Scientific Visualization, 2022, volume 14, number 5, pages 54 - 65, DOI: 10.26583/sv.14.5.04

Bibliometric Analysis and Visualization of Scientific Literature on Random Forest Regression

Sherin Babu[1,A,C], Binu Thomas[2,B,C]

[A] Department of Computer Science, Assumption College Autonomous,
Changanacherry, Kottayam, Kerala, India
[B] Department of Computer Applications, Marian College Autonomous,
Kuttikanam, Idukki, Kerala, India
[C] School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India

[1] ORCID: 0000-0002-7214-712X, sherinbabu@assumptioncollege.edu.in
[2] ORCID: 0000-0003-1594-2159, binu.thomas@mariancollege.org

Abstract

Random forest regression (RFR) is a versatile, easy-to-use and efficient tree based machine-learning algorithm that utilizes the power of multiple decision trees for making decisions. So random forest is a subject of a great deal of research, in many of the machine intelligence applications. The objective of this research is to investigate the scientific output of research based on RFR and to explore its hotspots and frontiers through bibliometric analysis for the years 2007 to 2019. The data are collected from the Web of Science database. The total number of publications, the citations, and types of publications, publication countries, productive authors, prominent journals, and keyword co-occurrence of RFR research are examined, using VOSviewer software. There are 516 papers, published in 299 journals, of which researchers from the USA published 162 articles. The most prolific author, with 6 publications and 240 citations, is Martin H. Teicher. The most cited article is the research article entitled "Genomic selection in wheat breeding using genotyping by sequencing". Among the journals, the most articles (41 publications) are published by the Remote Sensing journal.

Keywords: Bibliometric analysis, Machine learning, Random forest regression, Random forest, VOSviewer.

1. Introduction

Bibliometrics is the systematic application of mathematical and statistical methods to analyze research related books, articles and other publications (1–4). It is a quantitative study of different aspects of literature on a topic and is used to identify the pattern of publications, authorships and collaborations with the objective of ascertaining an insight into the process of growth of knowledge in the areas being studied (5,6). Bibliometric analysis serves as an important tool that can help to evaluate emerging developments in current scientific research (7,8). Bibliometrics is used to assess the characteristics of different types of academic outputs, to evaluate the contribution of researchers and institutes, to identify and forecast trends in the research fields concerned and also to find the amount of collaborative study (9–11). Bibliometric analysis methods, which are frequently used in the field of information science are aimed at measuring the properties of documents (12,13). Bibliometrics can be utilized to follow the evolution of a thought, and the nature of scholarly communications arising from it, inside and across disciplines (14). A bibliometric analysis using visualization software helps users to create visual representation of scientific research based on bibliographic data, which displays the associationship among scientific journals, authors, keywords, countries etc. (15–17).

Machine learning techniques have received considerable interest from researchers in every sector of life in the context of digitalization and automation. Machine learning approaches can recognize the underlying patterns and associations of broad and complex datasets and thus generate information from them (18). A random forest approach utilizes a collection of decision trees with enough power to handle complex nonlinear relationships within variable relationships (19–21). The random forest regressor (RFR) is comprised of an ensemble of decision trees, where each one independently predicts the value of a dependent variable based on several independent variables (22,23). Finally the predictions of all decision trees are merged by the means of majority voting or averaging method to get the best score and also to make the prediction more accurate and stable (24).

VOSviewer is a popular software tool for bibliometric data visualization developed by Van Eck and Waltman of Leiden University, Netherlands (25). VOSviewer is used for constructing and visualizing bibliometric maps. This software helps in the creation, visualization, and exploration of network maps based on bibliometric data (26). The Visualization of Similarities (VOS) mapping method was used by this program to measure and locate each subject in a two-dimensional map in such a way that the distance between two objects represents as accurately as possible the relativeness of the objects (27,28). The output results are displayed in multicolored clusters that visualize the existing correlations among the bibliometric data (29).

This study provides a comprehensive representation of the current status of research activities on the topic random forest regression, based on the data retrieved from WoS database for the years 2007 to 2019. In order to obtain a contemporary view of mainstream RFR studies, this study used a bibliometric visualization analysis tool called VOSviewer.

2. Materials and Methods

In this study, the scientific documents were retrieved from the core collection of the Web of Science (WoS) database of the Institute for Scientific Information (lSI) with indexes SCI-EXPANDED, SSCI, A&HCI. This search was performed on 12th October 2020. The search query was “random forest regression” OR “random forest regressor”. The period of publication of documents was from 2007 to 2019. In the search process, there were no language restrictions. All these data were saved as “Plain Text” files, which contained “Full Record and Cited References” content. The VOSviewer software version 1.6.14 was used to perform the bibliometric analysis in this study.

