EXPLORING DATA MINING TRENDS AND RESEARCH FOCUS IN INDIA (2018-2022): A COMPREHENSIVE BIBLIOMETRIC ANALYSIS

Abtract

This study comprehensively analyses "Data Mining" research in India from 2018 to 2022. The study utilises data from Scopus and employs R programming with the R-Studio package and Biblioshiny for data analysis. 5211 articles were analysed, and missing data were handled through imputation or removal. The research focuses on articles only and includes basic summary statistics to understand the dataset's characteristics. Applying Bradford's Law confirms its accuracy with an R-squared value of 0.972. The lattice plot of CiteScore publications from 2011 to 2019 shows temporal variations, indicating fluctuating impact for different journals. The grouped bar chart illustrates document distribution among affiliations, emphasising the leading contributors. Funding sponsor analysis highlights the significance of different sponsors in supporting research articles. The study offers meaningful interpretations and addresses research questions related to data mining in India. The results contribute to understanding data mining's impact and trends in the Indian context.

Keywords: Data Mining, Bibliometric Analysis, Data Mining, R Programming, Scopus, Research Trends, India.

Author

Manash Esh
Senior Information Scientist
University Library
University of North Bengal
Siliguri, India
manash@nbu.ac.in

I. INTRODUCTION

Data mining, a fundamental aspect of modern data science, is pivotal in extracting valuable insights and patterns from vast volumes of data. In recent years, India has emerged as a significant player in data mining, contributing to an ever-expanding body of knowledge and research. Understanding the trends and research focus in data mining within the Indian context is paramount to harnessing its potential for various domains, including business, healthcare, finance, and more.

This chapter presents a comprehensive analysis of data mining research in India from 2018 to 2022. Leveraging data from Scopus, a renowned scholarly database, The author has undertaken an extensive investigation to shed light on the development and evolution of data mining studies during these five years. In the digital age, data has emerged as the lifeblood of our interconnected world, driving decision-making processes and shaping the future of industries and economies. Extracting valuable insights from vast datasets has become crucial with the proliferation of information generated from various sources. This practice of extracting knowledge from large volumes of data is known as data mining. In the context of India, a country characterised by its diverse and dynamic landscape, data mining has taken centre stage as a catalyst for innovation, research, and societal development. With its colossal population and rapid digitisation, India is positioned as one of the world's leading data hubs. The country's technological prowess and the ever-expanding digital ecosystem have resulted in an astronomical surge in data generation across various sectors, including finance, healthcare, e-commerce, agriculture, and more. The abundance of data presents opportunities and challenges, making data mining an indispensable tool for harnessing its true potential. The domain of data mining in India extends beyond mere information extraction; it delves into predictive analytics, machine learning, artificial intelligence, and advanced statistical techniques. Researchers and experts in academia, industries, and government sectors are actively exploring novel methodologies and cutting-edge algorithms to make sense of the deluge of data, thereby providing actionable insights and informed decisions that propel progress. One of the critical areas where data mining has played a transformative role is research. Academic institutions and research organisations leverage data mining techniques to conduct studies across various disciplines. Researchers can uncover patterns, correlations, and trends that might have otherwise remained concealed by analysing large datasets. This newfound understanding catalyses groundbreaking discoveries, informs evidence-based policymaking, and fosters innovation across sectors. In healthcare, data mining has revolutionised medical research, allowing scientists to identify risk factors for diseases, design personalised treatment plans, and predict outbreaks of epidemics. In agriculture, data mining aids in optimising crop yields, managing resources efficiently and addressing food security challenges. Moreover, the finance sector benefits from data mining through risk assessment, fraud detection, and improved customer experience. The potential of data mining in India is boundless, and its applications span various domains with societal implications. However, alongside the remarkable opportunities, data mining in India also confronts formidable challenges. Data privacy, security, ethical concerns, and the need for skilled data scientists are among the key obstacles that researchers and policymakers must address. Striking a delicate balance between data utilisation and safeguarding individual rights is paramount to ensure that the benefits of data mining are accessible to all while upholding ethical standards. Throughout this research endeavor, It provides meaningful interpretations of analysis, culminating in addressing key research questions. By uncovering the trends and

research focus in data mining within the Indian context, it aims to contribute to the country's broader understanding of data science advancements.

