Article review: Statistics in Age of Big Data; An Overview of the Most Important Tools and Techniques

Hadiya H. Matrood
Esmaeel Ali Alselmawi
Department of Statistics, College of Administration and Economics, University of Wasit, Wasit, Iraq

مقال مراجعة: الإحصاء في عصر البيانات الضخمة؛ نظرة عامة على أهم الأدوات والتقنيات

م. م. هدية حسن مطرود
م. م. إسماعيل علي السلماوي
قسم الإحصاء – كلية الإدارة والاقتصاد – جامعة واسط

Abstract

This paper offered an overview of the concept of big data and the most significant statistical problems surrounding it, and contemporary methods applied to its analysis like systematic models, dimension reduction methods, and distributed analysis with Hadoop and Spark. Practical applications in many areas including medicine, digital marketing, banking, astronomy, and official statistics, were also considered in the review, where it was noted that modern statistical techniques have increasingly found themselves useful in helping in decision-making. In brief, the contemporary statistical tools in the big data analysis can be sorted into multiple major directions: advanced statistical models (e.g., generalized regression, multilevel models), inferences of automatic model selection, partial sample estimates, parallel and simultaneous procedures, a combination of statistics with interpretive machine learning, network and time analysis, and the utilization of privacy-protective tools. These approaches are highly capable in addressing the peculiarities of big data such as pseudo-correlations, data noise, complexity, data structuralism, and computational bottlenecks that enable statisticians and data scientists to derive accurate information and make sound decisions on the basis of sound scientific principles and multiple and massive data volumes. **Keywords:** Big Data, Statistics, Machine Learning, Hadoop and Spark, Predictive Modeling

Introduction

Over the last couple of decades, the world has been facing an unprecedented surge in the amount of data that has resulted in the development of the notion of big data as one of the most prominent ones in the sphere of data science and applied statistics. Manyika et al. (2011) have identified that data generation rate has grown exponentially as compared to storage and analysis capabilities of conventional systems. In this view, there is need to come up with new statistical tools and methods which can handle the quantity, speed as well as heterogeneity of current data. The purpose of this review was to investigate the most notable statistical instruments and methods that relate to big data, give some practical examples, and review peer-reviewed research on this topic.

Concept of big data

The emergence of information systems, the growing popularity of the internet and social media networks, has contributed to a substantial growth of the amount of data related to corporate operations. Thus, it is important that this available data is processed in order to assist in operational, tactical, and strategic decisions. This paper was written with the aim of introducing the notion of big data and major technologies that can be used to analyze voluminous amounts of data. The article discusses the possibilities of big data by analyzing nine fields of operation, including finance, retail, healthcare, transportation, and agriculture.

Energy, production, public sector, media and entertainment

Big data means that the data can be described as having a high volume, variety, and high velocity and these aspects are considered five dimensions (Almeida, 2018). Laney (2001) stated that these dimensions are the key to the comprehension of the difficulties in large-scale data analysis.

Statistical challenges in big data

Big data is a complicated world where the dimensional problems, diversity of sources, quality of data, and computational limitations come into conflict and thus statistical analysis is more challenging as compared to the traditional environment. According to Smith and Nichols (2018), such sophisticated programs like human brain imaging produce very large quantities of highly correlated data, which are problematic to correlation models and causal inference. According to Fan et al. (2014), the problem of dimensional height is even more acute in cases where the amount of variables far surpasses the amount of hits which results in unstable estimates and challenges in choosing the best models. Conversely, Chung (2018) notes that massive brain networks use statistical procedures that have the capability of processing nonlinear association, and intricate network structures.

The growing diversity of data demands a range of analytical methods which are able to include textual, visual, structural and unstructured data and the necessity to combine these types into one analytical system (Aggarwal, 2015). When the data contains varying degrees of noise or variation in the accuracy of measurements, the challenges are increased. Meanwhile, the issue of missing data is one of the most popular ones since it may be caused by sensor malfunctions, unfinished inputs, or transmission malfunctions that necessitate the use of advanced models to address missing data (Little and Rubin, 2019). The recent literature identifies the increasing necessity to create more efficient computation algorithms that can work with distributed settings and the necessity to implement the interpretable machine learning techniques that can provide the decisions that are premised on the clear scientific basis.

