A Review of Current Publications Trend on Missing Data Imputation Over Three Decades: Direction and Future Research

farah adibah adnan (farahadibah.adnan@gmail.com)
Universiti Malaysia Perlis https://orcid.org/0000-0003-2109-9460
Khairur Rijal Jamaludin
Universiti Teknologi Malaysia Razak Faculty of Technology and Informatics
Wan Zuki Azman Wan Muhamad
Universiti Malaysia Perlis
Suraya Miskon
Universiti Teknologi Malaysia - Main Campus Skudai: Universiti Teknologi Malaysia

Survey paper

Keywords: Bibliometrics, missing data, VOSViewer, machine learning

Posted Date: October 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-996596/v1

Abstract

Missing value or sometimes synonym as missing data, is an unavoidable issue when collecting data. It is uncontrollable and happen in almost any research fields. Hence, this study focused on identifying the current publications trend on missing data imputation techniques (1991- 2021) specifically in classification problems using bibliometric analysis. Most importantly, this research aims to uncover the potential missing data imputation methods. Two software were used; VOSViewer and Harzing Publish or Perish. Based on the Scopus database extracted in June 2021, the findings indicate an emerging trend in missing data imputation research to date, while there are two imputation methods that get the most attention; the random forest and the nearest neighbor methods.

1 Introduction

Missing data problem ubiquity encountered by researchers when analyzing real-world data. Usually, the real-world data contains many errors like incomplete data, inconsistent format (discrepancy in code), missing patterns and sometimes contain outliers. Most of the time, data scientists or researchers may spend lots of their time in data preprocessing. Data preprocessing has been indicted by researchers as a rudimentary stage in machine learning (ML) method (1). Many classification models (before applying any ML algorithms) are incapable of handling missing values directly. As a result, dealing with missing values in the data preprocessing step remains an important step in the classification process prior to estimation.

A well-known technique known as listwise deletion (or complete case analysis), had been extensively used to handle missing values during data preprocessing (2)(3). Ignoring or deleting instances with missing data is a common practice in some field (4)(5). However, it degrades the valuable information contain in the missing data and decrease statistical power as the sample size reduced (6)(7). Lin et al. (5) had experimented with the case deletion technique and he concluded that it can be used if the missing rate is small while their performance is parallel with imputation technique. But it depends on data type (categorical, numerical or mixed-type), missing mechanisms and the number of attributes or classes. The result is remarkably well in numerical dataset with missing rate up to 20% while in mixed-type dataset, the missing rate is up to 17%. The severe effect happens when the missing data substituted with zero or null value, where it produces biases in prediction and will interfere in decision making (8).

In contrast, missing data imputation technique replaces missing values with artificial estimates while maintaining data completeness (9). In the past three decades, multitude imputation approaches had been studied, range from statistical procedure to machine learning algorithms. The statistical procedure includes mean (10), mode and median (11), linear interpolation (12), regression (13) or by machine learning methods, such as K-nearest neighbors (14)(15), Fuzzy c-means (16), random forest (17), neural network (18), and decision trees (19). However, there is no solid conclusion in deciding which imputation model is the best because it depends on the type of data, missing proportion and also missing data mechanism (5). Missing data mechanism composed of missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (20). MCAR means every data in each attribute have

an equal chance to be missing, due to technical errors like machine breakdown or system failure. MAR relates with the missingness probability of an attributes depends on the observed information, but not depends on the missing data in that attribute. Whilst, MNAR is happened when the missingness probability of an attributes depends on the missing data in that attribute. Usually, researchers assume the missing data is either MCAR or MAR, while MNAR is complicated to identify. Before employed any imputation method, it is advised to identify the missing pattern either MCAR, MAR or MNAR.

Despite growing interest towards missing data imputation techniques, surprisingly, to the best of author's knowledge, there have been relatively very limited attempts in reporting the trend of prior works particularly those that used bibliometric approach. Only Adnan (21) reported using bibliometric analysis in studying on missing data covers 60 years (1960-2019) of research history, but the analysis is on general information (publication growth, document's language, subject area, and country of focus). Hence, this study expands the research by Adnan (21) which focused more on missing value imputation in classification problems. Moreover, this study aims to reveal the most impact authors, the most impact publications as well as the potential research gaps in missing data imputation method.

The next section discussed in details on the data source and methodology used in bibliometric analysis. Then, the analytical results are displayed in the form of graphs and tables, as well as visualization of the interconnection between keywords, authorship, and citations. Discussion and conclusion are presented in the last section.

2 Data And Methodology

2.1 Data Sources and Preprocessing

This study employed Scopus database as a basis to extract prior works on missing value-related matters. On 7th June 2021, a search was conducted with the keywords "missing data" or "missing values" or "missing value" or "incomplete data" and "imputation" and classification. To further specify relevant literature on missing data imputation, the search was based on article title, abstract and author's keywords and it returned 779 related papers. The result was later refined by comprising journal articles which covers from year 1991 to 2021. Finally, after screening, a total of 430 published journal articles were selected and included in the study.

2.2 Bibliometric Analysis

Bibliometric study or also known as scientometrics study, utilizes mathematical and statistical tools in the analysis in order to quantify and discover trends of the published materials. This study commences with a descriptive summary of the published documents by tabulating and graphing it by year, subject area, author, country, and document language. Next, an extensive bibliometric analysis comprises of the citation, authorship and keywords analysis can be achieved using VOSViewer and Harzing Publish or Perish software. The Harzing Publish or Perish software was used to show the citation metrics such as the total citations, document's average citations per year, document's average number of authors as well

as reveals the most impact researchers. Whilst for the VOSViewer, it was used to visualize the interconnection among authors, documents and keywords used by various authors.

3 Analysis And Findings

The analysis of extracted documents is divided into two phases; descriptive analysis and analysis of keywords, citations and authorship. The result also reveals the top 20 most cited articles in missing data related issue until June 2021.

3.1 Descriptive Analysis

3.1.1 Publication Growth

Based on the Scopus database, the first published journal article on missing data imputation in classification problems was in 1991 by Clogg, Rubin, Schenker, Schultz, and Weidman (22) where they studied on multiple imputation-based Bayesian logistic regression to generate new database (Tab. 1). They also listed on the top 20 most cited articles (Tab. 6) with 118 citations. According to Fig. 1, it shows a slow growth on the related publication from year 1991 until 2005 with the maximum number of publications is three. Following that, it shows a notable improvement with eight publications in 2006 and gradually increase since then. The possible reason for the publications in missing value was kicked started because of the popularization of data mining field (23). It is important to overcome missing values as it is the major drawback in data analysis. The highest number of publications was in 2020 with 63 articles or equivalent to 14.65% as in Tab. 1. It is anticipated that the number of publications will increase significantly in 2021, as there are already 36 papers in June 2021 when this article was extracted. Furthermore, an overall citation count of 9605 as in Tab. 5 confirms the relevance of the topic.

Table 1
Publication growth

Year	Frequency	% (N=430)	Cumulative Percent	Year	Frequency	% (N=430)	Cumulative Percent
1991	1	0.23	0.23	**2007**	12	2.79	7.91
1992	0	0.00	0.23	**2008**	4	0.93	8.84
1993	2	0.47	0.70	**2009**	14	3.26	12.09
1994	0	0.00	0.70	**2010**	14	3.26	15.35
1995	0	0.00	0.70	**2011**	6	1.40	16.74
1996	0	0.00	0.70	**2012**	29	6.74	23.49
1997	1	0.23	0.93	**2013**	25	5.81	29.30
1998	1	0.23	1.16	**2014**	22	5.12	34.42
1999	0	0.00	1.16	**2015**	31	7.21	41.63
2000	1	0.23	1.40	**2016**	26	6.05	47.67
2001	3	0.70	2.09	**2017**	38	8.84	56.51
2002	1	0.23	2.33	**2018**	44	10.23	66.74
2003	0	0.00	2.33	**2019**	44	10.23	76.98
2004	3	0.70	3.02	**2020**	**63**	**14.65**	91.63
2005	1	0.23	3.26	**2021**	36	8.37	100.00
2006	8	1.86	5.12	**Total**	**430**	**100%**	

3.1.2 Subject Area

This research categorizes the published documents based on their subject matter. It is obvious that computer science area dominated in this research with 212 publications (25.2%) followed by researchers from mathematics field, 115 publications (13.7%) and medicine field, 111 publications (13.2%) as indicated by Fig. 2 and Tab. 2 below.

Table 2
Publications by subject area

Subject Area	Frequency*	% (N=842)
Computer Science	212	25.2%
Mathematics	115	13.7%
Medicine	111	13.2%
Engineering	99	11.8%
Biochemistry, Genetics and Molecular Biology	62	7.4%
Decision Sciences	43	5.1%
Social Sciences	26	3.1%
Neuroscience	18	2.1%
Materials Science	17	2%
Agricultural and Biological Sciences	15	1.8%
Others	124	14.7%
Total	**842**	**100%**

*Some documents are classified in more than one subject area

3.1.3 Country Productivity

Researchers from 56 countries have expressed their interest in the study of missing data across various field. More and more countries begun to devote themselves in the research related with missing data such as to determine techniques in replacing the missing value (23)(24), understanding pattern of missing data (25)(26), and also evaluation of missing value imputation on classification accuracy (27). Fig. 3 describes the distribution of the top 10 country in the publication. Most of the articles were affiliated with researchers in the United States (110 documents) followed by India (53 documents) and China (48 documents).