II. LITERATURE REVIEW

Data mining is valuable for analysing and extracting information from diverse datasets. Several studies have explored its applications and benefits in various domains. **Karmakar (2018)** provides an overview of data mining in library science, emphasising its implications in participative librarianship. The study discusses the advantages of data mining and introduces bibliomining as a useful method for assigning library services. **Geng (2011)** focuses on data mining methods and their applications. The study explores classification, clustering, and their usage in network and business projects, highlighting the effectiveness of data mining technology. **E. Manjula (2016)** delves into data mining techniques for agriculture data analysis. The research aims to develop efficient techniques to solve complex agricultural problems using data mining, contributing to advancements in the field. **Pal and Patet (2015)** review time-series data mining, specifically addressing clustering and noise identification in time-series data. The study introduces the Possibilistic Fuzzy C-Means with Error Prediction (EP) method for effective data clustering. **Kabakchieva et al. (2010)** focus on educational data mining (EDM) and its potential impact on university management. The study aims to optimise data mining methods for analysing historical data, enhancing student behavior understanding and decision-making in education. **Arunachalam (2016)**surveys educational data mining techniques, analysing students' behavioral patterns to improve knowledge acquisition. The study suggests the best categories of EDM tools for practical use. **Govindasamy (2016)** explores learning data mining and its role in improving student performance. The research identifies data mining methods suitable for enhancing student learning and identifies optimal syllabus structures. **Aye Pwint and Khaing Khaing (2019)** discuss data mining's importance in modern manufacturing and education sectors. The paper emphasises data mining's role in optimising applications, including student performance and placements in the education domain. **Wang (2014)** examines data mining's application in personalised university library information systems. The study introduces classical algorithms like FP-growth and K-mean clustering for data mining tasks in library management. **Naheed and Shazia (2011)** focus on data mining techniques for identifying software defects. The research compares various data mining algorithms to find the most effective approach for defect prediction in software bug repositories. **Premysl (2012)**emphasises data mining as a problem-solving tool in science education. The study presents complex and partial data mining tools and their application in Physics Education. **Sarumathi et al. (2014)** undertake a comparative study of diverse data mining tools. The research highlights salient features and performance behavior of various tools, aiding researchers in choosing suitable tools for their tasks.

The reviewed studies demonstrate data mining's versatility and effectiveness in solving problems across different fields, such as library science, agriculture, education, and software development. Data mining techniques continue to evolve and contribute significantly to knowledge discovery and decision-making processes.

III. METHODOLOGY

To conduct the research study analysing the Scopus data on "Data Mining" in India from 2018 to 2022, we followed a step-by-step methodology using R programming with the R-Studio package and Biblioshiny. The search results were exported in a CSV format to 5211 articles on 05/07/2023for further analysis. The data was then loaded into R-Studio using the appropriate R function `read.csv()`. It checked for missing data, handled it accordingly using imputation or removing missing values, and ensured that the data types were appropriate for each variable. The data was filtered to include only articles specified in the document type. Basic summary statistics were computed to understand the dataset's characteristics, such as the total number of articles, mean, median, standard deviation, etc. A time series plot was created to visualise the trend of the number of articles over the five years. The research utilised Biblioshiny and other relevant packages to conduct bibliometric analysis on the articles. Various plots, including co-authorship networks, citation networks, keyword co-occurrence, and author productivity over time, were generated to gain insights into the research landscape. Metrics such as h-index, impact factor, and citation counts were calculated for individual articles and authors. It created various plots and visualisations using R packages like ggplot2, plotly, or lattice to present the analysis findings. It interpreted the results obtained from the analysis to answer the research questions and provided meaningful insights into the trends, patterns, and research focus on data mining in India during 2018-2022. By following this methodology and using the appropriate R code, It conducted a comprehensive research study on data mining in India and generated various plots and insights to answer the research questions effectively.

IV. RESEARCH QUESTIONS

These questions guide the research process and provide a clear focus for the study.

1. **RQ1:** Does the observed data on journal ranks and cumulative frequencies conform to Bradford's Law in scientific publishing? Precisely, how well does the calculated distribution factor and zone factor align with the established patterns of productivity distribution among journals, as described by Bradford's Law?

2. **RQ2:** To what extent do authors' distribution and document counts in the given dataset align with Lotka's Law? Specifically, does the observed number of authors for each number of documents follow the inverse square relationship predicted by Lotka's Law, and how well does the expected number of authors based on the law match the actual distribution?

3. **RQ3:** How do the CiteScore values for different journals vary from 2011 to 2019, and what insights can be drawn regarding the publication trends and impact of these journals based on the lattice plot analysis?

4. **RQ4:** How is the distribution of research documents among different affiliations, and what insights can be gained from the grouped bar chart analysis regarding the research output and contributions of various affiliations in the specified domain?

5. **RQ5:** What insights can be gained from the grouped bar chart analysis regarding the distribution of research documents among different funding sponsors and their research output?

V. ANALYSIS AND INTERPRETATION

It refers to examining data, information, or research findings to gain insights, conclude, and make meaningful sense of the presented information. It is crucial in research, data analysis, and decision-making across various fields.

Table 1: Journal Ranking and Frequency Statistics

Journal	Rank	Freq	Cumfreq
International journal of recent technology and engineering	1	274	274
International journal of innovative technology and exploring engineering	2	273	547
Journal of advanced research in dynamical and control systems	3	222	769
International journal of engineering and technology(uae)	4	157	926
International journal of engineering and advanced technology	5	139	1065
International journal of scientific and technology research	6	112	1177
International journal of advanced science and technology	7	90	1267
Journal of computational and theoretical nanoscience	8	68	1335
Ieee access	9	65	1400
Cluster computing	10	57	1457

The provided data represents a list of scientific journals and their ranks, frequencies, and cumulative frequencies. Let's analyse the data and perform some calculations.