Moreover, Chen and Zhang (2014) revealed that scalability was among the most critical concerns in big data environments when the volume of the data grows, the computational architecture must be flexible to process the data in a timely and efficient manner, thus necessitating the creation of parallel algorithms, which can be applied in distributed systems. Other scholars noted that one of the most inherent problems is the combination of data of multiple and heterogeneous sources, in particular, the combination of textual

data and numerical data or sensory data within one analytical system (Gandomi and Haider, 2015). This demands models which can comprehend the context and can resolve any discrepancies between various sources.

Conversely, Kitchin (2014) documented that data quality constitutes the key factor determining the success of any statistical analysis because the big data is likely to contain errors, redundancy, and inconsistency so that before actual analysis, advanced cleansing and conversion methods must be used. Combined, these issues demonstrate that big data management does not only entail using powerful computational instruments, but a comprehensive approach to methodology is also needed that guarantees the quality of inputs, data source uniformity, and scale log so that researchers can obtain detailed knowledge and base their decisions on sound scientific principles. Table 1 is a summary of comparison of big data overall statistical challenges.

Tools and techniques for distributed analysis of big data

One of the most significant pillars in the big data ecosystem is distributed storage and analysis systems, which include Hadoop and Apache Spark as Hadoop offers the framework of distributing data across multiple nodes via the Distributed File System (HDFS) enabling the processing of large amounts of data (Shvachko et al., 2010), and Spark is used alongside In-memory data, processing is identified to be faster at making advanced statistical computations (Zaharia et al., 2016). Other research indicated that due to its size, speed, and variety, the big data offers sufficient opportunities of developing new statistical and analytical models, yet is confronted with great challenges including high dimensions, more noise, pseudocortices, and storage bottlenecks (Fan et al., 2014; Wang et al., 2016).

Table 1: Comparison of general statistical challenges in big data

Challenge	Description	reference
Dimensional Height	The number of variables is greater than the number of hits	Fan et al. (2014)
noise	Incorrect or abnormal values	Aggarwal (2015)
Lack of data	Incomplete data that affects the analysis	Little and Rubin (2019)
Scalability	Data processing is difficult as volume, speed, and complexity increase	Chen and Zhang (2014)
Data Integration	The challenges of integrating different data sources into an integrated analytical structure	Gandomi and Haider (2015)
Data Quality	Unbalanced quality and the presence of incomplete or contradictory values affect the modeling	Kitchin (2014)
Computational complexity	The need for efficient algorithms that are capable of processing big data in a timely manner	Chen and Zhang (2014)
High	Highly correlated neural data that	Smith and Nichols

interdependence on data	makes statistical modeling difficult	(2018)
Excellent network complexity	Difficulty in Analyzing Large Networks with Nonlinear Relationships in Brain Data	Chung (2018)

Reality usages of big data like digital marketing-they show that they can potentially provide an in-depth insight into consumer behavior and tailor the marketing experience, yet there are other challenges, including how personalization affects consumer trust, that should be researched on a wider scale (Theodorakopoulos and Theodoropoulou, 2024). The significance of big data does not end with the business sector, as it has also had an effect on official statistics, whereby the necessity to combine varied and varied sources of big data has resulted in the role of the statistical institution being redefined, as well as the growing interaction between government and academia (Struijs et al., 2014).

Regarding the methodology, studies indicate that the big data setting has resulted in the change of the traditional methods of statistics as new aspects of high-dimensional data visualization, multi-tasks, network analysis, automated model selection, and multi-source integration have become a major component of modern analysis (Galeano and Peña, 2019). In practice, including sensor data analysis, traffic, and social networks, it is evident that the potential of big data is enormous yet needs advanced tools to clean, integrate and distributed analysis in the flexible computer architectures (Daas et al., 2015). The re-definition of statistical and computational behaviors in sectors is changing the work of big data, necessitating the creation of distributed algorithms, scalable models and computational structures to enable integration of different sources of data. In Table 2, we will be able to summarize the key concepts in the literature on big data.

