systems

MDPI

Article

Official Statistics and Big Data Processing with Artificial Intelligence: Capacity Indicators for Public Sector Organizations

Syed Wasim Abbas [1], Muhammad Hamid [2], Reem Alkanhel [3,*] and Hanaa A. Abdallah [3]

[1] Pakistan Bureau of Statistics, Lahore 54000, Pakistan; dr.wasim@pbs.gov.pk
[2] Department of Computer Science, Government College Women University, Sialkot 51310, Pakistan; mhamid@gcwus.edu.pk
[3] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia; haabdullah@pnu.edu.sa
* Correspondence: rialkanhal@pnu.edu.sa

Abstract: Efficient monitoring and achievement of the Sustainable Development Goals (SDGs) has increased the need for a variety of data and statistics. The massive increase in data gathering through social networks, traditional business systems, and Internet of Things (IoT)-based sensor devices raises real questions regarding the capacity of national statistical systems (NSS) for utilizing big data sources. Further, in this current era, big data is captured through sensor-based systems in public sector organizations. To gauge the capacity of public sector institutions in this regard, this work provides an indicator to monitor the processing capacity of the public sector organizations within the country (Pakistan). Some of the indicators related to measuring the capacity of the NSS were captured through a census-based survey. At the same time, convex logistic principal component analysis was used to develop scores and relative capacity indicators. The findings show that most organizations hesitate to disseminate data due to concerns about data privacy and that public sector organizations' IT personnel are unable to deal with big data sources to generate official statistics. Artificial intelligence (AI) techniques can be used to overcome these challenges, such as automating data processing, improving data privacy and security, and enhancing the capabilities of IT human resources. This research helps to design capacity-building initiatives for public sector organizations in weak dimensions, focusing on leveraging AI to enhance the production of quality and reliable statistics.

Keywords: artificial intelligence; big data; convex logistic principal component analysis; capacity indicator; sensor-based systems

Citation: Abbas, S.W.; Hamid, M.; Alkanhel, R.; Abdallah, H.A. Official Statistics and Big Data Processing with Artificial Intelligence: Capacity Indicators for Public Sector Organizations. *Systems* **2023**, *11*, 424. https://doi.org/10.3390/ystems11080424

Academic Editors: In-kee Kim, Avinash Kalyanaraman and Lakshmish Ramaswamy

Received: 29 May 2023
Revised: 8 August 2023
Accepted: 10 August 2023
Published: 13 August 2023

1. Introduction

Public sector organizations' adoption of advanced information technology, including artificial intelligence, triggers an increase in digital administrative data sources (DADs). This digitalization reveals the various processing, storage, and privacy issues of the host departments and the country's national statistical institutions (NSIs) [1]. Administrative data is one of the six main categories into which large datasets are classified; additionally, it is argued that administrative data fulfill the characteristics of big data if their recording velocity is high [2]. The united nations economics commission of europe (UNECE) defines three main classifications of the types of big data [3]. The first is social networks, i.e., human-sourced information, while the second is traditional business systems (process-mediated data) based on data produced by public agencies and businesses, which belongs to the class of "Administrative data" "we have earlier called DADs", and the third is the Internet of Things (IoT), i.e., machine-generated data [4,5]. The IoT connects the physical and digital worlds with gadgets through the Internet and several other protocols. The gadgets generate a significant amount of data that contain essential knowledge about the physical world. Wireless sensor networks (WSNs) are rich big data sources in the IoT, among many other

plausible data sources. Different sensor nodes produce big data in large-scale networks [6,7]. With the increasing number of IoT sensing devices available, data generated from public sector organizations are expected to grow exponentially; this digital administrative data fall under the definition of "Big data" [8,9]. DADs are vital for measuring the performances of a country or region's different economic and population measures. For an in-depth review of the literature regarding methodological tools and operationalization for big data in official statistics from national and international organizations, one may refer to Abbas et al. [10].

The demand to "Leave No One Behind" in the 2030 Agenda and its SDGs has increased the need for a variety of data and statistics to support informed policy and decision making in all countries [11]. While a strong official statistical system is vital for a country to formulate, implement, and monitor its development policies, it also assists the government in evidence-based planning [12]. United Nations (UN) agencies are also working to support countries in strengthening their statistical capacity to produce and use data, including artificial intelligence, for better policy formulation and approaches to implement the 2030 development agenda at the national level. Several capacity development projects are underway to empower the NSS of countries worldwide, including workshops, direct country assistance, training/guidance material development, establishment of specialized networks, and study visits [13]. The details of these projects are available on the UN Statistics Division's website.

Rogge, Agasisti, and De Witte [14] presented a global overview of statistical capacity development in their Paris 21's Partner Report on Support to Statistics (PRESS). A support of USD 623 million for statistical capacity building from the donors of development cooperation agencies (UNICEF, IMF, European Commission/Eurostat, UNFPA, World Bank) was provided in 2016 [10]. Lebada [15] argued that giving the data requirements of the SDG monitoring framework precedence over the development of NSS could be a miscalculation. Rather, countries should prioritize the development of an effective NSS that is sufficiently flexible, responsive, and cost-effective to meet the enormous demand of the SDG monitoring framework and national information needs. This enormous expansion of scope and scale raises serious concerns about the capacity of national statistical systems, or what others have termed the "Data eco-system", to implement such a massive monitoring framework [10]. The complexity and ambition of this challenge led Mogens Lykketoft, President of the UN General Assembly, to describe it as an "Unprecedented Statistical Challenge" [15].

The rising demand and importance of good-quality, independent official statistics provide a unique opportunity to make a real and long-lasting investment to improve NSS [16]. Moreover, it is important to improve the NSS at the grassroots level and empower this system with respect to grey areas. Assessing a country's NSS's strengths and weaknesses is vital before launching capacity-building activities. However, no scale, indicator, or index can be used to gauge and rate the efficiency of organizations in the public sector across a nation [11–17]. Only the World Bank's statistical capability indicator provides a nationwide overview of about 140 developing nations. Furthermore, their capacity indicator is based on a diagnostic methodology created to evaluate the nation's capability using metadata. It is typically accessible to most nations and used to track statistical capacity development over time. The framework of the World Bank's statistical capability indicator is composed of four elements: statistical methodology, data source, periodicity, and timeliness [18,19]. Besides this, we have proposed a micro indicator that provides a comprehensive capacitive assessment of the official statistics system inside a country in line with the World Bank's macro statistics capacity indicator. Several factors (Table 1) relating to the official statistics system and big data processing are considered in this study when measuring the capacity of public sector organizations.

Convex logistic principal component analysis PCA addresses data privacy concerns while playing a significant role in the field of AI. Convex logistic PCA allows companies to mine high-dimensional datasets for insightful information without jeopardizing the privacy of individual users. The secrecy of sensitive information is maintained while dimensionality

reduction and pattern recognition are possible. Convex logistic PCA ensures that data are kept anonymous, reducing the chance of privacy violations. Convex logistic PCA provides a workable approach for using AI while protecting privacy thanks to its capacity to strike a compromise between data value and privacy protection.