1. **Ranks:** The "Rank" column indicates the position or ranking of each journal in the list. The journals are sorted in descending order based on their frequencies.

2. **Frequencies:** The "Freq" column represents the number of occurrences or articles published in each journal. The frequencies are sorted in descending order, corresponding to the ranks.

3. **Cumulative Frequencies:** The "cumFreq" column shows the cumulative sum of frequencies up to a particular journal. For example, the cumulative frequency of the first journal is 274, which means that the first journal and the journals ranked below it have a total frequency of 274.

Calculations related to Bradford's Law:

1. **Core:** The core is the frequency of the most productive journal in a field. In this case, the core is the frequency of the first journal, 274.

2. **Distribution Factor:** The distribution factor is the ratio of the total number of articles to the core. It represents the average productivity of journals beyond the core. To calculate the distribution factor, sum up all the frequencies and divide it by the core: In this case, the total frequency is 1494 (sum of all the frequencies), and the distribution factor is approximately 5.45 (1494 / 274).

3. **Zone Factor:** The zone factor represents the decrease in productivity as we move away from the core. It is the distribution factor divided by the rank of the core journal. To calculate the zone factor: In this case, the zone factor is approximately 2.725 (5.45 / 2).

Table 2: Regression Model Parameters and Goodness-of-Fit

Parameter	Value
Slope	0.7061789
Intercept	5.763804
R-squared	0.972403

The above table 1 displays the estimated parameters obtained from the analysis. Additionally, here's a brief explanation of the other steps mentioned in the R code:

1. The coefficient of determination (R-squared) is calculated and found to be 0.972403. R-squared represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
2. A new rank sequence `new_rank` ranging from 1 to 30 is created for prediction purposes.
3. Using the estimated parameters (slope and intercept), the `predicted_cumFreq` is calculated based on the equation `exp(intercept) * new_rank^slope`.
4. A plot compares the observed and predicted cumulative frequencies. Black circles represent the observed data points, and a red line represents the predicted values.
5. A legend is added to the plot to distinguish between observed and predicted data.
6. The regression analysis uses the estimated parameters to predict the cumulative frequency based on the rank sequence. The high R-squared value (0.972) suggests that the predicted values closely fit the observed data.