Table 3
Publications by country

Country	Frequency	% (N=567)	Country	Frequency	% (N=567)
United States	111	19.6%	**Turkey**	4	0.7%
India	53	9.3%	**Indonesia**	3	0.5%
China	50	8.8%	**Iraq**	3	0.5%
United Kingdom	32	6.6%	**Russian Federation**	3	0.5%
Spain	27	4.8%	**Austria**	2	0.4%
Australia	24	4.2%	**Colombia**	2	0.4%
Canada	24	4.2%	**Egypt**	2	0.4%
Germany	16	2.8%	**Greece**	2	0.4%
South Korea	15	2.6%	**Norway**	2	0.4%
Netherlands	14	2.5%	**Sweden**	2	0.4%
Taiwan	14	2.5%	**Thailand**	2	0.4%
Italy	13	2.3%	**United Arab Emirates**	2	0.4%
South Africa	13	2.3%	**Algeria**	1	0.2%
France	13	2.3%	**Bangladesh**	1	0.2%
Iran	12	2.1%	**Bulgaria**	1	0.2%
Belgium	10	1.8%	**Ethiopia**	1	0.2%
Japan	10	1.8%	**Fiji**	1	0.2%
Brazil	9	1.6%	**Israel**	1	0.2%
Malaysia	8	1.4%	**Jordan**	1	0.2%
Finland	7	1.2%	**Kenya**	1	0.2%
Switzerland	6	1.1%	**Mexico**	1	0.2%
Hong Kong	5	0.9%	**Morocco**	1	0.2%
New Zealand	5	0.9%	**North Korea**	1	0.2%
Pakistan	5	0.9%	**North Macedonia**	1	0.2%
Poland	5	0.9%	**Romania**	1	0.2%
Saudi Arabia	5	0.9%	**Slovakia**	1	0.2%
Denmark	4	0.7%	**Tunisia**	1	0.2%

Country	Frequency	% (N=567)	Country	Frequency	% (N=567)
Portugal	4	0.7%	**Undefined**	5	0.9%
Viet Nam	4	0.7%	**Total**	567	100%

3.1.4 Document Language

Majority the journal articles retrieved from Scopus database are in English (420; 97.22%), while the remaining articles are either in Chinese, Russian, Spanish, German and Japanese language as in Tab. 4.

Table 4
Publications by language

Language	Frequency*	% (N=432)
English	420	97.22%
Chinese	4	0.93%
Russian	3	0.69%
Spanish	3	0.69%
German	1	0.23%
Japanese	1	0.23%
Total	**432**	**100%**

*Some documents were published in dual languages

3.1.5 Documents by Author

According to the Scopus database, the top ten authors in missing data related publications are shown in Fig. 4. Twala, B. had recorded the highest contribution with 8 articles, followed by Garcia-Laencina,P. J., and Sancho-Gomez, J. L., with 6 articles each. Among the contribution by Twala, B., is on the use of the neural networks in dealing with class imbalance and missing data problems (28), classification and regression trees in missing data with high attribute correlations (29), k-nearest neighbor (KNN) and support vector machines in missing data with higher complexity with limited number of instances (30). The second top ten authors, Garcia-Laencina, P. J., which is co-author with Sancho-Gomez, J. L., had proposed a novel KNN imputation with feature-weighted distance metric based on mutual information (MI) on solving classification task (15). In different research, he presented a new public software for missing data imputation, called Web IMPutation, that is linked to a computer cluster to perform high computational tasks. The software is free, where registered users can create, run, analyze and save simulations related to missing data imputation (31).

3.2 Analysis of Keywords and Citations Analysis

3.2.1 Keywords Analysis

A keyword analysis has been performed using the VOSViewer in order to evaluate the specifics debate on the missing data related publications. The analysis reveals that 3835 keywords were used within the papers. The number of keyword occurrence is set to be at least 8 times and resulting 135 items/selected keyword. From Fig. 5, it revealed the existence of three clusters, and it can be group according to the area of research; computer science with 58 items (Red Cluster), medicine with 42 items (Green Cluster), and mathematics/statistics with 35 items (Blue Cluster). This result is parallel as mentioned previously in the section 3.1.2 where computer science, mathematics/statistics and medicine area dominated in this study. The size of the nodes varies according to the importance of the element. For example (Fig. 5), on the keyword classification, missing data, imputation, classification (of information), and support vector machine appear to have big circle, hence it means most discussion with highest occurrence on this topic. In contrast, the smaller circle reflects less occurrence with low frequency on the keyword such as genetic algorithm. Each keyword is linked to another keyword. For instance, in Fig. 6, imputation keyword links with data mining, nearest neighbor search, neural networks, classification accuracy, learning systems, feature extraction, classification (of information), imputation methods, missing values, incomplete data, missing value imputations, optimization, cluster analysis, algorithms, data analysis, humans, article, priority journal, female, adult, middle age and aged keyword. The link shows the topic they are discussed together. The different in distance between two keywords indicates their relatedness of the keywords, the shorter the distance, the stronger their relatedness.

Overlay visualization as in Fig. 7 describes the keyword development over time. It is distinguished by colour, from dark blue to green to yellow. A colour bar in the bottom right corner explicates the colour; the dark blue colour indicates the keyword occurred mostly 2012 and below, while the colour transforms to yellow means it is the latest trend, 2018 onwards. For example, the dna microarray keyword had been discussed long time ago since 2012 and below, while the data imputation, nearest neighbor search and random forest keyword shows the current trend. Observing their size, even though these keywords having a small circle, there possibly still new in the research domain and have high chances to explore.

Fig. 7 Overlay visualization of keywords

Figure 8 depicts the density visualization of the keywords. The keywords in yellow colour area appear more frequently; meanwhile the keywords in green colour area appear less frequently. Density views are especially useful for understanding the overall structure of a map and drawing attention to the most important areas in the map. From Fig. 8, the hot topic discussed in the research are “missing data”, “classification”, “article” and “human” turn out to be important.

3.2.2 Citation Analysis by Documents/Articles

Table 5 summarizes the citation metrics obtained from Harzing Publish or Perish software where 430 articles were retrieved from Scopus database as on 7th June 2021. As indicated, there are 9605 citations

reported over 30 years (1991-2021) with average of 320 citation per year and with average of 4 authors per paper. In Tab. 6, discloses the top 20 most cited articles as reported by Scopus. The paper written by Stekhoven and Buhlmann (17) ranks first with 1023 citations or an average of 113.67 citation per year. They introduce new algorithm for missing data imputation, named missForest. The new algorithm can be found in R package missForest. The result showed the missForest algorithm outperformed KNN impute, multivariate imputation by chained equations (MICE) and Missingness Pattern Alternating Lasso (MissPALasso) since this method is robust to noisy data, do not rely upon distributional assumptions on the data, can work with multicollinearity and multidimensional data, and requires no tuning (32). Surprisingly, it can be used for categorical and numerical data simultaneously. Based on Tab. 6, among the methods that received much attention are KNN [15],[16] multiple imputation [6],[34],[35],[36],[41],42] and review methods [23],[27],[34],[45]. It should be noted that multiple imputation is a well-known method in medical research. Summary for the rest of the top 20 articles presented in Tab. 7.

The VOSViewer software was employed in order to comprehend thoroughly on citation analysis by documents. The citation analysis by documents was executed in order to measure the citation impact on certain documents and to investigate the expansion of an article. As an illustration, in Fig. 9, on the article “Missforest-Non-parametric missing value imputation for mixed-type data” by Stekhoven, D. J (2012) was referred by Huang, J (2017), Sahri, Z (2014), Lobato, F (2015), Xia, J (2017), Cevallos Valdiviezo, H (2015), Bertsimas, D (2018) and Wang, Z. X (2017). The rest of the other authors are not shown because of the setting is set to be at least 10 citations of a document. Only documents with 10 citations and above are appear in the map.

Table 5 Citation metrics

Metrics	Data
Publication years	1991-2021
Citation years	30(1991-2021)
Papers	430
Citations	9605
Cites/year	320.17
Cites/paper	22.34
Authors/paper	3.89
h-index	43
g-index	86
hI,norm	25
hI,annual	0.83
hA-index	15

By looking at the yellowish node, labelled Che, Z. (2018) in Fig. 9, it means this article had received much attention from scholars because the size of the node is moderate, while the yellow node reflects that this research is trending. His research entitles “Recurrent Neural Networks for Multivariate Time Series with Missing Values” having 394 citations (26). They emphasized on the significance of missing patterns in estimating missing values, and proposed a novel method namely Gated Recurrent Unit (GRU-D). In comparison, a small blue node, labelled Clogg, C.C (1991) indicates that this is an older study that has received less attention compared to Che, Z. (2018).