Table 1. OS and BD processing capacity measures.

Codes	DESCRIPTION
OS1	Collection/recording of data in an organization
OS2	Data collection for statistical purpose
OS3	Production of statistics from data
OS5	Dissemination of Data products officially/publicly
OS6	Data supply to statistical organizations
OS7	Have a framework to deal with privacy-related issues
OS9	Conduct self-data collection through surveys
OS10	Acquire data from statistical organizations periodically
OS11	Have unreported data sources
OS13	Unreported data sources are important
BD1	Electronic data recording
BD2J	Big data Recording
BD2JA	Accessible data storage
BD3J	Big data production
BD4	Big data awareness
BD5	Big data importance for POS
BD6J	Big data value in POS
BD7	Big data working
BD9	In future working with Big data
BD10J	Potential Big data is an administrative source
BD11	Well IT equipped and have enough resources
BD12	Have well trained IT Staff
BD13J	Have enough data processing IT skills
BD14J	Have statistical skills by the IT human resource
BD15J	Usage of Big data processing tools
BD16	Training needs for IT staff
BD17J	Training needs for Big data processing skills
BD18	Public-private partnership over data solution needs
BD20	Mutual data interests with other public departments
BD22	Public-public liaison over data solution needs
RC1	Statistical or data processing post exists in the department

The proposed capacity indicator gives a small picture of within-country public sector organizations to process large amounts of data through artificial intelligence and produces official statistics from their DADs. A census-based survey, "Survey of Official Statistics Pakistan (SOS-Pak)," was introduced at the national level in Pakistan to monitor capacity indicators. All the federal and provincial government organizations were contacted through post-mail inquiry with email and telephonic follow-up. A one-fourth response of 171 provincial and federal public sector organizations was received regarding different

aspects related to the processing of big data with official statistics. Convex logistic principal component analysis (PCA) was used to compute capacity scores as a dimensionality reduction tool. These scores were then transformed into relative capacity indicators (RCIs), which compare and assess the NSS on various dimensions at the organizational level.

In this paper, we have measured the statistical and big data processing capacity of Pakistan's public sector organizations through certain measures. The methodological aspects are covered in Section 2, in which data collection methodology is discussed along with the questionnaire tool and the measures used to compute the statistical capacity indicator. Section 3 contains the descriptive results of the survey used to determine measures in the study and scores were calculated using convex logistic PCA based on the collected measures. A statistical capability indicator was developed to determine the capability of public sector organizations. Finally, the paper is concluded in Section 4.

2. Methodology

In this article, we have developed statistical and big data processing capacity indicators in comparison with the World Bank's statistical capacity indicators. Our developed indicators are based on primary data collected from 171 public sector organizations. To collect data from public sector organizations, the survey used was the "Survey of Official Statistics Pakistan", i.e., SOS-Pak. A questionnaire was developed to collect these measures. The entire survey methodology is discussed in Section 2.1, explaining the key modules of the questionnaire with measures covered in each module. The data collection process is explained in Section 2.2, key official statistics and big data capacity measures with descriptive findings are discussed in Section 2.3, and the dimensionality reduction AI tools for the development of official statistics and big data processing capacity indicators are discussed in Section 2.4.

2.1. Survey Methodology

A questionnaire, SOS-Pak, was designed to obtain measures about the official statistics production and big data utilization for its production (see Supplementary Materials file S1). In the public sector organization operating under the NNS of Pakistan, this survey covered several topics, including data collection, data dissemination, data privacy concerns, reported and unreported data sources, big data literacy, working IT human resource and infrastructure, and rationalization of statistical and data processing human resources. SOS-Pak was carried out at the national level in collaboration with the Bureau of Statistics (BoS) in Punjab, Pakistan, to determine the capacity of public sector organizations in Pakistan to handle big data and to produce official statistics. As a survey tool, a questionnaire with four parts was employed. The self-explanatory questionnaire includes the following modules:

IP—Basic Information Panel
OS—Official Statistics Production Information
BD—Big data use in Official Statistics
RC—Rationalization of Statistical Cadre

The IP module includes the department's name, respondent's title, basic pay scale, contact information, and organization size. The second OS module, however, addresses issues with data privacy, reported and unreported data sources, and data collection/recording methods for data product dissemination. Abbas et al. [20] reviewed the key aspects of this module that dealt with disclosed and unreported data sources.

The third module, "BD," was created to examine how well public sector organizations might handle using big data sources to provide official statistics. Measures were designed by considering several studies conducted by different UN institutions, UNECE [21], UNSD [22,23], and business firms like Mark et al. [24,25].

The fourth module (RC) was designed to monitor public sector organizations' statistical and non-statistical human resources. A review of employment positions, approved and vacated posts, work duties, and activities performed by the relevant cadres is required to rationalize the statistical cadre and data processing human resources.

The target population of the study involved all the public sector Organizational Units (OUs) working under the Federal Government (FG) and Provincial (Punjab) Government (PG). A frame based on 472 OUs of the federal and 286 OUs of the provincial government was used to conduct this census-based survey.

2.2. Data Collection

Data were collected based on a survey where each OU of the frame was contacted through postal mail with email and telephonic follow-up (see Supplementary Materials file S2). A postage-paid return envelope was sent to all 758 OUs. The reference period for data collection was from March to August 2017. The questionnaire's IP, OS, and RC modules were requested to be filled in by any well-versed organization officer. In contrast, an IT professional of that responding organization was asked to fill in the BD module. The official statistics capacity indicator takes place through three dimensions with thirteen measures.

2.3. Official Statistics and Big Data Processing Capacity Measures

Using convex logistic PCA, this section has created scores based on several metrics relating to official statistics and big data processing capacity. Both the OS and BD capacity measurements' scores are created individually. These scores help investigate the potential that is accessible and the weak areas that require capacity growth in POS using contemporary data sources. The measures listed in Table 1 are used to create scores. Measures are divided into major categories to rationalize the whole OS and BD models into partially segmented models so that developed scores may explain the ideal deviation.

2.3.1. Official Statistics Capacity Measures

Here, the public sector organizations' competence for producing statistics effectively—from data collection and recording to data dissemination—is assessed. This capacity is separated into three partial groups.

(a) Data collection/recording for dissemination

This collection of measures includes those that have to do with creating and disseminating official data. The main metrics related to this first partial category of OS capability scores include data collecting, recording, storage, data gathering functionality, data generation, and product distribution.

(b) Liaison with other departments on data

One of the primary indicators included in this second subcategory of OS competency ratings is the extent to which organizations engage.

(c) Data Privacy

The safeguards for data privacy are covered in this category's section. Measures for the framework to address privacy issues, the degree of confidentiality, and the significance of these sources, as well as a collection of statistics based on needs, are all included in this dimension.