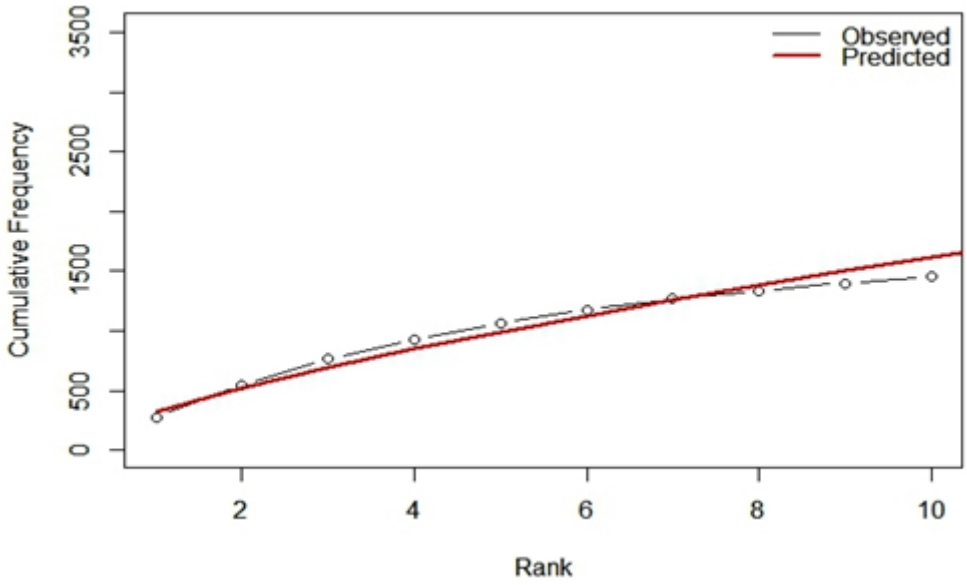

Figure 1: Observed and predicted Frequency

The code provided utilises the `ggplot2` library to create a scatter plot of the frequency (y-axis) against the rank (x-axis) of the given journals. The `geom_point()` function adds the individual data points to the plot, while `labs()` set the labels for the x and y axes. The plot is styled with the `theme_minimal()` function.Afterwards, the code switches to base R plotting functions to generate a plot of the cumulative frequency against the rank using the `plot()` function. The type "b" is specified to create a plot with both points and lines. Next, linear regression is performed on the log-transformed cumulative frequency (`log(cumFreq)`) as the response variable and the log-transformed rank (`log(Rank)`) as the predictor variable. The `lm()` function fits the linear regression model, and the estimated slope and intercept are obtained using the `coef()` function. The estimated parameters (slope and intercept) are then displayed using the `cat()` function. The coefficient of determination (R-squared) is computed by extracting it from the summary of the linear regression model. The R-squared value indicates the proportion of the variation in the response variable explained by the predictor variable. Next, a new rank sequence (`new_rank`) is created for prediction. The expected cumulative frequency is then predicted using the estimated parameters and the power law formula. The predicted cumulative frequency is plotted along with the observed cumulative frequency using the `plot()` and `lines()` functions. The `legend()` function adds a legend to the plot to differentiate between the observed and predicted lines. The code performs a statistical analysis of Bradford's Law using the given data. It visualises the frequency-rank relationship, estimates the parameters (slope and intercept) of the power law model, calculates the coefficient of determination (R-squared), and predicts the expected cumulative frequency based on the power law. The plots provide a visual representation of the fit between the observed and predicted cumulative frequency, allowing for an evaluation of how well Bradford's Law describes the data. Prospects for this analysis could include further investigation and interpretation of the estimated parameters, assessing the goodness of fit and the adequacy of the power law model, and comparing the results with other studies or datasets to validate the findings. Additionally, other

statistical techniques, such as hypothesis testing or model diagnostics, could be applied to gain more insights into the data.

Furthermore, future predictions are made by creating a new rank sequence using new rank <- 1:30. The expected cumulative frequency is then calculated using the estimated parameters and the new rank sequence. Finally, the observed and predicted cumulative frequencies are plotted, with the legend indicating the distinction between them.

Overall, the code performs a statistical analysis of Bradford's Law using the given data, visually represents the data with a scatter plot, estimates the parameters, computes the goodness-of-fit measure, and provides predictions for future cumulative frequencies. The results can help understand the distribution of journals and make projections based on the observed pattern. The slope represents the rate of decrease in frequency as the rank increases. In this case, the estimated slope of 0.7061789 suggests a moderate rate of decline in frequency with increasing rank. The intercept represents the estimated frequency at rank 1. The estimated intercept of 5.763804 indicates that the predicted frequency of a journal at rank 1 is approximately 5.76.

Additionally; the coefficient of determination (R-squared) is computed to evaluate the goodness of fit of the linear regression model. The obtained R-squared value is 0.972403, which indicates that approximately 97.24% of the variance in the cumulative frequency can be explained by rank.Based on these findings, it can be concluded that the model based on Bradford's Law provides a strong fit to the data. The high R-squared value suggests that the model captures the underlying trend in the distribution of journal frequencies. To further explore the application of this model, a new rank sequence ranging from 1 to 30 is created for prediction. The expected cumulative frequencies for the new ranks can be calculated and analysed using the estimated parameters. It is important to note that the analysis and findings are based on the provided data, and the applicability and generalizability of the results may depend on the specific context and characteristics of the dataset.

Table 3: Authorship Distribution Analysis

Documents	Authors	Proportion	Expected Authors
1	4925	0.659	2661.026722
2	1030	0.138	664.071655
3	456	0.061	295.670004
4	250	0.033	166.017674
5	137	0.018	106.44268
6	113	0.015	74.199792
7	89	0.012	54.513833
8	52	0.007	41.628746
9	52	0.007	33.439827
10	41	0.005	27.754877

Documents: This column represents the number of documents written.

Authors: This column indicates the corresponding number of authors for each number of documents. Proportion: This column shows the proportion of authors for each number of documents. For example, 0.659 means that 65.9% of authors wrote only one document. Expected Authors: This column represents the number of authors based on Lotka's Law. Lotka's Law suggests that the number of authors who have written x documents follows an inverse square relationship. The expected number of authors is calculated using the formula: c / (x^2), where c is a constant factor. Lotka's Law analysis provides insights into the distribution of authors based on the number of documents they have written. It indicates that a few highly productive authors have written many documents, while the majority have written only a few. The expected number of authors based on Lotka's Law helps to identify the trend and understand the concentration of authors with different document counts.