Table 6
The top 20 most cited articles

Authors	Title	Journal	Year	Cites	Cites per Year
D.J. Stekhoven, P. Bühlmann (17)	Missforest-Non-parametric missing value imputation for mixed-type data	Bioinformatics	2012	1069	118.78
Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu (26)	Recurrent Neural Networks for Multivariate Time Series with Missing Values	Scientific Reports	2018	394	131.33
F.M. Shrive, H. Stuart, H. Quan, W.A. Ghali (33)	Dealing with missing data in a multi-question depression scale: A comparison of imputation methods	BMC Medical Research Methodology	2006	360	24.00
A.I. Phipps, P.J. Limburg, J.A. Baron, A.N. Burnett-Hartman, D.J. Weisenberger, P.W. Laird, F.A. Sinicrope, C. Rosty, D.D. Buchanan, J.D. Potter, P.A. Newcomb (34)	Association between molecular subtypes of colorectal cancer and patient survival	Gastroenterology	2015	247	41.17
G.H. Kingsley, A. Kowalczyk, H. Taylor, F. Ibrahim, J.C. Packham, N.J. McHugh, D.M. Mulherin, G.D. Kitas, K. Chakravarty, B.D.M. Tom, A.G. O'keeffe, P.J. Maddison, D.L. Scott (35)	A randomized placebo-controlled trial of methotrexate in psoriatic arthritis	Rheumatology (United Kingdom)	2012	230	25.56
X. Zhu, S. Zhang, Z. Jin, Z. Zhang, Z. Xu (36)	Missing value estimation for mixed-attribute data sets	IEEE Transactions on Knowledge and Data Engineering	2011	209	20.90
A. Farhangfar, L. Kurgan, J. Dy (27)	Impact of imputation of missing values on classification error for discrete data	Pattern Recognition	2008	202	15.54
M. Saar-Tsechansky, F. Provost (37)	Handling missing values when applying classification models	Journal of Machine Learning Research	2007	192	13.71

Authors	Title	Journal	Year	Cites	Cites per Year
A. Elbaz, J. Clavel, P.J. Rathouz, F. Moisan, J.-P. Galanaud, B. Delemotte, A. Alperovitch, C. Tzourio (6)	Professional exposure to pesticides and Parkinson disease	Annals of Neurology	2009	185	15.42
S. Zhang, X. Li, M. Zong, X. Zhu, D. Cheng (38)	Learning k for kNN Classification	ACM Transactions on Intelligent Systems and Technology	2017	175	43.75
I.B. Aydilek, A. Arslan (16)	A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm	Information Sciences	2013	172	21.50
P.K. Shivaswamy, C. Bhattacharyya, A.J. Smola (39)	Second order cone programming approaches for handling missing and uncertain data	Journal of Machine Learning Research	2006	155	10.33
D. Buse, A. Manack, D. Serrano, M. Reed, S. Varon, C. Turkel, R. Lipton (40)	Headache impact of chronic and episodic migraine: Results from the American Migraine Prevalence and Prevention Study	Headache	2012	136	15.11
S. Leu, S. Von Felten, S. Frank, E. Vassella, I. Vajtai, E. Taylor, M. Schulz, G. Hutter, J. Hench, P. Schucht, J.-L. Boulay, L. Mariani (41)	IDH/MGMT-driven molecular classification of low-grade glioma is a strong predictor for long-term survival	Neuro-Oncology	2013	127	15.88
P.J. Garcia-Laencina, J.-L. Sancho-GÃ³mez, A.R. Figueiras-Vidal, M. Verleysen (15)	K nearest neighbours with mutual information for simultaneous classification and missing data imputation	Neurocomputing	2009	129	10.75
J. Luengo, S. Garcia, F. Herrera (23)	On the choice of the best imputation methods for missing values considering three groups of classification methods	Knowledge and Information Systems	2012	126	14.00

Authors	Title	Journal	Year	Cites	Cites per Year
Z.-G. Liu, Q. Pan, J. Dezert, A. Martin (42)	Adaptive imputation of missing values for incomplete pattern classification	Pattern Recognition	2016	122	24.40
C.C. Clogg, D.B. Rubin, N. Schenker, B. Schultz, L. Weidman (22)	Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression	Journal of the American Statistical Association	1991	118	3.93
G. Paleologo, A. Elisseeff, G. Antonini (43)	Subagging for credit scoring models	European Journal of Operational Research	2010	112	10.18
D. Jarquin, K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, G. Graef, A. Lorenz (44)	Genotyping by sequencing for genomic prediction in a soybean breeding population	BMC Genomics	2014	110	15.71

Table 7
Summary method of the top 20 most cited articles

Proposed Method / Best Method	Software	Data Type	Data Size	Missing Rate	References
Sequential Random Forest	missForest	Mixed-type	40-595	10%, 20%,30%	(17)
Novel Gated Recurrent Unit (GRU-D), deep learning	Phyton	Numerical, Categorical	2000-10000	1%-99%	(26)
Multiple imputation (review method)	SAS	Questionnaire (nominal, ordinal)	1580	10%, 20%,30%	(33)
*Multiple imputation	Not mentioned	Numerical, Categorical	706	Not mentioned	(34)
*Multiple Imputation by Chained Equations (MICE)	R statistical package	Numerical, Categorical	221	23%	(35)
Mixture kernel based iterative estimator	Not mentioned	Mixed-type	200-6000	10%, 20%,30%, 50%, 80%	(36)
Study an impact of missing value imputation on classification accuracy	Not mentioned	Discrete	47-28000	5%, 10%, 20%, 30%, 40%, 50%	(27)
Reduced-feature modeling	Not mentioned	Categorical, Continuous	270-20640	1.73 – 3.56 average missing features	(37)
*Multiple imputation using logistic regression	SAS PROC MI	Categorical	224	<5%	(6)
Correlation Matrix kNN (CM-kNN)	Not mentioned	All types (high-dimensional, low-dimensional, binary, multi-class, imbalance)	46-700	Not mentioned	(38)
Hybrid fuzzy c-mean with support vector regression and genetic algorithm	Matlab	Continuous	178-1489	1%, 5%, 10%, 15%, 20% 25%	(16)
Second order cone programming	Mosek solver	Binary	150	50%, 75%, 90%	(39)

**This paper applied multiple imputation in medical study*

Proposed Method / Best Method	Software	Data Type	Data Size	Missing Rate	References
*Multiple imputation	SAS PROC MI	Categorical, Continuous, Discrete	27 253	Not mentioned	(40)
*Multiple imputation	Mice package in R software	Categorical, Continuous, Discrete	210	159/210	(41)
KNN based mutual information (MI-KNNimpute)	-	Qualitative, Quantitative	50-871	5–40%	(15)
Review method (best imputation for different classifiers)	-	Nominal, Numeric, Mixed attribute	-	-	(23)
Credal classification with adaptive imputation (CCAI)	Matlab	-	155-1429	-	(42)
Bayesian Logistic Regression	SPSS	Nominal	200	-	(22)
Subagging decision trees	Spider library	Numerical, Categorical	11903	-	(43)
Comparison Naïve, random forest and haplotype-based imputation	-	Discrete	301	1% - <80%	(44)
**This paper applied multiple imputation in medical study*					

3.2.3 Citation Analysis by Authors

This section was designed to study an impact of authors based on citation. With at least three number of documents and 100 citations of an author, the most impact authors on the study of missing value are Zhang, S., Zhu, X., Herrera, F., Luengo, J., Twala, B., Li, X., Zhang, Z., Pan, Q., and Garcia-aencina, P. J. (Tab. 8). The relationship among the authors can be seen as in Fig. 10, where Garcia-laencina, P. J., Twala, B., Herrera, F., Luengo, J., and Pan, Q., were in the same cluster, Cluster 1, while Li, X., Zhang, Z., Zhang, S., and Zhu, X., were in Cluster 2.

Overlay visualization of citation analysis by authors in Fig. 11 mapped the authors involvement over time in the missing value related topic. The earliest study was by Herrera, F. and Luengo, J. and were referenced by Twala, B. and Pan, Q. Whilst, Twala, B. was cited by Zhang, X., and Pan, Q. Indirectly, we know that Pan, Q. follows Herrera, F., Luengo, J., Twala, B. and Li, X.

Table 8
Citation impact of authors

Author	Documents	Citations	Total link strength
Zhang, S.	6	554	15
Zhu, X.	3	409	11
Herrera, F.	3	217	7
Luengo, J.	3	217	7
Twala, B.	8	134	7
Li, X.	3	210	6
Zhang, Z.	3	275	6
Pan, Q.	3	153	4
Garcia-laencina, P. J.	6	322	1

4 Discussion

An extensive analysis had been done to investigate in what aspects other researchers cited the most impact articles and the most impact authors. First, according to the link in Fig. 9, seven authors have been cited the most impact article, “Missforest-Non-parametric missing value imputation for mixed-type data” by Stekhoven, D. J (2012) in the use of new theory in random forest algorithm, namely missForest in the capability of imputing missing values (17). Wang, Z. X (2017) had been demonstrated the same approach in prognostic nomograms of patients with gastric cancer based on metastatic lymph nodes (MLN), negative lymph nodes (LNR), and log odds of metastatic lymph nodes (LODDS) where used to forecast the 5-year survival in patients. With 15,320 samples of patients and the proportion of missing data from 0.3–34.2%, the result exhibited using Surveillance, Epidemiology, and End Results (SEER) database is comparable performances as with using the benchmark Chinese dataset (45). While a research by Xia, J (2017), improves the standard random forest method by proposing a novel random forest algorithm, called adjusted weight voting random forest (AWVRF) with modified surrogate splits that can address incomplete data without imputation. The experimental results show that the AWVRF algorithm can handle the classification task for incomplete data successfully. However, the method is successful applied in binary classification problems only (46). A study by Bathaeian, N.S. (2018) recommended to use random forest for classification and regression task because of the overall result indicates the best performance compared to MICE, KNN, tree and expectation maximization (24). Moreover, a research by Cevallos Valdiviezo, H. (2015) support the use of conditional random forest (CondRF) combine with multiple imputation by MICE for large proportion of missing data in missing not at random (MNAR) mechanism (47). While in the study by Sahri Z. (2014) cites Stekhoven, D. J. (2012) on the idea of using the normalized root mean squared error (NRMSE) as an evaluation accuracy, but he applied k-nearest neighbour in imputing the missing values (48).