3. Results

Based on the search criteria, 516 publications about random forest regression were collected from WoS database. The document collection contained the following types of documents: article, early access, review, proceeding paper, data paper, and meeting abstract. From 2007 to 2019, the number of published research papers showed a growing pattern. In 2019, the number of papers published increased from 2 in 2007 to 183 (Figure 1). The number of citations in these articles has also significantly increased.

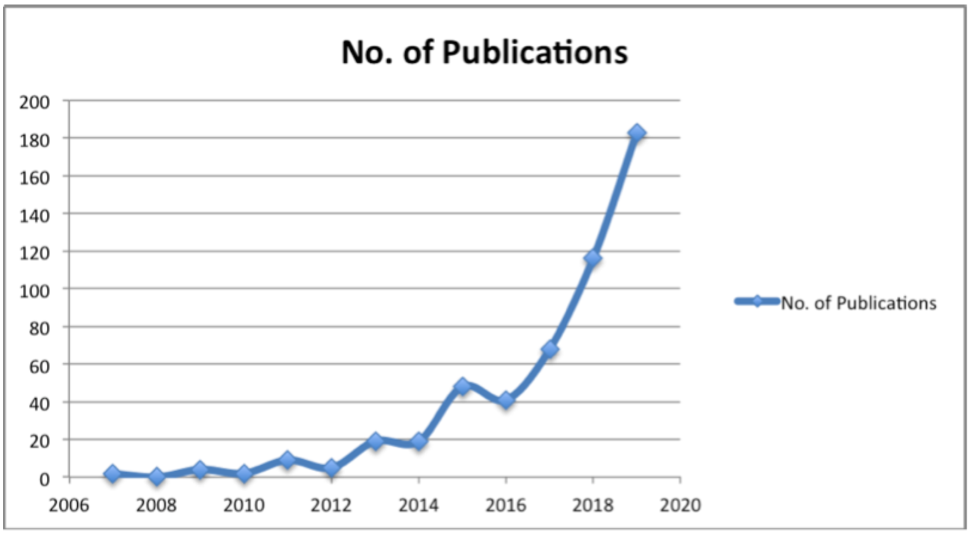

Figure 1 The number of publications on random forest regression from 2007 to 2019.

3.1. Publication output of countries

The country wise contribution of publications is estimated by the location of the affiliation of at least one author of the published documents. 71 countries contributed to the scientific production of RFR research. USA occupied the first position with a total of 162 documents, followed by People's Republic of China with 140 documents. The country wise visualization result is shown in Figure. 2.