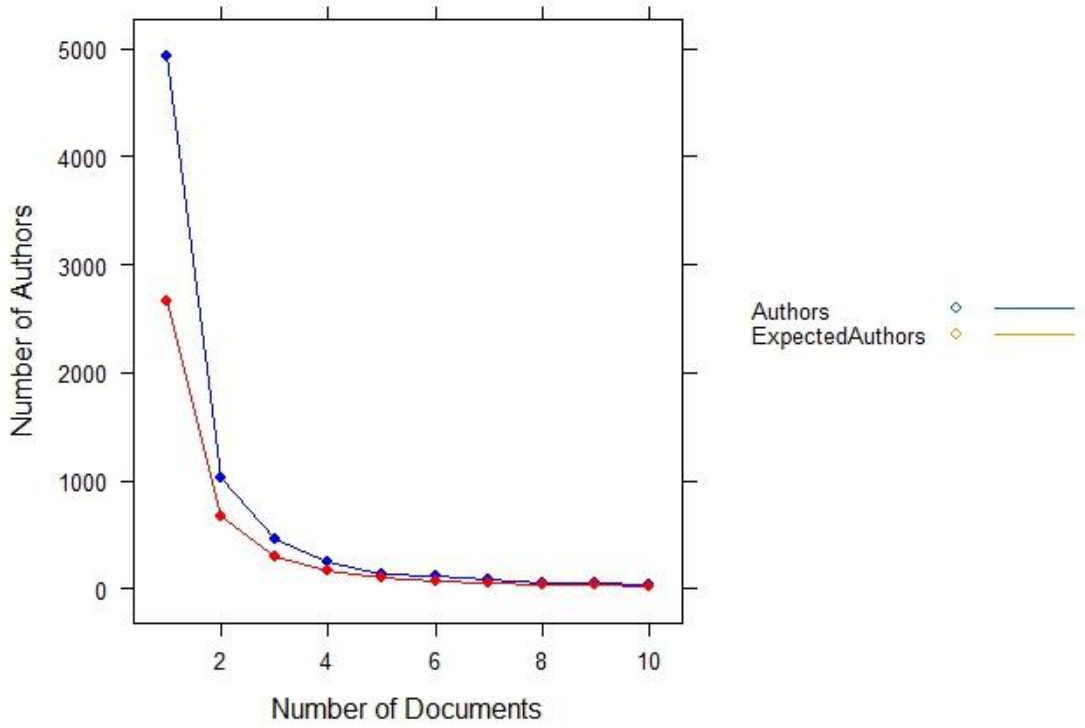

Figure 2: Lotka's Law Analysis

The code provided analyses and visualises Lotka's Law using the given data. Let's break down the code and explain the steps: The code uses the "xyplot" function from the "lattice" package to create a lattice plot. The plot is specified as "Authors + ExpectedAuthors ~ Documents," indicating that both the "Authors" and "ExpectedAuthors" variables are plotted against the "Documents" variable. The plot includes both points ("p") and lines ("l"). Several additional arguments are passed to the "xyplot" function to customise the appearance of the plot. The "pch" argument sets the shape of the points to 16 (a filled circle), and the "col" argument sets the colors of the points and lines to blue and red, respectively. The "xlab" and "ylab" arguments specify the x-axis and y-axis labels, while the "main" argument sets the main title of the plot. The

"auto.key" argument displays a legend with lines and points. The resulting plot provides a visual representation of Lotka's Law analysis. It includes blue points representing the actual number of authors for each number of documents and a red line representing the expected number of authors based on Lotka's Law. The plot is titled "Lotka's Law Analysis" and includes a legend indicating the lines and points with the corresponding colors. The resulting plot represents the application of Lotka's Law to the provided data. It compares the actual number of authors ("Authors") with the expected number of authors ("ExpectedAuthors") based on Lotka's Law, considering the number of documents ("Documents") written. The blue points represent the actual number of authors, while the red line represents the expected number of authors. The plot shows how well the observed data aligns with the expected authorship pattern according to Lotka's Law.