2.3.2. Big Data Processing Capacity Measures

Several indicators are used to assess how effectively public sector organizations can use big or large data sources to produce statistical products or to supply the data needed to produce those products. These indicators are divided into four subcategories here, which are detailed below.

(a) Big data 3Vs

The three key aspects of big data—volume, velocity, and variety—and data recording and storage are assessed to determine the extent of big data usage.

(b) Big data literacy

Scores that can be used to verify and compare the level of big data literacy in public sector organizations are produced using knowledge of big data, its value for POS, and its utility in organizational planning and decision-making.

(c) Big data workings

This dimension also includes looking for possible sources of big data, working together between the public and private sectors to process big data, and working with other groups to process and handle big data. This dimension includes how groups in the public sector use big data, and whether they do so now or plan to do so soon.

(d) Big data skills

This dimension is the main sub-component of the big data processing capability measures. Here, a measure for the computation of BD capacity scores is the availability of adequate IT infrastructure, IT human resources, statistical expertise of IT professionals, and training requirements for IT professionals in public sector organizations. Furthermore, the detailed rationalization of statistical cadre in public sector organizations is explained.

2.4. Official Statistics Capacity Indicator

This includes four dimensions with 18 measures:

1. Big data 3Vs (Volume, Velocity, Variety);
2. Big data literacy;
3. Big data workings;
4. Big data skills.

PCA is a well-known data visualization, feature analysis, and compression technique. A logistic PCA for the binary data version of this method was developed [26]. The PC scores in logistic PCA are, like in normal PCA, linear combinations of the saturated model's natural parameters. Both logistic PCA and convex logistic PCA are based on the same ideas. Logistic PCA minimizes rank k projection matrices, which is the only distinction. Contrarily, convex logistic PCA minimizes over the Fan tope, which is a convex hull of the rank k projection matrix.

The scores are generated individually for the OS and BD capacity models using convex logistic PCA. The scores are subsequently converted into RCIs via the vector transformation technique. By applying the linear transformation, the RCIs are calculated as follows.

$$\mathrm{RCI} = \frac{\tau}{\tau_{(m)}} \times 100 \tag{1}$$

$$\tau = \alpha k + m \ \text{ where } \alpha = (\pm 1) \tag{2}$$

where k is the PC scores vector of the ith entity, and m is the upper record value of the vector. Using this index, each organization can be generally evaluated compared to the one with the highest score. When the RCI is 100, it identifies the most competent organization among a group of OUs in terms of a given metric.

3. Results and Findings

This section contains key descriptive findings and results of AI-based dimensionality reduction tools for the development of official statistics and big data processing capacity indicators.

3.1. Key Descriptive Findings

This survey covers a total response of 23% of OUs, i.e., the results are presented based on a net response of 171 public sector OUs. The respondents of this survey were the top managers of their respective organizations. The net sample includes mixed-size OUs from small (<50) to large (>1000) employment sizes.

The following are the key findings related to the OS Module.

- OUs record/collect data for official or public use regularly 83% (142/171)
- OUs collect/record data for:Administrative use only 29% (40/139)
- Statistical use only 10% (14/139)
- Both (Admn./Statistical) 57% (79/139)
- OUs have electronic data recording 93% (150/160)
- Production of data products by the collected/recorded DADs 68% (97/142)
- OUs disseminate their data products officially or publicly 87% (89/102)
- OUs supply data to statistical organizations periodically 48% (68/142)
- OUs routinely obtain data from statistics organizations 23% (38/168)
- OUs have to conduct their collection to fulfill data needs 39% (66/169)
- OUs reported a confidentiality level
- Low 30% (40/135)
- Medium 36% (49/135)
- High confidentiality level 34% (46/135)
- OUs have an internal policy of 57% (78/137)
- OUs reported unreported data sources (data gaps) 27% (47/171)
- OUs ensure the importance of unreported data sources for POS 27% (47/171)

Module II (Big Data Use in Official Statistics)

In this section, we looked into and gathered data on several measures relating to Pakistani public sector organizations' capacity to process big data.

(a) Data recording and storage system

Most OUs collect, document, and store data in some way. The study's findings show that 94% of OUs record the data they acquire electronically, compared to only 6% of departments that do the same manually. The value of DADs depends on how easily their storage system is accessed. To analyze digital data storage in the public sector, respondents from the IT department were asked to explain the implemented system in their organizations. Figure 1 compares data saved in various ways on an FG and PG level.

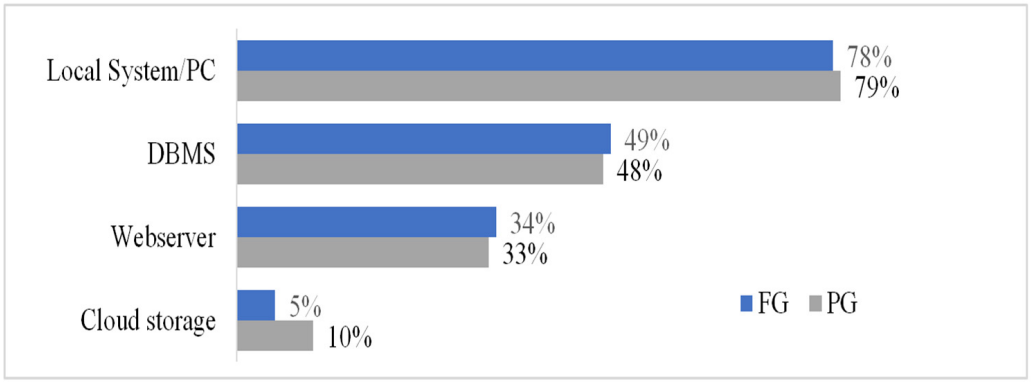

Figure 1. Data storage systems of FG and PG OUs.

(b) 3Vs of data

The 3Vs hold new opportunities for developing new official statistics and restructuring existing ones. High volume may produce more accurate and detailed statistics, high velocity may give more frequent and timely statistics, and high variety may lead to an official multidimensional statistic. Here, we have captured these characteristics in the public sector organizations of Pakistan. Table 2 can aid understanding of the public sector departments' volume and velocity of data. The Table displays a crosstab of the administered data's volume and velocity. Big data sources are absent from 23.6% of OUs overall, they are present in 33.8% of OUs with certainty, and 42.6% of OUs may have them but need further explanation.

Table 2. Big data inflow in FG and PG OUs (percentage of OUs).

		Volume of Data			
		<1 GB	**1–10 GB**	**>10 GB**	**Total**
Velocity of data	Daily	42.6	17.6	2.0	62.2
	Monthly	15.5	12.8	0.7	29.1
	Yearly	6.1	2.0	0.7	8.8
	Total	64.2	32.4	3.4	100.0
No OUs		OUs		Possible OUs	

The survey results regarding the variety of data revealed that 90% (141/157) of OUs reported recording numerical data, 82% (128/157) of OUs reported recording text data, 50% (79/157) of OUs recorded graphic data, and 4.5% (7/157) of OUs recorded other data types.

