Journal of Advances in Science and Technology
Vol. 20, Issue No. 1, March-2023, ISSN 2230-9659

The Impact of Big Data on Statistical Analysis Methods

Bimarsh Jha*

Student, Class: 12th, Welham Boys School

Abstract - Big data's revolutionary effects on statistical analysis are the subject of this dissertation, which also examines the cutting-edge methods and strategies required to realize its full potential and overcome its inherent challenges. The materials and approach engaged with involving enormous information for the factual examination have been talked about in this part. As far as materials, large information alludes to immense, confounded datasets that are regularly excessively tremendous for customary information handling methods to deal with. When dealing with massive amounts of data, distributed technologies and systems like Hadoop and Spark are frequently utilized. The scalability and parallel processing capability of big data technology are crucial to statistical analysis. The impact of big data on statistical analysis techniques is the subject of secondary analysis. The inclusion of big data in statistical analysis methods has revolutionized the field of data analysis.

Keywords - Big data, BDPA, RBV, chain management, Distributed systems, Statistical analysis, PL-SEM strategy, and data mining.

----------X----------

INTRODUCTION

The project is about Big data's revolutionary impact on statistical analysis that has given practitioners and researchers alike both opportunities and challenges. Huge volumes of data are being generated at an unprecedented rate due to the spread of digital technologies and the interconnection of the planet. Traditional statistical analysis techniques, however, face major methodological and computational problems due to the sheer amount, velocity, and variety of big data. This dissertation investigates the revolutionary effects of big data on statistical analysis, examining the cutting-edge procedures and techniques needed to maximize its potential while overcoming its inherent difficulties.

REVIEW OF LITERATURE

In this section, the discussion has been done on the various theory of the researchers from their research papers. According to Dubey *et al.* 2019, this study was inspired by the rapid rise in interest in the BDPA in the literature on manufacturing and operations management. Regardless of the consideration of the two scholastics and specialists, the hypothesis put together exploration with respect to the capability of BDPA in modern execution is as yet deficient. The research paper proposed a theoretical framework based on institutional theory and RBV in order to address the current limitations of RBV. The study's limitations necessitate that its conclusions be cautiously applied to various situations. Generality is necessary because it is extremely challenging, if not impossible, to collect a sample that accurately represents the entire population. is an issue for any overview-based research.

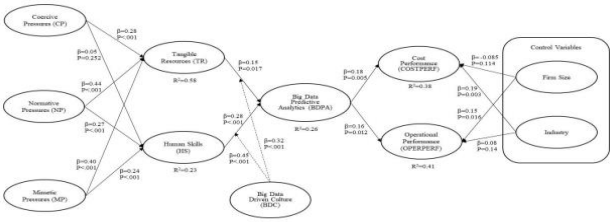

Figure 1: Final PLS Model

(Source: https://kar.kent.ac.uk/)

According to Wood *et al.* 2020, in this essay, researchers present evidence that the growth of employees and BDA talent capabilities are positively correlated. The findings of the research point to a link between organizational human capital, employee development, and sustainable supply chain outcomes. It is feasible with the help of both personnel-driven and data-driven techniques to maintain a sequence of actions in supply chain management. The primary motivator in this situation is the training component, since it may both build skills and close the skill gap among employees. Nowadays, technology is used to position or evaluate every action (BDA).

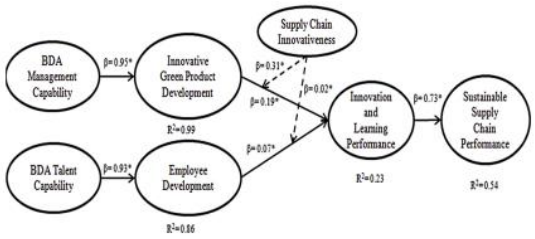

Fig. 2. Tested model (* indicates significance at 5% level).

Figure 2: Tested model (* indicates significance at 5% level)

(Source: https://e-tarjome.com/)

According to Hariri *et al*. 2019, the study has covered a wide range of big data analytics strategies as well as the effects of each technique's uncertainty. Each AI method is first classified as either ML, NLP, or CI. The second column shows how each approach is affected by uncertainty, including both the technique's own inherent uncertainty and the ambiguity in the data. The third column concludes by summarizing the suggested mitigation measures for each uncertainty problem.