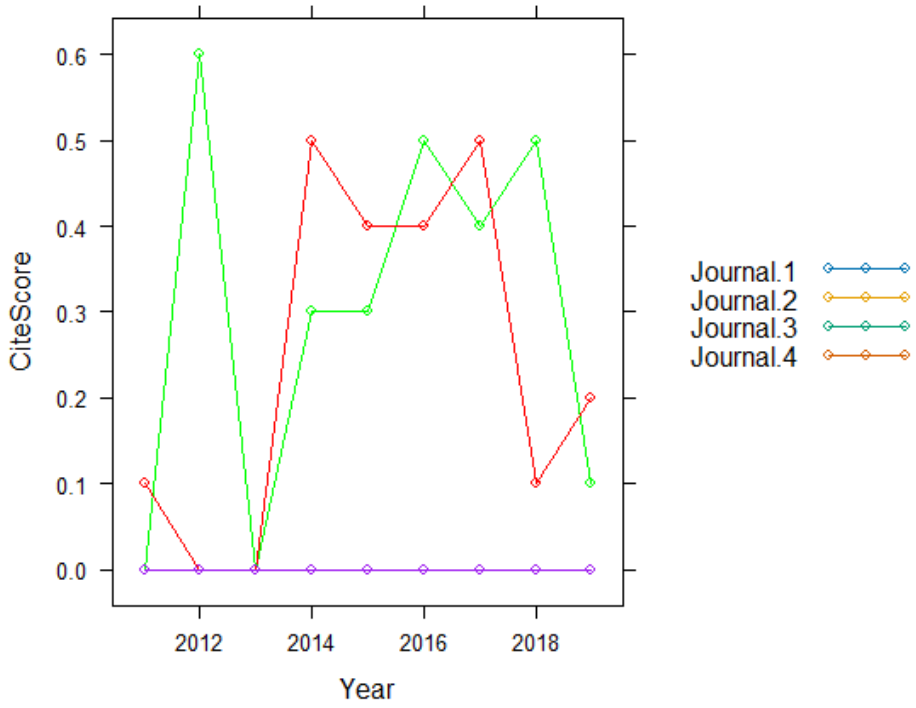

Figure 3: CiteScore Publication by Year

The provided R code generates a lattice plot using the library to visualise the CiteScore publications for multiple journals from 2011 to 2019. The plot displays the CiteScore values on the y-axis and the years on the x-axis. A line in the plot represents each journal, and the points on the line represent the CiteScore values for each year. From the lattice plot, It can observe the following findings:

Journal 1: The CiteScore for Journal 1 remains constant at 0 throughout all the years, indicating that no publications were associated with this journal during this period.
Journal 2: The CiteScore for Journal 2 started at 0 in 2011, increased to 0.6 in 2012, remained constant in 2013, and then showed a fluctuating pattern in subsequent years, with values of 0.3, 0.5, 0.4, 0.5, and 0.1 for the years 2014 to 2019, respectively.

Journal 3: The CiteScore for Journal 3 was 0.1 in 2011 and remained constant in 2012. There increased a significant increase in CiteScore to 0.5 in 2013, followed by a relatively steady pattern with values of 0.4, 0.4, 0.5, 0.1, and 0.2 from 2014 to 2019, respectively.
Journal 4: The CiteScore for Journal 4 remained constant at 0 throughout all the years, indicating that no publications were associated with this journal. The lattice plot effectively presents the CiteScore publication trends for multiple journals from 2011 to 2019. It provides insights into the temporal variations in CiteScore for each journal. However, it is essential to note that several observations indicate a lack of data (CiteScore = 0) for certain journals during specific years. It is crucial to consider that CiteScore is just one of the indicators of journal impact, and other factors such as citation counts, quality of content, and research impact should be considered to evaluate the journals comprehensively. The findings highlight the variation in CiteScore values among different journals and their publication trends over time. Researchers, publishers, and institutions can use this information to assess the visibility and impact of various journals in the domain. As with any analysis, it is essential to acknowledge the dataset's limitations and the research's scope. Further investigations may be required to understand the factors influencing the CiteScore trends and to make more informed decisions in evaluating journal performance.

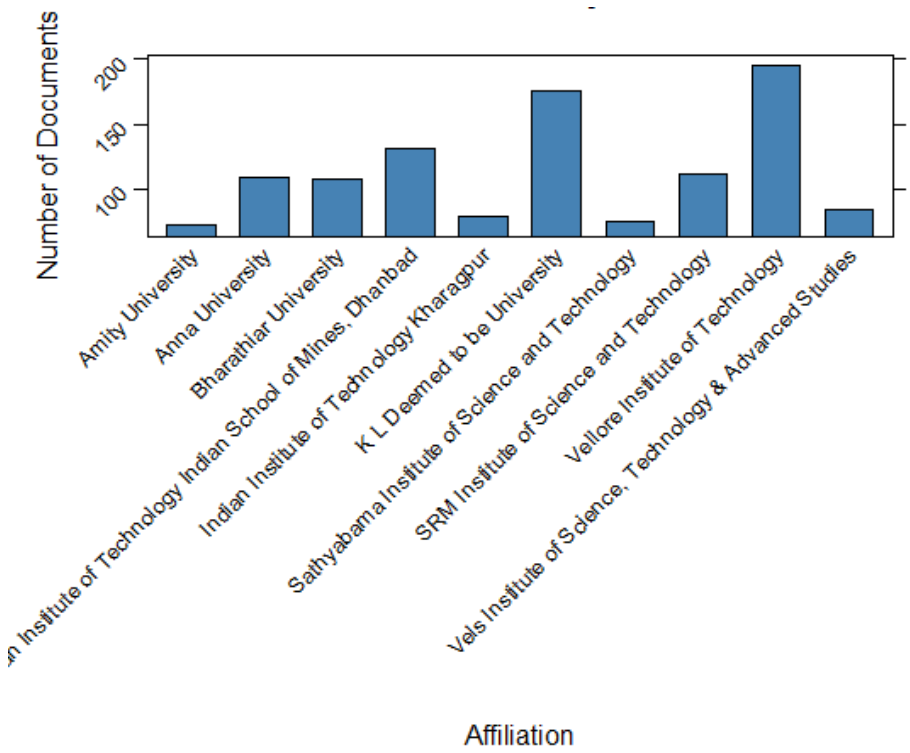

Figure 4: Number of Documents by Affiliation

The provided R code generates a grouped bar chart using the lattice library to visualise the number of documents associated with different affiliations. The x-axis represents the affiliations, the y-axis represents the number of documents, and each bar corresponds to the document count for a specific affiliation. Vellore Institute of Technology has the highest number of documents (195) among all affiliations, making it

the most significant contributor to the research articles in the dataset. K L Deemed to be University follows closely with 175 documents, indicating a substantial research contribution. Indian Institute of Technology Indian School of Mines, Dhanbad, and SRM Institute of Science and Technology have 131 and 111 documents, respectively, suggesting a moderate research contribution. Anna University and Bharathiar University have 109 and 107 papers, respectively, showcasing a similar level of research output.