Table 2: A Summary of key ideas in big data literature

axis	Brief Explanation	reference
Distributed analysis systems	Hadoop (HDFS) for distributed data processing Spark for in-memory processing	Shvachko et al. (2010); Zaharia et al. (2016)
Recent statistical developments	New approaches to complexity of big data	Wang et al. (2016)
Mathematical and statistical challenges	High dimensions, noise, memory limitation, false correlation	Fan et al. (2014)
Using big data in marketing	Models for understanding consumer behavior and personalization development: research gaps	Theodorakopoulos and Theodoropoulou (2024)
Big data and official Statistics	Changing the role of official statistical institutions and increasing cooperation	Struijs et al. (2014)
Changes in statistical methods	Network analysis, Miscellaneous Models, multi-experiments, Data source integration	Galeano and Peña (2019)

Practical applications in big data	Traffic data, sensors, and social networks	Daas et al. (2015)

Modern statistical techniques

The dimensional reduction tools that are provided by modern statistical approaches to the big data, like PCA and SVD, are necessary to decrease the involved computation (Jolliffe, 2011). Such regular models as LASSO and the ridge regression are also significant when it comes to variable selection and overgeneralization (Tibshirani, 1996). Chen et al. (2019) proved that the significance of the joint use of statistics and machine learning algorithms like Random Forest and XGBoost to obtain higher predictive performance.

Modern statistical methods used in big data

With big data, the conventional statistical tools are not adequate to handle the large volume of data, the complexity and diversity of resources, but the more modern and modernized techniques need to be used to process classical statistics, high-performance computing, and machine learning. Indicatively, Yoo et al. (2014) have shown how current statistical tools may be integrated with machine learning instead of big data Analyze medicine, in which predictive models are capable of showing true multivariate health patterns and the factors that drive this clinical outcome. Similarly, Lengauer (2020) mentions the usage of multilevel models, identical regression models, and automated model selection methods to address the issue of high correlation and overlap between variables that are typical of large multidimensional datasets.

As Wang et al. (2016) identified, there was the evolution of advanced statistical techniques such as generalized regression models, subset sampling techniques, and parallel techniques to expedite calculations in big data, and the combination of statistical algorithms with high-performance computing techniques to present decent processing time despite millions of observations and variables. Digital databases and statistical analysis and forecasting methods are applied in the domain of finance, where Jumayev (2025) also evaluates the credit risk of customers with high accuracy and takes the possibilities of modern models to anticipate the financial behavior, basing on all the numerous and more complex data. Moreover, Ajala et al. (2024) indicate the relevance of statistical means in complementing privacy-protective methods in the analysis of big data, particularly where sensitive or personal data are involved, in a manner such that models generate both the analysis of accuracy and protection of information. According to Faaique (2024), modern statistical analysis is used in some areas of astronomy, where some techniques are used to derive patterns and predict further phenomena out of large and heterogeneous data, such as observations of numerous telescopes. This study recommends that in the big data management engineering, modern statistical techniques should be used, such as integration of various data sources, time analysis of data, network analysis, and forecasting of models to enhance decision making in a complex environment (Theodorakopoulos et al., 2024).

Ethical aspects

A significant ethical concern that must be addressed with big data is the issue of privacy, data security, and bias in the algorithm. According to Verma (2017), the statistical models and machine learning methods can unconsciously foster discrimination or prejudice when they are performed without fairness and transparency, which is particularly risky when a set of databases is merged or predictive analysis of more sensitive data is provided. These issues are in line with Ajala et al. (2024) who have indicated that it is necessary to implement privacy-enhancing technologies in analyzing big data to ensure the personal information is not endangered, but the accuracy of the analysis is not affected. The General Data Protection Regulation (GDPR) provided by the European Union also promotes the need to take into consideration the rights of individuals, securing personal information, and its responsible use in particular purposes.