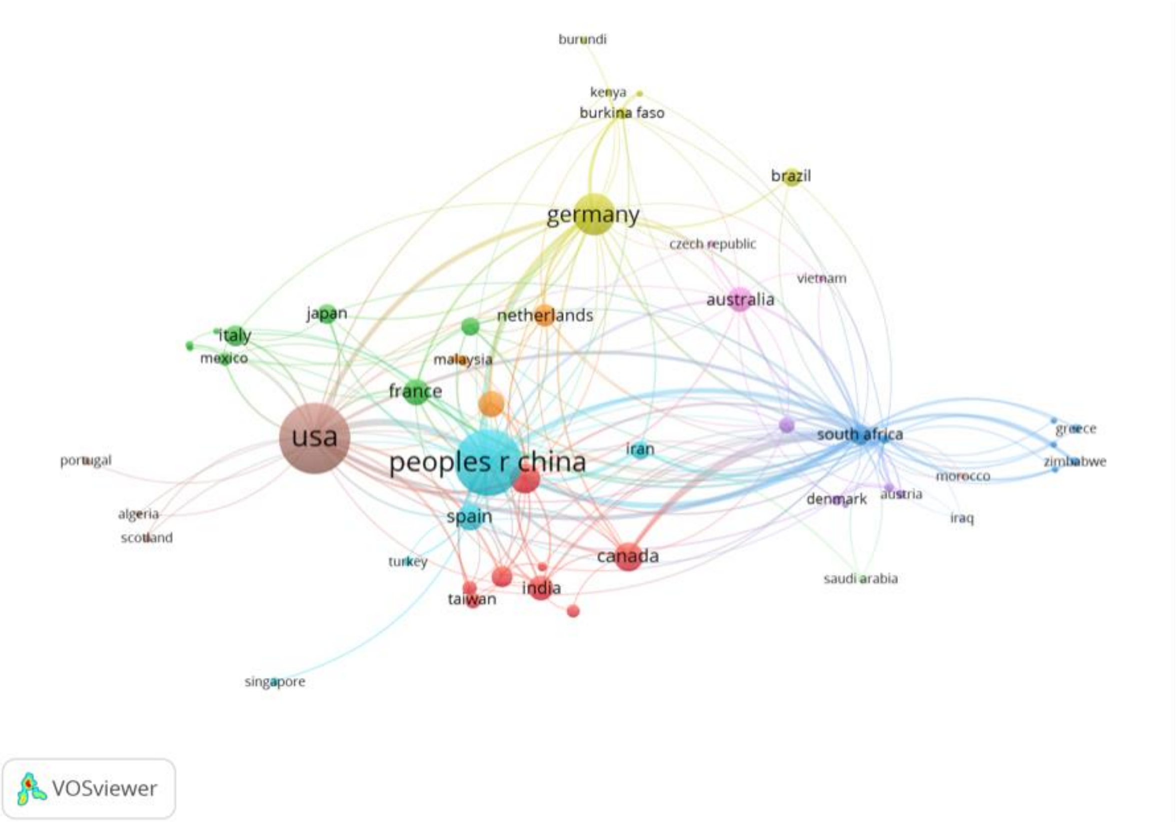

Figure 2 Visualization map of countries that published articles on RFR during 2007-2019

In terms of the amount of citations received, USA, People's Republic of China and Germany bagged the top three positions. The topmost 5 countries are listed in Table1.

Table 1 Topmost 5 countries ranked by the number of documents

No	Country	No. of Documents	No. of Citations
1	USA	162	4038
2	Peoples Republic of China	140	1857
3	Germany	57	1197
4	England	30	910
5	Canada	27	235

3.2. Most Productive Authors

The total number of authors publishing documents on RFR is 2475, of which only 3 authors published at least 5 documents. Table 2 shows the top five authors publishing literature on RFR ranked by total number of documents.

Table 2 Top 5 productive authors ranked by number of documents

No.	Author	No. of Documents	Citations
1	Martin H. Teicher	6	240
2	Onisimo Mutanga	5	528
3	Dinggang Shen	5	49
4	Deepak Bhatt	4	126
5	Vijay K Devabhaktuni	4	126

Martin H. Teicher was the most productive author with 6 documents and 240 citations. Onisimo Mutanga was the second productive author who contributed 5 documents and 528 citations. In the third position, the author Dinggang Shen came up with 5 documents and 49 citations. Deepak Bhatt secured the next place and Vijay K Devabhaktuni was placed in the fifth position with 4 documents each. The visualization network of authors is shown in Figure 3.

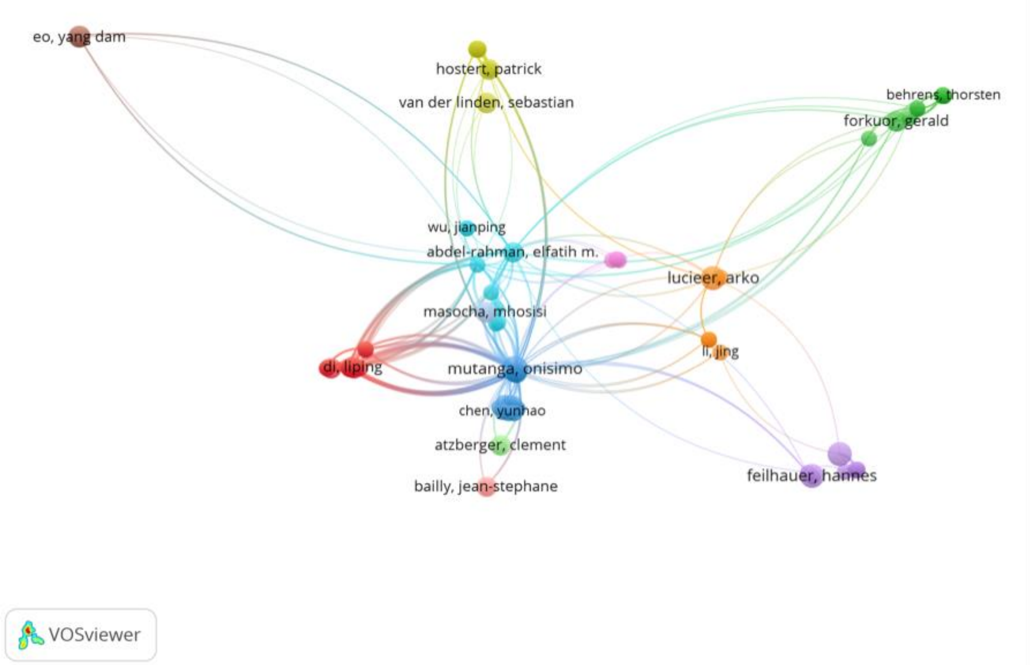

Figure 3 Network map of authors who published articles on the topic RFR during 2007-2019

3.3. Most Cited Articles

The highly cited literature is a very important indicator in the field of research, and it forms the foundation stone for exploring the research context and development direction. In citation networks, two items are linked if at least one cites the other. The top 5 highly cited articles are shown in Table 3.

