BigSurv18
BIG DATA MEETS SURVEY SCIENCE
OCTOBER 25 - 27, 2018 BARCELONA, SPAIN

Measuring the Official Statistics Capability of Public Sector Organizations in Presence of Big Data Sources

S. Wasim Abbas[1], Sajid Rasul[2] and Munir Ahmad[1]

[1]National College of Business Administration & Economics, Lahore, Pakistan
[2]Bureau of Statistics Punjab, Pakistan

ABSTRACT:

Big Data sources require modern tools for data processing with advanced statistical care and without this it is troublesome to optimally utilize these sources. Here, we have measured the Official Statistics Production (OSP) and Big Data Processing (BDP) capability of Public-sector organizations in Pakistan. A survey has been conducted at national level to capture the dimensions related to key indicators about OSP and BDP. Capability Indicators are developed to rank the public-sector organizations using Convex Logistic Principal Component Analysis as dimensionality reduction tool. Results can be used to explore the sectors /organizations need capacity building aside with to remove barriers and to resolve issues in the adoption of Big Data sources.

METHODOLOGY:

Census Based Survey
Questionnaire (Mail Inquiry)
March 17—Aug 17

Survey of Official Statistics Production Pakistan (SOS-Pak)

Population: 758
Respondents: 171 (22.6%)

Federal Government Organizations 49/472 (10%)

Provincial Government Organizations 122/286 (42%)

KEY SEGMENTS OF QUESTIONNAIRE

A. Informational Panel (IP)
B. Official Statistics Production Information (OS)
C. Big Data use in Official Statistics (BD)
D. Rationalization of Statistical Cadre (RC)

METHODOLOGY FOR OSP & BDP CAPABILITY INDICATORS:

Principal Component Analysis (PCA) is renowned technique helpful in data compression, visualization, and feature exploration. For binary data, this technique was modified as Logistic PCA (Landgraf and Lee 2015).

Using convex logistic PCA, the scores are developed separately for both OSP and BDP capability models for FG and PG departments. The scores are then transformed into Relative Capability Indicator (RCI) using vector transformation as.

$$\underline{\tau} = \alpha\underline{\kappa} + m$$

Here, κ be the vector of PC scores for the *ith* entity. m is the upper record value of PC scores and $\alpha = \pm 1$. RCI is calculated as:

$$RCI = \underline{\tau} / \tau_{(m)} \times 100$$

RCI=100 shows the topmost capable organization in a set of Organizational Entities (OEs) with respect to a given measure/dimension.

Key Indicators and Dimensions of OSP & BDP

A. Official Statistics Production (OSP)

1. Collection/Recording of Data to Dissemination of Data Products
2. Liaison with Statistical and Non-Statistical Departments on Data
3. Data Privacy and Confidentiality

