset. To fill these gaps, the Bureau made use of anonymized transaction data they sourced from a global payment technology company to be able to compile daily estimates of retail statistics by the different geographies. To retain the anonymity of the merchants and the customers, the Bureau receives only the anonymized data which is compiled according to their specifications therefore they do not receive the transaction level data. With the use of the transactional daily estimates can now be compiled with a lag of only a few days after the occurrence of a transaction. When comparing the transactions statistics with the official statistics, there is a similar pattern of monthly growth rates at national level (Aladangady et.al, 2019).

The Spanish National Institute of Statistics (INE) endeavoured to explore the possibility of compiling its retail trade index (RTI) using retail transactions of debit and credit card holders from one of Spain's largest banks. Using monthly transaction data from 2013/01 to 2016/12 INE was able to reproduce the dynamics of not only the aggregate official statistics but were also able to generate the regional retail trade index by retailer type. The exploration of this transactional data to compile RTI data highlighted some benefits which included real time availability of the data, higher frequency, and more granularity which allowed for more geographical and sectoral disaggregation. This exploration allowed for the compilation of the RTI for 50 of Spanish provinces where the current official statistics does not compile at this level (Bodas, et.al, 2018).

Scanner data: provides multiple opportunities to improve the accuracy of the data published for the Consumer Price Index (CPI). Scanner data consists of a complete coverage of items sold, the quantities of the items sold, and how much revenue was received by the retailer for these items. This information can be vital in improving the accuracy of the prices used to compile CPI, expand the sample of items,

and the information of the revenue can be used to weight items by their economic importance. Obtaining scanner data is one of the challenges that NSIs have had to deal with. Scanner data can be collected in two ways, NSIs may seek the data directly from the retailer or they can get the data from a third party. Both these data collection approaches have their benefits and challenges, the main challenge that has been experienced by some NSI is the bilateral negotiations between the NSI and the data provider. The Netherlands, Switzerland, and Australia noted that these negotiations were time consuming and that they covered a variety of topics including discussions on IT systems and confidentiality concerns (Graf, 2020; ESCAP, 2020).

Statistics Netherlands is one of the first countries to use scanner data in the compilation of their CPI. In 2002 Statistics Netherlands started to introduce supermarket scanner data from two supermarket chains into their CPI. It was in 2010 when the statistical office expanded the use scanner data into the compilation of CPI by increasing the number of supermarket chains to six, these six chains make up a market share of roughly 50% and account for more than 5% of the CPI-weights (Van der Grient and De Haan, 2010). The Australian Bureau of Statistics (ABS) started its exploration of scanner (point-of-sale) and online data in 2011 with the aim of incorporating the data into the compilation of its Consumer Price Index (CPI). The potential advantages that scanner data would have on official statistics include coverage, frequency, and granularity. The ABS sourced this data from major supermarkets in Australia, following extensive testing and experimentation scanner data was included in the production of Australia's official CPI from 2014 (ESCAP, 2020). Scanner data now contributes approximately 16% to the compilation of CPI (Australian Bureau of Statistics, 2021). The ABS senior staff had to put in significant effort in ensuring they obtain accurate and reliable scanner data from retailers. This included creating a solid relationship with the retailers so that assurances could be given the sensitivity of the nature of the data. Legal agreements were also signed to ensure that ABS would keep the name of the data supplier confident, a memorandum of understanding (MoU) was also signed ABS were it was agreed that this data would only be used for CPI, Retail Trade and National Accounts (ESCAP, 2020).

Sensor's data: Sensor data are collected from sensing devices, these devices collect information through the environment of that sensor, these include satellite imaging, road sensors and climate sensors. One of the biggest challenges that NSIs face when trying to incorporate this data source is the quality of the data. It has been found that the quality of the data is at times poor thus resulting in the statistician having to decide on its usability. This results in the need for NSIs to develop data cleaning processes before using sensory data, keeping in mind that cleaning big data will be different to cleaning small data. For Statistics Netherlands, they found that in some cases they had to check and clean large amounts of data therefore they found that they could only do this task by fully implementing an automated process.

Statistics Netherlands used large amounts of road senor data to produce official statistics. In the Netherland, about 60,000 road sensors can provide traffic images and minute-based vehicle counts at 24,000 sites. Statistics Netherlands uses this data to compile traffic intensity statistics. Between 2010 and 2014, 115 billion records had been collected, with the collected the issue of quality was a problem as the quality of the data was considered too poor to use in the compilation of official statistics. However, a carefully designed plan was implemented to handle the flaws identified in the data. The method applied to process the data was enough to improve the quality of the data to ensure that the

data was good enough to use in the production of official statistics. Statistics Netherlands was able to produce this data on their statistical database (Marco et al., 2019).

Mobile phone data: Mobile phone data is one of the big data sources that has the potential to be relevant in multiple official statistics including tourism, population, migration, and transportation statistics. A few countries have started to incorporate mobile phone data in the compilation of their official statistics, a recommendation is that before going the route of exploring the mobile phone data a NSI will need to ensure that there is a partnership amongst key stakeholders. One of the challenges associated with the use of mobile phone data for official statistics is the privacy and security of mobile phone records (UNSD, 2019).

From 2009 Bank Estonia has been investing the possibilities of using mobile phone data to compile their statistics. During the period 2009 – 2010 Estonia has been able to use mobile phone data to compile the inbound and outbound tourism statistics. These statistics have been compiled using what is referred to as mobile positioning which is, tracking the location of a mobile phone through space and time using the global positioning system (GPS). The process for compiling these statistics includes having an automatic quality control operators system which can filter out errors in the data. Another process included filtering and evaluating the data to ensure the data is representative and is at the required level of quality. One of the limitations highlighted from the experience of Estonia from using mobile phone data is on the sampling. Unlike with survey data, when using mobile phone data these is no overview of the total population (Kroon and Pank, 2012).

The Italian National Institute of Statistics (Istat) underwent an experiment by using mobile phone data to create high quality statistics for tourism, this experiment was done by comparing the mobile phone data with its current tourism survey to analyse if the mobile phone data could complement and enrich the tourism survey data. The mobile phone data was sourced from Vodafone analytics technology which collects data on mobile phone network activity. For the experiment Istat only focused on the areas that had strong tourist vocation and ended up selecting the province of Rimini and the municipality of Roma. The mobile phone data was processed in such a way as to ensure that the data collected fits into the international definition of tourism and ensuring that the processes that the mobile data would go through would meet the requirements of official statistics. Following this exercise, Istat found that there is potential in using mobile phone data in tourism statistics and that this data would complement the official tourism statistics. The advantage that the mobile data has over the official tourism data is that the mobile data allows for more timeliness and granularity of the statistics as compared to the official data source (Cavallo et al., 2022).

A few other big data sources have been used by NSIs in the compilation of their official statistics. The National Bank of Moldova compiles its residential property price indices using website data and land registry data. The data for the residential property price indices is sourced from the website of a well-known real estate listing company. The data is sources automatically twice a month through web scraping of both primary and secondary markets. The weights used for the primary and secondary market properties comes from the land registry (Statistica Moldovei, 2023).

The intensions of NSIs are not to replace traditional data sources with big data sources however, big data sources should be considered as an opportunity to supplement, in whole or partly information collected through the traditional data sources (United Nations, 2015).

## 2.4 Opportunities and Challenges of Big Data in Official Statistics

### 2.4.1 Opportunities

During discussions on the use of big data sources in official statistics the question on the opportunities and challenges is one that always comes up. NSIs have discovered some opportunities from exploring the use of big data in the statistical production processes. Opportunities that have been associated with the use of big data for compiling official statistics include timeliness, granularity, high frequency, and respondent burden reduction.

The big opportunity that big data sources will provide is their ability to improve official statistics either through complementing or verify existing datasets. For some datasets, big data source would allow for improvement of existing datasets by providing more granular data, these new data sources could even replace some datasets (Kitchin, 2015). Abbas et al. (2023) adds that the new data sources could aid in restructuring existing datasets and provide more accurate and detailed statistics.

Due to the nature of some of the big data sources, the data may provide more insights from the published statistics where currently it is not possible because of the standard traditional official statistics data sources (Hand, 2018). With the high variety offered by big data sources, frequent and official multidimensional statistics could become available (Abbas et al., 2023). By using big data sources this will also allow policymakers to answer specific, unforeseen questions that are distinctive to a particular crisis due to larger availability of data (Cajner et al.,2021). With the increase in large new data sources becoming available, NSIs have found that these data sources could be of use for improving and creating new official statistics (Lohr & Raghunathan, 2017).

The use of big data will allow for the timely release of statistics that would be reflected only later in statistics released by the statistical organisation through their current processes (Hand, 2018). Eurostat (2014) complements this and adds that these new data sources could potentially be an opportunity to produce more timely official statistics while also reducing the time it takes to compile the statistics on a continuous basis. More opportunities include real-time awareness and real-time feedback which is important for policy making decisions, the importance of real-time data was highlighted during the Covid-19 pandemic as policy makers needed the statistics to make informed policy decisions (Kitchin, 2015). In countries with limited resources use of traditional surveys is viewed cumbersome for both respondents and producers of official statistics, big data could be used for filling data gaps in official statistics without adding burden of survey respondents (Global Pulse, 2012).

Apart from the opportunities provided using big data for official statistics, various challenges have also been identified. Although new big data sources are available and advantages have been associated with these new data sources, there are some challenges that can be identified with incorporating these data sources in the production of official statistics.