Table 3 Top 5 articles on RFR ranked by total number of citations

No	Author	Title	Journal	No. of citations
1	Jesse Poland, Jeffrey Endelman, Julie Dawson, Jessica Rutkoski, Shuangye Wu,Yann Manes, Susanne Dreisigacker, José Crossa, Héctor SánchezVilleda, Mark Sorrells, Jean Luc Jannink (30)	"Genomic selection in wheat breeding using genotyping by sequencing."	The Plant Genome	583
2	Sally Archibald, David P. Roy, Brian W. Van Wilgen, Robert J. Scholes (31)	"What limits fire? An examination of drivers of burnt area in Southern Africa."	Global Change Biology	377
3	Mutanga, Onisimo, Elhadi Adam, and Moses Azong Cho (32)	"High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm."	International Journal of Applied Earth Observation and Geoinformation	347
4	T. Mitchell Aide, Matthew L. Clark, H. Ricardo Grau, David López-Carr, Marc A. Levy, Daniel Redo, Martha Bonilla-Moheno, George Riner, María J. Andrade-Núñez, María Muñiz (33)	"Deforestation and Reforestation of Latin America and the Caribbean (2001–2010)."	Biotropica	343
5	Eunkeu Oh, Rong Liu, Andre Nel, Kelly Boeneman Gemill, Muhammad Bilal, Yoram Cohen, Igor L. Medintz (34)	"Meta-analysis of cellular toxicity for cadmium-containing quantum dots."	Nature nanotechnology	190

The article by Poland, Jesse, et al. (30) entitled “Genomic selection in wheat breeding using genotyping by sequencing”, received the highest number of citations (n = 583). The next highly cited articles were (31), (32), (33) and (34). A network visualization map of citation relationship between documents is shown in Figure 4.

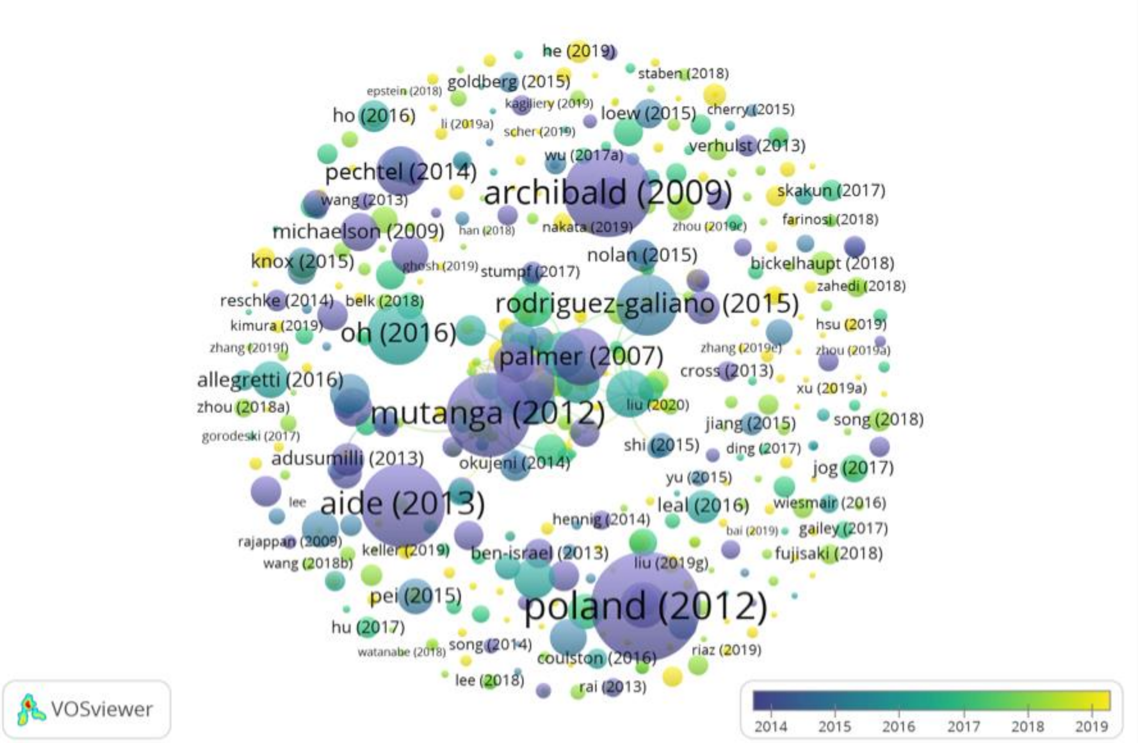

Figure 4 Citation network map of research articles on RFR published during 2007-2019

