



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

S. Wasim Abbas¹, Sajid Rasul² and Munir Ahmad¹

¹National College of Business Administration & Economics, Lahore, Pakistan

²Bureau 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

Federal Government Organizations
49/472 (10%)

Provincial Government Organizations
122/286 (42%)

Respondents: 171 (22.6%)

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, $\underline{\kappa}$ be the vector of PC scores for the i th entity. m is the upper record value of PC scores and $\alpha = \pm 1$. RCI is calculated as:

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

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)

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

B. Big Data Processing (BDP)

- Big Data 3Vs
- Big Data Literacy
- Big Data Workings
- 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

Big Data Literacy

66% Yes we have heard about Big Data before now.

74% Yes Big Data may support our organizational planning and decision making.

Saves Time in POS	5%	42%	52%
Improves Quality in POS	7%	38%	55%
Cost Saving in POS	12%	56%	32%

Legend: Not at all, To some extent, To great extent

Are you working with Big Data?

Yes 8%
No 92%

Have plan to work with Big Data in Future?

No 54%
Yes 46%

Have well trained IT staff to deal with modern data processing needs?

No 71%
Yes 29%

Potential Big Data Sources

Opinion records	19%
Sensors data	16%
Behavioral data	26%
Commercial / Transactional records	14%
Communication / Tracking devices data	25%
Administrative Data	72%

Data Processing Skills of IT professionals

NoSQL DBMS	5%
SQL DBMS	53%
Hadoop	1%
Spark	0%
R	5%
SAS	7%
SPSS	32%
Java	25%
Spread Sheet (Excel)	86%
Other (Specify)	21%

Well IT equipped and sufficient resources to deal with BD Processing needs?

Yes, 23%
No, 77%

Statistical Capacity of IT Human Resource

Machine Learning	12%
Supervised Learning	10%
Bayesian Techniques	4%
Neural Networks	5%
Decision Trees	12%
Data Visualization Methods	19%
Traditional Statistical Methods	59%
Others*	7%

* ACL Analytics, Image Processing, Local Network, Matrix, SQL Server

Currently or in future have plan to train IT staff?

No 24%
Yes 76%

IT Staff Training Preferences

NoSQL DBMS	18%
SQL DBMS	44%
Hadoop	12%
Spark	6%
R	14%
SAS	17%
SPSS	30%
Java	19%
Spread Sheet (Excel)	34%
Others*	16%

Reasons for Lacking Big Data Use

Limited awareness about Big Data	
Technological stagnation	
Limited resources	
Lack of research and research environment	
Political influences	
Legal constraints	
Socio-economic and developmental challenges	
Interests of the higher authority of the...	
Non-automation in organizational work	
Lack of statistically advanced analytical skills	
Lack of advanced data processing skills	
Uncertainty about its usefulness	
Privacy and policy concerns	
Others	

Legend: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree

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.