### 2.4.2 Challenges

In most countries NSIs have a legislative backing which allows then to collect data from public authorities, this is not the case with private organisations. Depending on the country specific law, private organisations are not obligated to share data with NSIs. This could be a challenge as most of the new data is collected by private institutions (UNECE, 2013). There is therefore a need for a legal

framework for these new type of data sources, the framework should cover regulations with regards to accessing and sharing of the data. This may require an amendment to existing legal frameworks or the replacement of the frameworks. According to ESCAP (2022) the legal framework should involve establishing a suitable partnership with the private companies that own the data. Struijs et al. (2014) mentions copyrights and ownership are also some other legal aspects that need to be considered, the data could be made available, but it could be inappropriate to use.

According to Badiee et al. (2017), the world is undergoing a "data revolution," which is resulting in advancements in how data is handled, exchanged, merged, analysed, and even broadcasted. Given the large dimension of these new data sources, important technological issues arise and NSIs will need the correct IT infrastructure to be able to accommodate new dimensions such as storage of data and processing of data (Virgillito et al., 2012). Being able to make use of Application Programme Interfaces (APIs) to receive data in real time, API will be useful for large datasets were sending details between parties in the traditional was will not be feasible (UNECE, 2013).

One of the biggest challenges that NSIs will encounter in the use of big data relates to human resources. The use of big data in the production process will be time consuming, and the use of big data will require staff that are trained, and this will a management and finance challenge for the management of NSIs (UNECE, 2013). Ensuring that there are organisational and management models in place and enough qualified employees (United Nations, 2014). This data revolution is something that cannot be ignored but rather embraced, but will require knowledge development, skills enhancement, and technological improvements within national statistical systems. the development of a national training system benefits from statistical education and practical experience (United Nations, 2014). Most NSIs traditional skill set do not include the skills and competences required to automate, analyse, and optimise such complex systems needed when working with big data (Ashofteh and Bravo, 2021). Struijs et.al. (2014) suggests that NSIs will need to open to collaborate with other NSIs that have knowledge and experience in using big data in their processes or have experimented with big data. NSIs could also reach out to commercial organisations such as Google and Facebook that have such human resources, to get advice from these organisations that have extensive knowledge in areas of IT services, security, storage, and cloud solutions. Upadhyaya and Kynclova (2017) adds that NSIs should also look to create partnerships with academia, this partnership should also include academic institutions creating programmes that would assist the current and future official statistics in data science.

Costs of acquiring the data could also be a management problem as costs from sourcing the data from a private source could be costly therefore, agreements need to be established (ESCAP, 2022). The cost of acquiring the data from a private organisation could provide to be a challenge. If there is no legal framework for the private institution to share data with the statistical office, the private institution could charge for the data (Struijs et.al, 2014).

Most of these big data sources might not have enough metadata, this is considered one of the biggest challenges as data without metadata cannot be used (Virgillito et al., 2012). Some of the data sources such as social media data that were not initially collected for data analysis could lack important metadata such as target population, structure, and quality. Struijs et al. (2014) stresses that methodological challenges cannot be ignored as they impact the quality of the data. The conceptual and methodological framework for statistical production, would include precise conceptual definition

of variables, and metadata documentation (United Nations, 2014). The official statistics that are produced by NSIs follows strict guidelines that are based on international statistical standards, these standards are essential as they allow for international comparisons of statistics amongst countries. The standards cover definitions, classifications, and other terms however, these new data sources to not conform to these standards. Considering the differences between traditional and big data sources there will be a need to consider making adjustment to the current methodologies (Upadhyaya & Kynclova, 2017).

The UNECE (2013) defines privacy as the right of individuals to have influence on how their private information gets disclosed. Regarding the issue of privacy when considering big data being used in official statistics, Upadhyaya & Kynclova (2017) state that NSIs need to note that there are two approaches of privacy they need to consider. The first is accessing private data and only use it for what was agreed upon and, the second is ensuring data privacy by avoiding data misuse. When it comes to respecting the privacy of individuals confidential information, NSIs need to ensure they do not ruin their credibility and lose the public's trust. Given that some of the data received from the private businesses could include private personal information, the national statistical institutes will need to establish guidelines that will ensure the anonymization of personal data (United Nations, 2014). The use of big data could put greater emphasis on national statistical institutions to adhere to ethical practices (ESCAP, 2022).

Eurostat (2022) adds that other challenges include identifying these new data sources, developing new platforms, skills development and adjustments being made to the institution's legal framework. Therefore, it is required that these challenges be addressed before incorporating the data into the production of official statistics. With some NSIs having started to explore the use of big data in the compilation of their official statistics, some have noted some of the challenges that were experienced. Although the challenges differ for each institution, the characteristics are similar. Some NSIs who have been exploring the use of big data sources in the compilation of their official statistics have shared some of the challenges they experienced. The challenges experienced by the Italian National Institute of Statistics include quality, time dependence and accessibility. According to Statistics Netherlands they experienced challenges with methodology, quality, privacy and legal issues, processing, storage of large datasets and volatility. The Austrian Bureau of Statistics experienced similar challenges and noted their challenges as privacy and public trust, data ownership and access, computational efficacy, and technological infrastructure (ESCAP, 2022).

## 2.5 The Future of Official Statistics

According to Hassani and MacFeely (2023) the future of official statistics will be faced with many opportunities and challenges due to the evolution of digital technologies and the consequent data deluge. It will be of importance to ensure that official statistics are strengthen through effective data governance to ensure reliable and quality statistics. For official statistics to remain a significant source of information in a future that will have an abundance of data sources and information sources, official statistics will need to ensure that they continue to provide authoritative and trustworthy information.

Tille et al. (2022) argues that the evolution and availability of new data source will not bring an end to the use of current data sources. Censuses and surveys as well as administrative data are trusted data sources in the world of official statistics and many NSIs will continue to use these data sources in the

future. According to Yung (2021) in future NSIs will need to ensure that they remain relevant, and this may include embracing alternative data sources. Currently the declining survey response rates are a challenge to NSIs, this together with the calls to reduce respondents' burden, NSIs will have to investigate alternative data sources.

Groshen (2021) is of the view that the future of official statistics is at a crossroad, the one path which involves the use of newly available big data could lead to more relevant, interoperable, granular, and timely statistics. This first path could lead to a future where NSIs achieve their goals while also adding a wealth of new statistics without adding burden to survey respondents. The second path is a bleak one where survey response rates continue to fall, and NSIs are left with a problem of data quality and thus jeopardising the trust of the public in official statistics. Yung (2021) disagrees with the notion of a bleak future for official statistics however, sees a very bright future for official statistics.

MacFeely (2016) adds that official statistics are confronted by the disruption caused by the growing number of data sources currently becoming available, thus a need for adaption may be warranted. Official statistics find themselves in a changing landscape and one that is rapidly changing. Today's digital world that is depended on technology, social media and electronic transactions is one that is leaving many digital prints. This availability is therefore challenging the monopoly that NSIs once had with regards to official statistics.

Yung (2021) adds that NSIs will continue to play a pivotal role in the national statistical system, NSIs will need to move with the times of the digital world. If NSIs to do not adapt to the new digital world they will lose to other organisations in the private sector. Private companies have also started to disseminate some statistics and the statistics they produce is at times released before NSIs can publish their official statistics. According to MacFeely (2016) these companies are disseminating statistics so that they can attract media attention or attention of its browsers. MacFeely highlights that ordinary uses at times are not able to differentiate what is official statistics and what is not, therefore NSIs find themselves having to differentiate the official statistics by highlighting the quality of official statistics that comes from methodological quality and transparency of data.

The growth in digital data sources such as sensor data, transactional data and social media data will pose a challenge for official statistics in future. Hassani and MacFeely (2023) caution that if statistical institutions are to leverage off this data, they need to adapt new methodologies, create infrastructure, and have good data governance practices in place to ensure data quality, privacy protection and data integration. Yung (2021) adds that as NSIs adapt to the digital data world they should not forget to continue to follow the principles and professional standards that they currently follow. In considering these new data sources for official statistics in the future is will also be important that NSI who plan to incorporate these new data sources into their production systems consider the sustainability of the data sources they choose. Bosch et al. (2018) argues and emphasises that some of these big data sources are not suitable for official statistics as they are not sustainable, and this is something that needs to be considered when thinking about the future of official statistics.

# CHAPTER THREE

# Research Methodology

## 3.1 The study of a science

A research field or a science is developed through the creation of scientific and technical knowledge that is written and shared. The knowledge of a research field is shared through publications which indicate a standard that specific research went through by being assessed, approved, and disseminated. Published research can be found in scholar journals, proceedings, patents, and books. Over the years the pace in which research has been published has increased. Thanks to the existence of bibliometric databases which index millions of publications, it has become possible and easier to evaluate a large volume of research in order to produce new knowledge on various topics. Bibliometric analysis makes use of metadata such as titles, abstracts, and keywords of publications retrieved from bibliometric databases to reveal the structure of the research field. The different types of metadata can reveal some specific information about the research field, the information on the authors could identify communities and lead researchers while the keywords can be used to uncover concepts in a research field (Glanzel et al., 2019). This chapter will present the types of bibliometric techniques that will be used in this study to identify the trends and study the evolution in the topic of big data sources in the compilation of official statistics. This chapter will include an introduction to bibliometrics analysis, the bibliometric techniques that will be used in this study and the data collected for the study.