3.4. Contribution of Organizations

In the scientific publication on RFR, 905 organizations around the globe were involved. The top 5 organizations involved in RFR research are listed in Table 4.

Table 4 Top 5 organizations involved in the RFR research, ranked by number of documents

No.	Author	No. of Documents	Citations
1	Chinese Academy of Sciences	26	347
2	University of Chinese Academy of Sciences	9	40
3	Peking University	8	74
4	University of KwaZulu-Natal	7	657
5	McLean Hospital	7	255

The organization Chinese Academy of Sciences ranked first with a total contribution of 26 documents. Followed by the University of Chinese Academy of Sciences and Peking University that were ranked second and third position with 9 and 8 documents respectively. The visualization network of the involved organizations is shown in Figure 5.

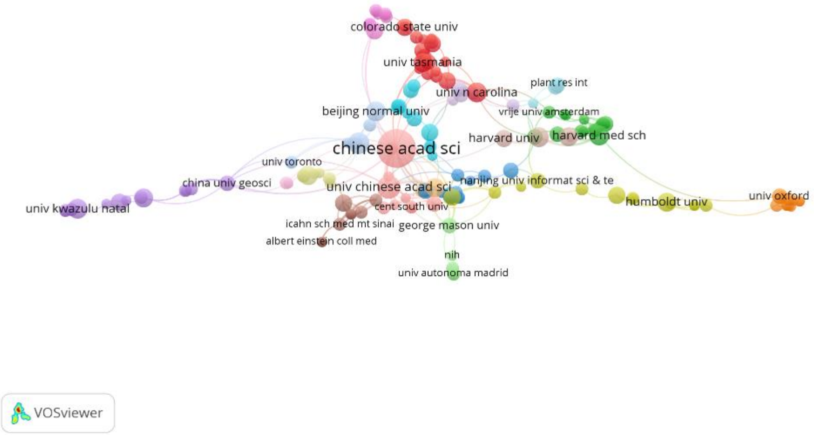

Figure 5 Network map of organizations involved in the RFR research during 2007-2019

3.5. Keyword Analysis

Using VOSviewer software, the keywords mentioned in all the 516 research documents were evaluated. A knowledge map of the co-occurrence of keywords represents the hot topics in the research field over a period of time. The minimum number of occurrences of a key word in titles and abstracts was set at 5 in this study. 3452 results were produced by the analysis of keywords related to random forest regression. Only 151 of them met the requirement of at least 5 co-occurrences. Six significant clusters were found. 29 items associated with regression and machine learning were found in Cluster 1 (blue points). 31 items made up Cluster 2 (red points), the majority of which were related to the topics of biomass, spectroscopy, leaves, and vegetation index. Climate change, land-cover change, temperature, and precipitation were the terms of greater relevance in Cluster 3 (green points), which included 26 items. Satellite and time-series were the two terms that stood out the most in Cluster 4 (yellow points), which contained 24 items. There were 19 items in Cluster 5 (violet points) that dealt with patterns, conservation, and biodiversity. Finally, cluster 6 included 12 components, with imaging, chlorophyl, and reflectance appearing most frequently. The knowledge map of the co-occurrence of keywords constructed is shown in Figure 6.

