From Big Data to Big Insights: The Power and Efficiency of Big Data and Data Science in Statistical Work

Ahlam Alrousan, Jordan Department of Statistics, Jordan, Ahlam.alrosan@dos.gov.jo

Abstract

This paper aims to study the factors involved in developing a policy for using non-traditional data in the work of statistical offices, assessing the impact of the diversity of sources, and developing research methodology on data quality and the needs of statisticians in supporting decision-makers. This study provides a comprehensive explanation of the benefits and limitations of using large and non-traditional data sources and following Best practices by researching many published studies and research available on the official search engines of international organizations, as well as reports issued by the Statistics Center. We also will emphasize in this research the importance of transparency, ethical standards, data protection, and security when publishing non-traditional data in addition of discuss some statistics that distinguish between traditional and non-traditional sources We also examine case studies where we touch on the importance of integrating traditional and non-traditional data sources and where they are used to support decision-making and study the challenges and opportunities for using these data sources. Finally, we propose to develop an international policy for the use of non-traditional data to ensure data accuracy and reliability in decision-making.

Introduction

Data is a crucial element for decision-makers, providing insights into economic, social, and political conditions. However, traditional methods of collecting official data such as surveys, censuses, and administrative data have become insufficient in meeting the demand for accurate and timely statistics, especially during disasters and challenges like the Covid19 pandemic. With the increasing number of non-traditional data sources such as social media, satellite imagery, and big data, it has become essential to utilize these sources to enrich statistical data and enhance the role of statistical offices in supporting policy decisions and keeping pace with the pace of global events.

The emergence of new and diverse data sources presents new challenges for statistical offices in terms of quality, privacy, and ethical considerations. The development of research methods and source diversity will affect the quality of data, requiring statisticians to adopt new strategies and follow new requirements. Integrating traditional and modern methods can provide more accurate and reliable results.

This study emphasizes the importance of the big data revolution and non-traditional data sources and the opportunities they create for statistical offices to improve their products' quality and provide real-time data to decision-makers and policy-makers. However, using non-traditional and big data requires a framework for public policy that addresses challenges and enables responsible use. This study covers statistical aspects from basic sources of information to the correct methodology for using big data, backed by results and recommendations.

Methodology

To Explore the Power of Big Data and How to Harness its Potential and Overcome Limitations, we delved into the political implications of using big data and non-traditional data sources in statistical work. we conducted a comprehensive literature review, analyzing relevant studies, articles, and reports to understand the potential policy implications of this cutting-edge technology.

We focused on studies that explored the use of big data and non-traditional data sources in statistical work and their potential policy implications. We also conducted a case study of the use of social media data for several examples and actual success stories, we researched the Advantages and limitations of big data and non-traditional data sources in statistical work.

To ensure the ethical use of this technology, we reached out of the necessity of putting best practices in using this type of data, One key recommendation is to use a combination of traditional and non-traditional data sources to provide a more complete picture of the phenomenon under study. Additionally, using appropriate statistical techniques, such as machine learning algorithms, natural language, and data science processing, can help to better analyze non-traditional data.

It's important to be aware of potential biases and limitations in the data and take steps to mitigate them.

Our research team also recommends developing clear policies and guidelines on the ethical use of personal data in statistical work, including obtaining informed consent and protecting the privacy of individuals.

With these best practices in mind, big data and non-traditional data sources have the potential to transform the way we approach statistics, and we are excited to continue exploring this cutting-edge technology.

Case studies

The summary was acquired from the Australian Bureau of Statistics. These images have been selected from among institutions affiliated with their respective fields. While the national statistical offices shared their experiences in participating in exhibitions.

" Information was collected from each organization through a combination of email exchanges, telephone interviews, and in-person meetings." [2] (unescap.org)

Australian Bureau of Statistics

Figure 3: Timeline of data collection methods (unescap.org)

The Australian Bureau of Statistics (ABS) has successfully integrated non-traditional data sources into its Consumer Price Index (CPI) since 2014. They used scanner data and web-scraped data and tested multilateral price index calculation methods to make better use of the data. About 40% of the CPI basket (by weight) now represents price movements calculated from non-traditional data sources, with half of that using scanner data. The ABS's success is largely attributed to its extensive testing and experimentation, collaboration with other National Statistical Offices (NSOs), academics, and other price index experts, as well as their significant consultation and communication efforts with stakeholders. Their use of non-traditional data sources has improved the quality of the CPI and allowed for better measurement of price changes in various sectors. [2] (unescap.org)

Social Media Data - In Canada, Statistics Canada used data from social media platforms such as Twitter and Facebook to analyze public opinion about the legalization of marijuana. The data was used to inform public policy and to develop targeted communication strategies. [5] Statistics Canada (2018)

Discussion and results

The use of non-traditional data sources is becoming increasingly important in statistical offices because it provides a more complete understanding of society and its needs, leading to better decision-making. The process of using this type of data in statistics can also speed up international decision-making during crises and other unexpected events, allowing for quicker response times and the development of more effective policies. Non-traditional data sources provide accurate and timely information that is difficult to obtain through traditional surveys or censuses, we can summarize the benefits of non-traditional data by

First, it allows organizations to make informed decisions based on real-time data, providing a competitive advantage in fast-paced markets.