These moral question issues are directly connected to the statistical and analytical problems mentioned above, and newer statistical tools (Yoo et al., 2014; Wang et al., 2016), distributed tools, including Hadoop and Spark (Shvachko, et al., 2010; Zaharia et al., 2016; Bhardwaj et al., 2025). They provide, the capability to process and analyze great quantities of data as fast and efficient as possible, and their negligent application may result in biased outcomes, structural prejudices or even the breach of privacy. Incorporating ethical considerations into the construction of the statistical model, the choice of distributed analysis methods, and the application of privacy-protective strategies have become the key to maintaining the precision, impartiality, and safety of the findings of big data.

Practical aspect

This part is split into three parts where we shall address R language codes of explanatory data extraction in tabular arrangements depending on its usage on various fields of application (Table 3). The aspect gives a summary of the uses of data analytics in many areas including medicine, digital marketing, banking, astronomy, official statistics etc. The nature of the data to be used, analysis purpose, statistical method or algorithm to be applied, and the appropriate reference in the science are established as far as each domain is concerned. Consumer behavior is understood by applying time series analysis and segmentation models in digital marketing. Besides, numerous applications and technologies emphasized the significance of data analytics in enhancing the decision-making and processes within the different sectors. The information analysis and its application to the decision-making process in the era of big data has become one of the most significant foundations of improvements in multiple spheres. In medicine, banking, astronomy, engineering, and many others, statistical methods and machine learning algorithms are applied to derive correct insights and enhance performance. This is just some of the notable uses of data analytics in different types of activities including credit risk forecasting, Vehicle performance monitoring highlights accurate medical diagnoses, and many more. Part of these applications are enumerated further, such as credit risk analysis in the era of big data, where financial institutions are turning to the latest technology to enhance the quality of credit ratings and financial decision-making.

Table 3: Practical applications of big data in various fields

Scope	Data Type	Analytical Purpose	Statistical Techniques / Algorithm	Reference
Medical and Healthcare	Patient Records, Genetic Data, Medical Images	Predicting disease risks and identifying healthy patterns	Predictive Models, Machine Learning, Regular Regression, Random Forest	Yoo et al. (2014)
Digital Marketing	Browsing data, purchase history, social media interactions	Analyze consumer behavior and personalize offers	Time Series Analysis, Segmentation Models, Machine Learning	Theodorakopoulos and Theodoropoulou (2024)
Banks and Credit	Customer Data, Financial Transactions, Digital Records	Credit Risk Forecasting and Decision Making	Logistic Regression, XGBoost	Jumayev (2025)
Stars	Telescopic data, optical time series	Pattern Detection and Prediction of Astronomical Phenomena	PCA, Predictive Models, Neural Networks	Faaique (2024)
Official Statistics	Government Big Data: Traffic, Health, Telecommunications	Improving the quality of official statistics	Time Data Analysis, Multi-Source Integration	Daas et al. (2015)
Vehicle Data Analysis	Automotive Sensor Data (IoT)	Condition monitoring and performance improvement	Spark Streaming, Instant Analysis	Alexakis et al. (2022)
Enterprise Data Engineering and Analysis	Operational data, sensor, network	Project Management and Performance Improvement	Network Analysis, Machine Learning, Interactive Models	Theodorakopoulos et al. (2024)
Data Security and Privacy	Sensitive and encrypted data	Privacy Protection in Analysis	Differential Privacy, PET Technologies	Ajala et al. (2024)

1. Credit risk analysis using big data

Through sophisticated algorithms, the banking sector today uses the outcome of analyzing the millions of existing records of customers in terms of credit risk to predict their risk status. In the example of Jumayev (2025), logistic regression, random forest, and XGBoost are only some of the techniques used. Developing a model that can forecast the default of customers with high accuracy up to 90% accuracy needs to be a big amount of time-related information regarding transactions, revenue and purchasing behavior.

2. Analyze vehicle sensor data using spark

Alexakis et al. (2022) demonstrated the way in which vehicle sensor data (GPS, speed, fuel consumption, and others) is processed with the help of a distributed architecture built on Apache Spark and Hadoop. It is beneficial as it will identify failures at an early stage, enhance the performance of a vehicle, and conserve energy.