In contrast to the study by Stekhoven, D. J (2012), Huang, J. (2017) said the KNN method is simpler and also free from parametric assumption. He proposed a novel incomplete-instance based KNN imputation technique, using cross-validation method, which could optimize the parameters setting of missing value, called (CVB*k*NN). This method outperforms other competing approaches such as mean imputation and other standard KNN imputation in overall imputation accuracy (1). However, Huang, J. (2017) did not perform comparison with missForest algorithm and also not include missing not at random (MNAR) scenario. Similarly, Lobato, F. (2015) argues with Stekhoven, D. J. (2012) as the method do not consider relationship between categorical and numerical variable. Hence, he proposed a novel multi-objective genetic algorithm, called MOGAImp, for data imputation based on the well-known evolutionary algorithm, Non-dominated Sorting Genetic Algorithm-II (NSGA-II) (49). Similar idea as with Huang, J. (2017), other study by Bertsimas, D. (2018) also treat missing values as an optimization problem. He used general first-order methods, named opt.impute and the results performs better than mean impute, KNN, iterative KNN, Bayesian PCA, and predictive-mean matching (50). Interestingly, Garcia-Laencina, P.J (2009) stressed on the important to identify the significance of input attributes to the target class variable before impute missing values. Hence, he proposed weighted distance metric based mutual information (MI), namely MI-KNNimpute, where it considers any relationship among the input variables before estimating missing values. The higher value of MI, indicates the strong relationship of the input attributes to the target class attributes or known as relevant features in classification task (15).

The second discussion is on the most cited author, Zhang, S. (Tab. 8), where he had been published 6 documents as reported in Scopus data base. Among the studied by Zhang, S. (2018) is on the use of mixture-kernel-based estimator for estimating missing values in mixed attribute dataset (36). Also, the use of correlation matrix in KNN, namely (CM-*k*NN), where the distribution in training data is used to obtain the best *k* value for the test data. The fixed constant value of *k* in nearest neighbor, will resulting low prediction rate in real classification application (38). Hence, in a different study, he proposed Sparse learning, called S-KNN, in order to obtain the optimal k value for each test sample (51). He also introducing new algorithm, called Shell Neighbors imputation (SNI), where used only left and right nearest neighbor from the missing instance in order to impute them. The size of the nearest neighbor is based on the cross-validation method (52). More recently, he designed a cost-sensitive method, called Date-drive Incremental imputation Model (DIM), where the top rank missing feature is imputing first based on the scoring rule. Then, the next missing features is imputing with information from the existing complete dataset and new complete dataset (the previous imputed dataset). The proposed model gain benefits in terms of prediction and classification accuracy (53).

6 Conclusion

This research has shown publication trends and other important information of earlier studies on missing data related publication using bibliometric analysis. From 430 publications retrieved from Scopus database covers from 1991 to 2021 as in July 2021, the results indicate a growing trend in this issue, with majority capturing attention from researchers in computer science, mathematics and medical area. Moreover, the top researchers are coming from United States (US) with English is the main medium

language. It should be noted that Twala, B. is the most productive author in this research with 8 publications until now, while Zhang, S. is the most impact author where had received the highest citations (554 citations). The most impact document is “Missforest-Non-parametric missing value imputation for mixed-type data” by Stekhoven, D. J. (2012) with total count of 1069 citations.

Based on the evolutional pathway performed in this study as in Fig. 7, surprisingly reveals two potential techniques in missing data imputation, they are random forest and nearest neighbor search (*k*NN) algorithm. Both methods appear to have the same strength in dealing with missing values include mixed-type attributes, MAR, MCAR, and MNAR missing mechanism, and belong to the same category; non-parametric method. These methods are robust; require no information on data distribution (5). However, previous researchers did not compare both methods in missing data imputation (17). Therefore, it is recommended future research directions to compare these two powerful methods in missing data imputation for evaluating their performances.

This research has several limitations. First, this research includes only journal articles from Scopus database. Other databases like the Web of Sciences (WOS) and other document types (conference paper, book chapter, review article) should considers in the future. The research query can still be improved by considering other keywords with the same meaning as imputing missing values in classification. Despite all these limitations, this study is among the first used bibliometric analysis in analyzing research progress and development trends in the publication related with missing data imputations.

Declarations

Acknowledgements

Not applicable.

Authors’ contributions

FAA conducted the literature search review, analyzed the extracted data obtained from Scopus database and write the first draft of the manuscript. KRJ and WZAWM provided direction for the bibliometrics review and criticize the contents. SM revised the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The papers analyzed in this study are available in Scopus database.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Huang J, Keung JW, Sarro F, Li YF, Yu YT, Chan WK, et al. Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study. J Syst Softw. 2017;132:226–52.
2. Chivers BD, Wallbank J, Cole SJ, Sebek O, Stanley S, Fry M, et al. Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. J Hydrol [Internet]. 2020;588(May):125126. Available from: https://doi.org/10.1016/j.jhydrol.2020.125126.
3. Vergara D, Gaudino R, Blank T, Keegan B. Modeling cannabinoids from a large-scale sample of Cannabis sativa chemotypes. PLoS One [Internet]. 2020;15(9 September):1–17. Available from: http://dx.doi.org/10.1371/journal.pone.0236878.
4. Marcaccio CL, Dumas RP, Huang Y, Yang W, Wang GJ, Holena DN. Delayed endovascular aortic repair is associated with reduced in-hospital mortality in patients with blunt thoracic aortic injury. J Vasc Surg [Internet]. 2018;68(1):64–73. Available from: https://doi.org/10.1016/j.jvs.2017.10.084.
5. Lin WC, Ke SW, Tsai CF. When should we ignore examples with missing values? Int J Data Warehous Min. 2017;13(4):53–63.
6. Elbaz A, Clavel J, Rathouz PJ, Moisan F, Galanaud JP, Delemotte B, et al. Professional exposure to pesticides and Parkinson disease. Ann Neurol. 2009;66(4):494–504.
7. Rubin LH, Witkiewitz K, Andre JS, Reilly S. Methods for handling missing data in the behavioral neurosciences: Don’t throw the baby rat out with the bath water. J Undergrad Neurosci Educ. 2007;5(2):71–7.
8. Sim J, Lee JS, Kwon O. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. Math Probl Eng. 2015;2015.
9. Song Q, Shepperd M, Chen X, Liu J. Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation. J Syst Softw. 2008;81(12):2361–70.
10. White KK, Reiter JP, Petrin A. Imputation in U.S. manufacturing data and its implications for productivity dispersion. Rev Econ Stat. 2018;100(3):502–9.
11. Geeitha S, Thangamani M. Integrating. HSICBFO and FWSMOTE algorithm-prediction through risk factors in cervical cancer. J Ambient Intell Humaniz Comput [Internet]. 2021;12(3):3213–25. Available from: https://doi.org/10.1007/s12652-020-02194-6.

12. Nguyen-Quoc H, Hoang VT. Rice seed image classification based on HOG descriptor with missing values imputation. Telkomnika (Telecommunication Comput Electron Control. 2020;18(4):1897–903.
13. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken Wiley Manly B F J McDonald T L and Amstrup S C. 2002.
14. Liu S, Zhang J, Xiang Y, Zhou W. Fuzzy-Based Information Decomposition for Incomplete and Imbalanced Data Learning. IEEE Trans Fuzzy Syst. 2017 Dec;25(6)(1):1476–90.
15. García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing. 2009;72(7–9):1483–93.
16. Aydilek IB, Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci (Ny) [Internet]. 2013;233:25–35. Available from: http://dx.doi.org/10.1016/j.ins.2013.01.021.
17. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012.
18. Rey-del-Castillo P, Cardeñosa J. Fuzzy min-max neural networks for categorical data: Application to missing data imputation. Neural Comput Appl. 2012;21(6):1349–62.
19. Deb R, Liew AWC. Missing value imputation for the analysis of incomplete traffic accident data. Inf Sci (Ny). 2016;339:274–89.
20. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.
21. Farah Adibah A, Mohd Hafiz Z, Safwati I. 60-year research history of missing data: A bibliometric review on Scopus database (1960-2019). 2020;9(1).
22. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. J Am Stat Assoc. 1991;86(413):68–78.
23. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst. 2012;Vol. 32:77–108 p.
24. Bathaeian NS. Using imputation algorithms when missing values appear in the test data in contrast with the training data. Int J Data Anal Tech Strateg. 2018;10(2):111–23.
25. García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: A review. Neural Comput Appl. 2010.
26. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci Rep [Internet]. 2018;8(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-018-24271-9.
27. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognit. 2008;41(12):3692–705.
28. Dogo EM, Nwulu NI, Twala B, Aigbavboa CO. Empirical Comparison of Approaches for Mitigating Effects of Class Imbalances in Water Quality Anomaly Detection. IEEE Access. 2020;8:218015–36.