B. Big Data Processing (BDP)

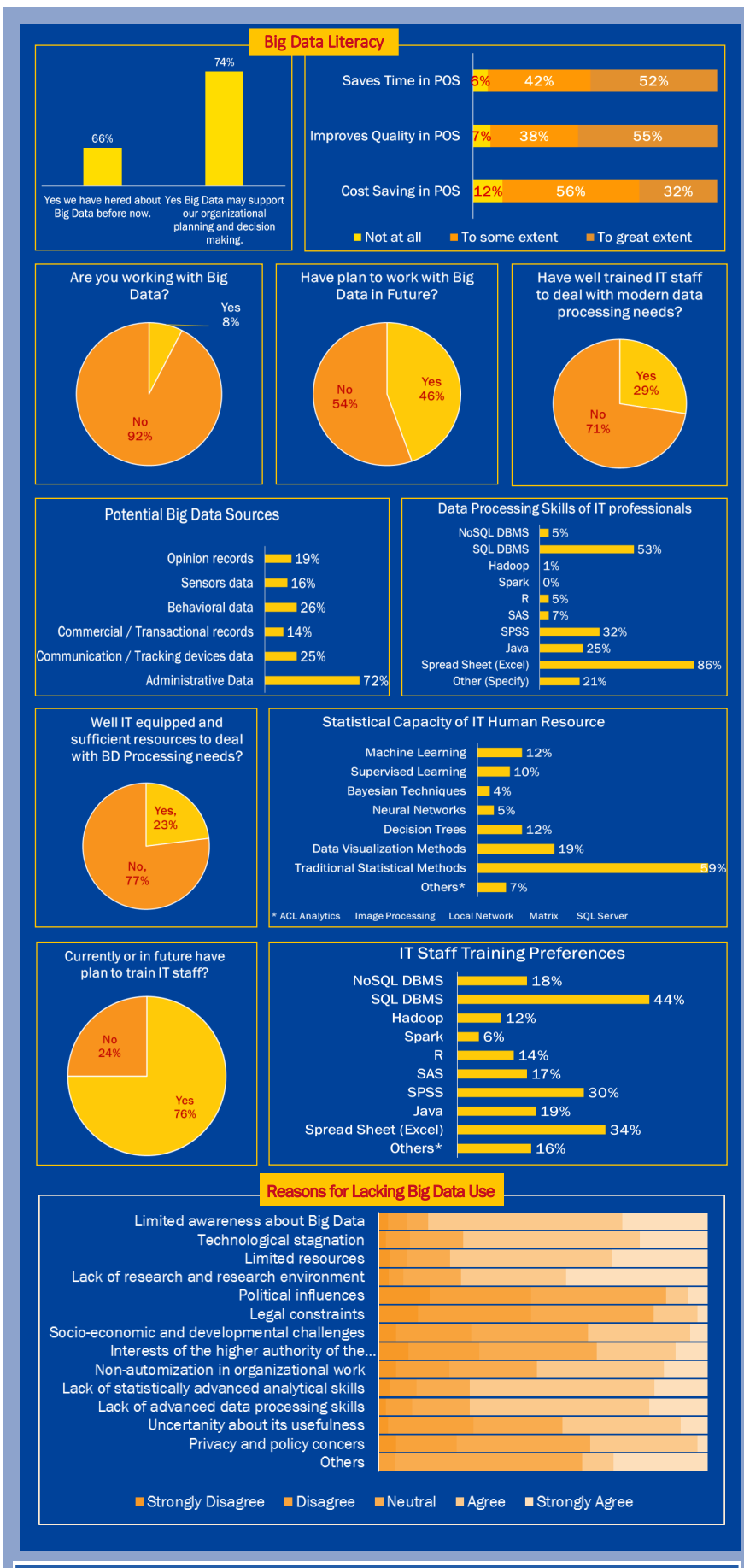
1. Big Data 3Vs
2. Big Data Literacy
3. Big Data Workings
4. Big Data Skills

SURVEY CHARACTERISTICS:

- **Level of Responding Organizations**
  12% Administrative | 45% Attached | 38% Autonomous
- **Respondents Level**
  75% response made by top tire of Public-Sector (BS>16)
- **Responding Organizations Employee size**
  15% (<50) | 33% (50-250) | 24% (250-1000) | 28% (>1000)

SURVEY FINDINGS:

- **Data Recording for Official/Public Use Regularly**
  83% (76% FG and 86% PG)
- **Purpose of Data Recording**
  29% Administrative Use | 10% Statistical Use | 57% Both Purposes
  04% Business/Research Purposes
- **Data Reporting and Dissemination**
  68% OEs Produce Data Products | 87% OEs Disseminate
- **Data Transmission among Statistical and Non Statistical Organizations**
  48% OEs supply data to statistical organizations periodically
  23% OEs acquire data from statistical organizations periodically
- **Self Data Collection Needs**
  39% Public-sector OEs conduct survey to meet data needs
- **Data Confidentiality**
  30% Low | 36% Medium | 34% high



Deviance Explained by Different Data Reduction Techniques for Overall capability indicators and for respective dimensions (k = 1)

Capability Indicators / Dimensions	Measures / Indicators Used (Nos.)	Deviance Explained (%)		
		Exponential Family PCA	Logistic PCA	Convex Logistic PCA
Full Model (OS + BD)	**31**	**18**	**16**	**39**
OS Full Model	13	28	24	55
1. Data Collection to Dissemination of Statistics	5	63	52	83
2. Liaison with other Departments on Data	4	52	41	89
3. Data Privacy and confidentiality	5	52	40	85
BD Full Model	20	23	21	50
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

ACKNOWLEDGEMENTS:

This research is based on indigenous scholarship by Higher Education Commission (HEC) of Pakistan to the first author. Authors are thankful to the secretariat departments of Federal and Punjab Govt. of Pakistan for cooperation in provision of required data.

Authors are thankful to **BigSurv18, ESRA and UPF** for providing an opportunity and offering Travel Grant to present this research in such a knowledge hub.

Authors are especially thankful to Dr. Antje Kirchner (Chair BigSurv18 Organizing Committee) for providing technical support in all aspects, throughout the conference program from day one.