(c) Big data literacy

IT personnel in public sector organizations were asked to complete a survey to determine their level of big data literacy. The respondents were given a broad questionnaire, and when asked if they had ever heard the word "big data" before, the question was phrased as, "Do they ever hear the term BIG DATA before now?" In total, 101 out of 154 OUs, or 66%, said they had heard of big data. The three-point Likert scale results in Figure 2 demonstrate that the public sector is fully aware of the value of using big data to save time and money while enhancing the accuracy of official production statistics (POS).

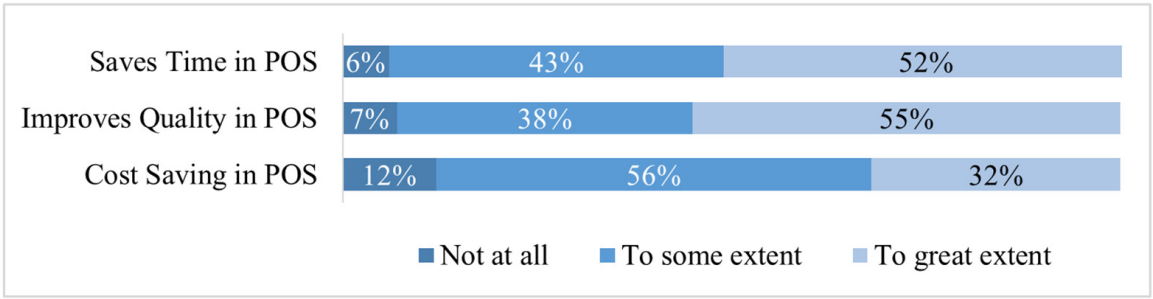

Figure 2. Significance of utilizing big data in POS from the perspective of IT professionals.

(d) Current and Upcoming work with big data

The questionnaire asked respondents to describe their past, present, and future use of big data. Currently, 8% (13/159) of OUs are utilizing both administrative and non-administrative big data sources. Below is a list of the preferences of the FG and PG OUs for potential big data sources as described by UNECE.

1. Administrative Data (72%, 115/160)
2. Behavioral Data (26%, 41/160)
3. Communication/Tracking Devices data (25%, 40/160)
4. Sensors Data (16%, 26/160)
5. Commercial/Transactional Records (14%, 22/160)
6. Opinion Records (19%, 31/160)

Due to its high potential, usability, and applicability in public sector functioning, the administrative record is considered to be a top preference of public sector organizations. It is important to note that the next five classifications aside from "Administrative Data" typically rely on unstructured data collections.

(e) IT personnel to handle Big Data

It is essential to have competent IT staff to manage big data sources. Out of the 153 OUs that responded, 44 indicated they had IT staff that were highly trained. The IT people resources in the separate organizations employ data processing tools wisely.

Figure 3 shows the percentage of different data processing technologies that IT HR possesses and employs.

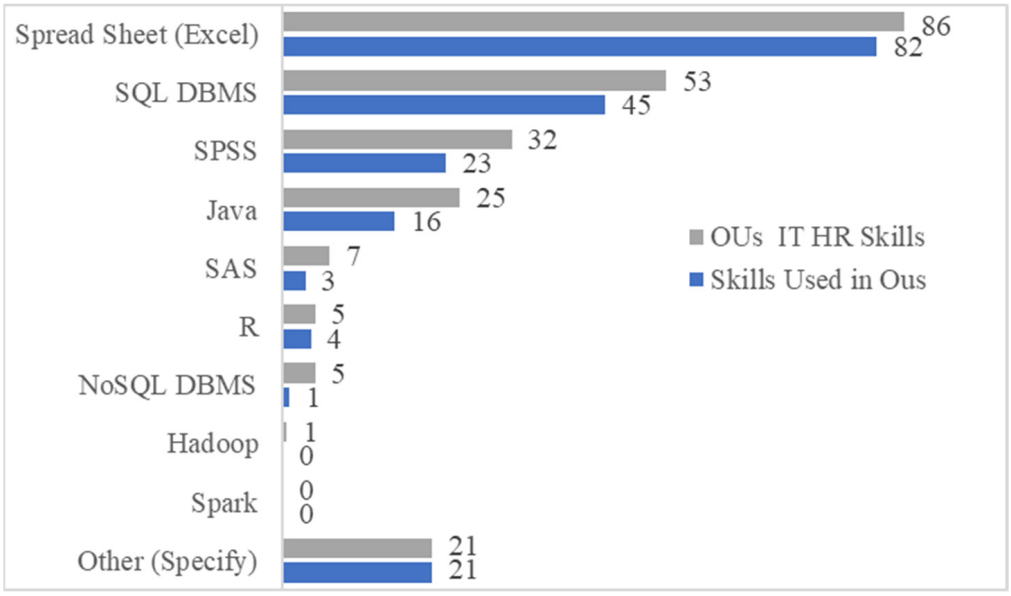

Figure 3. IT HR data processing tools vs. its utilization in Pakistani OUs (percentage).

(f) IT infrastructure to deal with big data

Adequate IT infrastructure and a knowledgeable IT staff are crucial to fulfill modern data processing requirements. Of the 153 OUs examined, 36 (23.5%) were found to have well-equipped IT infrastructure and sufficient resources to meet their big data processing needs. It was observed that most FG and PG OUs rely on structural databases (SQL DBMS).

(g) Statistical capacity of IT human resources

This question was posed to determine the statistical capacity of IT human resources in public sector organizations. The paucity of advanced statistical abilities is clearly illustrated by the ratio of various statistical skills among IT employees in public sector OUs (Figure 4). The first step towards data insights is data visualization, which only 19% of IT employees are proficient in. This circumstance highlights the urgent necessity for public sector organizations to strengthen the statistical capacity of IT HR.

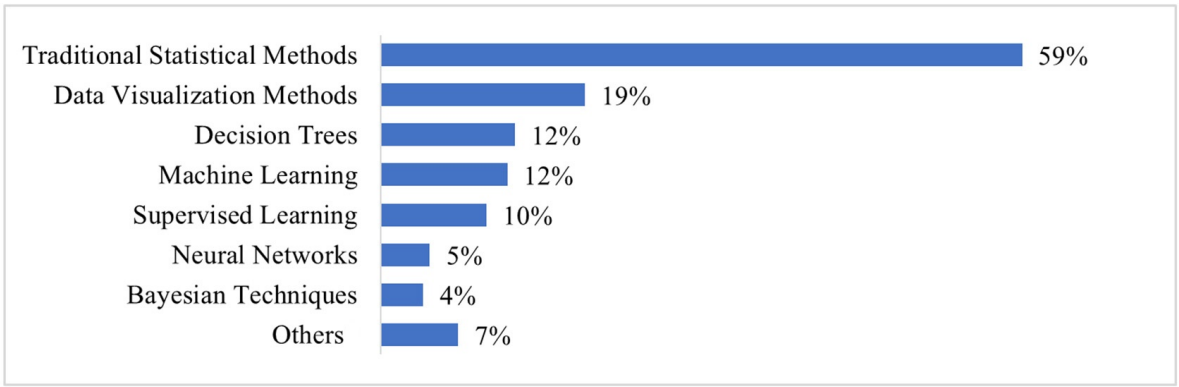

Figure 4. A statistical skill assessment of IT personnel in public sector organizations.