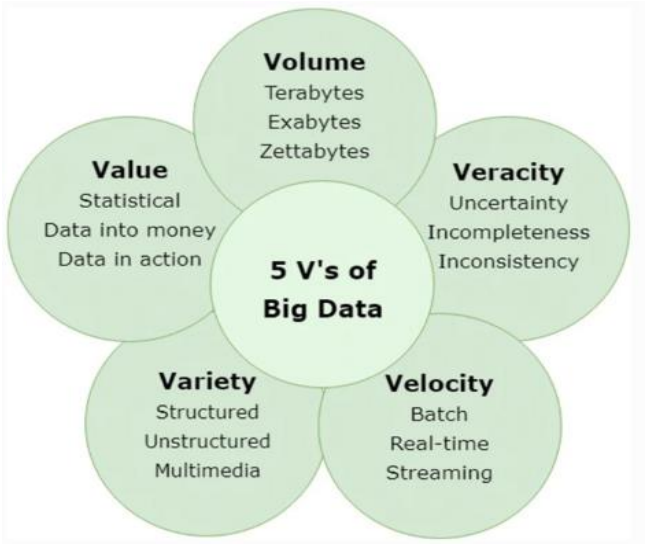

Figure 3: Combined big data characteristics

(Source: https://journalofbigdata.springeropen.com/)

Using an active learning strategy that only uses the data that have been determined to be the most important will help solve the issue of the lack of training data by overcoming this particular type of uncertainty.

MATERIALS AND METHODOLOGY

The emergence of big data has completely changed the area of statistical analysis, giving researchers and analysts the previously unheard-of opportunity to extract essential insights from enormous data sets. The materials and methodology involved in using big data for statistical analysis have been discussed in this section (Hariri *et al.* 2019). In terms of materials, big data refers to vast, complicated datasets that are frequently too huge for conventional data processing techniques to handle. These datasets may contain both structured and unstructured data from social media feeds, sensor readings, and written documents, as well as structured data from databases, spreadsheets, and transactional systems. Big data handling calls for strong computing power, storage space, and specialized software tools capable of quickly processing and analyzing such enormous amounts of information. The big data analysis approach includes a number of crucial components. Primarily, gathering data is vital. Distributed systems and technologies like Hadoop and Spark are frequently used to handle enormous amounts of data. The data is cleaned and transformed into a format that is appropriate for analysis during a preprocessing stage after it has been gathered. Techniques like feature engineering, data integration, and data cleaning can be used.

Statistical analysis heavily relies on big data technology's scalability and parallel processing capability. By distributing computation across numerous nodes or devices, these technologies allow analysts to handle and analyze massive datasets more quickly. Furthermore, the proper use of distributed storage systems guarantees that data can be retrieved and processed effectively even when working with enormous amounts of data (Mikalef *et al.* 2019). Big data is also ideally suited for statistical analysis because of its iterative nature. Before scaling up the analysis to the complete dataset, analysts can undertake exploratory analysis, hypothesis creation, and hypothesis validation on subsets of the data. Within big data, this iterative process aids in model improvement, pattern recognition, and the discovery of significant insights.

RESULTS AND DISCUSSION

The secondary analysis is done on the impact of big data on statistical analysis methods. For getting the result, a secondary analysis was done using the various research papers. In one of the research, for the data analysis, the PLS model is used. It has been made in such a way that met the standards for unwavering quality, legitimacy, and unidimensionality utilizing a three-stage consistent improvement cycle (Boura *et al.* 2019). It was used to check the validity of the constructs used in our study by looking at the average correlation between the items on the scale. The Cronbach's alpha values of the items and the scale were significantly higher than 0.6.18, as shown in the figure below.

Table 1: Inter-correlations among major constructs

	CP	NP	MP	TR	HS	BDPA	COST_PERF	OPER_PERF	BDC
CP	**0.77**								
NP	0.15*	**0.71**							
MP	0.29*	0.13*	**0.93**						
TR	0.15**	0.29*	0.17*	**0.85**					
HS	0.25*	0.20*	0.12*	0.21*	**0.87**				
BDPA	0.02***	-0.05***	-0.02***	-0.10***	-0.02***	**0.73**			
COST_PERF	0.28*	0.19**	0.24*	-0.09***	0.30*	0.07***	**0.71**		
OPER_PERF	0.24*	0.09***	0.08***	-0.12***	0.04***	0.01***	0.21*	**0.73**	
BDC	0.29*	0.21*	0.13**	0.30*	0.18**	-0.03***	0.29*	0.03***	**0.71**

(Source: https://kar.kent.ac.uk/)

At first, exploratory element examination (EFA) was utilized to decide the build legitimacy utilizing head part investigation with varimax turn. Since the number of constructs was determined prior to analysis through a comprehensive review of the literature, the precise number of components to be retrieved was revealed during the analysis (Figure 1). Cross-loaded items were eliminated, and Varimax rotation was then repeated until parsimonious factors were attained. Then, the evaluation was done on the concept of validity and unidimensionality using confirmatory factor analysis.