29. Twala B. When partly missing data matters in software effort development prediction. J Adv Comput Intell Intell Informatics. 2017.
30. Twala B, Phorah M. Predicting incomplete gene microarray data with the use of supervised learning algorithms. Pattern Recognit Lett [Internet]. 2010;31(13):2061–9. Available from: http://dx.doi.org/10.1016/j.patrec.2010.05.006.
31. Urda D, Subirats JL, García-Laencina PJ, Franco L, Sancho-Gómez JL, Jerez JM. WIMP: Web server tool for missing data imputation. Comput Methods Programs Biomed. 2012.
32. Huang J, Keung JW, Sarro F, Li YF, Yu YT, Chan WK, et al. Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study. J Syst Softw. 2017.
33. Shrive FM, Stuart H, Quan H, Ghali WA. Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. BMC Med Res Methodol. 2006;6:1–10.
34. Phipps AI, Limburg PJ, Baron JA, Burnett-Hartman AN, Weisenberger DJ, Laird PW, et al. Association between molecular subtypes of colorectal cancer and patient survival. Gastroenterology [Internet]. 2015;148(1):77-87.e2. Available from: http://dx.doi.org/10.1053/j.gastro.2014.09.038.
35. Kingsley GH, Kowalczyk A, Taylor H, Ibrahim F, Packham JC, McHugh NJ, et al. A randomized placebo-controlled trial of methotrexate in psoriatic arthritis. Rheumatol (United Kingdom). 2012;51(8):1368–77.
36. Zhu X, Zhang S, Jin Z, Zhang Z, Xu Z. Missing value estimation for mixed-attribute data sets. IEEE Trans Knowl Data Eng. 2011.
37. Saar-Tsechansky M, Provost F. Handling missing values when applying classification models. J Mach Learn Res. 2007;8:1625–57.
38. Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for kNN Classification. ACM Trans Intell Syst Technol. 2017;8(3).
39. Shivaswamy PK, Bhattacharyya C, Smola AJ. Second order cone programming approaches for handling missing and uncertain data. J Mach Learn Res. 2006;7:1283–314.
40. Buse D, Manack A, Serrano D, Reed M, Varon S, Turkel C, et al. Headache impact of chronic and episodic migraine: Results from the American Migraine Prevalence and Prevention Study. Headache. 2012;52(1):3–17.
41. Leu S, Felten S, Von, Frank S, Vassella E, Vajtai I, Taylor E, et al. DH/MGMT-driven molecular classification of low-grade glioma is a strong predictor for long-term survival. Neuro Oncol. 2013;15(4):469–79.
42. Liu ZG, Pan Q, Dezert J, Martin A. Adaptive imputation of missing values for incomplete pattern classification. Pattern Recognit [Internet]. 2016;52:85–95. Available from: http://dx.doi.org/10.1016/j.patcog.2015.10.001.
43. Paleologo G, Elisseeff A, Antonini G. Subagging for credit scoring models. Eur J Oper Res [Internet]. 2010;201(2):490–9. Available from: http://dx.doi.org/10.1016/j.ejor.2009.03.008.

44. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genom. 2014;15(1):1–10.
45. Wang ZX, Qiu MZ, Jiang YM, Zhou ZW, Li GX, Xu RH. Comparison of prognostic nomograms based on different nodal staging systems in patients with resected gastric cancer. J Cancer. 2017;8(5):950–8.
46. Xia J, Zhang S, Cai G, Li L, Pan Q, Yan J, et al. Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recognit. 2017;69:52–60.
47. Cevallos Valdiviezo H, Van Aelst S. Tree-based prediction on incomplete data using imputation or surrogate decisions. Inf Sci (Ny). 2015;311:163–81.
48. Sahri Z, Yusof R, Watada J. FINNIM: Iterative imputation of missing values in dissolved gas analysis dataset. IEEE Trans Ind Informatics. 2014;10(4):2093–102.
49. Lobato F, Sales C, Araujo I, Tadaiesky V, Dias L, Ramos L, et al. Multi-objective genetic algorithm for missing data imputation. Pattern Recognit Lett [Internet]. 2015;68:126–31. Available from: http://dx.doi.org/10.1016/j.patrec.2015.08.023.
50. Bertsimas D, Pawlowski C, Zhuo YD. From Predictive Methods to Missing Data Imputation: An Optimization Approach [Internet]. Vol. 18, Journal of Machine Learning Research. 2018. Available from: http://jmlr.org/papers/v18/17-073.html.
51. Zhang S, Cheng D, Deng Z, Zong M, Deng X. A novel kNN algorithm with data-driven k parameter computation. Pattern Recognit Lett [Internet]. 2018;109:44–54. Available from: https://doi.org/10.1016/j.patrec.2017.09.036.
52. Zhang S. Shell-neighbor method and its application in missing data imputation. Appl Intell. 2011;35(1):123–33.
53. Zhu X, Yang J, Zhang C, Zhang S. Efficient Utilization of Missing Data in Cost-Sensitive Learning. IEEE Trans Knowl Data Eng. 2021;33(6):2425–36.