3. Predicting diseases using big medical data

Yoo et al. (2014) employed statistical algorithms and machine learning to estimate the likelihood of some diseases in response to the data of thousands of the patients. The methods applied were regression, SVM and neural networks. This assists in enhancing the detection and identification of risk factors that cannot be seen using the traditional analyses.

4. Analysis of telescopes in the world

Faaique (2024) illustrates the methods applied to make astronomical signals out of high-noise big data using PCA and machine learning models.

5. Official statistics based on big data

To enhance population and economic activity estimates, the European statistical offices rely on traffic and social media data (Daas et al., 2015).

Practical example 1: Basic credit data analysis

R Code:

```
# Example of virtual data
set.seed (123)
credit_data <- data.frame (
Income = rnorm (200, 5000, 1500),
 Debt = rnorm (200, 2000, 700),
 Default = Sample [c(0,1), 200, Alternative = TRUE]
)
# Logistic Regression Model
Model < - GLM (default ~ income + debt, data = credit_data, family = binomial)
Summary (model)
# Drawing the relationship between income and debt
Plot of land ($credit_data income, debt credit_data$,
main="The relationship between income and debt",
xlab="income", ylab="debt", pch=19, col="blue")
```

Table 4: GLM Regression Results

Ice	Estimate	Traditional Error	Z value	Pr(>|z|)
(Interception)	-2.362e-01	6.731e-01	-0.351	0.726
Income	1.240e-05	1.003e-04	0.124	0.902
Debt	4.646e-05	2.037e-04	0.228	0.820

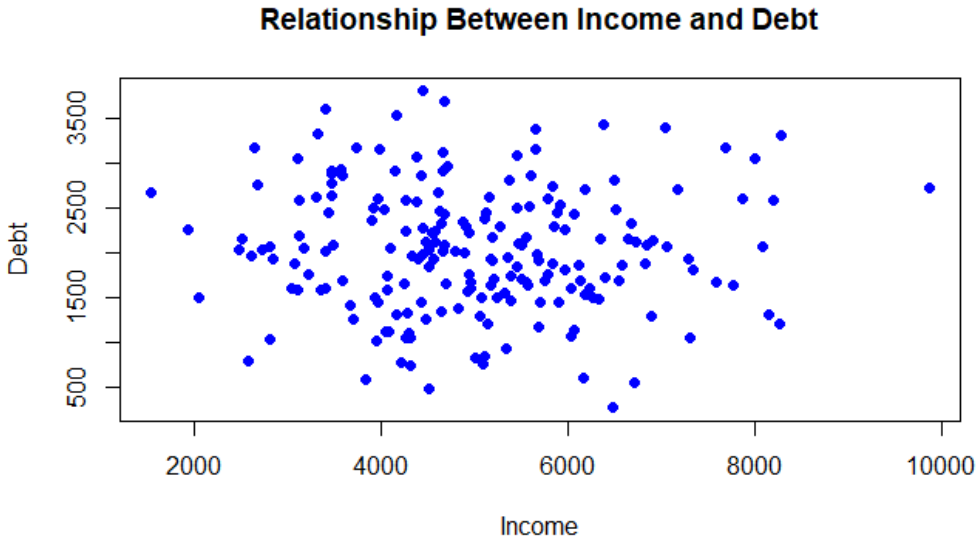

Figure 1: Analysis of the Relationship between Income and Religion

The Figure 1 demonstrates the correlation between the income and debt in thin distribution of data points where the horizontal axis (X) represents the level of income that ranges between 2,000 to 10, 000 units of currency and the vertical axis (Y) represents the amount of debt that varies between 500 to 3, 500 units of currency. It is obvious that the blue dots which reflect the data are random and deprived of patterns There is evident line or curve which means that there is no strong and direct correlation between increasing of income and increasing of debts and vice versa. This allocation of this randomness complies with the findings of the logistics model which demonstrated that both income and debt did not significantly affect the probability of default. This can be explained by other factors not stated in the data like credit history or occupation type which are bigger determiners of the amount of debt or potential default.