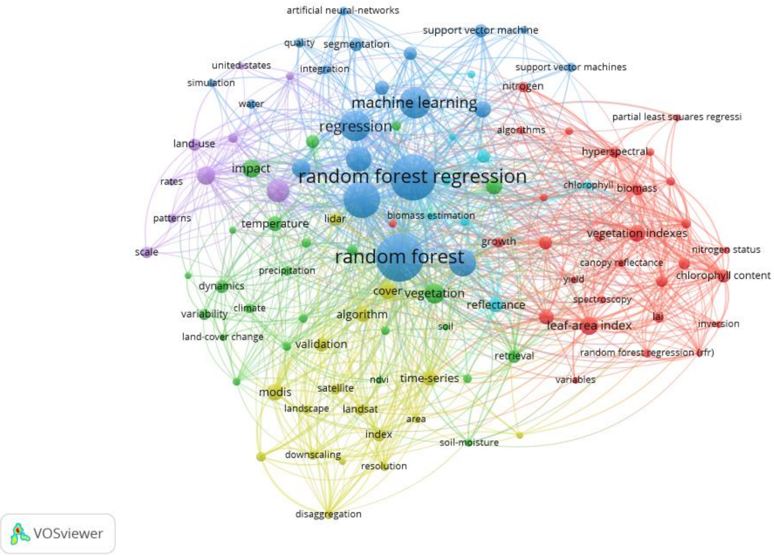

Figure 6 Co-occurrence map of keywords on random forest regression published during 2007-2019

The top 5 keywords by frequency count in publications from 2007 to 2019 are listed in Table 5.

Table 5 Top 5 keywords involved in the RFR research, ranked by occurrence count

No.	Keyword	Occurrence count
1	Random forest	104
2	Random forest regression	98
3	Classification	68
4	Machine learning	53
5	Regression	49

3.6. Most Collaborating Countries

The bibliographic coupling of the countries publishing works on the subject of RFR is depicted in Figure 7. The minimum publication threshold was set to five documents with link strength of 50. Of the 71 countries, 26 countries met the threshold. Different colors in the picture depict various clusters that were more frequently connected to one another. It indicated that studies from the same cluster of nations cite one another more frequently. There were 8 clusters in total. Out of these 8 clusters, there are only 3 clusters that contain at least 4 collaborating countries. Peoples Republic of China, USA, India, Japan, South Korea and Taiwan were located in the largest cluster (green color). In this cluster, the majority were Asian countries. The second largest cluster (red color) of collaborating countries included the European countries namely Germany, Switzerland, Finland, Spain, Poland and Italy. The third largest cluster (blue color) contained Australia, England, Sweden and Brazil.

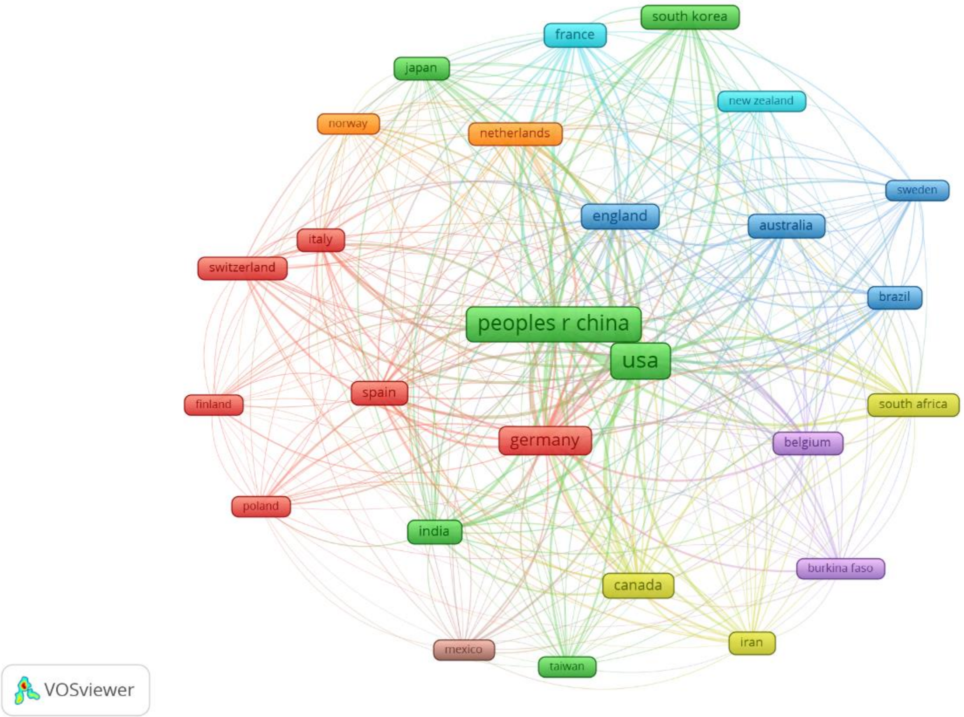

Figure 7 Bibliometric coupling of countries working on RFR during 2007-2019

4. Discussion

From 2007 to 2019, the number of global publications on random forest regression showed a positive growth. But in 2008, there was no publication on RFR. In the beginning years of study i.e., from 2007 to 2015, the rate of publication growth was sluggish. The overall number of publications on RFR has risen steadily since 2016. Graphing the data highlights that RFR research is a hotspot now.