(h) Training needs of IT human resources

Participants were asked to describe their demands by checking the desired answer to investigate the demand for training in data processing technologies by public sector organizations. In total, 76% (117/154) of OUs aim to train their IT employees in various

data processing tools in the near or distant future to meet the demands of today's data processing difficulties. Figure 5 illustrates the percentage distribution of training needs for various data processing software among public sector organizations. Other requirements were discovered for GIS, BI, ACL Analytics, Oracle, and STATA.

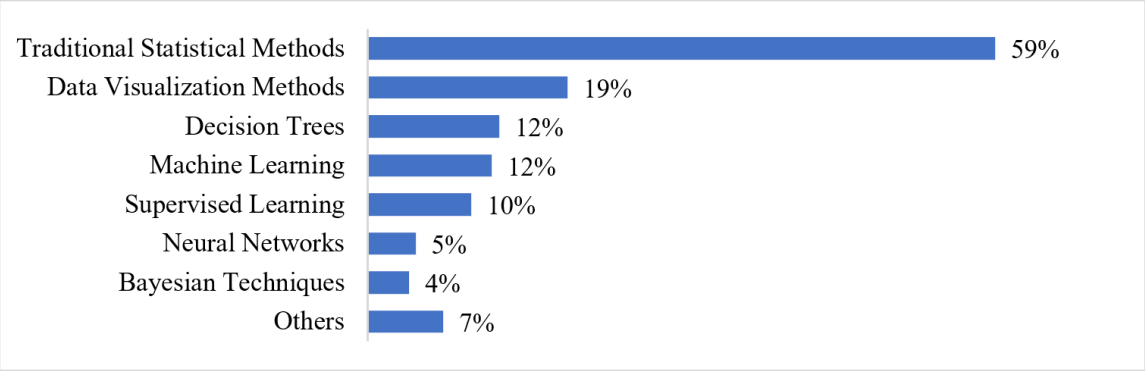

Figure 5. The public sector organizations' training needs for different data processing software, expressed as a percentage.

(i) IT and Statistical outsourcing

Organizations in the public sector were found to have lower IT and statistical capabilities than those in the private sector. Public–private partnerships are essential if the government system is to operate more effectively. We looked into this and discovered that 19% (29/153) of the organizations used to work with other IT companies to meet their data solution demands.

(j) Liaison with other departments on data

Liaison between various public sector OUs may assist them in improving their functionality and overcoming the challenges of contemporary data processing requirements. In this study, we discovered that 19.3% (29/150) of OUs work together with other statistical and non-statistical organizations to process data, collect data, compile it, store it, analyze it, write reports, and disseminate it. Most of the collaboration among different organizations was found with the statistical bodies.

(k) Reasons for lacking big data use

To identify potential causes of the decreased use of contemporary data sources, the IT staff of the participating public sector OUs were given a list of thirteen reasons, one of which was left open-ended, to rank on a five-point Likert scale (strongly disagree = 1 to strongly agree = 5). Figure 6 shows a stacked bar graph of the ratings.

The main causes for the slow adoption of big data sources include a lack of advanced statistical and data processing abilities, a lack of research and research environments, low levels of awareness of big data, and technical stagnation. By removing these obstacles, public sector organizations may be able to use contemporary data processing methods.

3.2. Selection of Optimum Data Reduction Approach

A variety of data reduction approaches were used over the complete and dimensional capacity groups to assess the ability of public sector OUs to process official statistics (OS) and big data (BD). R software was used to compare the logistic PCA, Convex Logistic PCA, and exponential family PCA previously described by Landgraf and Lee [18]. Table 3 displays the variance explained by different methods. The explained deviation shows that convex logistic PCA effectively calculates OS and BD capacity scores.

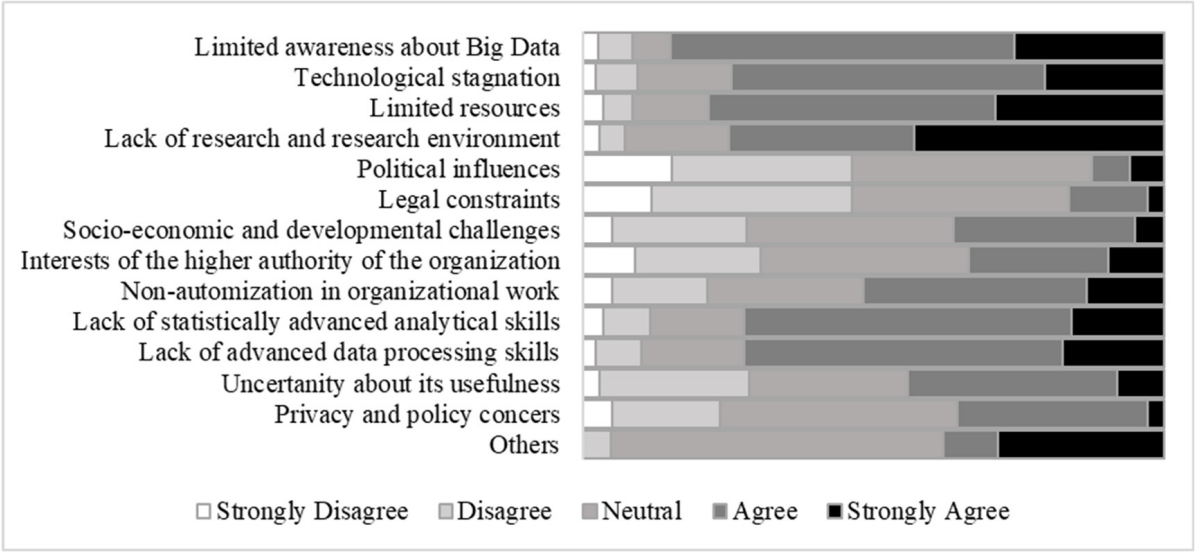

Figure 6. The reasons why public sector organizations do not use big data (FG + PG).

Table 3. Table of measures used, and deviance explained by different data reduction techniques over full and partial models (k = 1).