Figures

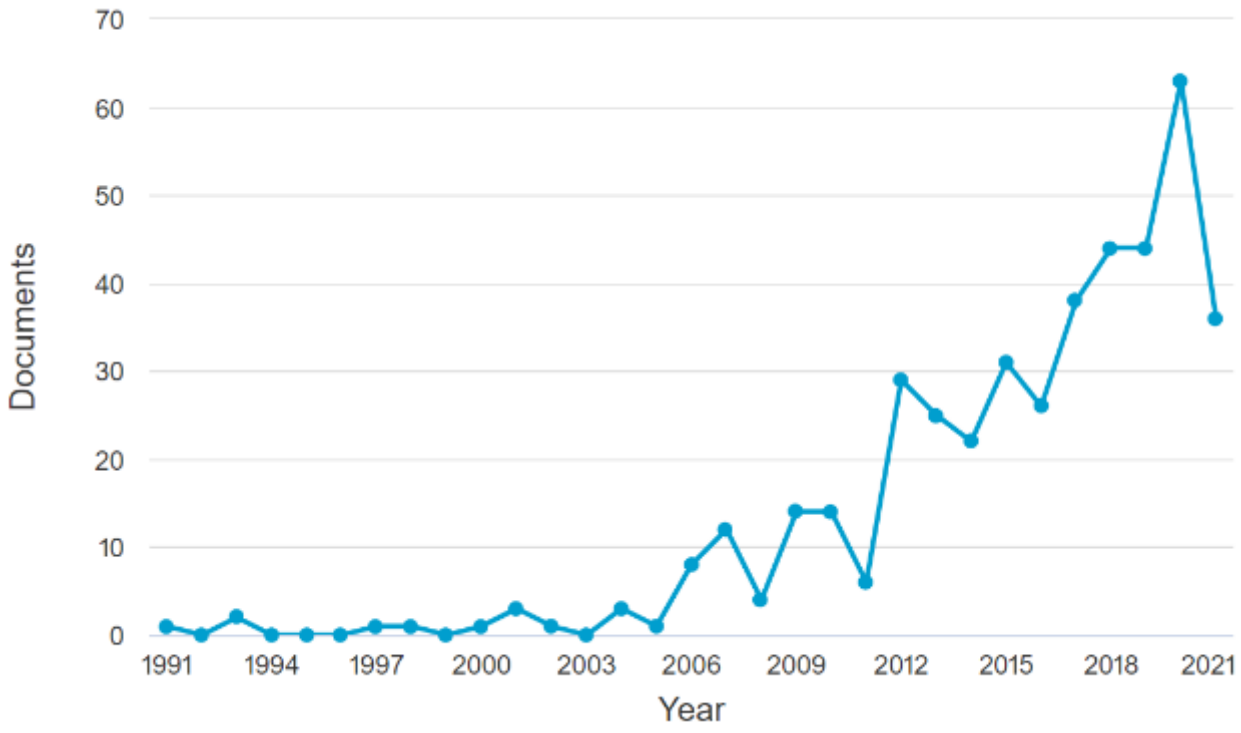

Figure 1

Documents by year

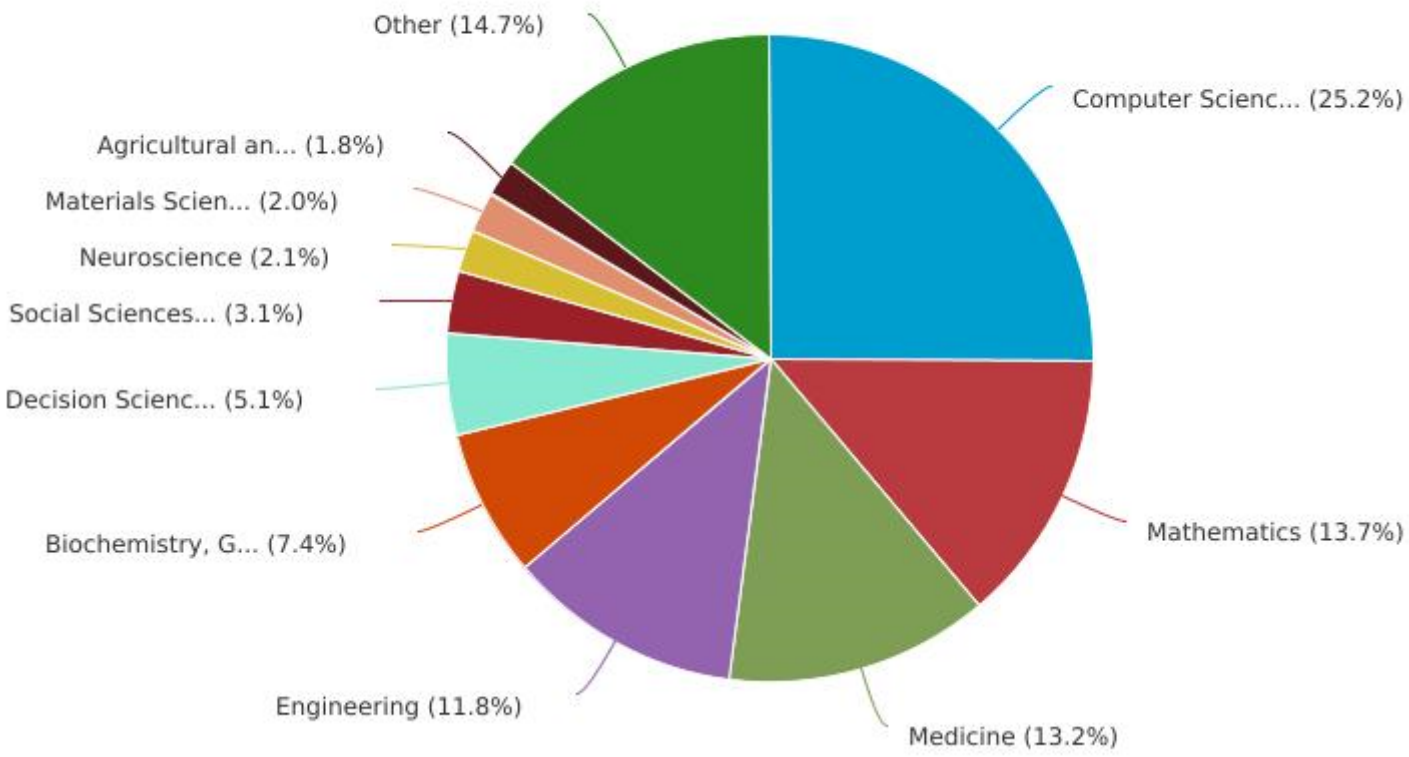

Figure 2

Documents classified by subject area

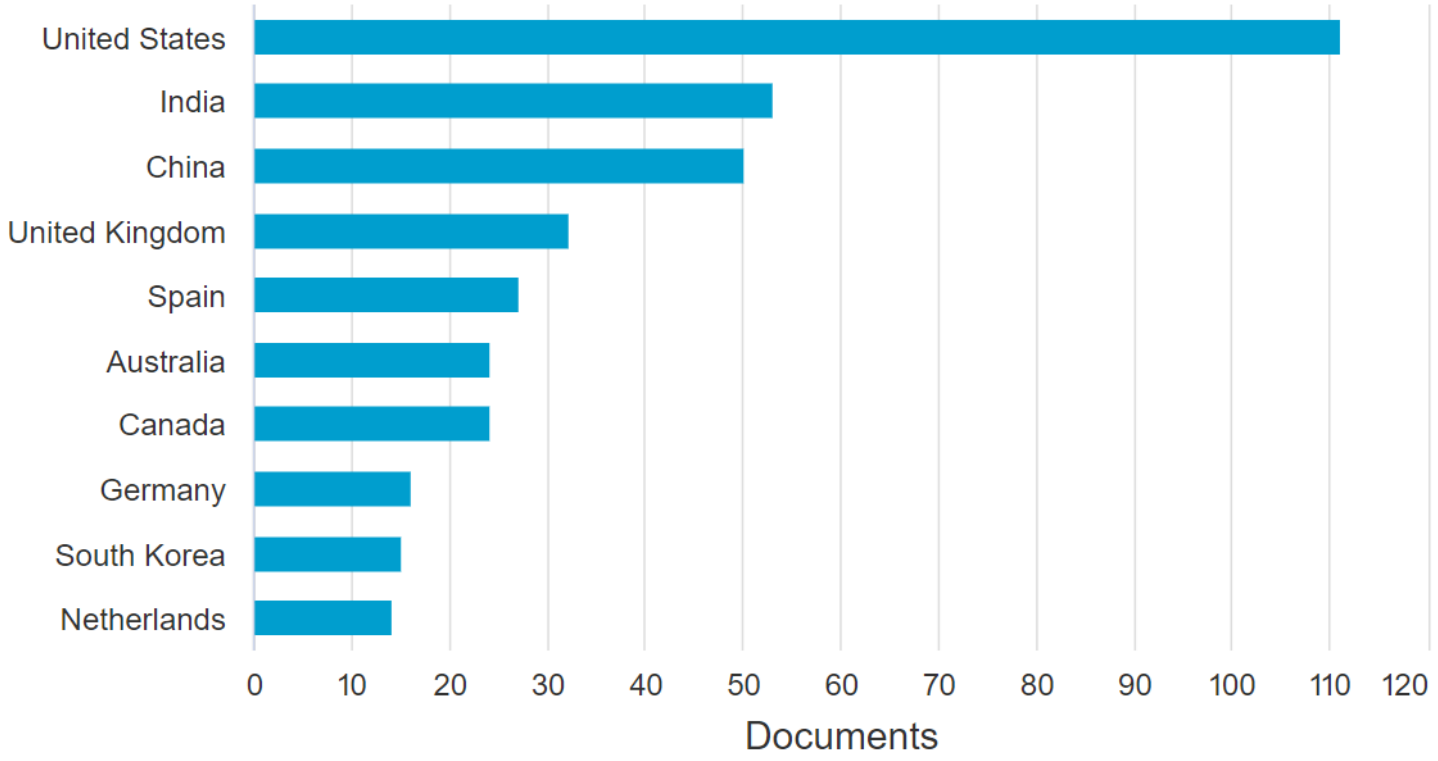

Figure 3

The top 10 country distribution in publications