Application example 2: Taking virtual medical data through PCA

This is an example where the Key Component Analysis (PCA) was used on hypothetical medical data using the R programming interpreter. Table (5) indicates that the first component (PC1) has the capacity to explain approximately 72.96 percent of the variance of the data and the second component (PC2) has the capacity to explain approximately 22.85 percent. A minor percentage of the variance (3.67% and 0.52% associated with the

third and fourth components (PC3 and PC4) are explained by the latter, indicating that the data would not be lost much in case of reducing it to the two primary components.

R Code:

```
#PCA Example
Data <- Iris (Iris)[,1:4]
pca_model <- prcomp(data, scale.= TRUE)
# View Abstract
Summary (pca_model)
# Diagram Design
Plot (pca_model, main="PCA Scree Plot")
```

The variance distribution of the four components is shown in figure 2. As we are able to observe, the largest value of variance is in the first component (PC1), then there is the second component (PC2) with a small percentage, and the contribution of the third and fourth component is small. This affirms that the bulk of the data in the information is condensed in the first two components and can be used in dimensional reduction in other analyses.

Table 5: Importance of Components (PCA)

	PC1	PC2	PC3	PC4
Standard Deviation	1.7084	0.9560	0.38309	0.14393
Variance ratio	0.7296	0.2285	0.03669	0.00518
Cumulative Ratio	0.7296	0.9581	0.99482	1.00000

The strategy will also be especially beneficial when it comes to medical big data analysis, as it aids in simplifying the management of various variables and enhancing the effectiveness of analytical models.

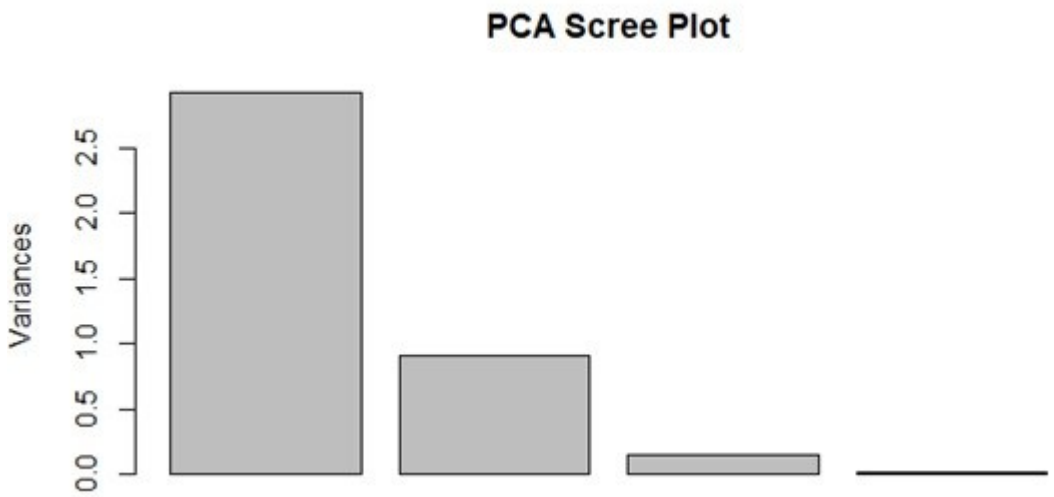

Figure 2: PCA Analysis Skreak Chart

Practical Example 3: Vehicle Sensor Data – Line Chart

In this case, the sensor information on the car was utilized to plot a line graph that indicated the variation in the speed of the car with time. This data was created rather by default with the use of the R programming language, such that the horizontal axis of time displays between 0 to 100 units, and the vertical axis displays the velocity of the vehicle between 0 to 40 units.