## 3.2 Bibliometrics analysis

Bibliometric analysis is a quantitative method used for analysing large amounts of information to identify main contributors, trends, themes and shifts in a specific field by presenting a big picture (Aria & Cuccurullo, 2017). Bibliometrics is an analysis technique that makes use of quantitative analysis to measure patterns seen in a research field of study. Bibliometric analysis is used to analyse research specific literature by making use of bibliometric data to provide insight into a fields evolution nuances and to identify emerging areas in that field. Bibliometric analysis makes use of bibliometric data that is collected and uses this data to allow for statistical analysis of the numeric data on published material. Bibliometric analysis provides an overview of the developments of a specific field and can be used to identify influential publications, authors, organisations, and countries (Rodriguez et.al, 2020; Ellegaard & Wallin, 2015; Lee et al., 2020; Manley, 2011). Bibliometrics combines quantitative analysis and statistics and applies it to publications such as journal articles together with their citation information. In recent years automated workflows have been created to aid in the complex that are entailed in making a bibliometric analysis (Leung et.al, 2017).

### 3.2.1 Bibliometrics analysis techniques

Two categories of bibliometric analysis techniques exist, the first is performance analysis. Performance analysis is one of the commonly used analysis techniques in bibliometric studies, this analysis is descriptive in nature. This technique is used to evaluate the productivity and popularity of authors, documents, and sources. The performance analysis technique examines the contributions made by research constituents of the research field and will provide an overview of the top productive authors, papers, countries, and keywords (Aria et al., 2022; Donthu et al. ,2021; Cucari et al., 2023).

The second technique, science mapping is focused on relationships found between research documents in a specific field. Science mapping is a technique that examines the relationships between disciplines, subjects, published documents, and authors (Cobo et al., 2011; Donthu et al. 2021). Science mapping which can also be referred to as bibliometric mapping is a technique which provides a representation of how disciplines, fields, specialities, and documents in a specific field relate to one another (Trybulska et al., 2017). Aria & Cuccurullo (2022) add that science mapping is a discipline that visualizes the structure and relations of a specific field.

This study will make use of both bibliometric analysis techniques, the performance analysis techniques will be used to provide the descriptive analysis of the collected bibliometric data. The results from the performance analysis technique will also be used to identify the productivity of the research field over the years and identify leading countries in the field. Science mapping will be used to analyse the impact that the collected documents have through the examination of citation analysis results, identify the conceptual structure by examining leading words and topics in the field, to identify the social structure to examine and analyse collaborations happening between countries contributing to this field.

## 3.3 Bibliometric analysis methods

This section of the chapter will be on the analysis methods that will be adopted in this study.

### 3.3.1 Citation analysis

Citation analysis is a bibliometric analysis method that measures influence. From this analysis the results will provide information regarding the most influential authors, documents, and journals. Citation analysis is a method which identifies links that exist in research using the same references. This method is also used to measure the amount of integration within a specific research field (Oberg, 2023; Zupic & Cater, 2015; Aria & Cuccurullo, 2022). Citation analysis is concerned with the total citation count of the authors, documents, and sources.

### 3.3.2 Co-word analysis

Co-word analysis which is used to create the conceptual structure which identifies links and the co-occurrence of words amongst the selected data. Co-word analysis is applied to identify the types of topics or words that are used in the literature of a specific topic (Aria & Cuccurullo, 2022). Chiang et.al (2022) add that co-word analysis identifies the degree of bibliometric relationship amongst selected literature by identifying keywords that co-exist. Co-word analysis makes use of text found in the titles, abstracts, and key words of the data to create a semantic map of the literature.

To identify the conceptual structure of a research topic, co-word analysis is used. Co-word analysis aims to firstly, construct the research domain or topics from common words found in the documents from the bibliometric data. Secondly, to highlight the most frequent topics that are covered and to also track trends which allows for the study of topic evolutions (Benomar et al., 2022; Aria & Cuccurullo, 2022). To identify the conceptual structure for this field of research the following bibliometric analysis methods will be used, thematic map analysis and thematic evolution analysis.

Thematic maps are created through the analysis that is done when keywords in different documents are analysed in order to create links between co-occurring words in order to identify main topics of a research field. The thematic map consists of bubbles that represent clusters which are made up of keywords that have the highest occurrence value. Each bubble will have the top keywords, the size of the bubble is relative to the number of occurrences of each word. The thematic map is analysed by assessing the position of the bubble, the bubble is positioned based on the density and centrality of the topics (Khare & Jain, 2022; Aria et al., 2022; Agbo et al., 2021).

Thematic maps are a great tool to use to analyse key themes or topics based on centrality and density however, for a more granular analysis on the chronological evolution of the topics then thematic evolution maps are more appropriate. Thematic evolution analysis is an approach that uses a combination of performance analysis and science mapping to detect and visualize general thematic areas and allows for quantification and visualization of a research fields thematic evolution. This method makes use of co-word analysis in a longitudinal mapping framework (Garfield, 1994; Cobo et al., 2011). The longitudinal thematic map shows how topics or themes have emerged, disappeared, merged, or even reappeared over time (Khare & Jain, 2022).

#### 3.3.3 Co-authorship analysis

Co-authorship is an analysis method that identifies the social structure of a specific research field. Co-authorship analysis is a method that analyses the collaborations between authors and countries, this method is important as it can be used to identify the collaborations between authors and countries in order to understand what kind of interactions are taking pace in the research world of the research field (Donthu et al. 2021; Fonseca et al., 2016).

### 3.4 Data collection

The data that will be used in the analysis for this study is retrieved from the Scopus database. The Scopus database is owned by Elsevier and has existed since 2004, this database is considered to have a high bibliometric coverage with over 1.7 billion cited references that go as far back as the 1970s (Zupic & Cater, 2015; Suominen & Hajikhani, 2021). The Scopus database has advantages such as providing a wider coverage compared to other databases, the database is supported by the most used bibliometric software programs including the Bibliometrix R package which will be used in this study. Scopus is also considered to be advantageous for making author-based citations and co-citation as the database contains all data from the cited references. Weidt (2011) adds that the Scopus database is sophisticated for citation analysis and for other research strategies.

The data that will be used in the study is collected following four steps as represented in figure 1. The first step entails querying the Scopus database and retrieving data, specific keywords are used to filter the database by searching the article title, abstract and keywords fields. Trawick and McEntyre (2003) advises that when conducting a query on a bibliographic database it is crucial to use concise and clear search terms as this will help is limiting the scope of the study as well as to ensure that core documents are included in the selected data.

The scope of this study is a narrow one therefore, in collecting the data it is essential to ensure that the focus of the documents that are going to be used in the analysis focus solely of the use of big data sources in the compilation of official statistics. To focus and limit the scope of the search an iterative approach is used to query the Scopus database where a string of words is used to search the database. In the search on the database the word “official statistics” is the commonly used term and in each

iteration the words are used together with official statistics. For the analysis of this study, the data collected is restricted to only research articles published between the years 2010 and 2023. Following a few iterations, the total number of exported metadata is 285 documents. These documents were the results from queries using “official statistics”, “big data”, “non-traditional data sources”, “new data sources”, “mobile data”, and “transactional data”.

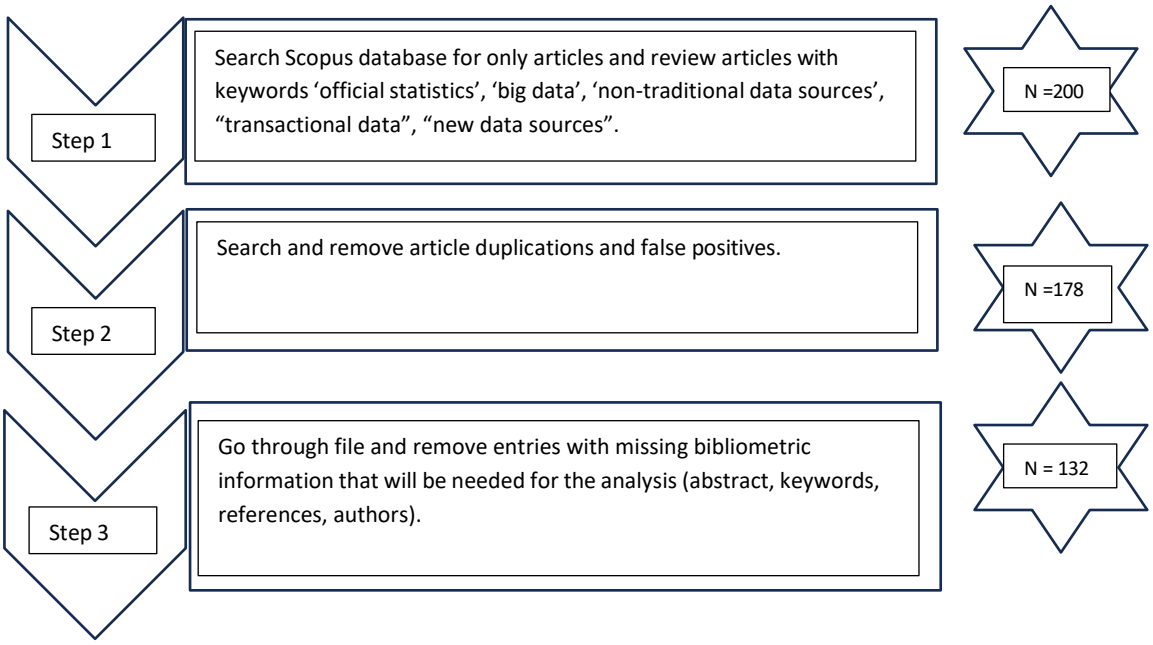


Figure 1: Steps followed in the data collection stage

From the exported data, article data that does not include author name(s), document title, year, source title, affiliations, abstract, author keywords, and references are removed. The mentioned fields are important for the bibliometric analysis therefore, it is important to get complete information so that the analysis can give accurate results. Following this step there is a total of 132 article data that will be used in the analysis for this study.