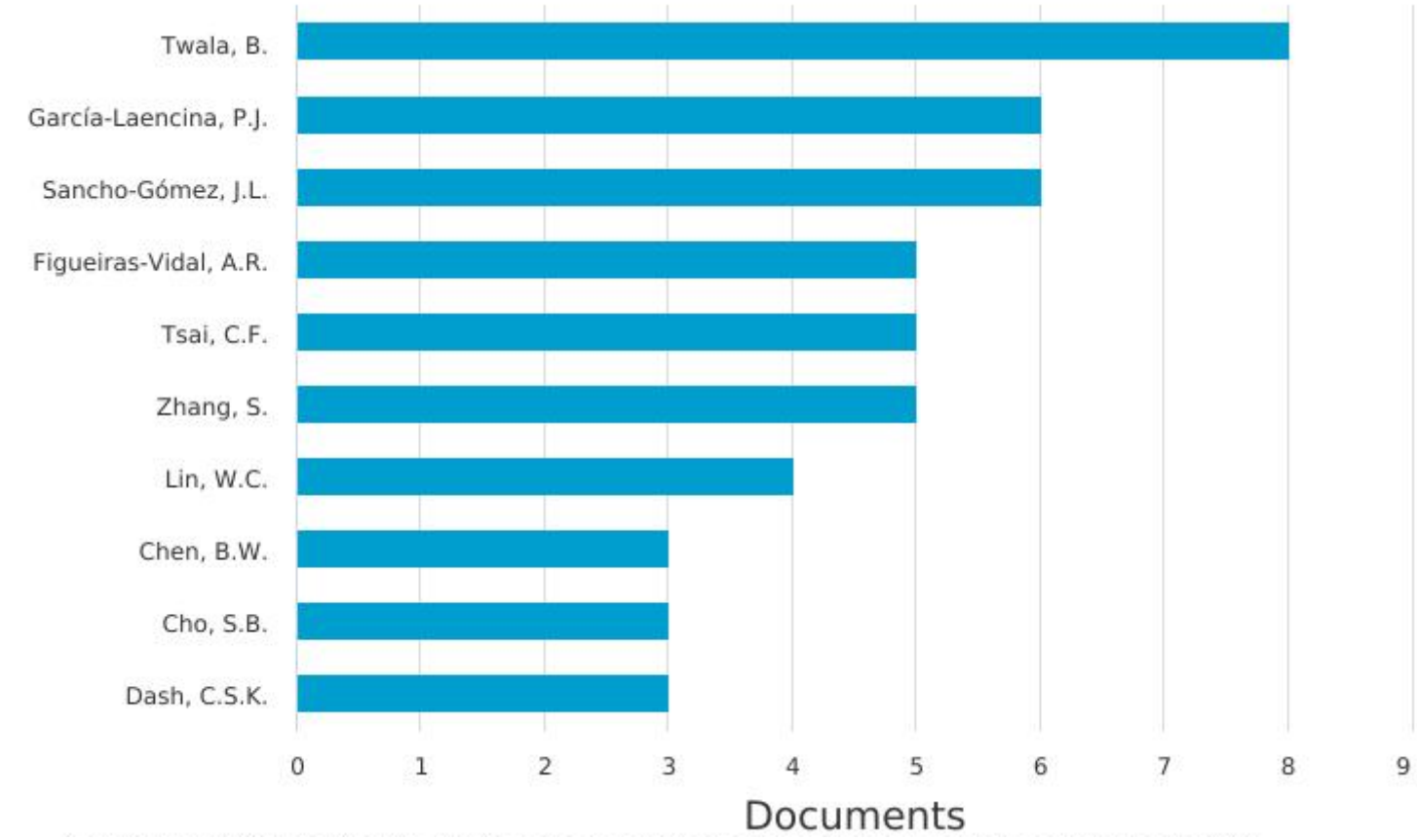

Figure 4

The top 10 authors in publications

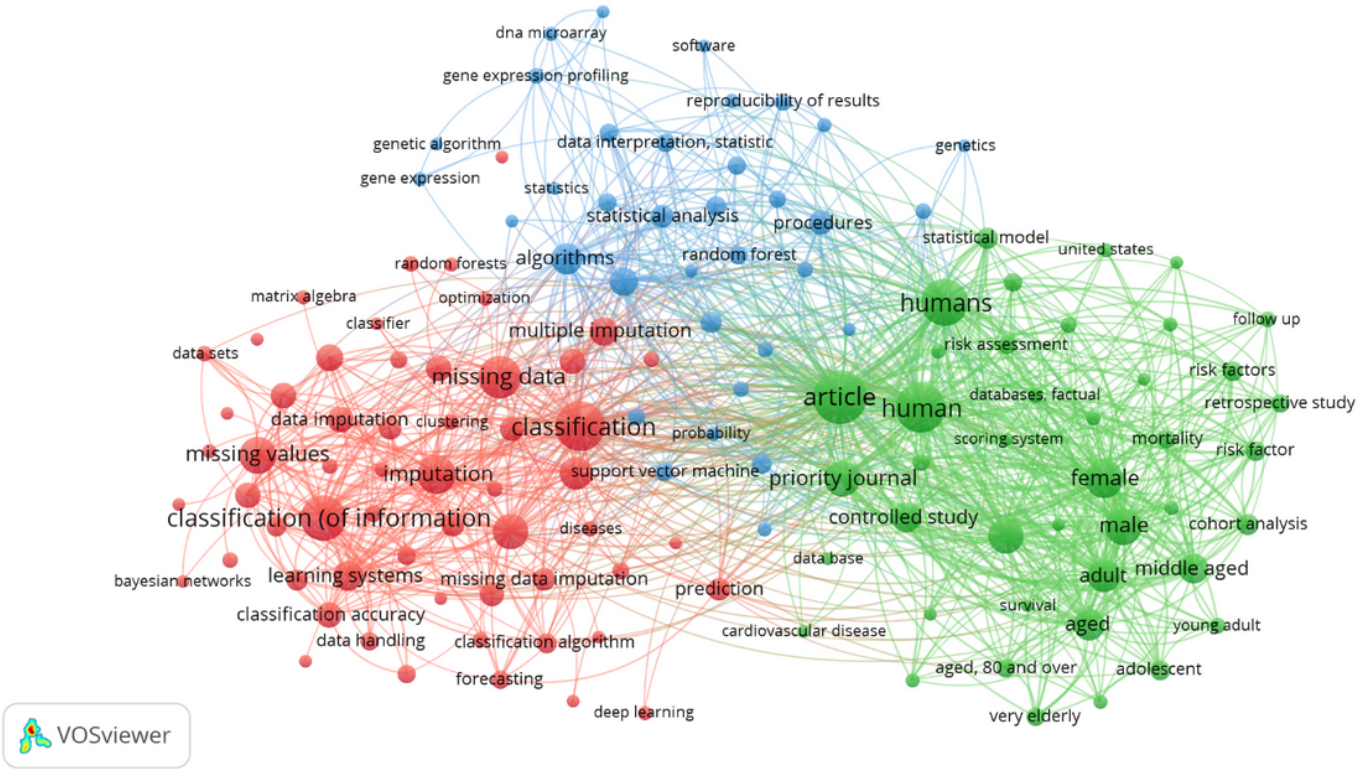

Figure 5

Network visualization of keywords

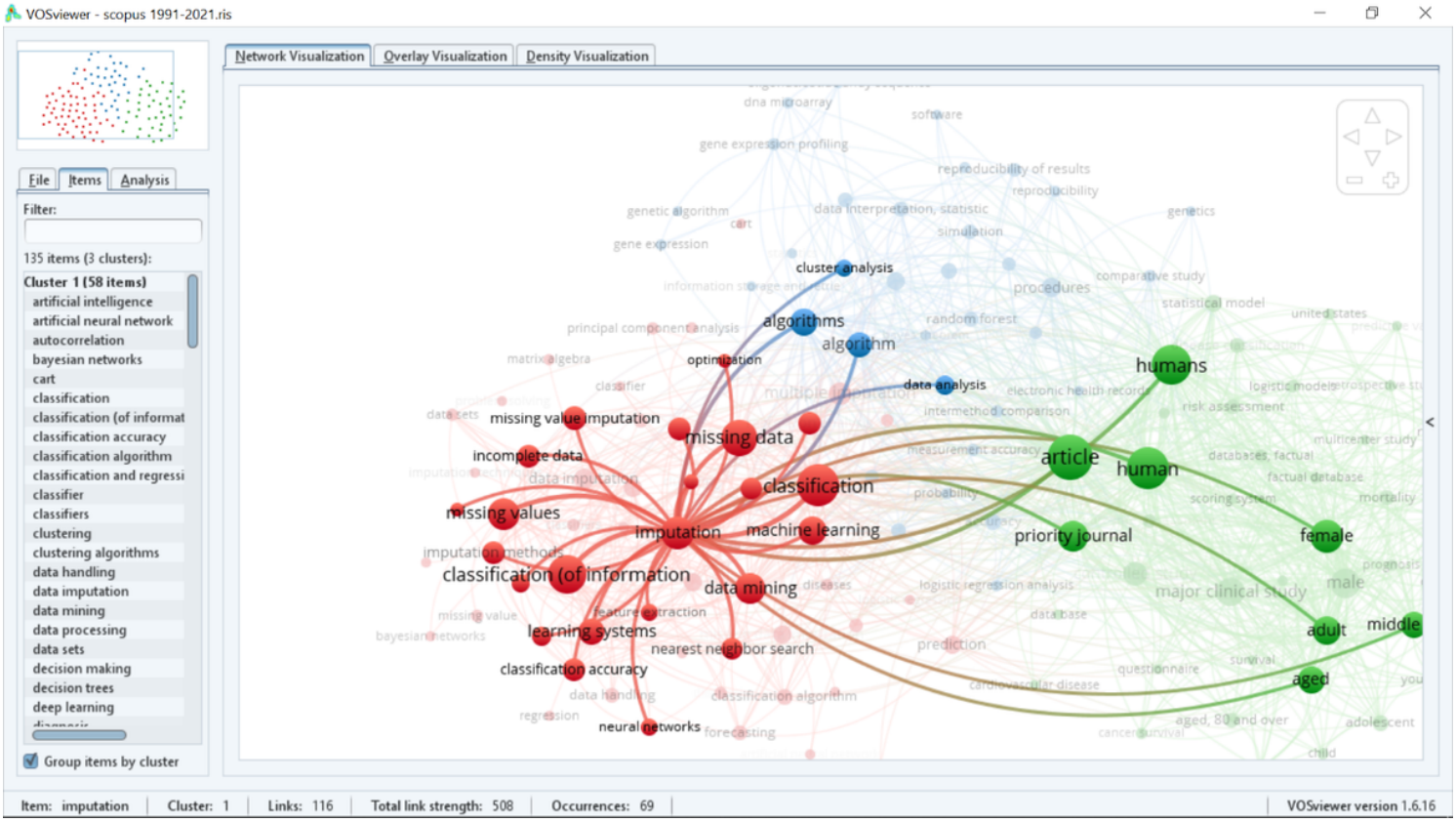

Figure 6

Network visualization of imputation keyword (zoom in)