R Code:

```
< Vehicle - data.frame(
 Time = 1:100,
 Velocity = cumsum(rnorm(100, 0.5, 1)))
story(vehicle$time, vehicle$speed, type="l",
 main="Vehicle sensor: speed over time",
 xlab="time"، ylab="speed")
```

Figure 3 depicts the line graph of increasing timewise speed of the vehicle and there are slight fluctuations that could be attributed to change in driving conditions or sensor response. The initial speed was nearly equal to zero and it reached approximately 42 points of speed at the terminal period of time (time =100).

Such an analysis can be applied to track the performance of the vehicle and define the presence of any anomalies in the data, i.e., the sudden increase in speed which can be a signal of technical issues or rather unstable driving conditions.

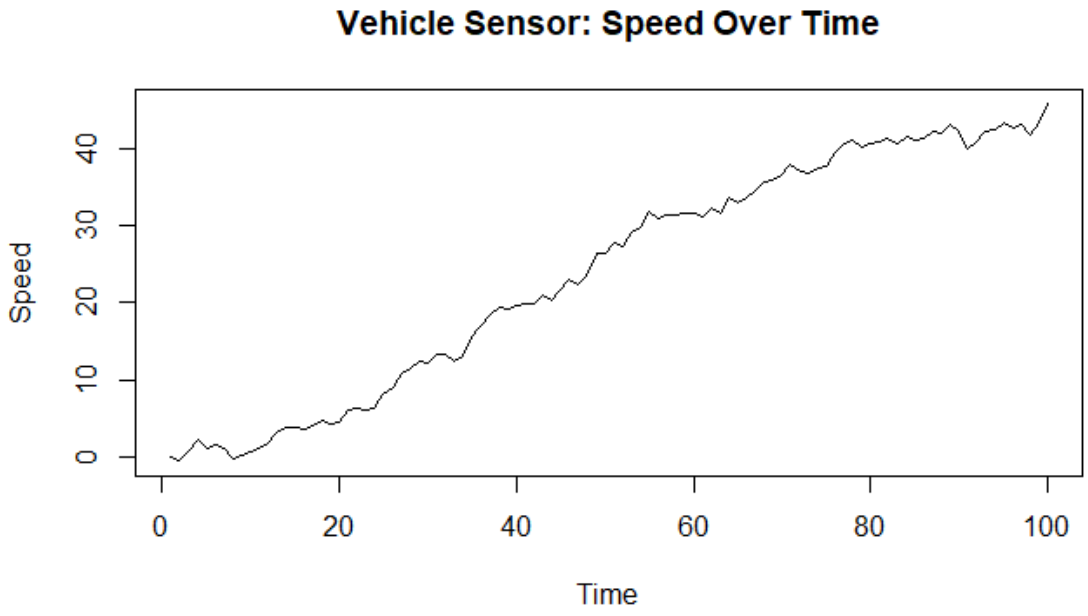

Figure 3: Vehicle speed change over time

Conclusions and recommendations

The review identified that big data possesses characteristics that cripple the performance of traditional practices especially high dimensions, and diversity of data sources, noise and computational complexity. This is also justified by the findings that the current algorithms like PCA, LASSO, and XGBoost assist to enhance the quality of predictive models. It is demonstrated that the employment of distributed architectures like Hadoop

and Spark can offer a friendly atmosphere to process the data on the large scale and the incorporation of the machine learning techniques enhances the efficiency of the forecasting and analysis in the various sectors. This study however suggested that there was a necessity to adopt the contemporary technologies that can be used to easily process big data particularly in the governments and financial institutions that needed to upgrade the abilities of researchers and analysts by training them in distributed analysis tools as well as machine learning methods. There is also the necessity to consider the ethical and privacy issues of data when collecting and analyzing data, particularly in medical and financial applications with the emerging models of the interpretable form that the decision that is made does not rely on the aura of transparency and accuracy. To analyze the data, the quality of the data and standardization of its resources must be improved, which can be done before the analysis.

References

Aggarwal, C.C. (2015). *Data mining: the textbook* (Vol. 1, No. 3). New York: springer.

Ajala, O.A., Arinze, C.A., Ofodile, O.C., Okoye, C.C., and Daraojimba, O.D. (2024). Reviewing advancements in privacy-enhancing technologies for big data analytics in an era of increased surveillance. *World Journal of Advanced Engineering Technology and Sciences*, *11*(1), 294-300.