Another research recommends that the information are broken down utilizing the fluctuation-based Fractional PL-SEM strategy. Multiple correlations between a wide range of variables, including latent variables, can be examined simultaneously using this multivariate data analysis technique. As it augments R2 for the endogenous parts, limits unexplained differences, and supports predominant hypothetical model development, PLS-SEM is reasonable for the exploratory examination of affiliations.

Table 2: Latent variable coefficients

	BMC	BTC	IGPD	ED	ILP	SSCM	SCI	SCI*IGPD	SCI*ED
R-squared			0.995	0.863	0.228	0.536			
Adj. R-squared			0.995	0.863	0.222	0.535			
Composite reliability	0.980	1.000	0.982	0.893	0.997	0.996	0.985	0.999	0.998
Cronbachs' alpha	0.975	1.000	0.975	0.834	0.996	0.996	0.976	0.999	0.998
Avg. var. extracted	0.892	1.000	0.850	0.680	0.976	0.972	0.955	0.984	0.977

(Source: https://e-tarjome.com/)

Latent variable coefficient results are shown in the picture below. The composite reliability was then assessed. It has been determined that the numbers are acceptable because they are over the threshold of 0.70. Based on the p-values in the model, the testing is done on each research hypothesis. In accordance with accepted standards, an alpha value of 0.05 (p 5%) is used to determine statistical significance. ***[Refer to Appendix 1]***

CONCLUSION AND FUTURE SCOPE

In conclusion, the area of data analysis has gone through a revolution thanks to the inclusion of big data in statistical analysis techniques. For statisticians, the sheer amount, speed, and variety of big data have presented challenges and opportunities. Big data has also facilitated the creation of more precise predictive models, and better decision-making techniques, and increased the general effectiveness of statistical analysis. The potential application of big data to statistical analysis techniques is seen in the future. Technological and data management developments will keep encouraging innovation, allowing statisticians to test out innovative methods, take on challenging issues, and draw deeper conclusions from ever-larger and more varied datasets.

RECOMMENDATIONS

By offering academics a wealth of data to go over, big data has completely changed the statistical analysis industry. Statistical techniques can now be used more effectively and precisely due to the ability to acquire and process enormous amounts of data (Wamba *et al.* 2019). In order to extract useful insights from large data, techniques like machine learning, data mining, and predictive modeling have become crucial tools. These developments have substantially improved our comprehension of complicated events and aided in the facilitation of evidence-based decision-making across a number of fields.

REFERENCE

1. Dubey, R., Gunasekaran, A., Childe, S.J., Blome, C. and Papadopoulos, T., 2019. Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. *British Journal of Management*, *30*(2), pp.341-361.
2. Bag, S., Wood, L.C., Xu, L., Dhamija, P. and Kayikci, Y., 2020. Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. *Resources, Conservation and Recycling*, *153*, p.104559.
3. Luo, J., Wang, Z., Xu, L., Wang, A.C., Han, K., Jiang, T., Lai, Q., Bai, Y., Tang, W., Fan, F.R. and Wang, Z.L., 2019. Flexible and durable wood-based triboelectric nanogenerators for self-powered sensing in athletic big data analytics. *Nature communications*, *10*(1), p.5147.
4. Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), pp.1-16.
5. Raut, R.D., Mangla, S.K., Narwane, V.S., Gardas, B.B., Priyadarshinee, P. and Narkhede, B.E., 2019. Linking big data analytics and operational sustainability practices for sustainable business management. *Journal of cleaner production*, *224*, pp.10-24.
6. Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data

analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), pp.1-16.
7. Wamba, S.F., Dubey, R., Gunasekaran, A. and Akter, S., 2020. The performance effects of big data analytics and supply chain ambidexterity: The moderating effect of environmental dynamism. *International Journal of Production Economics*, *222*, p.107498.
8. Mikalef, P., Krogstie, J., Pappas, I.O. and Pavlou, P., 2020. Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities. *Information & Management*, *57*(2), p.103169.
9. Mikalef, P., Boura, M., Lekakos, G. and Krogstie, J., 2019. Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. *British Journal of Management*, *30*(2), pp.272-298.

APPENDICES

Appendix 1: Latent variable coefficients

	BMC	BTC	IGPD	ED	ILP	SSCM	SCI	SCI*IGPD	SCI*ED
R-squared			0.995	0.863	0.228	0.536			
Adj. R-squared			0.995	0.863	0.222	0.535			
Composite reliability	0.980	1.000	0.982	0.893	0.997	0.996	0.985	0.999	0.998
Cronbachs' alpha	0.975	1.000	0.975	0.834	0.996	0.996	0.976	0.999	0.998
Avg. var. extracted	0.892	1.000	0.850	0.680	0.976	0.972	0.955	0.984	0.977

(Source: https://e-tarjome.com/)

Corresponding Author

Bimarsh Jha*

Student, Class: 12th, Welham Boys School