Second:" Big data and unconventional data provide a more complete picture of the situation, enabling decision-makers to make informed decisions based on a broader range of data." [6] (general , 2014)

Third, these data sources can help organizations predict future trends, enabling them to prepare proactively for potential challenges. However, statistical offices need to develop new skills and tools to handle and integrate this data into their workflows, including data cleaning and curation techniques, machine learning algorithms, and data visualization and analysis tools. [3] (Diego Zardetto)

Figure 1: Using Machine Learning Techniques for Big Data Analytics (Harding and Hersh, 2018)
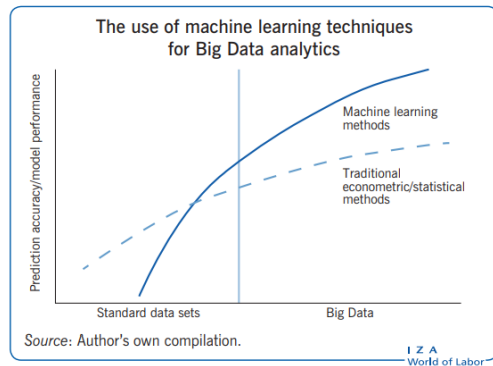
"Big data refers to larger, higher frequency datasets, and often more personalized information. Examples include data collected by smart sensors in homes or the aggregation of tweets on Twitter. On small data sets, traditional econometric methods tend to outperform more complex techniques. However, on large data sets, machine learning methods shine. New analytics approaches are needed to make the most of big data in economic analytics. Researchers and policymakers should pay close attention to recent developments in learning devices and technologies if they are to take full advantage of big data sources." [1] (Harding and Hersh ,2018)

Figure 2: Integrated use of big data and traditional data (from the author's conclusion, 2023)
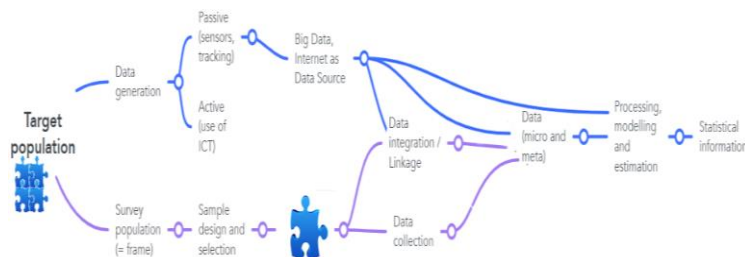
The integration of big data with traditional data by collaboratively using people, processes, suppliers, and technologies to retrieve data from disparate sources and reconcile them and make better use of them to support the decision making decision-making process more accurate, fast, and easy as big data is characterized by: size, speed, honesty, variance, value, and visualization. Big data is distinguished from traditional structured data managed by relational database systems in that the number of data sources is much higher than in the traditional approach to managing data inputs. In addition, the source of data increases the rate of data generation while data generation comes

from many sources with different formats and structures that are not formatted in traditional data. In addition to data reliability, not all data has value and high quality, however, challenges such as data quality must be addressed. privacy and legal issues. [2] (unescap.org)

Recommendations

The paper makes a compelling case for the importance of big data and data science in statistical work and the critical need for quality assurance policies to ensure accuracy and reliability, especially since non-traditional data is data that is not recognized at the international level and it doesn't comply with the usual standards of statistics, and since it has become a necessity that calls for turning to it because of the high-value data it contains to support decision-makers, Policymakers, and practitioners must recognize the potential of big data and data science to transform statistical work and prioritize the development and implementation of quality assurance policies that ensure ethical, transparent, and responsible use of data. [7] (EU,2021)

To achieve this, official statistical centers should cooperate with other concerned authorities to apply clear standards and agreed-upon conditions.

The policy should establish guidelines for data collection, analysis, and disclosure to ensure accuracy, relevance, and high quality also the policy should also ensure the achievement of strategic goals and statistical and ethical principles.

Conclusion

The emergence of modern sources of big data and non-traditional data has revolutionized the field of statistics and statistical analysis. The use of these data sources has provided new opportunities for research and insights into complex phenomena, transforming the way statistical analysis is done. However, the use of big data and non-traditional data sources also poses unique challenges, such as bias and privacy issues. To ensure the ethical and effective use of these data sources in statistical work, it is essential to adopt best practices and establish clear policies and guidelines. Policymakers and statisticians must collaborate to promote the ethical and effective use of big data and non-traditional data sources, encourage interdisciplinary collaboration, and invest in research and development to improve their quality and accuracy.

References

[1] Big Data in Economics, HARDING and HERSH,2018
https://wol.iza.org/uploads/articles/451/pdfs/big-data-in-economics.pdf
https://wol.iza.org/articles/big-data-in-economics.

[2]   Incorporating Non-traditional Data Source into Official Statistics - Lessons and experiences from Australia, Japan and New Zealand (unescap.org)

https://www.unescap.org/sites/default/d8files/knowledgeproducts/incorporating_non_traditional_sources_CPI.pdf

[3] The implication of Big Data for Official Statistics, Diego Zardetto.
https://circabc.europa.eu/sd/a/c35f00b3-3103-4873-a5a8-68386822af03/DAY%201_ITEM%202_The%20implication%20of%20big%20data.pdf

[4]   United Nations National Quality Assurance Frameworks Manual for Official Statistics, general, 2019
https://unstats.un.org/unsd/methodology/dataquality/

[5]   Statistics Canada (2018). Using social media to measure the impact of cannabis legalization in Canada:
https://www150.statcan.gc.ca/n1/pub/11-631-x/11-631-x2018003-eng.htm

[6]   Report of the Global Working Group on Big data for official statistics   ,   general,   2014
https://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf

[7] Guidelines on the use of electronic data collection technologies in population and housing censuses , EU , 2021.
https://unstats.un.org/unsd/demographic/standmeth/handbooks/guideline-edct-census-v1.pdf