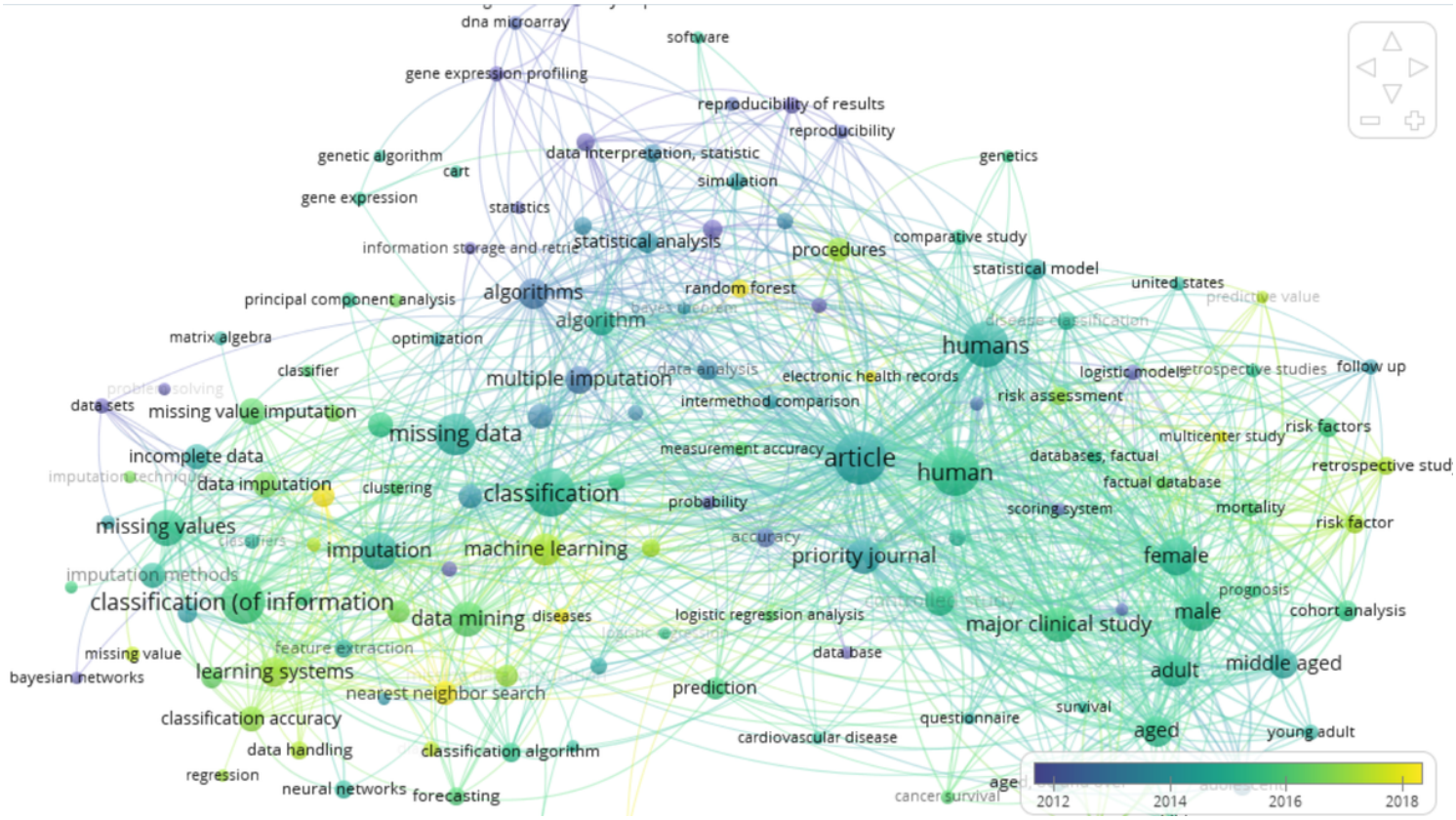

Figure 7

Overlay visualization of keywords

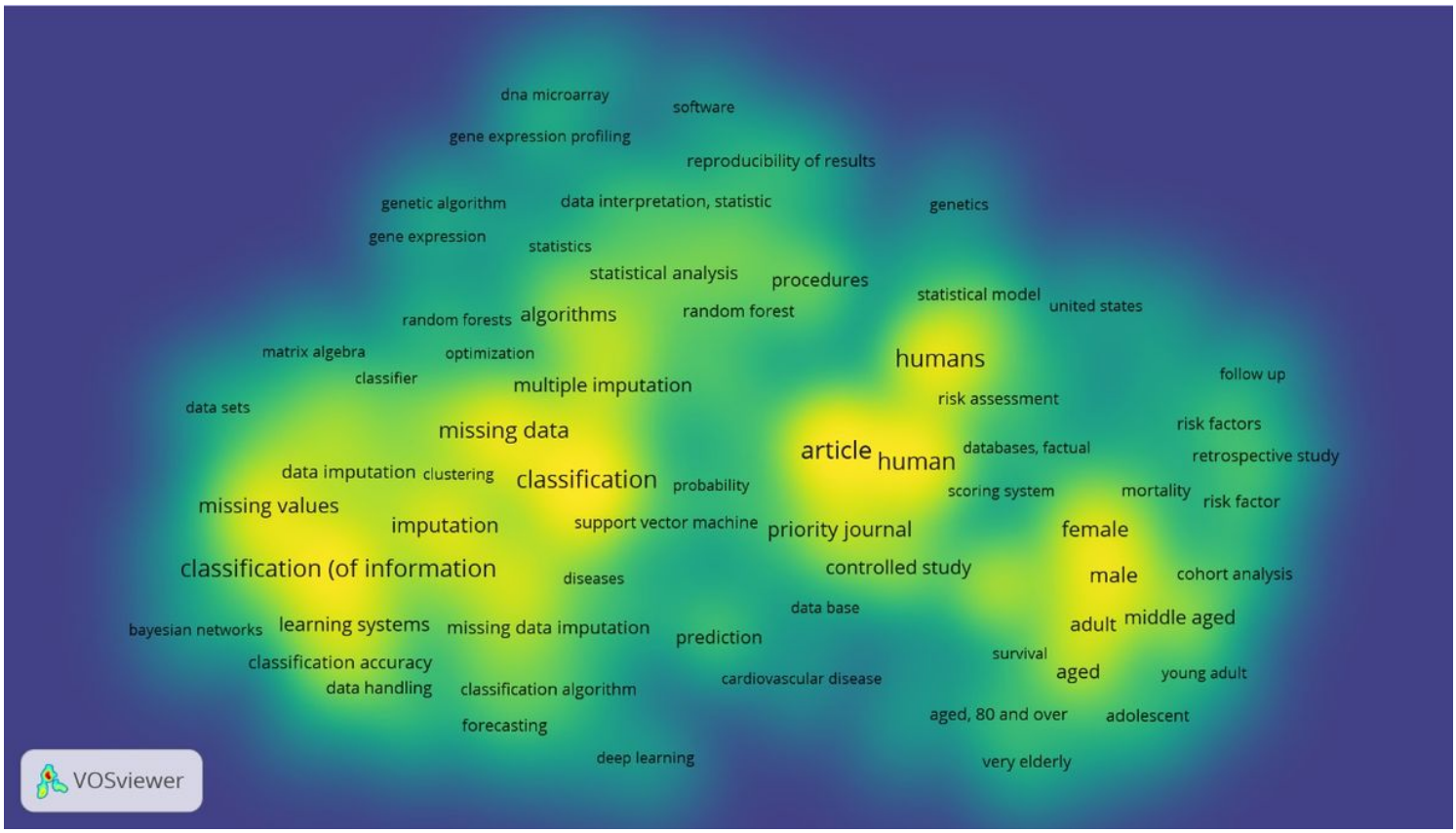

Figure 8

Density visualization of keywords

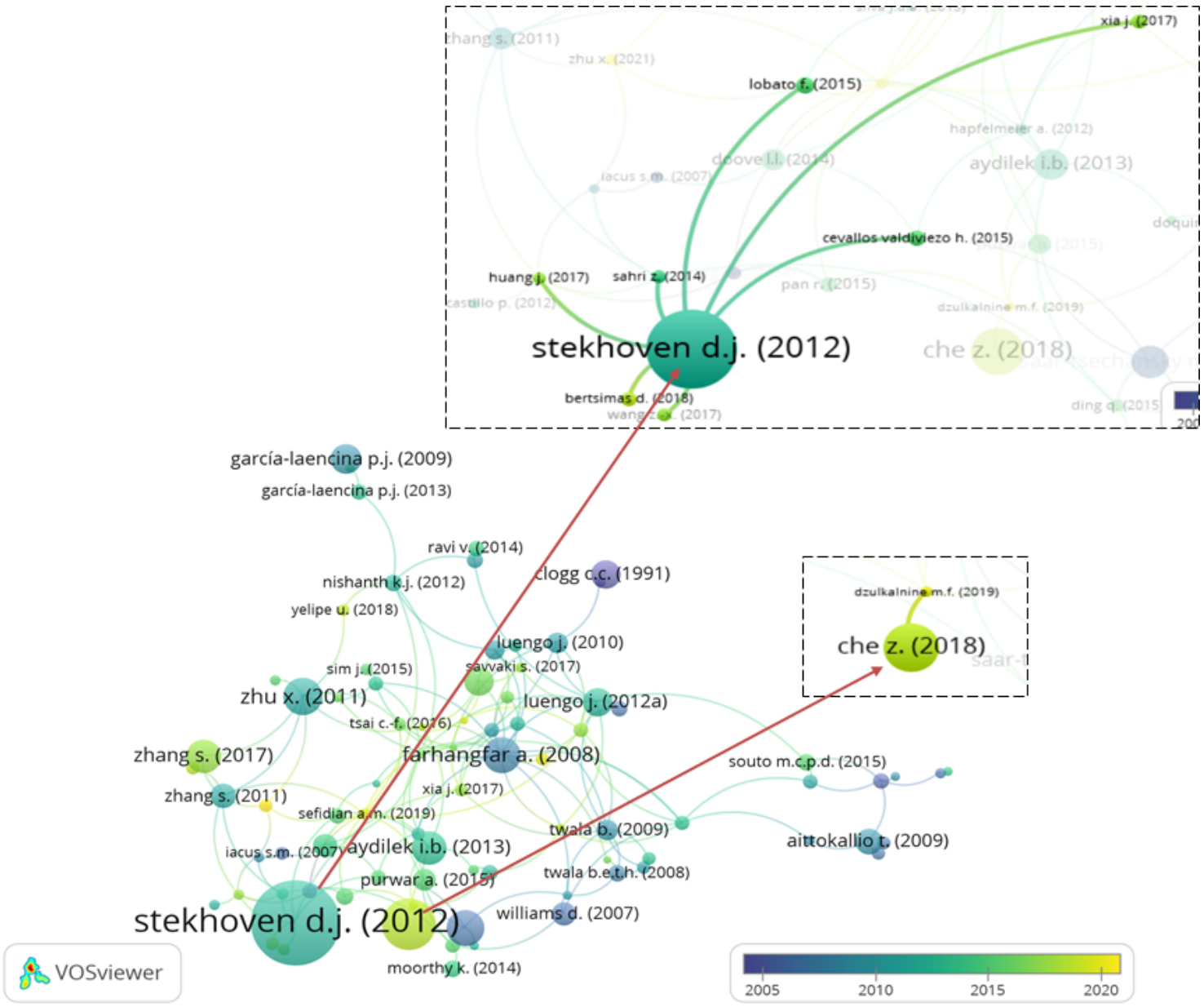

Figure 9

Overlay visualization of citation analysis by documents

Figure 10

Network visualization of citation analysis by authors

Figure 11

Overlay visualization of citation analysis by authors