Dimensions for Capacity Measures		Measures Used (Nos.)	Deviance Explained (%)		
			Exponential Family PCA	Logistic PCA	Convex Logistic PCA
Overall Capacity Indicator		31	18	16	39
Official Statistics Capacity Indicator		13	28	24	55
OS Sub-dimensions	1. Data Collection and Dissemination	5	63	52	83
	2. Liaison with other Depts. on Data	4	52	41	89
	3. Data Privacy	5	52	40	85
Big Data Processing Capacity Indicator		18	23	21	50
BD Sub-dimensions	1. Big data 3Vs	4	77	70	92
	2. Big data Literacy	3	64	49	95
	3. Big data Workings	6	44	36	77
	4. Big data Skills	7	42	33	78

Based on measurements gathered for various dimensions, convex logistic PCA was used to produce the OS and BD capacity scores in this case, and Equation (1) was used to calculate RCIs. The R package "logistic PCA" was used to calculate scores for both full and partial models.

3.2.1. Official Statistics Relative Capacity Indicator (OSRCI)

Using convex logistic PCA scores based on reported measures, the OSRCI was designed to assess public sector organizations' competence in producing official statistics for the overall and sub-dimensions presented in Table 3. When compared to the most active organization in the particular measure, an organization's OSRCI indicates how it stacks up relative to that organization. As an illustration, Tables 4 and 5, which exhibit data from a net sample of 171 OUs, illustrate ten federal and provincial government departments with the highest OSRCI.

Table 4. Official statistics relative capacity indicator (federal departments).

Department Name	OSRCI	Sub-D1 RCI	Sub-D2 RCI	Sub-D3 RCI
State Bank of Pakistan	100	100	100	100
PPARC Establishment Division	87.1	100	82.3	75.8
Pakistan Council for Science & Technology (PCST)	80.7	100	77.1	75.8
FBISE Islamabad	76.9	100	39.1	75.8
Gwadar Port Authority	64.6	75.7	46.7	100
Pakistan Bureau of Statistics (ACO Wing)	64.6	100	5.2	75.8
Directorate of Research and Statistics FBR	61.3	63.4	48.1	45.9
Ministry of Information Technology	56.2	100	5.2	85.3
Capital Development Authority	54.9	71.5	43.2	85.3
Civil Services Academy	54.5	34.8	77.1	37.9

Table 5. Official statistics relative capacity indicator (provincial departments).

Department Name	OSRCI	Sub-D1 RCI	Sub-D2 RCI	Sub-D3 RCI
Bureau of Statistics Punjab	100	100	100	100
Provincial Disaster Management Authority Punjab	92.8	100	100	75.8
Crop Reporting Service	85.7	100	100	49.9
Directorate of Industries (IPWM)	72.2	100	33.9	74.1
Punjab Vocational Training Council	71.8	100	5.2	100
Directorate General of Monitoring & Evaluation	70.0	100	77.1	75.8
Population Welfare Department Punjab	66.6	34.8	100	37.9
Literacy & Non-Formal Basic Education Department	64.7	100	43.2	100
Faisalabad Institute of Cardiology Faisalabad	64.6	100	5.2	75.8
Excise Taxation and Narcotics Control Department	64.1	100	48.1	75.8

Among all the FG public sector organizations, the State Bank of Pakistan (SBP) had the highest index score for official statistics capacity (OSRCI = 100). The Bureau of Statistics Punjab ranked first with an OSRCI of 100 among all provincial public sector organizations. The indication for these organizations' performance in the next three sub-dimensions also sits at 100, indicating that they are effectively meeting the POS standards. The RCI for Sub-D1 was 100, while it was 82.3 and 75.8 for Sub-D2 and D3, respectively, for PPARC, which had an OSRCI of 87.1. In terms of its official statistics capabilities, the ACO division of the Pakistan Bureau of Statistics (PBS) was comparably indexed as OSRCI = 64.6 compared to SBP. The fact that Sub-D1's RCI is 100 shows that this wing's methods for gathering and disseminating data are on par with those used by the SBP system. However, the RCI for Sub-D2 based on stated measures was incredibly low (RCI = 5.2), indicating that the wing's data-sharing with other departments (statistical or non-statistical) is insufficient.

The low RCI for Sub-D2 implies that the key Sub-D2 metrics of the organization require improvement or capacity growth.

As shown in Table 5, the PG OUs with low Sub-D2 ranks also have low OSRCI ratings. This includes the Department of Literacy and Non-Formal Basic Education, the Faisalabad Institute of Cardiology, the Excise Taxation Department, and the Narcotics Control Department. Both FG and PG government agencies lack Sub-D2. This suggests that OUs are not effectively sharing data with one another or with other OUs, whether those OUs are involved in statistics or not.

3.2.2. Big Data Processing Relative Capacity Indicator (BDRCI)

The BDRCI calculates an organizational unit's overall big data processing capacity by considering all eighteen associated metrics. The RCI for Sub-D1 refers to the 3Vs (volume, variety, and velocity) of an organization's data. Moreover, the RCI for Sub-D1 will be found within an organization that manages and produces a high volume, velocity, and variety of data. The OUs' big data literacy and comprehension level is also related to the RCI for Sub-D2. Through the RCI created under Sub-D3, the capacity of public sector organizations to work with big data sources at current or future positions is captured. Finally, the RCI obtained in Sub-D4 allows for the visualization of the organizations' big data processing capabilities.

The BDRCI demonstrates the ability of public sector organizations to manage big data sources, as well as their knowledge of big data workings and the presence of the necessary skills for processing big data sources. For the top 10 FG and PG OUs, respectively, the capacity scores derived from convex logistic PCA were utilized to build BDRCI, as shown in Tables 6 and 7.

Table 6. Big data processing relative capacity indicator (federal departments).

Department Name	BDRCI	Sub-D1 RCI	Sub-D2 RCI	Sub-D3 RCI	Sub-D4 RCI
State Bank of Pakistan	100	100	100	100	100
National Disaster Management Authority	89.1	100	100	94	79.7
Department of Auditor General of Pakistan	77.9	100	80.5	81.6	65.1
Provincial Election Commissioner Baluchistan	66.1	60.1	100	34.9	100
Gwadar Port Authority	58.1	100	34.8	25.8	67.2
Pakistan Bureau of Statistics (ACO Wing)	58	70.6	34.8	20.4	85
FBISE Islamabad	57.6	100	100	49.8	26.7
Pakistan Health Research Council	53.9	100	100	20.4	47.4
Directorate of Research and Statistics FBR	52.9	60.1	100	42.9	52.6
PPARC Establishment Division	49.5	18.3	100	45.8	44.7
Capital Development Authority	45.6	100	100	36.9	20.3

As stated in Table 6, the State Bank of Pakistan again tops the list in this instance of BDRCI, which is currently used as a yardstick to assess other FG organizations' big data processing capabilities. PPRCI, which formerly had a 100 OSRCI, is currently at a 49.5 BDRCI due to a weak position in terms of the measurements from Sub-D1, D3, and D4. Because of the poor performance found in Sub-D2 and D3, it was indexed at RCI = 58.0.