# CHAPTER FOUR

# ANALYSIS AND DISCUSSION OF RESULTS

## 4.1 An overview on published research articles on big data in official statistics

Table 1 provides a summary of the collected data on research articles published on the use of big data in official statistics. Although the Scopus search filtered for articles published from 2010, the first article published in this research field was published in 2014 therefore the analysis that will be carried out will be on articles published between the years 2014 – 2023. Within this period a total of 132 articles have been published that will be included in the analysis of this study. These 132 articles have been published across 58 source publications and have been written by 352 authors coming from NSIs and academia. The literature review mentioned the importance of collaboration between NSIs, the results of the summary analysis show that there is an international co-authorship of 21.21%. This shows that a majority of the articles are published by authors in the same countries.

Table 1: Data analysis

| Data Analysis | |
|---|---|
| Timespan | 2014 - 2023 |
| Sources | 58 |
| Articles | 132 |
| Annual Growth Rate % | 39.5 |
| Average citations per doc | 8.091 |
| References | 4670 |
| Total Keywords | 468 |
| Number of Authors | 352 |
| International co-authorships % | 21.21 |

Figure 2 shows the production of the articles published per year, a single article was published in 2014 which can be marked as the starting point on the research on the topic of big data and official statistics. This article relates some of the reasons why NSIs and international statistical organisations have ventured into investigating the possibilities of the use of big data sources in the compilation of official statistics. Following 2014 an upward trend can be seen starting from 2015 and where the number of published articles enters double digits starting in 2018. This upward trend in the number of published articles comes after the 45$^{th}$ session meeting held by the United Nations Statistical Commission, within this meeting the Statistical Commission highlighted that big data cannot be ignored as a possible source of information that can be used in official statistics.

The outcome of this meeting lead to the formulation of the Global Working Group on Big Data for Official Statistics. The mandate for this working group is based on strategic considerations that relate to the post-2015 development agenda, the data revolution initiative, and the Fundamental Principles of Official Statistics. Other such working groups were created within the official statistics community, the Conference on European Statisticians established the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) which is tasked with advancing the modernisation of official statistics.

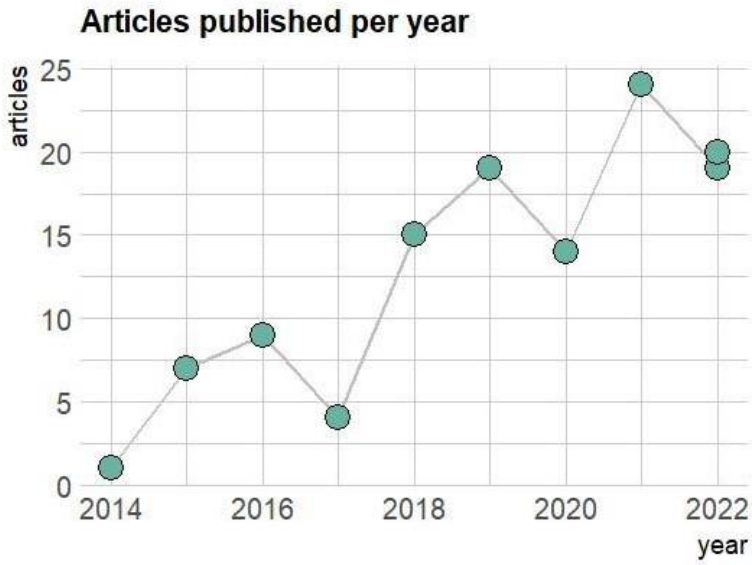


Figure 2: Published articles per year

A three-field plot representing the connections between the top ten countries (AU) left field, keywords (DE) middle field, and sources (SO) right field is shown in figure 3, the figure shows how these three fields relate to each other through a Sankey diagram. Each of the top ten authors, keywords, and sources are represented by the rectangular nodes where the height of the rectangle reflects the proportion of the frequency of occurrence of the author, keyword, and source.

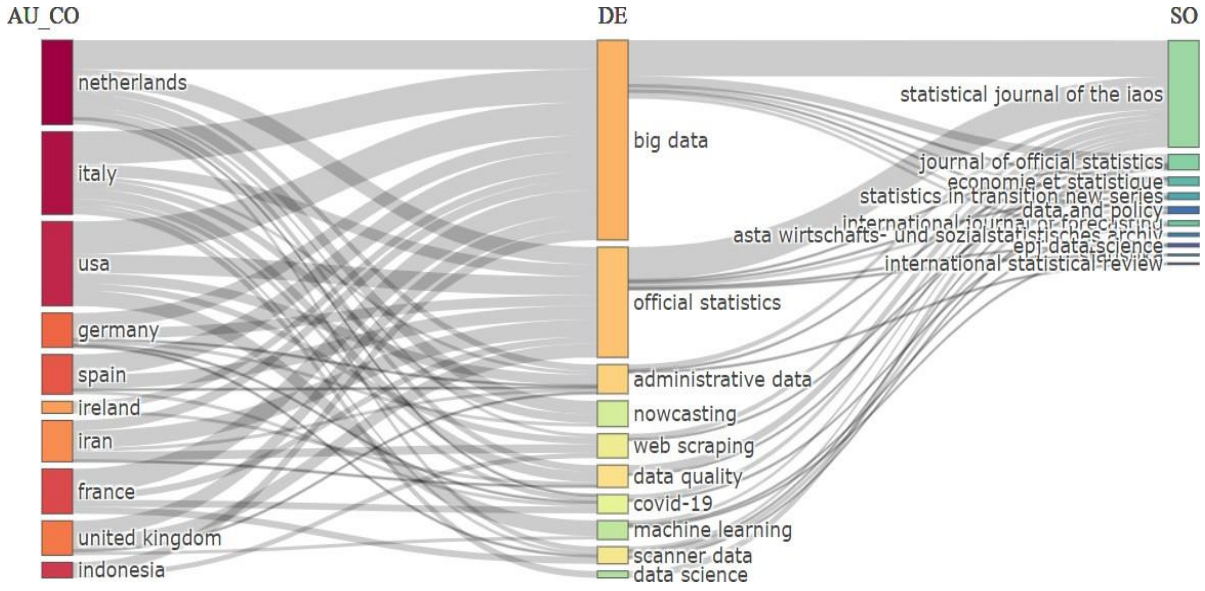


Figure3: Three-field plot

From the 468 keywords from all the articles, the top ten are found in the middle field of figure 3. As expected, big data and official statistics are the two most occurring keywords showing that most of the articles directly refer to these two words. The Netherlands, Italy, and the USA are the top three countries which have published articles with the keywords big data and official statistics where most of these articles are published in the Statistical Journal of the IAOS. Data quality is also one of the most occurring words within the research field, this is expected as data quality is one of the principles in the Fundamental Principles of Official Statistics. The quality of the big data sources is very important as it will have an impact on whether NSIs will be able to use the source or not therefore, data quality is also a very important terms within this research field. The other key words can be associated with the different avenues that are being explored in the research on big data and official statistics, these words also show the focus areas of the working groups on big data and official statistics. Figure 3 shows in which topic each country that is contributing to the research field is following.

## 4.2 Performance analysis

This section will look at the major contributors to the research field.

### 4.2.1 Top publication sources

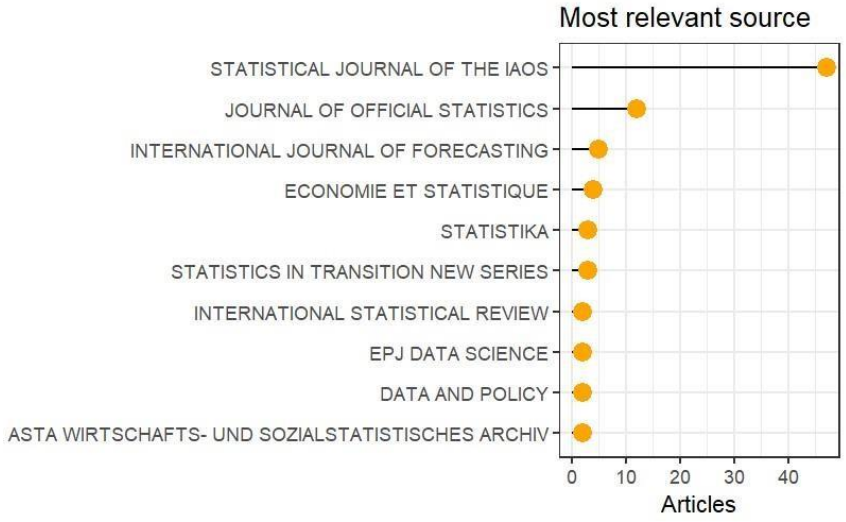


Figure 4: Most relevant source

Figure 4 shows the top ten journals in which the research articles appear. The top two journals have to do with official statistics while the rest cover topics in statistics and economics. The top two journals are the Statistical Journal of the IAOS with 47 publications and the Journal of Official Statistics with 12 publications, these two journals are the publication source of 44.6% of all the published articles. The aim of the Statistical Journal of IAOS is supporting the mission of the International Association of Official Statistics (IAOS) by publishing articles that contribute to the understanding and advancement of official statistics (IOS Press, 2023). The Journal of Official Statistics is a peer-reviewed journal that is published by Statistics Sweden and this journal publishes research articles that focuses on survey methodology and the production of official statistics (Statistics Sweden, 2023).