About 59% of the total number of publications resulted from USA and People's Republic of China. A total of 28 countries published only one research document on the topic RFR. USA, followed by People's Republic of China and Germany, were the top most productive countries. People's Republic of China is the only developing country in the list of the top five productive countries. Unsurprisingly, the most frequent collaborating countries included USA and People's Republic of China. Along with them, the Asian countries namely India, Japan, South Korea and Taiwan proved strong collaborative research works.

Out of the 299 journals that published research papers on RFR, only 277 journals got at least one citation for its documents. There were only 12 journals that had published at least five research documents on this topic and these journals contributed about 25% of the published outputs. Only three journals received more than 400 citations for their documents - Remote Sensing, International Journal of Applied Earth Observation and Geoinformation and Remote Sensing of Environment are those journals.

The total of 516 publications on RFR has resulted from 2475 authors. Among those 2475 authors, about 93% authors received at least 1 citation for their document. The author Jean-Luc Jannink with only 2 publications in his account received the highest number of citations i.e., 686 citations. The author Jesse Poland also had received 686 citations for his publications.

Among the 516 publications on RFR, 480 articles were cited at least once. The remaining 36 articles were never cited. It is evident from the citation analysis that the highly cited articles were those published during the years 2012 to 2016. Reading articles published at that time about RFR could greatly influence future research. The oldest cited paper "Random Forest Models To Predict Aqueous Solubility" was written by (35) in 2007. The most recently published paper that have got citation was published by (36) in December 2019.

Even though 905 organizations were involved in the RFR research contributions; only 25 organizations were able to produce at least 5 research documents during 2007 to 2019. Based on the number of publications, the top 3 organizations involved in RFR research belong to People's Republic of China. The organizations from South Africa and USA secured the next positions. An important point to be noted here is that the University of KwaZulu-Natal, Durban, South Africa outperformed all other organizations with a total of 657 citation counts.

The study of keywords helps to get an understanding of what kind of topics and trends the researchers have mainly concentrated on. From all the selected publications, a total of 3452 keywords were reported in this analysis. The keywords “random forest” and “random forest regression” occurred 202 times in the data set. In addition to these keywords, "machine learning," "classification," "regression," "model" and "prediction" are the most commonly used keywords. Based the correlation of all these keywords, it was identified that about 481 publications were dealt with modeling of machine learning based systems for classification and/or regression tasks.

5. Conclusion

This is the first bibliometric analysis that measures worldwide scientific productivity in the field of random forest regression research. The study shows that in the coming years, publications on RFR will continue to expand rapidly. With a large number of publications, USA is recognized as the most productive country. In RFR research, the Remote Sensing journal has published the maximum number of papers. In this scientific field, Martin H. Teicher is the most prolific scholar. The most cited literature is "Genomic selection in wheat breeding using genotyping by sequencing”. It seems that the Chinese Academy of Sciences is the main contributor to research papers on RFR. This study serves as a rich source of information for the research community interested in conducting future studies in the field of random forest regression.