Table 7. Big data processing relative capacity indicator (provincial departments).

Department Name	BDRCI	Sub-D1 RCI	Sub-D2 RCI	Sub-D3 RCI	Sub-D4 RCI
Provincial Disaster Management Authority Punjab	100	100	100	100	100
DG Public Relations Punjab Lahore	80.8	100	100	49.8	90.4
PITB Citizen Feedback Monitoring Program	79.1	70.6	100	49.8	100
Punjab Proc. Regularity Authority (PPRA)	77.7	89.8	100	68.2	85
Director General Health Punjab	75.9	94.5	100	72.2	67.2
Crop Reporting Service	71.1	28.8	100	94	49.9
Pakistan Kidney & Liver Institute	70.8	60.1	100	84.7	67.2
Livestock and Dairy Development Department	67.5	100	100	100	12.1
Literacy & Non-Formal Basic Education Dept.	65.7	100	80.5	66.2	52.6
Bureau of Statistics Punjab	63.6	60.1	80.5	78.7	67.2
Excise Taxation and Narcotics Control Dept.	63.3	70.6	100	22.4	82.3

According to the BDRCI for PG organizations (Table 7), the Provincial Disaster Management Authority (PDMA) has the largest big data processing capacity in the Punjab provincial government. PRP and CFM are weak in Sub-D3 with an RCI of 49.8. Tables 6 and 7 reveal that FG OUs are weak in Sub-D3 (large data workings) and PG Ous are weak in Sub-D4. Figure 7 displays the violin plot of RCI for overall capacity, official statistics capacity, big data processing capacity, and big data processing sub-capacity dimensions. This graphic presents boxplots and kernel densities for each statistic. The model's violin plot indicates that all FG and PG OUs have an average RCI of 40 and a normal kernel density. OS capacities have an RCI violin with a flat kernel density curve of 40. Big data processing capacity averages 50 with less variation. The lowest RCI for Sub-D2 OS modules indicates that data sharing between departments is limited. The Sub-D2 in the BD violin plot highlights the importance and value of big data sources to public sector organizations.

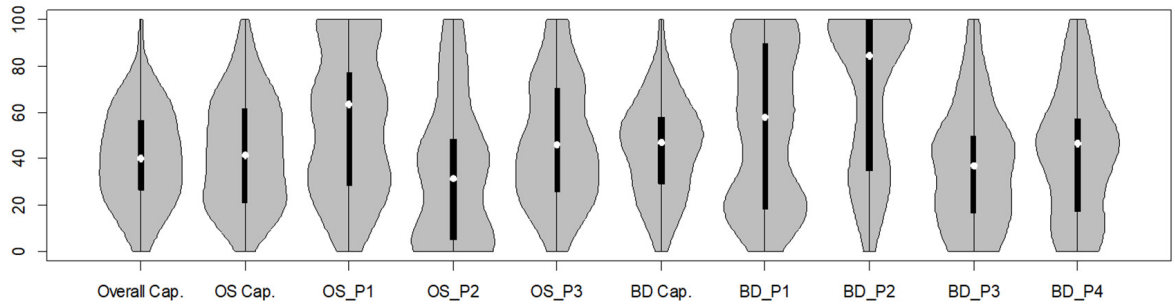

Figure 7. RCI violin plots for various full and partial capacity measures.

4. Conclusions

The growing importance and demand for high-quality, independent official statistics provide an excellent opportunity to invest in strengthening the national statistical systems in a meaningful and long-lasting way. To fulfill the modern data requirements for effective government administration, national statistical institutes are providing an increasing range

of official statistics from both conventional and contemporary sources, including the use of artificial intelligence. Additionally, it is more crucial to strengthen the NSS at the ground level and equip this system with all data producers, whether statistical or non-statistical institutions; in other words, to build the capacity of all (public sector) data producers in their grey areas. UN agencies are also working to strengthen their statistical capacity to produce and use data for better policy formulation and approaches to implement the 2030 development agenda.

In this case, a SWOT analysis of the national statistical system is required to examine the potential grey areas before beginning capacity-building programs. But there is no such scale that can be used to gauge and rank the nation's public sector organizations according to various criteria. Only the World Bank's statistical capacity index provides an overview of the statistical capacity of around 140 developing nations at the national level. In line with the macro statistics capacity indicator from the World Bank, we created a micro indicator that provides a comprehensive capacitive evaluation of the official statistics system within a nation by taking into account several dimensions related to the official statistics system and large data processing, including the use of artificial intelligence.

Pakistan launched a national census-based survey (SOS-Pak) to gather data from public sector organizations on several different metrics. Large or big digital administrative data sources exist within every third public sector organization. However, only 7% of organizations are currently utilizing these big data sources. Furthermore, it was discovered that the preferred data requirement for the public sector also includes organizational administrative data. This stems from its significant potential, applicability, and usability in governing governmental operations. Nevertheless, it has also been discovered that one-third of Pakistan's public sector organizations are worried about data disclosure controls, which account for 25% of the data that are produced or recorded remaining in storage rather than being used for national statistics. Consequently, this points to a general need for them to increase their capacity for advanced statistical and data disclosure control. In the future, about half of public sector organizations plan to work with big data sources, but on the other hand, roughly three out of every four organizations lack the IT infrastructure and human resources necessary to manage contemporary data sources. Because there is a statistical knowledge gap among current IT personnel, the statistical expertise of IT human resources has also raised concerns about the best use of digital data sources. Organizations that responded to the survey also stated their preferences for training requirements for both structured and unstructured databases. The study's findings also point to the urgent need for public sector organizations to work together on data-related projects. The utilization of big data sources and the production of quality official statistics are hindered by several issues, including insufficient awareness of big data, technological limitations, limited resources, and inadequate skills in advanced statistical and data processing techniques.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/systems11080424/s1, File S1: SOS-Pak Questionaire Final; File S2: final data for logistic regression; File S3: LogPCA Programm.

Author Contributions: Conceptualization, S.W.A., M.H., R.A. and H.A.A.; Data curation, S.W.A. and M.H.; Formal analysis, S.W.A., M.H., R.A. and H.A.A.; Funding acquisition, R.A. and H.A.A.; Investigation, S.W.A., M.H., R.A. and H.A.A.; Methodology, S.W.A., M.H., R.A. and H.A.A.; Project administration, S.W.A., M.H., R.A. and H.A.A.; Resources, S.W.A., M.H., R.A. and H.A.A.; Software, S.W.A., M.H., R.A. and H.A.A.; Supervision, S.W.A., M.H., R.A. and H.A.A.; Validation, S.W.A., M.H., R.A. and H.A.A.; Visualization, S.W.A., M.H., R.A. and H.A.A.; Writing—original draft, S.W.A., M.H., R.A. and H.A.A.; Writing—review and editing, S.W.A.,M.H., R.A. and H.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R323), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data Availability Statement: Supplementary Materials file S1 contains the questionnaire developed for the study. Supplementary Materials file S2 contains the microdata (cleaned and transformed into binary codes for Convex Logistic PCA) of the proposed study, collected from 171 public sector organizations in Pakistan. Supplementary Materials file S3 contains the R code for the proposed indicators.