Alexakis, T., Peppes, N., Demestichas, K., and Adamopoulou, E. (2022). A distributed big data analytics architecture for vehicle sensor data. *Sensors*, *23*(1), 357.

Almeida, F. (2018). Big data: concept, potentialities and vulnerabilities. *Emerging Science Journal*, *2*(1), 1-10.

Bhardwaj, V., Shareef, A.N., Singh, R., Gharban, H.A., and Essa, I.M. (2025). Improving IoT Service Quality with the Allocation of Dynamic Bandwidth. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 1276-1281). IEEE.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.

Chen, M., Mao, S., Zhang, Y., and Leung, V. C. (2014). Big data: related technologies, challenges and future prospects . Heidelberg: Springer. P. 100.

Chen, T., He, T., Benesty, M., and Khotilovich, V. (2019). Package 'xgboost'. *R version*, *90*(1-66), 40.

Chung, M.K. (2018). Statistical challenges of big brain network data. *Statistics and probability letters*, *136*, 78-82.

Daas, P.J., Puts, M.J., Buelens, B., and Hurk, P.A.V.D. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, *31*(2), 249-262.

Faaique, M. (2024). Overview of big data analytics in modern astronomy. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 96-113.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

Galeano, P., and Peña, D. (2019). Data science, big data and statistics. *Test*, *28*(2), 289-329.

Gandomi, A., and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Grover, P., and Kar, A.K. (2017). Big data analytics: A review on theoretical contributions and tools used in literature. *Global Journal of Flexible Systems Management*, *18*(3), 203-229.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science* (pp. 1094-1096). Springer, Berlin, Heidelberg.

Jumayev, B. (2025). Big data: customer credit analysis using digital banking database. *International Journal of Artificial Intelligence*, *1*(2), 1056-1059.

Kambatla, K., Kollias, G., Kumar, V., and Grama, A. (2014). Trends in big data analytics. *Journal of parallel and distributed computing*, *74*(7), 2561-2573.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big data and society*, *1*(1), 2053951714528481.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Lengauer, T. (2020). Statistical data analysis in the era of big data. *Chemie Ingenieur Technik*, *92*(7), 831-841.

Little, R.J., and Rubin, D.B. (2019). *Statistical analysis with missing data*. John Wiley and Sons.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

Rao, T.R., Mitra, P., Bhatt, R., and Goswami, A. (2019). The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems*, *60*(3), 1165-1245.

Shu, H. (2016). Big data analytics: six techniques. *Geo-spatial Information Science*, *19*(2), 119-128.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010, May). The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)* (pp. 1-10). Ieee.

Smith, S.M., and Nichols, T.E. (2018). Statistical challenges in “big data” human neuroimaging. *Neuron*, *97*(2), 263-268.

Struijs, P., Braaksma, B., and Daas, P.J. (2014). Official statistics and big data. *Big Data and Society*, *1*(1), 2053951714538417.

Theodorakopoulos, L., and Theodoropoulou, A. (2024). Leveraging big data analytics for understanding consumer behavior in digital marketing: A systematic review. *Human Behavior and Emerging Technologies*, *2024*(1), 3641502.

Theodorakopoulos, L., Theodoropoulou, A., and Stamatiou, Y. (2024). A state-of-the-art review in big data management engineering: Real-life case studies, challenges, and future research directions. *Eng*, *5*(3), 1266-1297.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Verma, S. (2019). Weapons of math destruction: how big data increases inequality and threatens democracy. *Vikalpa*, *44*(2), 97-98.

Wang, C., Chen, M.H., Schifano, E., Wu, J., and Yan, J. (2016). Statistical methods and computing for big data. *Statistics and its interface*, *9*(4), 399.

Yoo, C., Ramirez, L., and Liuzzi, J. (2014). Big data analysis using modern statistical and machine learning methods in medicine. *International neurourology journal*, *18*(2), 50.

Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A. and Stoica, I. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56-65.