References

1. Godin B. On the origins of bibliometrics. Scientometrics. 2006 Jul 1;68(1):109–33.
2. Hood WW, Wilson CS. The Literature of Bibliometrics, Scientometrics, and Informetrics. Scientometrics. 2001 Oct 1;52(2):291.
3. Zou X, Yue WL, Vu HL. Visualization and analysis of mapping knowledge domain of road safety studies. Accident Analysis & Prevention. 2018 Sep 1;118:131–45.
4. Jiang P, Wu H, Da Y, Sang F, Wei J, Sun X, et al. RFRCDB-siRNA: Improved design of siRNAs by random forest regression model coupled with database searching. Computer Methods and Programs in Biomedicine. 2007 Sep 1;87(3):230–8.
5. Heradio R, Torre L de la, Galan D, Cabrerizo FJ, Herrera-Viedma E, Dormido S. Virtual and remote labs in education: A bibliometric analysis. Computers & Education. 2016 Jul;98(1):14–38.
6. Wallin JA. Bibliometric Methods: Pitfalls and Possibilities. Basic & Clinical Pharmacology & Toxicology. 2005;97(5):261–75.
7. Hong T, Feng X, Tong W, Xu W. Bibliometric analysis of research on the trends in autophagy. PeerJ. 2019 Jun 5;7:e7103.
8. Zhou H, Tan W, Qiu Z, Song Y, Gao S. A bibliometric analysis in gene research of myocardial infarction from 2001 to 2015. PeerJ. 2018 Feb 12;6:e4354.
9. Belter CW, Seidel DJ. A bibliometric analysis of climate engineering research. WIREs Climate Change. 2013;4(5):417–27.
10. Keiser J, Utzinger J. Trends in the core literature on tropical medicine: a bibliometric analysis from 1952-2002. Scientometrics. 2005 Mar 1;62(3):351–65.
11. Wang L, Wei YM, Brown MA. Global transition to low-carbon electricity: A bibliometric analysis. Applied Energy. 2017 Nov 1;205:57–68.
12. Broadus RN. Toward a definition of “bibliometrics.” Scientometrics. 1987 Nov 1;12(5):373–9.
13. Cronin B. Bibliometrics and beyond: some thoughts on web-based citation analysis: Journal of Information Science [Internet]. 2016 Jul 1 [cited 2020 Sep 20]; Available from: https://journals.sagepub.com/doi/10.1177/016555150102700101
14. Borgman CL, Furner J. Scholarly communication and bibliometrics. Annual Review of Information Science and Technology. 2002;36(1):2–72.
15. Ellegaard O, Wallin JA. The bibliometric analysis of scholarly production: How great is the impact? Scientometrics. 2015 Dec 1;105(3):1809–31.
16. Yu Y, Li Y, Zhang Z, Gu Z, Zhong H, Zha Q, et al. A bibliometric analysis using VOSviewer of publications on COVID-19. Ann Transl Med [Internet]. 2020 Jul [cited 2020 Sep 21];8(13). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7396244/
17. Su M, Peng H, Li S. A visualized bibliometric analysis of mapping research trends of machine learning in engineering (MLE). Expert Systems with Applications. 2021 Dec 30;186:115728.
18. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture. 2020 Oct 1;177:105709.
19. Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.
20. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognition Letters. 2010 Oct 15;31(14):2225–36.
21. Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, et al. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environmental Pollution. 2019 Feb 1;245:746–53.

22. Babar B, Luppino LT, Boström T, Anfinsen SN. Random forest regression for improved mapping of solar irradiance at high latitudes. Solar Energy. 2020 Mar 1;198:81–92.

23. Liu D, Fan Z, Fu Q, Li M, Faiz MA, Ali S, et al. Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm. Journal of Cleaner Production. 2020 Mar 20;250:119468.

24. Malhotra S, Karanicolas J. A Numerical Transform of Random Forest Regressors corrects Systematically-Biased Predictions. arXiv:200307445 [cs, q-bio, stat] [Internet]. 2020 Mar 16 [cited 2020 Oct 23]; Available from: http://arxiv.org/abs/2003.07445

25. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010 Aug 1;84(2):523–38.

26. van Eck NJ, Waltman L. Text mining and visualization using VOSviewer. arXiv:11092058 [cs] [Internet]. 2011 Sep 9 [cited 2020 Sep 20]; Available from: http://arxiv.org/abs/1109.2058

27. van Eck NJ, Waltman L. VOS: A New Method for Visualizing Similarities Between Objects. In: Decker R, Lenz HJ, editors. Advances in Data Analysis. Berlin, Heidelberg: Springer; 2007. p. 299–306. (Studies in Classification, Data Analysis, and Knowledge Organization).

28. van Nunen K, Li J, Reniers G, Ponnet K. Bibliometric analysis of safety culture research. Safety Science. 2018 Oct 1;108:248–58.

29. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for Information Science and Technology. 2011;62(7):1382–402.

30. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, et al. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome. 2012;5(3):103–13.

31. Archibald S, Roy DP, Wilgen BWV, Scholes RJ. What limits fire? An examination of drivers of burnt area in Southern Africa. Global Change Biology. 2009;15(3):613–30.

32. Mutanga O, Adam E, Cho MA. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. International Journal of Applied Earth Observation and Geoinformation. 2012 Aug 1;18:399–406.

33. Aide TM, Clark ML, Grau HR, López-Carr D, Levy MA, Redo D, et al. Deforestation and Reforestation of Latin America and the Caribbean (2001–2010). Biotropica. 2013;45(2):262–71.

34. Oh E, Liu R, Nel A, Gemill KB, Bilal M, Cohen Y, et al. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. Nature Nanotechnology. 2016 May;11(5):479–86.

35. Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO. Random Forest Models To Predict Aqueous Solubility. J Chem Inf Model. 2007 Jan 1;47(1):150–8.

36. Williams HM, DeLeon RL. Deep learning analysis of nest camera video recordings reveals temperature-sensitive incubation behavior in the purple martin (Progne subis). Behav Ecol Sociobiol. 2019 Dec 26;74(1):7.