From the 58 different journals, these two journals are the only ones that have more than ten publications each. It should come as no surprise that these are the top two journals, given that most

of the authors of these articles are in the official statistics community therefore these journals are the leading two in this field.

#### 4.2.2 Most relevant authors

The most relevant author is measured by the number of articles published by each author. Daas P (Statistics Netherlands) and Buelens B (Flemish Institute for Technological Research) each with 5 articles are ranked as the two most relevant authors. Pramana S (Statistics Indonesia) and Saebo HV (Statistics Norway) are the next two authors in the list who make it into the top five with 4 articles each. These four authors De Broe S (Statistics Netherlands), Scannapieco M (Italian National Institute of Statistics), Ten Bosch O (Statistics Netherlands), Van Den Brakel J (Statistics Netherlands) come in with a tied 3 articles each. The top ten authors are made up of seven authors from Statistical Offices of which four are from Statistics Netherlands and three from Academic institutions. The dominance of the research done by Statistics Netherlands is evident from these results, this could be as a result of the Center for Big Data Statistics (CBDS) which is an initiative started by Statistics Netherlands in 2016. Statistics Netherlands is one of the leading players in the use of big data in the compilation of official statistics and being the first statistical office to publish traffic statistics using big data. The CBDS is in partnership with PhD students, university students and experts associated with various national and international parties including NSIs such as Statistics Korea and the Italian National Institute of Statistics.

#### 4.2.3 Most relevant countries

Figure 5 provides a snapshot on the international distribution of the knowledge domain of the research field. There are 31 countries that have contributed to the research on big data and official statistics. The Netherlands has the most published articles (14 articles) followed by Italy (12 articles) and the USA (11 articles). The articles in this research field span over 6 continents with Europe (75 articles) being the continent where most authors are established, accounting for over 56% of the published articles, followed by the Americas (14 articles), and Asia (11 articles). When analysing the international co-authorship, the Netherlands is the leading country in terms of international co-authorship with 5 articles followed by the USA with 2 international co-authorship articles.

#### 4.2.4 Most frequent keywords

Figure 6 shows a word tree with the most common occurring keywords from the sample of articles. Big data and official statistics are the top common occurring words highlighting the topic of this research field. Other words in the word tree identify the areas of interest in terms of research when it comes to big data and official statistics. The words administrative data and data quality can be linked to the word official statistics, administrative data is one of the traditional data sources that have and are being used in the compilation of official statistics. Data quality can be linked to official statistics however, the research is also linking it to the word big data as quality is a very important component when referring to official statistics.

It is interesting to note that the common keywords are linked to the work covered by the working groups on big data and official statistics, this shows that these working groups are setting the direction in which the research on big data and official statistics takes. This again highlights the aim of these

working groups in that they are trying to assist NSIs on how to approach and provide direction on the use of big data in official statistics.

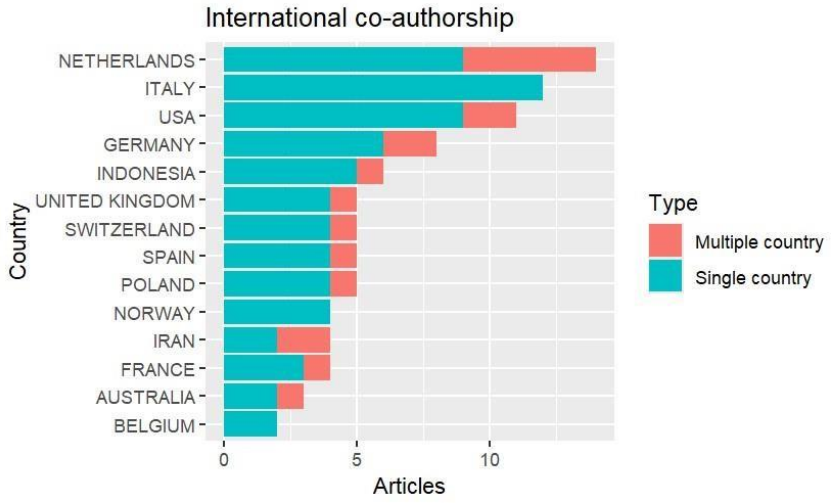


Figure 5: International co-authorship

Scanner data and web-scraping are examples of the two types of big data sources that have been used in experimental official statistics, these data sources were part of the initial topics that working groups on Big Data for Official Statistics was focused on back in 2014. The Working Group on Big Data for Official Statistics had a task team on scanner *data* whose objectives included, providing guidance of the use of alternative data sources to compile consumer price statistics, to advise on who to classify scanner data so that it can be used in the price index compilation, and to provide training and build capacity on using these new data sources to compile consumer prices.

A few countries including Australia, Japan, and New Zealand to name a few tested compiling their national CPI statistics using scanner data. Scanner data was also a data source that was investigated and part of one of the projects by the Sandbox in 2014, the Sandbox also had other projects relating to social media data, mobile phone data, traffic loops data.

NSIs that have ventured into exploring big data for the compilation of their official statistics have done so looking into the data sources that were highlighted by the working groups. Most of the topics of the research seem to be led by the work and recommendations that are coming from the working groups.


Figure 6: Most common keywords

## 4.3 Impact and conceptual knowledge

### 4.3.1 Performance vs impact

The previous section focused on the most influential players in the research on big data and official statistics, here we go further by not only looking at output (publications) but also impact (citations). The top ten based on citation is not only analysed based on the number of total publications but is based on the highest h-index. The h-index was introduced by Hirsch (2005), this index is calculated by taking the authors, document, or sources total published articles (h) and counting how many have been cited at least by other researchers (Donthu et al., 2021; Agbo et al., 2021).

Table 2 shows the top ten authors according to publications and citations. Note the difference in the ranking of the authors list when sorted by number of published articles (left-hand side) and total citation (right-hand side). With regards to the top ten according to the number of publications, these authors come from a concentrated number of countries, with only 5 countries represented. The same is seen on the right-hand side where the top ten authors represent 4 countries. Netherlands remains the dominant country in terms of the top ten authors for both top ten based on publication and citation.

### 4.3.2 Top published articles

The top 15 most cited articles that have contributed to the research on big data and official statistics. The geography of these articles is included in 7 countries, most of these articles are co-authorships and the affiliations of the authors are split between NSIs and Academia. 9 of the articles are from single country authors and the rest from multiple country authors.  The content of these articles can be

broadly clustered into two groups. The articles published before 2018 are the research of whether big data can be a data source for official statistics while, from 2018 onwards the research goes into the

Table 2: Top ten authors

| Top ten publications | | Top ten citations | | |
|---|---|---|---|---|
| **Author** | **Articles** | **Author** | **H-index** | **Total citation** |
| Bart Buelens | 5 | Bart Buelens | 4 | 95 |
| Piet Daas | 5 | Piet Daas | 3 | 95 |
| Setia Pramana | 4 | Sofie De Broe | 2 | 8 |
| Hans Viggo | | Ben laevens | 2 | 6 |
| Sofie De Broe | 3 | Setia Pranama | 2 | 8 |
| Monica scannapieco | 3 | Hans Viggo | 2 | 7 |
| Olav Ten Bosch | 3 | Siu-Ming Tam | 2 | 48 |
| Jan Van Den Brakel | 3 | Olav ten Bosch | 2 | 8 |
| Syed wasim Abbas | 2 | Arnout van Delden | 2 | 21 |
| Emanuele Baldacci | 2 | Jan Van Den Brakel | 2 | 24 |

actual big data sources and how these sources could produce official statistics. The research from 2018 can be considered as the start of research that could produce experimental official statistics.