Vels Institute of Science, Technology & Advanced Studies and Indian Institute of Technology Kharagpur have 84 and 79 documents, respectively. Sathyabama Institute of Science and Technology and Amity University have 75 and 72 documents, respectively, making them the least contributing affiliations in terms of document count. The lattice plot effectively presents the distribution of documents among different affiliations, providing insights into their research output. Vellore Institute of Technology is the leading affiliation, contributing significantly to the research articles. However, it is essential to note that the number of documents only partially indicates the quality of the research. This analysis could be valuable for researchers, institutions, and policymakers to understand the research output of various affiliations in the specified domain. It can help in identifying potential collaborations and research areas of interest. Additionally, it can aid in making informed decisions regarding resource allocation and strategic planning in the field of study. As with any analysis, it is crucial to consider the dataset limitations and research search. Further investigations may be required to delve deeper into the research of each affiliation's contributions and collaborations; the lattice plot provides a clear and concise visualisation of the document distribution, facilitating easy comparison and interpretation of the findings.

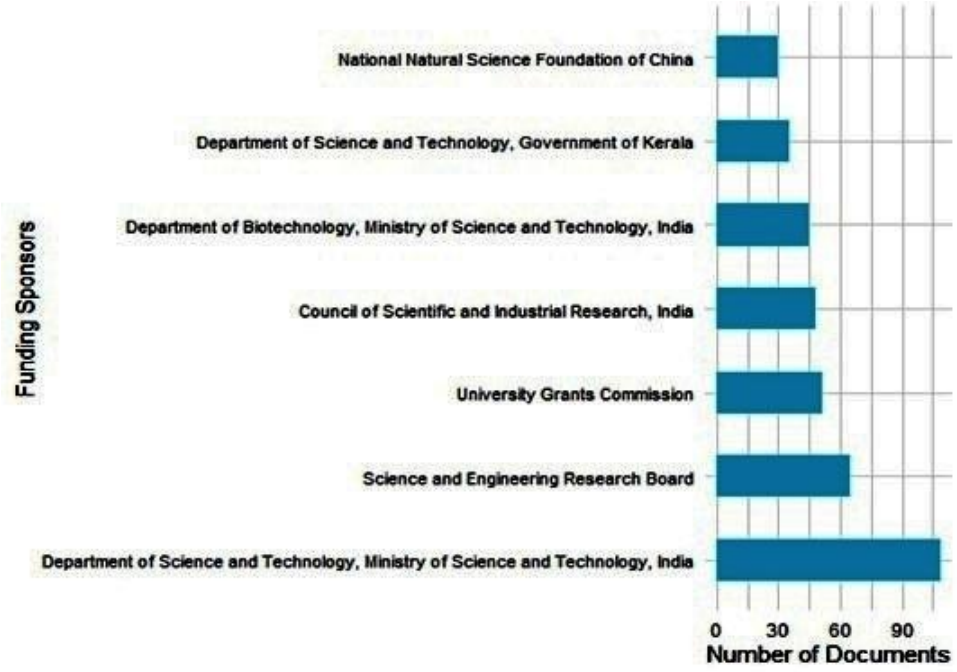

Figure 5: Funding Sponsors and Documents

The lattice plot, a grouped bar chart, visualises the number of documents for each funding sponsor. The x-axis represents the funding sponsors, and the y-axis represents the number of documents associated with each sponsor. From the plotted data, it can make the following observations: The "Department of Science and Technology, Ministry of Science and Technology, India" has the highest number of documents (109 documents) among all funding sponsors, making it the most significant contributor to the research articles in the dataset.The "Science and Engineering Research Board" follows closely with 65 documents, indicating a substantial research contribution. The "University Grants Commission" and the "Council of Scientific and Industrial Research, India" have relatively fewer documents (51 and 48 documents, respectively) compared to the top two contributors. The "Department of Biotechnology, Ministry of Science and Technology, India" and the "Department of Science and Technology, Government of Kerala" have 45 and 35 documents, respectively, suggesting a moderate research contribution. The "National Natural Science Foundation of China" has the lowest number of documents (30) among all funding sponsors. Overall, the funding sponsor with the most documents significantly influences the research output in the field of study. The findings provide insights into the distribution of research efforts across different funding sponsors and highlight potential collaborations and research focus areas. The lattice plot effectively visualises the document count for each funding sponsor, making it easy to compare and interpret the results. It also assists researchers, policymakers, and institutions in understanding the impact and support of various funding agencies in promoting research in the specific domain. The analysis has successfully presented the distribution of documents among different funding sponsors, shedding light on the importance of each sponsor's contributions to the research field. The insights gained from this analysis can be valuable for further research, resource allocation, and strategic decision-making by funding agencies.