Acknowledgments: The authors express their gratitude to Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R323), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, D.; Pee, L.G.; Pan, S.L.; Cui, L. Big Data Analytics, Resource Orchestration, and Digital Sustainability: A Case Study of Smart City Development. *Gov. Inf. Q.* **2022**, *39*, 101626. [CrossRef]
2. Andronie, M.; Lăzăroiu, G.; Karabolevski, O.L.; Ștefănescu, R.; Hurloiu, I.; Dijmărescu, A.; Dijmărescu, I. Remote Big Data Management Tools, Sensing, and Computing Technologies, and Visual Perception and Environment Mapping Algorithms in the Internet of Robotic Things. *Electronics* **2022**, *12*, 22. [CrossRef]
3. Pramanik, S.; Bandyopadhyay, S.K. Analysis of Big Data. In *Encyclopedia of Data Science and Machine Learning*; IGI Global: Hershey, PA, USA, 2022; pp. 97–115, ISBN 9781799892205.
4. Zhong, Y.; Chen, L.; Dan, C.; Rezaeipanah, A. A Systematic Survey of Data Mining and Big Data Analysis in the Internet of Things. *J. Supercomput.* **2022**, *78*, 18405–18453. [CrossRef]
5. Guo, J.; Liu, R.; Cheng, D.; Shanthini, A.; Vadivel, T. RETRACTED ARTICLE: Urbanization Based on IoT Using Big Data Analytics the Impact of Internet of Things and Big Data in Urbanization. *Arab. J. Sci. Eng.* **2022**, *48*, 4147. [CrossRef]
6. Ateya, A.A.; Sayed, M.S.; Abdalla, M.I. Multilevel Hierarchical Clustering Protocol for Wireless Sensor Networks. In Proceedings of the 2014 International Conference on Engineering and Technology (ICET), Cairo, Egypt, 19–20 April 2014; pp. 1–6.
7. Ateya, A.A.; Algarni, A.D.; Hamdi, M.; Koucheryavy, A.; Soliman, N.F. Enabling Heterogeneous IoT Networks over 5G Networks with Ultra-Dense Deployment—Using MEC/SDN. *Electronics* **2021**, *10*, 910. [CrossRef]
8. Wang, J.; Xu, C.; Zhang, J.; Zhong, R. Big Data Analytics for Intelligent Manufacturing Systems: A Review. *J. Manuf. Syst.* **2022**, *62*, 738–752. [CrossRef]
9. Rogge, N.; Agasisti, T.; De Witte, K. Big Data and the Measurement of Public Organizations' Performance and Efficiency: 450 The State-of-the-Art. *Public Policy Adm.* **2017**, *32*, 263–281. [CrossRef]
10. Abbas, S.W.; Ahmad, M.; Rasul, S. Leveraging Big Data for Official Statistics: Some Recent Developments. *Adv. Appl. Stat.* **2019**, *54*, 99–136. [CrossRef]
11. Mc Cartney, A.M.; Anderson, J.; Liggins, L.; Hudson, M.L.; Anderson, M.Z.; TeAika, B.; Geary, J.; Cook-Deegan, R.; Patel, H.R.; Phillippy, A.M. Balancing Openness with Indigenous Data Sovereignty: An Opportunity to Leave No One behind in the Journey to Sequence All of Life. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2115860119. [CrossRef] [PubMed]
12. Mills, D.; Pudney, S.; Pevcin, P.; Dvorak, J. Evidence-Based Public Policy Decision-Making in Smart Cities: Does Extant Theory Support Achievement of City Sustainability Objectives? *Sustainability* **2021**, *14*, 3. [CrossRef]
13. Telleria, J.; Garcia-Arias, J. The Fantasmatic Narrative of 'Sustainable Development'. A Political Analysis of the 2030 Global Development Agenda. *Environ. Plan. C Politics Space* **2022**, *40*, 241–259. [CrossRef]
14. Jutting, J. *Capacity Building, Yes—But How to Do It?* United Nations World Data Forum: New York, NY, USA, 2016.
15. Lebada, A.M. *Member States, Statisticians Address SDG Monitoring Requirements*; IISD Knowledge Sharing Hub: Winnipeg, MB, Canada, 2016; Volume 8.
16. Ardiansyah, A.; Ilyas, A.; Haeranah, H. Reformulation of Statistical Data Sources: Big Data New Data Sources Supporting Future Official Statistics. *Injurity* **2023**, *2*, 424–443. [CrossRef]
17. Alshahrani, A.; Dennehy, D.; Mäntymäki, M. An Attention-Based View of AI Assimilation in Public Sector Organizations: The Case of Saudi Arabia. *Gov. Inf. Q.* **2022**, *39*, 101617. [CrossRef]
18. Chohan, S.R.; Hu, G. Strengthening Digital Inclusion through E-Government: Cohesive ICT Training Programs to Intensify Digital Competency. *Inf. Technol. Dev.* **2022**, *28*, 16–38. [CrossRef]
19. Li, X.; Liu, H.; Wang, W.; Zheng, Y.; Lv, H.; Lv, Z. Big Data Analysis of the Internet of Things in the Digital Twins of Smart City Based on Deep Learning. *Future Gener. Comput. Syst.* **2022**, *128*, 167–177. [CrossRef]
20. Abbas, S.W.; Rasul, S.; Ahmad, M. Unreported Data Sources in Public Sector Organizations. *Stat. J. IAOS* **2019**, *35*, 359–370. [CrossRef]
21. *United Nations 460 Economic Commission for Europe-UNECE Questionnaire about the Skills Necessary for People Working with Big Data in the Statistical Organisations*; UN: New York, NY, USA, 2014.
22. *UN Global Working Group on Big Data for 462 Official Statistics-UNSD Analysis of Big Data Survey 2015 on Skills, Training and Capacity Building*; UN: New York, NY, USA, 2015.
23. *United Nations Sta-464 Tistics Division-UNSD Results of the UNSD/UNECE Survey on Organizational Context and Individual Projects of Big Data*; UN: New York, NY, USA, 2015.

24. Zekos, G.I. Risk Management Developments. In *Economics and Law of Artificial Intelligence*; Springer International Publishing: Cham, Switzerland, 2021; pp. 147–232, ISBN 9783030642532.
25. Ogrean, C. Relevance of Big Data for Business and Management. Exploratory Insights (Part II). *Stud. Bus. Econ.* **2019**, *14*, 169–180. [CrossRef]
26. Landgraf, A.J.; Lee, Y. Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. *J. Multivar. Anal.* **2020**, *180*, 104668. [CrossRef]