Table 3: Top cited articles

| Author(s) name | Title | Year | Total Citation | Main Country | Country | Affiliation |
|---|---|---|---|---|---|---|
| D'Amuri, F; Marcucci, J | The predictive power of Google searches in forecasting US unemployment | 2017 | 134 | Italy | Italy | Bank of Italy |
| Fatehkia, M; Kashyap, R; Weber, I | Using Facebook ad data to track the global digital gender gap | 2018 | 89 | USA | USA,UK,Qatar | Princeton University, University of Oxford, Qatar Computing Research Institute |
| Batini, C; Rula, A; Scannapieco, M; Viscusi, G | From data quality to big data quality | 2015 | 78 | Italy | Italy, Switerzerland | University of Milano-Bicocca, Istat |
| Daas, P; Puts, M; Buelens, B; van den Hurk, P | Big data as a source for official statistics | 2015 | 68 | Netherlands | Netherlands | Statistics Netherlands |
| Naccarato, A; Falorsi, S; Loriga, S; Pierini, A | Combining official and Google Trends data to forecast the Italian youth unemployment rate | 2018 | 50 | Italy | Italy | ISTAT - Italian National Institute of Statistics |
| Citro, C | From multiple modes for surveys to multiple data sources for estimates | 2014 | 46 | USA | USA | Committee on National Statistics |
| Tam, S; Clarke, F | Big data, official statistics and some initiatives by the Australian Bureau of statistics | 2015 | 44 | Australia | Australia | Australian Bureau of Statistics |
| Carroll, P; Murphy, T; Hanley, M; Dempsey, D; Dunne, J | Household Classification Using Smart Meter Data | 2018 | 41 | Ireland | Ireland | University College Dublin, Central Statistics Office |
| Severo, M; Feredj, A; Romele, A | Soft Data and Public Policy: Can Social Media Offer Alternatives to Official Statistics in Urban Policymaking? | 2016 | 30 | France | France, Portugal | UniversitÃ Paris, Universidade do Porto |
| Di Bella, E; Leporatti, L; Maggino, F | Big Data and Social Indicators: Actual Trends and New Perspectives | 2018 | 27 | Italy | Italy | University of Genoa, University of Florence |
| Jahani, E; Sundsa, P; Bjelland, J; Bengtsson, L; Pentland, A; De Montjoye, Y | Improving official statistics in emerging markets using machine learning and mobile phone data | 2017 | 23 | USA | USA, Norway, Stockholm | Massachusetts Institute of Technology, Telenor Research, Karolinska Institute |
| GÃtz, T; Knetsch, T | Google data in bridge equation models for German GDP | 2019 | 23 | Germany | Germany | Deutsche Bundesbank |
| De Waal, T; van Delden, A; Scholtus, S | Multi-source Statistics: Basic Situations and Methods | 2020 | 19 | Netherlands | Netherlands | Statistics Netherlands |
| Van Den Brakel, J; Sähler, E; Daas, P; Buelens, B | Social media as a data source for official statistics; the dutch consumer confidence index | 2017 | 17 | Netherlands | Netherlands | Statistics Netherlands, Maastricht University |
| Geiger, C; Frosio, G; Bulayenko, O | Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?: Legal Analysis and Policy Recommendations | 2018 | 17 | France | France, Germany | University of Strasbourg, Max Planck Institute for Innovation and Competition |

## 4.4 Conceptual structure

The section will investigate the conceptual structure of the research on big data and official statistics. The conceptual structure shows the relationship that exists between words or topics that have dominated in the set of analysed documents. These topics are grouped together based on how they occur together through co-occurrence in the documents. This is also known as the co-word network. The results below will show the most important topics in the field and the evolution of these topics over time.

### 4.4.1 Co-word network

Figure 7 illustrates the co-word network which shows the keywords that appear in all of the published articles, the connecting lines reveal the co-occurrence of concepts found in the literature. Big data and official statistics are the two enlarged and highlighted words in the network showing that these two words are the main topics of the literature, also note the thick connection line connecting the two words, this indicates that these two usually appear together in the published literature. More terms are shown around these two words where the distance between the words shows the co-occurrence of these words, the co-word network has shown the different clusters in which these terms fall. The big data cluster underlines the topics that relate to the big data sources (twitter, mobile phones, web scraping) and how these data sources can be used in official statistics (nowcasting, experimental statistics, CPI, sustainable development goals..). The official statistics cluster more on the broad concepts that need to be considered when viewing big data in the space of official statistics (data quality, framework, challenges, innovation..).

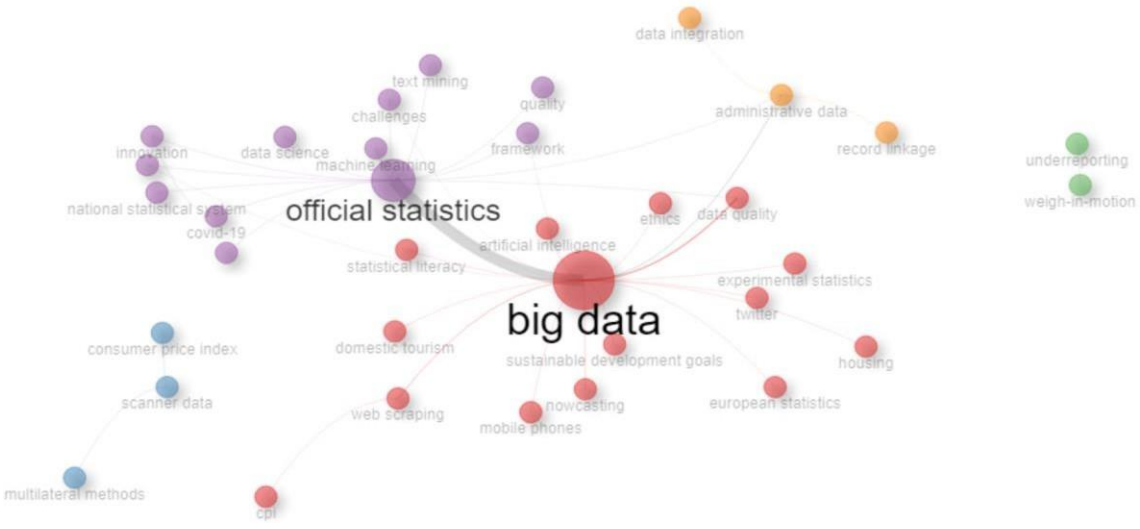


Figure 7: Co-word network

### 4.4.2 Thematic evolution

A more granular thematic map is one that analyses the themes at different points in time, below figure 8 shows the thematic evolution map which will show the main themes of the research field at different periods to better analyse how themes and topics have emerged or developed over time. For this study

the time periods range from 2014 to 2023 has been split into three periods, 2014 – 2017, 2018 – 2021 and 2022 – 2023.

In the first period from 2014 – 2017 big data, data quality and official statistics are identified as the three main topics during this period. This first period is at the initial starting years of the research field, these topics could therefore be considered as the foundation topics with regards to big data and official statistics. The first two words describe the basis of the research field therefore it is expected that these will remain relevant topics. The keyword data quality was very significant during, data quality is always a word to constantly comes up in the discussions of big data and official statistics given the importance of data quality in the field of official statistics. The second period from 2018 – 2021 shows an expansion in the trending topics compared to the previous period. Scanner data was one of the first big data sources to be considered for compiling official statistics, much of the research and experimental data that was produced during this period involved scanner data. The last period does not show much change in the main topics as compared to the 2018-2021 period. However, web scraping became one of the main research topics. It should be noted that web scraping is a very broad topic in that it is a method that can be used to collect data from online, figure 8 therefore does not show the type of data sources are collected and have been research in the field. From the published articles, web scraping is being used to collect data from multiple sources including credit cards data, mobile phones data however, the majority of the research is on the use of web scraping is for the collecting of data used to compile the Consumer Price Index (CPI). To further explore the evolution of topics on big data and official statistics a different view is analysed with the use of keywords from the abstracts of the research articles.

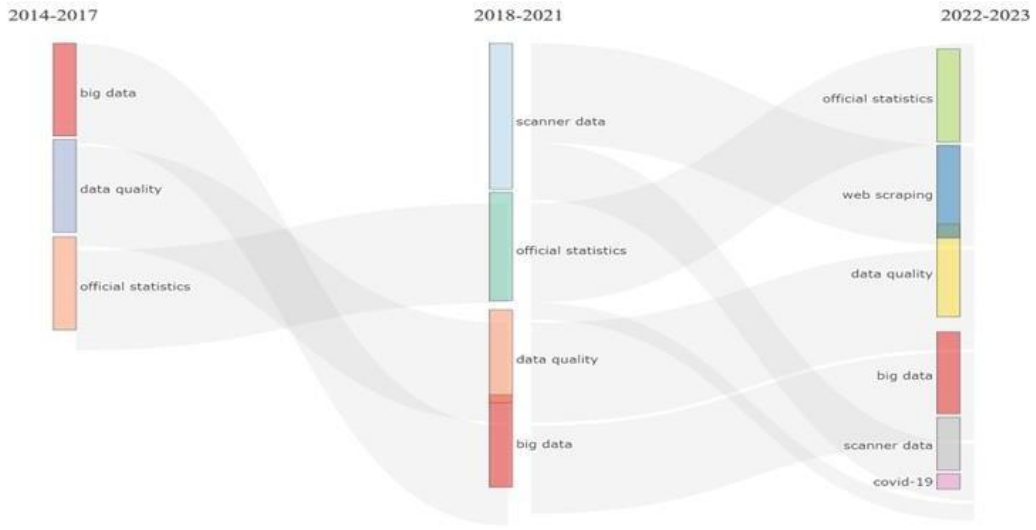


Figure 8: Thematic evolution

### 4.4.3 Keywords over time

Figure 9 analyses the content dynamics on the articles published on big data and official statistics, here a bi-gram (two words) of the most common words found in the abstracts of the different articles is analysed. The graph shows the most common abstract keywords and the evolution of the topic over

time, the dark colours represent the strength of the presence of the topic and the number of the tiles is the frequency at which the word appeared in the abstracts. The most re-occurring words in the abstracts are official statistics and data sources, these two words underline the objective of the entire research field which is broadly on data sources used for the compilation of official statistics.