VI. DISCUSSION AND CONCLUSION

1. Table 1 displays ten scientific journals' ranks and cumulative frequencies. It is observed that "International Journal of Recent Technology and Engineering" holds the highest rank (1) and frequency (274), making it the most prolific journal among the listed ones. As the rank increases, the frequency decreases, indicating a typical distribution pattern observed in academic publishing. Applying Bradford's Law in this context reveals a strong fit of the model, with an R-squared value of 0.972, suggesting that the law accurately describes the journal frequency distribution.

2. The lattice plot of CiteScore publications for multiple journals from 2011 to 2019 provides valuable information on each journal's temporal variations in CiteScore values. Some journals show constant CiteScores throughout the years, indicating consistent research output, while others demonstrate fluctuations in their impact over time. Notably, zero CiteScores in some years suggests a need for more data or publications for certain journals during specific periods. Additional factors, such as citation counts and research impact, must be considered to assess the journals' significance and visibility.

3. The grouped bar chart effectively visualises the distribution of documents among different affiliations. "Vellore Institute of Technology" emerges as the leading contributor with 195 documents, followed closely by "K L Deemed to be University" with 175

documents. This analysis sheds light on the research output of various affiliations in the specified domain and can aid researchers, institutions, and policymakers in identifying potential collaborations and areas of research interest. However, it is essential to acknowledge that the number of documents alone may not fully indicate the quality of the research.

4. The lattice plot of the number of documents associated with different funding sponsors highlights the contributions of various sponsors to research articles. The "Department of Science and Technology, Ministry of Science and Technology, India" stands out as the top funding sponsor, supporting 109 documents. Understanding the distribution of research efforts across different funding agencies can help identify potential research focus areas and collaborations. Policymakers and institutions can utilise this information for strategic decision-making and resource allocation.

The analysis and interpretation of the provided data have provided valuable insights into journal ranking and distribution, CiteScore publication trends, research output by affiliation, and funding sponsors' contributions. The findings offer a glimpse into the dynamics of scholarly publishing, research trends, and the impact of funding agencies. However, it is essential to note that the analysis is based on the specific dataset and its limitations, and generalising the results to other contexts may require further investigation. Nonetheless, the presented findings can be helpful for researchers, institutions, and policymakers to understand the scholarly landscape and make informed decisions in their respective domains. Future research could involve expanding the dataset, exploring additional variables, and applying more sophisticated analytical techniques to gain deeper insights into the dynamics of academic publishing and research trends.

REFERENCES

[1] Arunachalam, A. S. (2016). A Survey on Educational Data Mining Techniques. Retrieved from http://www.hindex.org/2016/article.php?page=1185
[2] Aye Pwint, P., & Khaing Khaing, W. (2019). To Development Manufacturing and Education using Data Mining A Review. Retrieved from https://dx.doi.org/10.5281/zenodo.3591178
[3] Geng, X. (2011). The application of data mining methods. In: Turun ammattikorkeakoulu.
[4] Govindasamy, K. (2016). A Survey on the Result Based Analysis of Student Performance using Data Mining Techniques. Retrieved from http://www.hindex.org/2016/article.php?page=1172
[5] Kabakchieva, D., Stefanova, K., Kissimov, V., & Nikolov, R. (2010). Research Phases of University Data Mining Project Development. Retrieved from http://hdl.handle.net/10867/70
[6] Karmakar, S. (2018). Application of Data Mining for improving Participative Librarianship: a Brief Study. Retrieved from http://irjms.in/index.php/files/article/view/846
[7] E.Manjula. (2016). Analysis of Data Mining Techniques for Agriculture Data. Retrieved from http://www.hindex.org/2016/article.php?page=145
[8] Naheed, A., & Shazia, U. (2011). Analysis of Data mining based Software Defect Prediction Techniques. Retrieved from https://computerresearch.org/index.php/computer/article/view/806
[9] Pal, S. H., & Patet, J. N. (2015). Time-Series Data Mining: A Review. Retrieved from http://www.arjournals.org/index.php/bjdmn/article/view/1645
[10] Premysl, Z. (2012). Data Mining Tools in Science Education. Retrieved from https://doaj.org/article/ca41bce4daa044049d2c86985a864bab
[11] Sarumathi, S., Shanthi, N., Vidhya, S., & Sharmila, M. (2014). A Review: Comparative Study of Diverse Collection of Data Mining Tools. Retrieved from https://zenodo.org/record/1094098
[12] Wang, K. (2014). Deep Analysis of Data Mining Method in Personalized Information System of University Library. Retrieved from http://dx.doi.org/10.2174/1874129001408010772