When analysing the entire graph, it is possible to note that these keywords can be grouped in themes. The first will be referred to official statistics principles (quality framework, data quality, data governance), these are topics that are related to the official statistics label. These are topics that will

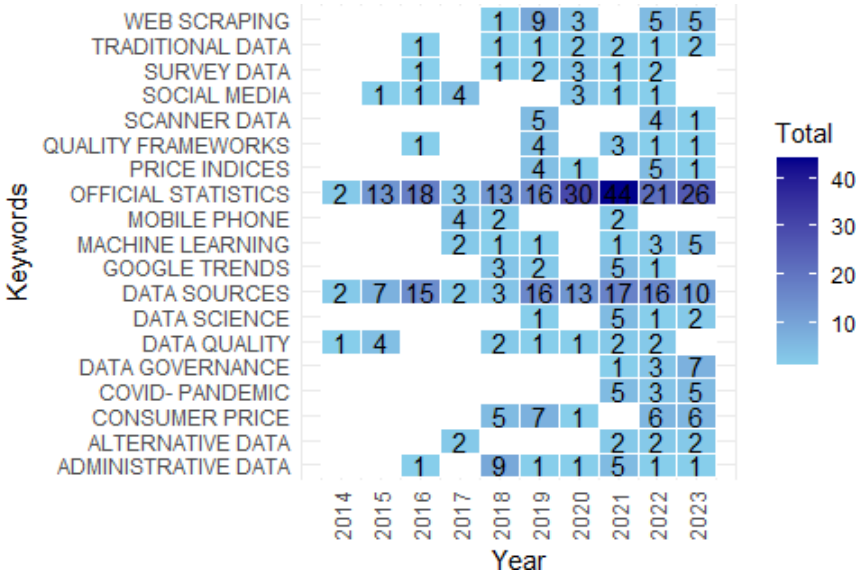


Figure 9: Keywords over time

also be included in any research that includes official statistics. The second theme is regarding the data sources (social media, scanner data, mobile phone, google trends, alternative data), this theme is on the different data sources that have thus far been tested in the compilation of official statistics. The final theme is on the technology or methods (web scraping, data science, machine learning), this theme covers topics that NSIs will need in order to be able to use big data in the context of official statistics.

# CHAPTER FIVE

# FINDINGS, CONSLUSIONS AND FUTURE WORKS

Big data has presented National Statistical Institutions (NSIs) with the opportunity to better and extend their official statistics. Even so, due to the quality label that is attached to official statistics. The official statistics community had to come together an explore ways in which big data to be applied to official statistics. From as early as 2009, international statistical organisations have created working and expert groups on big data for official statistics to facilitate collaborations between NSIs and private organisations. Many opportunities exist in this field of big data for NSI however, challenges still persist.

This study set out to examine and determine the effects of big data on official statistics and NSIs. It was revealed that the use of big data by NSIs will allow them to reap opportunities in the compilation of their official statistics such as timeliness, granularity, high frequency, and respondent burden reduction. However, NSI will have to overcome human resource, legal, technological, and methodological challenges.

The study showed evidence of the use of big data sources in the compilation of official statistics. Big data sources such as scanner data, mobile phone data, sensor data, and transactional data have been used institutions such as Statistics Netherlands, Bank Estonia, Italian National Institute of Statistics, US Census Bureau, and the Spanish National Institute of Statistics. Scanner data did seem to be the one big data source that has been used by a few institutions for the compilation of the Consumer Price Index (CPI) with Statistics Netherlands being the first to explore with this data source from 2002.

Additionally, the study identified some of the key topics relating to big data and official statistics. Big data, official statistics, nowcasting, machine learning are some of the common keywords that have been covered in this research field. When analysing the bi-gram (two-words) of the most commonly found abstract words, emerging and relevant topics include web scraping, official statistics, data sources, and data governance.

The findings highlighted that the main dominant publishing country was the Netherlands, with Italy and the United States of America also making sizeable contributions. Majority of the publications were from European countries, which highlights where most of the research done on big data and official statistics comes from.

This study had a few constraints with the first being attributed to the available bibliometric data, it is possible that more research has been done in this space however not all research is published in journals or available in bibliometric databases. The second constraint also relates to data, the only bibliometric database that was used in this study was Scopus therefore data from other databases such as Web of Science (WoS) were not investigated, this could have limitations impact if the Web of Science database had additional articles found to the ones on Scopus.

A recommended area for future research work would be on the analysis of the readiness of NSIs on the topic of big data sources in the compilation of official statistics. It will be important for all NSIs to also explore the use of big data in official statistics, given that the standards surrounding official

statistics are international more countries need to start experimenting and researching in this area. More international collaboration is also an area that needs further development in future.

## Bibliography


Abbas, S.W., Muhammad, H., Reem, A., and Hanaa, A. (2023). Official Statistics and Big Data Processing with Artificial Intelligence: Capacity Indicators for Public Sector Organizations. *Systems* 11, no. 8: 424. https://doi.org/10.3390/systems11080424

Alrousan, A. (2023). From Big Data to Big Insights: The Power and Efficiency of Big Data and Data Science in Statistical Work. B52 - Ahmad - The policy of using big data in official statistical centers to support decision-makers.pdf (un.org)

Aria M., Cuccurullo C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*. doi:10.1016/j.joi.2017.08.007.

Aria,M. & Cuccurullo, C. (2022). Science Mapping Analysis with Bibliometrix R-package: an example.

https://bibliometrix.org/documents/bibliometrix_Report.html

Ashofteh, A., & Bravo, J.M. (2021). Data Science Training for Official Statistics: a New Scientific Paradigm of Information and Knowledge Development in National Statistical Systems. Statistical Journal of the IAOS, 1-19. https://doi.org/10.3233/SJI-210841

Badiee, S. et al. (2017). The role of national statistical systems in the data revolution. Chapter 3. Development Co-Operation Report. OECD. https://doi.org/10.1787/dcr-2017-8-en

Baker, H.K., Kumar, S., & Pandey, N. (2021). Forty years of the Journal of Futures Markets: A bibliometric overview. *Journal of Futures Markets*, *41*(7), 1027-1054. https://doi.org/10.1002/fut.22211

Baldacci, E., Ricciato, F., & Wirthmann,A. (2021). A Reflection on the Re (Use) of New Data Sources for Official Statistics. Revista Indice, Número 83**.** Estadísticas en la sociedad digital. http://www.revistaindice.com/numero83/

Benomar, L., Elferjani, R., Hamilton, J. (2022). Bibliometric Analysis of the Structure and Evolution of Research on Assisted Migration. Curr Forestry Rep 8, 199–213. https://doi.org/10.1007/s40725-022-00165-y

Cajner,T., Crane,L., Decker,R., Hamins-Puertolas,A., & Kurz,C. (2020). Tracking Labor Market Developments during COVID-19 Pandemic: A Preliminary Assessment. Finance and Economics Discussion Series. Divisions of Research & Statistics and Monetary Affairs. Federal Reserve Board, Washington, D.C. https://doi.org/10.17016/FEDS.2020.030

Cajner,T., Feiveson,L., Kurz,C., & Tevlin,S. (2021). Lessons Learned from the use of Non-traditional data during COVID-19. In Endelberg,W., Sheiner,L., & Wessel,D. Recession Remedies: Lessons Learned from the U.S Economic Policy Response to COVID-19.

Chiang,LT., Kai-Yu,T.,Shu-Sheng,L.,Miao-Li,C.,Ying-Shao,H., (2022). A Co-word Analysis of Selected Science Education Literature: Identifying Research Trends of Scaffolding in Two Decades (2000–2019). Frontiers in Psychology, Sec. Educational Psychology Volume 13. https://doi.org/10.3389/fpsyg.2022.844425

Citro, C,F. (2014) From multiple modes for surveys to multiple data sources for estimates. Statistics Canada

Cobo, M. J., L´opez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics, 5*(1), 146–166. https://doi.org/10.1016/j.joi.2010.10.002

Cucari, N., Tutore, I., Montera, R., & Profita, S. (2023). A bibliometric performance analysis of publication productivity in the corporate social responsibility field: Outcomes of SciVal analytics. *Corporate Social Responsibility and Environmental Management*, 30(1), 1–16. https://doi.org/10.1002/csr.2346

Dehdarirad, T., Villarroya, A., and Barrios, M. (2014). Research trends in gender differences in higher education and science: a co-word analysis. *Scientometrics* 101, 273–290. https://doi.org/10.1007/s11192-014-1327-2

Dias, L.T. (2019). Post-crisis skills landscape: the emergence of "purple people". IFC Satellite Seminar on "Post-crisis data landscape: micro data for the macro world", volume 53, Bank for International Settlements.

Donthu,N. Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, Weng Marc Lim. (2021). How to conduct a bibliometric analysis: An overview and guidelines, Journal of Business Research, Volume 133, Pages 285-296, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2021.04.070.

Dumbacher,B., Morris,D., & Hougue,C., (2019). Using electronic transaction data to add geographic granularity to official estimates of retail sales. Journal of Big data. https://doi.org/10.1186/s40537-019-0242-z

Egghe,L .(2006). Theory and practice of the *g*-index. *Scientometrics*. 69 (1): 131–152.

https://doi.org/10.1007/s11192-006-0144-7

Ellegaard, O., Wallin, J.A. (2015). The bibliometric analysis of scholarly production: How great is the impact?. *Scientometrics* 105, 1809–1831. https://doi.org/10.1007/s11192-015-1645-z

ESCAP. (2022). The fundamental principles of official statistics and the use of mobile phone data for official statistics. United Nations Economic and Social Commission for Asia and the Pacific (ESCAP) https://www.unescap.org/sites/default/files/02_ESCAP_Fundamental_Principle_MPD_11-14Jun2019.pdf

ESCAP. (2022). Using Big Data for Official Statistics. Using Big Data for Official Statistics: Key considerations when using mobile phone data (unescap.org)

Eurostat. (2003). Methodological documents – Definition of quality statistics. Working Group , Assessment of quality in statistics. Luxembourg. European Commission. https://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf/f0fdc8d8-6a9b-48e8-a636-9a34d073410f

Eurostat. (2017). Beginners: Statistical concept-What are official statistics? https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-

Fonseca, B., Sampaio, R.B., Fonseca, M. (2016). Co-authorship network analysis in health research: method and potential use. *Health Res Policy Sys* 14, 34. https://doi.org/10.1186/s12961-016-0104-5

Ghirelli, C., Hurtado, S., Pérez, J.J., Urtasun, A. (2021). New Data Sources for Central Banks. In: Consoli, S., Reforgiato Recupero, D., Saisana, M. (eds) Data Science for Economics and Finance. Springer, Cham. https://doi.org/10.1007/978-3-030-66891-4_8

Glänzel, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators. Course Handouts.

Glanzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (Eds.). (2019). *Springer Handbook of science and technology indicators*. Berlin: Springer. https://link.springer.com/book/10.1007/978-3-030-02511-3

Global Pulse. (2012). Big data for development: Challenges and opportunities. Global Pulse, New York. https://unstats.un.org/unsd/trade/events/2014/beijing/documents/globalpulse/Big%20Data%20for%20Development%20-%20UN%20Global%20Pulse%20-%20June2012.pdf

Global Pulse. (2023). UN Global Pulse – facilitating the most creative look to the future. https://www.unglobalpulse.org/

Graf, B. (2020). Scanner Data 10. In *Consumer Price Index Manual, 2020*. USA: International Monetary Fund. Retrieved May 26, 2024, from https://doi.org/10.5089/9781484354841.069.ch010

Groshen, E.L. (2021). The Future of Official Statistics. Harvard Data Science Review, 3(4). https://doi.org/10.1162/99608f92.591917c6

Hand,D.J. (2018). Statistical challenges of administrative and transaction data. Journal of the Royal Statistical Society Series A, Royal Statistical Society, vol. 181(3), pages 555-605. https://doi.org/10.1111/rssa.12315

Hassani H, MacFeely S. (2023). Driving Excellence in Official Statistics: Unleashing the Potential of Comprehensive Digital Data Governance. *Big Data and Cognitive Computing*, 7(3):134. https://doi.org/10.3390/bdcc7030134

Hirsch J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

IOS Press. (2023). Official Statistics. Statistical journal of the IAOS. IOS Press. https://officialstatistics.com/aims-scope

Khare,A., Jain,A. (2022). Mapping the conceptual and intellectual structure of the consumer vulnerability field: A bibliometric analysis. Journal of Business Research, Volume 150,Pages 567-584, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2022.06.039.

Kitchin, R. (2015). The opportunities, challenges, and risks of big data for official statistics. Statistical Journal of the IAOS, vol. 31, no. 3, pp. 471-481. https://content.iospress.com/download/statistical-journal-of-the-iaos/sji906?id=statistical-journal-of-the-iaos%2Fsji906

Kroon, J. and Pank, E. (2012). Mobile positioning as a possible data source for international travel service statistics. United Nations Economic Commission for Europe. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP6.pdf

Lee I.S, Lee H., Chen Y.H., Chae Y. (2020). Bibliometric Analysis of Research Assessing the Use of Acupuncture for Pain Treatment Over the Past 20 Years. *J Pain Res*;13:367-376 https://doi.org/10.2147/JPR.S235047

Lohr, S,L., & Raghunathan, T,E. (2017). Combining Survey Data with Other Data Sources. Statistical Science. https://projecteuclid.org/journals/statistical-science/volume-32/issue-2/Combining-Survey-Data-with-Other-Data-Sources/10.1214/16-STS584.full

MacFeely,S.(2016).The Continuing Evolution of Official Statistics: Some Challenges and Opportunities. Journal of Official Statistics,32(4) 789-810. https://doi.org/10.1515/jos-2016-0041

Aria,M., & Cuccurullo,C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis, Journal of Informetrics, Volume 11, Issue 4, Pages 959-975, ISSN 1751-1577, https://doi.org/10.1016/j.joi.2017.08.007.

Mehrhoff,M. (2018). Demystifying big data in official statistics – it's not rocket science!. Ninth IFC Conference on Are post-crisis statistical initiatives completed. Bank for International Settlements

Oberg,C. (2023). Neuroscience in business-to-business marketing research: A literature review, co-citation analysis and research agenda. Industrial Marketing Management, Volume 113, Pages 168-179. https://doi.org/10.1016/j.indmarman.2023.06.004.

Persson, O. (1994). The Intellectual Base and Research Fronts of JASIS 1986-1990. Journal of the American Society for Information Science, 45, 31-38. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<31::AID-ASI4>3.0.CO;2-G

Pieter,. E, (2022). Standards as the Backbone for Official Statistics, How Well Do They Fit Within National and International Systems or Does the Obsession with Cross-national Comparisons Blind Us to the Weakly Implemented Standards Behind Theme 675 – 680.

Pons, P., Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In: Yolum, p., Güngör, T., Gürgen, F., Özturan, C. (eds) Computer and Information Sciences - ISCIS 2005. ISCIS 2005. Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11569596_31

Ricciato, F., Wirthmann, A., & Hahn, M. (2020). Trusted Smart Statistics: How new data will change official statistics. *Data & Policy*, *2*, e7. https://doi.org/10.1017/dap.2020.7

Rodríguez,R., Uribe-Toril,J., De Pablo Valenciano,J. (2020). Worldwide trends in the scientific production on rural depopulation, a bibliometric analysis using bibliometrix R-tool, Land Use Policy, Volume 97, https://doi.org/10.1016/j.landusepol.2020.104787

Rusydiana, A. S. (2021). Bibliometric analysis of journals, authors, and topics related to COVID-19 and Islamic finance listed in the Dimensions database by Biblioshiny. Science Editing, 8(1), 72–78. DOI: https://doi.org/10.6087/kcse.232

Salgado, D., & Oancea, B. (2020). On new data sources for the production of official statistics. https://arxiv.org/abs/2003.06797v1

Serena, J.M. et al. (2021). Use of big data sources and applications at central banks. Irving Fisher Committee on Central Bank Statistics.

Small, H., (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. Journal of the American Society for Information Science: Volume 24, Issue 4, Pages 265-269. https://doi.org/10.1002/asi.4630240406

Statistica Moldovei. (2023). Residential Property Price indices. National Bureau of Statistics of the Republic of Moldova. Residential Property Price indices (gov.md)

Statistical Commission. (2016). Report on the Global Working Group on Big Data for Official Statistics.

Statistics Sweden. (2023). Journal of Official Statistics. Sage Journals.

https://journals.sagepub.com/description/JOF

Struijs,P.,Braaksma,B.,&Daas,P.J. (2014). Official statistics and Big Data. Big Data & Society, 1(1).

https://doi.org/10.1177/2053951714538417

Suominen, A. & Hajikhani, A. (2021). Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. Policy & Internet. Volume13, Issue4. https://doi.org/10.1002/poi3.258

Surwase,G., Sagar,A., Kademani, B.S., Bhanumurthy, K. (2011). Co-citation Analysis: An Overview.

https://core.ac.uk/download/pdf/290485458.pdf

Tillé,Y.,Debusschere,M.,Luomaranta,H.,Axelson,M.,Elvers,E.,Holmberg,A. & Valliant,R.(2022).Some Thoughts on Official Statistics and its Future (with discussion). Journal of Official Statistics,38(2) 557-598. https://doi.org/10.2478/jos-2022-0026

UNECE. (2013). Big data for official statistics – strategies and some initial European applications. United Nations Economic Commission for Europe. Seminar on Statistical Data Collection https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP30.pdf

United Nations. (2008). How Should a Modern National System of Official Statistics Look: The relationship between international principles on systems of official statistics and national statistical legislation. https://unece.org/DAM/stats/documents/applyprinciples.e.pdf

United Nations. (2015b). Big Data. https://unece.org/statistics/ces/big-data

United Nations. (2023). The Global Working Group on Big Data for Official Statistics. https://unstats.un.org/bigdata/

UNSD. (2019). Handbook on the Use of Mobile Phone Data for Official Statistics. UN Global Working Group on Big Data for Official Statistics. United Nations. https://unstats.un.org/bigdata/task-teams/mobile-phone/MPD%20Handbook%2020191004.pdf

UNSTATS. (2014). Fundamental Principles of Official Statistics. United Nations. https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

Upadhyaya,S., Kynclova,P. (2017). Big Data – Its relevance and impact on industrial Statistics. United Nations industrial development organization. https://cepei.org/wp-content/uploads/2019/09/Big-Data-Its-Relevance-and-Impact-on-Industrial-Statistics.pdf

Van der Grient,H., & De Haan,J. (2010). The use of supermarket scanner data in the Dutch CPI. StatisticsNetherlands.
https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2010/zip.6.e.pdf

Virgillito, A., Scannapieco, M., & Zardetto, D. (2013). Placing Big Data in Official Statistics: A Big Challenge? Paper for the New Techniques and Technologies for Statistics conference.

Waye, A. et al. (2018). Transactional Data Processing System. Proceedings of Statistics Canada Symposium 2018. Statistics Canada. https://www.statcan.gc.ca/en/media/2293

Weidt, G. (2011). Bibliometrics: Publication evaluation and citation analysis. TUHH University. https://www.tub.tuhh.de/en/publishing/bibliometrics/

Leung,X.Y., Sun,J., & Bai,B. (2017). Bibliometrics of social media research: A co-citation and co-word analysis, International Journal of Hospitality Management, Volume 66, 2017, Pages 35-45, ISSN 0278-4319, https://doi.org/10.1016/j.ijhm.2017.06.012.

Zupic, I., & Cater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods*, *18*(3), 429-472. https://doi.org/10.1177/1094428114562629